Predicting Distributions of Waiting Times in Customer Service Systems using Mixture Density Networks

Majid Raeis, Ali Tizghadam and Alberto Leon-Garcia
Department of Electrical and Computer Engineering
University of Toronto, Toronto, ON, Canada
Emails: m.raeis@mail.utoronto.ca, ali.tizghadam@utoronto.ca and alberto.leongarcia@utoronto.ca

***Abstract*—Motivated by interest in providing more efficient services in customer service systems, we use statistical learning methods and delay history information to predict the conditional distribution of the customers' waiting times in queueing systems. From the predicted distributions, descriptive statistics of the system such as mean, variance and percentiles of the waiting times can be obtained, which can be used for delay announcements, SLA conformance and better system management. We model the distributions by mixtures of Gaussians, parameters of which can be estimated using Mixture Density Networks. We use the extensions of the Lindley's equation for multi-server queues to generate our datasets. The evaluations show that exploiting more delay history information can result in much more accurate predictions under realistic time-varying arrival assumptions.**

***Index Terms*—Queueing Systems, Waiting Time Prediction, Mixture Density Networks**

I. INTRODUCTION

Services are by definition intangible products that are experienced by the customers. Services may be defined in different contexts such as information technology (IT), telecommunication, transportation, healthcare, banking, etc. Because of the intangible nature of services, quality of service mainly depends on the customers' experience of the received service. One of the important measures of the quality of service is the delay experienced by the customers. Any service system has a limited service capacity and therefore is not capable of providing service to all customers at the same time, during high-demand periods. As a result, customers usually have to wait in queues in order to receive service. In other words, servers are analogous to shared resources among customers, which cannot be utilized by all the customers at the same time. In addition to the service capacity limitations of the system, time-variability and randomness of the demand are other important factors which can lead to queue formation.

Waiting time prediction for the new customers can be beneficial from both the customers' and the service providers' points of view. Service providers can use delay predictions for better management of the system by adaptive matching of the service capacity to the demand. On the other hand, waiting time predictions can be used for delay announcements to the customers, which will result in lower uncertainty about the waiting times, and therefore, higher customer satisfaction [1].

Call centers are one of the important examples of the customer service systems, which have been studied extensively in the operations research field. In these systems, uncertainty about the waiting time is high and the customers have no means to estimate the progress rate. Therefore, waiting time estimation becomes even more important in these systems, because of the invisible queue problem. IT help desk is another closely related example, which provides IT support to the employees in a company using a ticket management system. The same kind of problems can be observed in many other service contexts such as retail stores, manufacturing systems, hospital emergency rooms, etc.

Previous Work

Waiting time prediction in service systems can be classified into two categories: queueing-theoretic methods and data-based methods [1]. Let us first begin with the queueing-theoretic methods.

One of the earliest work on predicting customer's waiting time in a multi-server queue is [2]. This paper investigates the possibility of improving delay predictions by exploiting information about the system state, and the elapsed service time of the customers in service, under non-exponential service time assumptions. Following up on [2], Ibrahim and Whitt have studied the performance of alternative queue-length-based and delay-history-based predictors. Three types of delay history information which have been used widely in these papers are the delay of the last customer to enter service (LES), the elapsed waiting time of the customer at the head of the line (HOL) and the delay of the last customer to complete service (LCS). The real-time performance of delay-history-based predictors such as LES and HOL predictors are studied for the standard $GI/M/c$ queueing model in [3]. In [4], Ibrahim and Whitt extend their analysis to queueing models with *abandonments*. Particularly, they study the overloaded $GI/GI/c{+}GI$ model, where $+GI$ denotes i.i.d abandonment times with general distributions. It is shown that the queue-length predictor performs poorly in this regime, while HOL remains an effective estimator. The performance of the delay-history based predictors in multi-server queueing systems with *non-stationary* arrivals is studied in [5]. It is shown that the delay-history based predictors can have significant estima-

978-3-903176-24-9 © 2019 Crown

tion bias, particularly if the system goes through alternating overload and underload periods. Moreover, a refined delay estimator based on HOL is introduced in [5] to cope with time-varying arrivals in the $M(t)/GI/c+GI$ model, where $M(t)$ represents non-homogeneos Poisson arrivals. Finally, *time-varying capacity* is taken into account in [6], where four new predictors have been introduced for the $M(t)/M/c(t)+GI$ queueing model.

Limitations of the queueing-theoretic analysis have resulted in recent interest in data-based methods such as machine-learning algorithms and data-mining techniques. These methods have been used for waiting time prediction in different contexts and more realistic settings such as healthcare and transportation systems. For instance, machine learning techniques have been used in [7] to predict flight delays by exploiting available information from sensors in the airport. Combining process mining and queueing-theoretic results, [8] introduces queue-mining techniques for predicting waiting times in service systems. Ang et al. [9] propose a new predictor, called Q-Lasso, which combines the Lasso method from statistical learning and fluid models from queueing theory. The authors consider waiting time prediction in emergency departments and use datasets from four hospitals.

Both of these methods have their own advantages and shortcomings. One of the main disadvantages of the queueing-theoretic methods is that the analysis can easily become intractable by considering more realistic assumptions or including more information sources in the prediction process. Particularly, consider delay-history-based prediction, which is the focus of this paper. The existing queueing-theoretic methods based on delay history information are often limited to exponential service time assumptions and stationary arrival times. Moreover, the refined delay-history-based predictors that have been introduced for time-varying arrival settings, such as the ones proposed in [5], require knowledge of the model parameters including arrival rate, service rate and the number of servers, which might not be available and need to be estimated as well. Another shortcoming of the existing queueing-theoretic methods is their limitation in exploiting all the available information, such as delay history of multiple recent customers, instead of only focusing on a single customer's delay history (e.g., LES or LCS). On the other hand, the prediction method and the feature selection process in data-based predictions are usually specialized for a particular application and do not provide much insights about the importance of the features. Furthermore, uncertainty of the estimations and the distribution of the waiting times are other important pieces which are often missing in the literature. All these reasons motivated us to use statistical learning methods to study queueing models under more realistic assumptions, such as time-varying arrivals and non-exponential service times.

The remainder of this paper is organized as follows. In Section II, we will describe the queueing system model and formulate the problems that we are going to study in this paper. We explain the data set generation process in Section III. In Section IV, we briefly review mixture density networks and some problems associated with them. The evaluation of the proposed predictors are presented in Section V. Finally, Section VI presents the conclusions and the future work.

II. System Model and Problem Setting

Consider a multi-server queueing system with infinite queue size and C homogeneous servers with FCFS service discipline. We do not assume a specific distribution for service times or inter-arrival times and therefore, the arrival and service processes can have non-stationary distributions. Let w_i denote the observed waiting time (before entering service) of the i'th last customer who entered service. Based on this definition, w_1 represents the LES delay, i.e., $w_1 = w_{LES}$. Furthermore, the random waiting time of a new arrival conditional on observed delay history of the last h customers who entered service, i.e. $\mathbf{w}_h = (w_1, w_2, \cdots, w_h)$, is represented by $W(\mathbf{w}_h)$.

Our goal is to predict the distribution of a new arrival's waiting time, given that the customer has to wait and an observed delay history of $\mathbf{w}_h = (w_1, w_2, \cdots, w_h)$, i.e., we are aiming to predict the conditional distribution $P(W|\mathbf{w}_h)$. Furthermore, we are interested in predicting a single-value prediction for $W(\mathbf{w}_h)$, which will be denoted by $\widehat{W}(\mathbf{w}_h)$ in the rest of the paper. In the case where the delay predictor directly uses delay of the last customer to enter service as its prediction, i.e., $W(\mathbf{w}_h) \equiv w_1 = w_{LES}$, we get the LES estimator. However, we are primarily concerned with the MMSE predictions, which can be obtained as $\widehat{W}(\mathbf{w}_h) \equiv E[W(\mathbf{w}_h)]$. In other words, $E[W(\mathbf{w}_h)]$ minimizes the mean squared error (MSE) of the predictor, which is defined as

$$\text{MSE}(\widehat{W}(\mathbf{w}_h)) \equiv E\left[\left(W(\mathbf{w}_h) - \widehat{W}(\mathbf{w}_h)\right)^2\right]. \tag{1}$$

It should be mentioned that it is difficult to determine MMSE predictor theoretically, even for the simple case of $h=1$, and therefore, we use statistical learning methods to estimate the conditional mean of $W(\mathbf{w}_h)$.

Our approach for estimating the conditional distribution of the waiting time is to use mixture density networks (MDNs) [11]. In particular, the MDN uses a mixture of Gaussians to estimate the conditional distribution of the waiting time as follows

$$P(W|\mathbf{w}_h) = \sum_{k=1}^{K} \pi_k(\mathbf{w}_h) \mathcal{N}(W|\mu_k(\mathbf{w}_h), \sigma_k^2(\mathbf{w}_h)), \tag{2}$$

where $\pi_k(\mathbf{w}_h) \in (0,1)$, $\mu_k(\mathbf{w}_h)$ and $\sigma_k^2(\mathbf{w}_h)$ denote the the mixing coefficient, mean and variance of the k^{th} kernel, respectively, given delay history information $\mathbf{w}_h$. We will discuss this method in more detail in Section IV.

As we mentioned earlier, an important reason for estimating distribution of the delay is to obtain probabilistic bounds instead of just making predictions. More specifically, we can define a probabilistic lower-bound (w_{lb}) and upper-bound (w_{ub}) as follows:

$$P(W(\mathbf{w}_h) > w_{ub}) \leq \varepsilon_{ub}, \tag{3}$$

$$P(W(\mathbf{w}_h) < w_{lb}) \leq \varepsilon_{lb}, \tag{4}$$

where ε_{ub} and ε_{lb} are the violation probabilities for the upper-bound and the lower-bound, respectively. Since probabilistic upper bound provides a pessimistic estimation of the waiting time, it can be a good candidate for delay announcements to the customers. Finally, confidence interval is one of the other statistics that will be used in this paper to measure the amount of uncertainty for each prediction. Since the confidence intervals will be used along with the MMSE predictions, we define the confidence interval for random waiting time $W(\mathbf{w}_h)$ with confidence level P_{cl}, as an interval with endpoints $(E[W(\mathbf{w}_h)] - x, E[W(\mathbf{w}_h)] + x)$ such that:

$$P(E[W(\mathbf{w}_h)] - x < W(\mathbf{w}_h) < E[W(\mathbf{w}_h)] + x) = P_{cl} \quad (5)$$

III. Data set Generation

In this section, we describe the data set generation process and the features that will be used for the learning problem. We consider delay-history-based predictors and investigate the effect of increasing delay history information on the predictor's performance, as well as the distribution of the delay. The existing works on delay-history-based predictors only focus on delay history of a single customer such as LES or LCS, while exploiting delay history of a larger number of customers who have entered service or completed service might result in more accurate predictions, particularly in the case of time-varying arrivals. Since LES delay provides more up-to-date information than LCS in a multi-server system with large number of servers [3], we construct our data set based on the LES delays. We use the extension of the Lindley's equation to the multi-server systems to calculate departure times, waiting times and some other parameters for each customer. Based on the calculated parameters, we build our data set such that each data sample belongs to a particular customer and comprises h features corresponding to the delay history of the last h customers to enter service. The data set generator receives the arrival and service times for each customer as input, which can be generated using different distributions or fed from real data, and produces the data set. Now, let us explain the data set generation in more detail.

Lindley's equation [10] is a recursive equation in terms of consecutive waiting times in a single-stage single-server queue with FIFO discipline. Let ω_n denote the waiting time of the n^{th} customer, then ω_{n+1} can be easily obtained using Lindley's equation as follows:

$$\omega_{n+1} = \max\{0, \omega_n + s_n - \tau_n\}, \quad (6)$$

where s_n is the service time of the n^{th} customer, and τ_n represents the inter-arrival time between the n^{th} and $n + 1^{\text{th}}$ customer arrivals.

Since the multi-server queue model does not have the FIFO property, the recursive relationship between consecutive waiting times would not be as simple as Eq. (6). There exist multiple extensions of the Lindley's equation to the multi-server case. Kiefer and Wolfowitz [12] proposed an algorithmic recursive procedure, where computing the waiting time of a particular customer requires sorting a list of C elements, where C denotes the number of the servers. We compute departure times in a multi-server system using a similar approach to [12]. Now, let us briefly describe the recursive procedure for computing departure times starting with empty queue and servers. Consider a multi-server queue with C servers. This procedure keeps a sorted list of scheduled departure times, defined as $D = (d_1, d_2, \cdots, d_C)$, where d_i represents the i^{th} scheduled departure at a given time. Starting with an empty queue and servers, D is initialized with $[a_1 + s_1, a_2 + s_2, \cdots, a_C + s_C]$ and sorted in ascending order, since the first C customers enter servers without any delay. Therefore, $D[1]$ gives the 1^{st} customer departure time. In order to calculate the departure time of the second customer, the list should be updated as follows. Since the $(C+1)^{\text{th}}$ customer will enter service as soon as one of the servers becomes available, the departure time of customer $C + 1$ will be equal to $\max\{a_{C+1}, d_1\} + s_{C+1}$. This departure time will be appended to list D and then the list is sorted in ascending order. Now, the first element of the updated list, $D[1]$, gives the second customer departure time and the new list can be used in a similar procedure to obtain the next departure times. We can calculate LES delays for each customer using a similar approach presented in Algorithm 1.

Algorithm 1 Calculating departure times and service entrance times

```
$E \leftarrow [a_1, a_2, \cdots, a_C]$
$D_{temp} \leftarrow [a_1 + s_1, a_2 + s_2, \cdots, a_C + s_C]$
$D \leftarrow \text{sort}(D_{temp})$
for $i = 1$ to $N$ do
    $e_i \leftarrow E_1$ {$i^{th}$ service entrance time}
    $d_i \leftarrow D_1$ {$i^{th}$ service departure time}
    if $i \leq N - C$ then
        $e_{temp} \leftarrow \max\{a_{C+i}, d_i\}$
        $d_{temp} \leftarrow \max\{a_{C+i}, d_i\} + s_{C+i}$
        $E \leftarrow [E_2, E_3, \cdots, E_C, e_{temp}]$
        $D \leftarrow \text{sort}([D_2, D_3, \cdots, D_C, d_{temp}])$
    end if
end for
```

IV. Mixture Density Networks (MDNs)

The mixture density networks provide a general framework for approximating arbitrary conditional distributions. Using a mixture model as an approximation of the true conditional distribution, an MDN estimates the parameters of the mixture model using a fully-connected neural network. The output layer consists of three types of nodes which predict the parameters of the mixture model in Eq. (2). The first type uses the soft-max activation function to predict the mixing coefficients such that $0 \leq \pi_k \leq 1$ and $\sum_k \pi_k = 1$. The second group, which predict the variances of the kernels, use exponential activations to ensure non-negative values. The last group of the nodes use linear activations and compute the means of the kernels. Using a data set of N_{sample} observations and their corresponding target values, $\{(\mathbf{x}_n, \mathbf{y}_n) | 1 \leq n \leq N_{sample}\}$, the mixture density network learns the weights of the neural network by minimizing the error function, which is defined as the negative logarithm of the likelihood, i.e.,

$$E = -\sum_{n=1}^{N_{sample}} \ln\left\{P(\mathbf{y}_n|\mathbf{x}_n)\right\}. \tag{7}$$

It is worth mentioning that a standard feed-forward neural network with a single linear output unit trained by least squares, corresponds to maximum likelihood with a Gaussian distribution assumption [11]. As a result, the output of this simple neural network will approximate the conditional mean of the waiting time and therefore can be used to estimate the MMSE predictions.

One of the main challenges in implementing a mixture density network is its instability issue. In order to address this problem, various techniques and modifications have been proposed. An important problem associated with mixtures of Gaussians is the presence of singularities in the Likelihood function. In other words, in the case of having more than one Gaussian component, $K \geq 2$, the maximization of the Likelihood function is not a well-posed problem since the Likelihood can easily go to infinity whenever one of the Gaussian components lies on a specific data point and its variance goes to zero. For a more detailed discussion of the implementation issues related to mixture density networks refer to [13].

V. Evaluation and Results

A. Performance Measures

In order to evaluate the performance of the delay predictors, we consider two measures: *absolute bias* and *mean squared error* (MSE) for accuracy and precision, respectively. The absolute bias is defined as $\text{Bias}(\widehat{W}) = |E[W - \widehat{W}]|$ and will be approximated by

$$\widehat{\text{Bias}} = \left| \frac{1}{N_{sample}} \sum_{i=1}^{N_{sample}} (d_i - p_i) \right|, \tag{8}$$

where d_i and p_i are the ground-truth and predicted waiting times for the i^{th} data point.

The other measure is defined as $\text{MSE}(\widehat{W}) = E[(W - \widehat{W})^2]$ and will be approximated by the *average squared error* as follows:

$$\text{ASE} = \frac{1}{N_{sample}} \sum_{i=1}^{N_{sample}} (d_i - p_i)^2. \tag{9}$$

B. Simulation Experiments

In the rest of this section, we present our results on delay prediction and distribution estimation in multi-server queueing systems. Let us begin with a short description of the arrival and service processes that are used in the following experiments. First, we consider a deterministic ON-OFF arrival process which can be used for simple approximation of a system with batch arrivals, such as transport terminals, in which arrival and departures occur in batches based on schedules. The ON-OFF arrivals have a cycle length of 4 hours and a duty cycle equal to 75%. It should be noted that time is normalized to the mean service time in this paper. The second type of time-varying arrivals that are used in the experiments is the Non-homogeneous Poisson process (NHPP), which is a good fit for arrivals in hospitals and call centers [14]. We adopt the same model as in [5] with sinusoidal arrival rate to capture the daily cycles, i.e., we consider an arrival rate of $\lambda(t) = \bar{\lambda}(1+\alpha\sin(2\pi t/T))$, where $\bar{\lambda}$, α and T represent the average arrival rate, relative amplitude and the cycle length of the arrival rate. We have evaluated systems with exponential, lognormal and H2 service times (hyperexponential with coefficient of variation equal to 2), however, the results are only presented for lognormal service times, due to space considerations. Moreover, we use an MDN implementation which is based on [15] and uses the Keras deep learning library. The training data set consists of around 27000 samples and the test results are evaluated on 5000 sample customers. As mentioned earlier, since we are studying delay-history-based predictors, we only consider delay of the last h customers to enter service as the feature set. The simulation parameters are summarized in Table I.

TABLE I
Parameters of the simulation.

Notation	Definition	value
$E[s]$	Mean service time	1 (10 mins)
ρ	Traffic intensity	0.95
T	Cycle of NHPP arrivals	144 (1 day)
α	Relative amplitude of $\lambda(t)$	0.5
C	Number of servers	20

In the first experiment, we consider a multi-server queueing system with 20 servers, deterministic ON-OFF arrival and lognormal service times. We use a simple feed-forward neural network to approximate the conditional mean of the delay given previous history, i.e., $E[W(\mathbf{w}_h)]$, and use it as our predictor. Fig. 1 shows the scatter plots of the ground-truth and predicted delays for the LES predictor, and three other NN-based predictors with delay history lengths of $1, 25$ and 50. As can be seen in Fig. 1.a, there exists a time lag between the LES delays and the ground-truth delays, particularly when the system is busy and the ground-truth values are large. On the other hand, we can observe that even a simple feed-forward neural network which only relies on the LES delay can achieve a much better performance and reduce the MSE by around 65% compared to the LES predictor (see Fig. 1.b). Furthermore, the absolute bias has been reduced tremendously, which suggests that the systematic time lag does not exist anymore. It seems that increasing the delay history information does not have a noticeable impact on MSE in this experiment and might even result in overfitting and therefore a larger MSE (Fig. 1.d). From Fig. 1 we can observe that the predictors can have large errors when the ground-truth delay is very small. One explanation might be that large past delays can either imply large delays for the next customers if the ON period

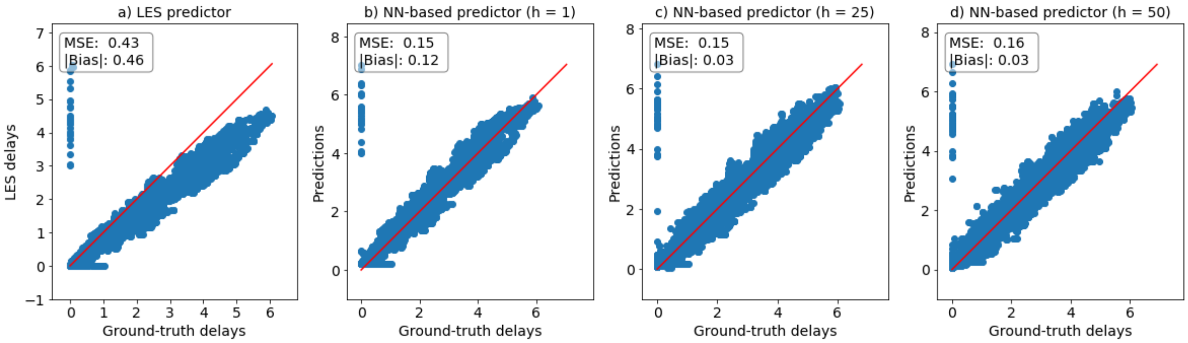

Fig. 1. Scatter plots of the predicted delays and ground-truth delays

continues, or suggest very small delays if the arrival process goes to the OFF state. The vertical group of the outliers around zero ground-truth delay represent this phenomenon.

Now, consider the same queueing system but with NHPP arrivals. Fig. 2 shows the scatter plots for the same set of predictors. Similar to Fig. 1, there is a time lag between the LES delays and the ground-truth values. Moreover, Fig. 2.a suggests that the conditional distribution of the delay given LES delay information should be multimodal. In other words, a particular LES delay can suggest both larger or smaller future delays, depending on whether the arrival rate and therefore system backlog, is in the increasing or decreasing phase. We observe that by increasing the length of the delay history to $h = 50$, the NN-based predictor can better learn whether the system backlog is in the increasing or decreasing phase and hence, it is able to reduce the MSE by around 73%.

As we mentioned earlier, single value predictions of the waiting times might not be very informative and we are interested in more descriptive statistics of the system. For instance, it is more desirable to obtain stochastic upper bounds or lower bounds on a customer's delay similar to Eqs. (3) and (4), rather than providing a single value prediction. In order to achieve this goal, we use MDNs as described in Section IV to estimate the conditional distribution of the waiting time given delay history of the last h customers to enter service. Fig. 3 shows the predicted conditional distribution of the waiting time in the previous experiment with NHPP arrivals, given LES delays equal to 5, 10 and 15, i.e., $P(W|\mathbf{w}_1 = 5)$, $P(W|\mathbf{w}_1 = 10)$ and $P(W|\mathbf{w}_1 = 15)$, respectively. The two modes in the estimated distribution function, correspond to the two groups of points in the scatter plot shown in Fig. 2.a, with the same LES delay value. Moreover, Fig. 3 suggests that given a particular LES delay, it is more likely to have a larger delay for the new arrival. This can be explained by the fact that most of the customers with the same LES delay should have entered the system in the increasing phase of the arrival rate, while the queue length is growing, and therefore they are more likely to experience larger delays than the LES customer.

Now, we use the estimated conditional distributions in a more effective way to obtain probabilistic bounds on the waiting time of a new arrival. Fig. 4 shows a sample path of the waiting times in the previous experiment with NHPP arrivals for a period of two days. The stochastic upper bounds and lower bounds calculated from Eqs. (3) and (4) are also shown in Fig. 4. The stochastic bounds are calculated to hold with a probability more than 0.95, i.e, $\varepsilon_{ub} = \varepsilon_{lb} = 0.05$. It should be mentioned that decreasing the violation probabilities can result in looser bounds. As mentioned earlier, the calculated upper bounds are good candidates for delay announcements. Closely related to these bounds, we can find the confidence intervals for each prediction using Eq. (5) and the predicted distributions. Fig. 5 shows the same sample path of the waiting times as in Fig. 4, along with the MMSE predictions and 95% confidence intervals. We should emphasize that the confidence intervals are calculated based on the predicted distributions, rather than a simple Gaussian assumption. It can be observed that the confidence intervals can be pretty large for longer waiting times, which shows the importance of considering other statistics instead of focusing on single value predictions.

VI. Conclusions

In this paper, we attempted to show the potential of the statistical learning methods in providing insights on service systems under realistic assumptions, such as time-varying arrivals and non-exponential service times. In particular, we studied the problem of delay prediction and distribution estimation in multi-server queueing systems. We showed that even a very simple NN-based predictor that only uses the delay history information of the previous customers can outperform the traditional LES predictor (decreasing MSE by 73%), without requiring any other information about the system parameters. More importantly, MDNs enable us to estimate the conditional distribution of the waiting time, which can be used to obtain much more informative statistics, such as probabilistic bounds, compared to the common single-value predictions.

Although the proposed NN-based methods are able to make good estimations for service systems under fixed model assumptions, we are interested in studying more realistic and complex cases, where model parameters such as the number of servers and service time distributions can change over time. As discussed in the previous sections, these systems can appear in many different contexts, such as call centers and emergency

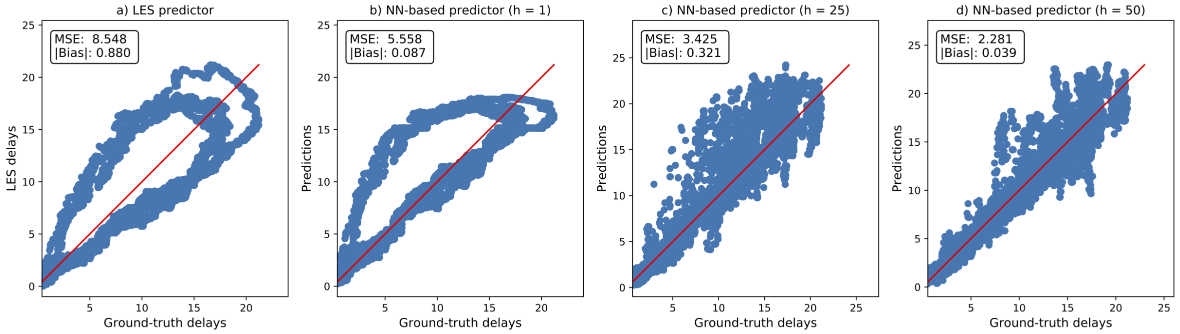

Fig. 2. Scatter plots of the predicted delays and ground-truth delays

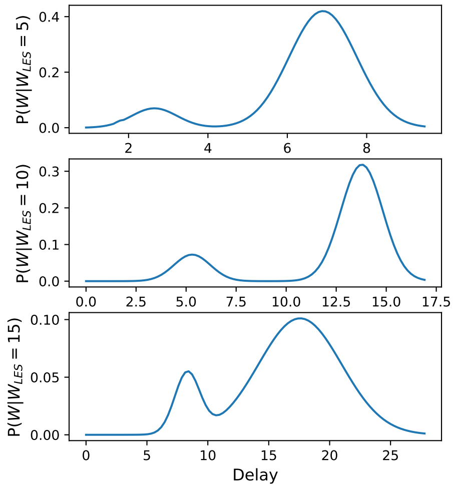

Fig. 3. Estimated probability density function $P(W|w_1)$, for $w_1 = 5, 10$ and 15.

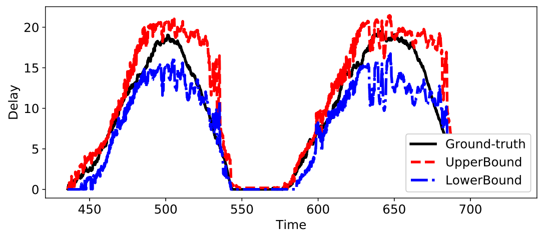

Fig. 4. Probabilistic upper bound and lower bound on the waiting times with less than 5% violation probabilities, given delay history of the last $h = 50$ customers who entered service.

departments, where theoretical analysis of the system under realistic assumptions gets intractable.

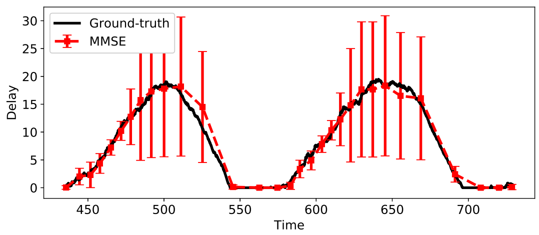

Fig. 5. MMSE predictions along with 95% confidence intervals, given delay history length of $h = 50$.

References

[1] Ibrahim, R.: Sharing delay information in service systems: a literature survey. Queueing Syst. (2018)

[2] Whitt, W.: Predicting queueing delays. Manag. Sci. 45(6), 870–888 (1999b)

[3] Ibrahim, R.,Whitt,W.: Real-time delay estimation based on delay history. Manuf. Serv. Oper.Manag. 11(3), 397–415 (2009a)

[4] Ibrahim, R., Whitt, W.: Real-time delay estimation in overloaded multiserver queues with abandonments. Manag. Sci. 55(10), 1729–1742 (2009b)

[5] Ibrahim, R., Whitt, W.: Real-time delay estimation based on delay history in many-server service systems with time-varying arrivals. Prod. Oper. Manag. 20(5), 654–667 (2011a)

[6] Ibrahim, R., Whitt, W.: Wait-time predictors for customer service systems with time-varying demand and capacity. Oper. Res. 59(5), 1106–1118 (2011b)

[7] Demir, E., Demir, V.B.: Predicting flight delays with artificial neural networks: case study of an airport. In: 2017 25th IEEE Signal Processing and Communications Applications Conference (SIU), pp. 1–4 (2017)

[8] Senderovich, A., Weidlich, M., Gal, A., Mandelbaum, A.: Queue mining–predicting delays in service processes. In: International Conference on Advanced Information Systems Engineering, pp. 42–57. Springer, Berlin (2014)

[9] Ang, E., Kwasnick, S., Bayati, M., Plambeck, E., Aratow, M.: Accurate emergency department wait time prediction. Manuf. Serv. Oper. Manag. 18(1), 141–156 (2015)

[10] D. V. Lindley. The theory of queues with a single server. Mathematical Proceedings of the Cambridge Philosophical Society, 48(2):277–289, 1952.

[11] C. Bishop. Mixture density networks. Technical report, 1994.

[12] J. Kiefer and J. Wolfowitz. On the theory of queues with many servers. Transactions of the American Mathematical Society, 78(1):1–18, 1955.

[13] A. Brando Guillaumes, “Mixture density networks for distribution and uncertainty estimation,” Master’s thesis, Universitat Politecnica ‘ de Catalunya, 2017.

[14] Kim, S. H. and Whitt, W. Are Call Center and Hospital Arrivals Well Modeled by Nonhomogeneous Poisson Processes? M&SOM 16 464–480, 2014.

[15] https://github.com/cpmpercussion/keras-mdn-layer