Construction of Optimal Prediction Intervals for Load Forecasting Problems

Abbas Khosravi, *Member, IEEE*, Saeid Nahavandi, *Senior Member, IEEE*, and Doug Creighton, *Member, IEEE*

***Abstract*—Short-term load forecasting is fundamental for the reliable and efficient operation of power systems. Despite its importance, accurate prediction of loads is problematic and far remote. Often uncertainties significantly degrade performance of load forecasting models. Besides, there is no index available indicating reliability of predicted values. The objective of this study is to construct prediction intervals for future loads instead of forecasting their exact values. The delta technique is applied for constructing prediction intervals for outcomes of neural network models. Some statistical measures are developed for quantitative and comprehensive evaluation of prediction intervals. According to these measures, a new cost function is designed for shortening length of prediction intervals without compromising their coverage probability. Simulated annealing is used for minimization of this cost function and adjustment of neural network parameters. Demonstrated results clearly show that the proposed methods for constructing prediction interval outperforms the traditional delta technique. Besides, it yields prediction intervals that are practically more reliable and useful than exact point predictions.**

***Index Terms*—Load forecasting, neural network, prediction interval.**

I. Introduction

TO REMAIN competitive in the privatized and deregulated markets of power generation, it is vital for companies to reduce their operating cost. Over estimation of loads may lead to excess supply, and consequently, increment of operational costs. On the other hand, under estimation may result in loss of reliability to supplying utilities. Therefore, formulating optimal strategies and schedules for generating power is of utmost important for utility companies. Such planning can potentially save millions of dollars per year for utility companies [1]. Furthermore, many operational activities within power systems, including, among others, unit commitment, economic dispatch, automatic generation control, security assessment, maintenance scheduling, and energy commercialization are usually scheduled on the basis of short-term load forecasting (STLF). The lead time of forecast may vary between minutes to days.

Motivated by these, a countless number of numerical studies have been reported in the scientific and industrial literature. Loosely speaking, all STLF methods can be divided into two broad categories: statistical methods (parametric), and artificial intelligence-based techniques (nonparametric). Statistical methods include regression models (linear or piecewise-linear) [2], Kalman filter [3], and time series (autoregressive moving average models) [4], [5]. The inherent complexity and nonlinearity of relationships between electric loads and their exogenous variables make application of these techniques for load forecasting problematic. Forecasters developed based on these techniques are often prone to bias [6].

On the contrary, artificial intelligent based techniques, and in particular neural networks (NNs), possess an excellent capability of learning and approximating nonlinear relationships to any arbitrary degree of accuracy (universal approximators) [7]. Applications of expert systems [8], [9], NNs [10]–[12] (and references therein), fuzzy systems [13], and neuro-fuzzy systems [14] have proliferated for STLF within the last two decades. It has been also stated that the majority of commercial STLF packages used by utility companies have been developed based on artificial intelligent-based techniques (mainly NNs) [15], [16]. A good review of NN-based STLF can be found in [11], [12], and references therein.

Recently reported reviewing studies indicate that in many engineering and science fields NNs significantly outperform their traditional rivals in term of prediction and classification accuracy [17]. There is, however, some skepticism related to the performance of NNs for STLF [12] (and references therein). It has been mentioned that in the majority of conducted studies, NN models have been 1) unnecessarily very large and 2) overfitted. The first problem can be easily managed through developing NNs in a *constructive* approach [18]; i.e., NN complexity is increased whenever it does not satisfy the prediction requirements. Practicing this principle satisfactorily guarantees minimality of NN size. Overfitting can be also avoided through using theoretically well-established methods such as Bayesian learning algorithm or weight decay cost function technique [7].

Despite countless reports on successful application of NNs for STLF, here we argue that modelers have often lost sight of a basic characteristic of NNs. NN models are theoretically deterministic [7], and by that, their application of predicting future of stochastic systems is always in doubt and questionable [19]. It is empirically very important to notice that loads often show completely nonlinear and in some cases chaotic behaviors. Their fluctuations through the time are erratic and influenced by many known or unknown factors. In either case, often information about influencing factors is uncertain. Unreliability of forecasts of weather conditions and temperature variations are often high. Although local system failures are compensable though considering power generation surplus, they may dramatically change system behavior and stability. Uncertainties and probabilistic

Manuscript received October 12, 2009. First published March 11, 2010; current version published July 21, 2010. This work was supported by the Centre for Intelligent Systems Research (CISR) at Deakin University. Paper no. TPWRS-00805-2009.

The authors are with the Centre for Intelligent Systems Research (CISR), Deakin University, Geelong, Vic 3117, Australia (e-mail: akhos@deakin.edu.au; nahavand@deakin.edu.au; dougc@deakin.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TPWRS.2010.2042309

events highly contribute to the degradation of performance of NN models for load forecasting. Negative consequences raised from the stochastic nature of power systems cannot be compensated solely through increasing NN size (neither hidden layers nor neurons) or repeating its training procedure. With the presence, occurrence, and accumulation of these uncertainties and probabilistic events, power systems look like a stochastic system with volatile behaviors in term of load demands in future. As there is more than one probable reality for future of these systems (load demands in future), any claim about accuracy of future prediction is dubious and untrustworthy.

Seeking to remedy these defects, construction of prediction intervals (PIs) has been proposed in literature. By definition, a PI with confidence level of $(1-\alpha)\%$ is a random interval developed based on past observations, $X = (x_1, x_2, \ldots, x_n)$ for future observations, $\mathrm{PI} = [L(X), U(X)]$, such that $\Pr(L(X) < x_{n+1} < U(X)) = 1-\alpha$. PIs indicate the expected error between the prediction and the actual targets. Furthermore, they convey more meaningful information than predicted point values. Of utmost importance is level of confidence, giving PIs an indication of their reliability. In literature, different schools of methods exist for construction of PIs: 1) delta technique [20], [21], 2) Bayesian technique [7], [22], 3) bootstrap [23], [24], and 4) mean-variance estimation [25]. The cornerstone of the delta technique lies in interpreting NNs as nonlinear regression models and linearizing them based on Taylor's series expansion [20]. The Bayesian technique interprets the NN parameter uncertainty in terms of probability distributions and integrates them to obtain the probability distribution of the target conditional on the observed training set [7]. The bootstrap technique is essentially a resampling method that its computation requirement is massive. The fourth school is implemented through developing two NNs for prediction of mean and variance of targets. Selection of any of these techniques for constructing PIs depends on problem domain, computation burden, number of available samples, and analysis purpose. Construction of PIs has been a subject of much attention in recent years. Examples are temperature prediction [26], travel time prediction in baggage handling system [27], [28], watershed simulation [29], solder paste deposition process [30], and time series forecasting [31].

To the best of our knowledge, power engineering field, and in particular STLF domain is void of information about supporting theories and applications of PIs. Motivated by these gaps in practical and scientific research, one fold of this study aims at applying the delta technique to the STLF problem. Instead of developing and exploiting NNs for yielding exact load forecasts, PIs with a high confidence level ($(1-\alpha)\%$) are constructed for future loads. In experiments with real data, it is demonstrated that PIs are empirically more useful and reliable than exact point predictions.

Another fold of this research concentrates on designing practical indices and measures for quantitative evaluation of PIs. Literature only offers a measure for evaluating coverage probability of PIs. Often discussion about length of PIs (and similarly for confidence intervals) is ignored or represented ambiguously [30], [32]–[34]. Here, we propose a new measure for quantitative evaluation of PIs that covers both aspects of PIs: length and coverage probability. With regard to this new index, a new cost function is developed for improving quality of PIs (squeezing PIs without compromising their coverage probability). Ample care is exercised in definition of the new cost function to keep fundamental assumptions of the delta technique valid. As calculation of mathematical characteristics of this new cost function is very problematic (if not impossible), gradient-based optimization methods are not applicable for its minimization. Therefore, stochastic optimization techniques should be employed for its minimization. In this study, simulated annealing (SA) is adopted for minimization of this cost function in order to adjust NN parameters. It is shown that PIs developed using the optimized NNs are effectively narrower with at least the same coverage probability like PIs constructed using NNs trained based on traditional techniques such as Levenberg-Marquardt technique [7].

The rest of this paper is organized as follows. Section II provides a brief review of fundamental theories of the delta technique. The new PI assessment measure is explained in Section III. Section IV represents the new cost function and its minimization procedure. Experimental results are demonstrated in Section V. Finally, Section VI concludes the paper with some remarks for further study in this domain.

II. Theory and Background

A. Delta Technique for PI Construction

The delta technique is based on representation and interpretation of NNs as nonlinear regression models. This allows applying standard asymptotic theory to them for constructing PIs. According to this, one may represent them as follows:

$$y_i = \Psi(X_i, \Theta^*) + \epsilon_i, \quad i = 1, 2, \ldots, n. \tag{1}$$

X_i and y_i are, respectively, the ith set of inputs (m independent variables) and the corresponding target (dependent variable). $\Psi(\cdot)$ with Θ^* is the nonlinear function representing the true regression function. n is also the number of observations. $\hat{\Theta}$, an estimate of Θ^*, can be obtained through minimization of sum of squared error (SSE) cost function

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

where $\hat{y}_i = \Psi(X_i, \hat{\Theta})$. A first-order Taylor's expansion of $\hat{y}_i$ around the true values of model parameters (Θ^*) can be expressed as

$$\hat{y}_i = \Psi(X_i, \Theta^*) + \nabla^T_{\Theta^*} \hat{y}_i (\hat{\Theta} - \Theta^*) \tag{3}$$

where $\nabla^T_{\Theta^*}$ is gradient of $\Psi(\cdot)$ (here NN models) with respects to its parameters, $\hat{\Theta}$, calculated for Θ^*. With the assumption that ϵ_i in (1) are independently and normally distributed $(N(0, \sigma^2)$, the $(1-\alpha)\%$ PI for $\hat{y}_i$ is

$$\hat{y}_i \pm t_{df}^{1-\frac{\alpha}{2}} s \sqrt{1 + \nabla^T_{\Theta^*} \hat{y}_i (J^T J)^{-1} \nabla_{\Theta^*} \hat{y}_i}. \tag{4}$$

$t_{df}^{1-(\alpha)/(2)}$ is the $\alpha/2$ quantile of a cumulative t-distribution function with df degrees of freedom. df here is the difference between number of training samples (n) and number of NN parameters (p). J is also the Jacobian matrix of the NN model with respect to its parameters.

The cost function defined in (2) is only related to the prediction errors and does not put any penalties on the network size or constrain the parameter magnitudes. This may result in singularity of matrix J^TJ, that in turn makes computed PIs less reliable. Inclusion of some weight decay terms in (2) can potentially solve this problem. The new cost function therefore will be (weight decay cost function)

$$\text{WDCF} = (Y - \hat{Y})^T(Y - \hat{Y}) + \lambda\hat{\Theta}^T\hat{\Theta} \tag{5}$$

where λ is the regularizing factor [7]. Adjusting NN parameters through minimization of this cost function often improves the NN generalization. Rebuilding PIs based on (5) will yield the following PIs:

$$\hat{y}_i \pm t_{df}^{1-\frac{\alpha}{2}} s\sqrt{1 + \nabla_{\Theta^*}^T\hat{y}_i \quad \Omega\, \nabla_{\Theta^*}\hat{y}_i} \tag{6}$$

where $\Omega = (J^TJ + \lambda I)^{-1}(J^TJ)(J^TJ + \lambda I)^{-1}$. Calculation of s in (6) is as follows:

$$s = \frac{\sqrt{\text{SSE}}}{n - \text{trac}(2\Gamma - \Gamma^2)} \tag{7}$$

where $\Gamma = J^T(J^TJ + \lambda I)^{-1}J$.

B. Simulated Annealing

SA is a gradient-free optimization technique first introduced in [35]. SA is based on the annealing of metals. If a metal is cooled slowly, its molecules enter a crystal structure. This crystal structure represents the minimum energy state. Essentially, SA is a Monte Carlo technique that can be used for seeking out the global minimum. The effectiveness of SA is attributed to the nature that it can explore the design space by means of neighborhood structure and escape from local minima by probabilistically allowing uphill moves. The primary virtues of the SA method for optimization are as follows: first, since no derivative information is needed during the search, SA performs well in conjunction with nondifferentiable cost functions, and secondly, SA is stochastic, thus it has better chances to explore the entire design space and reach the global optimum.

SA system is initialized at a temperature T_0 with a configuration (x_{old}) whose energy is evaluated to be E_{old}. A new configuration (x_{new} with new energy level E_{new}) is constructed by applying a random change. Decision about acceptance or rejection of the new configuration is made based on the difference in energy level $(\Delta E = E_{\text{new}} - E_{\text{old}} \leq 0)$. The new configuration is unconditionally accepted if it lowers the energy of the system $(\Delta E \leq 0)$. If the energy of the system is increased by the change, the new configuration is accepted with some random probability, $P = e^{(-(E_{\text{new}} - E_{\text{old}}))/(\kappa T)}$, where κ is the Boltzmann factor. If $P \geq r$, where r is a random number between 0 and 1, the new configuration is approved. This process is repeated sufficient times at the current temperature to sample the search space, and then the temperature is decreased based on a cooling schedule. This procedure continues until one of the stopping criterion is met.

Examples of cooling schedules are geometric and exponential. Generally, the higher the temperature, the more likely the acceptance of an uphill transition. This means that in early stages of optimization, SA behaves like a random walk. Mathematically, T_0 should be chosen so that $\forall(x_{\text{old}}, x_{\text{new}}),\ e^{(-(E_{\text{new}} - E_{\text{old}}))/(\kappa T_0)} \simeq 1$. As T decreases, SA becomes a greedy optimization search looking for global optimum. When $T = 0$, SA becomes totally greedy and only accepts good changes. Further information about SA and its fundamental theories can be found in [35] and [36].

III. Quantitative Measures for PI Assessment

As discussed before, literature does not offer a suitable measure for comprehensive assessment of PIs. In this section a new general examination measure is proposed that covers both important aspects of PIs: length and coverage probability. As the proposed measure is general and developed based on features of PIs (not the utilized method for constructing PIs), it can be applied in other relevant studies as well.

Theoretically, one can characterize PIs based on their length and coverage probability. One approach for quantitative assessment of PI lengths is to normalize each interval length with regard to range of targets. Following this, a measure called normalized mean prediction interval length (NMPIL) can be obtained as follows:

$$\text{NMPIL} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{U(X_i) - L(X_i)}{\xi}\right). \tag{8}$$

Normalization of PI length by the range of targets makes the objective comparison of PIs possible, regardless of techniques used for their construction or magnitudes of the underlying targets. The upper bound of NMPIL is one, obtained for the case that minimum and maximum of targets are considered as upper and lower bounds of PIs for all targets. Usually, the smaller the NMPIL, the more useful the PIs. The lower bound of NMPIL is model dependent and is dominated by mean squared error (MSE) of NN models. Assuming that in the ideal case, the gradient term in (4) and (6) vanishes for unobserved samples, one can obtain the lower bound of NMPIL for the delta technique as follows:

$$\text{NMPIL}_{\min} = \frac{2t_{df}^{1-\frac{\alpha}{2}} s}{\xi}. \tag{9}$$

Practically, achieving $\text{NMPIL}_{\min}$ for PIs is far remote. This stems from the fact that gradient terms in (4) and (6) are not ignorable. Indeed, they are often big for unobserved (test) samples, as these samples are not used in the training stage of NNs.

Empirically, it is desirable to have PIs such that their $\text{NMPIL}_{\min}$ is as small as possible. Although it is possible to

have a very small s for training samples (and by that minimizing $\mathrm{NMPIL}_{\min}$), this often leads to overfitting problems. While overfitting results in NNs with poor generalization (very high MSE for unobserved samples), it negatively contributes to coverage of PIs. While NMPIL relates to the length of PIs, another measure is required for monitoring coverage of PIs. If PIs are deliberately squeezed in favor of achieving smaller NMPIL, many targets may drop out of PIs. Therefore, another measure is required for quantification of this phenomenon. The PI coverage probability (PICP) indicates the probability that the underlying targets will lie within the constructed PIs. It can be calculated through counting the covered targets by PIs:

$$\mathrm{PICP} = \frac{1}{n}\sum_{i=1}^{n} c_i \tag{10}$$

where $c_i = 1$ if $y_i \in [L(X_i), U(X_i)]$; otherwise, $c_i = 0$.

Theoretically, PICP should be as close as possible to its nominal value, $(1-\alpha)\%$, the confidence level that PIs have been constructed based on. Unfortunately, in reality this often does not happen. Imperfectness of PICP is attributable to the presence of noise in samples and severe effects of uncertainty. Other issues such as under-fitting and over-fitting [which are direct results of using (very) small or big NNs] also contribute to the unsatisfactory smallness of PICP.

PIs whose PICP is the highest possible value are a matter of interest. Such high PICP can be simply achieved through considering target ranges as PIs for all samples. Needless to say, wide PIs like these ones are practically useless. This argument makes clear that judgment about PIs based on PICP without considering length of PIs (here, NMPIL) is always subjective and biased. It is essential to evaluate PIs simultaneously based on their both key measures: length (NMPIL) and coverage probability (PICP). Put in other words, these two measures should be read and interpreted in conjunction with each other.

Generally, PI lengths and PICP have a direct relationship. The wider the PIs, the higher the corresponding PICP. This means that as soon as PIs are squeezed, some targets will lie out of PIs, which results in a lower PICP. According to this discussion, the following coverage-length-based criterion (CLC) is proposed for comprehensive evaluation of PIs in term of their coverage probability and lengths

$$\mathrm{CLC} = \frac{\mathrm{NMPIL}}{\sigma(\mathrm{PICP}, \eta, \mu)} \tag{11}$$

where $\sigma(\cdot)$ is the sigmoidal function defined as follows:

$$\sigma(\mathrm{PICP}, \eta, \mu) = \frac{1}{1 + e^{-\eta(\mathrm{PICP}-\mu)}}. \tag{12}$$

η and μ are two controlling parameters determining how sharply and where the sigma function rises. The level of confidence that PIs have been constructed based on can be appropriately used as a guide for selecting hyperparameters of CLC. One reasonable principle is that we highly penalize PIs that their PICP is less than $(1-\alpha)\%$. This is based on the theory that the coverage probability of PIs in an infinite number of replicates will approach towards $(1-\alpha)\%$.

Generally, as η increases, the sigmoid function drops more sharply in higher values of PICP. The exact area of fall can be controlled by values of μ. The critical values of PICP are determined based on the confidence level of PIs, $(1-\alpha)\%$. For instance, if the confidence level is 90%, values of η and μ can be easily adjusted to guarantee sharp drop of the sigmoid function for $\mathrm{PICP} \leq 90\%$. Based on this, the CLC will highly increase, no matter what the length of PIs is. In this way, PIs with unsatisfactorily high coverage probability are heavily penalized. Generally, smallness of CLC is an indication of goodness of constructed PIs (simultaneously achieving small NMPIL and high PICP). Smallness or bigness of CLC is totally case-dependant. However, if PICP is sufficiently high, CLC and NMPIL will be almost the same.

IV. PI Optimization Procedure

As discussed in Section I, literature (with the exception of power engineering domain) is rich in applications of (4) and (6) for constructing PIs. Despite these reports, there are many issues left unarticulated in this domain. One issue, which is in fact the main motivation for conducting this research, is how PIs can be constructed to have the minimum length with the highest coverage probability. The motivating argument here is that PI construction in scientific literature has always been investigated from a point prediction perspective. As our focus here is on PIs, it is more reasonable to develop a cost function based on explanatory features of PIs (length and coverage probability). This new cost function then can be appropriately used for adjusting NN parameters. We believe that such attitude is one step forward in turning focus from point prediction to optimally constructed PIs.

The first problem in the definition of a new cost function is that the delta technique is based on minimization of the traditional cost functions defined in (2) and (5). All supporting theories of the delta technique are valid when NN parameters are adjusted based on these cost functions. For both of these cost functions, the designing principle is minimization of prediction error. To keep those theories valid, any effort for design of a new cost function needs to somehow cover the prediction error. With regards to this discussion and with the purpose of optimizing length and coverage probability of PIs, the following PI-error-based cost function (PICF) is introduced for training parameters of NNs

$$\mathrm{PICF} = \mathrm{CLC} + e^{(\mathrm{WDCF}_{\mathrm{opt}} - \mathrm{WDCF}_{\mathrm{trad}})}. \tag{13}$$

The first term in the right side of (13) has been defined in (11). It corresponds to the basic characteristics of PIs: NMPIL and PICP as defined in (8) and (10), respectively. The second term is an exponential term of the difference of the weight decay cost functions (5) calculated for two sets of NN parameters: Θ_{opt} obtained through minimization of (13) and Θ_{trad} obtained based on minimization of (5). The exponential terms in (13) converts small differences in WDCFs into big values [can be potentially much bigger than CLC in (13)]. Therefore, any action (here

any change in parameters of NNs) resulting in violation of the delta technique assumptions is highly penalized. This penalization guarantees that fundamental theories of the delta technique remain valid when training NNs through minimization of (13), instead of (2) and (5).

With the integration of the traditional WDCF into the new cost function (13), three purposes have been followed simultaneously: 1) keeping theoretical assumptions and conditions of the delta technique for constructing PIs valid, 2) guaranteeing better performance of the optimized NN in terms of WDCF, and 3) directly focusing on improving quality of PIs (rather than minimizing point prediction error). NN parameters, Θ, will be adjusted based on minimization of PICF

$$\Theta^* = \arg \min_{\Theta} \text{PICF}. \tag{14}$$

Minimization of PICF through mathematical analysis is not possible. This is mainly attributable to difficulties in calculation of characteristics of this function. PICF is not a differentiable function (due to presence of PICP). Furthermore, it is highly bumpy. Traditional optimization techniques, hence, are highly likely to be trapped in its local minima. Stochastic gradient-free techniques, such as SA introduced in Section II-B, are particularly useful in this regard.

Optimization is completed in two stages using two training sets: D_1 and D_2. First, NNs are trained using traditional techniques, such as Levenberg-Marquardt technique. The training objective is minimization of WDCF (according to the delta technique assumptions). Then the trained NN is used for calculation of MSE for D_1 and retrained based on minimization of PICF for D_2. After completion of this stage by SA (or any other stochastic optimization technique), PIs are constructed for the obtained NNs (traditional and new one). Performance of methods and quality of PIs can be easily assessed and compared through calculation of PICP, NMPIL, and CLC.

It does matter to emphasize differences between the proposed method and the traditional delta technique for construction of PIs. First of all, the traditional method is based on WDCF. WDCF directly deals with prediction error not measures related to PIs. In contrast, PICF well covers indexes related to quality of PIs: NMPIL and PICP. Therefore, its optimization directly improves quality of PIs. Secondly, PICF covers all aspects covered by WDCF. Its exponential term includes WDCF computed for the new and traditionally trained NNs. Finally, employment of two different datasets during training process significantly reduces chance of overfitting. If the NN is overfitted (trained by D_1), its PICF for D_2 will dramatically rise, and therefore that set of NN weights will be discarded automatically.

V. Experiments and Results

The proposed method for construction of PIs is here applied to the real electric load datasets introduced in [9] for STLF. The experiment consists of computing forecasts and constructing PIs of the hourly loads for two days ahead. The dataset includes records of consumed loads, weather condition, and calendar information for three years. D_1, D_2, and D_{Test} are generated through sampling from the main dataset. All variables are preprocessed to have zero mean and unit variance. This is done to avoid over-weighting some variables against others solely due to their magnitude.

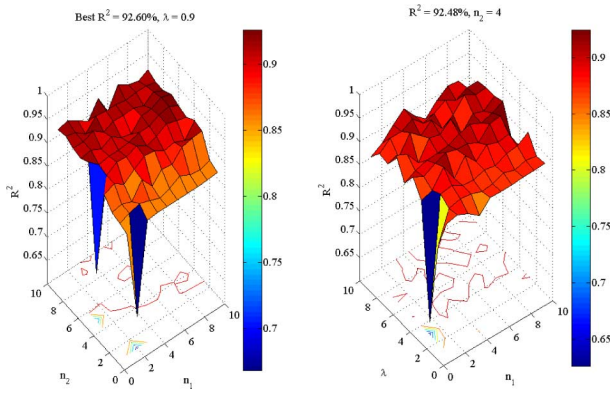

Fig. 1. Coefficient of determination for samples of D_{Test}. (Left) Grid composed on n_1 and n_2. (Right) Grid composed of n_1 and λ.

First, it is shown that NN performance for forecasting future loads is unsatisfactory, no matter what its size, structure, and parameters are. NNs are data driven techniques that their prediction performance highly depends on their initial parameters and structure. To avoid any subjective argument about their performance, two experiments are conducted. In both experiments, NNs have two layers with a variable number of neurons. In the first experiment, a grid of different structures is developed through changing number of neurons between one to ten in each layer (the regularizer factor, λ is fixed at 0.9). For each structure, NNs are initialized and trained five times using D_1. Then coefficient of determination (R^2) is calculated and averaged for samples of D_{Test}. In the second experiment, the regularizer factor and number of neurons in the first layer are changed within a grid (number of neurons in the second layer is set to 4). The same procedure explained for the previous experiment is applied here as well.

Fig. 1 shows averaged R^2 over the grids defined in two experiments. It has been plotted versus number of neurons in first layer (n_1), number of neurons in second layer n_2, and the regularizer factor λ. The best R^2 in first and second experiments are 92.60% and 92.48%, respectively. The imperfectness of these values indicates that NNs are not capable of yielding highly accurate forecasts of the future loads. This defect is not amendable through changing structure of NNs or retraining them. According to this, forecasts produced by NNs are unreliable and any decision made based on them may result in disastrous consequences.

According to the discussions made in Section II and IV, SA is chosen and adopted as the optimization engine in this work. The SA parameters used in the current study and some other quantities are summarized in Table I. All PIs are constructed with 90% confidence $(\alpha = 0.1)$. Values for η and μ have been chosen so that PIs with $\text{PICP} \leq 90\%$ are highly penalized. The cooling schedule is geometric with a factor set to 0.95. T_0 has been selected big enough to allow uphill transitions in the early iterations of the optimization procedure.

TABLE I
PARAMETERS USED IN EXPERIMENTS AND OPTIMIZATION PROCEDURE

α	0.1
η	200
μ	0.875
T_0	10
T_{Final}	10^{-2}
Geometric cooling schedule	$T_{k+1} = 0.95\ T_k$
D_1	40% of all samples
D_2	40% of all samples
D_{Test}	20% of all samples

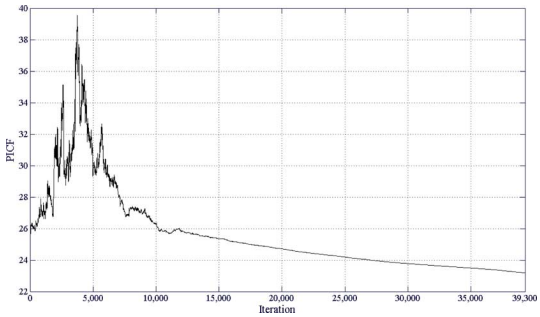

Fig. 2. Variation of PICF during the optimization process.

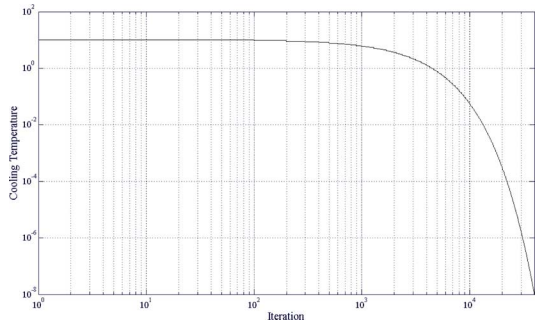

Fig. 3. Profile of cooling temperature.

In the experiments here, NNs have the following features: two-layered, $n_1 = 7$, $n_2 = 4$, and $\lambda = 0.9$. According to these results shown in Fig. 1, R^2 for an NN with these features is quite close to the best possible value of R^2. The procedure explained in Section IV is applied here to this NN for obtaining optimal PIs. In the experiments, it is demonstrated that NNs trained based on mechanism proposed in Section IV produce narrower PIs with higher coverage probability.

Fig. 2 represents variation of PICF for D_2 during the optimization process. This plot should be read in conjunction with the cooling temperature displayed in Fig. 3. In the early stages (up to 6000), T is high and allows in acceptance of inferior movements. Therefore, PICF widely fluctuates reflecting exploration of different corners of the weight space. Sharp jumps and falls of PICF reflect the bumpiness of the search space with respect to NN parameters. As temperature cools down, PICF takes a downward trend and gradually decreases. It finally settles at 23.13%, which is much below its original value taken from NNs trained by traditional cost functions.

After completion of the training stage, the optimized NN is utilized for constructing PIs for samples of D_{Test}. With the purpose of comparison, the NN trained using the traditional WDCF

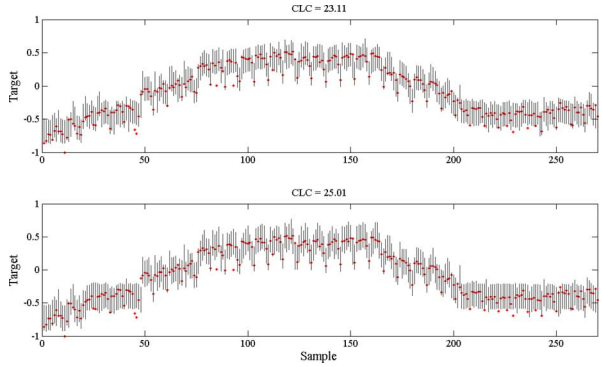

Fig. 4. PIs for test samples constructed using (top) NN_{opt} and (bottom) NN_{trad}.

TABLE II
COMPARING PERFORMANCE OF NN_{opt} AND NN_{trad} FOR CONSTRUCTING PIS FOR SAMPLES OF D_{Test}

	NN_{Opt}	NN_{trad}
$PICP$	94.12%	92.65%
$NMPIL$	23.13	25.01
CLC	23.13	25.01
$Improvement$	7.52%	-

is also employed here for developing PIs. Fig. 4 represents PIs for samples of D_{Test} built using NN_{opt} and NN_{trad}. Hereafter, we refer to these PIs as PI_{opt} and PI_{trad}. It is obvious that PI_{opt} are narrower with a higher coverage probability. A more quantitative comparison of performance of two methods has been given in Table II. CLC of PI_{opt} is 23.13 that is smaller than CLC of PI_{trad}. The obtained results indicate a 7.52% reduction in the length of PIs without compromising their coverage probability. According to Table II, NN_{opt} always outperforms NN_{trad} in terms of any measure related to PIs (PICP, NMPIL, and CLC). These results explicitly imply that squeeze of PIs through the proposed method is effective and it does enhance quality of constructed PIs.

It is very important to evaluate PIs based on their two key features: coverage probability (measured by PICP), and length (measured by NMPIL). As PICP of PIs reported in Table II is above the nominal confidence level (90%), there is no concern about coverage of PIs. In term of length, the minimum value for NMPIL [$\text{NMPIL}_{\text{min}}$ calculated using (9)] is equal to 22.53%. Again NMPIL shown in Table II does not much differ with this. The obtained results, hence, are satisfactory and acceptable.

It is also interesting to examine the effects of level of confidence on length of PIs. In a new experiment, the level of confidence is changed between 1% to 99%. PIs are constructed for four specific samples taken randomly from D_{Test}. Fig. 5 displays evolution of upper and lower bounds of PIs due to change in level of confidence $((1-\alpha)\%)$. Actual and predicted target have been also plotted in this figure. It is easy to conclude that PIs expand as confidence level rises. PIs constructed using the delta technique are symmetric. Predicted loads, therefore, are always in the middle of PIs. If predicted values are accurate, PIs with low confidence level well cover the actual values. Top-left plot in Fig. 5 corresponds to this case. All PIs with confidence level bigger than 17% cover the target. The

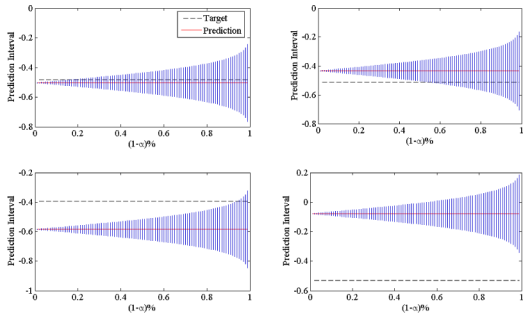

Fig. 5. Effects of level of confidence on length of PIs for four case studies.

problem appears when predictions are not accurate. In reality prediction errors are often (very) big (for instance, in three out of four cases predictions and target highly differ). Therefore, it is required to increase the confidence level of PIs. Top-right and down-right plots correspond to this case. The minimum level of confidence for covering targets are 57% and 88%, respectively. The down-right plot represents a case that point prediction error has been quite high. The error is so high that its effects are compensable through increasing level of confidence. Therefore, the actual value always remains beyond the bounds of PIs.

PIs shown in Fig. 5 can be interpreted and exploited in different ways. Traditionally, electricity generating schedules are developed based on the point prediction values (center of each PI). These predictions are either over-estimated (Fig. 5, down-left) or under-estimated (Fig. 5, top-left/right and down-right). Regardless of their accuracy, they also lack an indication of their accuracy. PIs, in contrast, not only have an indication of the accuracy (confidence level), but also provide more information for schedulers. It is totally up to the modeler to decide which value best satisfies the scheduling conditions, constraints, supply policies, and system reliability. Before going further, one should make note of the fact that all values within a PI may be realized in future with a high probability. If PIs are wide, it reflects presence of a high level of uncertainty in data. Therefore, making decisions solely based on point prediction values generated by NNs (or any other type of predictor) is potentially risky. For conservative planning, the upper bounds of PIs can be used for developing electricity generation schedules. With a high level of confidence attached to PIs, reliability of PIs is guaranteed. Selecting the lower bounds of PIs reflects an optimistic attitude in scheduling with more attention to over-supply avoidance.

VI. Conclusion

This paper aimed to investigate the short load forecasting problem from a prediction interval perspective. Instead of developing neural network models for predicting exact load values, prediction intervals were developed based on the delta technique. A new measure was proposed for quantitative assessment of prediction intervals. Length and coverage probability constitute the core of the proposed measure. According to this new measure, a new prediction interval-based objective function was designed for training neural networks. Ample care was taken at the design stage to guarantee validity of fundamental assumptions of the delta technique. As this new cost function is highly nonlinear and bumpy, simulated annealing was employed for its minimization. In experiments with real electric load data, it was demonstrated that the quality of prediction intervals developed for neural networks trained using the proposed method is significantly better than prediction intervals developed using the traditional delta technique. The main improvement is for reducing length of prediction intervals without compromising their coverage probability.

The obtained results in this paper can be interpreted, utilized, and extended in a variety of different manners. As the proposed method is generic, it can be applied to any prediction and forecasting problem in the power engineering domain. The most probable ones are midterm/long-term load and electricity price forecasting. Upper and lower bounds of prediction intervals can be appropriately used for electricity generation scheduling. Upon their availability, conservative and optimistic operational plans can be developed and applied for electricity generation. Such planning gives electricity suppliers more confidence in terms of reliability and profitability. An interesting point of the proposed method is that its computation mass after completion of optimization stage is nothing. This makes its application for real time planning possible.

References

[1] D. W. Bunn and E. D. Farmer, *Comparative Models for Electrical Load Forecasting.*. Chichester, U.K.: Wiley, 1985.

[2] A. Papalexopoulos and T. Hesterberg, "A regression-based approach to short-term system load forecasting," *IEEE Trans. Power Syst.*, vol. 5, no. 4, pp. 1535–1547, Nov. 1990.

[3] H. M. Al-Hamadi and S. A. Soliman, "Short-term electric load forecasting based on Kalman filtering algorithm with moving window weather and load model," *Elect. Power Syst. Res.*, vol. 68, no. 1, pp. 47–59, Jan. 2004.

[4] S.-J. Huang and K.-R. Shih, "Short-term load forecasting via ARMA model identification including non-Gaussian process considerations," *IEEE Trans. Power Syst.*, vol. 18, no. 2, pp. 673–679, May 2003.

[5] J. D. Cryer, *Time Series Analysis.*. Pacific Grove, CA: Duxbury Press, 1986.

[6] V. Ferreira and A. Alves da Silva, "Toward estimating autonomous neural network-based electric load forecasters," *IEEE Trans. Power Syst.*, vol. 22, no. 4, pp. 1554–1562, Nov. 2007.

[7] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford Univ. Press, 1995.

[8] K.-H. Kim, J.-K. Park, K.-J. Hwang, and S.-H. Kim, "Implementation of hybrid short-term load forecasting system using artificial neural networks and fuzzy expert systems," *IEEE Trans. Power Syst.*, vol. 10, no. 3, pp. 1534–1539, Aug. 1995.

[9] R. Barzamini, M. Menhaj, A. Khosravi, and S. Kamalvand, "Short term load forecasting for iran national power system and its regions using multi layer perceptron and fuzzy inference systems," in *Proc. IEEE Int. Joint Conf. Neural Networks (IJCNN'05)*, 2005, vol. 4, pp. 2619–2624.

[10] G.-C. Liao and T.-P. Tsao, "Application of a fuzzy neural network combined with a chaos genetic algorithm and simulated annealing to short-term load forecasting," *IEEE Trans. Evol. Comput.*, vol. 10, no. 3, pp. 330–340, Jun. 2006.

[11] K. Metaxiotis, A. Kagiannas, D. Askounis, and J. Psarras, "Artificial intelligence in short term electric load forecasting: A state-of-the-art survey for the researcher," *Energy Convers. Manage.*, vol. 44, no. 9, pp. 1525–1534, Jun. 2003.

[12] H. Hippert, C. Pedreira, and R. Souza, "Neural networks for short-term load forecasting: A review and evaluation," *IEEE Trans. Power Syst.*, vol. 16, no. 1, pp. 44–55, Feb. 2001.

[13] M. Tamimi and R. Egbert, "Short term electric load forecasting via fuzzy neural collaboration," *Elect. Power Syst. Res.*, vol. 56, no. 3, pp. 243–248, Dec. 2000.

[14] L.-C. Ying and M.-C. Pan, "Using adaptive network based fuzzy inference system to forecast regional electricity loads," *Energy Convers. Manage.*, vol. 49, no. 2, pp. 205–211, Feb. 2008.

[15] H. Hippert, D. Bunn, and R. Souza, "Large neural networks for electricity load forecasting: Are they overfitted?," *Int. J. Forecast.*, vol. 21, no. 3, pp. 425–434, Jul. 2005.

[16] A. Khotanzad, R. Afkhami-Rohani, and D. Maratukulam, "ANNSTLF-artificial neural network short-term load forecaster generation three," *IEEE Trans. Power Syst.*, vol. 13, no. 4, pp. 1413–1422, Nov. 1998.
[17] M. Paliwal and U. A. Kumar, "Neural networks and statistical techniques: A review of applications," *Expert Syst. Appl.*, vol. 36, no. 1, pp. 2–17, Jan. 2009.
[18] T.-Y. Kwok and D.-Y. Yeung, "Constructive algorithms for structure learning in feedforward neural networks for regression problems," *IEEE Trans. Neural Netw.*, vol. 8, no. 3, pp. 630–645, May 1997.
[19] R. A. Kilmer, A. E. Smith, and L. J. Shuman, "Computing confidence intervals for stochastic simulation using neural network metamodels," *Comp. Ind. Eng.*, vol. 36, no. 2, pp. 391–407, Apr. 1999.
[20] J. T. G. Hwang and A. A. Ding, "Prediction intervals for artificial neural networks," *J. Amer. Statist. Assoc.*, vol. 92, no. 438, pp. 748–757, Jun. 1997.
[21] R. D. d. Veaux, J. Schumi, J. Schweinsberg, and L. H. Ungar, "Prediction intervals for neural networks via nonlinear regression," *Technometrics*, vol. 40, no. 4, pp. 273–282, 1998.
[22] D. J. C. MacKay, "The evidence framework applied to classification networks," *Neural Comput.*, vol. 4, no. 5, pp. 720–736, 1992.
[23] B. Efron, "Bootstrap methods: Another look at the jackknife," *Ann. Statist.*, vol. 7, no. 1, pp. 1–26, 1979.
[24] T. Heskes, , T. P. M. Mozer and M. Jordan, Eds., "Practical confidence and prediction intervals," in *Neural Information Processing Systems*. Cambridge, MA: MIT Press, 1997, vol. 9, pp. 176–182.
[25] D. Nix and A. Weigend, "Estimating the mean and variance of the target probability distribution," in *Proc. 1994 IEEE Int. Conf. Neural Networks*, 1994, vol. 1, pp. 55–60.
[26] T. Lu and M. Viljanen, "Prediction of indoor temperature and relative humidity using neural network models: Model comparison," *Neural Comput. Appl.*, vol. 18, no. 4, pp. 345–357, May 2009.
[27] A. Khosravi, S. Nahavandi, and D. Creighton, "A prediction interval-based approach to determine optimal structures of neural network metamodels," *Expert Syst. Appl.*, vol. 37, pp. 2377–2387, 2010.
[28] A. Khosravi, S. Nahavandi, and D. Creighton, "Constructing prediction intervals for neural network metamodels of complex systems," in *Proc. Int. Joint Conf. Neural Networks*, 2009, pp. 1576–1582.
[29] Y. Jia and T. B. Culver, "Bootstrapped artificial neural networks for synthetic flow generation with a small data sample," *J. Hydrol.*, vol. 331, no. 3–4, pp. 580–590, Dec. 2006.
[30] S. Ho, M. Xie, L. Tang, K. Xu, and T. Goh, "Neural network modeling with confidence bounds: A case study on the solder paste deposition process," *IEEE Trans. Electron. Packag. Manuf.*, vol. 24, no. 4, pp. 323–332, Oct. 2001.
[31] A. M. Alonso and A. E. Sipols, "A time series bootstrap procedure for interpolation intervals," *Comput. Statist. Data Anal.*, vol. 52, no. 4, pp. 1792–1805, Jan. 2008.
[32] G. Papadopoulos, P. Edwards, and A. Murray, "Confidence estimation methods for neural networks: A practical comparison," *IEEE Trans. Neural Netw.*, vol. 12, no. 6, pp. 1278–1287, Nov. 2001.
[33] G. Yu, H. Qiu, D. Djurdjanovic, and J. Lee, "Feature signature prediction of a boring process using neural network modeling with confidence bounds," *Int. J. Adv. Manuf. Technol.*, vol. 30, no. 7, pp. 614–621, Oct. 2006.
[34] E. Zio, "A study of the bootstrap method for estimating the accuracy of artificial neural networks in predicting nuclear transient processes," *IEEE Trans. Nucl. Sci.*, vol. 53, no. 3, pp. 1460–1478, Jun. 2006.
[35] G. C. V. M. Kirkpatrick and S. , "Optimization by simulated annealing," *Science*, vol. 220, pp. 671–680, 1983.
[36] E. H. L. Laarhoven and P. J. M. van Aarts, *Simulated Annealing: Theory and Applications*. Norwell, MA: Kluwer, 1987.

Abbas Khosravi (M'07) received the B.Sc. degree in electrical engineering from Sharif University of Technology, Tehran, Iran, and the M.Sc. degree (hons.) in electrical engineering from Amirkabir University of Technology, Tehran, in 2002 and 2005, respectively. Currently, he is pursuing the Ph.D. degree in the Centre for Intelligent Systems Research (CSIR) at Deakin University, Geelong, Australia.

His major in those careers was control with specialization on artificial intelligence (AI). In 2006, he joined eXiT Group in the University of Girona, Girona, Spain, as a research academic in the area of AI applications. His primary research interests include development and application of AI techniques for (meta)modeling, analysis, and optimization of complex systems. Also he does research in the area of discrete event simulation of complex systems including manufacturing enterprisers.

Saeid Nahavandi (SM'07) received the B.Sc. (hons.), M.Sc., and Ph.D. degrees in automation and control from Durham University, Durham, U.K.

He is the Alfred Deakin Professor, Chair of Engineering, and the leader for the Intelligent Systems Research Centre, Deakin University, Geelong, Australia. He has published over 300 peer reviewed papers in various international journals and conferences. He designed the world's first 3-D interactive surface/motion controller. His research interests include modeling of complex systems, simulation-based optimization, robotics, haptics, and augmented reality.

Dr. Nahavandi was a recipient of the Young Engineer of the Year title in 1996 and six international awards in engineering. He is the Associate Editor of the IEEE SYSTEMS JOURNAL, an Editorial Consultant Board member for the *International Journal of Advanced Robotic Systems*, and an Editor (South Pacific Region) of the *International Journal of Intelligent Automation and Soft Computing*. He is a Fellow of Engineers Australia (FIEAust) and IET (FIET).

Doug Creighton (M'10) is a research academic and group coordinator in the Centre for Intelligent Systems Research (CSIR) at Deakin University, Geelong, Australia. His main research interests include discrete event simulation, simulation-based optimization, advanced OR techniques, distributed control, agent-based computing, and visualization interfaces.