Learning Quantile Function without Quantile Crossing for Distribution-free Time Series Forecasting

Youngsuk Park [1] **Danielle Robinson** [1] **Yuyang Wang** [1] **Jan Gasthaus** [1]

Abstract

Quantile regression is an effective technique to quantify uncertainty, and fit challenging underlying distributions. Generating full probabilistic predictions requires multiple quantile regressions over multiple quantile levels. As a result, quantile crossing is a common drawback to these approaches since it violates the desirable monotone property of the conditional quantile function. In this work, we propose the incremental quantile function (IQF), a general distribution-free quantile estimation framework with a simple neural network layer that resolves the issue of quantile crossing. Moreover, IQF is able to extrapolate to other quantile levels, resulting in predictions on arbitrary quantile levels that differ from the underlying training quantiles. We apply IQF to the NN-based times series forecasting use case, where this is particularly appealing because it saves the expensive re-training cost for non-trained quantile levels. We provide generalization error analysis with our proposed approach under the sequence-to-sequence forecasting setting. Experimental results demonstrate the effectiveness and resulting accuracy improvement of our method.

1. Introduction

Probabilistic time series forecasting methods are increasingly replacing point prediction techniques in practical applications, as it is crucial for downstream decision-making processes to take the uncertainties in predictions into consideration. A popular approach to probabilistic forecasting is to have a sequential model with a likelihood model that determines how to "emit" from the hidden or latent states to the observations. Examples include State-Space Models (SSMs), Gaussian processes (GPs), DeepAR (Salinas et al., 2020), etc. While full probabilistic predictions can be obtained from these models, a practical conundrum of which likelihood function to choose from arises. As an example, DeepAR, a RNN-based probabilistic forecaster, offers likelihood choices of normal, student-t, negative binomial, etc. It is desirable in both theory and practice to have an option that requires no assumption of the data generating process.

Fortunately, quantile regression (Koenker & Bassett Jr, 1978; Koenker, 2005), which has been successfully used for robustly modeling probabilistic outputs, comes to the rescue. The incorporation of the quantile regression component to various sequential neural network backbones has been shown to be particularly effective with recent advances in deep learning (Wen et al., 2017; Gasthaus et al., 2019; Lim et al., 2019; Eisenach et al., 2020). To obtain a full probabilistic prediction (i.e., one can query a forecast at any arbitrary quantile), usually requires generating multiple quantiles at once. However, when modeling multiple quantiles simultaneously, *quantile crossing*, i.e., the failure of the estimated conditional quantile function to obey the required monotonicity constraint, is a commonly noted concern beyond mere theoretical consistency. This phenomena particularly stands out in the case of sequence-to-sequence predictions, where inconsistency can propagate over the sequence of predictions. Various strategies (reviewed in Section 1.1) have been proposed to remedy quantile crossing in the general regression setting. Due to their perceived complexity and/or lack of solid principles, these techniques are not widely adopted in combination with recent deep probabilistic forecasting modelings. In particular, Wen et al. (2017); Lim et al. (2019) do not explicitly address quantile crossing.

Another issue of typical quantile regression approaches in the time series context is that they require the set of predicted quantile levels to be fixed for the training ahead of time. This restriction may require the expense of re-training the models when making a prediction at arbitrary quantile levels. This is a particular concern with the computational overhead for the class of heavy deep forecasting models with numerous training panels of time series.

In this paper, we propose a simple methodology, incremental quantile function (IQF), that quantifies prediction uncertainty in a distribution-free manner. IQF models the condi-

[*]Equal contribution [1]AWS AI Labs. Correspondence to: Youngsuk Park <pyoungsu@amazon.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

tional quantile function jointly for multiple quantiles in addition to resolving the quantile crossing and re-training issues. IQF consists of a family of conditional quantile functions that linearly interpolate between the given quantile levels and extrapolate in the parametric form beyond the extremal training quantiles using exponential tails. Our methodology is suitable for but not limited to sequence-to-sequence time series forecasting where both of aforementioned issues are raised mostly frequently. We analyze the generalization errors of IQF under multi-horizon and multi-quantile time series forecasting, characterized with important quantities, e.g. quantization error, Rademachar complexity, and discrepency measure. Through the experiments, we demonstrate the consistency and accuracy improvement of our methodology under IQF compared with the state-of-the-art sequence-to-sequence MQ-CNN (Wen et al., 2017) model.

1.1. Related Works

Sequence-to-sequence (seq2seq) Forecast Sequence-to-sequence (seq2seq) models directly map input sequences to output sequences of possibly different lengths. In the context of forecasting, Wen et al. (2017) proposes a seq2seq architecture for probabilistic forecasting. Their model uses an RNN as the encoder and an MLP as the decoder (MQ-RNN). The MLP is constructed to output a set of quantile levels for the entire forecast horizon. The advantage of this architecture is that the simultaneous output of all quantile levels for the entire forecast horizon is insensitive to error accumulation. However, the model needs to be retrained if other quantile levels than the trained ones are required by the user. The authors show that dilated causal convolutions can also be used as the encoder (MQ-CNN) (Wen et al., 2017). Wen & Torkkola (2019) extend the MQ-CNN model with a generative quantile copula. This model learns the conditional quantile function that maps the quantile index, which is a uniform random variable conditioned on the covariates, to the target. During training, the model draws the quantile index from a uniform distribution. This turns MQ-CNN into a generative, marginal quantile model. The authors combine this approach with a Gaussian copula to draw correlated marginal quantile index random values. They show that the Gaussian copula component improves the forecast at the distribution tails. Chen et al. (2020) proposes DeepTCN, another seq2seq model where the encoder is a dilated causal convolution with residual blocks, and the decoder is simply a MLP with residual connections. Structure-wise, DeepTCN is almost the same as the basic structure of MQ-CNN (Wen et al., 2017), i.e., without the local MLP component that aims to model spikes and events.

Quantile Regression There are various approaches to resolve the issue of quantile crossing that occurs when quantile estimations over multiple quantiles are learned jointly. Most of these works are heuristic as a sorting-based post-processing at the end (Kim et al., 2021) or expensive constrained optimization (Liu & Wu, 2009). Schmidt & Zhu (2016) proposed learning on the non-negative increment between quantile estimates on pre-determined quantile levels, and then stacking them. In addition, some linear/non-linear interpolation between quantile estimates allows for a form of quantile function in some range of quantile levels. SQF (Gasthaus et al., 2019) is another method that gives a functional form, and results in no quantile crossing. These works provide the functional form on the quantile levels trained on, but have limitations on the tails outside of these quantile levels.

Theoretical Analysis on Forecasting There are several recent theoretical analyses of time series forecasting with the tools from learning theory. Kuznetsov & Mohri (2015) investigates the theoretical analysis on the general scenario of non-stationary non-mixing stochastic processes in terms of a data-dependent measure of sequential complexity and a discrepancy measure. Zimin & Lampert (2017) studies the learnability of stochastic processes with respect to the conditional risk, focusing on analyzing scenarios, where the pairwise discrepancy is controllable. Mariet & Kuznetsov (2019) examines theoretical studies on general multivariate sequence-to-sequence setting in terms of discrepancy measure and mixing coefficient of the underlying stochastic process. Albeit the powerful intuitions and mathematical machinery provided, these works focus on general scenarios often only with auto-regressive (AR) models as examples, lacking more fruitful intuition our use case of multi-horizon and multi-quantile deep time series forecasting.

2. Background

For a random variable $Z \in \mathbb{R}$, we denote $F_Z(z)$ as its cumulative distribution function (CDF). Then, the α-quantile of Z is given as:

$$q_Z(\alpha) := F_Z^{-1}(\alpha) = \inf\{z \in \mathbb{R} : \alpha \leq F_Z(z)\},$$

where $\alpha \in (0, 1)$ denotes a quantile level. The quantile function $q(\cdot)$ is also called the percent-point function or inverse cumulative distribution function.

2.1. Quantile Regression

Let $(X, Z) \sim F_{(X,Z)}$ for a regression setting. Quantile regression seeks to estimate α-quantile of $Z \in \mathbb{R}$ conditioned on $X = x$ for some $\alpha \in [0, 1]$, i.e., $q(\alpha \mid x) = F^{-1}_{Z|X=x}(\alpha)$. Equivalently, the α-quantile is the solution of minimizing the expected quantile loss (Koenker & Bassett Jr, 1978):

$$q(\alpha \mid x) = \operatorname*{argmin}_{q \in \mathbb{R}} \mathbb{E}_{Z|X=x}(\rho_\alpha(Z - q)), \qquad (1)$$

where $\rho_\alpha(u) = u \times (\alpha - \mathbf{1}\{u < 0\})$ with $u = z - q$ denotes the quantile loss of $q \in \mathbb{R}$ w.r.t. any $z \in \mathbb{R}$.

Continuous Ranked Probability Score (CRPS). Rather than optimizing the single quantile loss, the CRPS is a *proper scoring rule*[1] that averages the quantile loss over all quantiles. Formally, given a fixed target $z \in \mathbb{R}$ and a family of quantile fuctions $\mathcal{Q}$, the CRPS $L : \mathcal{Q}, \mathbb{R} \to \mathbb{R}$ is given as:

$$L(q, z) = \int_{\alpha=0}^{1} 2\rho_\alpha(q(\alpha) - z)d\alpha. \tag{2}$$

Similarly, an approximated CRPS $\hat{L}$ is given as:

$$\hat{L}(q, z) = \sum_{\alpha} w(\alpha) 2\rho_\alpha(q(\alpha) - z)d\alpha, \tag{3}$$

where $w(\alpha)$ denotes a weight at quantile level α for the integral approximation.

2.2. Quantile Regression in Time Series Forecasting

Quantile regression settings can be extended to the common case of multiple horizon and multiple quantile estimations in time series forecasting as follows. Suppose we have m related time series data, each of which consists of observation $z_{i,t} \in \mathbb{R}$ with (optional) input covariates $\mathbf{x}_{i,t} \in \mathbb{R}^d$ at time t. In i-th time series forecasting, given T past target observations $z_{i,1:T}$ and all *future* covariates $\mathbf{x}_{i,1:T+\tau}$, we wish to make τ future quantile predictions at time T:

$$\{\hat{z}^\alpha_{i,T+1}, \dots, \hat{z}^\alpha_{i,T+\tau}\}_\alpha = F_\theta(z_{i,1:T}, \mathbf{x}_{i,1:T+\tau}), \tag{4}$$

where F_θ is a global[2] sequence-to-sequence quantile prediction. In short, we can express the α-quantile estimate at each time $t = T+1, \dots, T+\tau$ as:

$$\hat{z}^\alpha_{i,t} = q_\theta(\alpha \mid y^t_i) \tag{5}$$

where $y^t_i = (z_{i,1:T}, \mathbf{x}_{i,1:t})$ is *input* to the quantile function.

Finally, we formulate the Empirical Risk Minimization (ERM) with $\hat{L}$ in (3) as

$$\underset{\theta}{\text{minimize}} \frac{1}{m\tau} \sum_{i=1}^{m} \sum_{t=T+1}^{T+\tau} \hat{L}(q(\cdot \mid y^t_i), z_{i,t}), \tag{6}$$

where $\hat{L}(q(\cdot \mid y^t_i), z_{i,t}) = \sum_\alpha w(\alpha)\rho_\alpha(z_{i,t} - q(\alpha \mid y^t_i))$.

2.3. Estimation Inconsistency: Quantile Crossing

When tackling the problem of quantile learning naively, inconsistencies that violate the non-decreasing property of quantile estimate on α, called *quantile crossing* may arise. In other words, for $\alpha_1 > \alpha_2$, there exists x s.t.

$$q(\alpha_1 \mid x) < q(\alpha_2 \mid x),$$

where $q(\cdot \mid x)$ is the quantile estimate of the target conditioned on input x.

The quantile crossing issue occurs in many cases when quantile regression is applied separately over multiple quantiles $\{\alpha\}$ or with CRPS under naive parameterizations on quantile functions. Forecasting problems in the multi-horizon multi-quantile cases in (6) often report more severe quantile crossing, as illustrated in Figure 3a.

3. Incremental Quantile Function (IQF)

Most forecast models output quantile estimates on a fixed finite set $\{\alpha_k\}_{k=1}^K$, called *quantile knots*. In this section, we propose the Incremental Quantile Function (IQF), a family of conditional quantile functions. IQF estimates targets on any quantile $\alpha \in (0, 1)$, and overcomes quantile crossing. We first learn some basis quantile estimates on quantile knots, and then apply an interpolation and extrapolation strategy around the knots to make it accessible on any query quantile without the need to re-train.

For notational simplicity, we omit index i, t. Let $h \in \mathcal{H}$ be the last hidden variable $h = \mathbf{h}(y^t)$ in the model at each prediction time. With a slight notational override, we denote the last output layer as q_ϕ with parameter ϕ into which the last hidden variable h is fed into, i.e.,

$$q_\phi(\alpha \mid h) := q_\theta(\alpha \mid y).$$

3.1. Basis Estimates on Quantile Knots

On the quantile knots $\{\alpha_k\}_{k=1}^K$ with $\alpha_k < \alpha_{k+1}$, IQF first gives basis quantile estimates by cumulating non-negative amounts as the quantile level α_k increases:

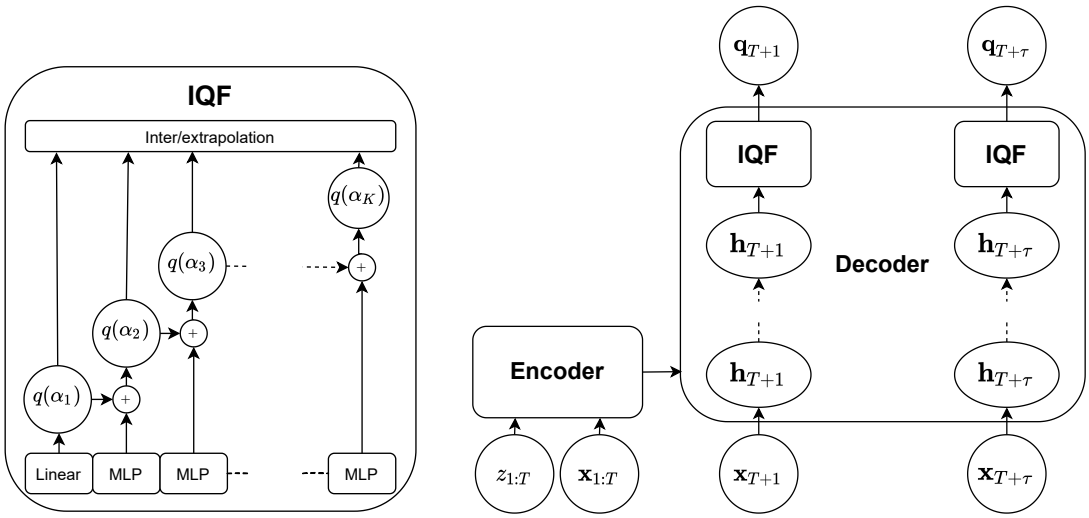
$$q_\phi(\alpha_k \mid h) = \begin{cases} \pi^{\text{Linear}}(h; \phi_1), & k = 1, \\ q_\phi(\alpha_{k-1} \mid h) + \pi^{\text{MLP}}(h; \phi_k), & k = 2, \dots, K, \end{cases} \tag{7}$$

where $\pi^{\text{Linear}} : \mathcal{H} \to \mathbb{R}$ is a linear function, $\pi^{\text{MLP}} : \mathcal{H} \to \mathbb{R}_+$ is a multilayer perceptron with a non-negative non-linear component, e.g., ReLu, Sigmoid.

The basis estimates on the knots are derived by solving the ERM in (6) under the training data. Note that the design of the IQF architecture makes quantile crossing issues never occur on these basis quantiles under whichever input y or hidden variable h is given. Figure 1a visualizes an IQF module on the basis estimates $\{q(\alpha_k \mid h)\}_{k=1}^K$ generation.

[1] $\mathbb{E}_{Z\sim F} L(F^{-1}, Z) \leq \mathbb{E}_{Z\sim F} L(G^{-1}, Z)$.

[2] Forecast models are the same across all time series i.



(a) IQF layer

(b) IQF for the final layers

Figure 1. A Seq2Seq framework combined with the IQF.

3.2. Inter/Extrapolation around Quantile Knots

In order to provide quantile estimates beyond basis estimates in (7), i.e., $q(\alpha \mid h)$ on $\alpha \notin \{\alpha_k\}$, IQF takes the interpolation and extrapolation strategies into account: non-parametric interpolation in the non-tail region and parametric extrapolation in the tails. To be specific, we consider a simple yet efficient linear interpolation and extrapolation decaying on tails, e.g., Gaussian or Exponential.

Let's assume the knot size $K \geq 2$. Then, the left and right quantiles are set to be the extreme quantile on the knots on both sides [3], i.e., $\alpha_{\text{tail}_L} = \alpha_1$ and $\alpha_{\text{tail}_R} = \alpha_K$.

Linear Interpolation. For the non-tail region $\alpha \in [\alpha_{\text{tail}_L}, \alpha_{\text{tail}_R}]$, linear interpolation on $\alpha_k \leq \alpha \leq \alpha_{k+1}$ interval gives:

$$\hat{q}(\alpha \mid h) = \frac{\alpha_{k+1} - \alpha}{\alpha_{k+1} - \alpha_k} q(\alpha_k \mid h) + \frac{\alpha - \alpha_k}{\alpha_{k+1} - \alpha_k} q(\alpha_{k+1} \mid h). \tag{8}$$

Note that interpolation $\hat{q}$ and basis q matches at every quantile in the knots $\{\alpha_k\}$, as desired.

Extrapolation with Parametric Tails. For tail regions, we can extrapolate the quantile estimate through some parametric modeling. We adopt the simplest approach best fit for the distribution-free principle by parameterizing with an exponential tail at both ends:

$$\alpha = \begin{cases} \exp(\beta_L * (\hat{q}(\alpha \mid h) - \gamma_L)), & \alpha \leq \alpha_{\text{tail}_L}, \\ 1 - \exp(-\beta_R * (\hat{q}(\alpha \mid h) - \gamma_R)), & \alpha \geq \alpha_{\text{tail}_R}. \end{cases} \tag{9}$$

Here, parameters (β_l, γ_l) and (β_r, γ_r) are uniquely chosen so that extrapolation $\hat{q}$ and basis q coincide on the two leftmost and two rightmost quantile knots α_1, α_2 and α_{K-1}, α_K, respectively. Then, this is equivalent to extrapolating:

$$\hat{q}(\alpha \mid h) = \begin{cases} \beta_L^{-1} \log \frac{\alpha}{\alpha_2} + q(\alpha_2 \mid h), & \alpha \leq \alpha_{\text{tail}_L}, \\ \beta_R^{-1} \log \frac{1-\alpha_{K-2}}{1-\alpha} + q(\alpha_{K-2} \mid h), & \alpha \geq \alpha_{\text{tail}_R}, \end{cases} \tag{10}$$

where

$$\beta_L = \frac{\log(\alpha_2/\alpha_1)}{q(\alpha_2 \mid y) - q(\alpha_1 \mid y)},$$
$$\beta_R = \frac{\log((1-\alpha_{K-2})/(1-\alpha_{K-1}))}{q(\alpha_{K-1} \mid y) - q(\alpha_{K-2} \mid y)}.$$

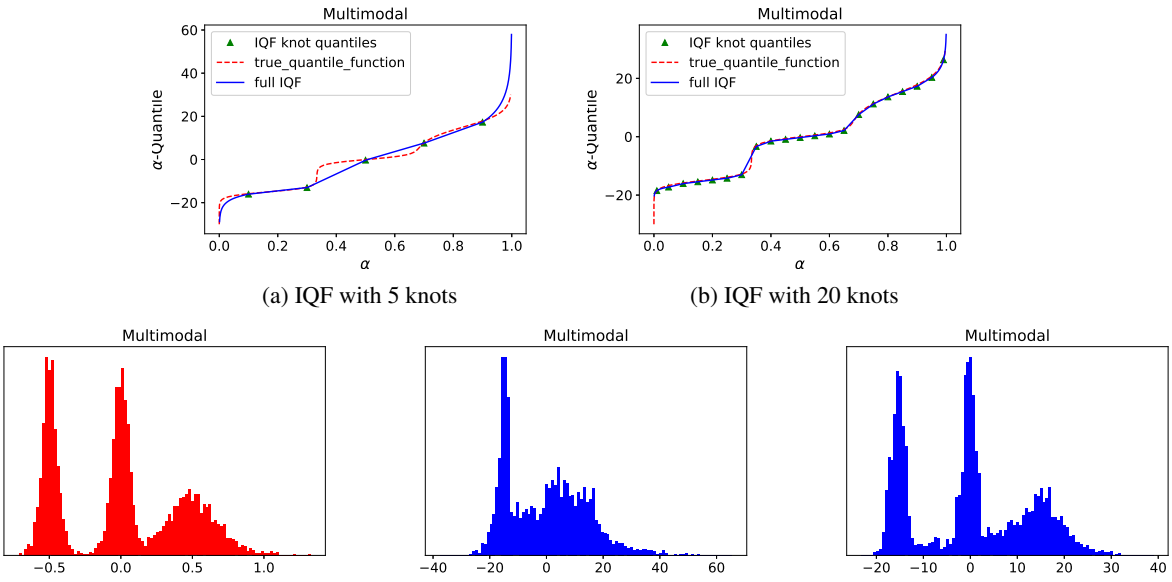
3.3. IQF as the Final Layer of Models

Aggregating interpolation (8) and extrapolation (10) in Section 3.2 based on basis estimates on knots in Section 3.1 completes IQF as follows:

$$\hat{q}(\cdot \mid y) = \begin{cases} (8), & \alpha \in [\alpha_{\text{tail}_L}, \alpha_{\text{tail}_R}], \\ (10), & \alpha \in (0, \alpha_{\text{tail}_L}] \cup [\alpha_{\text{tail}_R}, 1), \end{cases} \tag{11}$$

where y is an input to quantile function and $h = \mathbf{h}(y)$ is last hidden variable at prediction time t. This IQF can be

[3] The tail region can be selected more arbitrarily.



(a) IQF with 5 knots (b) IQF with 20 knots

(c) Empirical PDF from true QF. (d) Empirical PDF from IQF with 5 knots. (e) Empirical PDF from IQF with 20 knots.

Figure 2. Toy multimodal distribution with 3 peaks: IQF can interpolate and extrapolate around traning knots to approximate a multimodal distribution (c) without any distributional assumption, even more accurately as the number of knots increases from 5 ((a), (d)) to 20 ((b), (e)).

replaced with naive quantile output layer in the seq2seq forecast setting as depicted in Figure 1b.

Toward Sample Path Generations Now that IQF $\hat{q}(\cdot \mid y^t)$ allows to access on any quantile outside of knots, we can generate a sample path

$$[\hat{z}_{T+1}, \dots, \hat{z}_{T+\tau}] = [\hat{q}(\alpha \mid y_{T+1}), \dots, \hat{q}(\alpha \mid y_{T+\tau})],$$

by drawing α uniformly random, i.e., $\alpha \sim \mathrm{U}[0, 1]$.

Figure 2 shows the behaviors of IQF for a multi-modal distribution over different number of training knots[4]. It can learn without any distributional assumption. IQF also fits heavy tail distribution, e.g., Cauchy and Exponential. See the additional results in Appendix 7. Note that naive distributional assumption like Gaussianity fails in these examples. Even though IQF is in a distribution-free class, any additional information regarding to tail distribution for IQF can be enforced based on prior knowledge on the data, e.g., the choice of tail regions α_{tail_L} and α_{tail_R}, heavy tail distribution, Gaussian distribution, count data, etc.

Remark. Note that SQF approximates the conditional quantile function only through the family of piece-wise linear functions whereas IQF is the mixture of piece-wise linear in the middle and other parametric functions in tails. For specific choice of weights $w(\alpha)$ under the fixed knots, the optimization objectives (6) of SQF and IQF can be equivalent. Note that the whole interpolation and extrapolation let IQF provide the full function form under any choice of knots, whereas SQF may not be able to do.

[4]In toy examples, we assume that the quantiles learned at knots are exact.

4. Generalization Error

We transform available data $\mathcal{D} = \{z_{i,1:T}, \mathbf{x}_{i,1:T+\tau}\}$ into a tuple of input and output suitable for a seq2seq training and test framework, e.g., $\mathcal{D}_{\mathrm{train}}$ and $\mathcal{D}_{\mathrm{test}}$. In the stage of training, we pretend to forecast at $T-\tau$ for the next τ targets $\{z_{i,t}\}_{t=T-\tau+1}^{T}$ accessible in the data to compute a loss. On the other hand, the generalization takes the expected loss over next unknown τ targets $\{z_{i,t}\}_{t=T+1}^{T+\tau}$ from the actual forecast time T.

Formally, the *conditional generalized loss* is

$$\mathcal{L}(\theta \mid \mathcal{D}_{\mathrm{test}}) = \frac{1}{m\tau} \sum_{i=1}^{m} \sum_{t=T+1}^{T+\tau} \mathbb{E}\big[L(Z_{i,t}, q_\theta(\cdot \mid Y_i^t)) \mid Y_i^{T+\tau} = y_i^{T+\tau}\big], \tag{12}$$

under the CRPS L. Here $D_{\mathrm{test}} = \{y_i^t\}_{i=1,t=T+1}^{m,T+\tau}$ is the test data with test input $y_i^t = (z_{i,1:T}, \mathbf{x}_{i,1:t})$ to the quantile function. On the other hand, the *conditional empirical loss*

is given as:

$$\hat{\mathcal{L}}(\theta; \mathcal{D}_{\text{train}}) = \frac{1}{m\tau} \sum_{i=1}^{m} \sum_{t=T-\tau+1}^{T} \hat{L}(z_{i,t}, q_\theta(\cdot \mid y_i^t)), \quad (13)$$

under the approximated CRPS $\hat{L}$. Here $\mathcal{D}_{\text{train}} = \{z_{i,t}, y_i^t\}_{i=1,t=T-\tau+1}^{m,T}$ is the training data with input $y_i^t = (z_{i,1:T-\tau}, \mathbf{x}_{i,1:t})$ and corresponding observed target $z_{i,t}$ for $t = T-\tau+1, \ldots, T$.

We define

$$\theta^* = \text{argmin } \mathcal{L}(\theta \mid \mathcal{D}_{\text{test}}), \quad (14)$$

$$\hat{\theta} = \text{argmin } \hat{\mathcal{L}}(\theta; \mathcal{D}_{\text{train}}). \quad (15)$$

To begin with the analysis, we make the following assumptions.

Assumption 1. *For any $\bar{\delta} > 0$ and $y^t \in \mathcal{Y}$, there exist $D_2, D_3 \geq 0$ s.t.*

$$\Pr(|Z_t| \leq D_2 \mid Y^t = y^t) > 1 - \bar{\delta},$$

and

$$\Pr(\|Y^t\| \leq D_3) > 1 - \bar{\delta}.$$

Assumption 2. *The quantile estimates of IQF at knots $q_\theta(\alpha \mid y^t)$ is bounded by D_1, i.e.,*

$$|q_\theta(\alpha_1 \mid y^t)| \leq D_1, \ |q_\theta(\alpha_K \mid y^t)| \leq D_1.$$

for any fixed $y^t \in \mathcal{Y}$.

Assumption 1 is about bounded targets and input covariates with high probability. For the class of underlying distributions on unbounded domains, D_2 may be dependent on $\bar{\delta}$ but weakly. For example, when $Z_t \mid Y^t = y^t$ follows a sub-Gaussian[5] or sub-Exponential distribution, D_2 is $O\big(\sqrt{\log\left(1/\bar{\delta}\right)}\big)$ and $O(\log\left(1/\bar{\delta}\right))$, respectively. Regarding Assumption 2, the most of common quantile functions q_θ in neural network modeling frameworks, e.g., convolution, linear, MLP, ReLu, are bounded by some function of the norm of weight parameter and of input y^t and the total number of those components.

4.1. Conditional Generalization Error

Without loss of generality, we set equidistant quantile knots $\{\alpha_k\}_{k=1}^K$ and equal weights $w(\alpha)$ in IQF for the theoretical analysis purpose.

Definition 1. *The Rademacher complexity of $\mathcal{F}$ is defined as*

$$\mathfrak{R}_N(\mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left(\frac{1}{N} \sum_{i=1}^{N} \sigma_i f(Z_i)\right)\right],$$

where $\sigma_1, \ldots, \sigma_N$ are independent random variables uniformly chosen from $\{-1, 1\}$. Here the expectation is taken over both σ_i and data Z_i for $i = 1, \ldots, N$.

Lemma 1. *For the class of quantile function $\mathcal{Q}$ and CRPS loss L, the Rademacher complexity of $L \circ Q = \mathcal{F}$ with sample size N is upper bounded by that of Q, i.e.,*

$$\mathfrak{R}_N(\mathcal{F}) \leq \mathfrak{R}_N(\mathcal{Q}).$$

The proof is provided in Appendix 9. Lemma 1 indicates that the complexity of composite function class $\mathcal{F}$ only depends on that of quantile function class $\mathcal{Q}$. Note that standard neural network models have $\mathfrak{R}_N(\mathcal{Q})$ decreasing on sample size N, e.g., $O(\frac{1}{\sqrt{N}})$ and proportional to some max norm of weight parameters, input space, and number of layers.

Definition 2. *The temporal discrepancy over τ time difference is defined as*

$$\Delta_{\text{dis}}^{\tau}(\mathcal{D}_{\text{test}}) = \sup_{\theta} [\mathcal{L}(\theta \mid \mathcal{D}_{\text{test}}) - \mathcal{L}(\theta \mid \mathcal{D}_{\text{test}}^{-\tau})].$$

Here $\mathcal{D}_{\text{test}}^{-\tau} = \{y_i^t\}_{i=1,t=T-\tau+1}^{m,T}$ is the backtest data with τ-shifted input $y_i^t = (z_{i,1:T-\tau}, \mathbf{x}_{i,1:t})$, and its conditional generalized loss on $\mathcal{D}_{\text{test}}^{-\tau}$ is

$$\mathcal{L}(\theta \mid \mathcal{D}_{\text{test}}^{-\tau}) = \frac{1}{m\tau} \sum_{i=1}^{m} \sum_{t=T-\tau+1}^{T} \mathbb{E}\big[L(Z_{i,t}, q_\theta(\cdot \mid Y_i^t)) \mid Y_i^T = y_i^T\big].$$

Theorem 1. *IQF guarantees a conditional generalization error as follows. For any $\delta > \bar{\delta} > 0$,*

$$\mathcal{L}(\hat{\theta} \mid \mathcal{D}_{\text{test}}) - \mathcal{L}(\theta^* \mid \mathcal{D}_{\text{test}}) \leq 2\Delta_{\text{quant}} + 2\Delta_{\text{dis}}^{\tau}(\mathcal{D}_{\text{test}}) + 2\mathfrak{R}_{m\tau}(\mathcal{Q} \mid \mathcal{D}_{\text{test}}^{-\tau})) + 2\sqrt{\frac{2D^2}{m\tau} \log\left(\frac{1}{\delta - \bar{\delta}}\right)},$$

holds with at least $1 - \delta$ probability. Here, $D = \max(D_1, D_2, D_3)$, $\Delta_{\text{quant}} = O\left(\frac{D}{K}\right)$, and $\mathfrak{R}_{m\tau}(\mathcal{Q} \mid \mathcal{D}_{\text{test}}^{-\tau})$ is the Rademacher complexity of $\mathcal{Q}$ where the expectation on target data is conditioned on backtest input data $\mathcal{D}_{\text{test}}^{-\tau}$.

The proof is provided in Appendix 9. Theorem 1 says that the error consists of terms in the quantization error on CRPS, the discrepancy about stationarity, and the Rademachar complexity on quantile function, along with the additional high probability term but only with the logarithmic dependency. All of terms except discrepancy diminish to zeros as the sample size $m\tau$ and knot size K increase. Note that $\bar{\delta}$ can be arbitrarily smaller than δ by choice.

[5] $\Pr(|Z_t| > \epsilon) \leq c_1 e^{-c_2 \epsilon^2}$

Corollary 1. *For a stationary time series with sub-Gaussian tail distribution and any $\delta > 0$,*

$$\mathcal{L}(\hat{\theta} \mid \mathcal{D}_{\text{test}}) - \mathcal{L}(\theta^* \mid \mathcal{D}_{\text{test}}) \leq$$
$$2\mathfrak{R}_{m\tau}(\mathcal{Q} \mid \mathcal{D}_{\text{test}}^{-\tau}\}) + O\left(\left(\sqrt{\frac{1}{m\tau}} + \frac{1}{K}\right) \log^{\frac{3}{2}}\left(\frac{1}{\delta}\right)\right),$$

holds with at least $1 - 2\delta$ probability, where $\Delta_{\text{quant}} = O\left(\frac{1}{K} \log\left(\frac{1}{\delta}\right)\right)$.

The proof is immediate by setting $\delta = 2\bar{\delta}$ along with the fact that $\Delta_{\text{dis}}^{\tau}(\mathcal{D}_{\text{test}}) = 0$, $D = O \log\left(\sqrt{\frac{1}{\delta}}\right)$.

4.2. Unconditional Generalization Error

Unconditional generalization loss $\mathcal{L}$ is defined as

$$\mathcal{L}(\theta) = \frac{1}{\tau} \sum_{t=T+1}^{T+\tau} \mathbb{E}\big[L(Z_t, q_\theta(\cdot \mid Y^t))\big]. \tag{16}$$

Definition 3. *The expected temporal discrepancy at forecast time T on the next τ is*

$$\Delta_{\text{dis}}^{\tau} = \mathbb{E}[\Delta_{\text{dis}}^{\tau}(\mathcal{D}_{\text{test}})].$$

Assumption 3. *Time series are i.i.d., i.e.,*

$$Z_{i,1:T+\tau}, X_{i,1:T+\tau} \perp Z_{j,1:T+\tau}, X_{j,1:T+\tau},$$

for $i \neq j \in [m]$.

Corollary 2. *IQF guarantees a generalization error as follows. For any $\delta > \bar{\delta}$,*

$$\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta^*) \leq 2\Delta_{\text{quant}} + 2\Delta_{\text{dis}}^{\tau} + 2\mathfrak{R}_{m\tau}(\mathcal{Q}) + 2\sqrt{\frac{2D^2}{m\tau} \log\left(\frac{1}{\delta - \bar{\delta}}\right)},$$

with at least $1 - \delta$ probability.

The proof is immediate by taking expectation on Theorem 1 over $\mathcal{D}_{\text{test}}^{-\tau} = \{y_i^T\}$ under Assumption 3.

Remark. Even though a seemingly similar analysis exists, note that our theoretical results are arguably more sophisticated and informative, specifically targeted for the seq2seq quantile forecast setting with alleviated assumptions. Our error analysis, in particular, ERM (13) is designed with an approximated CRPS whereas the generalization (12) is defined over exact CRPS. The generalization errors are directly expressed with the complexity of quantile functionals $\mathcal{Q}$ by eliminating the loss (CRPS) dependency. Moreover, we consider the multi-horizon τ cases with more practical assumptions, along with examining its suitability on quantile models and distributions of our interest. On the other hand, existing works that depend on the complexity of both the loss function and models $\mathcal{F}$, exact CRPS, and several stronger assumptions on boundness that more general distribution classes do not satisfy.

5. Experiments

In the experiments, we show the impact of the incremental quantile function (IQF) on sequence-to-sequence forecasting models. We examine several properties of IQF including its no quantile crossing property, dynamic quantile query, and accuracy improvements. All experiments are done using AWS SageMaker (Liberty et al., 2020).

5.1. Setup

Datasets. We use the open-source benchmarking datasets from Dheeru & Karra Taniskidou (2017) as referenced in the below table. Both datasets have hourly frequencies, and the forecast is computed for prediction length $\tau = 24$.

DOMAIN	NAME	SUPPORT	FREQ	NO. TS
electrical load	`Elec`	$\mathbb{R}^+$	H	370
road traffic	`Traf`	$[0, 1]$	H	963

Forecast Models and Parameters. We select the MQ-CNN (Wen et al., 2017) model, which is a state-of-the-art time series quantile forecasting model in the sequence-to-sequence framework. This is available in the open-source GluonTS[6] (Alexandrov et al., 2020) package with the default hyper-parameters, e.g., context length = 4 * prediction length = 96 and epochs = 100. We use the following time features $\mathbf{x}_t \in \mathbb{R}^2$ as covariates: the day of the week and hour of the day.

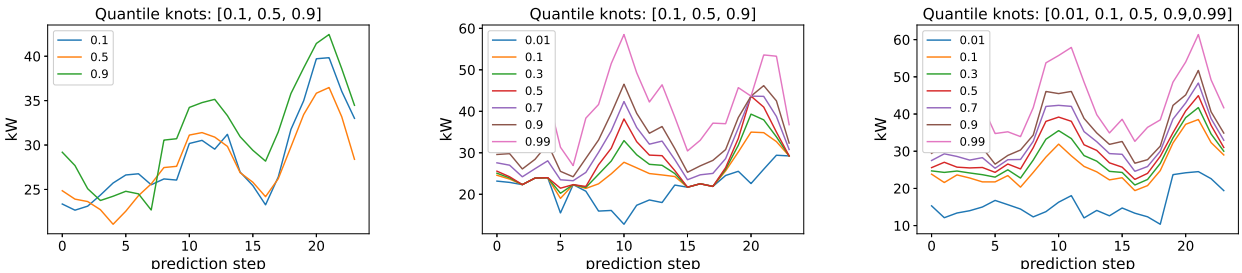
5.2. Results

Removal of Quantile Crossing. Figure 3 shows sample results of MQ-CNN with and without IQF layers on `Elec` data. MQ-CNN with our IQF layers in Figures 3b-3c show consistent predictions without any quantile crossing whereas original MQ-CNN in Figure 3a suffers severe quantile crossing.

Dynamic Quantile Query. The original MQ-CNN model can make an inference only on pre-determined quantile knots as shown in the case of 3 knots in Figure 3a. In other words, inferences on different quantiles on deployment are not immediately doable, and require an expensive re-training. MQ-CNN with IQF shown in Figures 3b-3c provides the full conditional quantile function queryable on any quantile. The more extreme case in Figure 3b also shows that MQ-CNN with IQF even performs well when training on only 3 knots and querying on 4 new quantiles.

Accuracy Improvements. Overcoming the issue of quantile crossing arguably leads to the overall test accuracy improvement. Table 1 shows that the improvement gain for

[6] https://github.com/awslabs/gluon-ts



(a) MQ-CNN with QF and 3 training knots (b) MQ-CNN with IQF and 3 training knots (c) MQ-CNN with IQF and 5 training knots

Figure 3. Experiment results of `Elec` dataset under the MQ-CNN forecaster with different quantile output layers and training quantile knots with prediction length = 24. Default quantile function (QF) is limited to query only on training quantiles, together with severe quantile crossing. On the other hand, IQF shows more consistent predictions with no quantile crossing over any quantile query.

Dataset	Strategy	training quantiles	wQL[0.05]	wQL[0.1]	wQL[0.5]	wQL[0.7]	wQL[0.9]	wQL[0.99]	mean_wQL
`Elec`	MQ-CNN-IQF	0.1, 0.5, 0.9	0.0316 (±6%)	0.0476 (±5%)	0.0816 (±2%)	0.0721 (±3%)	0.0377 (±3%)	0.0104 (±12%)	0.0468 (±2%)
		0.01, 0.1, 0.5, 0.9, 0.99	**0.0309** (±5%)	0.0482 (±5%)	**0.0808** (±3%)	**0.0718** (±4%)	**0.0366** (±4%)	**0.0093** (±3%)	**0.0463** (±3%)
	MQ-CNN	0.1, 0.5, 0.9	N/A	**0.0447** (±2%)	0.0864 (±1%)	N/A	0.0368 (±1%)	N/A	N/A
		0.01, 0.1, 0.5, 0.9, 0.99	N/A	0.0476 (±4%)	0.0934 (±4%)	N/A	0.0407 (±2%)	0.0094 (±8%)	N/A
`Traf`	MQ-CNN-IQF	0.1, 0.5, 0.9	0.1812 (±23%)	0.1823 (±21%)	0.1909 (±6%)	0.1952 (±7%)	0.1994 (±14%)	0.2014 (±17%)	0.1917 (±5%)
		0.01, 0.1, 0.5, 0.9, 0.99	**0.1115** (±9%)	0.1194 (±7%)	0.1828 (±4%)	**0.1840** (±1%)	0.1339 (±2%)	0.0478 (±4%)	**0.1299** (±2%)
	MQ-CNN	0.1, 0.5, 0.9	N/A	**0.0712** (±3%)	**0.1808** (±1%)	N/A	**0.1304** (±2%)	N/A	N/A
		0.01, 0.1, 0.5, 0.9, 0.99	N/A	0.0737 (±5%)	0.1893 (±4%)	N/A	0.1351 (±5%)	**0.0425** (±2%)	N/A

Table 1. Comparison of the accuracy measured by the weighted quantile losse (wQL) and training time for the base MQ-CNN model and our proposed MQ-CNN-IQF that resolves quantile crossing for prediction length = 24. The mean and standard deviation are computed over 5 runs, and the winning method is shown in bold.

IQF becomes more significant under a larger number of quantile knots, $[0.1, 0.5, 0.9]$ vs. $[0.01, 0.1, 0.5, 0.9, 0.99]$ in both `Elec` and `Traf` data. This implies that having more quantile knots can provide some regularization effects, by approximating the weighted quantile loss into the CRPS of our ultimate interest. Without IQF, however, this regularization effect is not observed as shown in `Traf` dataset. This leaves another concern to practitioners on hyperparameter tuning regarding the design of knot quantiles and the objective function accordingly. As shown in Table 1, IQF with 5 quantiles performs best overall except in one case of using 3 training knots for the `Traf` dataset without IQF.

6. Conclusions and Future Work

In this paper, we propose a distribution-free methodology that infers quantile estimates. Our approach, the incremental quantile function (IQF), is capable of forming a conditional quantile function that is accessible at any quantile query without any quantile crossing issue. As a by-product, our method saves us re-training costs, which are particularly expensive in large deep forecast models, when prediction quantiles are different from fixed training quantiles. We apply IQF to the scenarios of multi-horizon quantile time series forecasting, and derive the generalization error. In the experiments, we prove IQF resolves quantile crossing completely, and improves the forecast accuracy.

We leave for the future work various parametric extrapolation on top of current simple exponential tails, efficiently taking advantage of prior distribution information e.g., heavy tail, count data, non-negativeness. Developing the variant of IQF for multi-quantile prediction on multivariate time series is an open problem, alongside its theoretical analysis. While time series forecasting is mainly covered, each of our contributions open up our framework to new potential applications, providing additional benefits to future research.

References

Alexandrov, A., Benidis, K., Bohlke-Schneider, M., Flunkert, V., Gasthaus, J., Januschowski, T., Maddix, D. C., Rangapuram, S., Salinas, D., Schulz, J., Stella, L., Türkmen, A. C., and Wang, Y. GluonTS: Probabilistic and Neural Time Series Modeling in Python. *Journal of Machine Learning Research*, 21(116):1–6, 2020. URL `http://jmlr.org/papers/v21/19-820.html`.

Chen, Y., Kang, Y., Chen, Y., and Wang, Z. Probabilistic forecasting with temporal convolutional neural network. *Neurocomputing*, 399:491–501, 2020.

Dheeru, D. and Karra Taniskidou, E. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Eisenach, C., Patel, Y., and Madeka, D. Mqtransformer: Multi-horizon forecasts with context dependent and feedback-aware attention. *arXiv preprint arXiv:2009.14799*, 2020.

Gasthaus, J., Benidis, K., Wang, Y., Rangapuram, S. S., Salinas, D., Flunkert, V., and Januschowski, T. Probabilistic forecasting with spline quantile function rnns. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1901–1910. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/gasthaus19a.html.

Kim, T., Fakoor, R., Mueller, J., Smola, A. J., and Tibshirani, R. J. Deep quantile aggregation. *arXiv preprint arXiv:2103.00083*, 2021.

Koenker, R. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005. doi: 10.1017/CBO9780511754098.

Koenker, R. and Bassett Jr, G. Regression quantiles. *Econometrica: Journal of the Econometric Society*, pp. 33–50, 1978.

Kuznetsov, V. and Mohri, M. Learning theory and algorithms for forecasting non-stationary time series. In *NIPS*, pp. 541–549. Citeseer, 2015.

Liberty, E., Karnin, Z., Xiang, B., Rouesnel, L., Coskun, B., Nallapati, R., Delgado, J., Sadoughi, A., Astashonok, Y., Das, P., et al. Elastic machine learning algorithms in amazon sagemaker. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pp. 731–737, 2020.

Lim, B., Arik, S. O., Loeff, N., and Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *arXiv preprint arXiv:1912.09363*, 2019.

Liu, Y. and Wu, Y. Stepwise multiple quantile regression estimation using non-crossing constraints. *Statistics and its Interface*, 2(3):299–310, 2009.

Mariet, Z. and Kuznetsov, V. Foundations of sequence-to-sequence modeling for time series. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 408–417. PMLR, 2019.

Salinas, D., Flunkert, V., Gasthaus, J., and Januschowski, T. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.

Schmidt, L. and Zhu, Y. Quantile spacings: A simple method for the joint estimation of multiple quantiles without crossing. *Available at SSRN 2220901*, 2016.

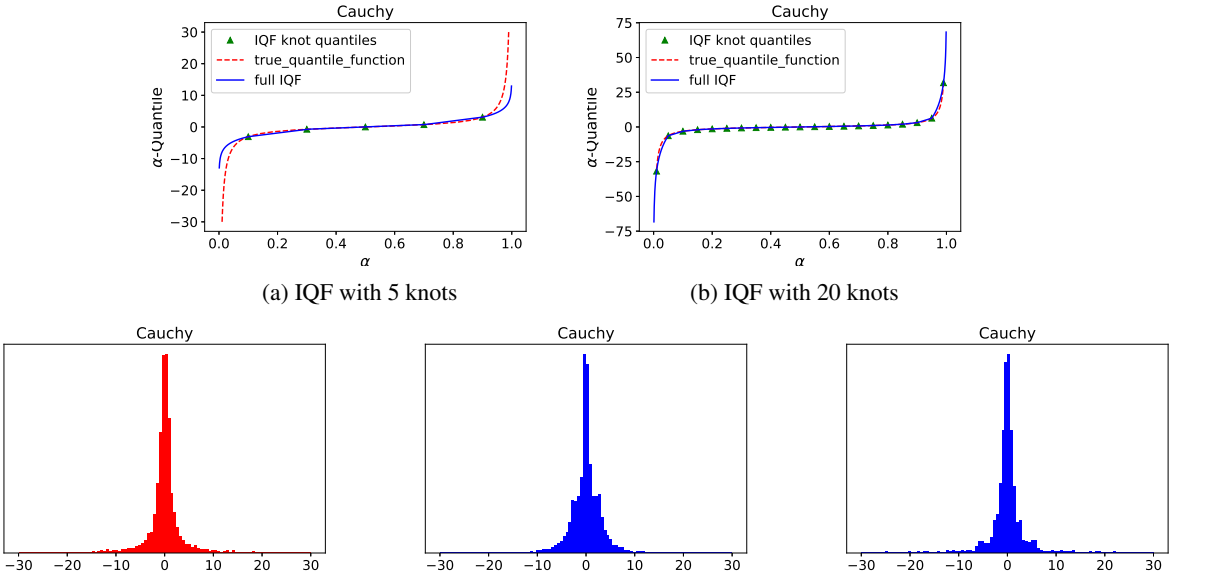
Wen, R. and Torkkola, K. Deep generative quantile-copula models for probabilistic forecasting. *arXiv preprint arXiv:1907.10697*, 2019.

Wen, R., Torkkola, K., and Narayanaswamy, B. A multi-horizon quantile recurrent forecaster. *arXiv preprint arXiv:1711.11053*, 2017.

Zimin, A. and Lampert, C. Learning theory for conditional risk minimization. In *Artificial Intelligence and Statistics*, pp. 213–222. PMLR, 2017.

(Supplementary Materials)

7. IQF for Cauchy Distribution



(a) IQF with 5 knots

(b) IQF with 20 knots

(c) Empirical PDF from true QF.

(d) Empirical PDF from IQF with 5 knots.

(e) Empirical PDF from IQF with 20 knots

Figure 4. IQF can fit a Cauchy distribution as a distribution-free approach with 5 training knots and 20 training knots.

8. The Proof about Rademachar Complexity

Lemma 2. *Let* $z_1, z_2 \in \mathcal{Z}_D$ *where* $\mathcal{Z}_D \subset \mathbb{B}_D := \{z \mid \|z\| \leq D\}$, *then*

$$|L(q, z_1) - L(q, z_2)| \leq 2D$$

Proof. Without loss of generality, let $z_1 \leq z_2$. Then, for a fixed α with $q := q(\alpha)$,

$$\rho_a(q - z_1) - \rho_a(q - z_2) = \begin{cases} \alpha(z_2 - z_1) & q \geq z_2 \\ q - z_2 + \alpha(z_2 - z_1) & z_2 \leq q \leq z_1 \\ (1 - \alpha)(z_2 - z_1) & q \leq z_1 \end{cases}$$

Since

$$(1 - \alpha)(z_1 - z_2) \leq q - z_2 + \alpha(z_2 - z_1) \leq \alpha(z_2 - z_1)$$

under $z_2 \leq q \leq z_1$,

$$|\rho_a(q - z_1) - \rho_a(q - z_2)| \leq \max(\alpha, 1 - \alpha)|z_2 - z_1| \leq |z_2 - z_1|$$

holds. Thus

$$\begin{aligned} \int \rho_\alpha(q(\alpha) - z_1) - \rho_\alpha(q(\alpha) - z_2) d\alpha &\leq \int |\rho_\alpha(q(\alpha) - z_1) - \rho_\alpha(q(\alpha) - z_2)| d\alpha \\ &\leq |z_2 - z_1| \\ &\leq 2D. \end{aligned}$$

□

Lemma 3. *CRPS $L(q, z)$ is 1-Lipschitz continuous w.r.t any quantile funciton $q \in \mathcal{Q}$ in L_1 norm, i.e., for any $z \in \mathcal{Z}_D$ and $q_1, q_2 \in \mathcal{Q}$,*

$$L(q_1, z) - L(q_2, z) \leq \|q_1 - q_2\|_1.$$

Proof. Similar to the proof of Lemma 2, for a fixed α and z,

$$|\rho_a(q_1(\alpha) - z) - \rho_a(q_2(\alpha) - z)| \leq |q_2(\alpha) - q_1(\alpha)|$$

holds. Thus

$$\begin{aligned}\int \rho_a(q_1(\alpha) - z) - \rho_a(q_2(\alpha) - z)d\alpha &\leq \int |\rho_a(q_1(\alpha) - z) - \rho_a(q_2(\alpha) - z)|d\alpha \\ &\leq \int |q_2(\alpha) - q_1(\alpha)|d\alpha \\ &= \|q_2 - q_1\|_1.\end{aligned}$$

□

Lemma 4. *The Rademacher complexity on $L \circ q \in \mathcal{F}$ is upperbounded by*

$$\mathfrak{R}_N(\mathcal{F}) \leq \mathfrak{R}_N(\mathcal{Q}),$$

where $q \in \mathcal{Q}$.

Proof. The proof is immediate from Lemma 3 and the propertiy of Rademacher complexity. □

9. The Proof of Main Theorems

Recall that $\mathcal{D}_{\text{test}} = \{y_i^{T+\tau}\}$ with $y_i^t = (z_{i,1:T}, \mathbf{x}_{i,1:T+\tau})$, $\mathcal{D}_{\text{test}}^{-\tau} = \{y_i^T\}$ with $y_i^t = (z_{i,1:T-\tau}, \mathbf{x}_{i,1:T})$, and $\mathcal{D}_{\text{train}} = \{z_i^T, y_i^T\}$.

Lemma 5. *Let $\{\bar{z}_i^T\}$ be the same as $\{z_i^T\}$ except for the one element $\bar{z}_{i,t} \neq z_{i,t}$ for some $T - \tau + 1 \leq t \leq T$ and $i \in [m]$. Then, for $\phi(\{z_i^T, y_i^T\}) := \sup_\theta[\mathcal{L}(\theta \mid \{y_i^T\}) - \bar{\mathcal{L}}(\theta; \{z_i^T, y_i^T\})]$,*

$$\phi(\{z_i^T, y_i^T\}) - \phi(\{\bar{z}_i^T, y_i^T\}) \leq \frac{2D}{m\tau}$$

holds.

Proof.

$$\begin{aligned}\phi(\{z_i^T, y_i^T\}) - \phi(\{\bar{z}_i^T, y_i^T\}) &\leq \bar{\mathcal{L}}(\theta; \{z_i^T, y_i^T\})] - \bar{\mathcal{L}}(\theta; \{\bar{z}_i^T, y_i^T\})] \\ &= \frac{1}{m}\frac{1}{\tau}\sum_{i=1}^{m}\sum_{t=T-\tau+1}^{T} L(z_{i,t}, q_\theta(\cdot \mid y_i^t)) - L(\bar{z}_{i,t}, q_\theta(\cdot \mid y_i^t)) \\ &\overset{(a)}{\leq} \frac{2D}{m\tau}\end{aligned}$$

where (a) holds due to Lemma 2. □

Lemma 6. *For $\phi(\{z_i^T, y_i^T\}) := \sup_\theta[\mathcal{L}(\theta \mid \{y_i^T\}) - \bar{\mathcal{L}}(\theta; \{z_i^T, y_i^T\})]$,*

$$\Pr\left(\phi(\{z_i^T, y_i^T\}) - \mathbb{E}\phi(\{z_i^T, y_i^T\}) > \epsilon\right) < e^{-\frac{m\tau}{2D^2}\epsilon^2}$$

holds where $\mathbb{E}$ is the expectation over Z_i^T conditioned on y_i^T for $i \in [m]$.

Proof. The results is immediate through applying Lemma 5 into McDiarmid's inequality. □

Assumption 4. *The quantile estimates of IQF at knots $q_\theta(\alpha \mid y^t)$ is bounded by D_1, i.e.,*

$$|q_\theta(\alpha_1 \mid y^t)| \le D_1,\ |q_\theta(\alpha_K \mid y^t)| \le D_1.$$

for any fixed $y^t \in \mathcal{Y}$.

Assumption 5. *For any $\bar{\delta} > 0$ and $y^t \in \mathcal{Y}$, there exist $D_2, D_3 \ge 0$ s.t.*

$$\Pr(|Z_t| \le D_2 \mid Y^t = y^t) > 1 - \bar{\delta},$$

and

$$\Pr(\|Y^t\| \le D_3) > 1 - \bar{\delta}.$$

Example 1. *For $Z \mid Y^t = y_t \sim \mathcal{N}(\mu, \sigma)$,*

$$\Pr(Z \in \mathcal{Z}_\mathcal{D} \mid Y^t = y_t) > 1 - \bar{\delta}$$

holds where $\mathcal{Z}_\mathcal{D} = \{z \mid |z - \mu| \le \sigma\sqrt{2\log\left(\frac{1}{\bar{\delta}}\right)})\}$. Note that $\mathcal{Z}_\mathcal{D} \subset \mathbb{B}_D$ where $D = |\mu| + \sigma\sqrt{2\log\left(\frac{1}{\bar{\delta}}\right)}$.

Proof. From the upperbound of Gaussian tail distribution $\Pr(|Z - \mu| > \sigma x) \le 2e^{-2x^2}$, it is immediate to see

$$\Pr(|Z - \mu| > \sigma\sqrt{2\log\left(\frac{1}{\delta}\right)}) \le \delta.$$

□

Lemma 7. *Let $\Delta_{\text{quant}} = \frac{2D+2}{K}$ and $D = \max(D_1, D_2, D_3)$. Then, under the Assumption 4 and 5,*

$$\Pr\left(|L(Z_t, q_\theta(\cdot \mid Y^t)) - \hat{L}(Z_t, q_\theta(\cdot \mid Y^t))| \ge \Delta_{\text{quant}}\right) < \bar{\delta}.$$

holds for any $z \in \mathcal{Z}$.

Proof. Let $\alpha_{\min} = \min\{\alpha_k\}$ and $\alpha_{\max} = \max\{\alpha_k\}$. Here, we drop the notation time t and θ in $q_\theta(\cdot \mid Y^t)$ and Z_t, and $\rho := \rho_\alpha$ for simplicity when obvious.

First, let's consider the case of $Z \in \mathcal{Z}_D$. If there exists $\alpha^* \in [\alpha_{\min}, \alpha_{\max}]$ s.t. $q(\alpha^* \mid Y) = Z$. For $w(\alpha) = \frac{2}{K}$ for $\alpha = \alpha_{\min}$ and $w(\alpha) = \frac{1}{K}$ otherwise, then

$$L(Z,q) - \hat{L}(Z,q) = \int_{\alpha_{\min}}^{\alpha_{\max}} \rho(Z - q(\alpha \mid Y)) - \sum_{\alpha \in \{\alpha_k\}_{k=1}^{K-1}} \frac{1}{K}\rho_\alpha(Z - q(\alpha \mid Y)) \tag{17}$$

$$+ \int_0^{\alpha_{\min}} \rho(Z - q(\alpha \mid Y)) - \frac{1}{K}\rho_{\alpha_{\min}}(Z - q(\alpha_{\min} \mid Y)) \tag{18}$$

$$+ \int_{\alpha_{\max}}^1 \rho(Z - q(\alpha \mid Y)) - \frac{1}{K}\rho_{\alpha_{\max}}(Z - q(\alpha_{\max} \mid Y)) \tag{19}$$

According to Riemmanian integral approximation,

$$\begin{aligned}
(17) &\overset{()}{\le} \frac{1}{K}\left(\alpha_{\max}(q(\alpha_{\max} \mid Y) - Z) - \alpha^*(q(\alpha^* \mid Y) - Z)\right) + \frac{1}{K}\left(\alpha_{\min}(q(\alpha_{\min} \mid Y) - Z) - \alpha^*(q(\alpha^* \mid Y) - Z)\right) \\
&\le \frac{1}{K}\left((\alpha_{\max} - \alpha^*)(q(\alpha_{\max} \mid Y) - Z)\right) \\
&\le \frac{2}{K}\max(D_q, D_Z)
\end{aligned}$$

where (a) holds due to the existence of α^* and its monotonicity on its left and right respectively.

$$(18) \leq \int_{\alpha_{\max}}^{1} (q(\alpha \mid Y) - Z)d\alpha - \frac{1}{K}\left(q(\alpha_{\max} \mid Y) - Z\right) = \int_{\alpha_{\max}}^{1} \left(q(\alpha \mid Y) - q(\alpha_{\max} \mid Y)\right) d\alpha \overset{()}{\leq} \frac{1}{K}$$

where (b) holds due to exponential tails of IQF. Likewise, $(19) \leq \frac{1}{K}$.

Now, for a given Z, let's consider the case that $\alpha^* \notin [\alpha_{\min}, \alpha_{\max}]$ s.t. $q(\alpha^* \mid Y)$. Then, $\rho_\alpha(Z - q)$ is monotonic on α. Again, according to Riemmanian integral approximation,

$$L(Z, q) - \hat{L}(Z, q) = \frac{D}{K} + \frac{2}{K} \tag{20}$$

To sum up, when $Z \in \mathcal{Z}_D$,

$$L(Z, q_\theta(\cdot \mid Y^t)) - \hat{L}(Z, q_\theta(\cdot \mid Y^t)) \leq \frac{2D + 2}{K}. \tag{21}$$

Let an event $E := |L(Z, q_\theta(\cdot \mid Y^t)) - \hat{L}(Z, q_\theta(\cdot \mid Y^t))| > \Delta_{\text{quant}}$ where $\Delta_{\text{quant}} := \frac{2D+2}{K}$. Then

$$\begin{aligned}
\Pr(E) &= \Pr(E \cap \{Z \in \mathcal{Z}_D\}) + \Pr(E \cap \{Z \notin \mathcal{Z}_D\}) && (22)\\
&\leq \Pr(E \mid \{Z \in \mathcal{Z}_D\}) + \Pr(\{Z \notin \mathcal{Z}_D\}) && (23)\\
&\overset{(c)}{=} \Pr(\{Z \notin \mathcal{Z}_D\}) && (24)\\
&\leq \delta_Z && (25)
\end{aligned}$$

where (c) follows from Equation (21).

□

Definition 4. *Given $\mathcal{D}_{\text{test}} = \{y_i^{T+\tau}\}$, the temporal discrepancy is defined as*

$$\Delta_{\text{dis}}(\mathcal{D}_{\text{test}}) = \sup[\mathcal{L}(\theta \mid \mathcal{D}_{\text{test}}) - \mathcal{L}(\theta \mid \mathcal{D}_{\text{test}}^{-\tau})].$$

Equivalently,

$$\Delta_{\text{dis}}(\{y_i^{T+\tau}\}) = \sup[\frac{1}{m\tau}\sum_{i=1}^{m}\sum_{t=T+1}^{T2\tau} \mathbb{E}\left[L(Z_{i,t}, q_\theta(\cdot \mid y_i^t)) - L(Z_{i,t-\tau}, q_\theta(\cdot \mid \bar{y}_i^{t-\tau})) \mid y_i^{T+2\tau}\right]]$$

where $y_i^t = (z_{i,1:T}, \mathbf{x}_{i,1:t})$ for test input and $\bar{y}_i^t = (z_{i,1:T-\tau}, \mathbf{x}_{i,1:t})$ for backtest input.

Note that due to sub-additivity of supremum,

$$\Delta_{\text{dis}}(\{y_i^{T+\tau}\}) \leq \sum_{t=T+1}^{T+\tau} \Delta_{\text{dis}}^t$$

where $\Delta_{\text{dis}}^t = \sup[\mathcal{L}(\theta \mid \{y_i^t\}) - \mathcal{L}(\theta \mid \{\bar{y}_i^{t-1}\})]$.

Theorem 2. *IQF with equidistant quantile knots $\{\alpha_k\}_{k=1}^K$ guarantees, with at least $1 - \delta$ probability,*

$$\mathcal{L}(\hat{\theta} \mid \mathcal{D}_{\text{test}}\}) - \mathcal{L}(\theta^* \mid \mathcal{D}_{\text{test}}\}) \leq 2\mathfrak{R}_{m\tau}(\mathcal{Q} \mid \mathcal{D}_{\text{test}}^{-\tau}) + 2\Delta_{\text{dis}}(\mathcal{D}_{\text{test}}^{-\tau}) + 2\Delta_{\text{quant}} + 2\sqrt{\frac{2D^2}{m\tau}\log\left(\frac{1}{\delta - \delta_D}\right)}$$

where $\Delta_{\text{quant}} = o\left(\frac{D}{K}\right)$.

Proof. Let $\mathcal{L}(\theta) := \mathcal{L}(\theta \mid \mathcal{D}_{\text{test}})$ and $\hat{\mathcal{L}}(\theta) = \hat{\mathcal{L}}(\theta; \mathcal{D}_{\text{train}})$. Then

$$\begin{aligned}\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta^*) &= \mathcal{L}(\hat{\theta}) - \hat{\mathcal{L}}(\hat{\theta}) + \hat{\mathcal{L}}(\hat{\theta}) - \hat{\mathcal{L}}(\theta^*) + \hat{\mathcal{L}}(\theta^*) - \mathcal{L}(\theta^*) \\ &\leq L(\hat{\theta}) - \hat{\mathcal{L}}(\hat{\theta}) + \hat{\mathcal{L}}(\theta^*) - \mathcal{L}(\theta^*) \\ &\leq 2 \sup_{\theta} |\mathcal{L}(\theta) - \hat{\mathcal{L}}(\theta)|.\end{aligned}$$

The sub-additivity of the supremum gives

$$\begin{aligned}\sup[\mathcal{L}(\theta \mid \mathcal{D}_{\text{test}}) - \hat{\mathcal{L}}(\theta; \mathcal{D}_{\text{train}})] \leq & \underbrace{\sup[\mathcal{L}(\theta \mid \mathcal{D}_{\text{test}}) - \mathcal{L}(\theta \mid \mathcal{D}_{\text{test}}^{-\tau})]}_{\Delta_{\text{dis}}(\mathcal{D}_{\text{test}}))} \\ & + \underbrace{\sup[\mathcal{L}(\theta \mid \mathcal{D}_{\text{test}}^{-\tau})) - \bar{\mathcal{L}}(\theta; \mathcal{D}_{\text{train}}))]}_{\phi(\mathcal{D}_{\text{train}}))} + \underbrace{\sup[\bar{\mathcal{L}}(\theta; \mathcal{D}_{\text{train}})) - \hat{\mathcal{L}}(\theta; \mathcal{D}_{\text{train}}))]}_{\psi(\mathcal{D}_{\text{train}}))}\end{aligned} \tag{26}$$

where

$$\begin{aligned}L(\theta \mid \mathcal{D}_{\text{test}})) &= \frac{1}{m\tau} \sum_{i=1}^{m} \left[\sum_{t=T+1}^{T+\tau} \mathbb{E}\left[L(Z_{i,t}, q_\theta(\cdot \mid Y_i^t)) \;\middle|\; Y_i^{T+\tau} = y_i^{T+\tau} \right] \right], \\ L(h \mid \mathcal{D}_{\text{test}}^{-\tau})) &= \frac{1}{m\tau} \sum_{i=1}^{m} \left[\sum_{t=T-\tau+1}^{T} \mathbb{E}\left[L(Z_{i,t}, q_\theta(\cdot \mid Y_i^t)) \;\middle|\; Y_i^{T} = y_i^{T} \right] \right].\end{aligned}$$

The first term on RHS follows from the definition Δ_{dis}.

For the second term $\phi(\mathcal{D}_{\text{train}})$, let's define

$$L^\gamma(h) = \frac{1}{m\tau} \sum_{i=1}^{m} \left[\sum_{t=T+1}^{\gamma} \mathbb{E}\left[L(Z_{i,t}, q_\theta(\cdot \mid Y_i^t)) \;\middle|\; Y_i^{T+\tau} = y_i^{T+\tau} \right] + \sum_{t=\gamma+1}^{T+\tau} L(z_{i,t}, q_\theta(\cdot \mid y_i^t)) \right]$$

for $\gamma = T - \tau + 1, \ldots, T$. Then telescoping the term gives

$$\mathcal{L}(h \mid \{y_i^T\}) - \bar{\mathcal{L}}(h; \{z_i^T, y_i^T\}) = L^T(h) - L^{T-\tau}(h) \tag{27}$$

$$= \sum_{\gamma=T-\tau+1}^{T} L^\gamma(h) - L^{\gamma-1}(h) \tag{28}$$

$$= \frac{1}{m\tau} \sum_{\gamma=T-\tau+1}^{T} \sum_{i=1}^{m} \mathbb{E}\left[L(Z_{i,\gamma}, q_\theta(\cdot \mid Y_i^\gamma)) \;\middle|\; Y_i^T = y_i^T \right] - L(z_{i,\gamma}, q_\theta(\cdot \mid y_i^\gamma)). \tag{29}$$

Recall that $\mathcal{D}_{\text{train}} = \{z_i^T, y_i^T\}$ Now the expectation of ϕ conditioned on $\{y_i^{T+\tau}\}$ is

$$\mathbb{E}\phi(\mathcal{D}_{\text{train}}) = \mathbb{E}\sup[-\mathcal{L}(\theta \mid \mathcal{D}_{\text{train}}^{-\tau}) - \bar{\mathcal{L}}(\theta; \mathcal{D}_{\text{train}})] \tag{30}$$

$$\overset{(a)}{\leq} \mathbb{E}\sup \frac{1}{m\tau} \sum_{i,t} f(\theta; i, t) - \mathbb{E} f(\theta; i, t) \tag{31}$$

$$\overset{(b)}{\leq} \mathfrak{R}_{m\tau}(\mathcal{F} \mid \{y_i^T\}) \tag{32}$$

where (a) holds by defining $f(\theta; i, t) = L(z_{i,t}, q_\theta(\cdot \mid y_i^t))$ and $\mathbb{E} f(\theta; i, t) = \mathbb{E}_{Z_{i,t} \mid y_i^t}\left[L(Z_{i,t}, q_\theta(\cdot \mid Y_i^t)) \;\middle|\; Y_i^T = y_i^T \right]$ and (b) holds due to standard symmetrization trick followed by the definition of Rademacher complexity on $f = L \circ q \in \mathcal{F}$.

For the third term $\psi(\{z_i^{T+\tau}, y_i^{T+\tau}\})$,

$$\bar{\mathcal{L}}(h; \{z_i^{T+\tau}, y_i^{T+\tau}\}) - \hat{\mathcal{L}}(h; \{z_i^{T+\tau}, y_i^{T+\tau}\}) = \frac{1}{N}\frac{1}{\tau}\sum_{i=1}^{m}\sum_{t=T}^{T+\tau} L(z_{i,t}, q_\theta(\cdot \mid y_i^t)) - \hat{L}(z_{i,t}, q_\theta(\cdot \mid y_i^t)) \overset{()}{\leq} \frac{2D+1}{K} \tag{33}$$

Aggregating Equation (26) and (33) gives

$$G_{m\tau} := \sup_\theta |\mathcal{L}(\theta) - \hat{\mathcal{L}}(\theta)| - \left(\Delta_{\text{dis}} + \mathbb{E}\phi(\{z_i^T, y_i^T\}) + \Delta_{\text{quant}}\right) \leq \phi(\mathcal{D}_{\text{train}}) - \mathbb{E}\phi(\mathcal{D}_{\text{train}}) + \psi(\mathcal{D}_{\text{train}}) - \Delta_{\text{quant}}$$

Writing $\phi := \phi(\mathcal{D}_{\text{train}})$ and $\psi := \psi(\mathcal{D}_{\text{train}})$ in shorthand, we have

$$\begin{aligned}
\Pr(G_{m\tau} > \epsilon) &\leq \Pr\left(\phi - \mathbb{E}\phi + \psi - \frac{2D+1}{K} \geq \epsilon\right) \\
&\leq \Pr\left(\phi - \mathbb{E}\phi + \psi - \frac{2D+1}{K} \geq \epsilon\right) \\
&\leq \Pr\left(\phi - \mathbb{E}\phi + \psi - \frac{2D+1}{K} \geq \epsilon, \psi - \frac{2D+1}{K} \leq 0\right) + \Pr\left(\phi - \mathbb{E}\phi + \psi - \frac{2D+1}{K} \geq \epsilon, \psi - \frac{2D+1}{K} \geq 0\right) \\
&\leq \Pr\left(\phi - \mathbb{E}\phi \geq \epsilon\right) + \Pr\left(\psi - \frac{2D+1}{K} \geq 0\right) \\
&\overset{(c)}{\leq} \Pr\left(\phi - \mathbb{E}\phi \geq \epsilon\right) + \delta_Z \\
&\overset{(d)}{\leq} e^{-\frac{m\tau}{2D^2}\epsilon^2} + \delta_Z
\end{aligned}$$

where (c) and (d) hold due to Lemma 6 and Lemma 7. Choosing $\delta > \bar{\delta}$ and setting

$$\epsilon = \sqrt{\frac{2D^2}{m\tau}\log\left(\frac{1}{\delta - \bar{\delta}}\right)}$$

guarantees the following w.p. $1 - \delta$:

$$\sup|\mathcal{L}(\theta) - \hat{\mathcal{L}}(\theta)| \leq \mathbb{E}\phi + \Delta_{\text{dis}}(\mathcal{D}_{\text{test}}) + \Delta_{\text{quant}} + \sqrt{\frac{2D^2}{m\tau}\log\left(\frac{1}{\delta - \delta_Z}\right)} \tag{34}$$

$$\overset{(f)}{\leq} \mathfrak{R}_{m\tau}(\mathcal{Q} \mid \mathcal{D}_{\text{test}}^{-\tau}) + \Delta_{\text{dis}}(\{y_i^{T+2\tau}\}) + \Delta_{\text{quant}} + \sqrt{\frac{2D^2}{m\tau}\log\left(\frac{1}{\delta - \bar{\delta}}\right)} \tag{35}$$

where (f) holds due to Equation (32) and Lemma 4. □