ESTIMATING STOCHASTIC POISSON INTENSITIES USING DEEP LATENT MODELS

Ruixin Wang

Prateek Jaiswal

Harsha Honnappa

School of Industrial Engineering
Purdue University
West Lafayette, IN 47906, USA

ABSTRACT

We present a new method for estimating the stochastic intensity of a doubly stochastic Poisson process. Statistical and theoretical analyses of traffic traces show that these processes are appropriate models of high intensity traffic arriving at an array of service systems. The statistical estimation of the underlying latent stochastic intensity process driving the traffic model involves a rather complicated nonlinear filtering problem. We develop a novel simulation method, using deep neural networks to approximate the path measures induced by the stochastic intensity process, for solving this nonlinear filtering problem. Our simulation studies demonstrate that the method is quite accurate on both in-sample estimation and on an out-of-sample performance prediction task for an infinite server queue.

1 INTRODUCTION

This paper introduces a simulation-based method for estimating the stochastic intensity process of a doubly stochastic Poisson process (DSPP), using sample path observations of the DSPP over a fixed time horizon and under the assumption of a stochastic differential equation (SDE) model of the intensity. DSPPs are widely acknowledged as an appropriate model of traffic arriving at a variety of service systems, including hospitals and call centers. Specifically, multiple statistical analyses (Jongbloed and Koole 2001; Avramidis, Deslauriers, and L'Ecuyer 2004; Avramidis and L'Ecuyer 2005; Maman 2007; Kim and Whitt 2014) show that the (estimated) index of dispersion (i.e., the ratio of the variance to the mean) of the arrival counts typically exceeds 1 at reasonable operational time-scales; for Poisson processes the index equals 1. Furthermore, the arrival intensity appears time-varying and there are temporal correlations between traffic counts across non-overlapping time intervals. These conditions strongly indicate that the traffic process is not a Poisson process with deterministic intensity. However, at smaller time-scales (on the order of inter-arrival times) it is not possible to reject the null hypothesis that the arrival counts over a fixed time interval are Poisson distributed (Kim and Whitt 2014). DSPPs can model the overdispersion, temporal correlations and time-varying nature of the intensity while remaining reasonably tractable to use for performance prediction and control/optimization tasks. A rigorous definition of DSPPs is provided in the next section.

The expansive definition of DSPPs allows for many models of the stochastic intensity process. A simple model advocated for modeling call center traffic in Whitt (1999) assumes that the uncertainty in the arrival rates is determined by a single random variable that determines the daily 'busyness' level. However, as noted in Zhang et al. (2014), the static nature of the intensity model implies it cannot account for the temporal correlation structure observed in many traffic traces. Zhang et al. (2014), in turn, suggest the use of a 'dynamic' (sic) intensity model. In the context of high intensity call center traffic they show, through a combination of theoretical and empirical analysis, that a Cox-Ingersoll-Ross (CIR) diffusion is appropriate.

Recall that the CIR process is defined as the solution to the SDE

$$dZ(t) = (\beta - Z(t))dt + \eta\beta^\alpha \sqrt{Z(t)}dW(t), \ \forall t \geq 0 \tag{1}$$

where $(W(t) : t \geq 0)$ is a standard Brownian motion process, (α, β, η) are positive constants that constitute the parameters of the model. Specifically, Zhang et al. (2014) present empirical evidence that the empirical distribution of the standardized arrival counts roughly follows a standard normal distribution, in time intervals where the mean arrival counts are 'large' . This empirical observation is supported by a rigorous central limit theorem (CLT) that holds for all $\alpha \in \left(0, \frac{1}{2}\right)$. Following Zhang et al. (2014), we assume that the stochastic intensity process is well-modeled by an SDE (though not necessarily (1)).

In practice, the stochastic intensity is *latent* (i.e., unobserved) and must be estimated from traffic traces. As noted in Cheng (2017), this estimation problem is challenging. In fact, it entails the solution of a nonlinear filtering problem where the underlying stochastic intensity process can be viewed as the 'signal,' and the arrival process is a noisy 'observation' of the intensity. The solution of the nonlinear filtering problem depends crucially on the computation of the *pathwise Kallianpur-Striebel formula* (see Van Handel (2007), Ch. 1)), which is remarkably complicated. More crucially, the computation of the filter assumes complete knowledge of the latent intensity model. In our setting, while a *structure* of the model might be assumed, model parameters are unknown and must be estimated themselves.

We present a computational method that simultaneously estimates the intensity model and solves the nonlinear filtering problem. We model the unknown drift and diffusion functions of this SDE using deep neural networks (DNNs), which are trained by maximizing a tight lower bound on the marginal log-likelihood of the traffic process. This is an instance of a so-called *deep latent model* (DLM); examples of such models includes variational autoencoders (VAEs) and generative adversarial networks (GANs) used to synthesize video and image samples (so-called 'deep fakes') in the artificial intelligence (AI) literature (Goodfellow et al. 2016, Ch. 20). To the best of our knowledge, this method has not been developed in the context of continuously observed stochastic processes where DNN training can be rather complicated. Recent work in (Tzen and Raginsky 2019a; Tzen and Raginsky 2019b) considers a more restrictive class of problems where the latent signal process is a diffusion over the interval $[0, 1]$ and the objective is to estimate the terminal marginal distribution using observations of a random variable dependent on the terminal marginal (latent) random variable.

In the subsequent sections we first present an overview of DSPPs in Section 2, followed by an extensive description of the statistical estimation problem and variational autoencoders in Section 3 and 4. We present our method in Section 5, where we derive the lower bound referenced above and the DNN training procedure we have developed, based on the theory of stochastic flows by Kunita (1984). Finally, in Section 6 we present simulation results that demonstrate the efficacy of our method. Specifically, we present results on a) in-sample estimation of the stochastic intensity process itself, and on b) out-of-sample 'run-through' experiments for predicting performance metrics in an infinite server queue. Section 7 concludes with a summary and some commentary on future work.

2 DOUBLY STOCHASTIC POISSON PROCESSES

Let $(X(t) : t \geq 0)$ be a non-decreasing $\mathbb{Z}_+$-valued point process, $(X|Z)$ represent the process conditioned on the stochastic process $(Z(t) : t \geq 0)$, and $\text{Poi}(\Lambda)$ represent a Poisson process with integrated intensity function $(\Lambda(t) : t \geq 0)$. Formally, a DSPP is defined as:

Definition 1 Let $(Z(t) : t \geq 0)$ be a non-negative stochastic process such that with probability one $t \mapsto Z(t)$ is locally integrable. Then, $(X(t) : t \geq 0)$ is a DSPP driven by $(Z(t) : t \geq 0)$ if $(X|Z) \sim \text{Poi}(\mathbf{Z})$, where $\mathbf{Z}$ is the integrated process defined as $\mathbf{Z}(s, t) := \int_s^t Z(r)dr$ for any $s < t$.

That is, for any set of points $\{t_0, t_1, \ldots, t_d\} \subset (0, \infty)$, where $0 < t_0 \leq t_1 \leq \cdots \leq t_d < \infty$, the finite dimensional distributions of $(X|Z)$ satisfy

$$\mathbb{P}(X(t_0) = k_0, X(t_1) = k_1, \ldots, X(t_d) = k_d | Z_{0:t_d}\}) \tag{2}$$

$$= \frac{\exp(-\mathbf{Z}(0, t_0))(\mathbf{Z}(0, t_0)^{k_0}}{k_0!} \prod_{i=0}^{d-1} \frac{\exp(-\mathbf{Z}(t_i, t_{i+1}))(\mathbf{Z}(t_i, t_{i+1}))^{k_{i+1}-k_i}}{(k_{i+1} - k_i)!},$$

where $Z_{0:t} \equiv (Z(s) : 0 \leq s \leq t)$. Formally, the path measure induced by $(X(t) : t \geq 0)$ is defined as $\int Poi(\mathbf{Z})dP(Z)$, where $P(\cdot)$ is the path measure induced by the stochastic process $(Z(t) : t \geq 0))$, so that the finite dimensional distribution of $X(t)$ (at any fixed $t \geq 0$) satisfies

$$\mathbb{P}(X(t_0) = k_0, \ldots, X(t_d) = k_d) = \int \mathbb{P}(X(t_0) = k_0, X(t_1) = k_1, \ldots, X(t_d) = k_d | Z_{0:t_d}\})dP(Z_{0:t_d}). \tag{3}$$

Note that we are deliberately being less than rigorous in our description of this path measure so as to avoid a heavier notational burden that distracts from the primary message of this paper.

3   THE STATISTICAL ESTIMATION PROBLEM

In the setting of a stochastic differential equation (SDE) model of the intensity, the estimation problem amounts to estimating the drift and diffusion coefficients. Suppose the drift and diffusion coefficients are parameterized by θ. To understand the complexity of the problem, consider the following formal argument for deriving the maximum (log-)likelihood estimator (MLE) of the marginal distribution of $X(t)$,

$$\log \mathbb{P}(X(t) = k) = \log \int \mathbb{P}(X(t) = k | Z_{0:t})dP_\theta(Z_{0:t}), \tag{4}$$

where P_θ is the path measure corresponding to the parameters θ. Observe that computing the MLE requires differentiating with respect to θ under this path measure. There are potentially two ways of doing this. First, suppose we are able to compute the distribution of $\mathbf{Z}(0, t)$ as a function of the parameters θ. Then, the gradient of the log-likelihood can be computed using the score function. However, while the distribution of $\mathbf{Z}$ could be computed with some effort for some instances (such as the CIR model (1)), this is unlikely to be true for arbitrary stochastic processes. On the other hand, suppose the path measure P_θ has a Radon-Nikodym density with respect to a reference path measure π_0; that is, there exists a real-valued potential function $\Phi(Z_{0:T}; \theta)$ such that $dP_\theta/d\pi_0(Z_{0:T}) \propto \exp(\Phi(Z_{0:T}; \theta))$, then the gradient can be computed by differentiating the potential function. In general, however, we are confronted by the question of the choice of an appropriate reference measure. Note that measures in infinite dimensional spaces have a strong tendency towards either singularity or equivalence, complicating this choice; for instance, standard Brownian motion is not a feasible reference measure for the CIR process. While this issue can be resolved in specific cases, we would like a method that works for arbitrary choices of the stochastic intensity process.

While the reference measure might not be known, we can introduce another measure into (4) that also has a density with respect to the reference measure, making it equivalent to P_θ. Observe that the conditional measure $P(Z_{0:t}|X(t) = k)$ is the "optimal" choice in the sense that we have

$$\log \mathbb{P}(X(t) = k) = \log \int \mathbb{P}(X(t) = k | Z_{0:t}) \frac{dP_\theta(Z_{0:t})}{dP(Z_{0:t}|X(t) = k)} dP(Z_{0:t}|X(t) = k) \tag{5}$$

$$= \int \log \left(\mathbb{P}(X(t) = k | Z_{0:t}) \frac{dP_\theta(Z_{0:t})}{dP(Z_{0:t}|X(t) = k)} \right) dP(Z_{0:t}|X(t) = k), \tag{6}$$

where the second equality follows from the fact that the term inside the $\log$ is precisely $\mathbb{P}(X(t) = k)$ (and therefore a constant with respect to the conditional measure). This formal calculation shows that computing the MLE of the count process amounts to solving a complex nonlinear filtering problem to compute the conditional measure, where the unobserved stochastic intensity function should be viewed as a 'signal' and the (conditionally) Poisson counts are noisy 'observations' of the signal. More precisely, the Doob-Meyer decomposition of the DSPP $(X(t) : t \geq 0)$ implies $X(t) = \int_0^t Z(s)ds + \eta(t)$, where $(Z(t) : t \geq 0)$ is the stochastic intensity process and $(\eta(t) : t \geq 0)$ is a martingale (see (Segall and Kailath 1975) as well). However, solving this filtering problem is remarkably hard. Observe that the density $dP(\cdot|X(t) = k)/dP_\theta(\cdot)$ is the pathwise Kallianpur-Striebel formula (Van Handel 2007, Ch. 1). Solving this nonlinear filtering problem, however, is no easier than the 'direct differentiation' methods for computing the MLE noted in the previous paragraph.

Revisiting the computation in (5), suppose we now introduce an arbitrary (but equivalent) measure $P_{\phi,k}$ (parameterized by ϕ and k). Then, Jensen's inequality implies that

$$\log \mathbb{P}(X(t) = k) \geq \int \log \left(\mathbb{P}(X(t) = k|Z_{0:t}) \frac{dP_\theta(Z_{0:t})}{dP_{\phi,k}(Z_{0:t})} \right) dP_{\phi,k}(Z_{0:t}). \tag{7}$$

While this is a lower bound, observe that the inequality can be tightened by maximizing it over both θ and ϕ. The objective, however, is highly non-concave in these parameters and consequently we can only guarantee the computation of a local optimum. Furthermore, the choice of parameterization will, in general, imply that the class of measures being optimized over may not include the 'true' measures, resulting in an approximation to the filtering distribution. Therefore, this procedure of optimizing over path measures is an example of *approximate inference*, used extensively in the machine learning literature for approximately solving high dimensional and large sample statistical inference problems, particularly with Bayesian models. In the next section, we briefly review approximate inference in a general setting.

4 APPROXIMATE INFERENCE

Consider an ensemble of n observations $\mathbf{Y}_n := \{Y_1, Y_2, \ldots, Y_n\}$, where each $Y_i \in \mathcal{A}$, an arbitrary topological space, and represents the available dataset. Each Y_i induces a distribution P_i that lies in some space of measures $\mathcal{P}$. The inference problem is to estimate a distribution over the sequence of unknown data generating distributions $\{P_1, P_2, \ldots, P_n\}$ given the observations $\mathbf{Y}_n$. In the Bayesian inference setting, we assume the existence of a sequence of 'prior' distributions $\{\Pi_1, \Pi_2, \ldots, \Pi_n\} \in \mathcal{P} \times \cdots \times \mathcal{P} =: \bigotimes_n \mathcal{P}$. Subsequently, $Y_i \in \mathbf{Y}_n$ is assumed to follow a *generative model* defined in the following hierarchical manner: (i) Generate $P_i \sim \Pi_i$ $i \in \{1, 2, \ldots n\}$; then (ii) sample $Y_i \sim P_i$. Now, using Bayes rule, observe that for any subset $B \subseteq \bigotimes_n \mathcal{P}$ the 'posterior' distribution satisfies

$$\Pi_n(B|\mathbf{Y}_n) = \frac{\int_B \prod_{i=1}^n d\Pi_i(P_i)P_i(Y_i)}{\int \prod_{i=1}^n d\Pi_i(P_i)P_i(Y_i)}. \tag{8}$$

Observe that we have assumed $\mathbf{Y}_n$ forms an independent ensemble; we will continue with this assumption in the remainder of the paper. In most of the high (or possibly infinite) dimensional settings computing this posterior distribution is intractable and consequently the problem of computing and performing statistical inference with the posterior is challenging. To address the intractability of the posterior, various sampling and optimization based methods have been proposed. We now describe a variational approach to do approximate inference that belongs to the latter category. In this framework, we first fix a class of measures $\mathcal{Q}_n \subseteq \bigotimes_n \mathcal{P}$ and then compute an approximation to the posterior (8) in the family $\mathcal{Q}_n$ by optimizing a lower bound to the 'model evidence' $\mathbb{P}(\mathbf{Y}_n) := \int \prod_{i=1}^n d\Pi_i(P_i)P_i(Y_i)$. Observe that for any sequence of

measures $\{Q_i\}_{1 \leq i \leq n} \in \mathcal{Q}_n$, Jensen's inequality implies that

$$
\begin{aligned}
\log \mathbb{P}(\mathbf{Y}_n) &\geq \int \prod_{i=1}^n dQ_i(P_i) \log \prod_{i=1}^n P_i(Y_i) - \int \prod_{i=1}^n dQ_i(P_i) \log \frac{\prod_{i=1}^n dQ_i(P_i)}{\prod_{i=1}^n d\Pi_i(P_i)} \\
&= \sum_{i=1}^n \left[\mathbb{E}_{Q_i} \left[\log P_i(Y_i) \right] - \frac{dQ_i}{d\Pi_i}(P_i) \right] ;
\end{aligned}
\tag{9}
$$

in the machine learning literature the right hand side (RHS) in (9) is popularly known as *evidence lower bound* (ELBO). Observe that (7) precisely corresponds to a single random variable in the sum on the RHS of (9), where the measure Q_i corresponds to $P_{\Theta,k}$ and Π_i corresponds to the measure P_Γ. In the variational framework, for a given sequence of prior distributions $\{\Pi_1, \Pi_2, \ldots, \Pi_n\}$ the ELBO is maximized over the distribution in $\mathcal{Q}_n$ using stochastic gradient descent methods to find the best $\{Q_i\}_{1 \leq i \leq n}$ in $\mathcal{Q}_n$. In particular, observe that the ELBO can be rewritten as

$$
\mathrm{ELBO} = -\mathrm{KL} \left(\prod_{i=1}^n dQ_i(P_i) \middle\| d\Pi_n(P_1, P_2 \ldots P_n | \mathbf{Y}_n) \right) + \log \mathbb{P}(\mathbf{Y}_n),
$$

where KL represents the Kullback-Leibler divergence. Therefore, the optimizer of ELBO is an approximation to the posterior distribution as defined in (8).

Deep latent models (DLMs) (Goodfellow et al. 2016, Ch. 19, 20) specialize this general presentation to the setting where the probability measures are parameterized by deep neural networks (DNNs). Variational autoencoders (VAEs) (Kingma and Welling 2019) are an example of DLMs in the multivariate setting where the sequence of prior distributions are known only up to the parameters of an appropriately chosen DNN modeling these parameters. In the VAE literature this sequence of prior distributions are also known as *decoders*. The approximating measures $\mathcal{Q}_n$, entitled *encoders* in the VAE literature, are also parameterized using DNNs. Given the ensemble $\mathbf{Y}_n$, the DNN parameters of both the encoder and decoder are estimated using stochastic gradient descent (SGD). Our current setting, of course, is far more complicated than the VAE setting since the DNNs model the drift and diffusion coefficients of SDEs leading to a complicated training procedure, as we will see.

5   DLMs FOR DSPPs

We assume access to n independent and identically distributed (i.i.d.) observations of a stochastic process $\{X(t), t \leq T\}$. In many service systems, such as hospitals and call centers, traffic counts are collected at fixed, regular intervals; for instance, in many large call centers, this is typically at intervals of length 30 seconds to 1 minute. As noted before, it has been observed (Zhang et al. 2014) that a DSPP with CIR-type ergodic diffusion process driving the intensity is an appropriate model of the traffic counts at operational time-scales (typically of the order of 10 minutes). The time interval $[0, T]$ in our model represents this operational time-scale.

For clarity of exposition, we will describe our method under two specific conditions: (i) the traffic counts are observed at the time epochs $T/2$ and T; and (ii) a single sample $n = 1$. These can be extended to more observation instants and samples at the expense of a more burdensome notation, but our method will not change. We model the unknown stochastic intensity process by the SDE

$$
dZ(t) = b(Z(t), t; \theta)dt + \eta \sqrt{Z(t)} dW(t), \quad t \leq T
\tag{10}
$$

where $\{W(t), t \geq 0\}$ is the standard Brownian motion, $b(\cdot, t; \theta) : C_b[0, T] \times [0, T] \mapsto \mathbb{R}$ is the drift and $\eta \sqrt{(\cdot)}$ with $\eta > 0$ is the diffusion coefficient. $C_b[0, T]$ denotes the space of all continuous and bounded function on the interval $[0, T]$. Here, the unknown drift function is modeled using a DNN parameterized
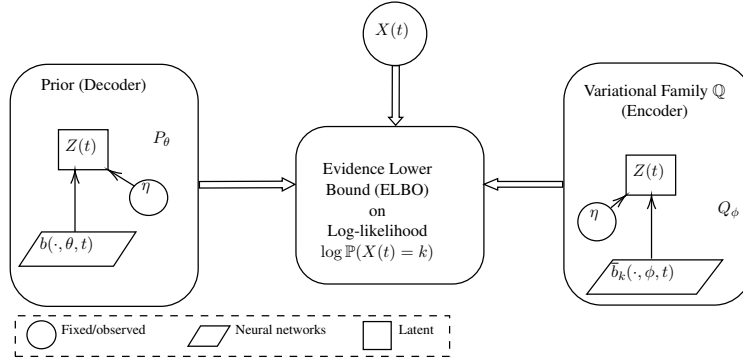
Figure 1: An illustration of the deep latent modeling framework.

by θ, and to avoid getting bogged down in technical detail, we assume the existence of a strong solution to (10). For technical reasons we will, for now, assume that the diffusion coefficient is known. We denote the measure induced by the solution of this SDE as $P_\theta(\cdot)$; this corresponds to the 'prior' measure in the previous section. The independent increments property implies that the joint distribution of the arrival count random vector $Y_1 := (X\left(\frac{T}{2}\right), X(T))$, conditional on the intensity process $Z_{0:T}$, can be expressed as:

$$\mathbb{P}(Y_1 = (k_1, k_2)|Z_{0:T}) = \frac{e^{-\int_0^{\frac{T}{2}} Z(t)dt} \left(\int_0^{\frac{T}{2}} Z(t)dt\right)^{k_1}}{k_1!} \frac{e^{-\int_{\frac{T}{2}}^T Z(t)dt} \left(\int_{\frac{T}{2}}^T Z(t)dt\right)^{k_2-k_1}}{(k_2 - k_1)!}. \quad (11)$$

5.1 A DLM for the Stochastic Intensity Process

By definition, the variational family $\mathcal{Q}$ must consist of measures that are absolutely continuous with respect to the 'prior' measure P_θ. In our current setting, $\mathcal{Q}$ is the class of equivalent measures induced by the solutions of SDEs that have the same diffusion coefficient as (10). To be precise, consider the SDE

$$dZ(t) = \bar{b}_k(Z(t), t : \phi)dt + \eta\sqrt{Z(t)}dW(t), \text{ for } t \leq T, \quad (12)$$

where for each $k \in \{0, 1, \ldots\}$ the drift function $\bar{b}_k(\cdot, \cdot; \phi)$ is modeled using a DNN with parameter ϕ. We denote the measure induced by the solution of this SDE as Q_ϕ. Figure 1 illustrates the use of deep latent models in defining measures P_θ and Q_ϕ and consequently ELBO. Next, we derive the ELBO for the observation random vector Y_1. The proof, omitted for space reasons, follows from Girsanov's theorem.

Theorem 1 Define $u_k(Z(t), t; \theta, \phi) := (\eta\sqrt{Z(t)})^{-1}\bar{b}_k(Z(t), t; \phi) - b(Z(t), t; \theta)$ and suppose that u_k satisfies a *strong* Novikov's condition, $\mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^T |u_k(Z(t), t; \theta, \phi)|^2 dt\right)\right] < +\infty \ \forall\theta, \phi$. Then,

$$\hat{W}_t := \int_0^t u_k(Z(s), s; \theta, \phi)ds + W(t) \quad (13)$$

is a Brownian motion w.r.t. Q_ϕ, $dZ(t) = b(Z(t), t; \theta)dt + \eta\sqrt{Z}(t)d\hat{W}_t$, and

$$\log \mathbb{P}(Y_1 = (k_1, k_2)) \geq \mathbb{E}_{Q_\phi}\left[\log \mathbb{P}(Y_1 = (k_1, k_2)|Z_{0:T}) - \frac{1}{2}\int_0^T u_k^2(Z(s), s; \theta, \phi)ds\right] := \text{ELBO}. \quad (14)$$

Notice that we must assume that Novikov's condition holds for all possible parameterizations of the functions $\bar{b}$ and b. This is a strong condition that is satisfied for the class of DNNs that we work with in this paper, since the output of the DNN is bounded by definition. However, more analysis is required on sufficient conditions for DNNs to satisfy Novikov's condition.

601

5.2 TRAINING THE DLM

Our objective is to train the neural networks $b(Z(t), t; \theta)$ and $u_k(Z(t), t; \theta, \phi)$ by maximizing the ELBO. We fix $u_k(Z(t), t; \theta, \phi)$ to be a deterministic neural network defined as $\tilde{u}(k, t; \beta)$ with parameters β. Combined with (13), this additional restriction imposed on $u_k(Z(t), t; \theta, \phi)$ ensures that the process $\hat{W}_t$ has independent increments. In the variational inference literature (Blei et al. 2017), this assumption is also known as the mean-field approximation, that is each partition of the unknown latent variable is independent of the other. A similar assumption on the latent process was used in (Tzen and Raginsky 2019a), where the authors call it a path-space analog of the mean-field approximation. Now substituting $u_k(Z(t), t; \theta, \phi) = \tilde{u}(k, t; \beta)$ in (13) and from the observation $dZ(t) = b(Z(t), t; \theta)dt + \eta\sqrt{Z}(t)d\hat{W}_t$, it follows that we can simulate the SDE using $W(t)$ instead of $\hat{W}(t)$; that is,

$$dZ(t) = b(Z(t), t; \theta)dt + \sqrt{Z(t)}\tilde{u}(k, t; \beta)dt + \sqrt{Z(t)}dW(t) \text{ and } Z(0) = 0. \tag{15}$$

We denote the measure induced by the above SDE as $Q_{\beta,\theta}$. For simplicity we fix $\eta = 1$.

We use stochastic gradient descent (SGD) to maximize the objective in (14) to learn the unknown neural network parameters θ and β. In order to use SGD, we first need to generate sample paths of the latent process $Z(t)$ (15), which we do using the Euler-Maruyama discretization method. We partition the time interval $[0, T]$ in N equal sub-intervals, denoted as $\{t_0, t_1, ... t_N\}$, with $t_0 = 0$ and $t_N = T$, set $Z(t_0) = Z(0)$, and simulate $\{Z(t_m)\}_{0 \le m \le N}$ using the recursive equation

$$Z(t_{m+1}) = Z(t_m) + b(Z(t_m), t_m, \theta)(t_{m+1} - t_m) + \sqrt{Z(t_m)}\tilde{u}(k, t_m; \beta)(t_{m+1} - t_m) + \sqrt{Z(t_m)}\Delta W_m$$

where $\{\Delta W_m := W(t_{m+1}) - W(t_m)\}_{0 \le m < N}$ are N i.i.d. standard Gaussians.

In order to use SGD we also need to compute the gradient of the objective function (9) with respect to the parameters θ and β. Notice that the expectation in ELBO is with respect to the measure induced by SDE in (15) denoted as $Q_{\beta,\theta}$. Observe that the only source of randomness in generating $Z(t)$ is from the Brownian motion $W(t)$, which does not depend on either β or θ. Therefore we interchange the differential operator with respect to the parameters and the expectation in (14). To make the dependence of $Z(t)$ on β and θ explicit, we write $Z(t)$ as $Z^{\beta,\theta}(t)$. In particular for given values of parameters θ and β_{-j} (all components of parameter β except β^j) observe that $\frac{\partial}{\partial \beta^j}\mathbb{E}\left[\log \mathbb{P}(Y_1 = (k_1, k_2)|Z_{0:T}^{\beta,\theta}) - \frac{1}{2}\int_0^T \tilde{u}^2(k, s; \beta)ds\right] =$

$$\mathbb{E}\left[\frac{\partial}{\partial \beta^j}\log\left(\frac{e^{-\int_0^{\frac{T}{2}} Z^{\beta,\theta}(t)dt}\left(\int_0^{\frac{T}{2}} Z^{\beta,\theta}(t)dt\right)^{k_1}}{k_1!}\frac{e^{-\int_{\frac{T}{2}}^T Z^{\beta,\theta}(t)dt}\left(\int_{\frac{T}{2}}^T Z^{\beta,\theta}(t)dt\right)^{k_2-k_1}}{(k_2 - k_1)!}\right)\right.$$
$$\left. - \frac{\partial}{\partial \beta^j}\int_0^T \tilde{u}^2(k, s; \beta)ds\right], \tag{16}$$

where we use the likelihood expression from (11). Also note that, to avoid any confusion, we have omitted subscript $Q_{\beta,\theta}$ from $\mathbb{E}[\cdot]$ above. Now applying straightforward product differentiation rule and subsequently interchanging the integral and $\frac{\partial}{\partial \beta^j}$, would result into an expression requiring us to compute the derivative of the process $Z^{\beta,\theta}(t)$ with respect to β^j. To compute the derivative process, it follows from Kunita (1984), Theorem 3.1) that under certain regularity condition on the drift and diffusion coefficient of the process $Z^{\beta,\theta}(t)$ (15) the derivative process $\frac{\partial}{\partial \beta^j}Z^{\beta,\theta}(t)$ is the solution of the following SDE

$$\frac{\partial Z^{\beta,\theta}(t)}{\partial \beta^j} = \int_0^t \left(\frac{\partial b(Z^{\beta,\theta}(s), s; \theta)}{\partial Z^{\beta,\theta}(s)}\frac{\partial Z^{\beta,\theta}(s)}{\partial \beta^j} + \frac{\tilde{u}(k, s; \beta)}{2\sqrt{Z^{\beta,\theta}(s)}}\frac{\partial Z^{\beta,\theta}(s)}{\partial \theta^j} + \sqrt{Z^{\beta,\theta}(s)}\frac{\partial u_s}{\partial \beta^j}\right)ds$$

$$+ \int_0^t \left(\frac{1}{2\sqrt{Z^{\beta,\theta}(s)}}\frac{\partial Z^{\beta,\theta}(s)}{\partial \beta^j}\right)dW_s \text{ and } \frac{\partial Z^{\beta,\theta}(0)}{\partial \beta^j} = 0.$$

We simulate the derivative process above using the Euler-Maruyama method in a similar manner as we did for $Z^{\beta,\theta}(t)$. Lastly, we use a similar procedure to generate the derivative of the latent process $Z^{\beta,\theta}(t)$ with respect to a component of θ for given values of the other parameters. We omit this for space reasons.

6 NUMERICAL EXPERIMENTS

We conducted a number of simple experiments to demonstrate both the in- and out-of-sample performance of the DLM. We start by describing the setting for the experiments. The code is written in Matlab with the Deep Learning Toolbox. The computation and space complexity of this method can be found in Tzen and Raginsky (2019a). In our specific case, the time complexity for each iteration of the gradient update is $\mathcal{O}\left(N((k+n)T(b) + T(\tilde{u}))\right)$, where k and n are number of parameters in β and θ, respectively, $T(f)$ is the time complexity for computing f, and N is the number of time steps in the time discretization.

6.1 Setting

Observe that training the neural network by maximizing the ELBO entails solving a stochastic optimization problem in (9). We use a sample average approximation (SAA) of (9) for which we simulate m independent sample paths of $(Z(t) : t \in [0, T])$:

$$\frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} \left[\log \left(P(X^i(T/2) = k_1^i, X^i(T) = k_2^i | Z_{0:T}^j) \right) - \frac{1}{2} \int_0^T \tilde{u}^2(k, s; \beta) ds \right]. \quad (17)$$

We integrate the SDEs using Euler-Maruyama discretization noted in the previous section. The architecture of the neural networks is

- $b(Z(t), t; \theta) : R^2 \to R$ is a feedforward neural network with 20 fully connected layers of size 10. The activation function is chosen as $tanh$. The inputs are time epoch and the current intensity.
- $\tilde{u}(k, t; \beta) : R^2 \to R$ is also a feedforward neural network with 20 fully connected layers of size 10. The activation function is chosen as $tanh$. The inputs are time epoch and the state at time T.

Notice that unlike Li et al. (2020), we do not require any specific architecture on the Neural network. One can always tune the hyperparameters to find as good or even better architecture.

We assume that the true latent intensity process is a standard CIR process:

$$dZ(t) = 0.3(80 - Z(t))dt + \eta\sqrt{Z(t)}dW(t), \quad (18)$$

where $Z(0) = 5$ and $\eta \in [0, 1]$ is the 'noise magnitude' of the model. Observe that if $\eta = 0$, the intensity process is the solution of an ordinary differential equation, and the arrival process is a NHPP. We set the simulation horizon to be $T = 4$, and uniformly partition the interval $[0, T]$ into the grid $\mathcal{P} = \{t_1, t_2, .., t_M\}$ with $t_{k+1} - t_k = 1/15$, $t_1 = 0$ and $t_M = 4$. The training data consists of $n = 200$ sample paths of the DSPP generated using the theoretical model (18). This data is further divided into 'mini-batches' of size 10 and then fed into the Adam solver (Kingma and Ba 2014). We run the code for 35 epochs (350 gradient updates in total). The learning rate for $b(Z(t), t; \theta)$ and $\tilde{u}(k, t; \beta) : R^2 \to R$ are both set to be 0.01.

We compare our method against the piece-wise linear maximum likelihood estimate (MLE) of the intensity assuming the traffic model is an NHPP, developed in Zheng and Glynn (2017). This estimator is quite robust when the objective is to predict a mean performance metric. While it can be very inaccurate in predicting higher moments, betraying the fact that the MLE is computed assuming no correlation structure in the count process, we believe the relative simplicity and the fact that the computation of mean performance metrics are frequently the focus of performance analysis make it a useful reference. As noted before, our experiment will use arrival counts at time $T/2$ and T, and therefore it suffices to consider a two piece

linear estimator in this experiment by maximizing the likelihood function,

$$\mathcal{L}_n(\hat{Z}(t)) = \frac{1}{n}\sum_{i=1}^{n}\left(X^i(T/2)\int_0^{\frac{T}{2}}\hat{Z}(t)dt + (X^i(T) - X^i(T/2))\int_{\frac{T}{2}}^{T}\hat{Z}(t)dt\right) - \int_0^{T}\hat{Z}(t)dt,$$

which follows from display (2) in Zheng and Glynn (2017).

6.2 Estimating the Intensity Process

Our first experiment focuses on the estimation of the 'true' latent intensity model (18) when $\eta = 1$. Figure 2(a) shows the results of an in-sample estimation of the average intensity (computed using 200 training samples of the 'true' model). Observe that both our method ('predicted') and the piece-wise linear model estimate the mean intensity process quite accurately. Figure 2(b) shows that the estimated mean integrated intensity, too, is almost identical to the 'true' model in either model. This is unsurprising: recall that the Poisson count distribution in the ELBO (14) is a function of the integrated intensity, and this plays a crucial role in constraining the estimation problem.
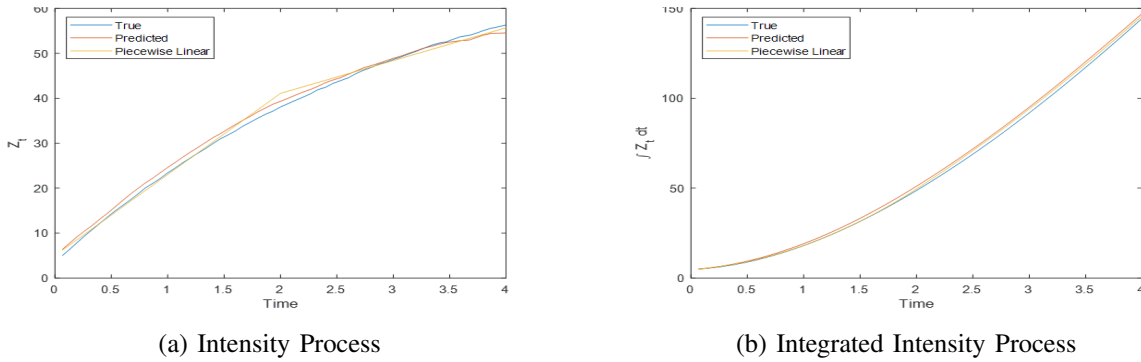

(a) Intensity Process

(b) Integrated Intensity Process

Figure 2: Learning result for Model (18).

6.3 Performance Prediction in an Infinite Server Queue.

In the second experiment, we focus on an out-of-sample performance prediction task for an infinite server queue. Specifically, we conduct 'run-through' experiments where traffic generated from a DSPP using estimated intensity processes is used as an input to a simulation of an infinite server queue. We start with a $G_t/M/\infty$ queue, where traffic is generated using the theoretical model (18), and the deep latent and piece-wise linear models estimated in the previous section.

We generate 500 sample paths of the number of occupied servers over $[0, T]$ with service rate $\mu = 2$. Observe from Table 1 that the estimated DLM gives a reasonable inference of both the mean and variance of the number of occupied servers at $\frac{T}{2}$ and T. Note that the variance is roughly in the ballpark of the variance of the 'true' model as estimated from the test dataset. On the other hand, the piece-wise linear model, underestimates the variance quite significantly.

Table 1: Simulation of the number of occupied servers for a $G_t/M/\infty$ queue.

Number of Occupied Servers	Test		DLM		PL	
	T/2	T	T/2	T	T/2	T
Mean±CI	30.73±0.68	72.97±1.33	31.18±0.75	71.75±1.51	31.88±0.60	73.14±1.05
Variance	61.2	229.93	73.25	296.2	47.54	143.76
CI	54.26 69.56	203.87 261.36	64.94 83.26	262.62 336.68	42.16 54.04	127.47 163.41

Next, we repeat the previous experiment on a $G_t/G/\infty$ system, with Erlang distributed service times, parameterized by $\lambda = 6$ and $k = 3$ (implying the mean service time remains $\frac{1}{2}$). The simulation is

604

summarized in Table 2. Again, the DLM model makes acceptable predictions. We note that while the DLM model tends to predict higher variance, and the estimates tend to have a larger confidence interval, we conjecture that the accuracy of the predictions can be improved with a more appropriate choice of neural network size and more Monte Carlo samples in the SAA approximation to the ELBO (recall we have used $m = 5$ throughout).

Table 2: Simulation of the number of occupied servers for a $G_t/G/\infty$ queue.

Number of Occupied Servers	Test		DLM		PL	
	T/2	T	T/2	T	T/2	T
Mean±CI	37.34±0.67	85.5±1.39	37.57±0.72	84.36±1.51	38.18±0.57	87.35±1.04
Variance	58.89	253.57	68.05	298.73	42.47	140.33
CI	52.21 66.94	224.83 288.23	60.33 77.35	338.73 264.22	37.65 48.27	124.42 159.51

6.4 Impact of the Noise Factor

The previous experiments demonstrate that the DLM model is robust on both mean and variance prediction tasks. To further explore this, in this experiment we demonstrate how the DLM predictions change when the 'noise factor' η in (18) increases from 0 to 1; here, $\eta = 0$ (formally) corresponds to a deterministic intensity and $\eta > 0$ to increasing levels of stochasticity in the intensity model. We conducted the same 'run-through' experiment from the previous section on a $G_t/M/\infty$ queue, albeit with different estimated traffic models under the different η factors. While this is might appear surprising, recall that the mean number of occupied servers under the 'annealed' measure (i.e., averaged over the stochastic intensity) of an infinite server queue with DSPP traffic depends only the mean intensity function. The PL estimate, even though it is based on the 'quenched' (i.e., conditioned on the intensity) measure, accurately estimates the mean intensity when averaged over the individual sample paths. For larger η, we observe that the DLM makes reasonable predictions on the mean number of occupied servers. On the other hand, Figure 3b shows that when η increases the DLM model significantly outperforms the piece-wise linear model in predicting the variance of the number of occupied servers. This is due to the fact that the DLM estimates the annealed measure of the traffic model, while the PL model only estimates the quenched measure.
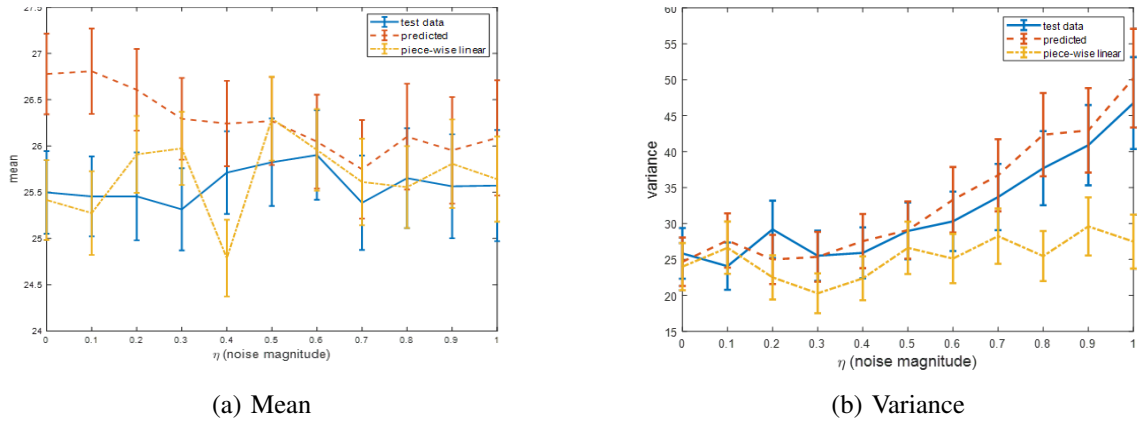

(a) Mean

(b) Variance

Figure 3: Statistical inference on the number of occupied servers as a function of the noise factor.

6.5 Estimating a Nonhomogeneous Poisson Intensity

We demonstrate the robustness of our method on estimating the intensity of a NHPP, with deterministic intensity. Consider an intensity function that is the solution of the ordinary differential equation (ODE) $\dot{Z}(t) = a(b - Z(t))$ with $a = 0.3$ and $b = 80$. Let d be the number of time intervals (or 'pieces') in the regressors, representing the number of degrees of freedom. We compare our method, using intensity

605

process $dZ(t) = a(b - Z(t))dt + d^{-1/2}\sqrt{Z(t)}dW(t)$, with the piecewise linear estimator (Zheng and Glynn 2017) and the nonparametric 'Gaussianization machine' method from Cai et al. (2019) ('GRP' in the table below). GRP uses a variance stabilizing transformation of the Poisson counts and Gaussian process regression on the transformed variables. Table 3 shows that our method is significantly better than GRP even with 50 degrees of freedom.

Table 3: Traffic count prediction for NHPP.

	Mean				Variance			
d	Test	Predicted	Piecewise Linear	GRP	Test	Predicted	Piecewise Linear	GRP
2	146.46	146.73±1.09	149.30±1.10	116.51±5.79	146.46	156.17 [138.46 177.51]	158.93 [140.91 180.65]	$4.34 * 10^3$ [$3.85 * 10^3$ $4.94 * 10^3$]
10	146.46	145.93±1.11	148.32±1.10	128.45±4.07	146.46	160.91 [142.67 182.90]	159.50 [141.42 181.30]	$2.15 * 10^3$ [$1.91 * 10^3$ $2.44 * 10^3$]
20	146.46	146.25±1.07	146.87±1.04	140.93±3.18	146.46	148.33 [131.52 168.60]	142.44 [126.30 161.91]	$1.31 * 10^3$ [$1.16 * 10^3 1.48 * 10^3$]
50	146.46	146.87±1.06	150.16±1.11	146.03±2.26	146.46	146.06 [129.50 166.02]	161.52 [143.21 183.59]	662.51 [587.41 753.06]

6.6 Estimating the Diffusion Coefficient

We presented our method under the assumption that the diffusion coefficient is known. However, Zhang et al. (2014) argue that the model in (1) is appropriate for modeling the stochastic arrival intensity in a range of service systems. Estimating this model necessitates consideration of the situation where both the diffusion and drift function are unknown. In this section we present numerical results showing that our method can work even in this situation. We assume that $\theta = 80$, $\eta = 1$ and $\alpha = \frac{1}{4}$ in the theoretical/true model. Table 4 below summarizes the results of the experiment.

We model the diffusion function by another neural network $\sigma(Z(t), t; \hat{\theta})$ with the same structure as $b(Z(t), t; \theta)$. Notice that the only difference in the training framework is that, following the definitions in Section 4, we must use $\sigma(Z(t), t; \hat{\theta})u_k(Z(t), t; \theta, \phi) = \bar{b}(Z(t), t; \phi) - b(Z(t), t; \theta)$ to define $u_k(Z(t), t; \theta, \phi)$.

Table 4: Traffic count prediction for DSPP with learnt diffusion coefficient.

Traffic Counts	Test		DLM		PL	
	T/2	T	T/2	T	T/2	T
Mean±CI	94.93±2.35	246.9±5.52	93.1±2.17	240.13±6.99	97.07±0.83	248.42±1.35
Variance	285.91	1569.52	243.26	2518.06	90.62	237.01
CI	253.50 324.99	1391.62 1784.6	215.68 276.50	2232.61 2862.2	80.35 103.01	210.15 269.41

7 Conclusions and Commentary

This paper presents a versatile computational method for estimating the stationary, ergodic stochastic intensity of a DSPP. We demonstrate our method by in-sample estimation of the intensity and out-of-sample run-through simulation experiments, both of which demonstrate accuracy of our method. We believe that the method presented in this paper demonstrates how machine learning can help enhance simulation and modeling, in the spirit of the observation made by Peter W. Glynn in his *Titans of Simulation* keynote lecture at the Winter Simulation Conference in 2019 (Glynn 2019). In future work we intend to extend our method to jump Markov intensities and self-exciting traffic models (such as the Hawkes process). In on-going work we are developing large sample statistical analyses of DLMs on general measure spaces (including asymptotic consistency and central limit theorems), and will be presented in future papers.

REFERENCES

Avramidis, A. N., A. Deslauriers, and P. L'Ecuyer. 2004. "Modeling daily arrivals to a telephone call center". *Management Science* 50(7):896–908.

Avramidis, A. N., and P. L'Ecuyer. 2005. "Modeling and simulation of call centers". In *Proceedings of the 2005 Winter Simulation Conference*, edited by M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 1805–1812. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe. 2017. "Variational inference: A review for statisticians". *Journal of the American statistical Association* 112(518):859–877.

Cai, T. T. et al. 2019. "Gaussianization Machines for Non-Gaussian Function Estimation Models". *Statistical Science* 34(4):635–656.

Cheng, R. 2017. "History of input modeling". In *Proceedings of the 2017 Winter Simulation Conference (WSC)*, edited by W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, 181–201. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Peter W. Glynn 2019, 12. "Uncertainty and Risk: Simulation, Stochastics, and Statistics".

Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep learning*. MIT press.

Jongbloed, G., and G. Koole. 2001. "Managing uncertainty in call centres using Poisson mixtures". *Applied Stochastic Models in Business and Industry* 17(4):307–318.

Kim, S.-H., and W. Whitt. 2014. "Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes?". *Manufacturing & Service Operations Management* 16(3):464–480.

Kingma, D. P., and J. Ba. 2014. "Adam: A method for stochastic optimization". *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., and M. Welling. 2019. "An Introduction to Variational Autoencoders". *Foundations and Trends® in Machine Learning* 12(4):307–392.

Kunita, H. 1984. "Stochastic differential equations and stochastic flows of diffeomorphisms". In *Ecole d'été de probabilités de Saint-Flour XII-1982*, 143–303. Springer.

Li, X., T.-K. L. Wong, R. T. Chen, and D. Duvenaud. 2020. "Scalable Gradients for Stochastic Differential Equations". *arXiv preprint arXiv:2001.01328*.

Maman, S. 2007. "Uncertainty in the Demand for Service: The Case of Call Centers and Emergency Departments M. Sc. Research Proposal".

Segall, A., and T. Kailath. 1975. "The modeling of randomly modulated jump processes". *IEEE Transactions on information theory* 21(2):135–143.

Tzen, B., and M. Raginsky. 2019a. "Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit". *arXiv preprint arXiv:1905.09883*.

Tzen, B., and M. Raginsky. 2019b. "Theoretical guarantees for sampling and inference in generative models with latent diffusions". *arXiv preprint arXiv:1903.01608*.

Van Handel, R. 2007. *Filtering, stability, and robustness*. Ph. D. thesis, California Institute of Technology.

Whitt, W. 1999. "Dynamic staffing in a telephone call center aiming to immediately answer all calls". *Operations Research Letters* 24(5):205–212.

Zhang, X., L. J. Hong, and J. Zhang. 2014. "Scaling and modeling of call center arrivals". In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk, S. D. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, 476–485. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Zheng, Z., and P. W. Glynn. 2017. "Fitting continuous piecewise linear poisson intensities via maximum likelihood and least squares". In *Proceedings of the 2017 Winter Simulation Conference*, edited by W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, 1740–1749. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

Ruixin Wang is a Ph.D. student in the School of Industrial Engineering at Purdue University. His research interests lie in approximate dynamic programming, electricity market modeling and stochastic simulation. His email address is wang2252@purdue.edu.

PRATEEK JAISWAL is a Ph.D. candidate in the School of Industrial Engineering at Purdue University. His research interests are in machine learning and stochastic optimization. His e-mail address is jaiswalp@purdue.edu.

HARSHA HONNAPPA is an assistant professor in the School of Industrial Engineering at Purdue University. His research interests are in applied probability, game theory, and machine learning. He is a member of INFORMS, IEEE, and SIAM, and serves as an associate editor for Operations Research and Operations Research Letters. His email address is honnappa@purdue.edu.