AI Foundation Model for Time Series with Innovations Representation

Lang Tong

Xinyi Wang

School of Electrical and Computer Engineering
Cornell University, Ithaca, NY, USA

Menlo Park, Santa Clara, CA, USA

Abstract

This paper introduces an Artificial Intelligence (AI) foundation model for time series in engineering applications, where causal operations are required for real-time monitoring and control. Since engineering time series are governed by physical, rather than linguistic, laws, large-language-model-based AI foundation models may be ineffective or inefficient. Building on the classical innovations representation theory of Wiener, Kallianpur, and Rosenblatt, we propose Time Series GPT (TS-GPT)—an innovations-representation-based Generative Pre-trained Transformer for engineering monitoring and control. As an example of foundation model adaptation, we consider Probabilistic Generative Forecasting, which produces future time series samples from conditional probability distributions given past realizations. We demonstrate the effectiveness of TS-GPT in forecasting real-time locational marginal prices using historical data from U.S. independent system operators.

I. Introduction

A. The Rise of AI Foundation Model

At the heart of the current AI revolution is the Foundation Model (FM), designed to overcome computational, data, and learning challenges of AI applications across a broad range of applications. Today, FMs power some of the most successful AI applications, such as ChatGPT, demonstrating high levels of comprehension, fluency in natural language, and impressive capabilities in information extraction, synthesis, and reasoning.

A defining feature of the architectural design of an FM is the partition of highly complex AI tasks into two processes: (i) FM pretraining, which generates critical latent features—a summary statistic—applicable across a wide range of tasks, and (ii) FM adaptation, which fine-tunes the model for specific and task-oriented applications. The genius of this partition lies in the "division of labor" concept of Adam Smith [1], making FM an attractive business model for AI technology. By separating the costly AI pretraining that requires vast amounts of data and immense computation power from the more application-specific FM adaptations, the FM architecture allows AI companies with AI expertise, computation resources, and financial strength to monetize FMs as commercial products from which experienced practitioners can specialize and adapt these models to meet application needs, maximizing both efficiency and applicability.

At a high level, the architecture of FM pretraining can be abstracted as an autoencoder shown in the upper layer of Fig 1. The encoder transforms the input into a latent representation of the input, and the decoder generates outputs that match the input according to some similarity measure. Trained with large datasets, FM is capable of producing autoencoder output that approximate input with the right underlying probabilistic structure: sentences with proper grammar and images with authentic appearance. The FM autoencoder is also an abstraction of the so-called *transformer* architecture [2] where a specific "attention mechanisms" are embedded in the encoder-decoder structure to capture temporal dependencies and other data characteristics.

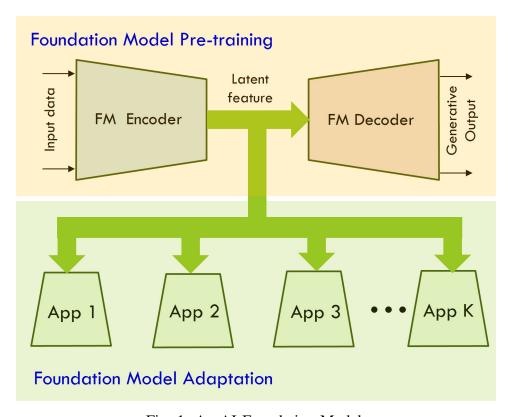


Fig. 1: An AI Foundation Model.

A key feature of the FM autoencoder is its randomization that makes the FM *generative*, capable of producing outputs outside the training samples used in the pretraining process. The generative feature of the FM plays a crucial role in enabling FM to evolve, to acquire new capabilities as it adapts to new environments, and to learn from new data.

Generative Pretrained Transformer, abbreviated as GPT, is the hallmark of modern AI architectures. Once pretrained, the underlying temporal dependencies of the input process are captured by the FM encoder neural network. The output of the FM encoder is the latent features representing the input, typically of lower dimensions, that can be used for various purposes: classification, prediction, language translation, interactive question-and-answer, and producing summary texts according to prompts.

Most prominent Foundation Models (FMs) are based on large language models (LLMs) [2], [3], which are trained and optimized for tasks involving text, speech, and image/video data. While there have been efforts to apply LLM-based FMs to physical systems, it remains an open question whether these language

models can effectively address real-time *causal* decision-making problems that rely on real-time data from systems governed by physical laws—the Ohm's, Kirchhoff's, and Maxwell's laws—other than the linguistic structures central to commercial FMs. The lack of a mathematical foundation and interpretability are critical barriers to FM adoption in many engineering fields, particularly for critical infrastructures and control systems where considerations of physical/cyber security, operational stability, and safety are paramount.

This article focuses on FMs for time series arising from physical systems—a causal time series GPT with a canonical autoencoder structure based on an innovations representation model—aimed at developing a time series GPT approach and an adaptation for generative probabilistic forecasting.

II. INNOVATIONS THROUGH THE LENS OF MODERN MACHINE LEARNING

A. The Wiener-Kallianpur Conjecture

In 1958, Norbert Wiener and Gopinath Kallianpur considered the problem of efficient representations of stationary random processes [4]. They postulated that a stationary random process $\boldsymbol{x} := (x_t)$ could be transformed by a causal encoder G to an independent and identically distributed random sequence $\boldsymbol{v}(v_t)$ with the uniformly distributed marginals on [0,1], herein referred to as IID-uniform. Furthermore, there exists a causal decoder H that maps the sequence \boldsymbol{v} back to the input sequence \boldsymbol{x} .

Specifically, there exist causal G and H such that

$$\begin{cases} v_t = G(x_t, x_{t-1}, \cdots), & v_t \stackrel{\text{IID}}{\sim} \mathcal{U}(0, 1) \\ \hat{x}_t = H(v_t, v_{t-1}, \cdots), & \hat{\boldsymbol{x}} \stackrel{\text{a.s.}}{=} \boldsymbol{x}. \end{cases}$$
(1)

The IID-uniform property of v and the almost-sure matching of x and \hat{x} imply that v_t is statistically independent of the past $x_{0:t-1} := (x_0, x_1, \dots, x_{t-1})$, thus the interpretation that v_t represents the new information at t—the innovation—and v the innovations sequence of x [5]-[7].

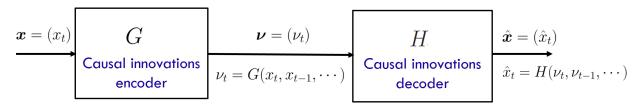


Fig. 2: Innovations autoencoder with causal encoder G and decoder H.

Through the lens of modern machine learning, the Wiener-Kallianpur conjecture suggests a universal and *causal* autoencoder architecture to represent stationary time series, as illustrated in Figure 3. The causality of the autoencoder makes the representation particularly suitable for real-time decision making.

We call the above autoencoder the Wiener-Kallianpur Autoencoder and (1) the Strong Innovations Representation¹ (SIR). If (G, H) exists, SIR is universal, nonparametric, and powerful. The Wiener-Kallianpur autoencoder "disentangles" complex time series models into the autoencoder (G, H) that

¹The word "strong" emphasizes that the output of the decoder matches the input on the realization basis. The innovation representation is *weak* if the matching is only in distribution.

captures the structural temporal dependencies of x and the latent sequence v that carries the incremental information of the input realizations. SIR is lossless, making the innovations sequence v a sufficient statistic for online decision-making; the optimal decision based on x is equivalent to that based on the much simpler IID-uniform innovations v.

Instances of innovations representations were studied before the Wiener-Kallianpur conjecture, starting from the work of Kolmogorov [8], Wiener [9], Bode and Shannon [10], Wiener and Masani [5]. In a sequence of contributions [11]–[14], Kailath and his co-authors made significant contributions that popularized the idea of innovations representation in engineering fields, particularly in control, communications, and signal processing. See [15], [16] for surveys of the field at the time.

For practical applications, SIR is the most powerful when the autoencoder (G,H) exists and can be computed. Most of the known cases in which a causal autoencoder can be obtained explicitly or from data involve Gaussian assumptions. The simplest case is the stationary Gaussian process, for which (G,H) can be computed from the (linear) minimum-mean-squared-error (MMSE) prediction error filter. By the orthogonality principle, the prediction error at every time is statistically independent of the past, making the prediction errors a sequence of innovations.

Kailath and Frost made a key contribution to the innovations representation of a particular type of non-Gaussian continuous-time random processes [17], [18]. In particular, they considered the class of additive white Gaussian noise (AWGN) processes defined as the sum of a (possibly) non-Gaussian stationary process and a white Gaussian noise (Wiener) process. Under mild assumptions that hold favorably in practical scenarios, they showed the remarkable result parallel to the Gaussian case: the innovations process is the output of the MMSE (nonlinear) prediction error filter. The Kailath-Frost innovations representation is especially appealing because it not only generalizes the SIR of Gaussian processes but also provides a method to extract the innovations process. Unfortunately, the SIR for the continuous-time AWGN processes does not translate directly to the discrete-time AWGN process. However, it is arguable that the nonlinear MMSE prediction error sequence can serve as a good approximation of the actual innovations sequence in practice. Indeed, the training of a time series foundation model with strong innovation represents developed in Sec. III follows the idea of nonlinear MMSE prediction.

B. Weak Innovations Representation

Powerful as SIR is, the generality of SIR was brought into question by Murray Rosenblatt, who constructed counterexamples for which SIR does not exist "even in the case of a finite-state purely nondeterministic Markov chain" [19], [20]. Although the existence of a causal mapping that extracts an IID-uniformly distributed random sequence is easily satisfied under mild assumptions, establishing the existence of causally invertible mappings is challenging. It remains an open problem to characterize the general existence conditions of SIR [20]–[22].

However, Rosenblatt considered a relaxation of the perfect reconstruction condition in (1) to the weaker version that requires the input and output of the autoencoder to match only in distribution²:

$$\begin{cases} v_t = G(x_t, x_{t-1}, \cdots), & v_t \stackrel{\text{IID}}{\sim} \mathcal{U}(0, 1) \\ \hat{x}_t = H(v_t, v_{t-1}, \cdots), & \hat{\boldsymbol{x}} \stackrel{\text{d}}{=} \boldsymbol{x}. \end{cases}$$
 (2)

²Rosenblatt credited Paul Levy for suggesting this relaxation in [23]

We shall call the representation (2) weak innovations representation (WIR), v the weak innovations sequence, and the autoencoder pair (G, H) a Weak Innovations AutoEncoder (WIAE). Rosenblatt showed that WIR exists for the purely nondeterministic finite-state stationary Markov chain—the cases in which SIR does not exist except for rare and nearly pathological scenarios.

The benefit of WIR for being applicable to broader classes of random processes comes with nontrivial costs. First, without requiring perfect reconstruction of the autoencoder input at its output, WIR representation cannot be used for certain applications, such as compression, when accurate recovery of the source is the objective.

Second, while a strong innovations sequence is a *sufficient statistic*, a weak innovations sequence is not in general. Therefore, decision-making based on weak innovations may be suboptimal. There are important exceptions, however. For the probabilistic forecasting problem discussed in Sec. IV, we show that the weak innovations sequence defined by (2) is *Bayesian sufficient* and that using the weak innovations sequence for probabilistic forecasting is optimal.

Finally, it is significant that the weak innovations sequence v does not have the same interpretation that v_t is statistically independent of the past $\mathcal{X}_{t-1} := (x_{t-1}, x_{t-2}, \cdots)$ as required by SIR (1). Instead, v_t is statistically independent of the past autoencoder *output* $\hat{\mathcal{X}}_{t-1} := (\hat{x}_{t-1}, \hat{x}_{t-2}, \cdots)$. In other words, the weak innovations are the (strong) innovations of \hat{x} rather than x.

III. A TIME SERIES GPT FOUNDATION MODEL

The strong and weak innovations representations, initially envisioned by Wiener, Kallianpur, and Rosenblatt, give perhaps the most succinct *causal* representations of general stationary processes with the autoencoder (G, H) capturing the statistical temporal dependencies and its latent process v the realized randomness of the underlying random process. Embedding a strong (or weak) innovations representation as the information processing engine of an AI foundation model for real-time decision-making is natural.

This section focuses on the Generative and Pretrained Transformer FM architecture for time series, illustrated in Fig. 1. Specifically, the pretraining of the FM involves using historical data to train the autoencoder (G, H) to extract the innovations sequence v as the latent feature of the realized input x. The autoencoder may include attention mechanisms, not necessarily the same in the same form of as the standard query-key-value implementations, that focus the attention to specific characteristics of the data such as spikiness and long range dependencies.

The generative feature of the time series FM is implemented by the decoder H. By replacing the innovations sequence v at the encoder output with independently generated *pseudo-innovations* \tilde{v} from IID-uniform distributions, the autodecoder H produces out-of-sample data \tilde{x} with the same distribution as but different realizations from the training data population.

A. Pretraining Foundation Model with Innovations Representation

Except for the Gaussian and AWGN (continuous-time) processes, there had not been computationally tractable solutions to construct strong or weak innovations autoencoders, severely limiting the applications of innovations representations. Here, we describe, at the architecture-level, the learning from data an

autoencoder pair (G, H) for the strong or weak innovations representation, based on a variation of the generative adversarial networks (GAN) learning strategy, following the first such development in [24].

The architecture of the learning of the strong and weak innovations representations is illustrated in Fig. 3, where the autoencoder (G, H) is realized by deep neural networks (G_{θ}, H_{η}) with coefficients (θ, η) , respectively. The dashed (red) and solid (black) lines are signal flow paths during pre-training. After training, only the solid (black) lines remain.

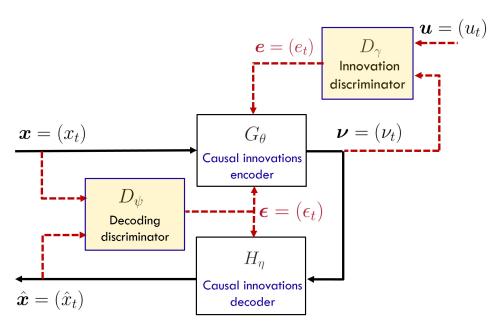


Fig. 3: A deep learning architecture of innovations representations.

At time t, the input of the causal encoder G_{θ} consists of the current and past samples $\boldsymbol{x}_t := (x_t, x_{t-1}, \cdots)$, and the output of G is the innovation sequence $\boldsymbol{v}_t := (v_t, v_{t-1}, \cdots)$. Likewise, the input of the decoder H_{η} is the innovation sequence (\boldsymbol{v}_t) , and the output $(\hat{\boldsymbol{x}}_t)$ the estimate of \boldsymbol{x}_t .

The objective of training the autoencoder pair (G_{θ}, H_{η}) toward its optimal setting is to drive the encoder G_{θ} to output an IID-uniform sequence as required by the innovations sequence and the decoder H_{η} to output \hat{x} that matches x, in distribution for WIR and in the mean-squared sense for SIR.

To generate training updates for the encoder G_{θ} , an innovations discriminator neural network D_{γ} compares the encoder output v with a synthetically generated IID-uniform sequence u and produces a Wasserstein distance error sequence e to update to update parameter e of the encoder e decoding discriminator e compares the autoencoder input e and output e and generates error sequence e to updates jointly e of the encoder e and decoder e and decoder e derives updates from the stochastic gradient of the mean squared error between e and e. For WIR, e derives updates from the stochastic gradient of the Wasserstein distance error as in e and e computation of Wasserstein distance error is standard.

Define the overall pretraining objective as the weighted sum of innovations sequence error at the output of the innovations discriminator and the decoding error at the output of the decoding discriminator:

$$L(\theta, \eta, \gamma, \psi) := \mathbb{E}[D_{\gamma}(\boldsymbol{v}, \boldsymbol{u})] + \lambda \mathbb{E}[D_{\psi}(\hat{\boldsymbol{x}}, \boldsymbol{x})].$$

Through the Kantorovich-Rubinstein duality [25], the neural network parameters $(\theta, \eta, \gamma, \psi)$ are obtained from the min-max optimization:

$$\min_{\theta,\eta} \max_{\gamma,\psi} L(\theta,\eta,\gamma,\psi). \tag{3}$$

Standard stochastic gradient descent algorithms can be used to train (D_{η}, D_{ψ}) . A pseudo-code with a detailed training procedure can be found in an unpublished arXiv article [26]. Our experience showed that the training of WIR appears easier than that of SIR, thanks to the use of the Wasserstein GAN in the decoding discriminator.

B. Structural Convergence

The existence of SIR and WIR requires the autoencoder (G, H) to take input samples from the infinite past. While a recurrent neural network autoencoder could be used, the training of such an autoencoder is challenging. If the autoencoder (G_{θ}, H_{η}) and related discriminators (D_{γ}, D_{ψ}) are implemented with large but finite-dimensional feedforward convolutional neural networks that input from only the finite past, can the learned autoencoder approximate the true strong/weak innovations well?

In [24], structural convergence is established for strong innovations representations. The same can be shown for the weak representations. Assume that all neural networks in Fig. 3 take as their input the past k data samples. Let the finite k+1-input dimensional autoencoder be $(G_{\theta}^{(k)}, H_{\eta}^{(k)})$, and

$$\begin{cases}
v_t^{(k)} &:= H^{(k)}(x_t, x_{t-1}, \dots, x_{t-k}), \\
\hat{x}_t^{(k)} &:= G^{(k)}\left(v_t^{(k)}, v_{t-1}^{(k)}, \dots, v_{t-k}^{(k)}\right).
\end{cases}$$
(4)

Let the finite input-dimensional discriminators $(G_{\theta}^{(k)}, H_{\eta}^{(k)})$ be similarly defined. It can be shown that, as $k \to \infty$, $\boldsymbol{v}^{(k)} \overset{\text{m.s.}}{\to} \boldsymbol{v}$ and $\boldsymbol{x}^{(k)} \overset{\text{m.s.}}{\to} \boldsymbol{x}$ in the mean-squared sense, assuming that the training of $(G_{\theta}^{(k)}, H_{\eta}^{(k)})$ based on (3) converges globally for every k.

IV. PROBABILISTIC FORECASTING GPT WITH INNOVATIONS REPRESENTATION

With the time series GPT discussed in the previous section, this section focuses on generative probabilistic forecasting of the time series at future instances. For notational simplicity, we consider only the prediction of a scalar random process (X_t) . Here, a capital letter X denotes a random variable and its lower case x the realization.

A. Generative Probabilistic Forecasting

Given observation $\mathbf{x}_{0:t} = (x_0, \dots, x_t)$ of a time series (X_t) up to time t, the classical T-step-ahead point forecasting is to produce an estimate $\hat{x}_{t+T} = f(\mathbf{x}_{0:t})$ of x_{t+T} . A conventional approach is to find the estimator f that minimizes the conditional (and unconditional) mean-squared or mean absolute error. Probabilistic Forecasting, in contrast, is to obtain an estimate $\hat{F}_{t+T|t}$ of the conditional distribution of

$$F_{t+T|t}(x|\boldsymbol{x}_{0:t}) := \Pr(X_{t+T} \le x|\boldsymbol{X}_{0:t} = \boldsymbol{x}_{0:t}).$$

Once the conditional distribution $F_{t+T|t}$ is estimated, point estimates can be computed, with the minimum mean-squared-error (MMSE) being the conditional mean and the minimum mean-absolute-error (MMAE) the conditional median.

Probabilistic forecasting is exceedingly difficult without parameterizing the distribution; standard techniques of nonparametric distribution estimation cannot be applied, because $F_{t+T|t}$ is a function of the past $x_{0:t}$. Given the trajectory $x_{0:t}$ of the time series up to time t, there is a single realization x_{t+T} associated with the observed history $x_{0:t}$, making it intractable to estimate $F_{t+T|t}$. A standard approach is to assume that the time series can be parameterized by a finite-dimensional parameter θ , say, the mean and variance of a stationary Gaussian process.

Generative Probabilistic Forecasting (GPF) is to produce realizations with the conditional distribution $F_{t_{t+T}|t}$. Instead of producing a single "best estimate" of the realized X_{t+T} , a GPF is to produce an ensemble of realizations of X_{t+T} given the past data $x_{0:t}$. In a way, GPF and probabilistic forecasting are estimating the same object—the conditional distribution $F_{t+T|t}$. Having obtained a probabilistic forecast $\hat{F}_{t+T|t}$, one can produce an ensemble of realizations via Monte Carlo simulations. On the other hand, having obtained a GPF, one can produce a distribution estimate by standard nonparametric techniques by applying the law of large numbers. A key difference between probabilistic forecasting and GPF is that, as we show next, obtaining a GPF from data is considerably simpler.

B. Structure of a Generative Probabilistic Forecaster

As shown in Fig. 4, given $X_{0:t} = x_{0:t}$, a generative probabilistic forecaster (GPF) is a randomized mapping K that maps past observations $x_{0:t}$ and a randomization vector $Z_t = (Z_{t+1}, \dots, Z_{t+T})$ to random variable \tilde{X}_{t+T} having the same conditional distribution as X_{t+T} , *i.e.*,

$$\tilde{X}_{t+T} = K(\boldsymbol{x}_{0:t}, \boldsymbol{Z}_t) \sim F_{t+T|t},$$

where the randomization vector $\mathbf{Z}_t = (Z_{t+1}, \cdots, Z_{t+T})$ follows some sampling distribution $F_{\mathbf{Z}_t | \mathbf{X}_{0:t}}$. Thus, the design of GPF is to find a mapping K and a sampling distribution $F_{\mathbf{Z}_t | \mathbf{X}_{0:t}}$.

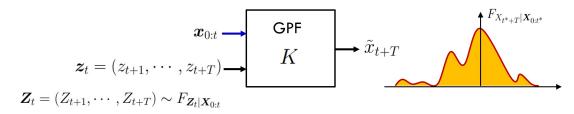


Fig. 4: Structure of Generative Probabilistic Forecaster (GPF).

A trivial but impractical GPF is to choose the sampling probability distribution directly as $F_{X_{t+T}|X_{0:t}}$, which is unknown, unfortunately. On the other hand, the innovations representation model immediately suggests a generative probabilistic forecaster, thanks to the IID-uniform nature of the innovations sequence that decouples the future from the current and the past. The lack of future realizations of the innovation sequence can be replaced by independently generated pseudo-innovations that are IID-uniform, as shown in Fig. 5. Specifically, given the realized time series $\mathbf{x}_{0:t}$ up to time t, the generative forecast \tilde{x}_{t+T} of x_{t+T} is given by

$$\tilde{x}_{t+T} = H_n(\boldsymbol{v}_{0:T}, \tilde{\boldsymbol{v}}_{t+1:t+T}), \tag{5}$$

where $v_{0:T} = (v_0, \dots, v_t)$ is the innovations sequence of $x_{0:t}$ and $\tilde{v}_{t+1:t+T} = (v_{t+1}, \dots, v_{t+T})$ the pseudo innovations of unrealized time series $x_{t:t+T}$.

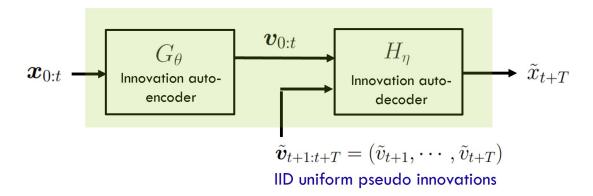


Fig. 5: Generative Probabilistic Forecaster (GPF).

The validity of the above construction can be established as follows. Here we assume the ideal training of the weak innovation autoencoder (G_{θ}, H_{η}) , such that the latent innovations sequence $\mathbf{V} = (V_t)$ is IID-uniform, and the input and output of the autoencoder match in distribution. In particular have

$$\Pr\left[X_{t+T} \le x \middle| \boldsymbol{X}_{0:t} = \boldsymbol{x}_{0:t}\right] = \Pr\left[\hat{X}_{t+T} \le x \middle| \boldsymbol{X}_{0:t} = \boldsymbol{x}_{0:t}\right]$$

Assume that the ideally trained autoencoder G_{θ} is injective—a theoretically limiting but practically reasonable for deep neural network representations. Let $\tilde{V}_t = (\tilde{V}_{t+1}, \cdots, \tilde{V}_{t+T})$ be the peseudo-innovations vector independent of the weak innovations $V = (V_t)$ and $\tilde{X}_{t+T} = H(V_{0:t}, \tilde{V}_t)$ the output of the autodecoder with innovations vector $V_{t+1:t+T}$ replaced by \tilde{V}_t . Then,

$$\Pr\left[\hat{X}_{t+T} \leq x \middle| \mathbf{X}_{0:t} = \mathbf{x}_{0:t}\right] = \Pr\left[\hat{X}_{t+T} \leq x \middle| \mathbf{V}_{0:t} = \mathbf{v}_{0:t}\right]$$

$$= \Pr\left[H_{\eta}(\mathbf{v}_{0:t}, \mathbf{V}_{t+1:t+T}) \leq x \middle| \mathbf{V}_{0:t} = \mathbf{v}_{0:t}\right]$$

$$= \Pr\left[H_{\eta}(\mathbf{v}_{0:t}, \tilde{\mathbf{V}}_{t+1:t+T}) \leq x \middle| \mathbf{V}_{0:t} = \mathbf{v}_{0:t}\right]$$

$$= \Pr\left[\tilde{X}_{t+T} \leq x \middle| \mathbf{V}_{0:t} = \mathbf{v}_{0:t}\right],$$
(6)

where we used the fact that $(V_{0:t}, V_{t+1:t+T}, \tilde{V}_t)$ are jointly independent and $(V_{0:t}, V_{t+1:t+T}) \stackrel{\text{d}}{=} (V_{0:t}, \tilde{V}_t)$. A few remarks on the above derivation are in order.

- Bayesian Sufficency: Equation (7) implies that the weak innovation sequence is Bayesian sufficient for decision involving any random variable X_{t+T} at a future time, in the sense that conditioning on innovations up to time t incurrs no loss comparing to conditioning on the raw data $X_{0:t}$.
- Validity of GPT: With (7), we show that the GPF shown in Fig. 5 gives samples with the correct conditional probability distribution, under the ideal pretraining of the innovations autoencoder. Note that the above GPF can easily generate an arbitrarily large number of independent samples, from which conditional probability distributions can be estimated using standard distribution estimation techniques.

C. From GPF to Point and Quantile Forecasting

GPF can produce an arbitrarily large number of independently generated samples of the conditional probability distribution, from which point and quantile forecasts can be readily computed.

Let $\left\{\tilde{x}_t^{(k)}, k=1,\cdots,K\right\}$ be the set of K independently generated samples (by a GPF) following the conditional probability distribution $F_{t+T|t}$ of the time series for X_{t+T} given past observations $\boldsymbol{x}_{0:t}$ up to time t. For the simplicity of mathematical expressions, we assume that $\left\{\tilde{\boldsymbol{x}}_t^{(k)}\right\}$ is sorted in an ascending order.

Some of the most popular point estimates are computed as follows:

• Minimum Mean-Squared-Error (MMSE) Forecast: The MMSE forecast is the mean of the conditional distribution. The MMSE forecast \hat{x}_t^{MMSE} by a GPF is given by the conditional sample mean

$$\hat{\boldsymbol{x}}_t^{\text{MMSE}} = \frac{1}{K} \sum_{k=1}^K \tilde{\boldsymbol{x}}_t^{(k)}.$$

• Minimum Mean-Absolute-Error (MMAE) Forecast: The MMAE forecast is the median of the conditional distribution. The MMAE forecast \hat{x}_t^{MMAE} by a GPF is given by the conditional sample median

$$\hat{\boldsymbol{x}}_t^{\text{MMAE}} = \begin{cases} \tilde{\boldsymbol{x}}_t^{((K+1)/2)}, & \text{if } K \text{ is odd} \\ 0.5 \left(\tilde{\boldsymbol{x}}_t^{(K/2)} + \tilde{\boldsymbol{x}}_t^{(K/2+1)} \right), & \text{if } K \text{ is even.} \end{cases}$$

• Quantile-Forecast: The forecast of q-quantile \hat{x}_t^{q-QT} is given by:

$$\hat{\boldsymbol{x}}_t^{q\text{-QT}} = \begin{cases} \tilde{\boldsymbol{x}}_t^{(qK)}, & \text{if } qK \text{ is an integer} \\ 0.5 \left(\tilde{\boldsymbol{x}}_t^{([qK])} + \tilde{\boldsymbol{x}}_t^{([qK]+1)} \right), & \text{otherwise}, \end{cases}$$

where [a] indicates the greatest integer not exceeding a.

D. Performance Measure: CRPS, CPE and ACPE

Evaluating the performance of probabilistic forecasting and GPF is nontrivial. The difficulty lies in that the ground truth of the forecast quantity—the conditional probability distribution $F_{t+T|t}$ of the future random variable X_{t+T} given the past $x_{0:t}$ —is unknown.

a) Continuous Ranked Probability Score (CRPS): A commonly used probabilistic forecasting metric is the Continuous Ranked Probability Score (CRPS). Let $\hat{F}_{t+T|t}$ be the estimated conditional CDF of X_{t+T} given $X_{0:t} = x_{0:t}$ and x_{t+T} the realized X_{t+T} . CRPS and the expected CRPS are defined by

$$\begin{aligned} \mathsf{CRPS}\big(\hat{F}_{t+T|t}, x_{t+T}\big) &:= \int_{-\infty}^{\infty} \Big(\hat{F}_{t+T|t}(z|\boldsymbol{x}_{0:t}) - \mathbb{I}\{x_{t+T} \leq z\}\Big)^2 dz \\ \overline{\mathsf{CRPS}}(\hat{F}_{t+T|t}) &:= \mathbb{E}\Big[\mathsf{CRPS}(\hat{F}_{t+T|t}, X_{t+T})\Big], \end{aligned}$$

where \mathbb{I} is the indicator function. It can be easily shown that $\overline{\mathsf{CRPS}}(\hat{F}_{t+T|t})$ is minimum (albeit not necessarily zero) when $\hat{F}_{t+T|t} = F_{t+T|t}$.

In practice, given a realization of the time series $(X_t = x_t)$ of length N, we approximate $\overline{\mathsf{CRPS}}(\hat{F}_{t+T|t})$ by the empirical sum

$$\widehat{\mathsf{CRPS}} = \frac{1}{N-T} \sum_{t=T-1}^{N} \int_{-\infty}^{\infty} \left(\hat{F}_{t|t-T}(z|\boldsymbol{x}_{0:t-T}) - \mathbb{I}\{x_t \leq z\} \right)^2 dz.$$

For generative probability forecasting that produces only generative forecast samples $\hat{\mathcal{X}} := \{\hat{x}_k\}, \hat{F}_{t+T|t}$ is estimated from $\hat{\mathcal{X}}$.

b) Coverage Probability Error (CPE): While CRPS measures the accuracy of a GPF over the domain of $F_{t+T|t}$, the Coverage Probability Error at level α , denoted by CPE_{α} , evaluates the accuracy of the predicted $\alpha \times 100\%$ coverage interval³ based on $\hat{F}_{t+T|t}$.

Let $[L_{t+T}^{\alpha}, U_{t+T}^{\alpha}]$ be the (random conditional) α -coverage interval of X_{t+T} under the ground truth conditional distribution $F_{t+T|t}$, *i.e.*,

$$\Pr\left(X_{t+T} \in [L_{t+T}^{\alpha}, U_{t+T}^{\alpha}] \middle| \boldsymbol{X}_{0:t} = \boldsymbol{x}_{0:t}\right) = \alpha.$$

Let $[\hat{L}_{t+T}^{\alpha}, \hat{U}_{t+T}^{\alpha}]$ be the *predicted* conditional α coverage interval of X_{t+T} under the predicted conditional distribution $\hat{F}_{t+T|t}$. The *Coverage Probability Error (CPE)* is defined by

$$\mathsf{CPE}_{\alpha}(\hat{F}_{t+T|t}) := \mathbb{E}\Big[\Pr\Big(X_{t+T} \in [\hat{L}_{t+T}^{\alpha}, \hat{U}_{t+T}^{\alpha}] \Big| \boldsymbol{X}_{0:t}\Big) - \alpha\Big],$$

which is zero if $\hat{F}_{t+T|t} = F_{t+T|t}$. The Absolute CPE at the coverage level α is defined by

$$\mathsf{ACPE}_{lpha}(\hat{F}_{t+T|t}) := \Big| \mathsf{CPE}_{lpha}(\hat{F}_{t+T|t}) \Big|.$$

Given the realizations of the time series and the sequence of predicted α -coverage intervals $\{[\hat{l}_{t+T}^{\alpha}, \hat{u}_{t+T}^{\alpha}], t=0:N-T\}$, the CPE and ACPE of a GPF are approximated by sample averages. In particular, the estimated ACPE $\overrightarrow{ACPE}_{\alpha}$ is computed by

$$\widehat{\mathsf{ACPE}}_{\alpha} := \left| \frac{1}{N-T} \sum_{t=T}^{N} \mathbb{1} \{ x_t \in [\hat{l}_{t-T}^{\alpha}, \hat{u}_{t-T}^{\alpha}] \} - \alpha \right|,$$

where the predicted coverage bounds $(\hat{l}_{t-T}^{\alpha}, \hat{u}_{t-T}^{\alpha})$ are computed from the generated samples $\{\tilde{x}_{t+T}^{(k)}, k = 1, \cdots, K\}$ by the generative forecaster.

V. APPLICATION: GPF OF REAL-TIME ELECTRICITY PRICES

We illustrate the application of the innovations representation model-based FM for generative probabilistic forecasting of real-time prices of electricity using published LMP data by the New York Independent System Operator (NYISO) to train the FM and evaluate to the forecasting performance. More extensive empirical study results can be found in [27].

³Typically, the α coverage interval is the symmetrical interval with respect to the mean with probability α .

a) Locational Marginal Price (LMP): In deregulated electricity markets in the U.S., electricity prices are referred to as locational marginal prices (LMP), set every five minutes based on bids from demands, offers from supplies, and locations of trades. Unlike time series of physical processes such as wind, solar, and aggregated electricity consumption, LMPs are computed from dual variables associated with power flow constraints of a convex optimization. Fig. 6 shows the actual electricity prices at the Long Island (LONGIL) and New York City (NYC) buses on July 26, 2023.

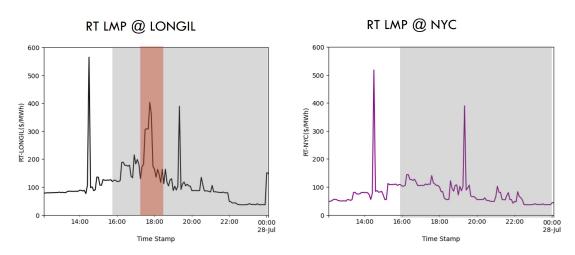


Fig. 6: Locational Marginal Prices at the Long Island and New York City buses on July 26, 2027.

Unlike time series of physical processes such as wind, solar, and aggregated electricity consumption, LMPs can be highly volatile with huge spikes that could jump hundreds, even thousands, of dollars per megawatt-hour. The volatility of LMP makes probabilistic price forecasting an essential risk-mitigating tool.

b) Dataset and Forecasting Settings: A typical deregulated real-time electricity market operates at the 5-minute timescale. NYISO publishes decades of 5-minute real-time LMPs for all its buses.

The setup of the GPF of real-time LMP follows the participation requirement of real-time electricity market, where bids and offers must be submitted 30 to 75 minutes ahead of the market clearing. We therefore set the forecast horizon T=12 (60 minutes). We assume that the FM is updated weakly using the past 30 days of LMP data.

c) Empirical Evaluations: Extensive performance evaluations are reported in [27]. Here we present a sample of results based on the 2023-2024 LMP and demand traces, where we selected four months, one in each season (July and October in 2023, and January and April in 2024) for evaluation. We choose the Long Island bus (LONGIL) for evaluating the probabilistic forecasting performance of 60-minute ahead forecaster based on X-minute past LMP values at LONGI and NYC and the day-ahead LMP at LONGIL and system demand.

We compared the GPF with the weak innovations autoencoder forecaster, abbreviated by WIAE and defined in (5), with several leading GPF benchmarks: TLAE [28], DeepVAR [29], and the more recent large language model (LLM) based BWGVT [30]. Fig. 7 shows the 60-minute ahead probabilistic forecasting

performance under CRPS and 50%-CPE metric, where WIAE demonstrated superior performance for most test cases.

	Summer (July)		Fall (October)		Winter (January)		Spring (April)	
	CRPS	ACPE (50%)	CRPS	ACPE (50%)	CRPS	ACPE (50%)	CRPS	ACPE (50%)
WIAE	13.2607 (1)	0.0264(1)	10.9462 (1)	0.0390 (2)	12.1949 (1)	0.0358 (2)	5.3111 (1)	0.0208 (1)
TLAE	23.3288 (4)	0.1492 (3)	19.1394 (4)	0.0395 (3)	20.7468 (4)	0.0369 (3)	15.3330 (4)	0.0273 (2)
DeepVar	14.0716 (2)	0.0495 (2)	12.6109 (2)	0.0328 (1)	13.1403 (2)	0.0160 (1)	5.9420 (2)	0.0423 (3)
BWGVT	15.0476 (3)	0.1646 (4)	14.1423 (3)	0.1763 (4)	15.2919 (3)	0.1804 (4)	14.6103 (3)	0.2436 (4)

Fig. 7: Performance comparisons of GPF benchmarks. Boldface numbers are the best scores. The numbers in brackets are rankings.

VI. CONCLUSION

Through the lens of modern machine learning and AI foundation models, this article aims to connect modern machine learning and AI methodology with the classical innovations representation theory and powerful model-based practical solutions such as Wiener and Kalman filtering, matched filtering, and optimal control, leading to data-driven approaches that are grounded sound mathematical principles. To this end, the innovations representation model can be a powerful engine for AI foundation models for time series applications.

REFERENCES

- [1] A. Smith, The Wealth of Nations, L. von Mises, Ed. Martino Publishing, 2015, originally published in 1776.
- [2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS). Curran Associates Inc., 2017, pp. 5998–6008. [Online]. Available: https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
- [3] R. Bommasani, D. A. Hudson, E. Adeli, et al., "On the opportunities and risks of foundation models," 2022. [Online]. Available: https://arxiv.org/abs/2108.07258
- [4] N. Wiener, *Nonlinear Problems in Random Theory*, ser. M.I.T. paperback series. Cambridge, MA: MIT Press, 1958. [Online]. Available: https://books.google.com/books?id=HQBRAAAAMAAJ
- [5] N. Wiener and P. Masani, "The prediction theory of multivariate stochastic processes: I. The regularity condition," *Acta Mathematica*, vol. 98, no. none, pp. 111 150, 1957. [Online]. Available: https://doi.org/10.1007/BF02404472
- [6] —, "The prediction theory of multivariate stochastic processes, II: The linear predictor," *Acta Mathematica*, vol. 99, no. none, pp. 93 137, 1958. [Online]. Available: https://doi.org/10.1007/BF02392423
- [7] P. Masani, "Wiener's contributions to generalized harmonic analysis, prediction theory and filter theory," *Bulletin of the American Mathematical Society*, vol. 72, no. 1.P2, pp. 73 125, 1966.
- [8] A. N. Kolmogorov, "Stationary sequences in Hubert space," in *Selected Works of A. N. Kolmogorov: Volume II Probability Theory and Mathematical Statistics*, A. N. Shiryayev, Ed. Dordrecht: Springer Netherlands, 1992, pp. 228–271.
- [9] N. Wiener, Extrapolation, Interpolation and Smoothing of Stationary Time Series: With Engineering Applications. Cambridge, MA: MIT Press, 1949, based on a report prepared for the Navy Department, Bureau of Ships, during World War II.
- [10] H. Bode and C. Shannon, "A simplified derivation of linear least square smoothing and prediction theory," *Proceedings of the IRE*, vol. 38, no. 4, pp. 417–425, 1950.

- [11] T. Kailath, "An innovations approach to least-squares estimation—part i: Linear filtering in additive white noise," *IEEE Transactions on Automatic Control*, vol. 13, no. 6, pp. 646–655, 1968.
- [12] T. Kailath and P. Frost, "An innovations approach to least-squares estimation—part ii: Linear smoothing in additive white noise," *IEEE Transactions on Automatic Control*, vol. 13, no. 6, pp. 655–660, 1968.
- [13] R. Geesey and T. Kailath, "Applications of the canonical representation to estimation and detection in colored noise," Proceedings of the 19th Polytechnic Institute of Brooklyn Symposium on Computer Processing in Communications, Polytechnic Press, U.S. Air Force Academy, Colorado Springs, CO, Technical Report, 1969.
- [14] T. Kailath, "A general likelihood-ratio formula for random signals in gaussian noise," *IEEE Transactions on Information Theory*, vol. 15, no. 3, pp. 350–361, 1969.
- [15] T. Kailath, "The Innovations Approach to Detection and Estimation Theory," *Proceedings of the IEEE*, vol. 58, no. 5, pp. 680–695, 1970.
- [16] ——, "A View of Three Decades of Linear Filtering Theory," *IEEE Transactions on Information Theory*, vol. 20, no. 2, pp. 146–181, 1974.
- [17] T. Kailath, "Some Extensions of the Innovations Theorem*," Bell System Technical Journal, vol. 50, no. 4, pp. 1487–1494, 1971.
- [18] P. Frost and T. Kailath, "An innovations approach to least-squares estimation—part iii: Nonlinear estimation in white gaussian noise," *IEEE Transactions on Automatic Control*, vol. 16, no. 3, pp. 217–226, 1971.
- [19] M. Rosenblatt, "Stationary processes as shifts of functions of independent random variables," *Journal of Mathematics and Mechanics*, vol. 8, no. 5, pp. 665–681, 1959. [Online]. Available: http://www.jstor.org/stable/24900682
- [20] —, "A comment on a conjecture of n. wiener," *Statistics & Probability Letters*, vol. 79, no. 3, pp. 347–348, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167715208004173
- [21] W. Wu, "Nonlinear system theory: Another look at dependence," *Proceedings of National Academy of Sciences*, vol. 102, no. 40, pp. 14150–14154, 2005. [Online]. Available: https://doi.org/10.1073/pnas.0506715102
- [22] —, "Asymptotic theory for stationary processes," Statistics and Its Interface, vol. 0, pp. 1–20, 01 2011.
- [23] P. Lévy, *Théorie de l'addition des variables aléatoires*, 1st ed., ser. Monographies des probabilités. Paris, France: Gauthier-Villars, 1937, vol. 1.
- [24] X. Wang and L. Tong, "Innovations autoencoder and its application in one-class anomalous sequence detection," *J. Mach. Learn. Res.*, vol. 23, no. 1, 2022.
- [25] C. Villani, The Wasserstein distances. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 93–111. [Online]. Available: https://doi.org/10.1007/978-3-540-71050-9_6
- [26] X. Wang, M. Lee, L. Tong, and Q. Zhao, "Novelty Detection in Time Series via Weak Innovations Representation: A Deep Learning Approach," 2022.
- [27] X. Wang, "Innovation representations and their applications in detection, estimation, and compression," Ph.D. dissertation, Cornell University, Stanford, California, 2024, ph.D. dissertation.
- [28] N. Nguyen and B. Quanz, "Temporal latent auto-encoder: A method for probabilistic multivariate time series forecasting," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, pp. 9117–9125, May 2021, number: 10.
- [29] D. Salinas, M. Bohlke-Schneider, L. Callot, R. Medico, and J. Gasthaus, "High-dimensional multivariate forecasting with low-rank gaussian copula processes," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.
- [30] J. Bottieau, Y. Wang, Z. De Grève, F. Vallée, and J.-F. Toubeau, "Interpretable transformer model for capturing regime switching effects of real-time electricity prices," *IEEE Transactions on Power Systems*, vol. 38, no. 3, pp. 2162–2176, 2023.