Optimization-based Quantification of Simulation Input Uncertainty via Empirical Likelihood

Henry Lam  
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027,  
khl2114@columbia.edu  
Huajie Qian  
Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, hqian@umich.edu

We study an optimization-based approach to construct statistically accurate confidence intervals for simulation performance measures under nonparametric input uncertainty. This approach computes confidence bounds from simulation runs driven by probability weights defined on the data, which are obtained from solving optimization problems under suitably posited averaged divergence constraints. We illustrate how this approach offers benefits in computational efficiency and finite-sample performance compared to the bootstrap and the delta method. While resembling robust optimization, we explain the procedural design and develop tight statistical guarantees of this approach via a multi-model generalization of the empirical likelihood method.

Key words: simulation input uncertainty, empirical likelihood, robust optimization

1. Introduction

Stochastic simulation relies on the propagation of the input variates, through the simulation logic, to generate outputs for decision-making; see, e.g., Banks et al. (2005) for an array of applications. Given that in practice the models that govern the input variates are often not fully known but only observed from limited data, the generated simulation outputs can be subject to input errors or uncertainty that adversely affects the decision. Handling this important source of errors has long been advocated and has gathered a fast growth of studies in recent years (see, e.g., the surveys Barton (2012), Henderson (2003), Chick (2006), Song et al. (2014) and Lam (2016a)).

In this paper, we consider the fundamental task of constructing confidence intervals (CIs) for simulation outputs that account for the input uncertainty, in addition to the noises in generating the random variates in the simulation process (known commonly as the stochastic or simulation uncertainty). We focus particularly on the nonparametric regime that makes no assumption on the specific parametric form of the input models. A common approach is the bootstrap (e.g., Barton and Schruben (1993, 2001)), which repeatedly generates resampled distributions to drive simulation runs and uses the quantiles of the simulated outputs to construct the CIs. Another approach is the delta method (e.g., Asmussen and Glynn (2007), Chapter III) that estimates the
asymptotic variance in the central limit theorem (CLT) directly. The latter has been considered mostly in the parametric setting (e.g., Cheng and Holland (1997, 1998, 2004)) but bears a straightforward analog in our considered nonparametric scenario (as we will illustrate later). Estimating this variance can also be conducted by bootstrapping (e.g., Cheng and Holland (1997), Song and Nelson (2015)).

Our focus in this paper is a new approach to construct input-induced CIs by using optimization as an underpinning tool. Our approach looks for a set of “maximal” and a set of “minimal” probability weights on the input data, obtained by solving a pair of convex optimization problems with constraints involving a suitably averaged statistical divergence. These weights can be viewed as “worst-case” representations of the input distributions which are then used to generate the input variates to drive the simulation, giving rise to upper and lower bounds that together form a CI on the performance measure of interest.

We will illustrate how this optimization-based approach offers benefits relative to the bootstrap and the delta method. Though very flexible, the bootstrap typically involves nested simulation due to the resampling step before simulation runs, which leads to a multiplicative computational requirement that can be substantial. At the same time, the performance of the (percentile) bootstrap can also be sensitive to the simulation budget allocation in the nested procedure, for which there is no known rigorous guidance as far as we know. A key element of our approach is to use optimization to replace the resampling step, thus offering a lighter and less sensitive computational requirement. However, akin to the delta method, our approach needs approximating gradient information. While our approach and the delta method have the same asymptotic behavior, we will demonstrate that ours tends to have more accurate finite-sample performance and curb the under-coverage issues often encountered in the delta method, with only a small computational overhead.

As our main technical contributions, we design and analyze procedures to achieve tight statistical coverage guarantees for the resulting optimization-based CIs. Our approach aligns with the recent surge of robust optimization (Ben-Tal and Nemirovski (2002), Bertsimas et al. (2011)) in handling decision-making under uncertainty, where decisions are chosen to perform well under the worst-case scenario among a so-called uncertainty or ambiguity set of possibilities. Our approach particularly resembles distributionally robust optimization (DRO) (e.g., Ben-Tal et al. (2013), Delage and Ye (2010), Goh and Sim (2010), Wiesemann et al. (2014)) where the uncertainty of the considered problem lies in the probability distributions, as our involved optimization formulation contains decision variables that are probability weights of the input distributions. However, contrary to the DRO rationale that postulates the uncertainty sets to contain the truth (including those studied recently in the simulation literature; Glasserman and Xu (2014), Lam (2016c).
Lam and Qian: Optimization-based Input Uncertainty via Empirical Likelihood

we will explain our procedures by viewing the constraints as log-likelihoods on the input data, and develop the resulting statistical guarantees from a multi-model generalization of the empirical likelihood (EL) method (Owen (2001)), a nonparametric analog of the celebrated maximum likelihood method in parametric statistics. Consequently, the form of our proposed constraint (i.e., the averaged statistical divergence constraint) differs drastically from previous DRO suggestions, and the guarantee is provably tight asymptotically. We mention that, though EL has appeared in statistics for a long time, its use in operations research has appeared only recently and is limited to optimization problems (e.g., Lam and Zhou (2017), Duchi et al. (2016), Lam (2016b), Blanchet and Kang (2016), Blanchet et al. (2016)). We therefore contribute by showing that a judicious use of this idea can offer new benefits in the equally important area of simulation analysis.

The rest of this paper is as follows. Section 2 reviews some related literature. Section 3 presents our procedure and main results on statistical guarantees. Section 4 explains the underlying theory giving rise to the approach and our statistical results. Section 5 discusses some strengths and weaknesses of our approach relative to the bootstrap. Section 6 shows some numerical results and compares them with previous approaches. The Appendix contains technical proofs.

2. Related Literature

We briefly survey three areas of related work, one on the problem domain and two on methodologies. The input uncertainty problem in simulation aims to compute CIs or closely related output variance decompositions. In the parametric case, Cheng and Holland (1997) studies both the delta method and the basic bootstrap for computing the variance due to the input noise. Cheng and Holland (1998) and Cheng and Holland (2004) study the so-called two-point method that reduces the total number of simulation runs in estimating the gradient, or the sensitivity coefficients, in applying the delta method. Under the Bayesian framework, Zouaoui and Wilson (2003) studies the variance decomposition and sampling of posterior output distribution. Barton et al. (2013), Xie et al. (2014, 2016) further study the construction of CIs built on Gaussian process metamodels. Beyond parametric uncertainty, Chick (2001) and Zouaoui and Wilson (2004) study Bayesian model averaging (BMA) under the choice of several candidate input parametric models. In the nonparametric regime (our focus in this paper), Barton and Schruben (1993, 2001) propose direct resampling (similar to sectioning; Asmussen and Glynn (2007), Chapter III), bootstrap resampling and the Bayesian bootstrap to construct CIs, where they use a single simulation run per bootstrap resample motivated from the overwhelming input noise in their problem setting. Song and Nelson (2015) studies a mean-variance model to capture the effect of input uncertainty and uses the bootstrap to approximate the input variance component. Finally, some recent work utilizes a risk perspective with
respect to model or distributional uncertainty (e.g., Glasserman and Xu (2014), Zhu and Zhou (2015), Lam (2016c, 2017)).

Our methodologies are related to several tools in statistics. First is the EL method. Initially proposed by Owen (1988) as a nonparametric counterpart of the maximum likelihood theory, the EL method has been widely studied in statistical problems like regression and hypothesis testing etc. (e.g., Owen (2001), Qin and Lawless (1994), Hjort et al. (2009)). Its use in operations research is relatively recent and is limited to optimization. Lam and Zhou (2017) investigates the use of EL in quantifying uncertainty in sample average approximation. Lam (2016b) uses EL to derive uncertainty sets for DRO that guarantees feasibility for stochastic constraints. Duchi et al. (2016) generalizes the EL method to Hadamard differentiable functions and obtains tight optimality bounds for stochastic optimization problems. Blanchet and Kang (2016), Blanchet et al. (2016) generalize the EL method to inference using the Wasserstein distance. In addition, our work also utilizes the influence function, which captures nonparametric sensitivity information of a statistic, and is first proposed by Hampel (1974) in the context of robust statistics (Hampel et al. (2011), Huber and Ronchetti (2009)) as a heuristic tool to measure the effect of data contamination. Influence function is also used in deriving asymptotic results for von Mises differentiable functionals which have profound applications in U-statistics (Serfling (2009)).

Lastly, our approach resembles DRO, which utilizes worst-case perspectives in stochastic decision-making problems under ambiguous probability distributions. In particular, our optimization posited over the space of input probability distributions has a similar spirit as the search for the worst-case distribution in the inner optimization in DRO. The DRO framework has been applied in various disciplines such as economics (Hansen and Sargent (2008)), finance (Glasserman and Xu (2014, 2013)), stochastic control (Petersen et al. (2006), Iyengar (2005), Nilim and El Ghaoui (2005), Xu and Mannor (2012)), queueing (Jain et al. (2010)) and dynamic pricing (Lim and Shanthikumar (2007)). Among them, constraints in terms of \( \phi \)-divergences, which include the Burg-entropy divergence appearing in our approach, have been considered in, e.g. Ben-Tal et al. (2013), Bayraksan and Love (2015), Jiang and Guan (2012), so are other types of statistical distances such as Renyi divergence (e.g., Dev and Juneja (2012), Blanchet and Murthy (2016b)) and the Wasserstein distance (e.g., Esfahani and Kuhn (2015), Blanchet and Murthy (2016a), Gao and Kleywegt (2016)), and other constraint types including moments and support (e.g., Delage and Ye (2010), Goh and Sim (2010), Wiesemann et al. (2014), Hu et al. (2012)). In simulation, the DRO idea has appeared in Glasserman and Xu (2014), Lam (2016c, 2017), Ghosh and Lam (2016) in quantifying model risks. Nonetheless, although our involved optimization looks similar to DRO, the underpinning statistical guarantees of our approach stem from the EL method. As we will explain, our constraints possess properties that are dramatically different from those studied in DRO, and their precise forms also deviate from any known DRO suggestions.
3. Optimization-based Confidence Intervals

This section presents our main procedure and the statistical guarantees. We start with our problem setting and some notations.

3.1. Problem Setting

We consider a performance measure in the form

\[ Z(P_1, \ldots, P_m) = \mathbb{E}_{P_1, \ldots, P_m}[h(X_1, \ldots, X_m)], \]

where each \( X_i = (X_i(1), \ldots, X_i(T_i)) \) is a sequence of \( T_i \) i.i.d. random variables under an independent input model (i.e., distribution) \( P_i \), and \( T_i \) is a deterministic run length. The function \( h \) mapping from \( \mathbb{X}_1^{T_1} \times \cdots \times \mathbb{X}_m^{T_m} \) to \( \mathbb{R} \) is assumed computable given the inputs \( X_i \)'s. In other words, given the sequence \( X_1, \ldots, X_m \), the value of \( h(X_1, \ldots, X_m) \) can be evaluated by the computer. The expectation \( \mathbb{E}_{P_1, \ldots, P_m}[\cdot] \) is taken over the independent i.i.d. sequences \( X_1, \ldots, X_m \), i.e., under the distribution \( P_1^{T_1} \times \cdots \times P_m^{T_m} \). We use \( X_i \) to denote a generic random variable distributed under \( P_i \).

As a simple example, \( X_1 \) and \( X_2 \) can represent respectively the sequences of inter-arrival times and service times in a queueing system. \( P_1 \) and \( P_2 \) represent the corresponding input distributions. \( h \) denotes the indicator function of the exceedance of some waiting time above a threshold. Then \( Z(P_1, P_2) \) becomes the waiting time tail probability.

Our premise is that there exists a true \( P_i \) that is unknown for each \( i \), but \( n_i \) i.i.d. data \( X_{i,1}, \ldots, X_{i,n_i} \) are available. The true value of (1) is therefore unknown even under abundant simulation runs. Our goal is to find a statistically accurate \((1 - \alpha)\) CI for the true performance measure, i.e., two numbers \( \mathcal{L}_\alpha \) and \( \mathcal{U}_\alpha \), derived from the data, such that

\[ P(\mathcal{L}_\alpha \leq Z(P_1, \ldots, P_m) \leq \mathcal{U}_\alpha) \approx 1 - \alpha \]

as the data size grows, where \( P \) refers to the probability under the distribution of the data.

3.2. Main Procedure

Algorithm 1 gives a full step-by-step description of our basic procedure for computing \( \mathcal{L}_\alpha \) and \( \mathcal{U}_\alpha \). The quantity \( \hat{G}_{i,j} \) introduced in Step 1 is the sample estimate of the influence function of \( Z \), which can be viewed as the gradient of \( Z \) taken with respect to the input distributions (see Assumption 1 and the subsequent discussion). Step 2 in Algorithm 1 outputs a minimizer and a maximizer of the optimization problems (2) in which “min / max” denotes a pair of minimization and maximization and the calibrating constant \( \chi^2_{1,1-\alpha} \) is the \( 1 - \alpha \) quantile of the chi-square distribution with degree of freedom one. Optimization (2) is a sample average approximation (SAA) (Shapiro et al. (2014)) on the influence function (which can be expressed as an expectation), with decision variables being
Algorithm 1 Basic Empirical-Likelihood-Based Procedure (BEL)

**Input:** Given data \(\{X_{i,1}, \ldots, X_{i,n_i}\}\) for each input model \(i = 1, \ldots, m\). Choose a target confidence level \(1 - \alpha\) and two positive integers \(R_1, R_2\).

**Procedure:**

1. **Influence Function Estimation:** Compute
   \[
   \hat{G}_{i,j} = \frac{1}{R_1} \sum_{r=1}^{R_1} \left[ (h(X^r_{i,1}, \ldots, X^r_{m}) - \hat{Z})(n_i \sum_{t=1}^{T_i} 1\{X^r_t(t) = X_{i,j}\} - T_i) \right], \quad \forall i, j
   \]
   where for each \(r = 1, \ldots, R_1\), \(X^r_i = (X^r_{i,1}, \ldots, X^r_{T_i})\) are i.i.d. variates drawn independently from the uniform distribution on \(\{X_{i,1}, \ldots, X_{i,n_i}\}\) for each \(i\), \(1\{\cdot\}\) is the indicator function, and \(\hat{Z} = \sum_{r=1}^{R_1} h(X^r_{i,1}, \ldots, X^r_{m}) / R_1\).

2. **Optimization:** Compute respective optimal solutions \((w^\text{min}_1, \ldots, w^\text{min}_m)\) and \((w^\text{max}_1, \ldots, w^\text{max}_m)\) of the following pair of programs
   \[
   \begin{align*}
   \min / \max & \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{G}_{i,j} w_{i,j} \\
   \text{subject to} & \quad -2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log n_i w_{i,j} \leq X^2_{1,1-\alpha} \\
   & \quad \sum_{j=1}^{n_i} w_{i,j} = 1, \text{ for all } i \\
   & \quad w_{i,j} \geq 0, \text{ for all } i, j
   \end{align*}
   \]

3. **Evaluation:** Compute
   \[
   \mathcal{L}^{\text{BEL}}_\alpha = \frac{1}{R_2} \sum_{r=1}^{R_2} h(X^\text{min}_1, \ldots, X^\text{min}_m), \quad \mathcal{U}^{\text{BEL}}_\alpha = \frac{1}{R_2} \sum_{r=1}^{R_2} h(X^\text{max}_1, \ldots, X^\text{max}_m)
   \]
   where for each \(r = 1, \ldots, R_2\), \(X^\text{min}_i = (X^\text{min}_{i,1}, \ldots, X^\text{min}_{i,T_i})\) and \(X^\text{max}_i = (X^\text{max}_{i,1}, \ldots, X^\text{max}_{i,T_i})\) are i.i.d. variates drawn independently according to weights \(w^\text{min}_{i,j}\) and \(w^\text{max}_{i,j}\), \(j = 1, \ldots, n_i\), respectively from \(\{X_{i,1}, \ldots, X_{i,n_i}\}\) for each \(i\).

**Output:** The CI \([\mathcal{L}^{\text{BEL}}_\alpha, \mathcal{U}^{\text{BEL}}_\alpha]\).

The probability weights \(w_i = (w_{i,j})_{j=1,\ldots,n_i}\) on the influence function evaluated at each data point of an input model.

Formulation (2) can be interpreted as two worst-case optimization problems over \(m\) independent input distributions, each on support \(\{X_{i,1}, \ldots, X_{i,n_i}\}\), subject to a weighted average of individual statistical divergences (Pardo (2005)). To explain, the quantity \(D_{n_i}(w_i) = -(1/n_i) \sum_{j=1}^{n_i} \log n_i w_{i,j}\) is the Burg-entropy divergence (Ben-Tal et al. (2013)) (or the Kullback-Leibler (KL) divergence).
between the probability weights \( w_i \) and the uniform weights. Thus, letting \( n = \sum_{i=1}^{m} n_i \) be the total sample size, we have

\[
-\frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log n_i w_{i,j} = \frac{1}{n} \sum_{i=1}^{m} n_i \left( -\frac{1}{n} \sum_{j=1}^{n_i} \log n_i w_{i,j} \right) = \frac{1}{n} \sum_{i=1}^{m} n_i D_{n_i}(w_i)
\]

which is an average of the Burg-entropy divergences imposed on different input models, each weighted by the proportion of the respective input data size over the total data size, \( n_i/n \). The first constraint in (2) can thus be written as

\[
\sum_{i=1}^{m} n_i \frac{D_{n_i}(w_i)}{n} \leq \chi_{1,1-\alpha}^2 / (2n)
\]

which constitutes a neighborhood ball of size \( \chi_{1,1-\alpha}^2 / (2n) \) measured by the averaged Burg-entropy divergence.

Finally, Step 3 in Algorithm 1 stipulates the use of the obtained optimal probability weights \( \{w_{i,j}^{\text{min}}\}_{i,j} \) and \( \{w_{i,j}^{\text{max}}\}_{i,j} \) to drive two independent sets of simulation runs in order to output the lower and upper confidence bounds.

An efficient method for solving the optimization of Step 2 is discussed in the following proposition.

**Proposition 1.** If it holds for program (2) that \( \hat{G}_{i_0,j_1} \neq \hat{G}_{i_0,j_2} \) for some \( 1 \leq i_0 \leq m \) and \( 1 \leq j_1 < j_2 \leq n_{i_0} \), then the minimizer \((w_1^{\text{min}}, \ldots, w_m^{\text{min}})\) of (2) can be obtained by

\[
w_{i,j}^{\text{min}} = \frac{2\beta^*}{\hat{G}_{i,j} + \lambda_i^*},
\]

where \((\beta^*, \lambda_1^*, \ldots, \lambda_m^*)\) is the unique solution, in the domain \(\{(\beta, \lambda_1, \ldots, \lambda_m) : \beta > 0, \lambda_i > -\min_j \hat{G}_{i,j} \text{ for all } i\}\), of the system of equations

\[
2 \sum_{i=1}^{m} \frac{\sum_{j=1}^{n_i} \log 2n_i \beta}{\hat{G}_{i,j} + \lambda_i} + \chi_{1,1-\alpha}^2 = 0, \quad \sum_{j=1}^{n_i} \frac{2\beta}{\hat{G}_{i,j} + \lambda_i} = 1 \text{ for all } i.
\]

The maximizer \((w_1^{\text{max}}, \ldots, w_m^{\text{max}})\) can be computed through the same procedure except that \( \hat{G}_{i,j} \) is replaced by \( -\hat{G}_{i,j} \).

Otherwise, if \( \hat{G}_{i,j} \) for \( j = 1, \ldots, n_i \) are all equal for each \( i \), then the program has a constant objective and hence becomes trivial.

The advantage of solving the system of equations (3) in place of directly dealing with optimization (2) is that the dimension of the decision space is now equal to the number of input models (plus one), which is more favorable in case of large data size. (3) can be efficiently solved by standard root-finding algorithms such as Newton’s method applied on \( \lambda_i \)'s together with bisection applied on \( \beta \).
Next we provide two variants of Algorithm 1 depicted as Algorithms 2 and 3 which differ only by the last step. The motivation (with more details in Section 3.3) is that Algorithm 1 tends to under-cover the true performance value because its last step only outputs the sample mean of the simulation drives and does not take full account of the stochastic uncertainty. Algorithm 2 takes care of this uncertainty by outputting the standard normal lower and upper confidence bounds. However, such an adjustment turns out to be too conservative and leads to over-coverage. This necessitates the refinement in Algorithm 3 that involves an estimate of the input-induced variance, $\hat{\sigma}_I^2$, and leads to accurate coverage performance as we will show.

**Algorithm 2** Evaluation-Adjusted Empirical Likelihood (EEL)

Follow Algorithm 1 until Step 3. Replace Step 3 by

$$L_{aEEL} = \hat{Z}_{\text{min}} - z_{1-\alpha/2} \frac{\hat{\sigma}_{\text{min}}}{\sqrt{R_2}}, \quad U_{aEEL} = \hat{Z}_{\text{max}} + z_{1-\alpha/2} \frac{\hat{\sigma}_{\text{max}}}{\sqrt{R_2}}$$

where

$$\hat{Z}_{\text{min}} = \frac{1}{R_2} \sum_{r=1}^{R_2} h(\mathbf{X}_{r,\text{min}}, \ldots, \mathbf{X}_{r,m}), \quad \hat{\sigma}^2_{\text{min}} = \frac{1}{R_2 - 1} \sum_{r=1}^{R_2} (h(\mathbf{X}_{r,\text{min}}, \ldots, \mathbf{X}_{r,m}) - \hat{Z}_{\text{min}})^2$$

and $\hat{Z}_{\text{max}}, \hat{\sigma}^2_{\text{max}}$ are defined accordingly. $z_{1-\alpha/2}$ is the $1 - \alpha/2$ critical value of the standard normal.

**Output:** The CI $[L_{aEEL}, U_{aEEL}]$.

**Algorithm 3** Fully Adjusted Empirical Likelihood (FEL)

Follow Algorithm 1 until Step 3. Replace Step 3 by

$$L_{aFEL} = \hat{Z}_{\text{min}} - z_{1-\alpha/2} \left( \sqrt{\hat{\sigma}_I^2 + \hat{\sigma}_{\text{min}}^2} - \hat{\sigma}_I \right), \quad U_{aFEL} = \hat{Z}_{\text{max}} + z_{1-\alpha/2} \left( \sqrt{\hat{\sigma}_I^2 + \hat{\sigma}_{\text{max}}^2} - \hat{\sigma}_I \right)$$

where $z_{1-\alpha/2}, \hat{Z}_{\text{min}}, \hat{\sigma}^2_{\text{min}}, \hat{\sigma}_{\text{max}}$ are the same as in Algorithm 2 and

$$\hat{\sigma}_I^2 = \sum_{i=1}^{m} \frac{1}{n_i} \left( \sum_{j=1}^{n_i} \frac{\hat{G}_{i,j}^2}{n_i} - \frac{n_i T_i \hat{\sigma}_I^2}{R_1} \right), \quad \text{with} \quad \hat{\sigma}^2 = \frac{1}{R_1 - 1} \sum_{r=1}^{R_1} (h(\mathbf{X}_r, \ldots, \mathbf{X}_m) - \hat{Z})^2$$

(4)

is computed from the $R_1$ replications generated in Step 1.

**Output:** The CI $[L_{aFEL}, U_{aFEL}]$. 
3.3. Statistical Guarantees

We present the statistical guarantees of Algorithms 1, 2, and 3. We assume the following:

**Assumption 1.** \(0 < \sum_{i=1}^{m} \text{Var}(G_i(X_i)) < +\infty\), where
\[
G_i(x) = \sum_{t=1}^{T_i} \mathbb{E}_{P_1,\ldots,P_m}[h(X_1,\ldots,X_m)|X_i(t) = x] - T_iZ(P_1,\ldots,P_m).
\]

**Assumption 2.** (Parameter \(k\)) For each \(i\) let \(I_i = (I_i(1),\ldots,I_i(T_i))\) be a sequence of indices such that \(1 \leq I_i(t) \leq T_i\), and \(X_i, I_i = (X_i(I_i(1)),\ldots,X_i(I_i(T_i)))\). Assume \(\mathbb{E}_{P_1,\ldots,P_m}[|h(X_{1,i},\ldots,X_{m,i},t)|^k] \) is finite for all such \(I_i\)'s.

The function \(G_i(x)\) in Assumption 1 is the influence function (Hampel, 1974, Hampel et al. (2011)) of the performance measure \(Z(P_1,\ldots,P_m)\) with respect to each input distribution, which measures the infinitesimal effect caused by perturbing \(P_i\) and represents the Gateaux derivative on \(Z\) in the sense
\[
\frac{d}{de} Z(P_1,\ldots,P_i-1, (1-\epsilon)P_i + \epsilon Q_i, P_{i+1},\ldots,P_m)\bigg|_{\epsilon=0^+} = \int G_i(x)dQ_i(x)
\]
for any distribution \(Q_i\) on \(\mathcal{X}\). Assumption 1 entails that the influence functions are non-degenerate at the true input distributions \(P_i\)'s. Note that the \(\hat{G}_{i,j}\) in Step 1 of Algorithm 1 is a sample version of \(G_i(X_{i,j})\).

Here we list notations that will be extensively referred to throughout the paper for ease of exposition. Define

\[
H_i(x,y) = \sum_{1 \leq t \leq T_i} \sum_{1 \leq s \leq T_i, s \neq t} \mathbb{E}_{P_1,\ldots,P_m}[h(X_1,\ldots,X_m)|X_i(t) = x, X_i(s) = y] \text{ for all } i
\]
\[
H_{i,i'}(x,y) = \sum_{1 \leq t \leq T_i} \sum_{1 \leq s \leq T_{i'}} \mathbb{E}_{P_1,\ldots,P_m}[h(X_1,\ldots,X_m)|X_i(t) = x, X_{i'}(s) = y] \text{ for all } i \neq i'.
\]

which can be viewed as the “second order influence functions” analogously as \(G_i(x)\). We further define

\[
M_i = \mathbb{E}(H_i(X_i, \tilde{X}_i) - \mathbb{E}[H_i(X_i, \tilde{X}_i)|X_i] - \mathbb{E}[H_i(X_i, \tilde{X}_i)] + \mathbb{E}[H_i(X_i, X_i)])^2 \text{ for all } i
\]
\[
M_{i,i'} = \mathbb{E}(H_{i,i'}(X_i, X_{i'}) - \mathbb{E}[H_{i,i'}(X_i, X_{i'})|X_i] - \mathbb{E}[H_{i,i'}(X_i, X_{i'})] + \mathbb{E}[H_{i,i'}(X_i, X_{i'})])^2 \text{ for all } i \neq i'
\]
where \(X_i, \tilde{X}_i\) are i.i.d. under \(P_i\) and \(X_i, X_{i'}\) are independent under \(P_i\) and \(P_{i'}\) respectively. Other frequently used notations include the feasible set of (2), the simulation variance, and the total run length:

\[
\mathcal{X}_\alpha = \left\{ (w_1,\ldots,w_m) \in \mathbb{R}^n : \sum_{j=1}^{n_i} \log n_i w_{i,j} \leq \chi^2_{1,1-\alpha} \right\}
\]
\[
\sigma^2 = \text{Var}_{P_1,\ldots,P_m}[h(X_1,\ldots,X_m)], \quad T = \sum_{i=1}^{m} T_i.
\]
For notational simplicity, we stick to $C(\alpha)$ for constants in CI error bounds that depend only on the desired coverage, which can take different values from one theorem to another.

We have the following statistical guarantees for using the three proposed algorithms to construct input-induced CIs:

**Theorem 1.** Suppose Assumptions 1 and 2 hold with $k = 4$, and that $\max_{i \in I} n_i \leq \gamma \min_{i \in I} n_i$ always holds for some constant $\gamma > 0$, where $I = \{i | \text{Var}(G_i(X_i)) > 0\}$. Then the outputs $\mathcal{L}_{\alpha}^{BEL}, \mathcal{Y}_{\alpha}^{BEL}$ of Algorithm 1 satisfy

$$\lim_{n_i \to \infty} P(\mathcal{L}_{\alpha}^{BEL} + E_l^{BEL} \leq Z(P_1, \ldots, P_m) \leq \mathcal{Y}_{\alpha}^{BEL} + E_u^{BEL}) = 1 - \alpha$$

where $E|E_l^{BEL}|, E|E_u^{BEL}| \leq C(\alpha) \left( \sqrt{\sum_{i \neq i'} M_{i,i'} \frac{1}{n_i n_{i'}}} + \sum_i \frac{M_i}{n_i^2} + \sqrt{\frac{2}{R_1}} \right) + \frac{\sqrt{\gamma}}{\sqrt{R_2}}$

**Theorem 2.** In addition to the conditions of Theorem 1, suppose Assumption 2 holds with $k = 8$. Then the outputs $\mathcal{L}_{\alpha}^{EEL}, \mathcal{Y}_{\alpha}^{EEL}$ of Algorithm 2 satisfy

$$\lim_{n_i, R_2 \to \infty} \inf P(\mathcal{L}_{\alpha}^{EEL} + E_l^{EEL} \leq Z(P_1, \ldots, P_m) \leq \mathcal{Y}_{\alpha}^{EEL} + E_u^{EEL}) \geq 1 - \alpha$$

where $E|E_l^{EEL}|, E|E_u^{EEL}| \leq C(\alpha) \left( \sqrt{\sum_{i \neq i'} M_{i,i'} \frac{1}{n_i n_{i'}}} + \sum_i \frac{M_i}{n_i^2} + \sqrt{\frac{2}{R_1}} \right)$

**Theorem 3.** In addition to the conditions of Theorem 2, suppose $\min_{i \in I} n_i = \omega(\sqrt{\min_{i \in I} n_i})$ and $R_1$ is chosen such that

$$R_1 = \omega(\min_{i \in I} n_i^{2/3}), R_1 \min_{i \in I} n_i^2 = \omega(\min_{i \in I} n_i^2).$$

Then the outputs $\mathcal{L}_{\alpha}^{FEL}, \mathcal{Y}_{\alpha}^{FEL}$ of Algorithm 3 satisfy

$$\lim_{n_i, R_2 \to \infty} \inf P(\mathcal{L}_{\alpha}^{FEL} + E_l^{FEL} \leq Z(P_1, \ldots, P_m) \leq \mathcal{Y}_{\alpha}^{FEL} + E_u^{FEL}) \geq 1 - \alpha$$

$$\lim_{n_i, R_2 \to \infty} \sup P(\mathcal{L}_{\alpha}^{FEL} + E_l^{FEL} \leq Z(P_1, \ldots, P_m) \leq \mathcal{Y}_{\alpha}^{FEL} + E_u^{FEL}) \leq 1 - \alpha + \frac{\alpha^2}{4}$$

where $E|E_l^{FEL}|, E|E_u^{FEL}| \leq C(\alpha) \left( \sqrt{\sum_{i \neq i'} M_{i,i'} \frac{1}{n_i n_{i'}}} + \sum_i \frac{M_i}{n_i^2} + \sqrt{\frac{2}{R_1}} \right)$

Assumption 1 is used to handle a linearization of the performance measure that is in turn needed to find the optimal probability weights. Assumption 2 is a moment condition that controls the error of the linearization and the simulation error. They hold trivially if $h$ is a bounded function for instance. The condition $\max_{i \in I} n_i \leq \gamma \min_{i \in I} n_i$ in Theorem 1 basically forces the sample sizes for all input models that have non-degenerate influence functions to grow at the same rate.

Theorem 1 states that Algorithm 1 generates an asymptotically exact CI for the true performance measure $Z(P_1, \ldots, P_m)$, up to errors $E_l^{BEL}$ and $E_u^{BEL}$ for the lower and upper limits. These errors consist of three terms. The first term corresponds to the combination of input errors with
the approximation of the performance measure using its influence function. The second term corresponds to the sample average approximation undertaken in Steps 1 and 2. The third term corresponds to the simulation errors from the final evaluation in Step 3. Theorem 2 states that the last error term goes away in Algorithm 2 due to the refinements of the final bounds, but the resulting CI is conservative and only a lower bound of the asymptotic coverage probability is guaranteed. Fortunately, the refinement in Algorithm 3 can recover the exact coverage up to a tiny error of $\alpha^2/4$ and remain unaffected by the last error term, as shown in Theorem 3. The $\hat{\sigma}_I^2$ in Algorithm 3 is the bias-corrected sample variance of the influence function (i.e., the input-induced variance). Bias correction is desirable here, because the naive sample variance $\sum_{i,j=1}^{n_i} \hat{G}_{i,j}/n_i$ for each input model is upward biased due to the simulation noise in each $\hat{G}_{i,j}$ and the introduction of the bias correction term $n_i T_i \hat{\sigma}_I^2/R_1$ significantly reduces the mean squared error. Thanks to the bias correction, $R_1$ growing sublinearly in data size, i.e., the condition (5) is sufficient to ensure that $\hat{\sigma}_I^2$ consistently estimates the input-induced variance. Note that the second requirement in (6) takes effect only if there is at least one input model that has a degenerate influence function.

The typical length of the CI is $O(1/\sqrt{\min_{i\in I} n_i})$. A negligible overall error in CI construction needs to be of order $o_p(1/\sqrt{\min_{i\in I} n_i})$. To maintain this error level, we need to choose $R_1, R_2 = \omega(\min_{i\in I} n_i)$ in Algorithm 1 and $R_1 = \omega(\min_{i\in I} n_i)$ for Algorithms 2 and 3. We will justify these claims in detail in Corollaries 2, 3 and 4 in the sequel. Since the total simulation effort is $R_1 + 2R_2$, the required total simulation effort is $\omega(\min_{i\in I} n_i)$.

4. Theory on Statistical Guarantees

This section explains Theorems 1, 2 and 3 and discusses some related results. To put our subsequent discussion into perspective, we point out first that despite the presented Burg-entropy divergence interpretation in Section 3.2 that ties the optimal solutions in Step 2 of the proposed algorithms to “worst-case” distributions in the DRO framework, the conceptual reasoning of the constraint in (2) that we discuss below is fundamentally different from DRO. The latter advocates the use of uncertainty sets that contain the true distribution with a certain confidence, so that the obtained optimization output value provides a confidence bound on the true performance measure of interest (e.g., Delage and Ye (2010), Ben-Tal et al. (2013)). Under such a framework, a divergence ball used as an uncertainty set must use a “baseline” distribution that is absolutely continuous to the true distribution, in order to have an overwhelming (or at least non-zero) probability of containing the truth (Jiang and Guan (2012), Esfahani and Kuhn (2015)). This condition is violated in formulation (2) when the true input distribution is continuous. As the baseline distribution in our divergence (namely the empirical distribution) is supported only on the data, the resulting ball does not contain any continuous distributions. Moreover, the use of weighted average and its
particular weights put on each of these empirically defined divergences is also an unnatural choice from a robust optimization perspective.

We will instead explain our procedure using the EL method. We start by looking at this method for a special case of our setting, and then build our theoretical analysis for more general cases.

4.1. Empirical Likelihood Theory for Sums of Means

First proposed by Owen (1988), the EL method can be viewed as a nonparametric counterpart of the maximum likelihood theory. Our first step to showing Theorem 1 is to develop such a method for a special class of performance measures.

Suppose we are given \( m \) independent sets of data \( \{X_{i1}, \ldots, X_{ini}\} \). For the \( i \)-th input model, we define its nonparametric likelihood, in terms of the probability weights \( w_{ij} \) over the support points of the data, to be \( \prod_{j=1}^{n_i} w_{ij} \). The multi-sample likelihood is \( \prod_{i=1}^{m} \prod_{j=1}^{n_i} w_{ij} \). By a simple convexity argument, it can be shown that assigning uniform weights \( w_{ij} = 1/n_i \) for each model maximizes \( \prod_{i=1}^{m} (1/n_i)^{n_i} \). Moreover, uniform weights still maximize even if one allows putting weights outside the support of data, in which case \( \sum_{j=1}^{n_i} w_{ij} < 1 \) for some \( i \), making \( \prod_{j=1}^{n_i} w_{ij} \) even smaller. Therefore, \( (1/n_i)^{n_i} \) can be viewed as the nonparametric maximum likelihood estimate for the \( i \)-th model, and \( \prod_{i=1}^{m} (1/n_i)^{n_i} \) is the multi-sample counterpart.

To proceed, we need to define a parameter of interest that is determined by the input models. In our case, the natural parameter of interest is the performance measure \( Z(P_1, \ldots, P_m) \). This is in general a complex nonlinear function of \( P_i \)'s. As a building block, we focus here on the special case that \( h(X_1, \ldots, X_m) \) is linearly separable among different input models, and each separated component is a simple function of \( X_i \). In other words, \( h(X_1, \ldots, X_m) = \sum_{i=1}^{m} h_i(X_i(1)) \) for some \( h_i : X_i \rightarrow \mathbb{R} \), with \( T_i = 1 \) for all \( i \). The parameter of interest is therefore simply the sum of means of random variables \( h_i(X_i(1)) \). For convenience, we treat \( h_i(X_i(1)) \) as an elementary random variable and upon notational replacement we consider estimating the true parameter \( \mu_0 = \sum_{i=1}^{m} \mathbb{E}X_i \) for a generic variable \( X_i \) under \( P_i \).

The key of the EL method is to establish limit theorems analogous to the celebrated Wilks' Theorem (Wilks 1938) in the maximum likelihood theory, which stipulates that a suitably defined logarithmic likelihood ratio converges to a \( X^2 \) random variable. In the EL setting, we use the so-called profile nonparametric likelihood ratio to carry out inference on parameters. To explain this, first, the nonparametric likelihood ratio is defined as the ratio between the nonparametric likelihood of a given set of weights and the uniform weights (i.e., the nonparametric maximum likelihood estimate). The profile nonparametric likelihood ratio is defined as the maximum among all ratios with weights giving rise to a particular value of \( \mu \), i.e.,

\[
R(\mu) = \max \left\{ \prod_{i=1}^{m} \prod_{j=1}^{n_i} n_i w_{ij} : \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} X_{ij} = \mu, \sum_{j=1}^{n_i} w_{ij} = 1 \text{ for all } i, w_{ij} \geq 0 \text{ for all } i, j \right\},
\]  

(7)
and is defined to be 0 if the optimization problem in (7) is infeasible. Profiling here refers to the categorization of weights that lead to the same value of $\mu$.

The quantity $R(\mu)$ satisfies the following asymptotic property:

**Theorem 4.** Let $X_i$ be a random variable distributed under $P_i$. Assume $0 < \sum_{i=1}^{m} \text{Var}(X_i) < \infty$, and $\max_{i \in I} n_i \leq \gamma \min_{i \in I} n_i$ always holds for some constant $\gamma > 0$, where $I = \{i| \text{Var}(X_i) > 0\}$. Then 

$$-2 \log R(\mu_0), \text{ where } \mu_0 = \sum_{i=1}^{m} \mathbb{E}X_i,$$

is the true parameter, converges in distribution to $\chi^2_1$, the chi-square distribution with degree of freedom one, as $n_i \to \infty$ for $i \in I$.

In other words, the logarithmic profile nonparametric likelihood ratio at the true value asymptotically follows a chi-square distribution with degree of freedom one. The degree of freedom here corresponds to the single target parameter $\mu$. From another perspective, treating the weights $w_i$ as parameters, the degree of freedom is the difference between the dimensions of the full and the constrained parameter space, and one constraint results in a loss of one dimension of the parameter space.

In Theorem 4 only the sample sizes of those models having positive variances are required to grow to infinity. To see the reason for this, note that the data of inputs with zero variance are always equal to the true mean so we can always assign them the uniform weights in (7). Other than these zero-variance inputs, the condition $\max_{i \in I} n_i \leq \gamma \min_{i \in I} n_i$ forces the sample sizes to grow at the same rate. Theorem 4 is a sum-of-mean generalization of the well-known empirical likelihood theorem (ELT) for single-sample mean:

**Theorem 5 (Owen (2001) Theorem 2.2).** Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. random variables distributed under some distribution $P$, with $0 < \text{Var}(Y_1) < \infty$. Then $-2 \log R(\mathbb{E}Y_1)$ converges in distribution to $\chi^2_1$, as $n \to \infty$. The function $R(\cdot)$ here is the same as that in (7) with $m = 1, n_1 = n, X_{1,j} = Y_j$.

Extensions of this theorem have been vastly studied in the literature (e.g., Owen (1990, 1991), Qin and Lawless (1994), Hjort et al. (2009)). The most relevant one is in the context of analysis-of-variance (ANOVA), in which the logarithmic profile nonparametric likelihood ratio at the true means of multiple independent samples are shown to converge to $\chi^2_m$, where $m$ is the number of samples (or groups). However, the argument for this result relies on viewing the multiple samples as a collection of heteroscedastic data and applying the triangular array ELT (Owen (1991)), which does not apply obviously to our case. Another related extension is the plug-in EL (Hjort et al. (2009)) which entails that, under $p$ estimating functions that possibly involve unknown nuisance parameters, the associated logarithmic profile likelihood ratio converges to a weighted sum of $p$ independent $\chi^2$’s, if “good enough” estimators of the unknown nuisance parameters are used in
evaluating the profile likelihood ratio. However, Hjort et al. (2009) focuses on the single-sample case, thus is not directly applicable. There have also been studies on applying EL to hypothesis testing of two-sample mean differences (Liu et al. (2008), Wu and Yan (2012)), but it appears that a fully rigorous proof is not available for our general multi-sample sum-of-means setting. In view of these, we provide a detailed proof of Theorem 4 in Appendix EC.1.

A sketch of the key idea is as follows. We first introduce the auxiliary variables \( \mu_i \) that represent the means of individual samples, so that the constraint \( \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{i,j} x_{i,j} = \mu \) is replaced by \( \sum_{j=1}^{n_i} w_{i,j} x_{i,j} = \mu_i, i = 1, \ldots, m \) and \( \sum_{i=1}^{m} \mu_i = \mu \). The Karush-Kuhn-Tucker (KKT) conditions then enforce the optimal weights to be

\[
w_{i,j}^* = \frac{1}{n_i + \lambda^* (X_{i,j} - \mu_i^*)}
\]

where \( \lambda^* \) is the Lagrange multiplier for the constraint \( \sum_{i=1}^{m} \mu_i = \mu \). An asymptotic analysis on the KKT conditions approximates \( \lambda^* \) as

\[
\lambda^* \approx \frac{\sum_{i=1}^{m} (\bar{X}_i - \mathbb{E} X_i)}{\sum_{i=1}^{m} \frac{\sigma_i^2}{n_i}}
\]

where \( \bar{X}_i = (1/n_i) \sum_{j=1}^{n_i} X_{i,j} \) is the sample mean and \( \sigma_i^2 \) is the variance for the \( i \)-th input model. Moreover, it also approximates \( \mu_i^* \approx \mathbb{E} X_i \). By Taylor’s expansion, the logarithmic profile nonparametric likelihood ratio can be approximated as

\[
-2 \log R(\mu_0) = 2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log \left( 1 + \frac{\lambda^*}{n_i} (X_{i,j} - \mu_i^*) \right)
\approx 2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left( \frac{\lambda^*}{n_i} (X_{i,j} - \mu_i^*) - \frac{\lambda^*^2}{2n_i^2} (X_{i,j} - \mu_i^*)^2 \right)
\approx 2 \sum_{i=1}^{m} \lambda^* (\bar{X}_i - \mu_i^*) - \sum_{i=1}^{m} \frac{\lambda^*^2 \sigma_i^2}{n_i}
\approx \left( \frac{\sum_{i=1}^{m} (\bar{X}_i - \mathbb{E} X_i)}{\sqrt{ \sum_{i=1}^{m} \frac{\sigma_i^2}{n_i} } } \right)^2
\Rightarrow \chi_i^2
\]

where “\( \Rightarrow \)” denotes convergence in distribution. This gives our result in Theorem 4.

4.2. Duality and Optimization-based Confidence Interval

From Theorem 4, a duality-type argument will give rise to a pair of optimization problems whose optimal values will serve as confidence bounds. Suppose that, as in the last section, the output function takes the simple separable form \( h(X_1, \ldots, X_m) = \sum_{i=1}^{m} h_i(X_i) \), where \( X_i \) is distributed under \( P_i \). The following asymptotically exact confidence guarantee is a consequence of Theorem 4:
Theorem 6. Assume $0 < \sum_{i=1}^{m} \text{Var}(h_i(X_i)) < \infty$, and $\max_{i \in I} n_i \leq \gamma \min_{i \in I} n_i$ for some constant $\mu > 0$, where $I = \{i | \text{Var}(h_i(X_i)) > 0\}$. Then

$$P\left( L_\alpha \leq \sum_{i=1}^{m} \mathbb{E} h_i(X_i) \leq U_\alpha \right) \rightarrow 1 - \alpha, \text{ as } n_i \rightarrow \infty \text{ for } i \in I,$$

where $L_\alpha/U_\alpha := \min / \max_{(w_1, \ldots, w_m) \in \alpha} \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{i,j} h_i(X_{i,j})$. Theorem 6 thus translates the asymptotic convergence in Theorem 4 into a confidence bound. The concept is similar to the construction of CI for maximum likelihood estimator in the parametric case, but here the CI is expressed in terms of optimization programs because of the profiling in the definition of $R(\cdot)$.

4.3. Influence Function and Linearization

In the last subsection, asymptotically accurate CIs are obtained for linearly separable output functions. Starting from this subsection, we discuss how to construct CIs for the general output function in (1) that is possibly nonlinear, and in particular, provide theoretical guarantees of our algorithms. The outline of our development consists of the following three stages:

**Stage 1.** Linearize the output $Z(P_1, \ldots, P_m)$ using the influence function taken at the (unknown) true input models, and analyze optimization programs to construct CIs using Theorem 6.

**Stage 2.** Investigate the error incurred by replacing the optimization programs in Stage 1, which are posited assuming full knowledge of the input distributions and expectation evaluation, by their simulatable empirical counterparts, i.e., Step 2 of Algorithm 1.

**Stage 3.** Investigate the error when evaluating the CI bounds using the output performance measure driven by optimal weights rather than the linearization directly, i.e., Step 3 of Algorithms 1, 2 and 3.

We will focus on Stage 1 in this subsection, and Stages 2 and 3 in subsequent subsections.

First, recall the definition of the influence function in (5). The structure of a finite-horizon performance measure $Z$ allows one to derive a precise expression of its influence function and hence a linearization, depicted as:

Proposition 2. Let $(Q^{1}_{1}, \ldots, Q^{1}_{m}), (Q^{2}_{1}, \ldots, Q^{2}_{m})$ be two sets of distributions such that for any $s_{i,t} \in \{1, 2\}$

$$\int |h(x_1, \ldots, x_m)| \prod_{i=1}^{m} \prod_{t=1}^{T_i} dQ^{s_{i,t}}(x_{i,t}) < +\infty,$$

where $x_i = (x_{i,t})_{t=1, \ldots, T_i}$, then

$$\lim_{\epsilon \rightarrow 0^+} \frac{1}{\epsilon} \left( Z(\epsilon Q^{1}_{1} + (1-\epsilon) Q^{2}_{1}, \ldots, \epsilon Q^{1}_{m} + (1-\epsilon) Q^{2}_{m}) - Z(Q^{1}_{1}, \ldots, Q^{1}_{m}) \right) = \sum_{i=1}^{m} \mathbb{E}_{Q^{2}_{i}}[G^{Q^{1}_{1}, \ldots, Q^{1}_{m}}_{i}(X)],$$
Moreover, $E_{Q_i}^i$ specifies the distribution of the $X$, and $G_{i}^{Q_1^i,\ldots,Q_m^i}(x)$ is the influence function with respect to the $i$-th distribution given by

$$G_{i}^{Q_1^i,\ldots,Q_m^i}(x) = \sum_{t=1}^{T_i} E_{Q_1^i,\ldots,Q_m^i} [h(X_1,\ldots,X_m)|X_i(t) = x] - T_i Z(Q_1^i,\ldots,Q_m^i).$$

Moreover, $E_{Q_i}^i[G_{i}^{Q_1^i,\ldots,Q_m^i}(X)] = 0$ for all $i$.

Proposition 2 can be shown by using techniques in the asymptotic analysis of von Mises statistical functionals (e.g., Serfling (2009)). Proposition 2 suggests the following linearization of $Z(Q_1^i,\ldots,Q_m^i)$ at $(Q_1^i,\ldots,Q_m^i)$

$$Z(Q_1^i,\ldots,Q_m^i) \approx Z(Q_1^i,\ldots,Q_m^i) + \sum_{i=1}^{m} E_{Q_i}^i[G_{i}^{Q_1^i,\ldots,Q_m^i}(X)].$$

Here the sum consists of expectations under $Q_i^2$ of influence functions, thus is linear in the distributions. If we take $Q_i^1 = P_i$, i.e., the true model, and $Q_i^2 = w_i$, where we abuse notation slightly here to denote $w_i$ as the discrete distribution that puts weight $w_{i,j}$ on the data point $X_{i,j}$, then

$$Z(w_1,\ldots,w_m) \approx Z(P_1,\ldots,P_m) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{i,j} G_i(X_{i,j}),$$

where $Z(P_1,\ldots,P_m)$ is the true performance measure, and the $G_i$'s are given in Assumption 1. Note that by the last claim in Proposition 2 $G_i$ satisfies $E_{P_i}[G_i(X_i)] = 0$. Therefore, as a special case of Theorem 6, the minimum and maximum values of the right hand side of (8) constitute an asymptotically exact CI, as described below:

**Theorem 7.** Under Assumption 1 if $\max_{i \in I} n_i \leq \gamma \min_{i \in I} n_i$ for some constant $\gamma > 0$, where $I = \{i | \text{Var}(G_i(X_i)) > 0\}$, then

$$P (L_\alpha \leq Z(P_1,\ldots,P_m) \leq U_\alpha) \to 1 - \alpha, \text{ as } n_i \to \infty \text{ for } i \in I,$$

where

$$L_\alpha/U_\alpha := \min / \max_{(w_1,\ldots,w_m) \in \mathcal{A}_\alpha} Z(P_1,\ldots,P_m) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{i,j} G_i(X_{i,j}).$$

### 4.4. Optimization under Sample Influence Function

The pair of programs (9) provide a valid CI for the true performance measure. However, $G_i$ is not precisely known and can only be estimated. We replace the coefficients $G_i(X_{i,j})$ in formulation (9) by their empirical counterparts $\hat{G}_i(X_{i,j})$, where $\hat{G}_i$ is the influence function taken at the empirical input distributions $P_i^{n_i}$ (i.e., uniform weight $1/n_i$ on each data point $X_{i,j}$) defined by

$$\hat{G}_i(x) = \sum_{t=1}^{T_i} E_{P_i^{n_1},\ldots,P_i^{n_m}} [h(X_1,\ldots,X_m)|X_i(t) = x] - T_i Z(P_i^{n_1},\ldots,P_i^{n_m}).$$
This expression can be estimated by simulation. 

Proposition 3 shows the scheme (see Ghosh and Lam (2016) for the proof).

**Proposition 3.** Given input data \( \{X_{i,j}\} \), the empirical influence function \( \hat{G}_i \) evaluated at data point \( X_{i,j} \) satisfies

\[
\hat{G}_{i,j}(X_{i,j}) = \text{Cov}_{P_1^{n_1}, \ldots, P_m^{n_m}}(h(X_1, \ldots, X_m), S_{i,j}(X_i)),
\]

where

\[
S_{i,j}(X_i) = \sum_{t=1}^{T_i} n_i 1 \{X_i(t) = X_{i,j}\} - T_i.
\]

Proposition 3 suggests the following sample covariance as an estimate of \( \hat{G}_i(X_{i,j}) \)

\[
\hat{G}_{i,j} = \frac{1}{R} \sum_{r=1}^{R} [(h(X'_1, \ldots, X'_m) - \hat{Z})(\sum_{t=1}^{T_i} n_i 1 \{X'_i(t) = X_{i,j}\} - T_i)],
\]

where \( X'_1, \ldots, X'_m, r = 1, \ldots, R \) are \( R \) independent replications of the \( m \) input processes generated under the uniform weights, and \( \hat{Z} = \sum_{r=1}^{R} h(X'_1, \ldots, X'_m) / R \) is the sample mean. Now by using (10) with \( R = R_1 \) as an estimate of the influence function \( G_i(X_{i,j}) \), and noting that the term \( Z(P_1, \ldots, P_m) \) in (9) does not depend on the decision variables, we arrive at the programs in (2), which are optimized to obtain the weights \( w_{i}^{\text{min}}, w_{i}^{\text{max}} \) in Step 2 of Algorithm 1.

To analyze the error of (2) relative to (9), we introduce the following three versions of linearized outputs, under the influence function, the empirical influence function, and the simulated empirical influence function respectively:

\[
Z_L(w_1, \ldots, w_m) = Z(P_1, \ldots, P_m) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{i,j} G_i(X_{i,j}),
\]

\[
\hat{Z}_L(w_1, \ldots, w_m) = Z(P_1^{n_1}, \ldots, P_m^{n_m}) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{i,j} \hat{G}_i(X_{i,j}),
\]

\[
\hat{\hat{Z}}_L(w_1, \ldots, w_m) = Z(P_1^{n_1}, \ldots, P_m^{n_m}) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{i,j} \hat{\hat{G}}_{i,j}.
\]

The error of \( \hat{\hat{Z}}_L \) as an approximation to \( Z_L \) comes from two sources: the statistical error between \( Z_L \) and its empirical counterpart \( \hat{Z}_L \) due to finite input data, and the simulation error between \( \hat{Z}_L \) and \( \hat{\hat{Z}}_L \) due to finite simulation runs. The magnitudes of these two sources of errors are investigated in the following proposition.
Proposition 4. Under Assumption 2 with \( k = 2 \), as all \( n_i \)'s are sufficiently large

\[
\mathbb{E}_{P_1, \ldots, P_m} \left[ \sup_{(w_1, \ldots, w_m) \in \mathcal{D}_p} |Z_L(w_1, \ldots, w_m) - \hat{Z}_L(w_1, \ldots, w_m)|^2 \right] \leq C(\alpha) \left( \sum_{i \neq i'} \frac{M_{i,i'}}{n_in_{i'}} + \sum_i \frac{M_i}{n_i^2} \right).
\]

If Assumption 2 holds with \( k = 4 \), then for sufficiently large \( n_i \)'s

\[
\mathbb{E} \left[ \sup_{(w_1, \ldots, w_m) \in \mathcal{D}_p} |\hat{Z}_L(w_1, \ldots, w_m) - \hat{\hat{Z}}_L(w_1, \ldots, w_m)|^2 \right] \leq C(\alpha) \sigma^2 \frac{T}{R_1},
\]

where the expectation is taken with respect to both input data and stochastic simulation.

From Proposition 4 and Theorem 7 we can compute a valid CI by solving the sample programs (2) and then plug in the optimal weights to evaluate \( \hat{\hat{Z}}_L \), where the involved \( Z(P_1^{n_1}, \ldots, P_m^{n_m}) \) can be estimated by averaging the \( R_1 \) replications generated in the influence function estimator (10). This is described in the following theorem.

**Theorem 8.** Suppose all assumptions in Theorem 7 hold. Replace the outputs in Step 3 of Algorithm 7 by

\[
\mathcal{L}_\alpha = \hat{Z} + \sum_{i=1}^m \sum_{j=1}^{n_i} \hat{G}_{i,j} w_{i,j}^{\min}, \quad \mathcal{U}_\alpha = \hat{Z} + \sum_{i=1}^m \sum_{j=1}^{n_i} \hat{G}_{i,j} w_{i,j}^{\max},
\]

where

\[
\hat{Z} = \frac{1}{R_1} \sum_{r=1}^{R_1} h(X_1^r, \ldots, X_m^r)
\]

is estimated from the same replications in Step 1. Then the outputs \( \mathcal{L}_\alpha, \mathcal{U}_\alpha \) satisfy

\[
\lim_{n_i \to \infty} P(\mathcal{L}_\alpha + E_l \leq Z(P_1, \ldots, P_m) \leq \mathcal{U}_\alpha + E_u) = 1 - \alpha,
\]

where \( \mathbb{E}|E_l|, \mathbb{E}|E_u| \leq C(\alpha) \left( \sqrt{\sum_{i \neq i'} \frac{M_{i,i'}}{n_in_{i'}}} + \sum_i \frac{M_i}{n_i^2} + \frac{\sigma^2 T}{R_1} + \frac{\sigma}{\sqrt{R_1}} \right).

Even though the approach suggested in Theorem 8 requires less simulation effort than Algorithm 7 (\( R_1 \) versus \( R_1 + 2R_2 \)), the latter is more beneficial in terms of finite-sample coverage accuracy because it alleviates the effect of linearization. One case of interest is when the performance measure elicits natural boundary (e.g., probabilities, which are between 0 and 1). In this case, using the approach in Theorem 8 which behaves like the delta method, may incur under-coverage issue since the CIs can lie significantly outside the range, whereas Algorithm 7 that outputs the final simulation evaluation is less susceptible to this concern.
4.5. Evaluation of CI Bounds

To establish the validity of Step 3 of Algorithm 1, we first need the following result that quantifies the error of the linearization under the weighted input empirical distributions.

**Proposition 5.** Under Assumption 2 with $k = 2$, as all $n_i$’s are sufficiently large

$$
E_{P_1, \ldots, P_m} \left[ \sup_{(w_1, \ldots, w_m) \in S_{\alpha_i}} |Z(w_1, \ldots, w_m) - Z_L(w_1, \ldots, w_m)|^2 \right] \leq C(\alpha) \left( \sum_{i \neq i'} \frac{M_{i,i'}}{n_i n_i'} + \sum_i \frac{M_i}{n_i^2} \right).
$$

By incorporating the errors from Propositions 4 and 5, we have the following guarantee for our CI estimate assuming $Z$ can be evaluated exactly under a given weight assignment.

**Theorem 9.** Suppose all the assumptions in Theorem 1 hold. Then

$$
\lim_{n_i \to \infty} P \left( Z(w_{i_{\text{min}}}^1, \ldots, w_{i_{\text{min}}}^m) + E_i \leq Z(P_i \ldots, P_m) \leq Z(w_{i_{\text{max}}}^1, \ldots, w_{i_{\text{max}}}^m) + E_u \right) = 1 - \alpha,
$$

where $w_{i_{\text{min}}}, w_{i_{\text{max}}}$ are the optimal weights computed in Step 2 of Algorithm 1, and $E_i, E_u$ satisfy

$$
E[E_i], E[E_u] \leq C(\alpha) \left( \sqrt{\sum_{i \neq i'} \frac{M_{i,i'}}{n_i n_i'} + \sum_i \frac{M_i}{n_i^2}} + \sqrt{\frac{\sigma^2}{R_2}} \right).
$$

To go from Theorem 9 to our main guarantee in Theorem 1, one only needs to incorporate the simulation error of the output evaluation in Step 3 of Algorithm 1. The proof of this is left to the Appendix.

Notice that the error term in Theorem 1 suggests that to make the CI error negligible in comparison to its width, both $R_1, R_2$ have to grow at a faster rate than the data size. Alternately, Algorithms 2 and 3 allow relatively small $R_2$ by adjusting against the simulation uncertainty in these runs, at the cost of slightly weaker coverage guarantees. The following corollary of Theorem 4 indicates the appropriateness of the adjustments in Algorithms 2 and 3.

**Corollary 1.** Let $\bar{X}_i = \sum_{j=1}^{n_i} X_{i,j}/n_i$, $\hat{\sigma}_i^2 = \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_i)^2/(n_i - 1)$ be the sample mean and variance of the $i$-th sample, and $z$ be a fixed constant. Under the same conditions of Theorem 4

$$
-2 \log R(\sum_{i=1}^{m} \bar{X}_i + z \sqrt{\sum_{i=1}^{m} \hat{\sigma}_i^2/n_i}) \to z^2 \text{ in probability as } n_i \to \infty \text{ for } i \in I.
$$

The main message of Corollary 1 is that the EL-based confidence bounds are asymptotically the same as those based on central limit theorems (e.g., the delta method), in the sense that the difference between them are negligible as the data size grows. Therefore we can adjust the EL bounds to incorporate simulation uncertainty in the same way as what we would do with the delta method. Taking the lower confidence bound for instance, the quantity $\sqrt{\hat{\sigma}_i^2 + \sigma_{\text{min}}^2/R_2}$ is the gross standard deviation that accounts for both input and simulation uncertainty, whereas $\hat{\sigma}_i$ alone accounts for the input uncertainty only. Hence their difference is the appropriate adjustment for incorporating simulation uncertainty. This leads to Algorithm 3. Note that Algorithm 3 requires
the estimation of the input-induced variance. Algorithm 2 builds on Algorithm 3 but provides a way to avoid this estimation at the cost of a more conservative CI. The adjustment of Algorithm 2 is based on the observation that $\hat{\sigma}_{\text{min}}/\sqrt{R_2}$ consistently upper bounds $\sqrt{\hat{\sigma}^2 + \hat{\sigma}_{\text{min}}^2/R_2 - \hat{\sigma}}$, and hence Algorithm 2 always generates wider CIs than Algorithm 3. The merit of these two approaches is that it suffices to choose $R_2$ as large as required by the standard CLT, regardless of the input data size. Hence we recommend these alternatives in case of limited simulation budget.

4.6. Confidence Interval under Sufficient Simulation Budget

It is established in Theorems 1, 2 and 3 that the CIs computed from our algorithms are asymptotically accurate up to an error consisting of three terms (or the first two for Theorems 2 and 3), i.e., the linearization error, the sample average approximation error, and the final evaluation error. When the available simulation budget dominates the input data size, the aggregate error is negligible relative to the width of CI. The following three corollaries of Theorems 1, 2 and 3 state precisely this phenomenon.

**Corollary 2.** In addition to the conditions of Theorem 1, if it further holds that $\min_{i \notin I} n_i = \omega(\sqrt{\min_{i \in I} n_i})$ and $R_1, R_2 = \omega(\min_{i \in I} n_i)$, then the outputs $L^{BEL}_\alpha, U^{BEL}_\alpha$ of Algorithm 1 constitute an asymptotically accurate CI

$$\lim_{n_i \to \infty} P \left(L^{BEL}_\alpha \leq Z(P_1, \ldots, P_m) \leq U^{BEL}_\alpha\right) = 1 - \alpha.$$ 

**Corollary 3.** In addition to the conditions of Theorem 2, if it further holds that $\min_{i \notin I} n_i = \omega(\sqrt{\min_{i \in I} n_i})$ and $R_1 = \omega(\min_{i \in I} n_i)$, then the outputs $L^{EEL}_\alpha, U^{EEL}_\alpha$ of Algorithm 2 satisfy

$$\lim_{n_i, R_2 \to \infty} P \left(L^{EEL}_\alpha \leq Z(P_1, \ldots, P_m) \leq U^{EEL}_\alpha\right) \geq 1 - \alpha.$$ 

**Corollary 4.** In addition to the conditions of Theorem 3, if it further holds that $R_1 = \omega(\min_{i \in I} n_i)$, then the outputs $L^{FEL}_\alpha, U^{FEL}_\alpha$ of Algorithm 3 satisfy

$$\lim_{n_i, R_2 \to \infty} P \left(L^{FEL}_\alpha \leq Z(P_1, \ldots, P_m) \leq U^{FEL}_\alpha\right) \geq 1 - \alpha$$

$$\limsup_{n_i, R_2 \to \infty} P \left(L^{FEL}_\alpha \leq Z(P_1, \ldots, P_m) \leq U^{FEL}_\alpha\right) \leq 1 - \alpha + \frac{\alpha^2}{4}.$$ 

Note that in general the condition $\min_{i \notin I} n_i = \omega(\sqrt{\min_{i \in I} n_i})$ is required. Although these input models have zero first-order influence functions, their high-order influence functions could be non-degenerate and hence contribute to the gross error of CI bounds. In fact, if the $l$-th input model has zero influence function up to order $d-1$ but non-zero $d$-th order, then the error incurred will be of order $n_i^{-d/2}$, hence $n_i = \omega(\min_{i \in I} n_i^{1/d})$ suffices.
5. Comparisons with Bootstrap Resampling

We compare our approach justified by Theorems 1, 2 and 3 with the widely used bootstrap resampling in constructing input-induced CIs. The bootstrap is a general technique to approximate sampling distribution of a statistic from which CIs or other statistical quantities can be constructed (Efron (1992)). One common approach, in the context of simulation input uncertainty, is the percentile bootstrap used in Barton and Schruben (1993, 2001). Given \( m \) input samples \( \{X_{i,1}, \ldots, X_{i,n_i}\}, \ldots, \{X_{m,1}, \ldots, X_{m,n_m}\} \), the percentile bootstrap proceeds as follows. First choose \( B \), the number of bootstrap resamples of the input empirical distributions, and \( R_b \), the number of simulation replications for each bootstrap resample. For each \( l = 1, 2, \ldots, B \), draw a simple random sample of size \( n_i \) with replacement, denoted by \( \{X_{i,l,1}, \ldots, X_{i,l,n_i}\} \), for each input model \( i \). Then generate \( R_b \) replications of the output function \( h \) with each variate uniformly generated over \( \{X_{i,l,1}, \ldots, X_{i,l,n_i}\} \), and take their average \( Z_l \). In other words, drive the \( R_b \) simulation runs using the bootstrapped empirical distributions formed by \( \{X_{i,1}, \ldots, X_{i,n_i}\}, i = 1, \ldots, m \). Finally output the 0.025\((B + 1)\)-th and 0.975\((B + 1)\)-th order statistics of \( \{Z_l\}_{l=1}^B \) and \( Z_{\lfloor 0.975(B+1) \rfloor} \), as the lower and upper limits of the CI. Barton and Schruben (1993, 2001) also proposed what they called the method of uniformly randomized empirical distribution function (EDF), which entails building resamples of input empirical distributions by allocating probability weights on all data points that are randomly assigned from an ordered sequence of uniform distributions. This technique, equivalent to the so-called Bayesian bootstrap (Rubin (1981)), is structurally more similar to our method since they both involve assigning probability weights on the support of the data. Barton and Schruben (1993, 2001) used \( R_b = 1 \), as the scale of the input noise dominated that of the stochastic noises in the example systems they considered. They concluded from their experiments that the percentile bootstrap and the uniformly randomized EDF gave rise to similar numerical performances and computational loads.

The total simulation load of the bootstrap is thus of order \( BR_b \). Suppose we want to run enough simulation so that the stochastic noise becomes negligible relative to the input uncertainty. In the bootstrap case one would need \( R_b = \omega(\min_i n_i) \). In the EL case, Theorem 3 in particular suggests to choose \( R_1 = \omega(\min_i n_i) \), \( R_2 = \omega(1) \). Therefore, in overall, bootstrap resampling requires \( BR_b \omega(B \min_i n_i) \) simulation load, whereas EL requires \( R_1 + 2R_2 = \omega(\min_i n_i) \) simulation load and a small effort for the root finding routine. Since \( B \) is typically a big number, one would expect that the EL method is more efficient. In our numerical examples (Section 6), we observe superior performances of EL that can be attributed to the lighter simulation requirement in achieving stable coverage.

Besides computational considerations, there is an additional benefit in using EL regarding the parameter configuration of the algorithm. For a given total simulation budget, it is not easy to...
figure out a reasonable choice of $B$ and $R_b$ for the bootstrap as it can highly depend on the input data sizes and the magnitude of the stochastic uncertainty relative to that of the input uncertainty. Indeed, our experiments in the next section indicates that the coverage of the bootstrap CIs is quite sensitive to the allocations of $B$ and $R_b$. When $B$ and $R_b$ are not appropriately chosen, the bootstrap CI tends to over-cover the truth. On the other hand, for the EL method, particularly FEL, it is safe to set $R_2$ to be some fixed moderately large number (say 50) and invest all other simulation budget to $R_1$ regardless of the data size, as the numerics in the next section show. Despite the strengths of our approach, the bootstrap possesses more flexibility in at least two aspects. First, the bootstrap provides a collection of resampled simulation replications. Under appropriate conditions and manipulations, these replications can approximate the distributions of other statistics, such as higher moments, of the performance measures. They can also be used to construct CIs at different confidence levels, without re-running the bootstrap procedure again. On the contrary, our EL-based optimization approach is more rigid and needs re-optimization and re-evaluation for each new confidence level or statistic of interest. Second, as mentioned in the introduction, the bootstrap does not require the estimation of gradient information, whereas our approach requires so and in this sense is more closely related to the delta method.

6. Numerical Experiments

We present our simulation experiments. Section 6.1 investigates the validity and statistical accuracy of the proposed approach. Section 6.2 compares our performance with the bootstrap and the delta method. Throughout this section, we consider the following two queueing systems.

**System §1:** We consider a canonical M/M/1 queue with arrival rate 0.95 and service rate 1. The system is empty when the first customer comes in. We set our target performance measure as the expected waiting time of the 10-th customer. To put it in the form of (1), let $A_t$ be the inter-arrival time between the $t$-th and $(t+1)$-th customers, $S_t$ be the service time of the $t$-th customer, and

$$h(A_1, A_2, \ldots, A_9, S_1, S_2, \ldots, S_9) = W_{10},$$

where the waiting time $W_{10}$ is calculated via the Lindley recursion

$$W_1 = 0, W_{t+1} = \max\{W_t + S_t - A_t, 0\}, \text{ for } t = 1, \ldots, 9.$$

**System §2:** The second system we consider is a G/G/1 queue with an unreliable server. Suppose at time zero the queue is empty and the server is ready for use. The inter-arrival time (including the time of the first arrival) follows Gamma(2, 1) distribution and the service time follows Gamma(1.5, 1) distribution. Independent of anything else, the system can suddenly fail, and the time to failure, i.e., the period from the last resumption to the next breakdown (the first failure
is counted from the initial time), is exponential with rate 0.2. Service is interrupted upon failure, and resumes after a repair period which is uniform over [0, 4]. The performance measure of interest is the probability that the queue length (including the one being served, if the server is functioning properly) ever exceeds 6 within the first 15 units of time. Mathematically, let $N(t), t \geq 0$ be the process representing the number of customers in the system at time $t$. Then the performance measure is the expected value of

$$ h = 1\{\max_{t \in [0,15]} N(t) \geq 6\}. $$

Strictly speaking, this performance measure does not satisfy our technical assumptions in this paper because of the random number of total arrivals during the period. However, one can modify the problem to include a fixed upper bound on the permitted total arrivals, which will cause little effect on our experiment.

Methods to be tested here include the bootstrap, the delta method, BEL (Algorithm 1), EEL (Algorithm 2) and FEL (Algorithm 3). We pretend that the input distributions as well as the parametric families they belong to are unknown, but data are accessible. Specifically, we simulate data from the prescribed input distributions, and compute a 95% CI for the target performance measure using any of these five methods. For the delta method, the input variance estimate in Algorithm 3 is used. To be specific, we use the empirical input distributions to generate $R_d$ i.i.d. replications of the output $h^r = h(X^r_1, \ldots, X^r_m)$, use them to estimate the gradient $\hat{G}_i(X_{i,j})$, and then compute the CI that accounts for both input and simulation uncertainty

$$ \hat{Z} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}^2}{R_d} + \sum_{i=1}^{m} \frac{1}{n_i} \left( \frac{1}{n_i} \sum_{j=1}^{n_i} \hat{G}^2_{i,j} - \frac{n_i T_i \hat{\sigma}^2}{R_d} \right)}, $$

where $\hat{Z}$ and $\hat{\sigma}^2$ are respectively the sample mean and variance of $\{h^r\}_{r=1}^{R_d}$, and $n_i T_i \hat{\sigma}^2 / R_d$ is the bias correction term for the input-induced variance estimate that removes the extra variance component due to simulation noise.

Each of Tables 1-4 below contains a summary of the simulation results under various input data sizes with relatively sufficient simulation budgets. To explore how the allocation of simulation budget affects the performance of each method, we vary the parameters $B, R_b$ for the bootstrap and $R_1, R_2$ for our methods within an acceptable range while keeping the total simulation size constant. In each setup, 1000 i.i.d. samples are drawn from the true input distributions, and then a CI is constructed from each of them, from which the coverage probability, mean CI length and standard deviation of CI length are estimated. The word “overshoot” means the case that the CI limits exceed the natural bounds of the performance measure. Specifically, it refers to the lower limit being below 0 for system #1 since waiting time must be non-negative, and either the lower limit being below 0 or the upper limit being above 1 for system #2 since the target performance measure is a probability.
6.1. Accuracy of the Optimization-based Confidence Interval

We test the coverage probabilities of the optimization-based CIs. For each data size, we compute a “benchmark” coverage of each method by generating 5000 CIs each of which consumes $5 \times 10^4$ simulation runs, to approximate the simulation-error-free coverage for comparison (shown underneath the name of each method in the tables). We observe first that the benchmark coverages of our optimization-based CIs are close to the nominal value 95% in all cases (92-93% for System ♯1 and 93-94% for System ♯2), which provides a sanity check for the validity of the EL method in our setting. Second, under the simulation budget of the experiments, Tables 1-4 show that in general BEL under-covers compared to the benchmark, EEL over-covers, whereas FEL is accurate. For instance, in Table 1 where the benchmark coverage of the EL method is 91.8%, BEL stays around 90%, EEL ranges from 93% to 96%, whereas FEL ranges from 91% to 92%. This phenomenon is in line with Theorems 1, 2 and 3 since, as we have discussed in Sections 3.3 and 4.5, BEL does not take into account the stochastic uncertainty in the final evaluation, EEL captures the stochastic uncertainty but in a conservative manner, while FEL is provably accurate. The under-coverage issue of BEL and the over-coverage issue of EEL become more severe when $R_2$ is chosen small, while FEL delivers accurate coverage for all considered parameter values. Therefore we recommend FEL over the other two when the user has a limited simulation budget.

| Table 1 | System ♯1. $n_1 = 30, n_2 = 25$. Total simulation budget 2000. Run times (second/CI): delta method $1.0 \times 10^{-2}$, bootstrap $0.8 \sim 1.6 \times 10^{-2}$, three EL methods $1.0 \sim 1.2 \times 10^{-2}$. |
|---|---|
| methods & parameters | coverage estimate mean CI length std. CI length % of overshoot |
| BEL (91.8%*) | $R_1 = 1000, R_2 = 500$ | 89.6% | 4.76 | 2.17 | 0% |
| | $R_1 = 1500, R_2 = 250$ | 90.7% | 4.72 | 1.99 | 0% |
| | $R_1 = 1800, R_2 = 100$ | 88.7% | 4.76 | 2.15 | 0% |
| | $R_1 = 1900, R_2 = 50$ | 89.2% | 4.79 | 2.24 | 0% |
| EEL (91.8%*) | $R_1 = 1000, R_2 = 500$ | 93.1% | 5.21 | 2.19 | 0% |
| | $R_1 = 1500, R_2 = 250$ | 94.1% | 5.38 | 2.21 | 0% |
| | $R_1 = 1800, R_2 = 100$ | 95.1% | 5.67 | 2.42 | 0% |
| | $R_1 = 1900, R_2 = 50$ | 96.0% | 6.16 | 2.64 | 0.1% |
| FEL (91.8%*) | $R_1 = 1000, R_2 = 500$ | 90.5% | 4.72 | 2.06 | 0% |
| | $R_1 = 1500, R_2 = 250$ | 91.9% | 4.83 | 2.07 | 0% |
| | $R_1 = 1800, R_2 = 100$ | 91.9% | 4.93 | 2.08 | 0% |
| | $R_1 = 1900, R_2 = 50$ | 91.5% | 5.06 | 2.20 | 0% |
| bootstrap (91.0%*) | $B = 50, R_b = 40$ | 91.2% | 4.90 | 2.23 | 0% |
| | $B = 100, R_b = 20$ | 93.5% | 4.98 | 2.02 | 0% |
| | $B = 400, R_b = 5$ | 96.9% | 6.09 | 2.28 | 0% |
| | $B = 1000, R_b = 2$ | 99.2% | 7.74 | 2.82 | 0% |
| delta method (86.6%*) | $R_d = 2000$ | 84.9% | 4.66 | 2.08 | 54% |

* denotes the benchmark coverage with approximately no simulation noise.
Table 2 System #2. \( n_1 = 30, n_2 = 25, n_3 = 30, n_4 = 25 \). Total simulation budget 2500. Run times (second/CI): delta method \( 7.6 \times 10^{-2} \), bootstrap \( (6.2 \sim 7.9) \times 10^{-2} \), three EL methods \( (7.1 \sim 8.2) \times 10^{-2} \).

| methods & parameters | coverage estimate | mean CI length | std. CI length | % of overshoot |
|----------------------|------------------|----------------|----------------|---------------|
| BEL (93.1%) \( \star \) | \( R_1 = 1500, R_2 = 500 \) | 90.7% | 0.354 | 0.143 | 0% |
| | \( R_1 = 2000, R_2 = 250 \) | 89.0% | 0.364 | 0.148 | 0% |
| | \( R_1 = 2300, R_2 = 100 \) | 90.7% | 0.373 | 0.147 | 0% |
| | \( R_1 = 2400, R_2 = 50 \) | 88.2% | 0.371 | 0.160 | 0% |
| EEL (93.1%) \( \star \) | \( R_1 = 1500, R_2 = 500 \) | 94.4% | 0.413 | 0.150 | 10% |
| | \( R_1 = 2000, R_2 = 250 \) | 96.7% | 0.449 | 0.155 | 22% |
| | \( R_1 = 2300, R_2 = 100 \) | 97.7% | 0.496 | 0.172 | 42% |
| | \( R_1 = 2400, R_2 = 50 \) | 98.1% | 0.538 | 0.190 | 51% |
| FEL (93.1%) \( \star \) | \( R_1 = 1500, R_2 = 500 \) | 91.9% | 0.364 | 0.140 | 0% |
| | \( R_1 = 2000, R_2 = 250 \) | 94.1% | 0.381 | 0.140 | 0% |
| | \( R_1 = 2300, R_2 = 100 \) | 92.7% | 0.394 | 0.148 | 0% |
| | \( R_1 = 2400, R_2 = 50 \) | 92.4% | 0.414 | 0.151 | 0% |
| bootstrap (93.7%) \( \star \) | \( B = 50, R_b = 50 \) | 93.8% | 0.436 | 0.180 | 0% |
| | \( B = 100, R_b = 25 \) | 95.9% | 0.469 | 0.169 | 0% |
| | \( B = 250, R_b = 10 \) | 98.8% | 0.529 | 0.184 | 0% |
| | \( B = 1250, R_b = 2 \) | 99.8% | 0.817 | 0.243 | 0% |
| delta method (86.8%) \( \star \) | \( R_d = 2500 \) | 85.8% | 0.400 | 0.180 | 93% |

Table 3 System #1. \( n_1 = 120, n_2 = 100 \). Total simulation budget 8000. Run times (second/CI): delta method \( 5.3 \times 10^{-2} \), bootstrap \( (2.9 \sim 3.8) \times 10^{-2} \), three EL methods \( (3.0 \sim 5.0) \times 10^{-2} \).

| methods & parameters | coverage estimate | mean CI length | std. CI length | % of overshoot |
|----------------------|------------------|----------------|----------------|---------------|
| BEL (93.7%) \( \star \) | \( R_1 = 4000, R_2 = 2000 \) | 92.6% | 2.47 | 0.597 | 0% |
| | \( R_1 = 7000, R_2 = 500 \) | 92.4% | 2.46 | 0.606 | 0% |
| | \( R_1 = 7800, R_2 = 100 \) | 91.9% | 2.48 | 0.713 | 0% |
| | \( R_1 = 7900, R_2 = 50 \) | 89.6% | 2.45 | 0.787 | 0% |
| EEL (93.7%) \( \star \) | \( R_1 = 4000, R_2 = 2000 \) | 95.7% | 2.66 | 0.626 | 0% |
| | \( R_1 = 7000, R_2 = 500 \) | 97.7% | 2.90 | 0.678 | 0% |
| | \( R_1 = 7800, R_2 = 100 \) | 98.0% | 3.50 | 0.870 | 0% |
| | \( R_1 = 7900, R_2 = 50 \) | 98.8% | 3.94 | 1.04 | 0% |
| FEL (93.7%) \( \star \) | \( R_1 = 4000, R_2 = 2000 \) | 93.7% | 2.45 | 0.588 | 0% |
| | \( R_1 = 7000, R_2 = 500 \) | 92.8% | 2.50 | 0.619 | 0% |
| | \( R_1 = 7800, R_2 = 100 \) | 94.1% | 2.74 | 0.705 | 0% |
| | \( R_1 = 7900, R_2 = 50 \) | 93.6% | 2.87 | 0.882 | 0% |
| bootstrap (94.2%) \( \star \) | \( B = 50, R_b = 160 \) | 92.7% | 2.56 | 0.675 | 0% |
| | \( B = 100, R_b = 80 \) | 96.4% | 2.64 | 0.613 | 0% |
| | \( B = 400, R_b = 20 \) | 98.8% | 3.19 | 0.658 | 0% |
| | \( B = 100, R_b = 8 \) | 100% | 4.19 | 0.800 | 0% |
| delta method (91.5%) \( \star \) | \( R_d = 8000 \) | 92.0% | 2.45 | 0.560 | 0% |
6.2. Numerical Comparisons with the Bootstrap and the Delta Method

We compare our methods with the bootstrap and the delta method in terms of coverage accuracy and algorithmic configuration.

The benchmark coverages of our methods and the bootstrap appear to be quite similar in all considered cases (within 1% in all tables except Table 4, which is 2%). Our optimization-based methods also appear to perform similarly favorably as the bootstrap if their respective parameters, namely $R_1, R_2$ for our methods and $B, R_b$ for the bootstrap, are optimally tuned, in the sense that the coverage estimates get close to the benchmark coverages. However, optimal tuning is not easy to achieve in hindsight, and our FEL appears to show more robust performance with respect to these allocations than the bootstrap. In the latter, when $R_b$ is chosen large relative to the data size and $B$ is set around 50, the coverages of the CIs are close to the benchmark coverages in all cases. However, as $R_b$ decreases, the coverage probabilities of bootstrap CIs quickly rise towards 100%. This over-coverage issue is due to not using a large enough $R_b$, and consequently leading to a higher variability, as discussed in Barton et al. (2002) and Barton (2007). In contrast, the coverage probabilities of FEL stay almost unchanged under various budget allocations (including the case that $R_2$ is as small as 50). FEL thus seems easy to use in terms of algorithmic configuration; in particular, merely setting $R_2 = 50$ appears doing well.

FEL also appears to require less simulation budget to stabilize, or in other words, to achieve the benchmark coverage. In each of Tables 4 and 5, the first row shows the coverage estimates of

| methods & parameters | coverage estimate | mean CI length | std. CI length | % of overshoot |
|-----------------------|-------------------|----------------|----------------|---------------|
| BEL (93.5%*)          |                   |                |                |               |
| $R_1 = 4000, R_2 = 2000$ | 91.3%            | 0.178          | 0.046          | 0%            |
| $R_1 = 7000, R_2 = 500$ | 92.6%            | 0.186          | 0.048          | 0%            |
| $R_1 = 7800, R_2 = 100$ | 88.2%            | 0.186          | 0.064          | 0%            |
| $R_1 = 7900, R_2 = 50$  | 81.6%            | 0.186          | 0.083          | 0%            |
| EEL (93.5%*)          |                   |                |                |               |
| $R_1 = 4000, R_2 = 2000$ | 95.4%            | 0.207          | 0.050          | 0%            |
| $R_1 = 7000, R_2 = 500$ | 97.0%            | 0.245          | 0.059          | 0%            |
| $R_1 = 7800, R_2 = 100$ | 98.3%            | 0.318          | 0.081          | 15%           |
| $R_1 = 7900, R_2 = 50$  | 99.1%            | 0.371          | 0.100          | 46%           |
| FEL (93.5%*)          |                   |                |                |               |
| $R_1 = 4000, R_2 = 2000$ | 92.4%            | 0.176          | 0.044          | 0%            |
| $R_1 = 7000, R_2 = 500$ | 92.4%            | 0.191          | 0.049          | 0%            |
| $R_1 = 7800, R_2 = 100$ | 93.8%            | 0.232          | 0.071          | 0%            |
| $R_1 = 7900, R_2 = 50$  | 94.4%            | 0.260          | 0.084          | 0%            |
| bootstrap (95%*)       |                   |                |                |               |
| $B = 50, R_b = 160$  | 95.9%            | 0.228          | 0.058          | 0%            |
| $B = 100, R_b = 80$   | 98.2%            | 0.247          | 0.054          | 0%            |
| $B = 400, R_b = 20$   | 100%             | 0.352          | 0.071          | 0%            |
| $B = 1000, R_b = 8$   | 99.9%            | 0.449          | 0.096          | 0%            |
| delta method (92.1%*) |                   |                |                |               |
| $R_d = 8000$          | 92.6%            | 0.197          | 0.048          | 0.1%          |
the bootstrap and FEL under particular allocations of the same overall simulation budget. Both appear to be close to their respective benchmark coverages shown in Tables 1 and 2. However, the coverages of the bootstrap can be illusory in these cases since, as the number of resamples increases with \( R_b \) fixed, the coverage rises from 91% to 95% for System \( \#1 \) and from 94% to 97% for System \( \#2 \) as shown in the following rows. These deviate from the benchmark coverages, and indicate that neither \( B \) nor \( R_b \) is large enough for the bootstrap to work properly. In contrast, the coverage of FEL appears quite stable and remains around its benchmark under increasing values of \( R_1 \) and \( R_2 \). Given the similar benchmark coverages between the bootstrap and our methods, FEL seems to elicit in overall a more robust and lighter simulation effort than the bootstrap.

| Parameters | Coverage | Estimate |
|------------|----------|----------|
| \( B = 40, R_b = 15 \) | 90.9% | 90.3% |
| \( B = 100, R_b = 15 \) | 92.4% | 91.9% |
| \( B = 200, R_b = 15 \) | 93.6% | 90.2% |
| \( B = 500, R_b = 15 \) | 94.7% | 90.8% |

| Parameters | Coverage | Estimate |
|------------|----------|----------|
| \( B = 37, R_b = 30 \) | 93.8% | 92.7% |
| \( B = 100, R_b = 30 \) | 96.6% | 93.1% |
| \( B = 200, R_b = 30 \) | 96.7% | 91.5% |
| \( B = 400, R_b = 30 \) | 96.9% | 92.1% |

Compared to the delta method, our optimization-based CIs possess better coverages, especially in the situation of limited input data size. When the data size is less than 30 for each input model, e.g., Tables 1 and 2, the coverage probabilities of the delta-method CIs are around 86%, while our methods are around 92% (for the benchmark). The unsatisfactory coverage of the delta-method CI could be attributed to the overshoot issue. Tables 1 and 2 show that most of the time the delta-method CI exceeds the natural bounds of the target performance measure, which renders its effective length shorter and hence an inferior coverage. The coverage gets much better for the delta-method CI when input data size rises above 100, e.g., Tables 3 and 4, which gets close to, but still falls short of, our optimization-based counterparts.
7. Conclusion

We have proposed an optimization-based approach to construct CIs for simulation output performance measures that account for the input uncertainty from finite data. This approach relies on solving a pair of optimization programs posited over distributions supported on the data, with a constraint expressed in terms of the weighted average of empirically defined Burg-entropy divergences. It then uses the solutions to define probability weights that subsequently drive simulation runs. We present several related procedures under this approach and analyze their statistical performances using a generalization of the EL method. Compared to the bootstrap, our approach requires less simulation budget to achieve stable coverage and is less sensitive to the allocation choices, as explained both theoretically and shown by our numerical experiments. The numerical results also reveal that our approach tends to curb the under-coverage issues encountered in the delta method. The last of our procedures, FEL, seems particularly attractive compared to both the bootstrap and the delta method in terms of finite-data finite-simulation performance.

Acknowledgments

A preliminary conference version of this paper appeared in the Winter Simulation Conference 2016. We gratefully acknowledge support from the National Science Foundation under grants CMMI-1542020, CMMI-1523453 and CAREER CMMI-1653339.

References

Asmussen S, Glynn PW (2007) Stochastic Simulation: Algorithms and Analysis, volume 57 (Springer Science & Business Media).

Banks J, CARSON II JS, Nelson BL, Nicol DM (2005) Discrete-event system simulation (Pearson), fourth edition.

Barton RR (2007) Presenting a more complete characterization of uncertainty: Can it be done. Proceedings of the 2007 INFORMS Simulation Society research workshop (INFORMS Simulation Society).

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

Barton RR, Chick SE, Cheng RC, Henderson SG, Law AM, Schmeiser BW, Leemis LM, Schruben LW, Wilson JR (2002) Panel discussion on current issues in input modeling: panel on current issues in simulation input modeling. Proceedings of the 34th conference on Winter simulation: exploring new frontiers, 353–369 (Winter Simulation Conference).

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

Barton RR, Schruben LW (1993) Uniform and bootstrap resampling of empirical distributions. Proceedings of the 25th conference on Winter simulation, 503–508 (ACM).
Barton RR, Schruben LW (2001) Resampling methods for input modeling. Proceedings of the Winter Simulation Conference, 2001, volume 1, 372–378 (IEEE).

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

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

Ben-Tal A, Nemirovski A (2002) Robust optimization—methodology and applications. Mathematical Programming 92(3):453–480.

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

Blanchet J, Kang Y (2016) Sample out-of-sample inference based on wasserstein distance. arXiv preprint arXiv:1605.01340.

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

Blanchet J, Murthy K (2016a) Quantifying distributional model risk via optimal transport. arXiv preprint arXiv:1604.01446.

Blanchet J, Murthy KR (2016b) On distributionally robust extreme value analysis. arXiv preprint arXiv:1601.06858.

Boyd S, Vandenberghe L (2004) Convex optimization (Cambridge university press).

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

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

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

Chick SE (2006) Bayesian ideas and discrete event simulation: Why, what and how. Proceedings of the 2006 Winter Simulation Conference (WSC), 96–106 (IEEE).

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

Dey S, Juneja S (2012) Incorporating fat tails in financial models using entropic divergence measures. arXiv preprint arXiv:1203.0643.
Duchi J, Glynn P, Namkoong H (2016) Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*.

Durrett R (2010) *Probability: theory and examples* (Cambridge university press).

Efron B (1992) *Bootstrap Methods: Another Look at the Jackknife* (Springer).

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

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

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

Ghosh S, Lam H (2016) Robust analysis in stochastic simulation: Computation and performance guarantees. *arXiv preprint arXiv:1507.05609*.

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

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

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

Hampel FR (1974) The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 69(346):383–393.

Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (2011) *Robust statistics: the approach based on influence functions*, volume 114 (John Wiley & Sons).

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

Henderson SG (2003) Input modeling: Input model uncertainty: Why do we care and what should we do about it? *Proceedings of the 35th Winter Simulation Conference*, 90–100 (IEEE).

Hjort NL, McKeague IW, Van Keilegom I (2009) Extending the scope of empirical likelihood. *The Annals of Statistics* 1079–1111.

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

Huber PJ, Ronchetti EM (2009) Robust statistics. *International Encyclopedia of Statistical Science*, 1248–1251 (Springer), second edition.

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

Jain A, Lim AE, Shanthikumar JG (2010) On the optimality of threshold control in queues with model uncertainty. *Queueing Systems* 65(2):157–174.
Jiang R, Guan Y (2012) Data-driven chance constrained stochastic program. *Mathematical Programming* 1–37.

Lam H (2016a) Advanced tutorial: Input uncertainty and robust analysis in stochastic simulation. *Winter Simulation Conference (WSC), 2016*, 178–192 (IEEE).

Lam H (2016b) Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *arXiv preprint arXiv:1605.09349*.

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

Lam H (2017) Sensitivity to serial dependency of input processes: A robust approach. *Management Science* Articles in advance, URL [http://dx.doi.org/10.1287/mnsc.2016.2667](http://dx.doi.org/10.1287/mnsc.2016.2667).

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

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

Liu Y, Zou C, Zhang R (2008) Empirical likelihood for the two-sample mean problem. *Statistics & Probability Letters* 78(5):548–556.

Nilim A, El Ghaoui L (2005) Robust control of markov decision processes with uncertain transition matrices. *Operations Research* 53(5):780–798.

Owen A (1990) Empirical likelihood ratio confidence regions. *The Annals of Statistics* 90–120.

Owen A (1991) Empirical likelihood for linear models. *The Annals of Statistics* 1725–1747.

Owen AB (1988) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75(2):237–249.

Owen AB (2001) *Empirical likelihood* (CRC press).

Pardo L (2005) *Statistical inference based on divergence measures* (CRC Press).

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

Qin J, Lawless J (1994) Empirical likelihood and general estimating equations. *The Annals of Statistics* 300–325.

Rao RR (1962) Relations between weak and uniform convergence of measures with applications. *The Annals of Mathematical Statistics* 659–680.

Rockafellar RT (2015) *Convex analysis* (Princeton university press).

Rubin DB (1981) The Bayesian bootstrap. *The Annals of Statistics* 9(1):130–134.

Serfling RJ (2009) *Approximation Theorems of Mathematical Statistics*, volume 162 (John Wiley & Sons).
Shapiro A, Dentcheva D, et al. (2014) *Lectures on stochastic programming: modeling and theory*, volume 16 (SIAM).

Song E, Nelson BL (2015) Quickly assessing contributions to input uncertainty. *IIE Transactions* 47(9):893–909.

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

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

Wilks SS (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics* 9(1):60–62.

Wu C, Yan Y (2012) Empirical likelihood inference for two-sample problems. *Stat. Interface* 5:345–354.

Xie W, Nelson BL, Barton RR (2014) A bayesian framework for quantifying uncertainty in stochastic simulation. *Operations Research* 62(6):1439–1452.

Xie W, Nelson BL, Barton RR (2016) Multivariate input uncertainty in output analysis for stochastic simulation. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 27(1):5.

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

Zhu H, Zhou E (2015) Risk quantification in stochastic simulation under input uncertainty. *arXiv preprint arXiv:1507.06015*.

Zouaoui F, Wilson JR (2003) Accounting for parameter uncertainty in simulation input modeling. *IIE Transactions* 35(9):781–792.

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

EC.1. Proofs of Theorem 4 and Corollary 1.

Proof of Theorem 4. If some $X_i$ has zero variance, consider the quantity defined as

$$R_I(\sum_{i \in I} \mathbb{E} X_i) = \max \left\{ \prod_{i \in I} n_i \prod_{j=1}^{n_i} w_{i,j} \left| \sum_{i \in I} n_i \sum_{j=1}^{n_i} w_{i,j} X_{i,j} = \sum_{i \in I} \mathbb{E} X_i, w_i \in \mathcal{P}^{n_i}, \text{ for } i \in I \right\}.$$ 

By letting $w_{i,j} = \frac{1}{n_i}$ for $i \notin I$, it is clear that $R_I(\sum_{i \in I} \mathbb{E} X_i)$ is always equal to $R(\sum_{i=1}^{m} \mathbb{E} X_i)$. So we can always assume that $\text{Var}(X_i) > 0$ for each $i$, i.e. $I = \{1, 2, \ldots, m\}$. Moreover, without loss of generality, we can assume $\sum_{i=1}^{m} \mathbb{E} X_i = 0$ by translating the first sample $\{X_{1,1}, \ldots, X_{1,n_1}\}$ to $\{X_{1,1} - \sum_{i=1}^{m} \mathbb{E} X_i, \ldots, X_{1,n_1} - \sum_{i=1}^{m} \mathbb{E} X_i\}$. For simplicity, we use $R$ in place of $R(\sum_{i=1}^{m} \mathbb{E} X_i)$ throughout the proof. Introducing extra variables $\mu_i$ for each $\sum_{j=1}^{n_i} w_{i,j} X_{i,j}$ shows clearly that the maximization in (7) is equivalent to the following convex program

$$\min_{w_1, \ldots, w_m, \mu} -\sum_{i=1}^{m} \sum_{j=1}^{n_i} \log(n_i w_{i,j})$$

subject to $\sum_{j=1}^{n_i} w_{i,j} X_{i,j} = \mu_i, i = 1, \ldots, m$ 

$$\sum_{j=1}^{n_i} w_{i,j} = 1, i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \mu_i = 0$$

(Ec.1)

where $w_i = (w_{i,1}, \ldots, w_{i,n_i})$ and $\mu = (\mu_1, \ldots, \mu_m)$. The non-negativity constraints $w_{i,j} \geq 0$ are dropped since they are implicitly imposed in the objective function.

As the first step, we show that, with probability tending to one, Slater’s condition holds for optimization problem (Ec.1), i.e. (Ec.1) has at least one feasible solution $(w_1, \ldots, w_m, \mu)$ such that each $w_{i,j} > 0$. To this end, it suffices to show that, with probability tending to one, there exists $w_{i,j} > 0$ such that $\sum_{j=1}^{n_i} w_{i,j} X_{i,j} = \mathbb{E} X_i$ for each $i$. $\text{Var}(X_i) > 0$ implies $P(X_i > \mathbb{E} X_i) > 0, P(X_i < \mathbb{E} X_i) > 0$. By strong law of large numbers, with probability one

$$\frac{1}{n_i} \sum_{j=1}^{n_i} 1\{X_{i,j} > \mathbb{E} X_i\} \rightarrow P(X_i > \mathbb{E} X_i),$$

$$\frac{1}{n_i} \sum_{j=1}^{n_i} 1\{X_{i,j} < \mathbb{E} X_i\} \rightarrow P(X_i < \mathbb{E} X_i).$$
where $1\{\cdot\}$ is the indicator function. Hence with probability tending to one

$$\frac{1}{n_i} \sum_{j=1}^{n_i} 1\{X_{i,j} > \mathbb{E}X_i\} > 0,$$

$$\frac{1}{n_i} \sum_{j=1}^{n_i} 1\{X_{i,j} < \mathbb{E}X_i\} > 0.$$ 

This implies that $\mathbb{E}X_i$ lies in the open interval $(\min_j X_{i,j}, \max_j X_{i,j})$, and hence there must exist a set of positive weights $w_{i,j} > 0$, $\sum_{j=1}^{n_i} w_{i,j} = 1$ such that

$$\sum_{j=1}^{n_i} w_{i,j} X_{i,j} = \mathbb{E}X_i.$$ 

Since there are $m$ samples, which is finite, the same thing is true for all $i$ with probability tending to one.

Second, we show that an optimal solution exists for both (EC.1) and its Lagrangian dual, which together satisfy the KKT conditions. Based on the first step, we can from now on pretend that Slater’s condition always holds for (EC.1). We will also need that (EC.1) indeed has an optimal solution with a finite optimal value. Notice that each $-\log(n_i w_{i,j})$ is bounded below by $-\log n_i$, so when $w_{i,j} \to 0$ for some $i, j$, $-\log(n_i w_{i,j})$ approaches $+\infty$, and hence the objective value approaches $+\infty$. Therefore, we can restrict the domain to $w_{i,j} \geq \epsilon$ for some small enough $\epsilon > 0$. Since the set $\{(w_1, \ldots, w_m) : \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{i,j} X_{i,j} = 0, \sum_{j=1}^{n_i} w_{i,j} = 1, w_{i,j} \geq \epsilon\}$ is compact, an optimal solution $(w^*_1, \ldots, w^*_m, \mu^*)$ exists for (EC.1). By Section 5.2.3 of Boyd and Vandenberghe (2004), if a convex program satisfies Slater’s condition, and has an optimal solution with finite value, then the duality gap is zero and an optimal solution exists for its Lagrangian dual program. Let $(\lambda^*_1, \lambda^*_2, \lambda^*_3)$ be an optimal solution to the dual. By Section 5.5.3 of Boyd and Vandenberghe (2004), any pair of primal and dual optimal solutions with zero duality gap satisfies the KKT conditions, i.e.

$$\sum_{j=1}^{n_i} w^*_{i,j} X_{i,j} = \mu^*_i, \text{ for } i = 1, \ldots, m$$

$$\sum_{j=1}^{n_i} w^*_{i,j} = 1, \text{ for } i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \mu^*_i = 0$$

$$-\frac{1}{w^*_{i,j}} + \lambda^*_1 X_{i,j} + \lambda^*_2 = 0, \forall i, j \quad \text{(EC.2)}$$

$$-\lambda^*_1 + \lambda^*_3 = 0, \text{ for } i = 1, \ldots, m$$

Some manipulation to these equations gives $\lambda^*_2 = n_i - \lambda^*_1 \mu^*_i$, $\lambda^*_3 = \lambda^*_1$, and so it follows from (EC.2) that

$$w^*_{i,j} = \frac{1}{n_i} + \lambda^*(X_{i,j} - \mu^*_i) \quad \text{(EC.3)}$$
where we substitute $\lambda^*_3$ with $\lambda^*$ for convenience, and $\lambda^*, \mu^*_i$ satisfy

$$
\sum_{j=1}^{n_i} \frac{X_{i,j} - \mu^*_i}{n_i + \lambda^*(X_{i,j} - \mu^*_i)} = 0, \text{ for } i = 1, \ldots, m
$$

(EC.4)

$$
\sum_{i=1}^{m} \mu^*_i = 0.
$$

(EC.5)

The third step is to bound the magnitude of $\lambda^*$. Write (EC.3) as

$$
\frac{1}{n_i + \lambda^*(X_{i,j} - \mu^*_i)} = \frac{1}{n_i} \left(1 - \frac{\frac{\lambda^*}{n_i}(X_{i,j} - \mu^*_i)}{1 + \frac{\lambda^*}{n_i}(X_{i,j} - \mu^*_i)}\right)
$$

(EC.6)

and plugging into (EC.4) gives

$$
\bar{X}_i - \mu^*_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\lambda^*}{n_i}(X_{i,j} - \mu^*_i)^2
$$

(EC.7)

where $\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{i,j}$. Multiply both sides by $\text{sign}(\lambda^*)$ to make the right hand side positive

$$
\text{sign}(\lambda^*)(\bar{X}_i - \mu^*_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{|\lambda^*|}{n_i}(X_{i,j} - \mu^*_i)^2.
$$

(EC.8)

Since $w_{i,j}$ must be strictly positive, we must have $1 + \frac{\lambda^*}{n_i}(X_{i,j} - \mu^*_i) > 0, \forall i, j$. Also note that $|\mu^*_i| = \left|\sum_{j=1}^{n_i} w_{i,j} X_{i,j}\right| \leq \sum_{j=1}^{n_i} w_{i,j} |X_{i,j}| \leq \max_{j=1, \ldots, n_i} |X_{i,j}|$. Recall that $\max_i n_i \leq \gamma \min_i n_i$, so $\min_i n_i > cn$ for $c = \frac{1}{m}$, $n = \sum_{i=1}^{m} n_i$. A lower bound of RHS of (EC.8) can be derived as follows

$$
\frac{1}{n_i} \sum_{j=1}^{n_i} \frac{|\lambda^*|}{n_i}(X_{i,j} - \mu^*_i)^2 \geq \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{|\lambda^*|}{n_i}(X_{i,j} - \mu^*_i)^2
$$

$$
\geq \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{|\lambda^*|}{n_i}(X_{i,j} - \mu^*_i)^2
$$

$$
\geq \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{|\lambda^*|}{n_i}(X_{i,j} - \mu^*_i)^2
$$

$$
\geq \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{|\lambda^*|}{n_i}(X_{i,j} - \mu^*_i)^2
$$

$$
\geq \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{|\lambda^*|}{n_i}(X_{i,j} - \mu^*_i)^2
$$

= \frac{1}{1 + \frac{|\lambda^*|}{cn} \cdot 2Z_n} \left(\hat{\sigma}^2 + 2(\bar{X}_i - EX_i)(EX_i - \mu^*_i) + (EX_i - \mu^*_i)^2\right)
$$

EC.9

where $Z_n = \max_{i=1, \ldots, m, j=1, \ldots, n_i} |X_{i,j}|$, $\hat{\sigma}^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (X_{i,j} - EX_i)^2$. Applying Lemma 11.2 of (Owen (2001)) to $X_{i,j}$ gives $\max_j |X_{i,j}| = o(n_i^{1/2})$ for each $i$, hence $Z_n = o(n^{1/2})$ and $\mu^*_i = o(n_i^{1/2})$. Also by
central limit theorem, \( \bar{X}_i - \mu_i = O_p\left(n_i^{-\frac{1}{2}}\right) \). By plugging the lower bound (EC.9) into (EC.8), we have

\[
\left(1 + \frac{|\lambda^*|}{cn} \cdot 2Z_n\right) \text{sign}(\lambda^*) \left(\bar{X}_i - \mu_i^*\right) \geq \frac{|\lambda^*|}{n} \left(\hat{\sigma}_i^2 + O_p(n^{-\frac{3}{2}})o(n^{\frac{1}{2}})\right)
\]

\[
= \frac{|\lambda^*|}{n} \left(\hat{\sigma}_i^2 + o_p(1)\right)
\]

(EC.10)

Summing up both sides of (EC.10) over \( i = 1, \ldots, m \), and using (EC.5) and \( Z_n = o(n^{\frac{1}{2}}) \) gives

\[
\left(1 + \frac{|\lambda^*|}{n} o(n^{\frac{1}{2}})\right) \text{sign}(\lambda^*) \sum_{i=1}^{m} \bar{X}_i \geq \frac{|\lambda^*|}{n} \left(\sum_{i=1}^{m} \hat{\sigma}_i^2 + o_p(1)\right).
\]

Rearrange it to

\[
\frac{|\lambda^*|}{n} \left(\sum_{i=1}^{m} \hat{\sigma}_i^2 + o_p(1) + o(n^{\frac{1}{2}}) \sum_{i=1}^{m} \bar{X}_i\right) \leq \sum_{i=1}^{m} \bar{X}_i.
\]

(EC.11)

Notice that \( \hat{\sigma}_i^2 \rightarrow \sigma_i^2 = Var(X_i) \) a.s. and \( \sum_{i=1}^{m} \bar{X}_i = 0 \), hence

\[
\sum_{i=1}^{m} \bar{X}_i = \sum_{i=1}^{m} (\bar{X}_i - \mu_i) = \sum_{i=1}^{m} O_p(n_i^{-\frac{1}{2}}) = O_p(n^{-\frac{1}{2}}).
\]

So if \( \sum_{i=1}^{m} \sigma_i^2 > 0 \), (EC.11) implies

\[
\frac{|\lambda^*|}{n} \leq \frac{O_p(n^{-\frac{1}{2}})}{\sum_{i=1}^{m} \sigma_i^2 + o_p(1)}.
\]

Therefore \( \frac{|\lambda^*|}{n} = O_p(n^{-\frac{1}{2}}) \).

Fourthly, before we can have an expression for \( \lambda^* \), we have to show \( \mu_i^* - \mu_i = o_p(1) \). From (EC.6) it follows that

\[
\bar{X}_i - \mu_i^* = \sum_{j=1}^{n_i} \left(\frac{1}{n_i} - w_{i,j}\right) X_{i,j}
\]

\[
= \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\lambda_{i,j}^*}{n_i} (X_{i,j} - \mu_i^*) X_{i,j}.
\]

(EC.12)

We have shown in the third step that \( Z_n = o(n^{\frac{1}{2}}) \) and \( \frac{|\lambda^*|}{n} = O_p(n^{-\frac{1}{2}}) \), hence max \( \left|\frac{\lambda_{i,j}^*}{n_i} (X_{i,j} - \mu_i^*)\right| \leq \frac{|2\lambda^*|}{cn} Z_n = O_p(n^{-\frac{1}{2}}) o(n^{\frac{1}{2}}) = o_p(1) \), therefore

\[
|\bar{X}_i - \mu_i^*| \leq \frac{1}{n_i} \sum_{j=1}^{n_i} \left|\frac{\lambda_{i,j}^*}{n_i} (X_{i,j} - \mu_i^*)\right| |X_{i,j}| \leq \frac{1}{n_i} \sum_{j=1}^{n_i} \left|\frac{o_p(1)}{1 - \min(o_p(1), 1)}\right| |X_{i,j}| \leq \left|\frac{o_p(1)}{1 - \min(o_p(1), 1)}\right| \frac{1}{n_i} \sum_{j=1}^{n_i} |X_{i,j}| = o_p(1).
\]
On the other hand, $\bar{X}_i - \mathbb{E}X_i = O_p(n^{-\frac{1}{2}})$ by central limit theorem, hence $\mu^*_i - \mathbb{E}X_i = o_p(1)$.

As the fifth step, we derive the expression for $\lambda^*$. Rewrite (EC.7) as

$$
\bar{X}_i - \mu^*_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \left[ \frac{\lambda^*}{n_i} (X_{i,j} - \mu^*_i)^2 - \frac{(\frac{\lambda^*}{n_i})^2 (X_{i,j} - \mu^*_i)^3}{1 + \frac{n}{n_i} (X_{i,j} - \mu^*_i)} \right],
$$

$$
= \frac{\lambda^*}{n_i} \left[ \frac{1}{n_i} \sum_{j=1}^{n_i} (X_{i,j} - \mu^*_i)^2 \right] - \frac{\lambda^*}{n_i} \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{(X_{i,j} - \mu^*_i)^3}{1 + \frac{n}{n_i} (X_{i,j} - \mu^*_i)}.
$$

(EC.13)

The last term in (EC.13) can be bounded as

$$
\left| \frac{\lambda^*}{n_i} \frac{1}{n_i} \sum_{j=1}^{n_i} |X_{i,j} - \mu^*_i|^3 \right| \leq \left| \frac{\lambda^*}{n_i} \right|^2 \frac{2Z_n}{1 - \min(o_p(1), 1)} \cdot \frac{1}{n_i} \sum_{j=1}^{n_i} |X_{i,j} - \mu^*_i|^2.
$$

$$
= O_p(n^{-1})o_p(n^{\frac{1}{2}})O_p(1)
$$

$$
= o_p(n^{-\frac{1}{2}})
$$

where in passing from $\frac{1}{n_i} \sum_{j=1}^{n_i} |X_{i,j} - \mu^*_i|^2$ to $O_p(1)$ we use $\mu^*_i - \mathbb{E}X_i = o_p(1)$. And the first term in (EC.13)

$$
\frac{1}{n_i} \sum_{j=1}^{n_i} (X_{i,j} - \mu^*_i)^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (X_{i,j} - \mathbb{E}X_i)^2 + 2(\bar{X}_i - \mathbb{E}X_i)(\mathbb{E}X_i - \mu^*_i) + (\mathbb{E}X_i - \mu^*_i)^2
$$

$$
= \hat{\sigma}_i^2 + o_p(1)
$$

$$
= \sigma_i^2 + o_p(1)
$$

(EC.14)

where $\mu^*_i - \mathbb{E}X_i = o_p(1)$ is used again. Hence (EC.13) can be written as

$$
\bar{X}_i - \mu^*_i = \frac{\lambda^*}{n_i} \sigma_i^2 + o_p(n^{-\frac{1}{2}}).
$$

(EC.15)

Summing (EC.15) over $i = 1, \ldots, m$ and using (EC.5) gives

$$
\sum_{i=1}^{m} \bar{X}_i = \lambda^* \sum_{i=1}^{m} \frac{\sigma_i^2}{n_i} + o_p(n^{-\frac{1}{2}}).
$$

Therefore the expression for $\lambda^*$ is

$$
\lambda^* = \frac{\sum_{i=1}^{m} \bar{X}_i + o_p(n^{-\frac{1}{2}})}{\sum_{i=1}^{m} \frac{\sigma_i^2}{n_i}}.
$$

(EC.16)

Finally, we plug in the expression of $\lambda^*$ to derive the leading term of the taylor expansion of $-2 \log R$ and conclude the convergence.

$$
\log(1 + \frac{\lambda^*}{n_i} (X_{i,j} - \mu^*_i)) = \frac{\lambda^*}{n_i} (X_{i,j} - \mu^*_i) - \frac{\lambda^*^2}{2n_i^2} (X_{i,j} - \mu^*_i)^2 + \eta_{i,j},
$$
where $\eta_{i,j} = \frac{1}{3(1 - \theta_{i,j} \frac{\lambda}{n_i} (X_{i,j} - \mu_i^*))^2} \left( \frac{\lambda}{n_i} (X_{i,j} - \mu_i^*) \right)^3$, $\theta_{i,j} \in (0, 1)$, so the log empirical likelihood ratio can be expressed as

$$-2 \log R = 2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log \left(1 + \frac{\lambda^*}{n_i} (X_{i,j} - \mu_i^*) \right)$$

$$= 2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left( \frac{\lambda^*}{n_i} (X_{i,j} - \mu_i^*) - \frac{\lambda^*^2}{2n_i^2} (X_{i,j} - \mu_i^*)^2 + \eta_{i,j} \right)$$

$$= 2 \sum_{i=1}^{m} \lambda^* (\bar{X}_i - \mu_i^*) - \sum_{i=1}^{m} \frac{\lambda^*^2}{n_i} \cdot \frac{1}{n_i} \sum_{j=1}^{n_i} (X_{i,j} - \mu_i^*)^2 + \sum_{i=1}^{m} \sum_{j=1}^{n_i} 2\eta_{i,j} \quad (\text{EC.17})$$

$$= 2 \lambda^* \sum_{i=1}^{m} \bar{X}_i - \sum_{i=1}^{m} \frac{\lambda^*^2}{n_i} (\sigma_i^2 + o_p(1)) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} 2\eta_{i,j} \quad (\text{EC.18})$$

The equality between (EC.17) and (EC.18) follows from (EC.5) and (EC.14). To bound the last term in (EC.18)

$$\left| \sum_{i,j} 2\eta_{i,j} \right| \leq \frac{C}{(1 - \min(o_p(1), 1))^3} \left| \frac{\lambda^*}{n} \right|^3 nZ_n \sum_{i=1}^{m} \frac{1}{n_i} \sum_{j=1}^{n_i} (X_{i,j} - \mu_i^*)^2$$

for some constant $C > 0$

$$= O_p(1) O_p(n^{-\frac{3}{2}}) n o_p(n^{-\frac{1}{2}}) O_p(1)$$

$$= o_p(1).$$

Hence using the above estimates and (EC.16), the log likelihood ratio (EC.18) becomes

$$-2 \log R = 2 \lambda^* \sum_{i=1}^{m} \bar{X}_i - \lambda^*^2 \sum_{i=1}^{m} \frac{\sigma_i^2}{n_i} + o_p(1)$$

$$= \frac{(\sum_{i=1}^{m} \bar{X}_i)^2}{\sum_{i=1}^{m} \frac{\sigma_i^2}{n_i}} + o_p(1). \quad (\text{EC.19})$$

So it remains to show that the leading term in (EC.19) is asymptotically $\chi^2$. The leading term can be written as

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{(X_{i,j} - \bar{X}_i)^2}{n_i} \left( \sum_{i=1}^{m} \frac{\sigma_i^2}{n_i} \right)^{-1} \cdot (\text{EC.20})$$

Here we use $\sum_{i=1}^{m} \bar{X}_i = 0$ again. It suffices to show that the sum in (EC.20) is asymptotically standard normal. We apply the Lindeberg-Feller theorem to

$$Y_{n,k} = \frac{X_{I(k), J(k)} - \bar{X}_I(k)}{n_i \sqrt{\sum_{i=1}^{m} \frac{\sigma_i^2}{n_i}}}$$

where $I(k)$ runs through $1, \ldots, m$, and $J(k)$ runs through $1, \ldots, n_i$ for each $i$. The independence and mean zero conditions are obviously met, and

$$\sum_{k=1}^{n} EY_{n,k}^2 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} E \left[ \frac{(X_{i,j} - \bar{X}_i)^2}{n_i^2 \sum_{i=1}^{m} \frac{\sigma_i^2}{n_i}} \right] = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{\sigma_i^2}{n_i} \sum_{i=1}^{m} \frac{\sigma_i^2}{n_i} = 1.$$
For any $\epsilon > 0$,

$$
\sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathbb{E} \left[ \left( \frac{X_{i,j} - \mathbb{E} X_i}{\sqrt{\sum_{i=1}^{m} \sigma_i^2 / n_i}} \right)^2 ; \left| \frac{X_{i,j} - \mathbb{E} X_i}{\sqrt{\sum_{i=1}^{m} \sigma_i^2 / n_i}} \right| > \epsilon \right] = \sum_{i=1}^{m} n_i \mathbb{E} \left[ \left( \frac{X_{i,1} - \mathbb{E} X_i}{\sqrt{\sum_{i=1}^{m} \sigma_i^2 / n_i}} \right)^2 ; \left| \frac{X_{i,1} - \mathbb{E} X_i}{\sqrt{\sum_{i=1}^{m} \sigma_i^2 / n_i}} \right| > \epsilon \right]
$$

$$
\leq \sum_{i=1}^{m} C_i \mathbb{E} \left[ (X_{i,1} - \mathbb{E} X_i)^2 ; |X_{i,1} - \mathbb{E} X_i| > \epsilon C_2 \sqrt{n} \right] \rightarrow 0.
$$

The convergence to zero follows from dominated convergence theorem. Therefore the assumptions of the Lindeberg-Feller theorem hold for $Y_{n,k}$, and convergence to $X_2^1$ holds, as $n_i \rightarrow \infty$. \hfill \square

**Proof of Corollary 2.** The proof is the same as that of Theorem 4 except the constraint $\sum_{i=1}^{m} \mu_i = 0$ in (EC.1) is replaced by $\sum_{i=1}^{m} \mu_i = \sum_{i=1}^{m} \bar{X}_i + z \sqrt{\sum_{i=1}^{m} \sigma_i^2 / n_i}$. \hfill \square

### EC.2. Proofs of Propositions 4 and 5

We will separate the entire proof into two parts. The first part deals with the linearization error and statistical error due to input data, which includes the first half of Proposition 4 and Proposition 5, whereas the second part focuses on the simulation error in sample average approximation, namely the second half of Proposition 4.

#### EC.2.1. Part 1: Linearization Error and Statistical Error

We prove two lemmas first.

**Lemma EC.1.** All feasible solution $(w_1, \ldots, w_m) \in \mathcal{A}_\alpha$ satisfies

$$
\frac{l(\alpha)}{n_i} \leq w_{i,j} \leq \frac{u(\alpha)}{n_i}, \forall i = 1, \ldots, m, j = 1, \ldots, n_i,
$$

where $0 < l(\alpha) < 1 < u(\alpha) < +\infty$ are the two solutions of the equation $xe^{1+\frac{x^2}2} - x = 1$.

**Proof of Lemma EC.1.** By Jensen’s inequality, for each $i$

$$
- \sum_{j=1}^{n_i} \frac{1}{n_i} \log n_i w_{i,j} \geq - \log \sum_{j=1}^{n_i} w_{i,j} = 0,
$$

thus

$$
-2 \sum_{j=1}^{n_i} \log n_i w_{i,j} \leq -2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log n_i w_{i,j} \leq \chi^2_{1,1-\alpha}.
$$

Hence

$$
\prod_{j} n_i w_{i,j} \geq e^{-\frac{\chi^2_{1,1-\alpha}}{2}}. \quad (EC.21)
$$
For any $s = 1, \ldots, n_i$, we take $n_i w_{i,s}$ out of the product and then apply the inequality $x_1 \cdots x_k \leq \left( \frac{x_1 + \cdots + x_k}{k} \right)^k$, $x_i \geq 0$ to $\prod_{j \neq s} n_i w_{i,j}$ to get

$$n_i w_{i,s} \left( 1 + \frac{1 - n_i w_{i,s}}{n_i - 1} \right)^{n_i - 1} \geq n_i w_{i,s} \prod_{j \neq s} n_i w_{i,j} \geq e^{-\frac{\chi^2_{1,1-\alpha}}{2}}.$$ 

Applying $e^x \geq 1 + x$ to $1 + \frac{1 - n_i w_{i,s}}{n_i - 1}$ gives

$$n_i w_{i,s} e^{1 - n_i w_{i,s}} \geq e^{-\frac{\chi^2_{1,1-\alpha}}{2}}.$$  \(\text{(EC.22)}\)

Some simple calculations show that the function $xe^{1-x}$ strictly increases over $(0, 1)$ and decreases over $(1, +\infty)$. So it follows from (EC.22) that $n_i w_{i,s}$ must fall between the two solutions of $xe^{1-x} = e^{-\frac{\chi^2_{1,1-\alpha}}{2}}$.

\[
\text{Lemmas EC.2.} \quad \text{Let } l(\alpha), u(\alpha) \text{ be the constants in Lemma EC.1. Any feasible solution } (w_1, \ldots, w_m) \in \mathcal{W}_\alpha \text{ satisfies}
\]

$$\sum_{i=1}^{m} n_i^2 \sum_{j=1}^{n_i} \left( w_{i,j} - \frac{1}{n_i} \right)^2 \leq u(\alpha)^2 \chi^2_{1,1-\alpha}. \quad \text{EC.1}$$

Moreover, for those such that $-2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log n_i w_{i,j} = \chi^2_{1,1-\alpha}$, the following lower bound holds

$$\sum_{i=1}^{m} n_i^2 \sum_{j=1}^{n_i} \left( w_{i,j} - \frac{1}{n_i} \right)^2 \geq l(\alpha)^2 \chi^2_{1,1-\alpha}. \quad \text{EC.2}$$

\[
\text{Proof of Lemma EC.2.} \quad \text{Expand the log-likelihood ratio constraint at the uniform distributions, and use mean value theorem}
\]

$$-2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log n_i w_{i,j} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left( 0 - 2 n_i (w_{i,j} - \frac{1}{n_i}) + (\theta_{i,j} w_{i,j} + (1 - \theta_{i,j}) \frac{1}{n_i})^{-2} (w_{i,j} - \frac{1}{n_i})^2 \right)$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\theta_{i,j} w_{i,j} + (1 - \theta_{i,j}) \frac{1}{n_i})^{-2} (w_{i,j} - \frac{1}{n_i})^2.$$

where $0 \leq \theta_{i,j} \leq 1$ for each $i, j$. Lemma EC.1 implies $\frac{l(\alpha)}{n_i} \leq \theta_{i,j} w_{i,j} + (1 - \theta_{i,j}) \frac{1}{n_i} \leq \frac{u(\alpha)}{n_i}$, hence

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{n_i^2}{u(\alpha)^2} \left( w_{i,j} - \frac{1}{n_i} \right)^2 \leq \chi^2_{1,1-\alpha}.$$  \(\text{EC.1}\)

Moving $u(\alpha)^2$ to the right side gives the upper bound. The lower bound can be justified by a similar argument involving $l(\alpha)$.

Now we start to prove the propositions.
Proof of the first half of Proposition 4 and Proposition 5. We will prove the upper bound for $|Z - Z_L|^2$ and $|Z - \hat{Z}_L|^2$, then the same bound automatically applies to $|Z_L - \hat{Z}_L|^2$ thanks to the following inequality

$$
\sup |Z_L - \hat{Z}_L|^2 \leq \sup \left( 2 |Z - Z_L|^2 + 2 |Z - \hat{Z}_L|^2 \right) \leq 2 \sup |Z - Z_L|^2 + 2 \sup |Z - \hat{Z}_L|^2.
$$

The entire proof is divided into two parts. We will first show the uniform error bound of the linear approximation at the true distributions, and then that of the linear approximation at the empirical distributions. We start the analysis with expressing $Z(w_1, \ldots, w_m)$ as

$$
Z(w_1, \ldots, w_m) = \int h(x_1, \ldots, x_m) \prod_{i=1}^{m} \prod_{t=1}^{T_i} dw_i(x_{i,t}).
$$

(RC.23)

Rewrite $dw_i$ as $d(w_i - P_i^{n_i} + P_i^{n_i} - P_i + P_i)$, where $P_i^{n_i}$ is the empirical distribution of the $i$-th sample, and distribute $w_i - P_i^{n_i}$, $P_i^{n_i} - P_i$ and $P_i$ in (RC.23)

$$
Z(w_1, \ldots, w_m)
= \sum_{\mathcal{T}_1, \mathcal{T}_2} \int \prod_{i=1}^{m} \prod_{t \notin \mathcal{T}_1 \cup \mathcal{T}_2} dP_i(x_{i,t}) \prod_{i=1}^{m} \prod_{t \in \mathcal{T}_1} d(P_i^{n_i} - P_i)(x_{i,t}) \prod_{i=1}^{m} \prod_{t \in \mathcal{T}_2} d(w_i - P_i^{n_i})(x_{i,t})
= \sum_{d=0}^{T} \sum_{|\mathcal{T}_1|+|\mathcal{T}_2|=d} \int \prod_{i=1}^{m} \prod_{t \notin \mathcal{T}_1 \cup \mathcal{T}_2} dP_i(x_{i,t}) \prod_{i=1}^{m} \prod_{t \in \mathcal{T}_1} d(P_i^{n_i} - P_i)(x_{i,t}) \prod_{i=1}^{m} \prod_{t \in \mathcal{T}_2} d(w_i - P_i^{n_i})(x_{i,t})
$$

(AC.24)

where for each $i$, $\mathcal{T}_1, \mathcal{T}_2$ are two disjoint ordered (possibly empty) subsets of $\{1, 2, \ldots, T_i\}$ that specifies the second subscript $t$ of the argument $x_{i,t}$, $|\cdot|$ denotes the cardinality of a set, and $T = \sum_{i=1}^{m} T_i$.

The desired conclusion can be achieved upon completing the following three tasks: (1) show that the terms with $d = 0, 1$ above give the linear approximation, (2) terms with $d = 2$ constitute the leading error term that is of the desired order of magnitude, (3) all others with $d \geq 3$ are negligible when input sizes are sufficiently large.

**Task one:** $d = 0, 1$

The only summand with $d = 0$ is

$$
\int h(x_1, \ldots, x_m) \prod_{i=1}^{m} \prod_{t \in \mathcal{T}_i} dP_i(x_{i,t}) = Z(P_1, \ldots, P_m),
$$

and summands with $d = 1$ are all of the following type

$$
\int h(x_1, \ldots, x_m) \prod_{i \neq r \text{ or } t \neq s} dP_r(x_{i,t})d(P_r^{n_r} - P_r)(x_{r,s}), \text{ for } r = 1, \ldots, m, s = 1, \ldots, T_i
$$

$$
\int h(x_1, \ldots, x_m) \prod_{i \neq r \text{ or } t \neq s} dP_r(x_{i,t})d(w_r - P_r^{n_r})(x_{r,s}), \text{ for } r = 1, \ldots, m, s = 1, \ldots, T_i
$$
which sum up to
\[
\sum_{r=1}^{m} \sum_{s=1}^{T_i} \int h(x_1, \ldots, x_m) \prod_{i \neq r \text{ or } t \neq s} dP_r(x_{i,t})d(w_r - P_r)(x_{r,s}) = \sum_{r=1}^{m} \sum_{j=1}^{n_i} w_{r,j}G_r(X_{r,j}).
\]

This concludes that the summands with \(d = 0\) sum up to the linear approximation
\[
Z(P_1, \ldots, P_m) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{i,j}G_i(X_{i,j}).
\]

Now we start to deal with terms with \(d \geq 2\). Since second moments of \(h\) will frequently pop up, we denote their upper bound by
\[
\mathcal{M} = \max_{i_1, \ldots, i_m} \mathbb{E}_{P_1, \ldots, P_m} [h(X_{1,i_1}, \ldots, X_{m,i_m})^2],
\]
where each \(I_i \in \{1, 2, \ldots, T_i\}\). Note that \(\mathcal{M}\) is finite under Assumption 2 with \(k = 2\).

**Task two:** \(d = 2\)

For any two input arguments \(x_{i,t}, x_{i',t'}\) of the function \(h\) such that \(1 \leq i, i' \leq m, 1 \leq t \leq T_i, 1 \leq t' \leq T_{i'}\), we denote the conditional expectation
\[
h(i,t)(i',t')(x,y) = \mathbb{E}_{P_1, \ldots, P_m}[h(X_{1,i}, \ldots, X_{m,i})|X_{i,t} = x, X_{i,t'} = y]
\]
\[
\hat{h}(i,t)(i',t')(x,y) = h_i(x,y) - \mathbb{E}_{X_{i'}}[h(i,t)(i',t')(x,X_{i'})] - \mathbb{E}_{X_i}[h(i,t)(i',t')(X_i,y)] + Z(P_1, \ldots, P_m).
\]

We write \((i,t) < (i',t')\) if \(i < i'\) or \(i = i'\) and \(t < t'\). Terms with \(d = 2\) in \([EC.24]\) can be grouped into the following three sums
\[
R_{pp} = \sum_{(i,t) < (i',t')} \int h(i,t)(i',t')(x,y)d(P_{i_1}^{n_1} - P_i)(x)d(P_{i'}^{n'} - P_{i'})(y)
\]
\[
R_{wp} = \sum_{(i,t) \neq (i',t')} \int h(i,t)(i',t')(x,y)d(w_i - P_i^{n_i})(x)d(P_{i'}^{n'} - P_{i'})(y)
\]
\[
R_{ww} = \sum_{(i,t) < (i',t')} \int h(i,t)(i',t')(x,y)d(w_i - P_i^{n_i})(x)d(w_{i'} - P_{i'}^{n'}) (y).
\]

If we distribute the integral in \(R_{pp}\) as \(dP_i dP_{i'}^{n'} - dP_i^{n_i} dP_{i'} - dP_i dP_{i'}^{n'} + dP_i dP_{i'}\), we get
\[
R_{pp} = \sum_{(i,t) < (i',t')} \sum_{j \leq n_i, j' \leq n_{i'}} \frac{1}{n_i n_{i'}} \hat{h}(i,t)(i',t')(X_{i,j}, X_{i',j'}).\]

Note that for any integral in \(R_{wp}\), \(R_{ww}\), inserting any function that depends on only \(x\) or \(y\) into the integrand does not change its value, because the measure governing each argument has a zero sum. Therefore one can replace the function \(h(i,t)(i',t')\) there by \(\hat{h}(i,t)(i',t')\), and arrive at
\[
R_{wp} = \sum_{(i,t) \neq (i',t')} \sum_{j \leq n_i, j' \leq n_{i'}} (w_{i,j} - \frac{1}{n_i}) \frac{1}{n_{i'}} \hat{h}(i,t)(i',t')(X_{i,k}, X_{i',k'})\]
\[
R_{ww} = \sum_{(i,t) < (i',t')} \sum_{j \leq n_i, j' \leq n_{i'}} (w_{i,j} - \frac{1}{n_i})(w_{i',j'} - \frac{1}{n_{i'}}) \hat{h}(i,t)(i',t')(X_{i,j}, X_{i',j'}).\]
Our goal is to show

\[
\mathbb{E}[\sup_{x \in \alpha} |R_{pp}|^2] \leq C(\alpha) \left( \sum_{i \neq i'} \frac{M_{i,i'}}{n_i n_{i'}} + \sum_i \frac{M_i}{n_i^2} \right). \tag{EC.26}
\]

Since the proofs are a bit involved, it would take up too much space to put down all three of them. We'll detail the proof for \( R_{wp} \) only, and then highlight the similar ideas and techniques for the other two. The proof hinges on the following property of \( \hat{h}_{(i,t)(i',t')} \)

\[
\mathbb{E}_{X_i} \left[ \hat{h}_{(i,t)(i',t')}(X_i, y) \right] = 0, \text{ for any } y, \quad \mathbb{E}_{X_{i'}} \left[ \hat{h}_{(i,t)(i',t')}(x, X_{i'}) \right] = 0, \text{ for any } x. \tag{EC.27}
\]

We rewrite the sum \( R_{wp} \) with the outer layer of summation going through the input data points, i.e.

\[
R_{wp} = \sum_{i,j \leq n_i} n_i (w_{i,j} - \frac{1}{n_i}) \sum_{t, (i', t')} \frac{1}{n_i n_{i'}} \hat{h}_{(i,t)(i',t')} (X_{i,j}, X_{i',j'})
\]

where we leave out the requirement that \( (i, t) \neq (i', t') \) and use the default \( \hat{h}_{(i,t_i),(i,t_i)} \equiv 0 \). By Cauchy Schwartz inequality and Lemma \( \text{EC.2} \)

\[
\sup_{x \in \alpha} |R_{wp}|^2 \leq \sum_{i,j \leq n_i} n_i^2 (w_{i,j} - \frac{1}{n_i})^2 \left( \sum_{t, (i', t')} \frac{1}{n_i n_{i'}} \hat{h}_{(i,t)(i',t')} (X_{i,j}, X_{i',j'}) \right)^2 \tag{EC.28}
\]

\[
\leq C(\alpha) \sum_{i,j \leq n_i} \left( \sum_{t, (i', t')} \frac{1}{n_i n_{i'}} \hat{h}_{(i,t)(i',t')} (X_{i,j}, X_{i',j'}) \right)^2. \tag{EC.29}
\]

Therefore it finally boils down to bounding the expectation of

\[
SR_{wp}^{i,j} = \left( \sum_{t, (i', t')} \frac{1}{n_i n_{i'}} \hat{h}_{(i,t)(i',t')} (X_{i,j}, X_{i',j'}) \right)^2, \text{ for each } i, j. \tag{EC.29}
\]

In view of \( \text{EC.27} \), one can expect that most terms in the expansion of the above square vanish upon taking expectation. The expansion is

\[
SR_{wp}^{i,j} = \sum_{t, (i', t'), j', (i', t'), j''} \frac{1}{n_i^2 n_{i'} n_{j'}} \hat{h}_{(i,t)(i',t')} (X_{i,j}, X_{i',j'}) \hat{h}_{(i',t')(i'',t'')} (X_{i',j}, X_{i'',j''})
\]

Due to \( \text{EC.27} \), the product of two \( \hat{h} \) functions in the expansion has non-zero expectation only if \( i' = i'', j' = j'' \). Hence

\[
\mathbb{E}[SR_{wp}^{i,j}] = \sum_{t, (i', t'), j', (i', t'), j''} \frac{1}{n_i^2 n_{i'} n_{j'}} \mathbb{E}[\hat{h}_{(i,t)(i',t')} (X_{i,j}, X_{i',j''}) \hat{h}_{(i',t')(i'',t'')} (X_{i',j}, X_{i'',j'})]
\]
\begin{align*}
&= ( \sum_{(i',j') = (i,j)} \sum_{t, t'} \frac{1}{n_i n_{j'}} \mathbb{E} \left[ \tilde{h}_{(i,t),(i',t')} (X_{i,j}, X_{i',j}) \right] + \sum_{(i',j') \neq (i,j)} \sum_{t, t'} \frac{1}{n_i n_{j'}} \mathbb{E} \left[ \tilde{h}_{(i,t),(i',t')} (X_{i,j}, X_{i',j}) \right] ) \\
&= \sum_{t, t'} \sum_{i, i'} \frac{1}{n_i n_{j'}} \mathbb{E} \left[ \tilde{h}_{(i,t),(i',t')} (X_{i,j}, X_{i',j}) \right] + \sum_{i, i'} \frac{1}{n_i n_{j'}} \mathbb{E} \left[ \tilde{h}_{(i,t),(i',t')} (X_{i,j}, X_{i',j}) \right].
\end{align*}

Starting from [EC:25], one can show, by the law of total variance and Young’s inequality $2ab \leq a^2 + b^2$, that $\mathbb{E}_{X_i} [\tilde{h}_{(i,t),(i',t')} (X_{i,j}, X_{i})]^2 \leq 16M$ for all $i, t, t'$. Hence applying Cauchy Schwartz inequality to the expectations in the first sum above gives

\begin{align*}
\mathbb{E} [SR_{wp}^{ij}] &\leq 16T^4 M \frac{1}{n_i} + \sum_{(i',j') \neq (i,j)} \sum_{t, t'} \frac{1}{n_i n_{j'}} \mathbb{E} \left[ \tilde{h}_{(i,t),(i',t')} (X_{i,j}, X_{i',j}) \tilde{h}_{(i,t),(i',t')} (X_{i,j}, X_{i',j}) \right] \\
&\leq 16T^4 M \frac{1}{n_i} + \sum_{i, i'} \frac{1}{n_i n_{i'}} \mathbb{E} \left[ \tilde{h}_{(i,t),(i',t')} (X_{i,j}, X_{i',j}) \tilde{h}_{(i,t),(i',t')} (X_{i,j}, X_{i',j}) \right] \\
&= 16T^4 M \frac{1}{n_i} + \sum_{i, i'} \frac{1}{n_i n_{i'}} \mathbb{E} \left[ \tilde{h}_{(i,t),(i',t')} (X_{i,j}, X_{i',j}) \tilde{h}_{(i,t),(i',t')} (X_{i,j}, X_{i',j}) \right],
\end{align*}

where $X_{i}, \tilde{X}_{i'}$ are independent and distributed under $P_i, P_{i'}$ respectively ($i, i'$ can be the same). Now we are able to sum up $\mathbb{E} [SR_{wp}^{ij}]$ over $i, j$ to get

\begin{align*}
&\frac{1}{C(\alpha)} \mathbb{E} [\sup_{\theta_{\alpha}} R_{wp}^2] \\
&\leq \sum_{i, j \leq n_i} \mathbb{E} [SR_{wp}^{ij}] \\
&\leq \sum_i 16T^4 M \frac{1}{n_i^3} + \sum_{i, i'} \frac{1}{n_i n_{i'}} \mathbb{E} \left[ \tilde{h}_{(i,t),(i',t')} (X_{i,j}, \tilde{X}_{i'}) \tilde{h}_{(i,t),(i',t')} (X_{i,j}, \tilde{X}_{i'}) \right] \\
&\leq O \left( \sum_i n_i^{-3} \right) + \sum_{i \neq i'} \frac{1}{n_i n_{i'}} \mathbb{E} \left[ \tilde{h}_{(i,t),(i',t')} (X_{i,j}, \tilde{X}_{i'}) \sum_{i, i'} \tilde{h}_{(i,t),(i',t')} (X_{i,j}, \tilde{X}_{i'}) \right] \\
&+ \sum_{i} \frac{1}{n_i^2} \mathbb{E} \left[ \tilde{h}_{(i,t),(i,t')} (X_{i,j}, \tilde{X}_{i}) \sum_{i, i'} \tilde{h}_{(i,t),(i',t')} (X_{i,j}, \tilde{X}_{i}) \right] \\
&\leq O \left( \sum_i n_i^{-3} \right) + \sum_{i \neq i'} \frac{1}{n_i n_{i'}} \mathbb{E} \left[ \left( \sum_{i, t} \tilde{h}_{(i,t),(i',t')} (X_{i,j}, \tilde{X}_{i'}) \right)^2 \right] + \sum_i \frac{1}{n_i^3} \mathbb{E} \left[ \left( \sum_{i, t} \tilde{h}_{(i,t),(i',t')} (X_{i,j}, \tilde{X}_{i}) \right)^2 \right] \\
&\leq \sum_{i \neq i'} \frac{1}{n_i n_{i'}} M_{i,i'} + \sum_i \frac{1}{n_i^2} M_i + O \left( \sum_i n_i^{-3} \right).
\end{align*}

Here the substitution in the last step comes from the definition of $M_{i,i'}, M_i$. This completes the proof for $R_{wp}$. Proofs for $R_{pp}, R_{ww}$ are similar. To highlight the main strategy, note that the key steps in the above proof are (1) first use Cauchy Schwartz inequality to separate the factors that
depend on $w_i$'s from those that don't (in (EC.28)), so that the supremum can be eliminated thanks to Lemma EC.2 (2) then upper bound the expectations of the squared terms (in (EC.29)) using the property (EC.27). For $R_{ww}$, we will separate $(w_{i,j} - \frac{1}{n_i}) (w_{i',j'} - \frac{1}{n_{i'}})$ from $\sum_{i,t'} h_{i,i'} (X_{i,j} X_{i',j'})$, and then the problem boils down to bounding expectation of the square of the latter. Whereas for $R_{pp}$, nothing depends on $w_i$, so we simply square the entire sum and bound its expectation with the help of property (EC.27), without using Cauchy Schwartz inequality first. It turns out that they are all bounded above by the same quantity (EC.26) (up to the factor $C(\alpha)$).

We would also like to mention that the bound (EC.26) is in fact tight. First note that all the inequalities in the proof for $R_{wp}$, except the Cauchy Schwartz inequality (EC.28), are effectively equalities as far as the leading term is concerned. This tightness remains true when bounding $R_{pp}$ and $R_{ww}$. Since Cauchy Schwartz inequality is not used to bound $R_{pp}$, (EC.26) is tight for $R_{pp}$, on one hand. On the other hand, when uniform weights $w_{i,j} = 1/n_i$ are taken, $R_{wp}, R_{ww}$ will vanish and $R_{pp}$ will be the only second order term. Hence (EC.26) is tight for $|Z - Z_L|$ as well.

**Task three:** $d \geq 3$

Now we deal with the summand for which $d \geq 3$. Theoretically speaking, tight bounds like (EC.26) can be derived for higher order remainders as well. However, we limit out goal to showing that they are negligible when compared with the second order remainder. So the proof will be similar to the case $d = 2$ but the bounds will be loose. We focus on a generic summand of the form

\[ R(\mathcal{T}^1, \mathcal{T}^2) = \int h(x_1, \ldots, x_m) \prod_{i=1}^m \prod_{t \in \mathcal{T}^1 \cup \mathcal{T}^2} dP_i(x_{i,t}) \prod_{i=1}^m \prod_{t \in \mathcal{T}^1} d(P_i^n_i - P_i)(x_{i,t}) \prod_{i=1}^m \prod_{t \in \mathcal{T}^2} d(w_i - P_i^n_i)(x_{i,t}), \]

(EC.30)

where $\mathcal{T}^1 = (\mathcal{T}^1_1, \ldots, \mathcal{T}^1_m)$, $\mathcal{T}^2 = (\mathcal{T}^2_1, \ldots, \mathcal{T}^2_m)$. Each $\mathcal{T}^1_i$ is a set of second subscripts of $x_{i,t}$ arranged in ascending order. So is each $\mathcal{T}^2_i$ and $\mathcal{T}^1_i \cap \mathcal{T}^2_i = \emptyset$. It is prescribed that $\sum_{i=1}^m (|\mathcal{T}^1_i| + |\mathcal{T}^2_i|) = d$. We use $\mathcal{T}^1_i(t)$ (or $\mathcal{T}^2_i(t)$) to denote its $t$-th element. Our goal is to show that

\[ \mathbb{E}\left[ \sup_{\alpha} |R(\mathcal{T}^1, \mathcal{T}^2)|^2 \right] = O\left( \prod_{i=1}^m n_i^{-(|\mathcal{T}^1_i| + |\mathcal{T}^2_i|)} \right). \]

(EC.31)

First, we separate factors in the form of $(w_{i,j} - 1/n_i)$ from others. Define conditional expectations of $h$ for given subscripts $\mathcal{T}^1 = (\mathcal{T}^1_1, \ldots, \mathcal{T}^1_m), \mathcal{T}^2 = (\mathcal{T}^2_1, \ldots, \mathcal{T}^2_m)$

\[ h_{\mathcal{T}^1, \mathcal{T}^2}(x_1, \ldots, x_m) = \mathbb{E}_{P_1, \ldots, P_m} \left[ h(X_1, \ldots, X_m) \left| X_i(\mathcal{T}^1_i(t)) = x_{i,t}, \forall i, 1 \leq t \leq |\mathcal{T}^1_i| \right. \right. \]

\[ \left. \left. X_i(\mathcal{T}^2_i(t - |\mathcal{T}^1_i|)) = x_{i,t}, \forall i, |\mathcal{T}^1_i| + 1 \leq t \leq |\mathcal{T}^1_i| + |\mathcal{T}^2_i| \right] \right], \]

where $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,|\mathcal{T}^1_i|+|\mathcal{T}^2_i|})$ whose length is equal to the number of variables of the $i$-th input that are conditioned on. By distributing the product $\prod_{i=1}^m \prod_{t \in \mathcal{T}^1_i \cup \mathcal{T}^2_i} dP_i(x_{i,t}) \prod_{i=1}^m \prod_{t \in \mathcal{T}^1_i} d(P_i^n_i - P_i)(x_{i,t})$, (EC.30) can be expressed as

\[ R(\mathcal{T}^1, \mathcal{T}^2) = \int h_{\mathcal{T}^1, \mathcal{T}^2}(x_1, \ldots, x_m) \prod_{i=1}^m \prod_{t \in \mathcal{T}^1_i} dP_i^n_i(x_{i,t}) \prod_{i=1}^m \prod_{t = |\mathcal{T}^1_i| + 1} d(w_i - P_i^n_i)(x_{i,t}), \]

(EC.32)
where

\[
\tilde{h}_{1_1, 1_2}(x_1, \ldots, x_m) = \sum_{\tilde{T}_1 < \tilde{T}_1} (-1)^{\|\tilde{T}_1\|_1} \mathbb{E}_{P_1 \ldots P_m} [h_{1_1, 1_2}(X_1, \ldots, X_m) | X_i(t) = x_{i,t}, \forall i, t \text{ s.t. } t > |T_1|] \text{ or } \tilde{T}_1(t) \notin \tilde{T}_1.
\]  

(EC.33)

Now let \( X_{i,j} = i, \ldots, m, j = 1, \ldots, n_i \) be the data, and for each \( i \) let

\[
J_i^1 = (J_i^1(1), \ldots, J_i^1(|T_1^i|)) \in \{1, 2, \ldots, n_i\}^{T_1^i},
\]

\[
J_i^2 = (J_i^2(1), \ldots, J_i^2(|T_2^i|)) \in \{1, 2, \ldots, n_i\}^{T_2^i}
\]

be two sequences of indices (if \( T_1^i \) or \( T_2^i \) is empty, then \( J_i^1 \) or \( J_i^2 \) is empty accordingly) that specify the second subscript of data \( X_{i,j} \), then the summation version of (EC.32) is

\[
R(T^1, T^2) = \sum_{J_1^1, J_1^2} \left[ \prod_{j=1}^{m} \left( w_{i,j_1(j)} - \frac{1}{n_i} \right) \right] \sum_{J_1^1, J_1^2} \left[ \prod_{i=1}^{n_i} \frac{1}{|T_1^i|} \sum_{J_1^1, J_1^2} \tilde{h}_{1_1, 1_2}(X_1, J_1^1, X_1, J_1^2, \ldots, X_m, J_1^m, X_m, J_2^m) \right],
\]  

(EC.34)

where each \( X_{i,j} = (X_{i,j^1}, \ldots, X_{i,j^2((|T_1^i|))}, s = 1, 2 \) contains the input data specified by \( J_i^s \). We will proceed in the same way as the case \( d = 2 \), i.e. bound the squared remainder using Cauchy Schwartz inequality so as to remove the supremum

\[
\sup_{s \alpha} |R(T^1, T^2)|^2 \leq \sup_{s \alpha} \left[ \sum_{J_1^1, J_1^2} \left( \prod_{j=1}^{m} \left( w_{i,j_1(j)} - \frac{1}{n_i} \right) \right)^2 \right] \sum_{J_1^1, J_1^2} \left[ \prod_{i=1}^{n_i} \frac{1}{|T_1^i|} \sum_{J_1^1, J_1^2} \tilde{h}_{1_1, 1_2} \right]^2
\]

\[
= \sup_{s \alpha} \left[ \sum_{J_1^1, J_1^2} \left( \prod_{j=1}^{m} \left( w_{i,j_1(j)} - \frac{1}{n_i} \right) \right)^2 \right] \sum_{J_1^1, J_1^2} \left[ \prod_{i=1}^{n_i} \frac{1}{|T_1^i|} \sum_{J_1^1, J_1^2} \tilde{h}_{1_1, 1_2} \right]^2
\]

\[
= \sup_{s \alpha} \prod_{i=1}^{m} \left( \sum_{j=1}^{n_i} \left( w_{i,j} - \frac{1}{n_i} \right)^2 \right)^{|T_1^i|^2} \sum_{J_1^1, J_1^2} \left[ \prod_{i=1}^{n_i} \frac{1}{|T_1^i|} \sum_{J_1^1, J_1^2} \tilde{h}_{1_1, 1_2} \right]^2
\]

\[
\leq C(\alpha, d) \prod_{i=1}^{m} n_i^{-|T_2^i|^2} \left[ \sum_{J_1^1, J_1^2} \left( \prod_{i=1}^{n_i} \frac{1}{|T_1^i|} \sum_{J_1^1, J_1^2} \tilde{h}_{1_1, 1_2} \right)^2 \right].
\]  

(EC.35)

where we suppress the arguments of \( \tilde{h}_{1_1, 1_2} \), and the last inequality follows from Lemma (EC.2).

The second task is to bound the expectation of quantities of the form for each fixed \( J_1^1, \ldots, J_2^m \)

\[
\left( \prod_{i=1}^{n_i} \frac{1}{|T_1^i|} \sum_{J_1^1, J_1^2} \tilde{h}_{1_1, 1_2}(X_1, J_1^1, X_1, J_1^2, \ldots, X_m, J_1^m, X_m, J_2^m) \right)^2.
\]

To do so, we need the following properties of \( \tilde{h}_{1_1, 1_2} \). The first property is that, for any \( i \) and \( t \leq |T_1^i| \), the marginal expectation is zero

\[
\int \tilde{h}_{1_1, 1_2}(x_1, \ldots, x_m) dP_i(x_{i,t}) = 0.
\]  

(EC.36)
The second property is the second moment bound. By the law of total variance, one can show for any \( m \) sequences of indices \( I_i = (I_i(1), \ldots, I_i(|T_i^1| + |T_i^2|)) \in \{1, 2, \ldots, |T_i^1| + |T_i^2|\} \)

\[
\mathbb{E}_{P_1, \ldots, P_m} [h_{T_1^1, T_2^1}(X_{1, I_1}, \ldots, X_{m, I_m})] \leq \mathcal{M},
\]

where \( X_{i, I_i} = (X_i(I_i(1)), \ldots, X_i(I_i(|T_i^1| + |T_i^2|))) \). \( \text{(EC.33)} \) tells us that \( \tilde{h}_{T_1^1, T_2^1} \) is the sum of \( 2|T_1^1| \) such conditional expectations. By using the law of total variance again, along with Young’s inequality \( 2ab \leq a^2 + b^2 \)

\[
\mathbb{E}_{P_1, \ldots, P_m} [\tilde{h}_{T_1^1, T_2^1}(X_{1, I_1}, \ldots, X_{m, I_m})] \leq 4|T_1^1| \mathcal{M}. \quad \text{(EC.37)}
\]

Now we are able to proceed

\[
\mathbb{E}_{P_1, \ldots, P_m} \left( \prod_i \frac{1}{n_i} \sum_{J_1^1, \ldots, J_m^1} \tilde{h}_{T_1^1, T_2^1}(X_{1, J_1^1}, X_{1, J_2^1}, \ldots, X_{m, J_m^1}, X_{m, J_m^1}) \right)^2
\]

\[
= \prod_i \frac{1}{2|T_i^1|} \sum_{J_1^1, \ldots, J_m^1} \sum_{J_1^2, \ldots, J_m^2} \mathbb{E}_{P_1, \ldots, P_m} [\tilde{h}_{T_1^1, T_2^1}(X_{1, J_1^1}, X_{1, J_2^1}, \ldots, X_{m, J_m^1}, X_{m, J_m^1}) \cdot \tilde{h}_{T_1^1, T_2^1}(X_{1, J_1^2}, X_{1, J_2^2}, \ldots, X_{m, J_m^2}, X_{m, J_m^2})]. \quad \text{(EC.38)}
\]

Note that because of property \( \text{(EC.36)} \), the expectation in \( \text{(EC.38)} \) is zero if there is some index \( i^* \in \{1, \ldots, m\} \), and \( j^* \in \{1, \ldots, n_i\} \) such that \( X_{i^*, j^*} \) does not appear in \( X_{i^*, J_i^*} \), and shows up exactly once among \( X_{i^*, J_i^1}, X_{i^*, J_i^2} \). It can be shown that, for each fixed \( i = 1, \ldots, m \), the number of choices of \( J_i^1, J_i^2 \) that avoid this is no more than \( C(|T_i^1|, |J_i^2|)n_i^{|T_i^1|} \), where \( C(|T_i^1|, |J_i^2|) \) is some constant depending on \( |T_i^1|, |J_i^2| \) only, and \( |J_i^2| \) denotes the number of distinct indices in \( J_i^2 \). So the total number of choices of \( J_i^1, J_i^2, i = 1, \ldots, m \) that can possibly produce a nonzero expectation in \( \text{(EC.38)} \) is at most

\[
\prod_{i=1}^m C(|T_i^1|, |T_i^2|) \prod_{i=1}^m n_i^{|T_i^1|}. \quad \text{(EC.39)}
\]

On the other hand, applying Cauchy Schwartz inequality and the upper bound \( \text{(EC.37)} \) to the expectation in \( \text{(EC.38)} \), we get

\[
\left| \mathbb{E}_{P_1, \ldots, P_m} [\tilde{h}_{T_1^1, T_2^1}(X_{1, J_1^1}, X_{1, J_2^1}, \ldots, X_{m, J_m^1}, X_{m, J_m^1}) \cdot \tilde{h}_{T_1^1, T_2^1}(X_{1, J_1^2}, X_{1, J_2^2}, \ldots, X_{m, J_m^2}, X_{m, J_m^2})] \right| \leq 4|T_1^1| \mathcal{M}
\]

regardless of choices of \( J_i^1, J_i^2, i = 1, \ldots, m \). We conclude from \( \text{EC.38}, \text{EC.39} \) and the above bound that

\[
\mathbb{E}_{P_1, \ldots, P_m} \left( \prod_i \frac{1}{n_i} \sum_{J_1^1, \ldots, J_m^1} \tilde{h}_{T_1^1, T_2^1}(X_{1, J_1^1}, X_{1, J_2^1}, \ldots, X_{m, J_m^1}, X_{m, J_m^1}) \right)^2 \leq \frac{C(\alpha, d) \mathcal{M}}{\prod_i n_i^{|T_i^1|}} \quad \text{(EC.40)}
\]

uniformly for all choices of \( J_i^2, i = 1, \ldots, m \).
Finally, we go back to the inequality (EC.35) to arrive at the desired conclusion

\[
E[\sup_{\alpha} R(T^1, T^2)^2] \leq C(\alpha, d) \prod_{i=1}^{m} n_i^{-2|\mathcal{T}_i^1|} \cdot \left[ \sum_{j_2^1, \ldots, j_m^1} E\left( \prod_{i} \frac{1}{n_i^{\alpha}} \sum_{j_i^1} h_{\mathcal{T}_i^1, \mathcal{T}_i^2} \right)^2 \right] \\
\leq C(\alpha, d) M \prod_{i=1}^{m} n_i^{-|\mathcal{T}_i^1| - |\mathcal{T}_i^2|}.
\]

Since \( \sum_i (|\mathcal{T}_i^1| + |\mathcal{T}_i^2|) = d \geq 3 \), high-order remainders like \( R(T^1, T^2) \) are indeed negligible. Note there are only finite number of them, hence they all together are negligible as well. This part shows that \( E[\sup_{\alpha} |Z - \hat{Z}_L|^2] \leq C(\alpha) \left( \sum_{i \neq t} M_{i,t} + \sum_i M_i \right) \) as input sizes \( n_i \) are sufficiently large.

Now let’s turn to the second part of this proof, i.e. the uniform approximation of linear approximation at the empirical distributions \( |Z - \hat{Z}_L|^2 \). Since the proof in this part requires similar analysis, we will at some points refer back to the established results from the first part to avoid repetition.

The approach is still to expand the integral form of \( Z(w_1, \ldots, w_m) \), but near \( P^n_i \) instead of \( P_i \)

\[
Z(w_1, \ldots, w_m) = \sum_{d=0}^{T} \sum_{\sum_i |\mathcal{T}_i^2| = d} \int h(x_1, \ldots, x_m) \prod_{i=1}^{m} \prod_{i \notin \mathcal{T}_i^2} dP^n_i(x_{i,t}) \prod_{i=1}^{m} \prod_{i \in \mathcal{T}_i^2} d(w_i - P^n_i)(x_{i,t}),
\]

where each \( \mathcal{T}_i^2 \) is again an ordered subset of \( \{1, 2, \ldots, T_i\} \) that specifies the second subscript \( t \) of the argument \( x_{i,t} \). Similarly, summation with \( d = 0 \) gives the linear approximation at the empirical distributions, i.e. \( \hat{Z}_L(w_1, \ldots, w_m) \). So all summation with \( d \geq 2 \) will be part of the error. To bound such summation, we rewrite \( P^n_i \) as \( P^n_i - P_i + P_i \), and distributes the measure product as we did before. For each \( i \), denote by \( \mathcal{T}_i^1 \) the set of second subscripts of all \( x_{i,t} \)'s for which \( P^n_i - P_i \) is picked. Recalling the definition of \( R(T^1, T^2) \), where \( T^s = (\mathcal{T}_1^s, \ldots, \mathcal{T}_m^s), s = 1, 2 \), gives the following representation for summation with \( d \geq 2 \)

\[
\sum_{\sum_i |\mathcal{T}_i^2| = d} \int h(x_1, \ldots, x_m) \prod_{i=1}^{m} \prod_{i \notin \mathcal{T}_i^2} dP^n_i(x_{i,t}) \prod_{i=1}^{m} \prod_{i \in \mathcal{T}_i^2} d(w_i - P^n_i)(x_{i,t}) = \sum_{\mathcal{T}_i^1, \mathcal{T}_i^2, |\mathcal{T}_i^2| = d} R(T^1, T^2).
\]

From the result (EC.31), we know that in this sum only terms with \( |\mathcal{T}_i^1| = 0, |\mathcal{T}_i^2| = 2 \) will enter the leading error term. Moreover, the sum of all such terms are equal to \( R_{uw} \), which has been shown in the first part to satisfy the desired bound. This shows that \( E[\sup_{\alpha} |Z - \hat{Z}_L|^2] \leq C(\alpha) \left( \sum_{i \neq t} M_{i,t} + \sum_i M_i \right) \).

**EC.2.2. Part 2: Simulation Error**

First we present three lemmas.

**Lemma EC.3.** Suppose Assumption 2 holds with \( k = 2 \) for the function \( h \), then as all \( n_i \)'s are sufficiently large.

\[
E_{P_1, \ldots, P_m} \left[ \sup_{(w_1, \ldots, w_m) \in \mathcal{A}_\alpha} |Z(w_1, \ldots, w_m) - Z(P_1, \ldots, P_m)|^2 \right] \leq O \left( \sum_{i=1}^{m} \frac{1}{n_i} \right).
\]
Proof of Lemma EC.3. This is a corollary of the proof for Proposition 4 and 5. Note that (EC.41) holds for any \( T_1, T_2 \), although it’s proved particularly for \( d \geq 3 \). Using this observation concludes the proof. □

**Lemma EC.4.** Let \( \sigma^2 = \text{Var}_{P_1, \ldots, P_m}[h(X_1, \ldots, X_m)] \). Under Assumption 2 with \( k = 4 \), as all \( n_i \)'s grow to \( \infty \)

\[
\mathbb{E}_{P_1, \ldots, P_m} \left[ \sup_{(w_1, \ldots, w_m) \in \mathcal{A}_\alpha} \left| \text{Var}_{w_1, \ldots, w_m}(h) - \sigma^2 \right| \right] = O\left( \sqrt{\sum_i \frac{1}{n_i}} \right).
\]

Proof of Lemma EC.4. We will denote \( Z(P_1, \ldots, P_m) \) by simply \( Z \) and \( Z(w_1, \ldots, w_m) \) by \( Z_w \). Applying Lemma EC.3 to \( h \) and \( |h - Z|^2 \) (\( k = 4 \) required) gives

\[
\mathbb{E}_{P_1, \ldots, P_m} \left[ \sup_{(w_1, \ldots, w_m) \in \mathcal{A}_\alpha} \left| Z_w - Z \right|^2 \right] = O\left( \sum_i \frac{1}{n_i} \right).
\]

\[
\mathbb{E}_{P_1, \ldots, P_m} \left[ \sup_{(w_1, \ldots, w_m) \in \mathcal{A}_\alpha} \left| \mathbb{E}_{w_1, \ldots, w_m}[h - Z]^2 - \sigma^2 \right|^2 \right] = O\left( \sum_i \frac{1}{n_i} \right).
\]

Since

\[
\sup_{\mathcal{A}_\alpha} \left| \text{Var}_{w_1, \ldots, w_m}(h) - \sigma^2 \right| \\
= \sup_{\mathcal{A}_\alpha} \left| \mathbb{E}_{w_1, \ldots, w_m}(h - Z + Z_w)^2 - \sigma^2 \right| \\
= \sup_{\mathcal{A}_\alpha} \left| \mathbb{E}_{w_1, \ldots, w_m}(h - Z)^2 - \sigma^2 - (Z - Z_w)^2 \right| \\
\leq \sup_{\mathcal{A}_\alpha} \left| \mathbb{E}_{w_1, \ldots, w_m}(h - Z)^2 - \sigma^2 \right| + \sup_{\mathcal{A}_\alpha} (Z - Z_w)^2,
\]

the expectation can be bounded as

\[
\mathbb{E} \left[ \sup_{\mathcal{A}_\alpha} \left| \text{Var}_{w_1, \ldots, w_m}(h) - \sigma^2 \right| \right] \\
\leq \mathbb{E} \left[ \sup_{\mathcal{A}_\alpha} \left| \mathbb{E}_{w_1, \ldots, w_m}(h - Z)^2 - \sigma^2 \right| \right] + \mathbb{E} \left[ \sup_{\mathcal{A}_\alpha} (Z - Z_w)^2 \right] \\
= O\left( \sqrt{\sum_i \frac{1}{n_i}} \right) + O\left( \sum_i \frac{1}{n_i} \right) = O\left( \sqrt{\sum_i \frac{1}{n_i}} \right),
\]

where we apply Jensen’s inequality to the first expectation term. □

**Lemma EC.5.** Under Assumption 2 with \( k = 4 \), when all \( n_i \)'s are sufficiently large the gradient estimator \( \hat{G}_{i,j} \) with \( R = R_1 \) satisfies

\[
\mathbb{E} \left[ \sum_{i=1}^m \frac{1}{n_i^2} \sum_{j=1}^{n_i} (\hat{G}_{i,j} - \hat{G}_i(X_{i,j}))^2 \right] \leq \frac{\sigma^2 T}{R_1},
\]

where \( T = \sum_{i=1}^m T_i \), and the expectation \( \mathbb{E} \) is take with respect to randomness from both the simulation procedure and input data.
Proof of Lemma EC.5. Since \( \hat{G}_{i,j} \) differs from the standard sample covariance by only a factor of \( \frac{R_1 - 1}{R_1} \), its bias can be easily identified as \( \hat{G}_i(X_{i,j})/R_1 \). We will use \( \xi \) to denote the randomness in gradient simulation. By the variance formula for the standard sample covariance

\[
\text{Var}_\xi(\hat{G}_{i,j}) = \frac{(R_1 - 1)^2}{R_1^3}\left[\left(\text{Var}[h]\right)\left((S_{i,j}(X_i))^2\right) + \frac{1}{R_1 - 1}\text{Var}[h]\text{Var}[S_{i,j}(X_i)] - \frac{R_1 - 2}{R_2 - 1}\left(\hat{G}_i(X_{i,j})\right)^2\right].
\]

Hence the mean squared error

\[
\sum_{j=1}^{n_i} \text{E}_\xi(\hat{G}_{i,j} - \hat{G}_i(X_{i,j}))^2
\]

\[
\leq \sum_{j=1}^{n_i} \left(\frac{(R_1 - 1)^2}{R_1^3}\left[\left(\text{Var}[h]\right)\left((S_{i,j}(X_i))^2\right) + \frac{1}{R_1 - 1}\text{Var}[h]\text{Var}[S_{i,j}(X_i)]\right]\right)
\]

\[
\leq \frac{1}{R_1}\sum_{j=1}^{n_i} \text{E}_\xi((h - \text{E}_\xi[h])^2(S_{i,j}(X_i))^2) + \frac{n_i^2\text{T}^2}{R_1^2}\text{Var}[h].
\]  

(EC.42)

To attack the first term

\[
\sum_{j=1}^{n_i} \text{E}_\xi((h - \text{E}_\xi[h])^2(S_{i,j}(X_i))^2)
\]

\[
= \text{E}_\xi\left[\sum_{j=1}^{n_i} (h - \text{E}_\xi[h])^2 \left(T_i^2 + n_i^2\left(\sum_{t=1}^{T_i} \mathbf{1}\{X_i(t) = X_{i,j}\}\right)^2 - 2T_in_i\sum_{t=1}^{T_i} \mathbf{1}\{X_i(t) = X_{i,j}\}\right)\right]
\]

\[
= T_i^2n_i\text{Var}[h] - 2T_i^2n_i\text{Var}[h] + \text{E}_\xi\left[\sum_{j=1}^{n_i} (h - \text{E}_\xi[h])^2n_i^2\left(\sum_{t=1}^{T_i} \mathbf{1}\{X_i(t) = X_{i,j}\}\right)^2\right]
\]

\[
= - T_i^2n_i\text{Var}[h] + \text{E}_\xi\left[\sum_{j=1}^{n_i} (h - \text{E}_\xi[h])^2n_i^2\left(\sum_{t=1}^{T_i} \mathbf{1}\{X_i(t) = X_{i,j}\}\right) + \sum_{s \neq t} \mathbf{1}\{X_i(t) = X_i(s) = X_{i,j}\}\right]
\]

\[
\leq n_i^2T_i\text{Var}[h] + n_i^2\text{E}_\xi[(h - \text{E}_\xi[h])^2]\mathbf{1}\{X_i(t) = X_i(s)\}
\]

\[
= n_i^2T_i\text{Var}[h] + n_i^2\sum_{s \neq t} P_t(X_i(t) = X_i(s))\text{E}_\xi[(h - \text{E}_\xi[h])^2|X_i(t) = X_i(s)]
\]

\[
= n_i^2T_i\text{Var}[h] + n_i\sum_{s \neq t} \text{E}_\xi[(h - Z^*)^2|X_i(t) = X_i(s)]
\]

\[
\leq n_i^2T_i\text{Var}[h] + 2n_i\sum_{s \neq t} \text{E}_\xi[(h - Z^*)^2|X_i(t) = X_i(s)] + 2n_iT_i(T_i - 1)(\text{E}_\xi[h] - Z^*)^2,
\]

where in the first inequality we leave out the negative term \(-T_i^2n_i\text{Var}[h]\) and in the second we use \((h - \text{E}_\xi[h])^2 \leq 2(h - Z^*)^2 + 2(\text{E}_\xi[h] - Z^*)^2\) with \(Z^*\) being the true performance measure \(Z(P_1, \ldots, P_m)\). One can rewrite the conditional expectation

\[
\text{E}_\xi[(h - Z^*)^2|X_i(t) = X_i(s)] = \text{E}_\xi[(h_{i,s,t} - Z^*)^2],
\]

\[
\text{E}_\xi[(h_{i,s,t} - Z^*)^2].
\]
where \( h_{i,s,t} = h(X_{1,i_1}, \ldots, X_{m,t_m}) \) with \( I_r = (1, \ldots, T_r) \) for \( r \neq i \) and \( I_i = (1, \ldots, \max(s,t) - 1, \min(s,t), \max(s,t), \ldots, T_i - 1) \). Apply Lemma EC.3 to \( (h_{i,s,t} - Z^*)^2 \) and \( h \) respectively to get

\[
E_{P_1,\ldots,P_m} \left[ \mathbb{E}_\xi \left[ (h_{i,s,t} - Z^*)^2 \right] \right] = E_{P_1,\ldots,P_m} \left[ \mathbb{E}_\xi \left[ (h_{i,s,t} - Z^*)^2 \right] \right] + O \left( \sqrt{\sum_i \frac{1}{n_i}} \right),
\]

\[
E_{P_1,\ldots,P_m} \left[ \mathbb{E}_\xi \left[ h \right]^2 \right] = O \left( \sum_i \frac{1}{n_i} \right).
\]

Lemma EC.3 indicates \( E_{P_1,\ldots,P_m} \left[ \text{Var}_\xi \left[ h \right] \right] = \sigma^2 + O \left( \sqrt{\sum_i \frac{1}{n_i}} \right) \).

Now we take expectation of \( \text{EC.32} \) with respect to the randomness from data and use the above upper bounds

\[
E_{P_1,\ldots,P_m} \left[ \mathbb{E}_\xi \left[ \sum_{j=1}^{n_i} (\hat{G}_{i,j} - \hat{G}_i(X_{i,j}))^2 \right] \right] = \frac{1}{R_i} E_{P_1,\ldots,P_m} \left[ (1 + 1/R_i)n_i^2 T_i \text{Var}_\xi \left[ h \right] + 2n_i \sum_{s \neq t} E_{\xi} \left[ (h_{i,s,t} - Z^*)^2 \right] + 2n_i T_i (T_i - 1) (E_{\xi} \left[ h \right] - Z)^2 \right] \leq \frac{n_i^2 T_i \sigma^2}{R_i}.
\]

Dividing each side by \( n_i^2 \) and summing up over \( i = 1, \ldots, m \) completes the proof. \( \blacksquare \)

Now we have the tools for proving the rest of Proposition 4.

**Proof of the second half of Proposition 4.** The second part of Proposition 2 entails that expectation of an influence function must be zero. That is, \( \sum_{j=1}^{n_i} \hat{G}_i(X_{i,j}) = 0 \) for all \( i \). Note that the estimator \( \mathbb{E}_{\hat{L}} \) with arbitrary \( R \) also has this property, i.e. \( \sum_{j=1}^{n_i} \hat{G}_{i,j} = 0 \) for all \( i \). Hence

\[
\sup_{(\hat{w}_1, \ldots, \hat{w}_m) \in \mathcal{A}_0} \left| \hat{L} - \hat{L} \right| = \sup_{(\hat{w}_1, \ldots, \hat{w}_m) \in \mathcal{A}_0} \left| \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{i,j} (\hat{G}_{i,j} - \hat{G}_i(X_{i,j})) \right| \leq \sqrt{u(\alpha^2)} \mathcal{L}_{\alpha,1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{1}{n_i^2} (\hat{G}_{i,j} - \hat{G}_i(X_{i,j}))^2,
\]

where in the second equality we use the zero-sum property, and in the last inequality we use Lemma EC.2. Square both sides of the above inequality, take expectation and use Lemma EC.5 to get the desired conclusion. \( \blacksquare \)
EC.3. Proof of Theorem 8

We need the following useful lemma.

**Lemma EC.6.** Consider two generic optimization problems that share the same arbitrary feasible set $\mathcal{D}$, and have possibly different objectives $f_1(x), f_2(x)$. Suppose

$$\sup_{x \in \mathcal{D}} |f_1(x) - f_2(x)| \leq d$$

and optimal values are attained. For $i = 1, 2$, let $x_i^{\text{min}}$ be a minimizer and $x_i^{\text{max}}$ a maximizer of $f_i(x)$. Then

$$f_2(x_1^{\text{min}}) - f_2(x_2^{\text{min}}) \leq 2d, \quad f_2(x_2^{\text{max}}) - f_2(x_1^{\text{max}}) \leq 2d.$$ 

**Proof of Lemma EC.6.** It suffices to prove the minimizer case, since the maximizer case can be reduced to minimizer case by replacing $f_i$ with $-f_i$. By optimality of $x_i^{\text{min}}$ and the condition of this lemma we have

$$f_2(x_1^{\text{min}}) \leq f_1(x_1^{\text{min}}) + d \leq f_1(x_2^{\text{min}}) + d \leq f_2(x_2^{\text{min}}) + 2d.$$

This completes the proof. □

**Proof of Theorem 8.** Denote by $\mathcal{L}_\alpha, \mathcal{U}_\alpha$ the minimum and maximum value of $Z_L(w_1, \ldots, w_m)$ over the feasible set $\mathcal{A}_\alpha$. Theorem 7 states that $P(\mathcal{L}_\alpha \leq Z(P_1, \ldots, P_m) \leq \mathcal{U}_\alpha) \to 1 - \alpha$. Now applying Lemma EC.6 to objectives $Z_L$ and $\hat{Z}_L$ with the feasible set being $\mathcal{A}_\alpha$ gives

$$Z_L(w_1^{\text{min}}, \ldots, w_m^{\text{min}}) - \mathcal{L}_\alpha \leq 2 \sup_{x \in \mathcal{A}_\alpha} |Z_L - \hat{Z}_L| \leq (EC.43)$$

$$\mathcal{U}_\alpha - Z_L(w_1^{\text{max}}, \ldots, w_m^{\text{max}}) \leq 2 \sup_{x \in \mathcal{A}_\alpha} |Z_L - \hat{Z}_L|. \quad (EC.44)$$

It is also clear that the outputs $\mathcal{L}_\alpha, \mathcal{U}_\alpha$ computed from this particular procedure satisfy

$$|\mathcal{L}_\alpha - Z_L(w_1^{\text{min}}, \ldots, w_m^{\text{min}})| = |\mathcal{U}_\alpha - Z_L(w_1^{\text{max}}, \ldots, w_m^{\text{max}}) = |\hat{Z} - Z(P_1^{n_1}, \ldots, P_m^{n_m})|$$

where $\hat{Z}$ is the sample mean estimate of $Z(P_1^{n_1}, \ldots, P_m^{n_m})$ based on $R_1$ i.i.d. replications. Therefore, the error of the resulting CI bounds

$$|E_{l}| = |\mathcal{L}_\alpha - \mathcal{L}_\alpha| \leq 2 \sup_{x \in \mathcal{A}_\alpha} |Z_L - \hat{Z}_L| + |\hat{Z} - Z(P_1^{n_1}, \ldots, P_m^{n_m})|$$

$$|E_{u}| = |\mathcal{U}_\alpha - \mathcal{U}_\alpha| \leq 2 \sup_{x \in \mathcal{A}_\alpha} |Z_L - \hat{Z}_L| + |\hat{Z} - Z(P_1^{n_1}, \ldots, P_m^{n_m})|.$$ 

Moreover, by Proposition 4 and Lemma EC.4

$$\mathbb{E}[|E_{l}|] \leq 2 \mathbb{E} \sup_{x \in \mathcal{A}_\alpha} |Z_L - \hat{Z}_L| + \mathbb{E}[|\hat{Z} - Z(P_1^{n_1}, \ldots, P_m^{n_m})|]$$
By using Proposition 4 and 5, the lower bound error trivially holds

\[ \sum \sum E_{\alpha} \leq 2E \left[ \sup_{x_{\alpha}} |Z_L - \hat{Z}_L| + \sup_{x_{\alpha}} |\hat{Z}_L| \right] + \sqrt{E \left[ (Z - Z(P_1^{n_1}, \ldots, P_m^{n_m}))^2 \right]}
\]

\[ \leq 2\sqrt{E \left[ \sup_{x_{\alpha}} |Z_L - \hat{Z}_L|^2 \right] + 2E \sup_{x_{\alpha}} |\hat{Z}_L - \hat{Z}_L|^2} \leq \frac{1}{R_1} \left( \text{Var}_{P_1^{n_1}, \ldots, P_m^{n_m}}(h) \right)
\]

Obviously the same bound can be established for \( E_u \) as well.

\[ \text{EC.4. Proofs of Theorems 9 and 1}
\]

\[ \text{Proof of Theorem 9.} \] The proof resembles the one for Theorem 8. The optimality gaps (EC.33) (EC.44) of the optimization pair continue to hold. On the other hand the following bound trivially holds

\[ |Z(w_1^{min}, \ldots, w_m^{min}) - Z_L(w_1^{min}, \ldots, w_m^{min})| \leq \sup_{x_{\alpha}} |Z - Z_L| \]

\[ |Z(w_1^{max}, \ldots, w_m^{max}) - Z_L(w_1^{max}, \ldots, w_m^{max})| \leq \sup_{x_{\alpha}} |Z - Z_L|.
\]

By using Proposition 4 and 5 the lower bound error \( E_l = Z(w_1^{min}, \ldots, w_m^{min}) - L_a \) now satisfies

\[ E[|E_l|] \leq 2E \left[ \sup_{x_{\alpha}} |Z_L - \hat{Z}_L| + \sup_{x_{\alpha}} |Z - Z_L| \right]
\]

\[ \leq 2\sqrt{E \left[ \sup_{x_{\alpha}} |Z_L - \hat{Z}_L|^2 \right] + 2E \sup_{x_{\alpha}} |\hat{Z}_L - \hat{Z}_L|^2} \leq \frac{1}{R_1} \left( \text{Var}_{P_1^{n_1}, \ldots, P_m^{n_m}}(h) \right)
\]

Again the same bound holds for \( E_u = Z(w_1^{max}, \ldots, w_m^{max}) - U_a \).

\[ \text{Proof of Theorem 1.} \] The outputs \( \mathcal{L}_a^{BEL}, \omega_a^{BEL} \) of Algorithm 1 is subject to one more layer of evaluation error (Step 3) in comparison with the CI in Theorem 9. To establish an upper bound for the total error, it suffices to quantify the evaluation error \( \mathcal{L}_a^{BEL} - Z(w_1^{min}, \ldots, w_1^{min}), \omega_a^{BEL} - Z(w_1^{max}, \ldots, w_1^{max}) \) only in view of Theorem 9

\[ E \left[ |\mathcal{L}_a^{BEL} - Z(w_1^{min}, \ldots, w_1^{min})| \right] \leq \sqrt{E \left[ (\mathcal{L}_a^{BEL} - Z(w_1^{min}, \ldots, w_1^{min}))^2 \right]}
\]

\[ = \sqrt{\left( \frac{1}{R_2} \text{Var}_{w_1^{min}}(h) \right) \left( \frac{1}{R_2} \text{Var}_{w_1^{min}}(h) - \sigma^2 \right) + \sigma^2}
\]
\[
\begin{align*}
\sqrt{\frac{1}{R_2}} \mathbb{E}_{P_1, \ldots, P_m} \left[ \sup_{\mathcal{F}_0} \text{Var}(\mathbf{w}_1, \ldots, \mathbf{w}_1(h) - \sigma^2) \right] & \leq 2 \sqrt{\frac{\text{Var}(\mathbf{U})}{n_0}} + o_p\left( \sqrt{\frac{n_0}{n}} \right),
\end{align*}
\]

where in the last inequality we use Lemma EC.4. Summing up this error and that of Theorem 9 gives the desired conclusion for \( E_{i!}^{BEL} \). The same bound is true for \( E_{n!}^{BEL} \). \( \square \)

**EC.5. Proofs of Theorems 2 and 3 and Corollaries 2, 3 and 4**

The following lemma explains in detail the implication of Corollary 1.

**Lemma EC.7.** Under the same conditions of Theorem 7, the optimal values of program 9 satisfy

\[
\mathcal{L}_\alpha/\mathcal{U}_\alpha = Z(P_1, \ldots, P_m) + \sum_{i=1}^m \bar{G}_i \mp z_{1-\alpha/2} \sqrt{\sum_{i=1}^m \frac{\text{Var}(G_i)}{n_i}} + o_p\left( \sqrt{\frac{1}{n_i}} \right), \text{ as } n_i \to \infty \text{ for } i \in I,
\]

where \( \bar{G}_i = \sum_{j=1}^{n_i} G_i(X_{i,j})/n_i \) and \( \text{Var}(G_i) = \sum_{j=1}^{n_i} (G_i(X_{i,j}) - \bar{G}_i)^2/(n_i - 1) \) are the sample mean and variance, \( I = \{ i | \text{Var}(G_i(X_i)) > 0 \} \), and \( z_{1-\alpha/2} \) is the \( 1 - \alpha/2 \) critical value of the standard normal.

**Proof of Lemma EC.7** Define the profile likelihood ratio

\[
R_G(\mu) = \max \left\{ \prod_{i=1}^m \prod_{j=1}^{n_i} n_i w_{i,j} \sum_{i=1}^m \sum_{j=1}^{n_i} w_{i,j} G_i(X_{i,j}) = \mu, \sum_{j=1}^{n_i} w_{i,j} = 1 \text{ for all } i, w_{i,j} \geq 0 \text{ for all } i, j \right\}.
\]

Corollary 1 entails that for any fixed \( \epsilon > 0 \)

\[
P\left( -2 \log R_G \left( \sum_{i=1}^m G_i - (z_{1-\alpha/2} - \epsilon) \sqrt{\sum_{i=1}^m \text{Var}(G_i)/n_i} \right) < A_{1-\alpha}^2 \right) \to 1, \quad (EC.45)
\]

\[
P\left( -2 \log R_G \left( \sum_{i=1}^m G_i - (z_{1-\alpha/2} + \epsilon) \sqrt{\sum_{i=1}^m \text{Var}(G_i)/n_i} \right) > A_{1-\alpha}^2 \right) \to 1. \quad (EC.46)
\]

Since the event \( EC.45 \) implies \( \mathcal{L}_\alpha < Z(P_1, \ldots, P_m) + \sum_{i=1}^m \bar{G}_i - (z_{1-\alpha/2} - \epsilon) \sqrt{\sum_{i=1}^m \text{Var}(G_i)/n_i} \) and the event \( EC.46 \) implies \( \mathcal{L}_\alpha > Z(P_1, \ldots, P_m) + \sum_{i=1}^m \bar{G}_i - (z_{1-\alpha/2} + \epsilon) \sqrt{\sum_{i=1}^m \text{Var}(G_i)/n_i} \), it holds that

\[
P\left( |\mathcal{L}_\alpha - Z(P_1, \ldots, P_m) - \sum_{i=1}^m \bar{G}_i + z_{1-\alpha/2} \sqrt{\sum_{i=1}^m \text{Var}(G_i)/n_i} | \leq \epsilon \sqrt{\sum_{i=1}^m \text{Var}(G_i)/n_i} \right) \to 1.
\]

Sending \( \epsilon \) to 0 gives the desired conclusion. The proof for \( \mathcal{U}_\alpha \) is similar. \( \square \)

The next two lemmas concern the consistency of input-induced variance estimate 11.
**Lemma EC.8.** Under Assumption [1] and Assumption [2] with \( k = 2 \), if \( \min_{i \in I} n_i = \omega(\sqrt{\min_{i \notin I} n_i}) \), where \( I = \{i | \text{Var}(G_i(X_i)) > 0 \} \), then the sample influence function \( \hat{G}_i(X_{i,j}) \) yields a consistent variance estimate
\[
\sum_{i=1}^{m} \frac{1}{n_i^2} \sum_{j=1}^{n_i} \hat{G}_i(X_{i,j})^2 / \sum_{i \in I} \frac{1}{n_i} \text{Var}(G_i(X_i)) \to 1 \text{ in probability, as all } n_i \to \infty.
\]

**Proof of Lemma EC.8.** Since the standard sample variance \( \sum_{i=1}^{m} \sum_{j=1}^{n_i} G_i(X_{i,j})^2 / n_i^2 \) is consistent, it suffices to show that \( \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{G}_i(X_{i,j})^2 / n_i^2 \) differs by only \( o_p(\sum_{i \in I} 1/n_i) \). By following the same line of the proof for Proposition [4] and [5] one can show the following
\[
\mathbb{E}[(\hat{G}_i(X_{i,j}) - G_i(X_{i,j}))^2] \leq C(h) \sum_{i \in I} \frac{1}{n_i}, \text{ for all } i, j
\]
where \( C(h) \) is some finite constant that depends on the performance measure \( h \). Hence
\[
\mathbb{E}[(\sum_{i=1}^{m} \frac{1}{n_i^2} \sum_{j=1}^{n_i} \hat{G}_i(X_{i,j})^2 - \sum_{i=1}^{m} \frac{1}{n_i^2} \sum_{j=1}^{n_i} G_i(X_{i,j})^2)]
\]
\[
= \mathbb{E}[(\sum_{i=1}^{m} \frac{1}{n_i^2} \sum_{j=1}^{n_i} (\hat{G}_i(X_{i,j}) + G_i(X_{i,j}))(\hat{G}_i(X_{i,j}) - G_i(X_{i,j})))]
\]
\[
\leq \sum_{i=1}^{m} \frac{1}{n_i^2} \sum_{j=1}^{n_i} (\mathbb{E}[(\hat{G}_i(X_{i,j}) - G_i(X_{i,j}))^2] + 2\mathbb{E}[(G_i(X_{i,j})(\hat{G}_i(X_{i,j}) - G_i(X_{i,j})))]
\]
\[
\leq C(h) \left( \sum_{i \in I} \frac{1}{n_i} \right)^2 + \sum_{i=1}^{m} \frac{2}{n_i} \sqrt{\text{Var}(G_i(X_i))} C(h) \sum_{i \in I} \frac{1}{n_i}
\]
\[
= C(h) \left( \sum_{i \in I} \frac{1}{n_i} \right)^2 + \sqrt{C(h) \sum_{i=1}^{m} \frac{1}{n_i} \sum_{i \in I} 2\sqrt{\text{Var}(G_i(X_i))}.}
\]
When \( n_i = \omega(\sqrt{\min_{i \notin I} n_i}) \) for all \( l \notin I \), the difference is indeed \( o_p(\sum_{i \in I} 1/n_i) \). \( \square \)

**Lemma EC.9.** Under Assumption [1] and Assumption [2] with \( k = 4 \), if \( \min_{i \in I} n_i = \omega(\sqrt{\min_{i \notin I} n_i}) \), where \( I = \{i | \text{Var}(G_i(X_i)) > 0 \} \), and \( R_1 \) is chosen such that [6] holds, then the input-induced variance estimate [1] is consistent, i.e.
\[
\hat{\sigma}_i^2 / \sum_{i \in I} \frac{1}{n_i} \text{Var}(G_i(X_i)) \to 1 \text{ in probability.}
\]

**Proof of Lemma EC.9.** Because of Lemma EC.8 it suffices to verify the following
\[
\hat{\sigma}_i^2 - \sum_{i=1}^{m} \frac{1}{n_i^2} \sum_{j=1}^{n_i} \hat{G}_i(X_{i,j})^2 = o_p(\sum_{i \in I} 1/n_i).
\]
(47)
Consider the following surrogate of \( \hat{G}_{i,j} \)
\[
\tilde{G}_{i,j} = \frac{1}{R_1} \sum_{r=1}^{R_1} \left[ (h(X_1^r, \ldots, X_m^r) - Z(P_1^{n_1}, \ldots, P_m^{n_m}))(n_i \sum_{t=1}^{T_i} 1\{X_i^r(t) = X_{i,j}\} - T_i) \right],
\]
which replaces the sample mean \( \hat{Z} \) in \( \hat{G}_{i,j} \) by the true mean \( Z(P_1^{n_1}, \ldots, P_m^{n_m}) \), and the surrogate of \( \sigma_i^2 \)

\[
\tilde{\sigma}_i^2 = \sum_{i=1}^{m} \frac{1}{n_i} \left( \sum_{j=1}^{n_i} \left( \hat{G}_{i,j}^2 - \frac{n_i T_i \sigma^2}{R_1} \right) \right),
\]

\[
\sigma^2 = \frac{1}{R_1 - 1} \sum_{r=1}^{R_1} \left( h(X_1^r, \ldots, X_m^r) - \hat{Z} \right)^2.
\]

We will break up the task of proving (EC.47) into two steps, i.e. first showing \( \tilde{\sigma}_i^2 - \sigma_i^2 = o_p(\sum_{i=1}^{m} 1/n_i) \) and then \( \tilde{\sigma}_i^2 - \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{G}_i(X_{i,j})^2 / n_i^2 = o_p(\sum_{i=1}^{m} 1/n_i) \).

For the first step note that by Cauchy Schwartz

\[
|\tilde{\sigma}_i^2 - \sigma_i^2| = \left| \sum_{i=1}^{m} \frac{1}{n_i^2} \sum_{j=1}^{n_i} (\hat{G}_{i,j} + \tilde{G}_{i,j}) (\hat{G}_{i,j} - \tilde{G}_{i,j}) \right|
\]

\[
\leq \sum_{i=1}^{m} \frac{1}{n_i^2} \sum_{j=1}^{n_i} (\hat{G}_{i,j} + \tilde{G}_{i,j})^2 \sum_{i=1}^{m} \frac{1}{n_i^2} \sum_{j=1}^{n_i} (\hat{G}_{i,j} - \tilde{G}_{i,j})^2. \tag{EC.48}
\]

Similar calculation to the proof of Lemma (EC.5) shows that

\[
\sum_{i=1}^{m} \frac{1}{n_i^2} \sum_{j=1}^{n_i} (\hat{G}_{i,j} - \tilde{G}_i(X_{i,j}))^2 = O_p \left( \sum_{i=1}^{m} \frac{T_i \sigma^2}{R_1} \right),
\]

\[
\sum_{i=1}^{m} \frac{1}{n_i^2} \sum_{j=1}^{n_i} (\tilde{G}_{i,j} - \tilde{G}_i(X_{i,j}))^2 = O_p \left( \sum_{i=1}^{m} \frac{T_i \sigma^2}{R_1} \right).
\]

Therefore the first sum in (EC.48)

\[
\sum_{i=1}^{m} \frac{1}{n_i^2} \sum_{j=1}^{n_i} (\hat{G}_{i,j} + \tilde{G}_{i,j})^2 \leq 8 \sum_{i=1}^{m} \frac{1}{n_i^2} \sum_{j=1}^{n_i} \left( \frac{\hat{G}_{i,j} + \tilde{G}_{i,j}}{2} - \tilde{G}_i(X_{i,j}) \right)^2 + 8 \sum_{i=1}^{m} \frac{1}{n_i^2} \sum_{j=1}^{n_i} \tilde{G}_i(X_{i,j})^2
\]

\[
= O_p \left( \sum_{i=1}^{m} \frac{T_i \sigma^2}{R_1} \right) + \sum_{i=1}^{m} \frac{1}{n_i}.
\]

To handle the second sum in (EC.48) we notice

\[
\hat{G}_{i,j} - \tilde{G}_{i,j} = \frac{1}{R_1} \sum_{r=1}^{R_1} \left[ (\hat{Z} - Z(P_1^{n_1}, \ldots, P_m^{n_m})) (n_i N_{i,j} - T_i) \right]
\]

\[
= (\hat{Z} - Z(P_1^{n_1}, \ldots, P_m^{n_m})) (n_i \bar{N}_{i,j} - T_i),
\]

where \( N_{i,j} = \sum_{t=1}^{T_i} 1\{X_i(t) = X_{i,j}\} \) and \( \bar{N}_{i,j} \) is the sample mean over the total \( R_1 \) simulation runs.

\[
\sum_{i=1}^{m} \frac{1}{n_i^2} \sum_{j=1}^{n_i} (\hat{G}_{i,j} - \tilde{G}_{i,j})^2 = (\hat{Z} - Z(P_1^{n_1}, \ldots, P_m^{n_m}))^2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\bar{N}_{i,j} - \frac{T_i}{n_i})^2
\]

\[
= O_p \left( \frac{\sigma^2}{R_1} \right) \cdot O_p \left( \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{T_i}{n_i R_1} (1 - \frac{1}{n_i}) \right) = O_p \left( \sum_{i=1}^{m} \frac{T_i \sigma^2}{R_1^2} \right).
\]

Now we go back to (EC.48) and arrive at the following bound

\[
|\tilde{\sigma}_i^2 - \sigma_i^2| \leq \sqrt{O_p \left( \sum_{i=1}^{m} \frac{T_i \sigma^2}{R_1} \right) + \sum_{i \in l} \frac{1}{n_i}} \cdot O_p \left( \sum_{i=1}^{m} \frac{T_i \sigma^2}{R_1^2} \right) = \sqrt{O_p \left( \frac{1}{R_1^3} + \frac{1}{R_1^2} \sum_{i \in l} \frac{1}{n_i} \right)}.
\]
When the simulation effort $R_1$ is set to be $\omega(\min_{i\in I} n_{i}^{2/3})$, we conclude $|\hat{\sigma}^2_t - \sigma^2_t| = o_p(\sum_{i\in I} 1/n_i)$. This completes the first step.

To perform the second step, i.e. to show $\hat{\sigma}^2_t - \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{G}_i(X_{i,j})^2/n_i^2 = o_p(\sum_{i\in I} 1/n_i)$, we conduct a direct analysis of the mean square error of $\hat{\sigma}^2_t$. Let $\xi$ denote the randomness introduced by simulation. Then $E_\xi[\hat{G}_i^2] = \hat{G}_i^2(X_{i,j}) + Var_\xi((h - E_\xi[h])(n_iN_{i,j} - T_i))/R_1$, where $N_{i,j} = \sum_{t=1}^{T_i} 1\{X_i(t) = X_{i,j}\}$. By the proof of Lemma [EC.5], the variance term satisfies

$$\sum_{j=1}^{n_i} Var_\xi((h - E_\xi[h])(n_iN_{i,j} - T_i)) = n_i^2 T_i Var_\xi[h] + O_p(n_i T_i^2).$$

The $\hat{\sigma}^2_t$ in the bias correction term is the sample variance of $h$, hence unbiased for $Var_\xi[h]$. So the bias of $\hat{\sigma}^2_t$ can be characterized as

$$E_\xi[\hat{\sigma}^2_t] - \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{G}_i^2(X_{i,j})/n_i^2 = O_p\left(\sum_{i=1}^{m} T_i^2/n_i R_1\right) = O_p\left(\frac{1}{R_1}\sum_{i=1}^{m} \frac{1}{n_i}\right). \quad (EC.49)$$

To compute its variance, we study its second moment. Somewhat tedious calculations show that when Assumption 2 holds with $k = 4$

$$E_\xi\left[\sum_{j=1}^{n_i} \hat{G}_i^2\right] = \sum_{j,k} E_\xi[\hat{G}_{i,j} \hat{G}_{i,k}]$$

$$= \frac{1}{R_1^4} \sum_{j,k} \sum_{r_1,r_2,r_3,r_4} E_\xi\left[\prod_{s=1}^{4} (h_{r_s} - E_\xi[h]) \cdot (n_iN_{i,j} - T_i)(n_iN_{i,j}^2 - T_i)(n_iN_{i,j}^3 - T_i)(n_iN_{i,j}^4 - T_i)\right]$$

$$= \frac{R_1 - 1}{R_1^3} \left(\sum_{j=1}^{n_i} E_\xi[(h - E_\xi[h])^2(n_iN_{i,j} - T_i)^2]\right) + \frac{(R_1 - 1)(R_1 - 2)(R_1 - 3)}{R_1^4} \left(\sum_{j=1}^{n_i} \hat{G}_i^2(X_{i,j})\right)^2 + \frac{2(R_1 - 1)(R_1 - 2)}{R_1^4} \sum_{j,k=1}^{n_i} E_\xi[(h - E_\xi[h])^2(n_iN_{i,j} - T_i)^2]\hat{G}_i^2(X_{i,k}) + O_p\left(\frac{n_i^4}{R_1^3} + \frac{n_i^3}{R_1^3} + \frac{n_i^2}{R_1}\right).$$

Hence the variance

$$Var_\xi\left[\sum_{j=1}^{n_i} \hat{G}_i^2\right] = E_\xi\left[\left(\sum_{j=1}^{n_i} \hat{G}_i^2\right)^2\right] - (E_\xi\left[\sum_{j=1}^{n_i} \hat{G}_i^2\right])^2$$

$$= \frac{1}{R_1^4} \left(\sum_{j=1}^{n_i} E_\xi[(h - E_\xi[h])^2(n_iN_{i,j} - T_i)^2]\right) + \frac{4R_1 - 6}{R_1^4} \left(\sum_{j=1}^{n_i} \hat{G}_i^2(X_{i,j})\right)^2 - \frac{4(R_1 - 1)}{R_1^4} \sum_{j,k=1}^{n_i} E_\xi[(h - E_\xi[h])^2(n_iN_{i,j} - T_i)^2]\hat{G}_i^2(X_{i,k}) + O_p\left(\frac{n_i^4}{R_1^3} + \frac{n_i^3}{R_1^3} + \frac{n_i^2}{R_1}\right)$$

$$= O_p\left(\frac{n_i^4}{R_1^3} + \frac{n_i^3}{R_1^3} + \frac{n_i^2}{R_1}\right).$$

It remains to study the variance of the bias correction $n_i T_i \hat{\sigma}^2_t / R_1$. Since $\hat{\sigma}^2_t$ is the sample variance, we have $Var_\xi[\hat{\sigma}^2_t] \leq E_\xi[(h - E_\xi[h])^4]/R_1$, which is $O_p(1/R_1)$ when Assumption 2 holds with $k = 4$.

The total variance of $\hat{\sigma}^2_t$ can thus be characterized as

$$Var_\xi[\hat{\sigma}^2_t] \leq (m + 1) \left(\sum_{i=1}^{m} Var_\xi\left[\frac{\sum_{j=1}^{n_i} \hat{G}_i^2}{n_i^2}\right] + Var_\xi\left[\frac{\sum_{i=1}^{m} T_i \hat{\sigma}^2_t}{R_1}\right]\right).$$
(m + 1) \left( \sum_{i=1}^{m} \frac{1}{n_i} O_p \left( \frac{n_i}{R_i^3} \right) + \frac{1}{R_i} \right) \\
= O_p \left( \frac{1}{R_i^3} \sum_{i=1}^{m} \frac{1}{n_i} + \frac{1}{R_i} \sum_{i=1}^{m} \frac{1}{n_i^2} \right). 

(\text{EC.50})

From \text{(EC.49)} and \text{(EC.50)} we obtain the mean squared error

\[ \mathbb{E}_\xi \left[ \left( \hat{\sigma}^2 - \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{G}_i(X_{ij})^2/n_i^2 \right)^2 \right] = O_p \left( \frac{1}{R_i^3} \sum_{i=1}^{m} \frac{1}{n_i} + \frac{1}{R_i} \sum_{i=1}^{m} \frac{1}{n_i^2} \right). \]

When \( R_1 \) is chosen according to \( \text{(3)} \), one can easily see that the mean squared error of \( \hat{\sigma}^2 \) is \( o_p((\sum_{i=1}^{m} 1/n_i)^2) \), which guarantees \( \hat{\sigma}^2 - \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{G}_i(X_{ij})^2/n_i^2 = o_p(\sum_{i=1}^{m} 1/n_i) \).

Up to now we have seen the asymptotic equivalence of confidence bounds based on EL and the delta method, and shown that consistent estimate of the input-induced variance is available. To finally justify the adjustment we need the following two standard results on weak convergence.

**Lemma EC.10.** (Uniform convergence of measures, adapted from Theorem 4.2 of \textit{Rad} (1962)) Let \( \mu^*, \{\mu_n\}_{n=1}^{\infty} \) be probability measures on \( \mathbb{R}^d \). If \( \mu^* \) is absolutely continuous with respect to the Lebesgue measure on \( \mathbb{R}^d \), then \( \mu_n \) converges weakly to \( \mu^* \) if and only if

\[ \lim_{n \to \infty} \sup_{C \in \mathcal{C}} |\mu_n(C) - \mu^*(C)| = 0, \]

where \( \mathcal{C} \) denotes the set of all measurable convex sets.

**Lemma EC.11.** (Berry-Esseen bound, adapted from \textit{Durrett} (2010)) Let \( \{Y_i\}_{i=1}^{\infty} \) be a sequence of i.i.d. variables such that \( \mathbb{E}[Y_1] = 0, \mathbb{E}[Y_1^2] = \sigma_Y^2, \mathbb{E}[Y_1^3] = \rho_Y < \infty, \) and \( S_N = \sum_{i=1}^{N} Y_i \). Then for any \( t \in \mathbb{R} \) and \( N \geq 9\sigma_Y^2 t^2/(16\rho_Y^3) \) it holds

\[ |\mathbb{E}[e^{itS_N/(\sigma_Y \sqrt{N})}] - e^{-t^2/2}| \leq \left( \frac{2|t|^3}{9} + \frac{t^4}{18} \right) e^{-t^2/4} \frac{3\rho_Y}{4\sigma_Y^3 \sqrt{N}}. \]

Now we proceed to the main proof.

**Proof of Corollary 4.** Denote \( Z^* = Z(P_1, \ldots, P_m), \ Z_{\min} = Z(w_{1, \min}, \ldots, w_{m, \min}) \), and \( Z_{\max} = Z(w_{1, \max}, \ldots, w_{m, \max}) \). The coverage guarantee is based on the following relation

\[ P(\mathcal{L}^\text{FEL}_\alpha \leq Z^* \leq \mathcal{H}^\text{FEL}_\alpha) \]

\[ = P(\mathcal{L}^\text{FEL}_\alpha \leq Z^*) + P(Z^* \leq \mathcal{H}^\text{FEL}_\alpha) - P(\mathcal{L}^\text{FEL}_\alpha \leq Z^* \text{ or } Z^* \leq \mathcal{H}^\text{FEL}_\alpha) \]

\[ = P(\mathcal{L}^\text{FEL}_\alpha \leq Z^*) + P(Z^* \leq \mathcal{H}^\text{FEL}_\alpha) - 1 + P(\mathcal{H}^\text{FEL}_\alpha < Z^* < \mathcal{L}^\text{FEL}_\alpha). \]

(\text{EC.51})

To compute the probabilities above, consider the following representation which can justified by Lemma \text{EC.7}, Theorem \text{9} and the conditions on \( n_i, R_1 \)

\[ \mathcal{L}^\text{FEL}_\alpha = \hat{Z}_{\min} - z_{1-\alpha/2} \left( \hat{\sigma}^2 + \frac{\hat{\sigma}_{\min}^2}{R_2} - \hat{\sigma}_1 \right) \]
\[ \mathcal{L}_\alpha + \hat{Z}_{\min} - Z_{\min} - z_{1-\alpha/2} \left( \sqrt{\hat{\sigma}_1^2 + \frac{\hat{\sigma}_{\min}^2}{R_2}} - \hat{\sigma}_I \right) + o_p\left( \sqrt{\sum_{t \in I} \bar{u}_i} \right) \]

\[ = Z^* + \sum_{i=1}^m \tilde{G}_i + (\hat{Z}_{\min} - Z_{\min}) - z_{1-\alpha/2} \left( \sqrt{\hat{\sigma}_1^2 + \frac{\hat{\sigma}_{\min}^2}{R_2}} - \hat{\sigma}_I + \sigma_I \right) + o_p(\sigma_I), \]

where \( \sigma_I^2 = \sum_{i=1}^m \text{Var}(G_i(X_i))/n_i \) for convenience. We know \( \hat{\sigma}_I^2/\sigma_I^2 \to_p 1 \) according to Lemma EC.8, EC.9. We would like to argue that \( \hat{\sigma}_{\min}^2 \to_p \sigma^2 \). Denote by \( \sigma_{\min}^2, \mu_{\min}^4 \) the variance and fourth central moment of \( h \) under input models \( \{w_i^{\min}\} \). When Assumption \( \Box \) holds with \( k = 8 \), Lemma EC.3 implies that \( \mu_{\min}^4 \to_p \mathbb{E}_P \ldots \mathbb{E}_P (h - Z^{*4}) \). Therefore the error \( \hat{\sigma}_{\min}^2 - \sigma_{\min}^2 = O_p(\sqrt{\mu_{\min}^4/R_2}) = O_p(1/\sqrt{R_2}) \) on one hand. On the other hand \( \sigma_{\min}^2 \to_p \sigma^2 \) according to Lemma EC.4 hence \( \hat{\sigma}_{\min}^2 \to_p \sigma^2 \) results. Because of the consistency of \( \hat{\sigma}_I^2 \) and \( \sigma_{\min}^2 \), the above representation can be simplified as

\[ \mathcal{L}_{\alpha}^{FEL} = Z^* + \sum_{i=1}^m \tilde{G}_i + (\hat{Z}_{\min} - Z_{\min}) - z_{1-\alpha/2} \sqrt{\hat{\sigma}_1^2 + \frac{\sigma_{\min}^2}{R_2}} + o_p\left( \sqrt{\frac{\sigma_I^2 + 1}{R_2}} \right). \]

The counterpart for the upper bound is

\[ \mathcal{U}_{\alpha}^{FEL} = Z^* + \sum_{i=1}^m \tilde{G}_i + (\hat{Z}_{\max} - Z_{\max}) + z_{1-\alpha/2} \sqrt{\hat{\sigma}_1^2 + \frac{\sigma_{\max}^2}{R_2}} + o_p\left( \sqrt{\frac{\sigma_I^2 + 1}{R_2}} \right). \]

The whole proof hinges on these two representations and the following joint weak limit

\[ (X_G, \hat{X}_{\min}, \hat{X}_{\max}) := \left( \frac{\sum_{i=1}^m \tilde{G}_i}{\sigma_I}, \frac{\sqrt{R_2}(\hat{Z}_{\min} - Z_{\min})}{\sigma_{\min}}, \frac{\sqrt{R_2}(\hat{Z}_{\max} - Z_{\max})}{\sigma_{\max}} \right) \to \mathcal{N}(0, \mathbf{I}_3). \]

We justify the weak limit by checking point-wise convergence of the characteristic function. Denote by \( \rho_{\min}, \rho_{\max} \) the third absolute central moments of output \( h \) under input models \( \{w_i^{\min}\} \) and \( \{w_i^{\max}\} \). When Assumption \( \Box \) holds with \( k = 6 \), applying Lemma EC.3 to \( |h - Z^{*3}| \) shows that \( \rho_{\min}, \rho_{\max} \to_p \mathbb{E}_P \ldots \mathbb{E}_P |h - Z^{*3}| \). Similarly \( \sigma_{\min}^2, \sigma_{\max}^2 \to_p \sigma^2 \). Denote by \( \xi_1 \) the randomness in Step 1 of Algorithm \( \Box \). Using conditional independence of \( \hat{X}_{\min}, \hat{X}_{\max} \) and Lemma EC.11 gives

\[ \mathbb{E}[e^{it_1 X_G + t_2 \hat{X}_{\min} + t_3 \hat{X}_{\max}}] = \mathbb{E}\left[ \mathbb{E}[e^{it_1 X_G + t_2 \hat{X}_{\min} + t_3 \hat{X}_{\max}} | \text{data}, \xi_1] \right] = \mathbb{E}\left[ e^{it_1 X_G(e^{-t_2^2/2} + O_p(R_2^{-1/2}))(e^{-t_3^2/2} + O_p(R_2^{-1/2}))} \right] \rightarrow e^{-(t_1^2 + t_2^2 + t_3^2)/2}. \]

The following joint weak limit then follows from Slutsky’s theorem

\[ (X_G, X_{\min}, X_{\max}) := \left( \frac{\sum_{i=1}^m \tilde{G}_i}{\sigma_I}, \frac{\sqrt{R_2}(\hat{Z}_{\min} - Z_{\min})}{\sigma_{\min}}, \frac{\sqrt{R_2}(\hat{Z}_{\max} - Z_{\max})}{\sigma_{\max}} \right) \to \mathcal{N}(0, \mathbf{I}_3). \]

Expressed in terms of \( (X_G, X_{\min}, X_{\max}) \), the confidence bounds are

\[ \mathcal{L}_{\alpha}^{FEL} = Z^* + \sigma_1 X_G + \frac{\sigma}{\sqrt{R_2}} X_{\min} - z_{1-\alpha/2} \sqrt{\frac{\sigma_1^2 + \sigma_{\min}^2}{R_2}} + o_p\left( \sqrt{\frac{\sigma_I^2 + 1}{R_2}} \right) \]

\[ \mathcal{U}_{\alpha}^{FEL} = Z^* + \sigma_1 X_G + \frac{\sigma}{\sqrt{R_2}} X_{\max} + z_{1-\alpha/2} \sqrt{\frac{\sigma_1^2 + \sigma_{\max}^2}{R_2}} + o_p\left( \sqrt{\frac{\sigma_I^2 + 1}{R_2}} \right). \]
Now we are able to compute the probabilities in (EC.51)

\[
P(\mathcal{L}_{\alpha}^{FEL} \leq Z^*) = P\left(\frac{\sigma I}{\sqrt{\sigma_I^2 + \sigma^2/R_2}} X_{\min} - z_{1-\alpha/2} \sqrt{\frac{\sigma_I^2}{\sigma_I^2 + \sigma^2/R_2}} + o_p(1) \leq 0 \right)
\]

\[
= P\left(\frac{\sigma I}{\sqrt{\sigma_I^2 + \sigma^2/R_2}} X_{\min} - \frac{\sigma}{\sqrt{\sigma_I^2 + \sigma^2/R_2}} \sqrt{\frac{\sigma_I^2}{\sigma_I^2 + \sigma^2/R_2}} \leq z_{1-\alpha/2}\right).
\]

Note that the \(o_p(1)\) term can be absorbed into \((X_G, X_{\min}, X_{\max})\) without affecting its convergence to standard normal \(\mathcal{N}(0, I_3)\), so one can neglect it when computing limits of probabilities. Although the coefficients of \(X_G, X_{\min}\) can vary as \(n_i, R_2\) grows, Lemma EC.10 saves the validity of the following limit

\[
P(\mathcal{L}_{\alpha}^{FEL} \leq Z^*) \to P\left(\frac{\sigma I}{\sqrt{\sigma_I^2 + \sigma^2/R_2}} Z_1 + \frac{\sigma}{\sqrt{\sigma_I^2 + \sigma^2/R_2}} Z_2 \leq z_{1-\alpha/2}\right) = P\left(\mathcal{N}(0, 1) \leq z_{1-\alpha/2}\right) = 1 - \frac{\alpha}{2}
\]

where \(Z_1, Z_2\) are independent standard normal variables. Similarly \(P(Z^* \leq \mathcal{L}_{\alpha}^{FEL}) \to 1 - \alpha/2\). To treat the last probability in (EC.51), let \(Z_1, Z_2, Z_3\) be independent standard normal variables

\[
P(\mathcal{L}_{\alpha}^{FEL} < Z^* < \mathcal{L}_{\alpha}^{FEL}) \to P\left(\frac{\sigma I}{\sqrt{\sigma_I^2 + \sigma^2/R_2}} Z_1, -\frac{\sigma}{\sqrt{\sigma_I^2 + \sigma^2/R_2}} Z_3 > z_{1-\alpha/2}\right) \to 1 - \frac{\alpha}{2}
\]

where \((\tilde{Z}_1, \tilde{Z}_2)\) is the joint normal \(\mathcal{N}\left(0, \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}\right)\), \(\rho = \sigma^2/(\sigma^2_I + \sigma^2/R_2) > 0\). It’s well known that the conditional distribution \(\tilde{Z}_2|\tilde{Z}_1\) is \(\mathcal{N}(-\rho \tilde{Z}_1, 1 - \rho^2)\), therefore

\[
P(\tilde{Z}_1 > z_{1-\alpha/2}, \tilde{Z}_2 > z_{1-\alpha/2}) = \int_{z_{1-\alpha/2}}^{\infty} \phi(x) P(\mathcal{N}(-\rho x, 1 - \rho^2) > z_{1-\alpha/2}) dx \leq \frac{\alpha}{2} \int_{z_{1-\alpha/2}}^{\infty} \phi(x) dx = \frac{\alpha^2}{4}.
\]

Here \(\phi\) denotes the density of standard normal. This establishes

\[
\limsup P(\mathcal{L}_{\alpha}^{FEL} < Z^* < \mathcal{L}_{\alpha}^{FEL}) \leq \frac{\alpha^2}{4}.
\]

Now we go back to (EC.51) and conclude that

\[
\liminf P(\mathcal{L}_{\alpha}^{FEL} \leq Z^* \leq \mathcal{L}_{\alpha}^{FEL}) \geq 1 - \alpha, \quad \limsup P(\mathcal{L}_{\alpha}^{FEL} \leq Z^* \leq \mathcal{L}_{\alpha}^{FEL}) \leq 1 - \alpha + \frac{\alpha^2}{4}.
\]

This completes the proof. \(\square\)

Theorems 2 and 3 and Corollaries 2 and 3 are all straightforward consequences of Lemma EC.7 or Corollary 4. Corollary 2 follows from Theorem 1, Lemma EC.7 and the delta method. Corollary 3 holds because the bounds proposed there are consistently more conservative than what Corollary 4 proposes, hence the lower bound of coverage probability remains valid. Theorems 2 and 3 are consequences of Corollaries 3 and 4 respectively after including the stochastic error incurred in Step 1 of Algorithm 1.
EC.6. Other Proofs

Proof of Proposition 1. It suffices to prove the theorem for the minimization problem. Since $w_{i,j} = \frac{1}{n_i}$ is in the interior of the positive orthant $[0, +\infty)^{\sum_{i=1}^{m} n_i}$, and the constraints are satisfied with strict inequality at this point, Slater’s conditions holds for program 2. It is also clear, by a compactness argument, that the optimal value of the program is finite and can be attained. By Corollary 28.3.1 of Rockafellar (2015), ($a$ compactness argument, that the optimal value of the program is finite and can be attained. By Corollary 28.3.1 of Rockafellar (2015), ($w_1^{\min}, \ldots, w_m^{\min}$) is a minimizer if and only if there exist $\beta^*, \lambda_i^* \in \mathbb{R}, i = 1, \ldots, m$ such that the KKT conditions are satisfied

$$2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log n_i w_{i,j}^{\min} + \lambda_{i,1}^{2} \geq 0, \beta^* \geq 0$$

$$\beta^* \left( 2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log n_i w_{i,j}^{\min} + \lambda_{i,1}^{2} \right) = 0$$

$$\sum_{j=1}^{n_i} w_{i,j}^{\min} = 1 \text{ for all } i$$

$$\hat{G}_{i,j} + \lambda_i^* - \frac{2 \beta^*}{w_{i,j}^{\min}} = 0 \text{ for all } i, j$$

When $\hat{G}_{i_0,j_1} \neq \hat{G}_{i_0,j_2}$ for some $1 \leq i_0 \leq m$ and $1 \leq j_1 < j_2 \leq n_{i_0}$, the objective is a non-constant linear function and thus any minimizer must lie on the boundary of the feasible set, i.e. $2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log n_i w_{i,j}^{\min} + \lambda_{i,1}^{2} = 0$. Since the constraint $-2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log n_i w_{i,j}^{\min}$ is strictly convex, the minimizer must be unique. Moreover, we shall show that $\beta^*$ must be strictly positive in this case. Suppose $\beta^* = 0$ then the last equation of KKT conditions requires $\hat{G}_{i,j} = -\lambda_i^*$ for all $i, j$, which is a contradiction. Note that the minimizer have $w_{i,j}^{\min} > 0$, so it is characterized by

$$w_{i,j}^{\min} = \frac{2 \beta^*}{\hat{G}_{i,j} + \lambda_i^*}, \beta^* > 0, \hat{G}_{i,j} + \lambda_i^* > 0 \text{ for all } i, j,$$

$$2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log \frac{2 n_i \beta^*}{\hat{G}_{i,j} + \lambda_i^*} + \lambda_{i,1}^{2} = 0, \sum_{j=1}^{n_i} \frac{2 \beta^*}{\hat{G}_{i,j} + \lambda_i^*} = 1 \text{ for all } i$$

This justifies the procedure proposed in the proposition. To show that such $(\beta^*, \lambda_1^*, \ldots, \lambda_m^*)$ is also unique, let $i_0, j_1, j_2$ be the indices mentioned in the theorem, then \[ \text{stipulates } w_{i_0,j_1}^{\min} / w_{i_0,j_2}^{\min} = (\hat{G}_{i_0,j_2} + \lambda_{i_0}^*) / (\hat{G}_{i_0,j_1} + \lambda_{i_0}^*) \]. Since the right hand side is strictly monotone in $\lambda_{i_0}^*$, the uniqueness of $w_{i,j}^{\min}$ implies the uniqueness of $\lambda_{i_0}^*$, which in turn implies the uniqueness of $\beta^*$ and other $\lambda_i^*$ due to the second equation of line 5.\[ \square \]

Proof of Proposition 2. Let $x_i = (x_{i,1}, \ldots, x_{i,T})$. First we rewrite the objective as a multiple integral

$$Z((1 - \epsilon)Q_1^1 + \epsilon Q_1^2, \ldots, (1 - \epsilon)Q_m^1 + \epsilon Q_m^2)$$
\[= \int h(x_1, \ldots, x_m) \prod_{i=1}^{m} \prod_{t=1}^{T_i} d(Q_i^t + \epsilon(Q_i^t - Q_i^1))(x_{i,t})\]
\[= Z(Q_1^1, \ldots, Q_m^1) + \sum_{i=1}^{m} \sum_{t=1}^{T_i} \epsilon \int h(x_1, \ldots, x_m) \prod_{r \neq i \text{ or } s \neq t} dQ_i^r(x_{r,s}) \cdot d(Q_i^2 - Q_i^1)(x_{i,t}) + R\]

where we distribute \(Q_i^1\) and \(\epsilon(Q_i^2 - Q_i^1)\) in the measure product, and the remainder \(R\) includes all the terms that have an \(\epsilon^k, k \geq 2\). The absolute integrability condition guarantees that all the integrals above, including those in \(R\), are finite. Thus the limit of the difference quotient can be identified as

\[
\sum_{i=1}^{m} \sum_{t=1}^{T_i} \int h(x_1, \ldots, x_m) \prod_{r \neq i \text{ or } s \neq t} dQ_i^r(x_{r,s}) \cdot d(Q_i^2 - Q_i^1)(x_{i,t})
\]
\[= \sum_{i=1}^{m} \sum_{t=1}^{T_i} \left( \int h(x_1, \ldots, x_m) \prod_{r \neq i \text{ or } s \neq t} dQ_i^r(x_{r,s}) \cdot dQ_i^2(x_{i,t}) - Z(Q_1^1, \ldots, Q_m^1) \right)
\]
\[= \sum_{i=1}^{m} \sum_{t=1}^{T_i} \left( \int h(x_1, \ldots, x_m) \prod_{r \neq i \text{ or } s \neq t} dQ_i^r(x_{r,s}) - Z(Q_1^1, \ldots, Q_m^1) \right) dQ_i^2(x_{i,t})
\]
\[= \sum_{i=1}^{m} \int \sum_{t=1}^{T_i} \left( \int h(x_1, \ldots, x_m) \prod_{r \neq i \text{ or } s \neq t} dQ_i^r(x_{r,s}) - Z(Q_1^1, \ldots, Q_m^1) \right) dQ_i^2(x_{i,t})
\]
\[= \sum_{i=1}^{m} \int G_i^{Q_1^1, \ldots, Q_m^1}(x) dQ_i^2(x) = \sum_{i=1}^{m} \mathbb{E}_{Q_i^1} \mathbb{E}_{Q_i^1, \ldots, Q_m^1}[h(X_1, \ldots, X_m)|X_i(t)] = Z(Q_1^1, \ldots, Q_m^1)
\]

for all \(t = 1, \ldots, T_i\).

**Proof of Theorem 7.** By applying Theorem 6 to \(h_i(X_{i,j})\), we know that the set \(\{\mu \in \mathbb{R} | -2 \log R(\mu) \leq X_i^2 \_\alpha\}\) contains the true sum \(\sum_{i=1}^{m} \mathbb{E}[h_i(X_i)]\) with probability \(1 - \alpha\) asymptotically. Note that this set can be identified as

\[\mathcal{V} = \left\{ \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{i,j} h_i(X_{i,j}) \mid -2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log n_i w_{i,j} \leq X_i^2 \_\alpha, \sum_{j=1}^{n_i} w_{i,j} = 1 \text{ for all } i, w_{i,j} \geq 0 \text{ for all } i, j \right\}.
\]

It is obvious that \(L_\alpha/U_\alpha = \min / \max \{\mu | \mu \in \mathcal{V}\}\), and that the feasible set \(\mathcal{A}_\alpha\) is compact. So if the set \(\mathcal{V}\) is convex, then \(\mathcal{V} = [L_\alpha, U_\alpha]\), and we conclude the theorem. To show convexity, it is enough to notice that \(\mathcal{A}_\alpha\) is convex, and the objective is linear in \(w_{i,j}\).

**Proof of Theorem 8.** This is a straightforward application of Theorem 6.