Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

A SIMULATION-BASED PREDICTION FRAMEWORK FOR STOCHASTIC SYSTEM DYNAMIC RISK MANAGEMENT

Wei Xie

Mechanical and Industrial Engineering
Northeastern University
360 Huntington Avenue, 334 SN, Boston, MA 02115, USA

Pu Zhang

Industrial and Systems Engineering
Rensselaer Polytechnic Institute
110 8th St.
Troy, NY 12180, USA

Ilya O. Ryzhov

Decision, Operations and Information Technologies
University of Maryland
4322 Van Munching Hall
College Park, MD 20742-1815, USA

ABSTRACT

We propose a simulation-based prediction framework which can quantify the prediction uncertainty of system future response and further guide operational decisions for complex stochastic systems. Specifically, by exploring the underlying generative process of real-world data streams, we first develop a nonparametric input model which can capture the important properties, including non-stationarity, skewness, component-wise and time dependence. It can improve the prediction accuracy, and the posterior predictive distribution can quantify the prediction uncertainty accounting for both input and stochastic uncertainties. Then, we propose the simulation-based prediction framework which can efficiently search for the optimal operational decisions hedging against the prediction uncertainty and minimizing the expected cost occurring in the planning horizon. The empirical study demonstrates that our approach has promising performance.

1 INTRODUCTION

The proposed simulation-based prediction framework can be applied to many application domains, such as manufacturing, health care and service. In this paper, we use the bio-pharmaceutical supply chain management as an illustration example. There are various challenges for biopharma supply chain management (Ma 2011). First, there exists *high uncertainty* in supply, production, testing and demand. The quality and ordering lead time of some key raw materials often have large variation. The production yield and cycle time have high fluctuation due to contamination and cross-contamination. The demand of clinical products is hard to predict. Second, the bio-pharmaceutical supply chains tend to be *complex* since many clinical and commercial products with totally different demand patterns share the same inventory, production and testing resources. Third, there is *rapid change* in the technology and market. The product lifetime is usually short and new products are frequently launched. At the same time, the internet-connected data collection devices can result in the availability of rich real-world data streams, such as sensor and barcode data, which have the potential to provide the current status of supply chains and production processes.

Since the bio-pharmaceutical manufacturing is a life saving industry, it requires a close to 100% service level. For the complex and dynamic biopharma supply chains with huge uncertainty in supply, testing, production and demand, *coherent and fast* decision making in inventory control, testing and production scheduling are extremely important in order to hedge against the impact of the uncertainties and guarantee the on-time product delivery. Since it is challenging to analytically assess the random behaviors of complex stochastic systems, the simulation has become an important tool for the design of supply chains and

production processes. However, the simulation methodologies developed to guide the dynamic operational decisions are still open.

To support the dynamic decision making, the simulation can be used to *predict* the system future response. To provide the reliable and cost-efficient decisions, it is important to improve the prediction accuracy of system performance in the planning horizon and correctly quantify the prediction uncertainty. The stochastic behaviors of outputs depend on the choice of input models, defined as stochastic processes used to drive simulation experiments. In the many real applications, we often have input processes with component-wise and time dependence. For example, in the bio-pharmaceutical supply chains, the demands of different drugs often depend on each other. Also, there exists the time dependence in the demand process since patients usually take the drug for several cycles. To improve the prediction of system performance, we need to faithfully capture the important properties of real-world data streams.

Several approaches have been proposed in the simulation literature to capture the dependence. For example, Cario and Nelson (1998) proposed an Autoregressive-to-Anything (ARTA) Processes to model the stationary time dependence. Biller and Nelson (2005) fitted ARTA processes with the Johnson translation system (JST) marginal distributions to the moments of real-world data. Biller and Nelson (2008) extended ARTA to Vector Autoregressive-to-Anything (VARTA), which can model both the component-wise and time series dependence. However, the stationary VARTA process can not capture the non-stationarity in the real-world data streams. In many situations, ignoring the non-stationarity could lead to unfounded estimation of the system response (Harrod and Kelton 2006).

Various approaches have been proposed to model non-stationary input models with time dependence. For example, Kuhl and Wilson (2001) introduced a non-homogeneous Poisson Process (NHPP) to model the non-stationary arrival process. The long-term trend and nested cyclic behavior within small cycle length are modeled. Gerhardt and Nelson (2009) proposed the non-stationary and non-Poisson (NSNP) arrival process, and they modeled the non-stationarity by a rate function with pre-specified forms. They further proposed the non-stationary and nonrenewal (NSNR) arrival process in Nelson and Gerhardt (2011). While the non-stationarity is captured by these approaches, they mainly focus on arrival processes.

The non-stationary dynamic behaviors of input data are often induced by some latent states or factors. For example, in bio-pharma supply chain, different drugs could have similar effects on the treatment for certain diseases, such as flu, which could occur periodically and seasonally. Under different spread status of diseases, the demands for these drugs have different statistical behaviors. There exist both time and component-wise dependence among the drug demands. First, the course of treatment usually lasts a certain period, which introduces the time series dependence of the demands. Second, since the drugs could have similar effects on the treatment, their demands could be highly correlated with each other.

In addition, the input models are often estimated by finite real-world data. For example, since the product life time in the bio-pharmaceutical industry is usually short and there are limited demand data, the estimation uncertainty of input models, called *the input uncertainty*, could be large. Ignoring the input uncertainty could lead to unsound prediction of system future response. Thus, it is necessary to correctly quantify the impact from both input and stochastic uncertainties.

In this paper, to improve the prediction accuracy and correctly quantify the prediction uncertainty, we first propose a nonparametric input model, called *the Infinite Markov Switching Vector Autoregressive (IMS-VAR)*. A Markov process is used to model the state evolution, which can be interpreted as "global" time dependence. Under each state, Vector Autoregressive (VAR) is used to model the "local" dependence. IMS-VAR can capture the rich dynamic behaviors of real-world input data streams, including skewness, multimodality and dependence, and it can improve the prediction accuracy. Further, the posterior distributions of the flexible input models can correctly quantify the input uncertainty. *The posterior predictive distribution can quantify the prediction uncertainty induced by both input and stochastic uncertainties.*

Then, to quickly find the optimal operational decisions hedging against the prediction uncertainty, we propose a simulation-based prediction framework to guide the real-time operational decisions for complex stochastic systems. Since each simulation run could be computationally expensive, motivated by Fu (2006), *we propose a mini-batch gradient descent (GD) method that can efficiently employ the simulation resource*

to search for the optimal operational decisions. The simultaneous perturbation stochastic approximation (SPSA) proposed by Spall (1998) is used to efficiently estimate the gradient.

The contributions of this paper are summarized as follow.

- The nonparametric IMS-VAR input model can capture the important properties in the real-world data streams, including non-stationarity, skewness, component-wise and time series dependency. It can improve the prediction accuracy.
- The posterior distribution of flexible input models is used to quantify the input uncertainty, and then the posterior predictive distribution can correctly quantify the prediction risk induced by both input and simulation uncertainties.
- We propose the simulation-based prediction framework that can efficiently employ the simulation resource to search for the optimal operational decisions. The simulation experiments are driven by the posterior predictive distribution. Then, the mini-batch gradient descent method is used to quickly search for the optimal operational decision hedging against the prediction uncertainty.

In the next section, we provide the problem description and briefly introduce the simulation-based prediction framework. In Section 3, we first present the nonparametric Bayesian IMS-VAR model, and provide the inference and sampling procedure to generate scenarios of future inputs. Then, built on the simulation-based probabilistic prediction, a mini-batch stochastic gradient descent approach is introduced to efficiently and quickly find the optimal decisions. We study the finite sample performance of our input forecast model and simulation-based prediction framework in Section 4, and conclude this paper in Section 5.

2 PROBLEM DESCRIPTION AND PROPOSED APPROACH

We use the bio-pharmaceutical supply chain risk management as an illustrative example to describe the problem of interest. It is challenging to manage the biopharma supply chains because there exists high uncertainty in supply, testing, production and demand, and the systems need to evolve fast to be competitive. Thus, to construct a cost-efficient and reliable supply chain, it is critically important to find the real-time operational decisions hedging against various sources of uncertainty. Here, we want to find the optimal decisions minimizing the expected overall cost occurring during the planning horizon, including the inventory holding cost, the backorder cost for unsatisfied demands, and the production cost. Suppose that the current time period is T, and the planning horizon length is τ. Denote the overall cost occurring in the planning horizon as $\sum_{h=1}^{\tau} C_{T+h}(\mathbf{X}_{T+h}, \boldsymbol{\mu})$, where $\mathbf{X}_{T+h}$ represents the demands of d drugs realized in the $(T+h)$-th time period, $C_{T+h}(\mathbf{X}_{T+h}, \boldsymbol{\mu})$ is the cost and $\boldsymbol{\mu} \equiv \{\mu_1, \ldots, \mu_L\}$ denotes the operational decision. For simplification, suppose that the decision $\boldsymbol{\mu}$ is fixed during the planning horizon.

The future demand $\mathbf{X}_{T+h}$ can be predicted by using the historical data, denoted by $\mathbf{X}_{[1:T]} = (\mathbf{X}_1, \ldots, \mathbf{X}_T)$. Specifically, the unknown "correct" input model, denoted by F^c, can be estimated by using the historical data with the posterior distribution $p(F|\mathbf{X}_{[1:T]})$ quantifying the input uncertainty. Then, given any input model estimate F, the prediction distribution $p(\mathbf{X}_{[T+1:T+\tau]}|F)$ quantifies the prediction risk induced by stochastic uncertainty. Thus, the prediction uncertainty characterized by the posterior predictive distribution,

$$p(\mathbf{X}_{[T+1:T+\tau]}|\mathbf{X}_{[1:T]}) = \int p(\mathbf{X}_{[T+1:T+\tau]}|F) p(F|\mathbf{X}_{[1:T]}) dF,$$

can account for both input and stochastic uncertainties. Notice that we do not separately handle input and stochastic uncertainties because the estimation over underlying input models is mainly for predicting the future demands based on the information extracted from the historical data.

To hedge against both stochastic and input uncertainties, we present *a data-driven stochastic optimization*. Given any decision $\boldsymbol{\mu}$, a predictive distribution of the overall cost is

$$p\left(\sum_{h=1}^{\tau} C_{T+h}(\mathbf{X}_{T+h}, \boldsymbol{\mu}) \middle| \mathbf{X}_{[1:T]}\right) = \int p\left(\sum_{h=1}^{\tau} C_{T+h}(\mathbf{X}_{T+h}, \boldsymbol{\mu}) \middle| F\right) p(F|\mathbf{X}_{[1:T]}) dF.$$

Then, the objective is to find the optimal operational decision minimizing the expected cost occurring in the planning horizon,

$$\underset{\boldsymbol{\mu}\in\boldsymbol{\Omega}}{\text{minimize}} \quad \mathrm{E}\left(\sum_{h=1}^{\tau} C_{T+h}(\mathbf{X}_{T+h},\boldsymbol{\mu})\middle|\mathbf{X}_{[1:T]}\right)$$

where $\boldsymbol{\Omega}$ is a continuous and convex feasible region. Notice that differing with existing empirical stochastic optimization (which takes the input model estimate as the true one), robust optimization, and distributionally robust optimization, our approach can lead to decisions equally hedging against both input and stochastic uncertainties.

Since the bio-pharma supply chains could be complex with numerous uncertainty and there is no closed form expected future cost, simulation is used to predict the future system response. In this paper, we propose *a simulation-based prediction framework* which can quantify the overall prediction uncertainty for the future system response, and further provide the stochastic gradient based optimization approach that can efficiently employ the simulation resource to search for the optimal decisions hedging against the prediction uncertainty induced by both stochastic and input uncertainties. Specifically, we propose a non-parametric Bayesian approach that can capture the important properties of the real-world data streams. The posterior distribution of flexible input models, $p(F|\mathbf{X}_{[1:T]})$, can correctly quantify the input uncertainty. The scenarios of $\mathbf{X}_{[T+1:T+\tau]}$ generated from the posterior predictive distribution $p(\mathbf{X}_{[T+1:T+\tau]}|\mathbf{X}_{[1:T]})$ can quantify the prediction risk induced by both input and stochastic uncertainties, and they are used to drive the simulation experiments in the prediction framework. Then, the mini-batch stochastic gradient descent approach is used to efficiently and quickly search for the optimal real-time operational decision, denoted by $\boldsymbol{\mu}^{\star}$, minimizing the expected future cost.

3 A SIMULATION-BASED PREDICTION FRAMEWORK

In order to provide the cost-efficient and reliable operational decisions, we need to improve the future input forecasting and correctly quantify the prediction uncertainty. In many situations, the dynamic behaviors of input models are induced by some latent factors or states. For example, in bio-pharmaceutical supply chains, the product demands depend on states, such as the spread level of diseases. The demands under different states could have different dynamic behaviors, and the spread level also evolves with time. Thus, to improve the prediction accuracy, it is necessary to model the stochastic processes of latent states and also the dynamic behaviors of input data under each state.

Building on the univariate wind energy forecasting model in Xie et al. (2018), in Section 3.1, we first present a nonparametric Infinite Markov Switching Vector Autoregressive (IMS-VAR) model to capture the important properties in the real-world data streams, including non-stationarity, dependence, skewness and multi-modality. Given the historical data $\mathbf{X}_{[1:T]}$, we provide the Bayesian inference and a sampling procedure to generate the posterior samples of input model. Then, in Section 3.2, we introduce the simulation-based prediction framework. We generate the future scenarios for $\mathbf{X}_{[T+1:T+\tau]}$ from the posterior predictive distribution $p(\mathbf{X}_{[T+1:T+\tau]}|\mathbf{X}_{[1:T]})$ to drive the simulation experiments. After that, we develop the mini-batch gradient descent method that can efficiently employ the tight computational resource to search for the optimal operational decision $\boldsymbol{\mu}^{\star}$.

3.1 A Nonparametric IMS-VAR Input Model and Input Uncertainty Quantification

Let s_t be the latent state at time period t. Since there could be infinite potential state values and the current state typically depends on the previous one, we use an infinite hidden Markov model (IHMM) to model the state transition process. At each state, the dynamic behaviors of $\{\mathbf{X}_t\}$ is modeled by VAR time series with order p. Thus, given the historical data $\mathbf{X}_{[1:t]}$ and input model F, the probabilistic prediction density function of $\mathbf{X}_{t+1}$ is

$$f(\mathbf{X}_{t+1}|\mathbf{X}_{[1:t]},F)=\sum_{i=1}^{+\infty} p(s_{t+1}=i|\mathbf{X}_{[1:t]})h(\mathbf{X}_{t+1}|\boldsymbol{\theta}_{s_{t+1}},\mathbf{X}_{[1:t]},s_{t+1}=i)$$

$$
\begin{aligned}
&= \sum_{i=1}^{+\infty}\sum_{j=1}^{+\infty} p(s_{t+1}=i|s_t=j)p(s_t=j|\mathbf{X}_{[1:t]})h(\mathbf{X}_{t+1}|\boldsymbol{\theta}_{s_{t+1}},\mathbf{X}_{[1:t]},s_{t+1}=i)\\
&= \sum_{i=1}^{+\infty}\sum_{j=1}^{+\infty} p_{ij}p(s_t=j|\mathbf{X}_{[1:t]})h(\mathbf{X}_{t+1}|\boldsymbol{\theta}_{s_{t+1}},\mathbf{X}_{[1:t]},s_{t+1}=i)
\end{aligned}
$$

where the VAR time series, denoted by $h(\cdot)$, is specified by parameters $\boldsymbol{\theta}_i$ at state i and p_{ij} is the transition probability from state i to j. The flexible nonparametric IMS-VAR input model can capture the rich properties in the real-world data streams, including non-stationarity, multi-modality, skewness, component-wise and time series dependence.

Following Xie et al. (2018), in order to support the inference and implementation, a hierarchical Dirichlet process (HDP) is used to represent IHMM; see the introduction on HDP in Teh et al. (2006). A global Dirichlet Process (DP) is used to model the prior distribution of latent state, denoted by $G_0 \sim \mathrm{DP}(\alpha, G_{\boldsymbol{\theta}})$, where α is the concentration parameter and $G_{\boldsymbol{\theta}}$ represents the prior distribution for parameters $\boldsymbol{\theta}$. Let $G_0 = \sum_{\ell=1}^{+\infty}\pi_\ell\delta_{\boldsymbol{\theta}_\ell}$, where π_ℓ is the probability staying in state ℓ which has the dynamic behaviors of $\{\mathbf{X}_t\}$ characterized by VAR with parameters $\boldsymbol{\theta}_\ell$, and $\delta_{\boldsymbol{\theta}_\ell}$ denotes a Dirac function at $\boldsymbol{\theta}_\ell$. Thus, our prior belief on $\boldsymbol{\pi}$ has a stick-breaking representation, $\boldsymbol{\pi} \sim \mathrm{Stick}(\alpha)$. Then, since the possible state values are shared by variables s_t with $t = 1,2,\ldots$, a set of *state-conditional DP*, $G_i|G_0,\eta \sim \mathrm{DP}(\eta, G_0)$, is used to model the prior transition probabilities from current state i to the next state, denoted by $\boldsymbol{p}_i$, where η denotes the concentration parameter. Thus, the IMS-VAR model can be represented as

$$
\begin{aligned}
&\mathbf{X}_t = \boldsymbol{\phi}(s_t)\tilde{\mathbf{X}}_t + \boldsymbol{\varepsilon}_t\\
&\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \Sigma(s_t))\\
&\boldsymbol{\theta}_{s_t} \equiv \{\boldsymbol{\phi}(s_t), \Sigma(s_t)\} \sim G_{\boldsymbol{\theta}}\\
&s_t|s_{t-1}, \{\boldsymbol{p}_i\}_{i=1}^{+\infty} \sim \boldsymbol{p}_{s_{t-1}}\\
&\boldsymbol{p}_i|\boldsymbol{\pi} \sim \mathrm{DP}(\eta, \boldsymbol{\pi})\\
&\boldsymbol{\pi} \sim \mathrm{Stick}(\alpha)
\end{aligned}
\tag{1}
$$

where $\tilde{\mathbf{X}}_t = [1, \mathbf{X}_{t-1}^\top, \ldots, \mathbf{X}_{t-p}^\top]^\top$ is a vector with length $k = 1 + dp$, and $\boldsymbol{\theta}_{s_t}$ is the VAR parameters under the state s_t with $\boldsymbol{\phi}(s_t)$ denoting a $d \times k$ coefficient matrix and $\Sigma(s_t)$ denoting a $d \times d$ covariance matrix.

The Bayesian posterior inference for IMS-VAR input model in (1) is similar to that of IMS-AR proposed in Xie et al. (2018). Specifically, given the historical data $\mathbf{X}_{[1:T]}$, there are M active states, defined as those states visited by $\mathbf{X}_{[1:T]}$. Let $T_i = \{t : s_t = i\}$ be all the time periods when the state is i, let $\boldsymbol{c} = (c_1, \ldots, c_M)$ denote the counter vector with c_i recording the number of visits to state i, and let $\boldsymbol{N}$ denote the transition matrix with N_{ij} recording the number of transitions from state i to state j for $i, j = 1, \ldots, M$.

The conditional posterior for s_t is

$$
\begin{aligned}
&p(s_t = i|\mathbf{X}_{[1:T]}, \boldsymbol{\pi}, s_{t-1}, s_{t+1}, \boldsymbol{\theta}_i)\\
&= C_0 \times p(s_t = i|\boldsymbol{\pi}, s_{t-1})p(s_{t+1}|\boldsymbol{\pi}, s_t = i)p(\mathbf{X}_t|s_t = i, \boldsymbol{\theta}_i, \mathbf{X}_{[1:t-1]})\\
&= \begin{cases} C_0 \times \frac{\eta\pi_i + N_{s_{t-1},i}}{\eta + c_{s_{t-1}}} \frac{\eta\pi_{s_{t+1}} + N_{i,s_{t+1}}}{\eta + c_i} f_i(\mathbf{X}_t|\boldsymbol{\theta}_i), & 1 < t < T \\ C_0 \times \frac{\eta\pi_i + N_{s_{t-1},i}}{\eta + c_{s_{t-1}}} f_i(\mathbf{X}_t|\boldsymbol{\theta}_i), & t = T \end{cases}
\end{aligned}
\tag{2}
$$

and $p(s_1 = 1) = 1$, where $f_i(\mathbf{X}_t|\boldsymbol{\theta}_i) = \frac{1}{(2\pi)^{d/2}\Sigma(i)^{1/2}} \exp\left[-\frac{1}{2}\left(\mathbf{X}_t - \boldsymbol{\phi}(i)\tilde{\mathbf{X}}_t\right)^\top \Sigma(i)^{-1}\left(\mathbf{X}_t - \boldsymbol{\phi}(i)\tilde{\mathbf{X}}_t\right)\right]$ and C_0 is a normalizing constant shared by all $s_t = i$ to guarantee that $\sum_{i=1}^{M+1} p(s_t = i|\mathbf{X}_{[1:T]}, \boldsymbol{\pi}, s_{t-1}, s_{t+1}, \boldsymbol{\theta}_i) = 1$.

Given the prior $\boldsymbol{\pi} \sim \mathrm{Stick}(\alpha)$, the conditional posterior for $\boldsymbol{\pi}$ is derived by following (Teh, Jordan, Beal, and Blei 2006),

$$
\boldsymbol{\pi}|\mathbf{s}_{[1:T]} \sim \mathrm{Dirichlet}(c_1, \ldots, c_M, \alpha), \tag{3}
$$

where $\mathbf{s}_{[1:T]} = (s_1, \ldots, s_T)$.

Let $\boldsymbol{X}_{[T_i]}$ denote a $c_i \times d$ matrix of $\boldsymbol{X}_t$ with $t \in T_i$, and $\tilde{\boldsymbol{X}}(i)$ denote the corresponding $c_i \times k$ regressors matrix having each row to be $\tilde{\boldsymbol{X}}_t$ with $t \in T_i$. For the parameters $\boldsymbol{\theta}_i$ of the VAR model i, suppose the prior has the form of a normal inverted Wishart,

$$\begin{aligned} vec(\boldsymbol{\phi}(i))|\Sigma(i) &\sim \mathcal{N}(vec(\mu_\Phi), V_\Phi \otimes \Sigma(i)) \\ Sigma(i) &\sim \text{Inv-Wishart}(v, \Psi). \end{aligned}$$

The posterior can be derived following by Banbura et al. (2010),

$$\begin{aligned} vec(\boldsymbol{\phi}(i))|\mathbf{X}_{[1:T]}, \Sigma(i) &\sim \mathcal{N}(vec(D), A^{-1} \otimes \Sigma(i)) \\ \Sigma(i) &\sim \text{Inv-Wishart}(v + c_i, C) \end{aligned} \tag{4}$$

where $A = \tilde{\boldsymbol{X}}^\top(i)\tilde{\boldsymbol{X}}(i) + V_\Phi^{-1}$, $D = (\boldsymbol{X}_{[T_i]}^\top \tilde{\boldsymbol{X}} + \mu_\Phi V_\Phi^{-1})A^{-1}$, $C = \Psi + (\boldsymbol{X}_{[T_i]} - \tilde{\boldsymbol{X}}(i)D^\top)^\top(\boldsymbol{X}_{[T_i]} - \tilde{\boldsymbol{X}}(i)D^\top) + (D - \mu_\Phi)V_\Phi^{-1}(D - \mu_\Phi)^\top$.

With the conditional posterior distributions for input model parameters in (2), (3) and (4), following by the Gibbs sampling procedure provided in Xie et al. (2018), we can generate B posterior samples of input models $F^{(b)} \sim p(F|\mathbf{X}_{[1:T]})$ for $b = 1, \ldots, B$ to quantify the input model estimation uncertainty. At each $F^{(b)}$, we can generate scenarios of future inputs $\mathbf{X}_{[T+1:T+\tau]}$.

3.2 A Simulation-Based Prediction Framework for Dynamic Risk Management

In this section, we propose a simulation-based prediction framework to guide the operational decision which can hedge against the prediction risk induced by both input and stochastic uncertainties. Considering the different computational cost required to generate the posterior samples of input model and draw the random variates from each input model estimate, the proposed data-driven stochastic optimization approach combines the sample average approximation (SAA) and the stochastic gradient descent (SGD) method to efficiently find the optimal operational decision.

Denote the expected cost in the planning horizon with

$$g(\boldsymbol{\mu}) \equiv \mathrm{E}\left(\sum_{h=1}^{\tau} C_{T+h}(\mathbf{X}_{T+h}, \boldsymbol{\mu}) \middle| \mathbf{X}_{[1:T]}\right) = \int_F \mathrm{E}\left(\sum_{h=1}^{\tau} C_{T+h}(\mathbf{X}_{T+h}, \boldsymbol{\mu}) \middle| \mathbf{X}_{[1:T]}, F\right) p(F|\mathbf{X}_{[1:T]}) dF.$$

Following the Bayesian inference and Gibbs sampling procedure in Section 3.1, we can generate the posterior samples, $F^{(b)}$ for $b = 1, \ldots, B$, quantifying the input uncertainty. Comparing with drawing a sample path of $\mathbf{X}_{[T+1:T+\tau]}$ from $F^{(b)}$, it is computationally more expensive to generate each posterior sample of input model. Thus, according to Jian and Henderson (2015), SAA can be employed to approximate the expected cost in the planning horizon. Given B posterior samples of input model, the optimization problem is approximated as

$$\min_{\boldsymbol{\mu}} \bar{g}(\boldsymbol{\mu}) \equiv \frac{1}{B}\sum_{b=1}^{B} \mathrm{E}\left(\sum_{h=1}^{\tau} C_{T+h}(\mathbf{X}_{T+h}, \boldsymbol{\mu}) \middle| \mathbf{X}_{[1:T]}, F^{(b)}\right).$$

where $\mathrm{E}\left(\sum_{h=1}^{\tau} C_{T+h}(\mathbf{X}_{T+h}, \boldsymbol{\mu}) | \mathbf{X}_{[1:T]}, F^{(b)}\right)$ is the expected cost at $F^{(b)}$. As B goes to infinity, the objective $\bar{g}(\boldsymbol{\mu})$ converges to $g(\boldsymbol{\mu})$ under some regularity conditions.

Since the feasible region is continuous, according to Chau and Fu (2015), we consider a stochastic gradient descent based approach to quickly search for the optimal solution. At any decision $\boldsymbol{\mu}$, the gradient of $\bar{g}(\boldsymbol{\mu})$ is

$$\nabla \bar{g}(\boldsymbol{\mu}) = \frac{1}{B}\sum_{b=1}^{B} \nabla \mathrm{E}\left(\sum_{h=1}^{\tau} C_{T+h}(\mathbf{X}_{T+h}, \boldsymbol{\mu}) \middle| \mathbf{X}_{[1:T]}, F^{(b)}\right).$$

The gradient provides the direction to iteratively search for the optimal solution. Denote the decision obtained in the n-th iteration as $\boldsymbol{\mu}^n$. According to Bottou (2010), the gradient descent update is,

$$\boldsymbol{\mu}^n = \boldsymbol{\mu}^{n-1} - \eta_n \nabla \bar{g}(\boldsymbol{\mu}^{n-1}) \tag{5}$$

where η_n is the step size in the n-th update. According to Chee and Toulis (2017), the step size η_n could be constant or decrease over n. We use a constant step size in the empirical study.

Since there is no closed form expected cost in general and each simulation run could be computationally expensive, we estimate the gradient by simultaneous perturbation stochastic approximation (SPSA) introduced in Spall (1998). Only two simulation runs are needed to estimate the gradient, which is especially suitable for complex systems with high-dimensional decision space. Specifically, at the n-th iteration, given any posterior sample $F^{(b)}$, a sample path of the future inputs $\mathbf{X}(F^{(b)}) \equiv \mathbf{X}_{[T+1:T+\tau]}$ can be generated from the distribution $p(\mathbf{X}_{[T+1:T+\tau]}|F^{(b)}, \mathbf{X}_{[1:T]})$. Two simulation runs at decisions $(\boldsymbol{\mu}^{n-1} + c_n\boldsymbol{\delta}_n)$ and $(\boldsymbol{\mu}^{n-1} - c_n\boldsymbol{\delta}_n)$ are used to estimate the gradient $\nabla \mathrm{E}\left(\sum_{h=1}^{\tau} C_{T+h}(\mathbf{X}_{T+h}, \boldsymbol{\mu}^{n-1}) \middle| \mathbf{X}_{[1:T]}, F^{(b)}\right)$ with the ℓ-th component for $\ell = 1, \ldots, L$

$$\frac{\widehat{\partial} \mathrm{E}\left(\sum_{h=1}^{\tau} C_{T+h}(\mathbf{X}_{T+h}, \boldsymbol{\mu}^{n-1}) \middle| \mathbf{X}_{[1:T]}, F^{(b)}\right)}{\partial \mu_\ell} = \frac{y(\mathbf{X}(F^{(b)}), \boldsymbol{\mu}^{n-1} + c_n\boldsymbol{\delta}_n) - y(\mathbf{X}(F^{(b)}), \boldsymbol{\mu}^{n-1} - c_n\boldsymbol{\delta}_n)}{2c_n\delta_{n,\ell}} \tag{6}$$

where the response $y(\mathbf{X}(F^{(b)}), \boldsymbol{\mu}) \equiv \sum_{h=1}^{\tau} C_{T+h}(\mathbf{X}_{T+h}, \boldsymbol{\mu})$ is obtained from the simulation output, $\boldsymbol{\delta}_n$ is an L-dimensional random vector with $\delta_{n,\ell}$ representing the ℓ-th element, c_n decays over n. Following Spall (1998), in the empirical study, each component $\delta_{n,\ell}$ follows a Bernoulli ± 1 distribution with probability of $1/2$ and we select $c_n = 1/n^{1/6}$. In addition, to reduce the gradient estimation variance, the common random number is applied at the two simulation runs in the gradient estimate in (6) (Kleinman et al. 1999).

Thus, we can efficiently employ the computational resource and speed up the operational decision making. Suppose that the total budget is R simulation runs. If the gradient in (5) is estimated by using all B posterior samples of input models, only $N = \lfloor R/2B \rfloor$ decision updates can be performed, which could lead to a large optimality gap. According to Ruder (2016), we consider a mini-batch gradient descent method where each update uses only part of B posterior samples of input models to estimate the expectation and gradient. Specifically, in each iteration, we randomly select a batch of $B_0 \leq B$ samples from $F^{(1)}, \ldots, F^{(B)}$. Denote the index set of selected B_0 samples as $\mathbf{S}$, and use these samples to estimate the gradient,

$$\widehat{\nabla} \bar{g}(\boldsymbol{\mu}^{n-1}) = \frac{1}{B_0} \sum_{b \in \mathbf{S}} \widehat{\nabla} \mathrm{E}\left(\sum_{h=1}^{\tau} C_{T+h}(\mathbf{X}_{T+h}, \boldsymbol{\mu}^{n-1}) \middle| \mathbf{X}_{[1:T]}, F^{(b)} \right) \tag{7}$$

where the gradient estimate $\widehat{\nabla}\mathrm{E}$ can be obtained by applying (6). In the mini-batch gradient descent method with batch size B_0, there are $N = \lfloor R/2B_0 \rfloor$ updates. There is a trade-off. With large B_0, the estimate of the gradient is more accurate in each update and less updates can be performed.

Then, by replacing $\nabla \bar{g}(\boldsymbol{\mu})$ in (5) with the estimated gradient $\widehat{\nabla} \bar{g}(\boldsymbol{\mu})$ in (7), the update at the n-th iteration can be performed,

$$\boldsymbol{\mu}^n = \boldsymbol{\mu}^{n-1} - \eta_n \widehat{\nabla} \bar{g}(\boldsymbol{\mu}^{n-1}). \tag{8}$$

Notice that the updated decision $\boldsymbol{\mu}^n$ could be outside the feasible set $\boldsymbol{\Omega}$. In such situation, according to Calamai and Moré (1987), the solution can be projected back to $\boldsymbol{\Omega}$ by $\boldsymbol{\mu}^n = \Pi(\boldsymbol{\mu}^n)$, where $\Pi(\boldsymbol{\mu}) = \underset{z \in \boldsymbol{\Omega}}{\operatorname{argmin}} ||z - \boldsymbol{\mu}||$.

Algorithm 1 describes the detailed procedure to find the optimal decision based on mini-batch SGD, where the gradient is estimated by simulation accounting for both input and stochastic uncertainties. The algorithm starts with initialization of a feasible solution and select the appropriate step sizes. During each iteration, B_0 posterior samples of input models are randomly drawn from $F^{(1)}, \ldots, F^{(B)}$ without replacement. At each $F^{(b)}$ of selected input models, a sample path of $\mathbf{X}_{[T+1:T+\tau]}$ is generated, and use it to drive two simulation runs at decisions $(\boldsymbol{\mu}^{n-1} + c_n\boldsymbol{\delta}_n)$ and $(\boldsymbol{\mu}^{n-1} - c_n\boldsymbol{\delta}_n)$. Then, the gradient component can be

estimated by following (6). By taking average of the gradient at all B_0 selected input model samples according to (7), we get the estimated gradient to update the decision by applying (8), and project it to $\mathbf{\Omega}$ if necessary. Repeat this procedure until reaching to the simulation budget R and record the optimal solution $\widehat{\boldsymbol{\mu}}^{\star}$.

Algorithm 1: Procedure to Find the Optimal Operational Decision

Randomly initialize a feasible solution $\boldsymbol{\mu}^0$, select the batch size B_0, a sequence of step sizes $\{\eta_1,\ldots,\eta_N\}$ and $\{c_1,\ldots,c_N\}$, where the number of iterations $N=\lfloor R/2B_0\rfloor$.
for $n=1,\ldots,N$ **do**
  Randomly select a mini-batch with size B_0 from the B posterior samples of input models, denote the set of indices as $\mathbf{S}_B$.
  Randomly generate a permutation vector $\boldsymbol{\delta}_n$.
  for $b\in\mathbf{S}_B$ **do**
    Generate a sample path of input variates $\mathbf{X}_{[T+1:T+\tau]}$ by using $F^{(b)}$, denoted by $\mathbf{X}(F^{(b)})$.
    Run simulations at decisions $(\boldsymbol{\mu}^{n-1}+c_n\boldsymbol{\delta}_n)$ and $(\boldsymbol{\mu}^{n-1}-c_n\boldsymbol{\delta}_n)$ driven by $\mathbf{X}(F^{(b)})$. Record the outputs $y(\mathbf{X}(F^{(b)}),\boldsymbol{\mu}^{n-1}+c_n\boldsymbol{\delta}_n)$ and $y(\mathbf{X}(F^{(b)}),\boldsymbol{\mu}^{n-1}-c_n\boldsymbol{\delta}_n)$.
    for $j=1,\ldots,L$ **do**
      Estimate the gradient component $\frac{\partial \mathrm{E}\left(\sum_{h=1}^{\tau} C_{T+h}(\mathbf{X}_{T+h},\boldsymbol{\mu}^{n-1})\middle|\mathbf{X}_{[1:T]},F^{(b)}\right)}{\partial \mu_j}$ according to (6).
    end
  end
  Estimate the gradient $\widehat{\nabla}\bar{g}(\boldsymbol{\mu})$ by applying (7).
  Update the decision according to (8), get $\boldsymbol{\mu}^n$, and project it to $\mathbf{\Omega}$ if necessary.
end
Let $\widehat{\boldsymbol{\mu}}^{\star}=\boldsymbol{\mu}^N$ and record it as the final solution.

4 EMPIRICAL STUDY

In this section, we provide the empirical study to evaluate the finite sample performance of the proposed simulation-based prediction framework. In Section 4.1, we first simulate the clinical demands according to the physical process described in Chen et al. (2012), and use the simulated data to compare the prediction performance of IMS-VAR with the commonly used Autogressive (AR) model. Then, in Section 4.2, we study the performance of the simulation based prediction framework by using a bio-pharmaceutical production scheduling example.

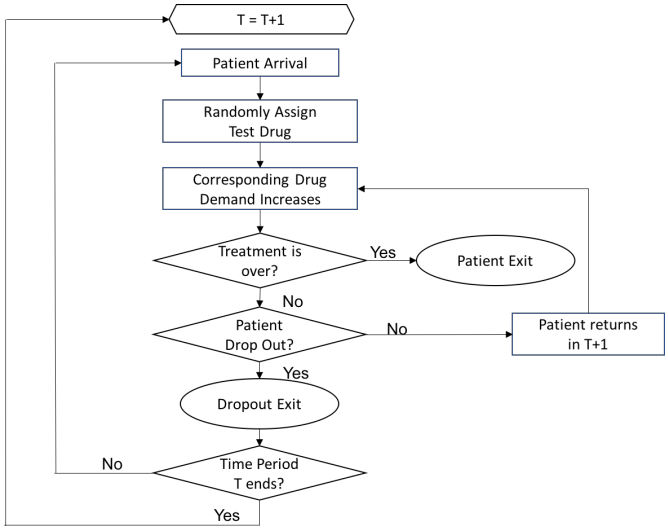
4.1 IMS-VAR for Predicting the Future Clinical Demand

Here, we study the prediction performance of IMS-VAR by using the clinical demands in bio-pharmaceutical supply chains. The demand data are simulated according to the physical process described in Chen et al. (2012). The patients' arrivals follow a non-stationary Poisson process, and the arrival rate varies over time. Suppose that there are three levels of arrival rate, $\lambda=5,10,50$. The level switches following a Markov process with the transition probability

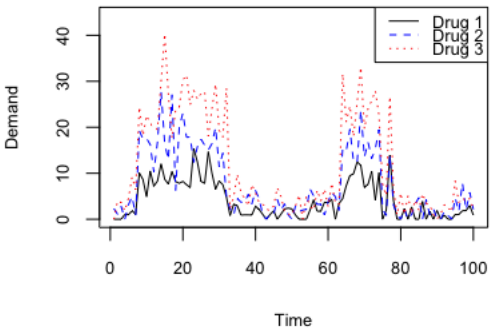
$$P=\begin{bmatrix}0.9 & 0.1 & 0\\ 0.1 & 0.8 & 0.1\\ 0 & 0.1 & 0.9\end{bmatrix}.$$

The logic of this simulation is shown in Figure 1(a). Each arrived patient is randomly assigned to one of $d=3$ clinical test drugs with probability $0.2,0.3,0.5$. The test has a probability $p_1=0.9$ to be successful at each time period and the treatment is over. Otherwise, the patient has a probability $p_2=0.5$

to drop out the treatment. If the patient continues the treatment, the same clinical drug will be used to treat the patient in the next time period. The dosage of drugs consumed in each treatment follows the normal distribution, $\mathcal{N}(1,0.1)$. The clinical demands can be simulated to obtain the historical data $\mathbf{X}_{[1:T]}$. Figure 1.b shows a representative clinical demand data for three drugs in $T = 100$ time periods. There exist the non-stationarity, component-wise and time series dependence in the demands of different drugs.



(a) Clinical Demand Simulation Flowchart



(b) The Representative Demand Sample Paths for Three Clinical Products

Figure 1: Clinical Trial Demands Simulation Flowchart and Demand Data.

Given the historical data $\mathbf{X}_{[1:T]}$, we compare the prediction performance of Bayesian IMS-VAR with the commonly used AR model. For the IMS-VAR, we let the order $p = 1$. We use flat priors with hyper-parameters $\alpha = 1, \eta = 1$, μ_Φ is a $d \times k$ matrix with all elements as 0, V_Φ is a $k \times k$ diagonal matrix with diagonal terms as 100, $\nu = 1$ and Ψ is $d \times d$ identity matrix. To evaluate the prediction performance of Bayesian IMS-VAR, we compare the predictive distribution $p(\mathbf{X}_{[T+1:T+\tau]}|\mathbf{X}_{[1:T]})$ with the prediction distribution $p(\mathbf{X}_{[T+1:T+\tau]}|F^c, \mathbf{X}_{[1:T]})$ when the underlying true input model is known or the sample paths of $\mathbf{X}_{[1:T+\tau]}$ are directly generated from the "physical" simulation system in Figure 1(a). For the commonly used AR approach, we construct a separate AR model for each drug and the model selection is based on the AIC criteria, where the input model estimation uncertainty is ignored. Let $\widehat{F}$ denote the estimated AR model. Then, the predictive distribution obtained from this approach is $p(\mathbf{X}_{[T+1:T+\tau]}|\widehat{F}, \mathbf{X}_{[1:T]})$. According to Harrison et al. (2015), to evaluate the performance of the h step-ahead forecasting, we record the average KS distance over the d components,

$$D_T(h) \equiv \frac{1}{d}\sum_{j=1}^{d} \sup_{x\in\Re}(|F^c_{h,j}(x) - F^p_{h,j}(x)|)$$

for each $h = 1,\ldots,\tau$, where $F^c_{h,j}(x)$ is the c.d.f of the marginal $p(X_{T+h,j}|F^c, \mathbf{X}_{[1:T]})$, and $F^p_{h,j}(x)$ is the c.d.f of $p(X_{T+h,j}|\mathbf{X}_{[1:T]})$ for Bayesian IMS-VAR or $p(X_{T+h,j}|\widehat{F}, \mathbf{X}_{[1:T]})$ for AR, where $X_{T+h,j}$ denotes the j-th component of $\mathbf{X}_{T+h}$ for $j = 1,\ldots,d$.

There is no closed form predictive distribution obtained from IMS-VAR. To assess the performance of predictive distribution, we generate 1000 sample paths of $\mathbf{X}_{[T+1:T+\tau]}$ at each posterior sample of the input model $F^{(b)}$ and the corresponding state $s_T^{(b)}$ for $b = 1,2,\ldots,B$. Here, we let $B = 100$. The samples of $\mathbf{X}_{T+h}$ generated at all B posterior samples of input models are used to calculate the KS distance. The results of $D_T(h)$ for $T = 50, 100, 500$ and $h = 1,2,3$ are recorded in Table 1, which are based on 100 macro-replications. The Bayesian IMS-VAR gives smaller and more robust prediction than AR, especially

when $h = 3$. Both IMS-VAR and AR improve their prediction performance as the amount of historical data T increases.

Table 1: KS Distance of the Predictive Distribution, $D_T(h)$ (inside the brackets are standard deviations).

KS Statistics		$h=1$	$h=2$	$h=3$
$T=50$	IMS-VAR	0.308 (0.032)	0.327 (0.034)	0.362 (0.039)
	AR	0.345 (0.036)	0.383 (0.040)	0.438 (0.045)
$T=100$	IMS-VAR	0.249 (0.027)	0.268 (0.028)	0.302 (0.031)
	AR	0.288 (0.030)	0.325 (0.033)	0.366 (0.038)
$T=500$	IMS-VAR	0.171 (0.018)	0.196 (0.021)	0.226 (0.024)
	AR	0.230 (0.025)	0.254 (0.027)	0.293 (0.031)

4.2 Simulation-based Prediction Framework for Production Scheduling

In this section, we evaluate the performance of the proposed simulation-based prediction framework by using a production scheduling problem. Suppose that the system produces $d = 3$ drugs to meet the clinical demands described in Section 4.1. The decision $\boldsymbol{\mu} \equiv \{\mu_1, \mu_2, \mu_3\}$ with $L = d = 3$, denotes the daily production for the 3 drugs. The inventory for each drug is kept to meet the patients' demands. The overall cost includes the backorder, inventory, and production costs. The backorder cost is denoted by $\boldsymbol{c}_b = 20, 15, 10$ with $c_{b,j}$ representing the backorder cost for each unit of drug j for $j = 1, \ldots, d$. The inventory cost per unit, denoted by $c_i = 1$, is the same for all three drugs. The production cost per unit is $c_p = 1$ for all three drugs. We want to find the optimal production decision $\boldsymbol{\mu}^\star$ to minimize the expected total cost,

$$\begin{aligned} \underset{\boldsymbol{\mu}}{\text{minimize}} \quad & \mathrm{E}\left[\sum_{h=1}^{\tau}\left(c_{b,j}\sum_{j=1}^{d} b_{T+h,j} + c_i \sum_{j=1}^{d} I_{T+h,j} + c_p \sum_{j=1}^{d} \mu_j\right)\right] \\ \text{subject to} \quad & b_{t,j} = (X_{t,j} - \mu_j - I_{t-1,j})^+ \\ & I_{t,j} = (I_{t-1,j} + \mu_j - X_{t,j})^+ \\ & \mu_j \geq 0 \ \forall j = 1, \ldots, d \end{aligned}$$

where $X_{t,j}$ is the random demand for drug j on day t, $I_{t,j}$ is the inventory of drug j on day t, $b_{t,j}$ is the amount of backorder drug.

At the current time period, suppose the inventory for each drug $I_{T,j}$ is 0. Given the historical demands $\mathbf{X}_{[1:T]}$ with $T = 50, 100, 500$, we first generate $B = 100$ posterior samples of IMS-VAR input models according to the sampling procedure in Section 3.1. Then, we apply Algorithm 1 to find the optimal solution. The simulation budget is $R = 10000$ in term of the number of simulation runs. We use different batch size B_0. When $B_0 = 100$, we use all the B posterior samples of input model to estimate the gradient. When $B_0 = 1$, we randomly select one among the B samples. Thus, for $B_0 = 1, 10, 100$, we have $N = \lfloor R/2B_0 \rfloor$ updates respectively. The initial solution is set to be $\boldsymbol{\mu}^0 = \{1, 1, 1\}$. The empirical study indicates that the selection of the initial solution does not have the significant impact on the performance. In the n-th iteration, if $\boldsymbol{\mu}^n$ is outside of the feasible region, we project $\boldsymbol{\mu}^n$ to the feasible region, $\mu_j^n = \max(0, \mu_j^n)$ for $j = 1, \ldots, L$. Denote the optimal solution obtained by IMS-VAR as $\widehat{\boldsymbol{\mu}}_I^\star$. For the AR model, the input model estimate $\widehat{F}$ is used in the stochastic gradient procedure, and we perform $N = R$ updates. In each update, we predict the future demands by using $p(\mathbf{X}_{[T+1:T+\tau]}|\widehat{F}, \mathbf{X}_{[1:T]})$. Denote the optimal solution obtained by the AR based prediction as $\widehat{\boldsymbol{\mu}}_A^\star$. In the empirical study, the step size for the stochastic gradient is chosen as $\eta_n = 0.01$.

To study the performance of $\widehat{\boldsymbol{\mu}}_I^\star$ and $\widehat{\boldsymbol{\mu}}_A^\star$, we compare the difference between the expected costs $E_I = \mathrm{E}\left(\sum_{h=1}^{\tau} C_{T+h}(\mathbf{X}_{T+h}, \widehat{\boldsymbol{\mu}}_I^\star)\right)$ and $E_A = \mathrm{E}\left(\sum_{h=1}^{\tau} C_{T+h}(\mathbf{X}_{T+h}, \widehat{\boldsymbol{\mu}}_A^\star)\right)$ with the expectation over F^c. Both expected costs are estimated by using 10^5 scenarios from F^c. We record the mean and standard error for the difference $\Delta E = E_I - E_A$ in Table 2, which are based on the results from 100 macro-replications. The Bayesian IMS-VAR based simulation prediction framework can lead to the optimal decision with significantly smaller expected cost than the AR when $B_0 = 1, 10$. When $B_0 = 100$, since there are limited updates in Algorithm 1, the performance of IMS-VAR does not show clear advantage.

Table 2: Mean $\pm$ Standard Error of ΔE.

	$T = 50$	$T = 100$	$T = 500$
$B_0 = 1$	-40.9 ± 13.4	-35.6 ± 4.7	-33.8 ± 6.0
$B_0 = 10$	-49.2 ± 9.3	-48.0 ± 7.7	-43.9 ± 6.2
$B_0 = 100$	-14.7 ± 12.5	-10.9 ± 8.6	-6.4 ± 7.5

5 CONCLUSION

In this paper, we propose a simulation-based prediction framework to guide operational decision making. By exploring the underlying generative process, the nonparametric IMS-VAR input model can capture the important properties in the real-world data streams, including non-stationary, skewness, component-wise and time series dependence. The posterior distribution of flexible input model can correctly quantify the input uncertainty. Then, the posterior predictive distribution is used to characterize the prediction uncertainty accounting for both input and stochastic uncertainties. After that, the mini-batch stochastic gradient descent method can efficiently employ the simulation resources to search for the optimal operational decision. The empirical study of the bio-pharmaceutical supply chain management demonstrates that the proposed framework can improve the prediction accuracy of system future response, and lead to the cost-efficient and reliable operational decisions hedging against the prediction uncertainty.

REFERENCES

Banbura, M., D. Giannone, and L. Reichlin. 2010. "Large Bayesian Vector Auto Regressions". *Journal of Applied Econometrics* 25(1):71–92.

Biller, B., and B. L. Nelson. 2005. "Fitting Time-Series Input Processes for Simulation". *Operations Research* 53(3):549–559.

Biller, B., and B. L. Nelson. 2008. "Evaluation of the ARTAFIT Method for Fitting Time-Series Input Processes for Simulation". *INFORMS Journal on Computing* 20(3):485–498.

Bottou, L. 2010. "Large-Scale Machine Learning with Stochastic Gradient Descent". In *Proceedings of COMPSTAT'2010*, edited by Y. Lechevallier and G. Saporta, 177–186. Heidelberg: Physica-Verlag HD.

Calamai, P. H., and J. J. Moré. 1987. "Projected Gradient Methods for Linearly Constrained Problems". *Mathematical Programming* 39(1):93–116.

Cario, M. C., and B. L. Nelson. 1998. "Numerical Methods for Fitting and Simulating Autoregressive-to-Anything Processes". *INFORMS Journal on Computing* 10(1):72–81.

Chau, M., and M. C. Fu. 2015. *An Overview of Stochastic Approximation*, 149–178. New York, NY: Springer New York.

Chee, J., and P. Toulis. 2017. "Convergence Diagnostics for Stochastic Gradient Descent with Constant Step Size". arXiv:1710.06382.

Chen, Y., L. Mockus, S. Orcun, and G. V. Reklaitis. 2012. "Simulation-Optimization Approach to Clinical Trial Supply Chain Management with Demand Scenario Forecast". *Computers & Chemical Engineering* 40:82–96.

Fu, M. C. 2006. "Chapter 19 Gradient Estimation". In *Handbooks in Operations Research and Management Science - Simulation*, edited by S. G. Henderson and B. L. Nelson, Volume 13, 575–616. Elsevier.

Gerhardt, I., and B. L. Nelson. 2009. "Transforming Renewal Processes for Simulation of Nonstationary Arrival Processes". *INFORMS Journal on Computing* 21(4):630–640.

Harrison, D., D. Sutton, P. Carvalho, and M. Hobson. 2015. "Validation of Bayesian Posterior Distributions Using a Multidimensional Kolmogorov–Smirnov Test". *Monthly Notices of the Royal Astronomical Society* 451(3):2610–2624.

Harrod, S., and D. W. Kelton. 2006. "Numerical Methods for Realizing Nonstationary Poisson Processes with Piecewise-Constant Instantaneous-Rate Functions". *Simulation* 82(3):147–157.

Jian, N., and S. G. Henderson. 2015. "An Introduction to Simulation Optimization". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz et al., 1780–1794. Piscataway, New Jersey: IEEE.

Kleinman, N. L., J. C. Spall, and D. Q. Naiman. 1999. "Simulation-Based Optimization with Stochastic Approximation Using Common Random Numbers". *Management Science* 45(11):1570–1578.

Kuhl, M. E., and J. R. Wilson. 2001. "Modeling and Simulating Poisson Processes Having Trends or Nontrigonometric Cyclic Effects". *European Journal of Operational Research* 133(3):566–582.

Ma, Y. 2011. *Risk Management in Biopharmaceutical Supply Chains*. Ph. D. thesis, University of California, Berkeley.

Nelson, B. L., and I. Gerhardt. 2011. "Modelling and Simulating Non-stationary Arrival Processes to Facilitate Analysis.". *J. Simulation* 5(1):3–8.

Ruder, S. 2016. "An Overview of Gradient Descent Optimization Algorithms". *arXiv* :1609.04747.

Spall, J. C. 1998. "Implementation of the Simultaneous Perturbation Algorithm for Stochastic Optimization". *IEEE Transactions on Aerospace and Electronic Systems* 34(3):817–823.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei. 2006. "Hierarchical Dirichlet Processes". *J. Amer. Stat. Assoc.* 101(476):1566–1581.

Xie, W., P. Zhang, R. Chen, and Z. Zhou. 2018. "A Nonparametric Bayesian Framework for Short-Term Wind Power Probabilistic Forecast". *IEEE Transactions on Power Systems*. Accepted.

AUTHOR BIOGRAPHIES

WEI XIE is an assistant professor in the Department of Mechanical and Industrial Engineering at Northeastern University. She received her M.S. and Ph.D. in Industrial Engineering and Management Sciences at Northwestern University. Her research interests are in computer simulation, data analytics, and stochastic optimization for cyber-physical system risk management. Her email address is xieweirosie6@gmail.com.

PU ZHANG is a Ph.D. candidate of the Department of Industrial and Systems Engineering at Rensselaer Polytechnic Institute. His research interests are input modeling and uncertainty quantification in stochastic simulation. His email address is zhangp5@rpi.edu.

ILYA O. RYZHOV is an Associate Professor of Operations Management and Management Science in the Department of Decision, Operations & Information Technologies at the Robert H. Smith School of Business, University of Maryland. His research primarily focuses on simulation optimization and statistical learning, with applications in business analytics, revenue management, and nonprofit/humanitarian operations. He is a coauthor of the book Optimal Learning (Wiley, 2012). His work was recognized in WSC's Best Theoretical Paper Award competition on three separate occasions (winner in 2012, finalist in 2009 and 2016), and he received I-SIM's Outstanding Publication Award in 2017. His email address is iryzhov@rhsmith.umd.edu.