Characteristic Functionals and the Innovations Approach to Stochastic Processes (With Applications to Random Neural Networks)

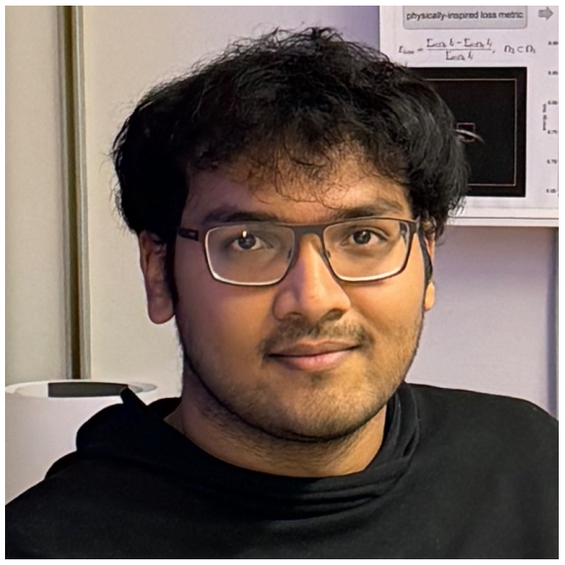
Rahul Parhi
UCSD ECE

UCSD Probability Seminar
6 February 2025

Outline

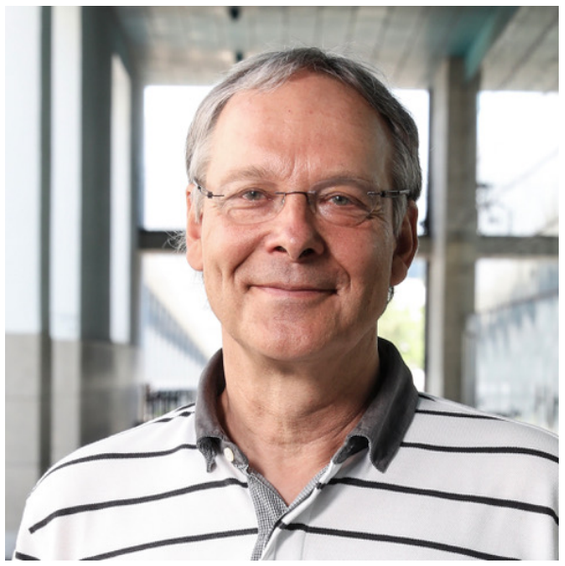
- Classical Formulation of Lévy Processes
- Innovation Model (Bode, Shannon, and Kailath, *ca.* 1950–1970)
- (Lévy) White Noises
- Generalized Stochastic Processes (Itô and Gelfand, *ca.* 1955)
- Generalized Lévy Processes and Stochastic Differential Equations
- Exact Characterization of the Law of Random Neural Networks

Joint Work With...



Pakshal Bohra

Mehrsa Pourya



Michael Unser

Ayoub El Biari

Classical Formulation of Lévy Processes

Definition (Lévy process)

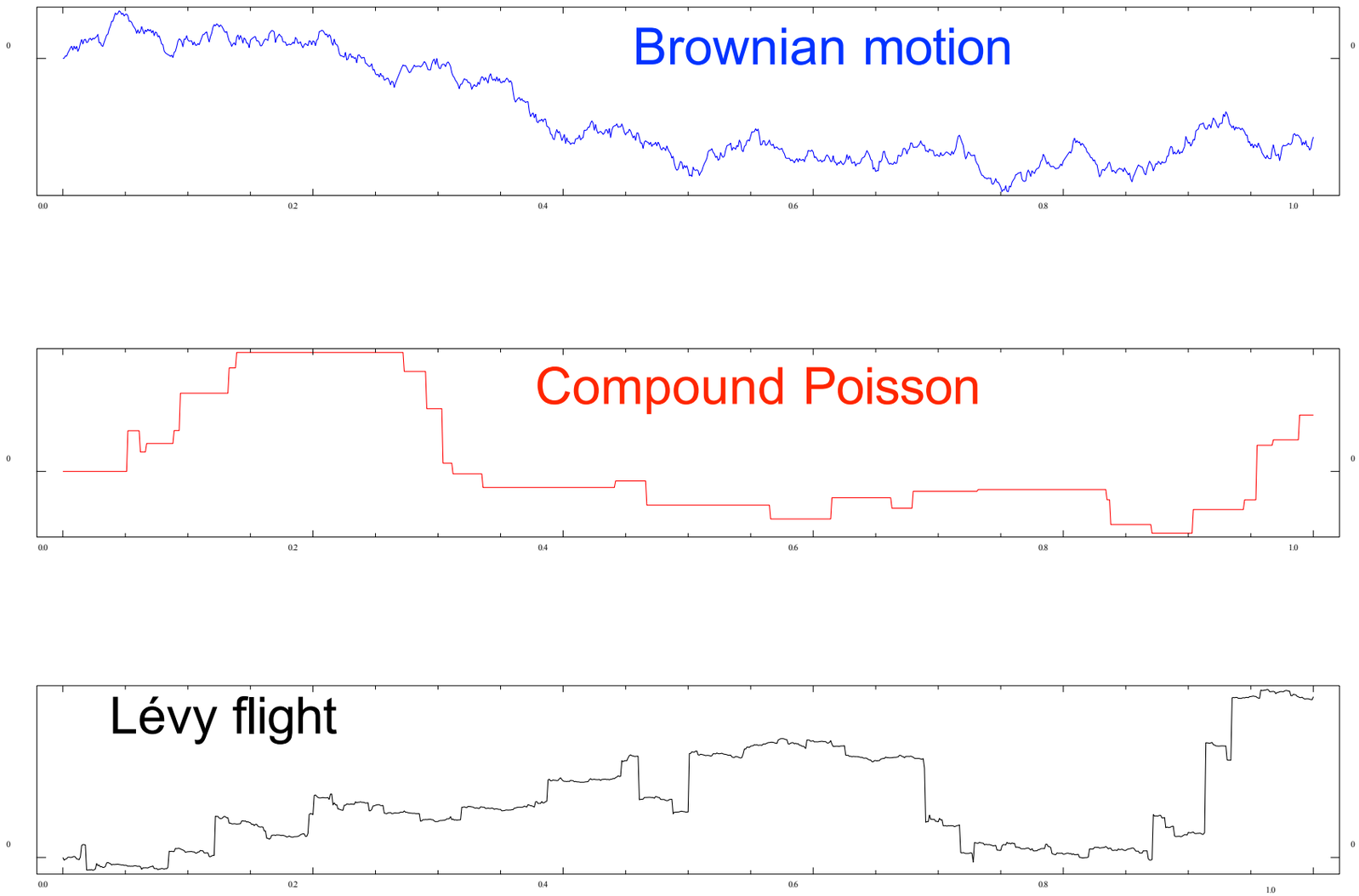
A stochastic process $(S_t)_{t\geq 0}$ is called a *Lévy process* if

1. $S_0 = 0$ almost surely;
2. Independence of increments: For any $0 \leq t_1 < \cdots < t_n < \infty$, $X_{t_2} - X_{t_1}, \cdots, X_{t_n} - X_{t_{n-1}}$ are mutually independent;
3. Stationary increments: For any $t_1 < t_2$, $S_{t_2} - S_{t_1} \overset{\mathscr{L}}{=} S_{t_1 - t_2}$;
4. Stochastic continuity: For any $\varepsilon > 0$ and $t \geq 0$

$$\lim_{h \to 0} \mathbf{P}(|S_{t+h} - S_t| > \varepsilon) = 0.$$

Lévy process are càdlàg and tightly linked to infinite divisibility

Examples of Lévy Processes



Wiener 1923

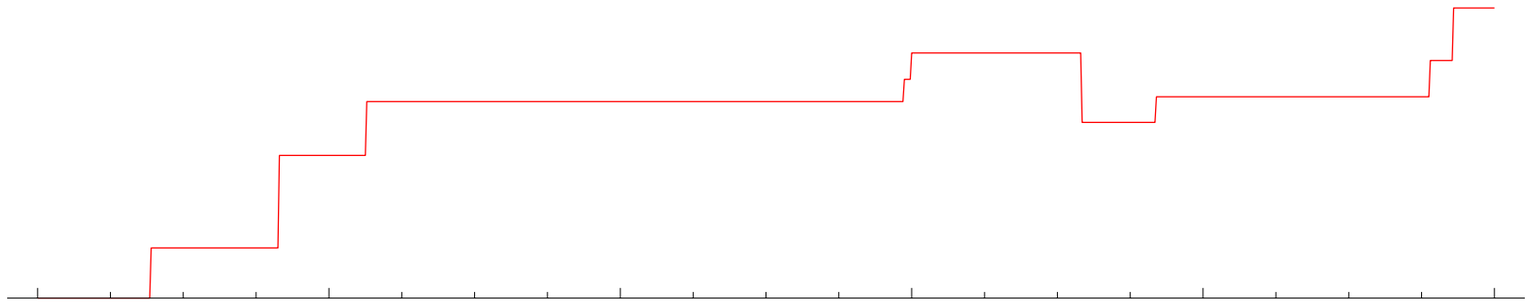
Lévy process are often used to model various components of signal processing and communication systems

Lévy, *ca.* 1930

Decoupling the Underlying Randomness

Is there a way to decouple the correlation properties of a Lévy process from its underlying randomness?

Example: Compound Poisson process $s(t)$ (rate $\lambda > 0$ and jump law $\mathbf{P}_V$)



$$\mathrm{D}s(t) = \dot{s}(t) = \sum_k v_k \, \delta_{b_k}$$

innovation

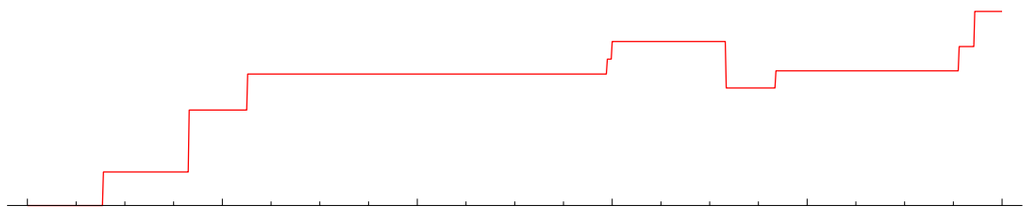
$v_k \sim \mathbf{P}_V$ i.i.d.

$\{b_k\}_k$ is a Poisson point process with rate λ

The underlying randomness of a compound Poisson process is completely determined by its innovation process

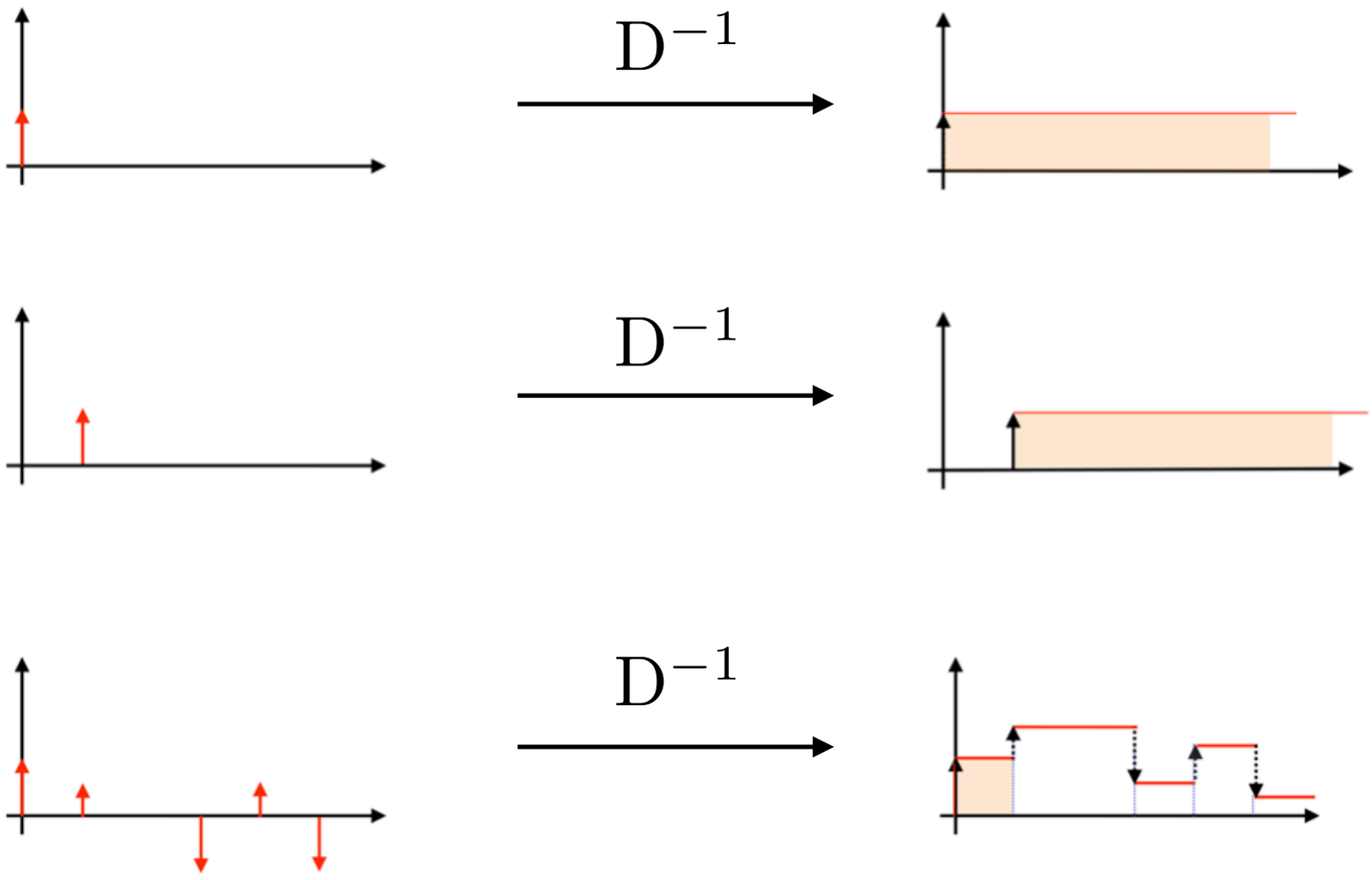
Innovation-Based Synthesis

$$\sum_k v_k\,\delta(\cdot - b_k) \quad \underset{\mathrm{D}}{\overset{???}{\rightleftarrows}}$$



Integrate the innovation process?

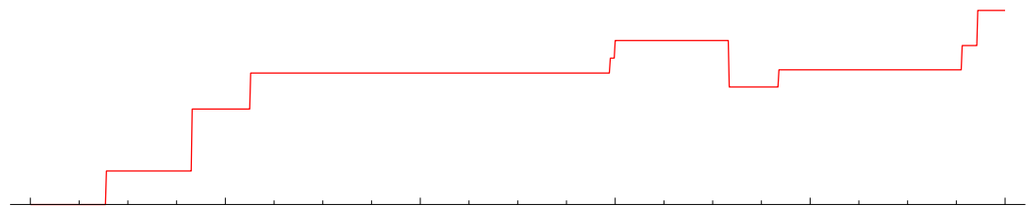
Recall: $s(0) = 0$ a.s.



Compound Poisson Process

- SDE formulation

$$\mathrm{D}s \overset{\mathscr{L}}{=} w \quad \text{s.t.} \quad s(0) = 0$$



Innovation: $w = \sum_k v_k \, \delta(\cdot - b_k)$

Poisson innovation, or
Poisson white noise

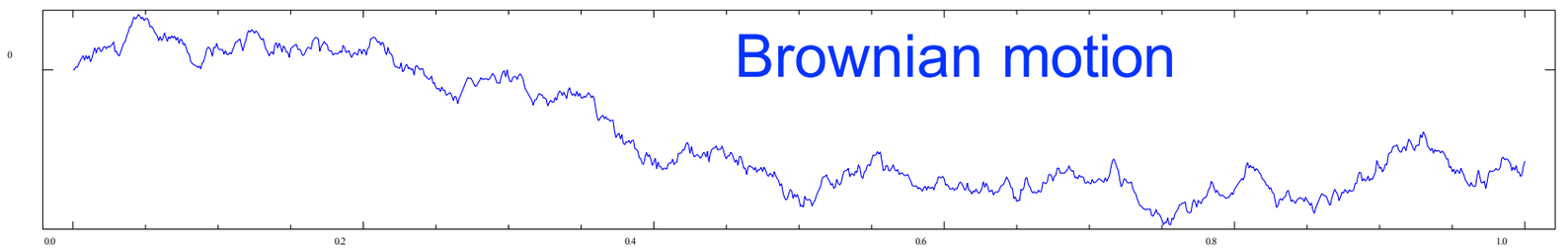
- "Formal" solution

$$s(t) = \mathrm{D}_0^{-1} w(t) = \sum_k v_k \, \mathrm{D}_0^{-1}\{\delta(\cdot - b_k)\}(t)$$

$$= \sum_k v_k \Big(H(t - b_k) - H(-b_k) \Big)$$

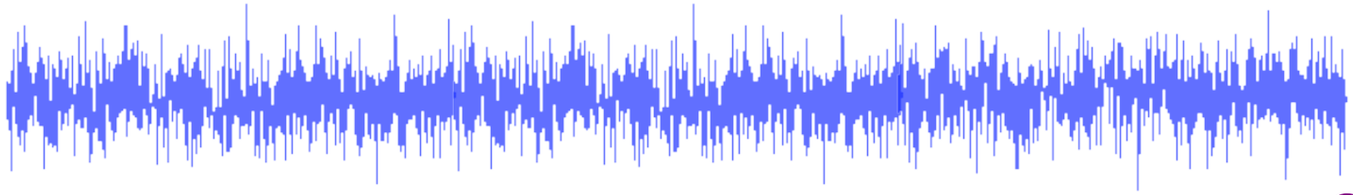
imposing boundary condition

Brownian Motion and Gaussian White Noise



$b(t)$

D



$w(t)$

Gaussian innovation, or
Gaussian white noise

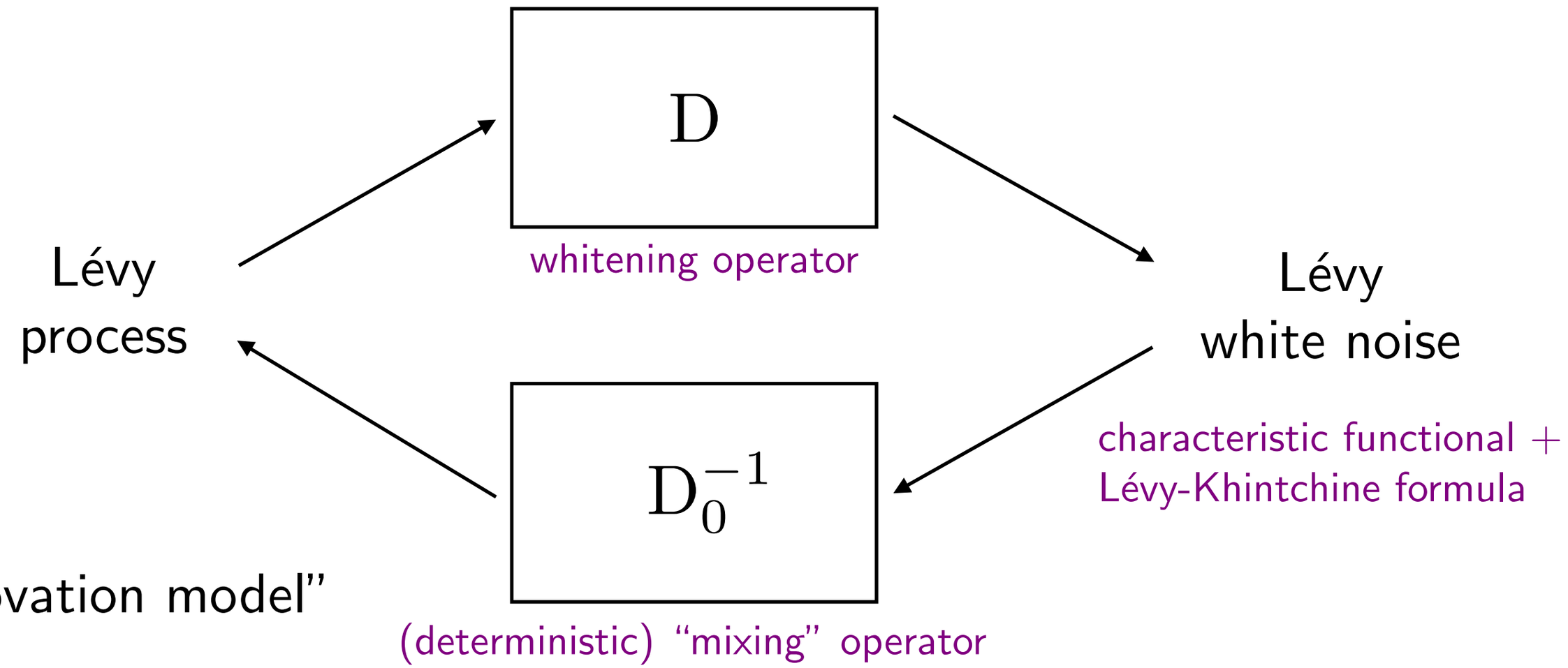
Gaussian white noise does not admit a pointwise representation

$w(t)$ is a random measure, or,
a random distribution (generalized function)

$$\mathbf{P}(w \in \mathcal{D}'(\mathbb{R})) = 1$$
$$\mathbf{P}(w \in \mathcal{S}'(\mathbb{R})) = 1$$

(Gelfand, 1955)

Innovations Approach to Lévy Processes



"innovation model"

$$\text{mixing} = \text{whitening}^{-1}$$

It is much easier to study white noise processes than it is to study general random processes.

Bode and Shannon (1950); Kailath (1970s)

Generalized Stochastic Processes

Kiyosi Itô

Israel Gelfand

Framework of generalized stochastic processes

ca. 1950s

Stochastic counterpart to Schwartz' theory of distributions

MEMOIRS OF THE COLLEGE OF SCIENCE, UNIVERSITY OF KYOTO, SERIES A
Vol. XXVIII, Mathematics No. 3, 1953.

Stationary random distributions

By

Kiyosi ITÔ

(Received April 15, 1954)

In the same way as the concept of distributions by L. Schwartz [11][1] was introduced as an extended one of functions, we may define stationary random distributions as an extension of stationary random functions viz. stationary processes. Such consideration will enable us to establish a unified theory of stationary processes, Brownian motion processes, processes with stationary increments and other

1.

Generalized random processes

Dokl. Akad. Nauk SSSR **100** (1955) 853–856. Zbl. **64**:111

1. Usually, a random process is defined by probability distributions of random variables $(x(t_1), \ldots, x(t_n))$ for n arbitrary moments of time. However, one can give many examples of random processes which are important in practice for which such probability distributions do not exist. An example of such a process is the white noise that can be obtained, roughly speaking, as a superposition of all frequencies with random amplitudes; these amplitudes are independent, identically distributed Gaussian variables. In this note we introduce a description of random processes that appears to cover all practically important examples. If $x(t)$ is a random process, then any "linear apparatus" gives us a probability distribution

Warm-Up: Classical Probability Theory

$\boldsymbol{X} \in \mathbb{R}^d$ is a random vector with law $\mathbf{P}_{\boldsymbol{X}}$

This means that

$$\boldsymbol{X} : \underbrace{(\Omega, \mathcal{F}, \mathbf{P})}_{\substack{\text{complete} \\ \text{probability space}}} \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \text{ is measurable.}$$

The law of $\boldsymbol{X}$ is given by the *pushforward measure*

$$\mathbf{P}_{\boldsymbol{X}}(A) = (\boldsymbol{X}_\sharp \mathbf{P})(A) := \mathbf{P}(\boldsymbol{X}^{-1}(A)) = \mathbf{P}(\{\omega \in \Omega \,:\, \boldsymbol{X}(\omega) \in A\})$$
$$= \mathbf{P}(\boldsymbol{X} \in A).$$

The *characteristic function* of $\boldsymbol{X}$ is given by

$$\widehat{\mathbf{P}}_{\boldsymbol{X}}(\boldsymbol{\xi}) = \mathbf{E}[\mathrm{e}^{\mathrm{i}\boldsymbol{X}^\mathsf{T}\boldsymbol{\xi}}], \quad \boldsymbol{\xi} \in \mathbb{R}^d.$$

Once you have the characteristic function, you have everything.

Bochner's and Lévy's Theorem

Bochner's Theorem

A function $\widehat{\mathbf{P}}$ is the characteristic function of a random variable $\boldsymbol{X} \in \mathbb{R}^d$ if and only if $\widehat{\mathbf{P}}$ is continuous, positive-definite, and satisfies $\widehat{\mathbf{P}}(\mathbf{0}) = 1$.

Lévy's Continuity Theorem

Let $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ and $\boldsymbol{X}$ be random variables in $\mathbb{R}^d$. The sequence $\boldsymbol{X}_n$ converges in law to $\boldsymbol{X}$ if and only if, for all $\boldsymbol{\xi} \in \mathbb{R}^d$,

$$\widehat{\mathbf{P}}_{\boldsymbol{X}_n}(\boldsymbol{\xi}) \xrightarrow[n\to\infty]{} \widehat{\mathbf{P}}_{\boldsymbol{X}}(\boldsymbol{\xi}).$$

Generalized Stochastic Processes

A generalized stochastic process is a random variable that takes values in the dual of a nuclear space.

Let $(\mathcal{N}, \mathcal{N}')$ denote a nuclear space and its dual. A generalized stochastic process s is a random variable, i.e., a measurable map

$$s : (\Omega, \mathcal{F}, \mathbf{P}) \to (\mathcal{N}', \mathcal{B}_c(\mathcal{N}')).$$

The law of s is the probability measure $\mathbf{P}_s := s_\sharp \mathbf{P}$, which is defined on $\mathcal{B}_c(\mathcal{N}')$. The *characteristic functional* of s is given by

$$\widehat{\mathbf{P}}_s(\varphi) = \mathbf{E}[\mathrm{e}^{\mathrm{i}\langle s, \varphi \rangle_{\mathcal{N}' \times \mathcal{N}}}], \quad \varphi \in \mathcal{N}.$$

Examples: $(\mathbb{R}^d, \mathbb{R}^d)$, $(\mathcal{D}(\mathbb{R}^d), \mathcal{D}'(\mathbb{R}^d))$, $(\mathcal{S}(\mathbb{R}^d), \mathcal{S}'(\mathbb{R}^d))$, etc.

Once you have the characteristic functional, you have everything.

Generalized Stochastic Processes

This formulation allows us to study classical stochastic processes $(s(\boldsymbol{x}))_{\boldsymbol{x}\in\mathbb{R}^d}$ as well as those that do not admit a pointwise representation such as white noise.

Example (Gaussian Processes)

A generalized stochastic process s that takes values in $\mathcal{N}'$ is called *Gaussian* if its characteristic functional is of the form

$$\widehat{\mathbf{P}}_s(\varphi) = \exp\left(\mathrm{i}\mu_s(\varphi) - \frac{1}{2}\Sigma_s(\varphi,\varphi)\right), \quad \varphi\in\mathcal{N},$$

where $\mu_s(\varphi) = \mathbf{E}[\langle s,\varphi\rangle]$ denotes the mean functional and $\Sigma_s(\varphi_1,\varphi_2) = \mathbf{E}[(\langle s,\varphi_1\rangle - \mu_s(\varphi_1))(\langle s,\varphi_2\rangle - \mu_s(\varphi_2))]$ denotes the covariance functional of the process.

$\Longrightarrow$ Backwards compatible with space-indexed Gaussian processes.

Hida and Ikeda (1967); Duttweiler and Kailath (1973)

Is Nuclearity Necessary?

Bochner–Minlos Theorem

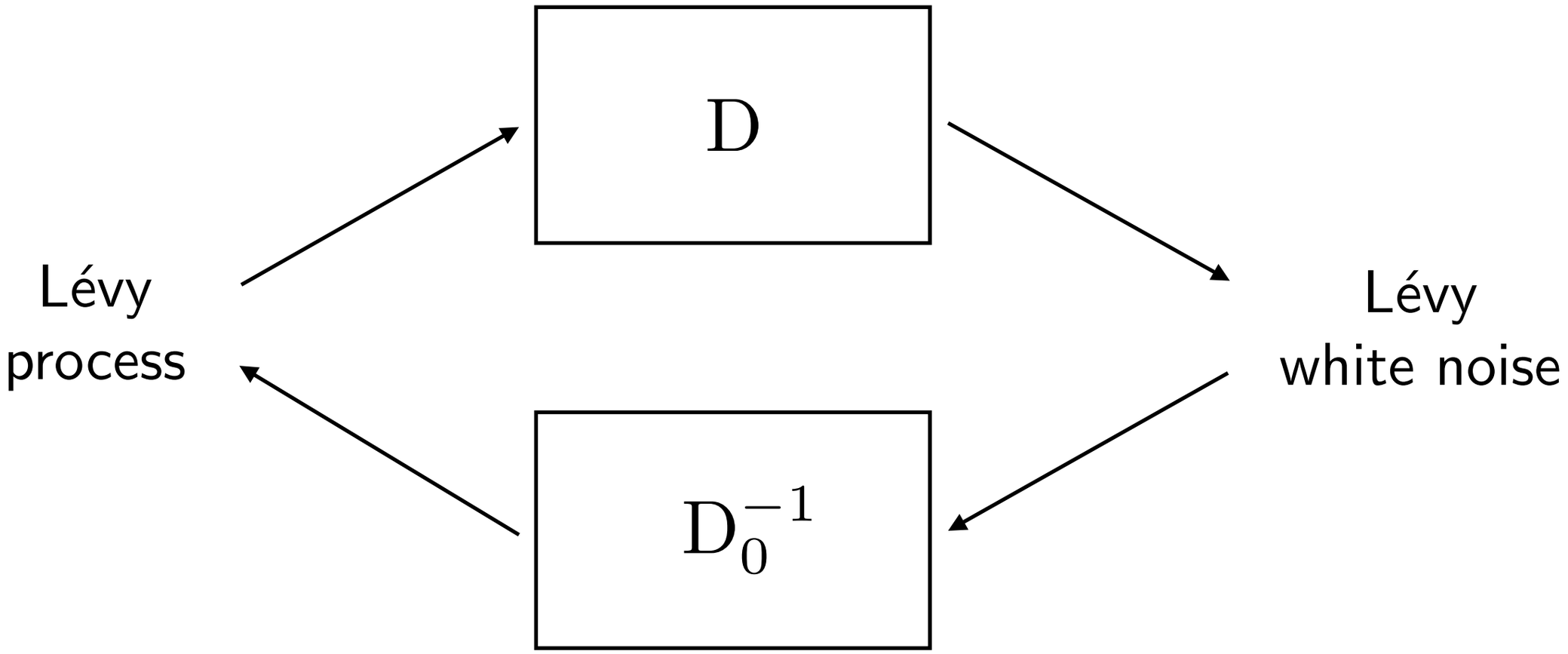
A function $\widehat{\mathbf{P}}$ is the characteristic function of a stochastic process $s \in \mathcal{N}'$ if and only if $\widehat{\mathbf{P}}$ is continuous, positive-definite, and satisfies $\widehat{\mathbf{P}}(0) = 1$.

Lévy–Fernique Continuity Theorem

Let $(s_n)_{n\in\mathbb{N}}$ and s be stochastic processes in $\mathcal{N}'$. The sequence s_n converges in law to s if and only if, for all $\varphi \in \mathcal{N}$,

$$\widehat{\mathbf{P}}_{s_n}(\varphi) \xrightarrow[n\to\infty]{} \widehat{\mathbf{P}}_{s}(\varphi).$$

Classical Lévy Processes: A New Perspective



- SDE formulation

$$\mathrm{D}s \overset{\mathscr{L}}{=} w \quad \text{s.t.} \quad s(0) = 0$$

- "Formal" solution

$$s = \mathrm{D}_0^{-1} w$$

What is the characteristic functional of a Lévy white noise?

Lévy White Noise Characteristic Functional

Theorem (Gelfand and Vilenkin, 1964)

Let $s(t)$ be a Lévy process on $\mathbb{R}$ and let $w = \mathrm{D}s$. Then, $w \in \mathcal{D}'(\mathbb{R})$ almost surely and

$$\widehat{\mathbf{P}}_w(\varphi) = \exp\left(\int_{\mathbb{R}} f(\varphi(x))\,\mathrm{d}x\right), \quad \varphi \in \mathcal{D}(\mathbb{R}),$$

where

$$f(\xi) = \mathrm{i}\mu\xi - \frac{\sigma^2\xi^2}{2} + \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}\xi t} - 1 - \mathrm{i}\xi \mathbf{1}_{[-1,1]}(t)\,\mathrm{d}\mathbf{P}_V(t),$$

with $\mu \in \mathbb{R}$, $\sigma^2 \geq 0$, and $\mathbf{P}_V$ is a Lévy measure, i.e.,

$$\int_{\mathbb{R}} \min\{1, t^2\}\,\mathrm{d}\mathbf{P}_V(t) < \infty \quad \text{and} \quad \mathbf{P}_V(\{0\}) = 0.$$

$(\mu, \sigma^2, \mathbf{P}_V)$ is called a **Lévy triplet** and f is a **Lévy exponent**

Examples

- Poisson white noise w_{Poi} on $\mathbb{R}$ with rate λ and jump law $\mathbf{P}_V$

$$\widehat{\mathbf{P}}_{w_{\mathrm{Poi}}}(\varphi) = \exp\left(\lambda \int_{\mathbb{R}} \int_{\mathbb{R}} \left(\mathrm{e}^{\mathrm{i}v\varphi(t)} - 1\right) \mathrm{d}t \, \mathrm{d}\mathbf{P}_V(v)\right)$$

Duttweiler and Kailath (1973)

cf., $\xi \mapsto \exp\left(\lambda(\mathrm{e}^{\mathrm{i}\xi} - 1)\right)$

- Gaussian white noise w_{Gauss} with unit variance

$$\widehat{\mathbf{P}}_{w_{\mathrm{Gauss}}}(\varphi) = \exp\left(-\frac{\|\varphi\|_{L^2}^2}{2}\right)$$

Gelfand and Vilenkin (1964)

cf., $\xi \mapsto \exp\left(-\frac{|\xi|^2}{2}\right)$

Lévy Process Characteristic Functional

$$\widehat{\mathbf{P}}_s(\varphi) = \mathbf{E}[\mathrm{e}^{\mathrm{i}\langle s,\varphi\rangle}], \quad \varphi \in \mathcal{D}(\mathbb{R})$$

$$s \overset{\mathscr{L}}{=} \mathrm{D}_0^{-1} w$$

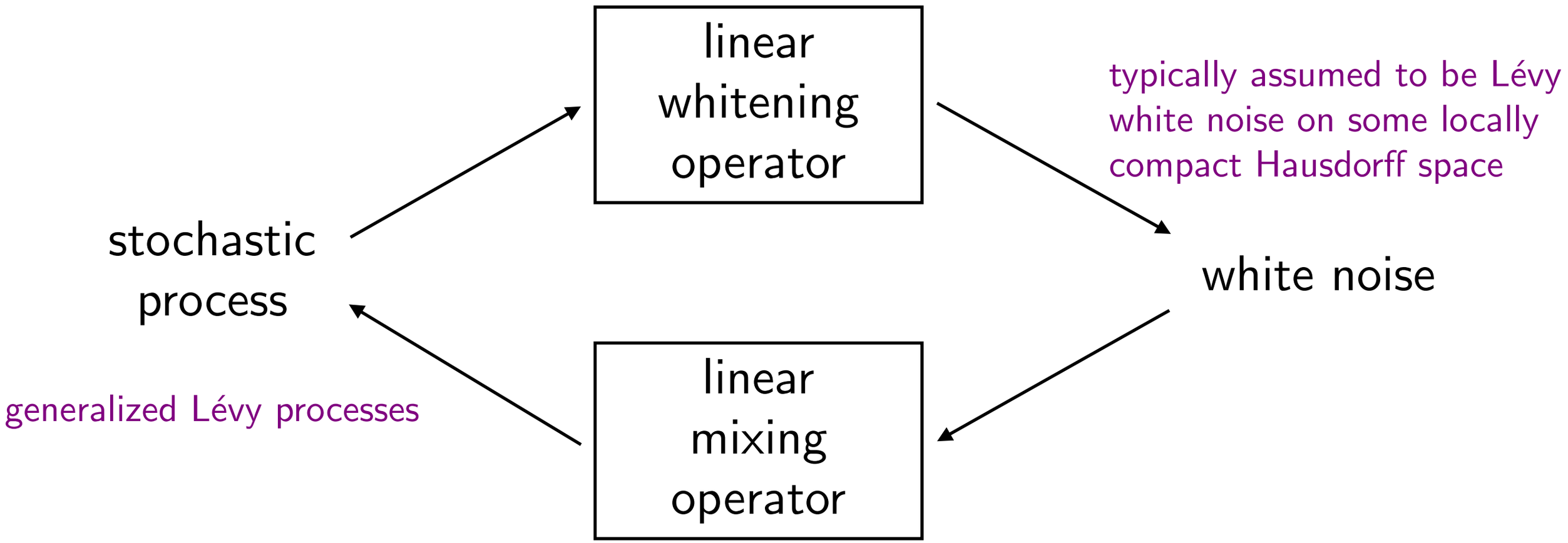
$$\begin{aligned}\langle s, \varphi\rangle &\overset{\mathscr{L}}{=} \langle \mathrm{D}_0^{-1} w, \varphi\rangle \\ &= \langle w, \mathrm{D}_0^{-1*}\varphi\rangle\end{aligned}$$

Fageot and Humeau (2021)

$$\begin{aligned}\widehat{\mathbf{P}}_s(\varphi) = \mathbf{E}[\mathrm{e}^{\mathrm{i}\langle s,\varphi\rangle}] &= \mathbf{E}[\mathrm{e}^{\mathrm{i}\langle w, \mathrm{D}_0^{-1*}\varphi\rangle}] \\ &= \widehat{\mathbf{P}}_w(\mathrm{D}_0^{-1*}\varphi)\end{aligned}$$

Given the characteristic functional of the innovation process, you automatically have the characteristic functional of the original process.

A More Abstract Innovation Model



$$\text{mixing} = \text{whitening}^{-1}$$

existence of suitable operators is an active area of research

$$\widehat{\mathbf{P}}_s(\varphi) = \widehat{\mathbf{P}}_w(\mathrm{L}^{-1*}\varphi)$$

Tempered Lévy White Noise

Theorem (Dalang and Humeau, 2017)

Let w be a Lévy white noise on $\mathbb{R}^d$, i.e., $w \in \mathcal{D}'(\mathbb{R}^d)$ and

$$\widehat{\mathbf{P}}_w(\varphi) = \exp\left(\int_{\mathbb{R}^d} f(\varphi(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}\right), \quad \varphi \in \mathcal{D}(\mathbb{R}^d),$$

where

$$f(\xi) = \mathrm{i}\mu\xi - \frac{\sigma^2\xi^2}{2} + \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}\xi t} - 1 - \mathrm{i}\xi\mathbf{1}_{[-1,1]}(t)\,\mathrm{d}\mathbf{P}_V(t),$$

with Lévy triplet $(\mu, \sigma^2, \mathbf{P}_V)$. Then, w is tempered (a.s.) if and only if there exists $\varepsilon > 0$ such that

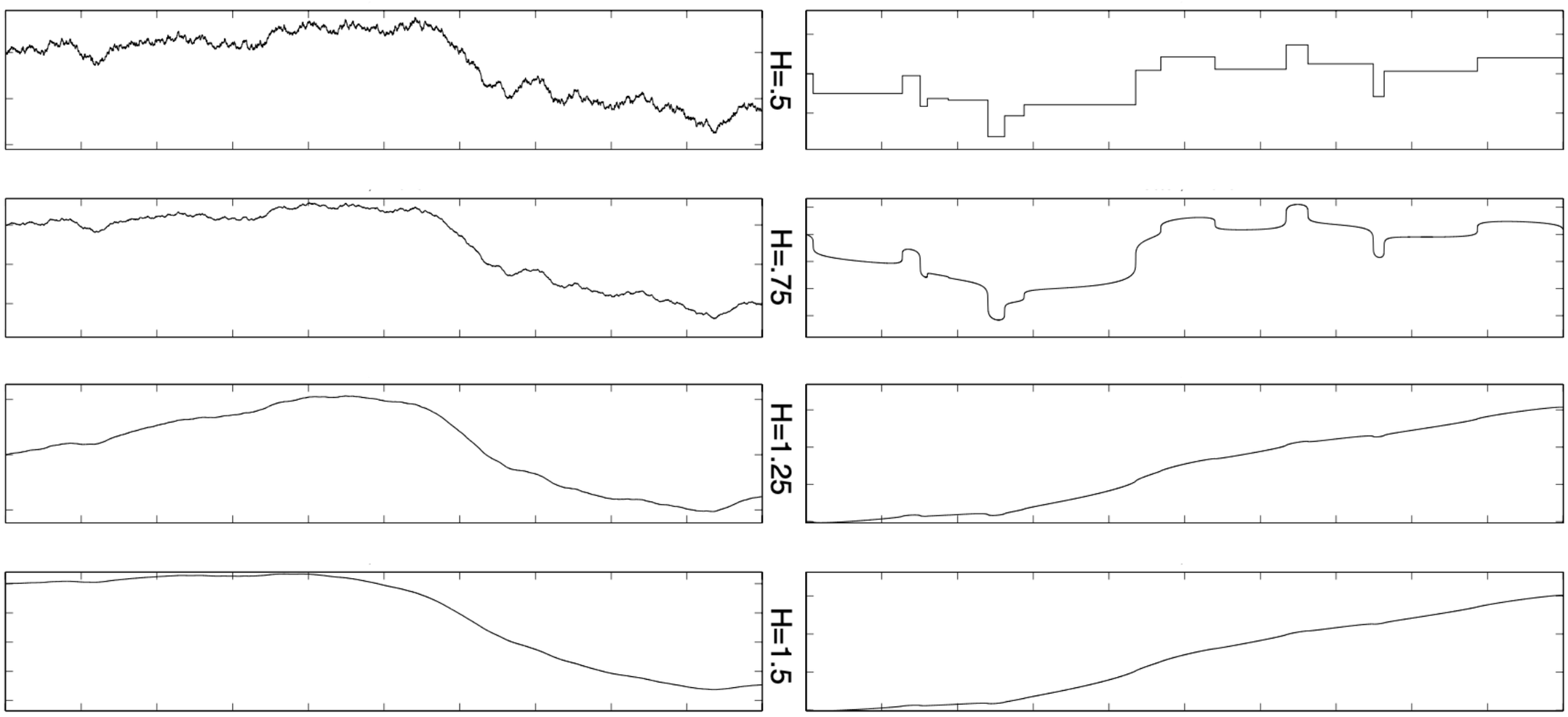
$$\mathbf{E}[|V|^\varepsilon] < \infty, \quad V \sim \mathbf{P}_V.$$

$$\mathbf{P}(w \in \mathcal{S}'(\mathbb{R}^d)) = 1 \Leftrightarrow \mathbf{P}(w \in \mathcal{D}'(\mathbb{R}^d) \setminus \mathcal{S}'(\mathbb{R}^d)) = 0 \Leftrightarrow \mathrm{supp}(\mathbf{P}_w) \subset \mathcal{S}'(\mathbb{R}^d)$$

Fractional Order Processes

$$\mathrm{D}^{H+\frac{1}{2}} s \overset{\mathscr{L}}{=} w$$

Blu and Unser (2007)

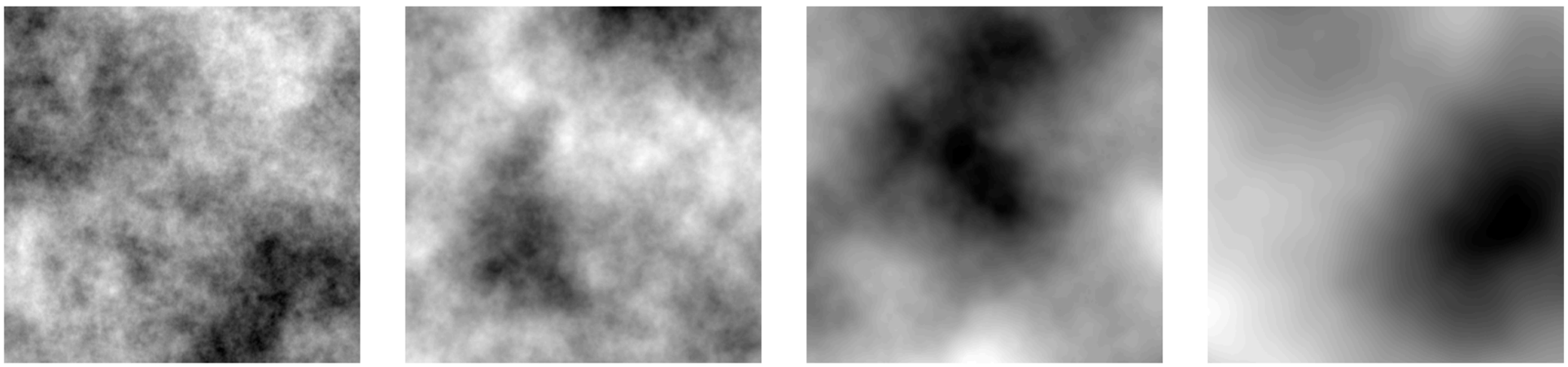


Mandelbrot and Van Ness (1968)

Fractional Order Processes

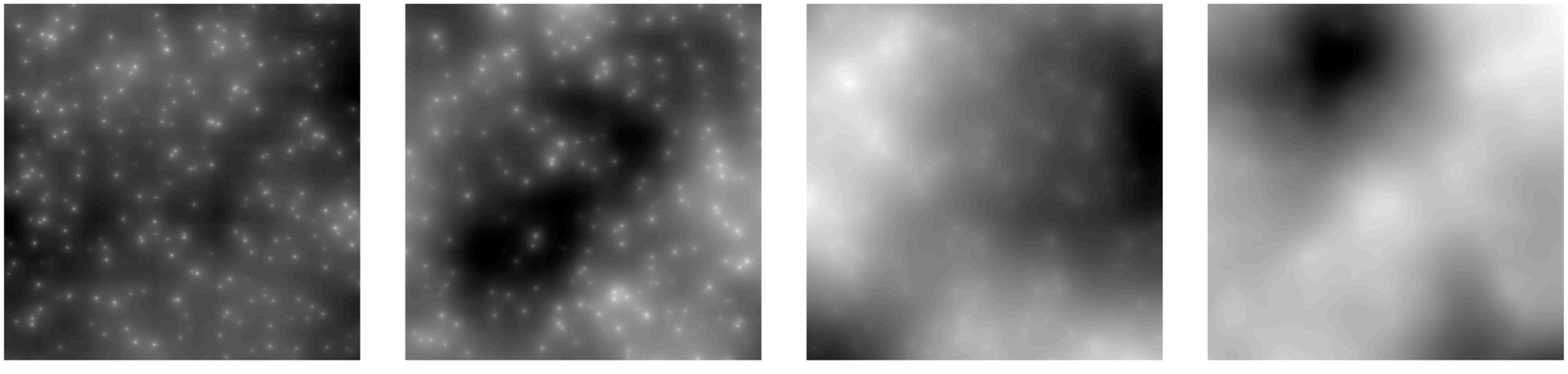
$$(-\Delta)^{\frac{H+1}{2}} s \overset{\mathscr{L}}{=} w$$

Gaussian

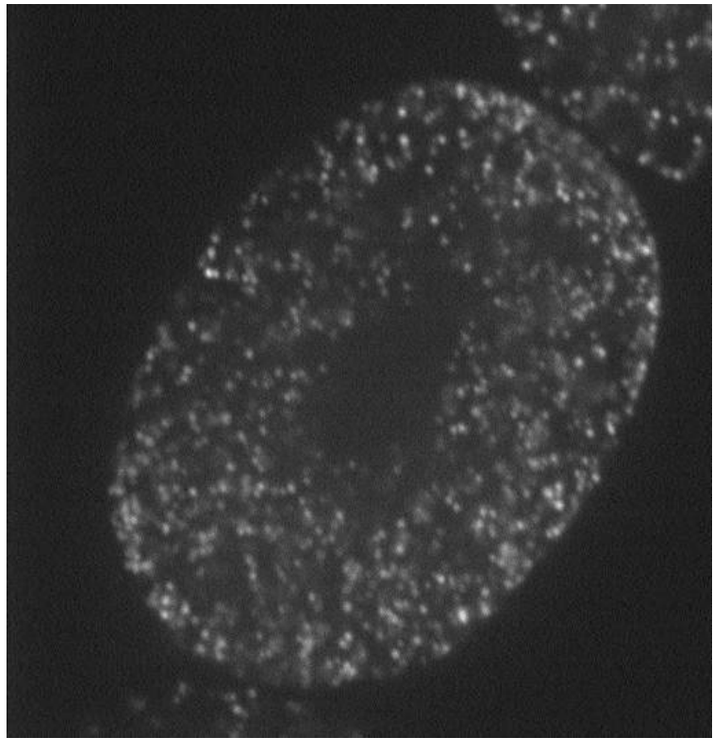


H=.5 H=.75 H=1.25 H=1.75

Sparse (generalized Poisson)



Sparse Stochastic Processes are Good Models

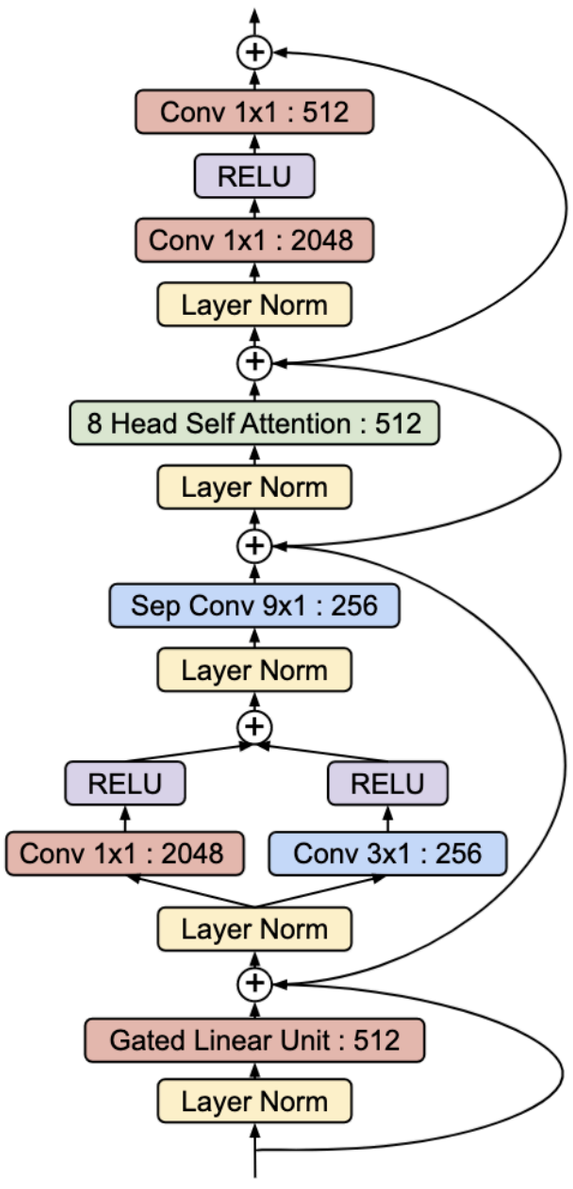


Applications to Random Neural Networks

Deep Neural Network Architectures

The Evolved Transformer

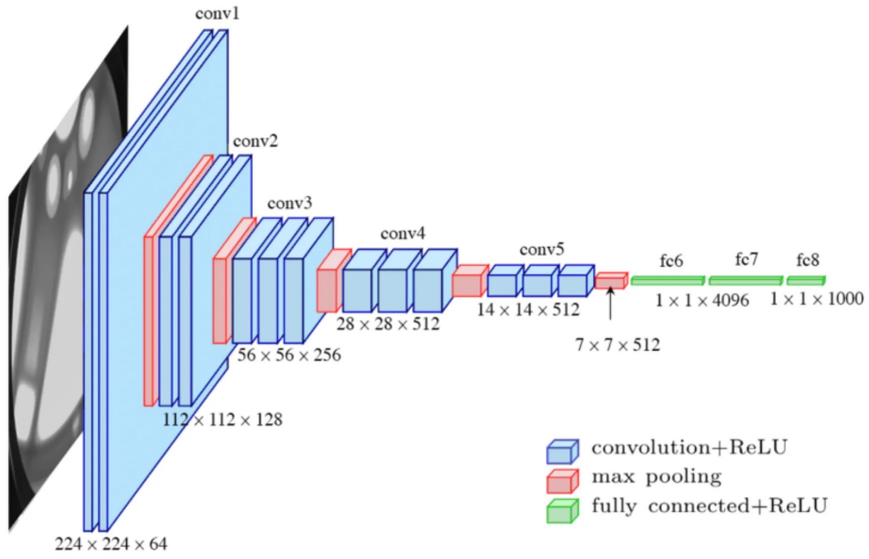
David So, Quoc Le, Chen Liang *Proceedings of the 36th International Conference on Machine Learning*, PMLR 97:5877-5886, 2019.



Google DeepMind 2023-10-26

ConvNets Match Vision Transformers at Scale

Samuel L Smith[1], **Andrew Brock**[1], **Leonard Berrada**[1] **and Soham De**[1]
[1]Google DeepMind

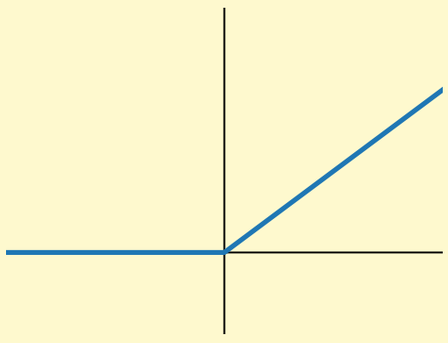


Very deep convolutional networks for large-scale image recognition

K Simonyan, A Zisserman

In this work we investigate the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting. Our main contribution is a thorough evaluation of …

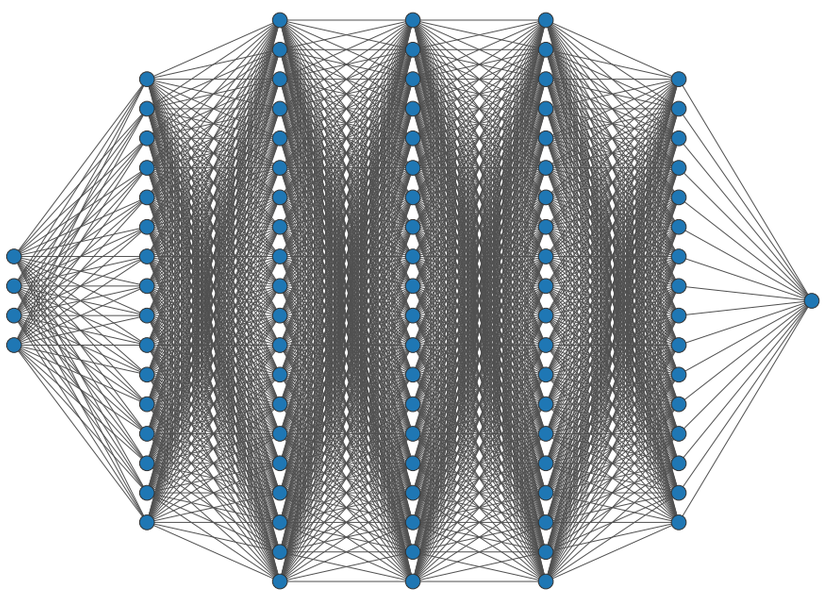
☆ Save 99 Cite Cited by 112161 Related articles »



Rectified Linear Unit (ReLU)

$$\mathrm{ReLU}(t) = \max\{0, t\} = t_+$$

Neural Network Training

$\boldsymbol{x}$ 

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\sigma}(\mathbf{W}_L \boldsymbol{\sigma}(\mathbf{W}_{L-1} \boldsymbol{\sigma}(\cdots \boldsymbol{\sigma}(\mathbf{W}_1 \boldsymbol{x}))))$$

parameterized by a vector $\boldsymbol{\theta} \in \mathbb{R}^P$ of neural network **weights**

Neural network training problem for the data $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$.

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^P} \underbrace{\sum_{n=1}^{N} \mathcal{L}(y_n, f_{\boldsymbol{\theta}}(\boldsymbol{x}_n))}_{\text{data fidelity}} + \underbrace{\frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2}_{\text{regularization}}$$

Neural Network Training

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^P} \underbrace{\sum_{n=1}^{N} \mathcal{L}(y_n, f_{\boldsymbol{\theta}}(\boldsymbol{x}_n))}_{\mathscr{L}(\boldsymbol{\theta})} + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2$$

Gradient descent update on θ_i

$$\theta_i^{t+1} = \theta_i^t - \tau \left(\left. \frac{\partial \mathscr{L}}{\partial \theta_i} \right|_{\theta_i = \theta_i^t} + \lambda \theta_i^t \right)$$

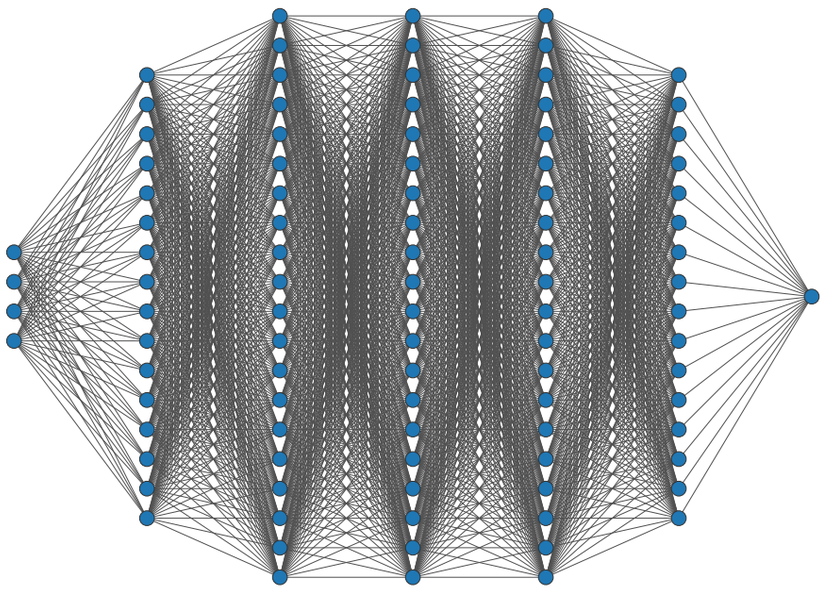
step size
"learning rate"

How do we choose θ_i^0?

randomly!

Hanson and Pratt (1988, NeurIPS)
Krogh and Hertz (1990, NeurIPS)

Random Neural Networks

$\boldsymbol{x}$ 

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\sigma}(\mathbf{W}_L \boldsymbol{\sigma}(\mathbf{W}_{L-1} \boldsymbol{\sigma}(\cdots \boldsymbol{\sigma}(\mathbf{W}_1 \boldsymbol{x}))))$$

parameterized by a vector $\boldsymbol{\theta} \in \mathbb{R}^P$
of neural network **weights**

A random neural network is a random function.

$$f_{\boldsymbol{\theta}} \qquad \boldsymbol{\theta} \sim \mathbf{Q}$$

A random neural network is a stochastic process.

$$(f_{\boldsymbol{\theta}}(\boldsymbol{x}))_{\boldsymbol{x} \in \mathbb{R}^d}$$

Machine Learning Folklore

Folklore Theorem

The parameters of neural networks trained with GD do not move far from their initialization.

If we can understand neural networks at initialization, then we can understand everything!

DEEP NEURAL NETWORKS AS GAUSSIAN PROCESSES

Jaehoon Lee[*†], **Yasaman Bahri**[*†], **Roman Novak**, **Samuel S. Schoenholz**, **Jeffrey Pennington**, **Jascha Sohl-Dickstein**

Google Brain
{jaehlee, yasamanb, romann, schsam, jpennin, jaschasd}@google.com

ICLR 2018
1300+ citations

Neural Tangent Kernel: Convergence and Generalization in Neural Networks

Arthur Jacot
École Polytechnique Fédérale de Lausanne
arthur.jacot@netopera.net

Franck Gabriel
Imperial College London and École Polytechnique Fédérale de Lausanne
franckrgabriel@gmail.com

Clément Hongler
École Polytechnique Fédérale de Lausanne
clement.hongler@gmail.com

NeurIPS 2018
3800+ citations

Spawned a Burgeoning Industry of Research

WIKIPEDIA
The Free Encyclopedia

Create account Log in

Neural network Gaussian process

3 languages

Article Talk | Read Edit View history Tools

From Wikipedia, the free encyclopedia

A **Neural Network Gaussian Process** (NNGP) is a Gaussian process (GP) obtained as the limit of a certain type of sequence of neural networks. Specifically, a wide variety of network architectures converges to a GP in the infinitely wide limit, in the sense of distribution.[1][2][3][4][5][6][7][8] The concept constitutes an intensional definition, i.e., a NNGP is just a GP, but distinguished by how it is obtained.

WIKIPEDIA
The Free Encyclopedia

Create account Log in

Neural tangent kernel

2 languages

Article Talk | Read Edit View history Tools

From Wikipedia, the free encyclopedia

In the study of artificial neural networks (ANNs), the **neural tangent kernel** (**NTK**) is a kernel that describes the evolution of deep artificial neural networks during their training by gradient descent. It allows ANNs to be studied using theoretical tools from kernel methods.

In general, a kernel is a positive-semidefinite symmetric function of two inputs which represents some notion of similarity between the two inputs. The NTK is a specific kernel derived from a given neural network; in general, when the neural network parameters change during training, the NTK evolves as well. However, in the limit of large layer width the NTK becomes constant, revealing a duality between training the wide neural network and kernel methods: gradient descent in the infinite-width limit is fully equivalent to kernel gradient descent with the NTK. As a result, using gradient descent to minimize least-square loss for neural networks yields the same mean estimator as ridgeless kernel regression with the NTK. This duality enables simple closed form equations describing the training dynamics, generalization, and predictions of wide neural networks.

The NTK was introduced in 2018 by Arthur Jacot, Franck Gabriel and Clément Hongler,[1] who used it to study the convergence and generalization properties of fully connected neural networks. Later works[2][3] extended the NTK results to other neural network architectures. In fact, the phenomenon behind NTK is not specific to neural networks and can be observed in generic nonlinear models, usually by a suitable scaling[4].

Contradicting Machine Learning Folklore

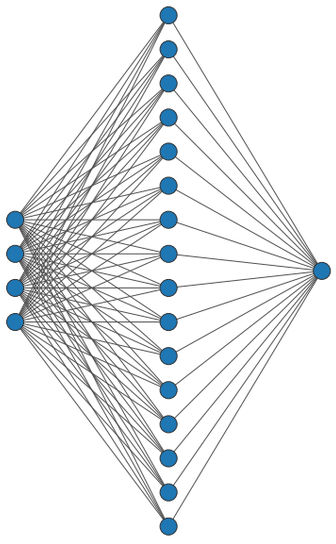
Folklore Theorem

Neural networks initialized with random parameters (with any law) converge to Gaussian processes in wide limits.

- Shallow neural networks with ReLU activation functions.
- Special kind of initialization of the network parameters.
- Finite-width networks are non-Gaussian processes.
- Gaussian and non-Gaussian processes in wide limits.

We provide a **complete characterization** of the statistical distribution of these neural network processes via their **characteristic functional**.

Shallow ReLU Neural Networks

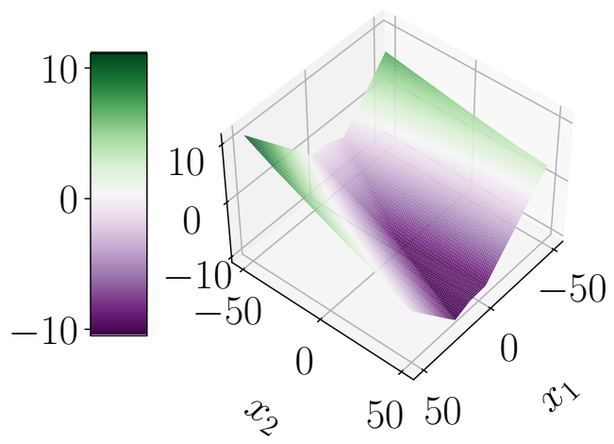


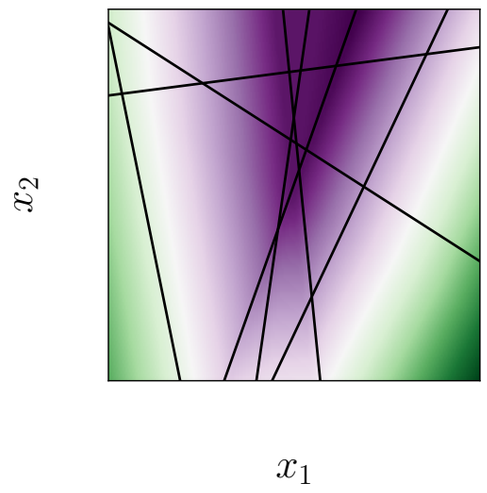
$$\boldsymbol{x} \mapsto \sum_{k=1}^{K} v_k \boxed{(\boldsymbol{w}_k^\mathsf{T} \boldsymbol{x} - b_k)_+}$$

ReLU neurons

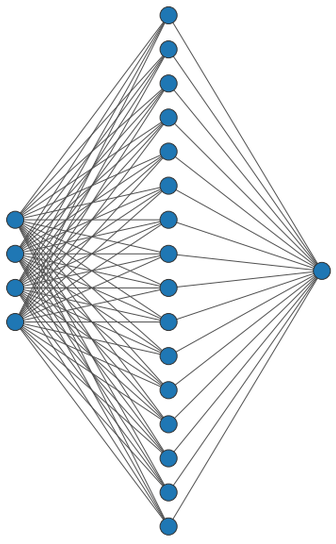
K is the "width" of the network

ReLU network





Shallow ReLU Neural Networks



$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^{K} v_k(\boldsymbol{w}_k^{\mathsf{T}}\boldsymbol{x} - b_k)_+$$

$$\boldsymbol{\theta} \sim \mathbf{Q}$$

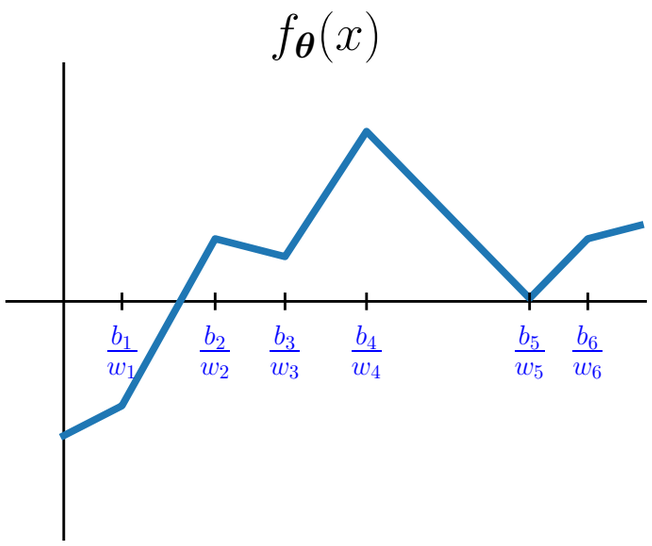
How do we study these random neural networks?

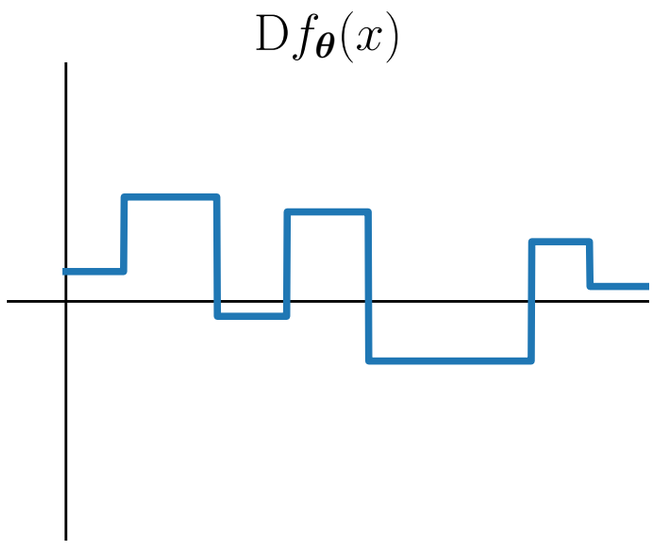
How do we derive a form of its **characteristic functional**?

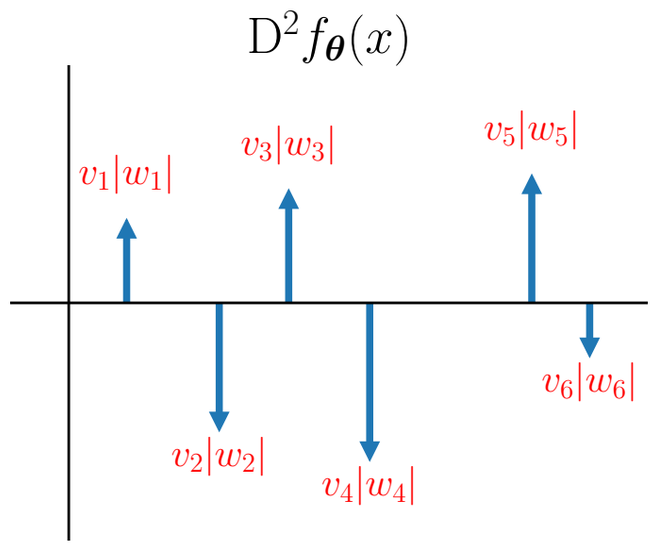
We need to find a way to **whiten** the neural network.

Univariate Shallow ReLU Neural Networks

$$f_{\boldsymbol{\theta}}(x) = \sum_{k=1}^{K} v_k (w_k x - b_k)_+$$



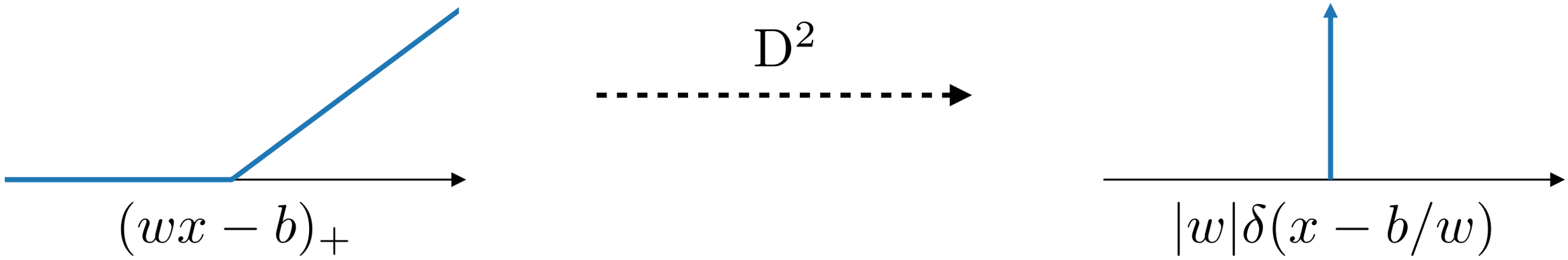


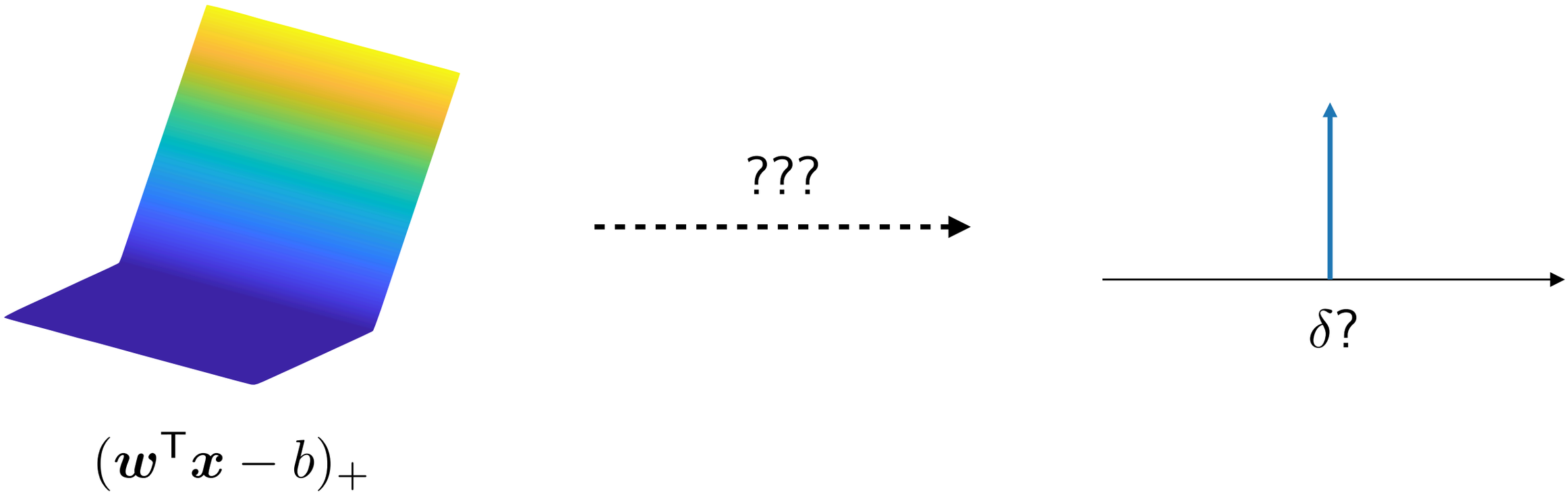


"looks like" a Poisson white noise

Second derivatives "whiten" univariate neural networks

What About the Multivariate Case?

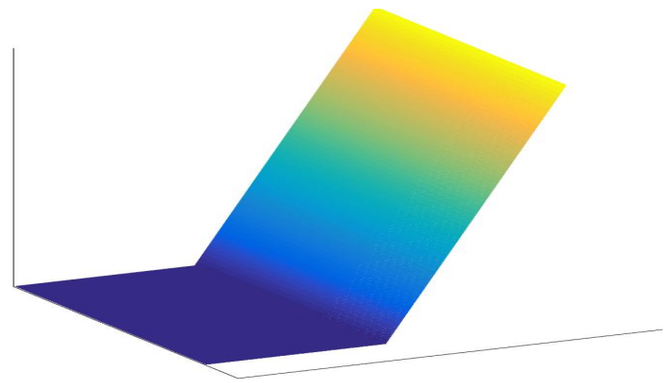




Whitening Operator of Shallow ReLU Networks

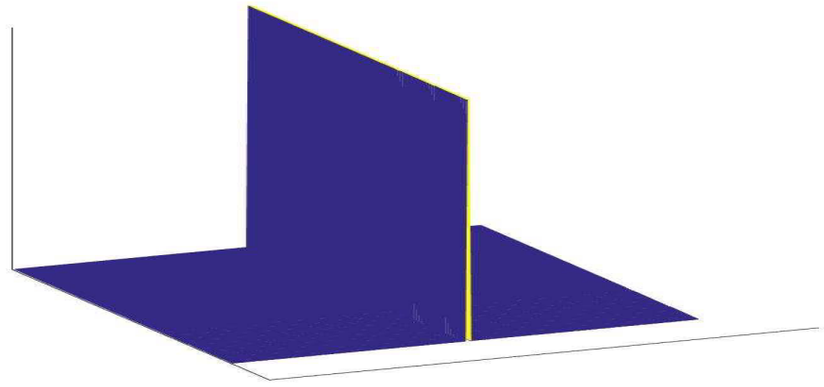
ReLU Neuron: $(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)_+$

$$= \|\boldsymbol{w}\|_2 \left(\left[\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2} \right]^\mathsf{T} \boldsymbol{x} - \frac{b}{\|\boldsymbol{w}\|_2} \right)_+$$



assume $\boldsymbol{w} \in \mathbb{S}^{d-1}$

$$\Delta\{(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)_+\} = \delta(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)$$



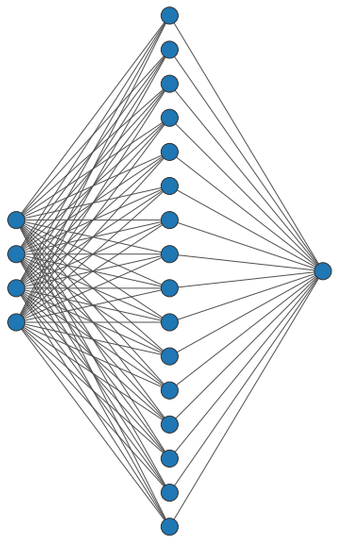
$$\mathrm{K}\mathscr{R}\Delta\{(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)_+\} = \delta^{\mathsf{e}}_{(\boldsymbol{w},b)}$$

$$\widehat{\mathrm{K}g}(\omega) \propto |\omega|^{d-1}\widehat{g}(\omega)$$

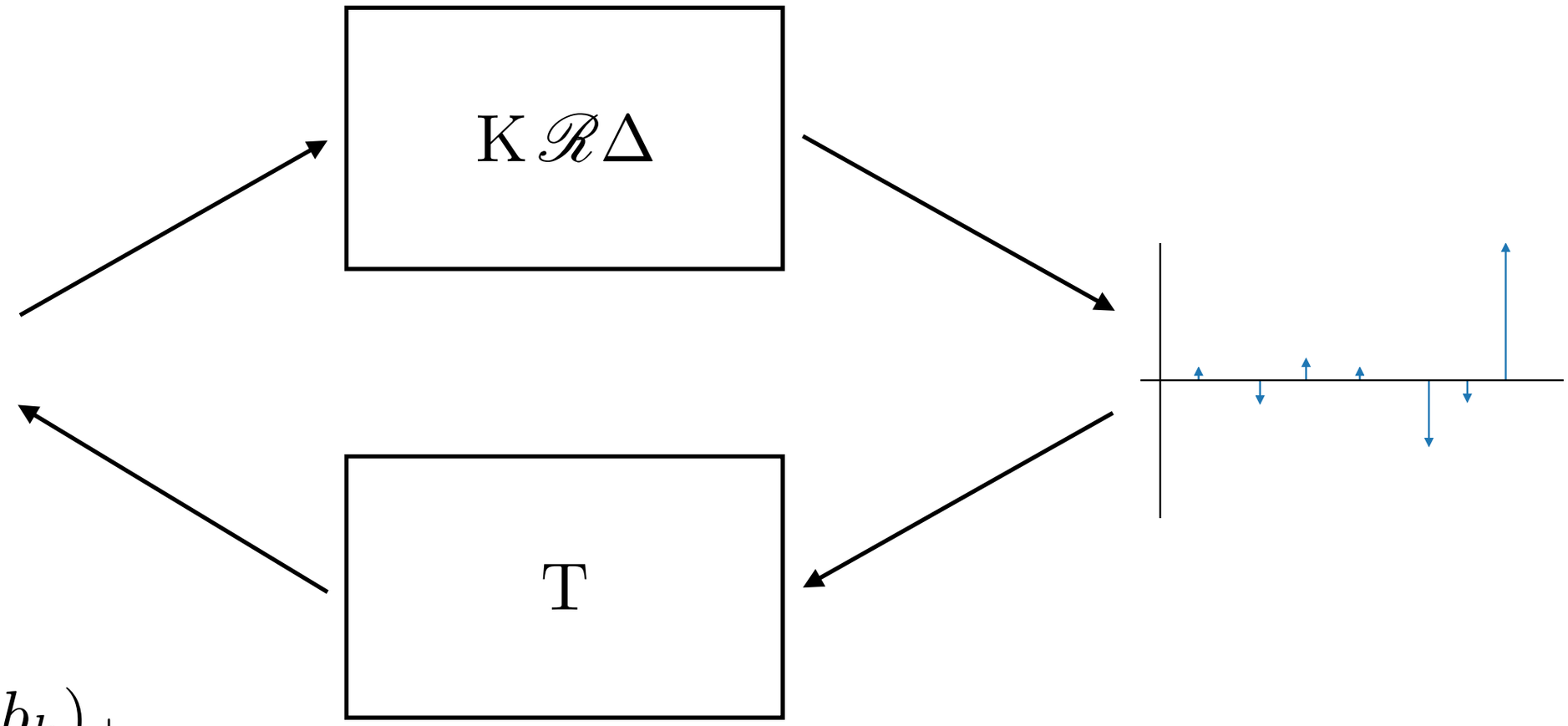
Radon-domain Dirac centered at $(\boldsymbol{w}, b)$

Ongie et al. (2020); **P.** and Nowak (2021); **P.** and Unser (2024)

Whitening Operator of Shallow ReLU Networks



$$\sum_{k=1}^{K} v_k(\boldsymbol{w}_k^\mathsf{T}\boldsymbol{x} - b_k)_+$$



Does such a well-behaved inverse exist?

$$\mathrm{T}\{\psi\}(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} k_{\boldsymbol{x}}(\boldsymbol{u}, t)\psi(\boldsymbol{u}, t)\,\mathrm{d}\boldsymbol{u}\,\mathrm{d}t$$

$$k_{\boldsymbol{x}}(\boldsymbol{u}, t) = (\boldsymbol{u}^\mathsf{T}\boldsymbol{x} - t)_+ - \frac{(\boldsymbol{u}^\mathsf{T}\boldsymbol{x} - t)}{2} - \frac{|t|}{2} + \boldsymbol{u}^\mathsf{T}\boldsymbol{x}\,\mathrm{sgn}(t)$$

P. and Nowak (2021); **P.** and Unser (2024)

Random Generation of Shallow ReLU Networks

$$\sum_k v_k(\boldsymbol{w}_k^\mathsf{T}\boldsymbol{x} - b_k)_+$$

1. Generate $(v_k, \boldsymbol{w}_k, b_k)$ according to some point process on $\mathbb{S}^{d-1} \times \mathbb{R}$.

$$w = \sum_k v_k \, \delta_{(w_k, b_k)}$$

2. Compute the characteristic functional $\widehat{\mathbf{P}}_w$ of this point process

3. The characteristic functional of the random neural network is $\widehat{\mathbf{P}}_w(\mathrm{T}^*\varphi)$

Random Generation of Shallow ReLU Networks

$$\sum_k v_k(\boldsymbol{w}_k^\mathsf{T}\boldsymbol{x} - b_k)_+$$

The v_k are drawn i.i.d. with respect to $\mathbf{P}_V$. The $(\boldsymbol{w}_k, b_k)$ are drawn such that

finite absolute moment

1. The *activation thresholds* are mutually independent.
2. The expectation of the number of thresholds that intersect a finite volume in $\mathbb{R}^d$ is a constant proportional to a rate $\lambda > 0$.
3. For every finite volume in $\mathbb{R}^d$, the thresholds are i.i.d. and "uniform" in the volume.

$(\boldsymbol{w}_k, b_k)$ form a homogeneous Poisson point process on $\mathbb{S}^{d-1} \times \mathbb{R}$

Write $s \sim \mathcal{RP}(\lambda; \mathbf{P}_V)$ (ReLU Process) to denote this randomness.

Characteristic Functional of ReLU Processes

Theorem (**P**BEPU, 2024)

The characteristic functional of random ReLU neural network

$$s \sim \mathcal{RP}(\lambda; \mathbf{P}_V)$$

is given by

$$\varphi \in \mathcal{S}(\mathbb{R}^d)$$

$$\widehat{\mathbf{P}}_s(\varphi) = \exp\left(\lambda \int_{\mathbb{R}} \int_{\mathbb{R}} \int_{\mathbb{S}^{d-1}} \left(e^{\mathrm{i}v\mathrm{T}^*\{\varphi\}(\boldsymbol{u},t)} - 1\right) \mathrm{d}\boldsymbol{u}\, \mathrm{d}t\, \mathrm{d}\mathbf{P}_V(v)\right).$$

$\widehat{\mathbf{P}}_s$ gives us the full statistical distribution of s.

The law $\mathbf{P}_s$ of s is the inverse Fourier transform of $\widehat{\mathbf{P}}_s$.

Properties of ReLU Processes

Theorem (PBEPU, 2024)

Let $s \sim \mathcal{RP}(\lambda; \mathbf{P}_V)$ and $V \sim \mathbf{P}_V$. Then,

- The process is non-Gaussian.
- The autocovariance of the process is

$$C_s(\boldsymbol{x}, \boldsymbol{y}) \propto \mathbf{E}[V^2] \left(\|\boldsymbol{x} - \boldsymbol{y}\|_2^3 - \|\boldsymbol{x}\|_2^3 - \|\boldsymbol{y}\|_2^3 + 3\boldsymbol{x}^\mathsf{T}\boldsymbol{y}(\|\boldsymbol{x}\|_2 + \|\boldsymbol{y}\|_2) \right).$$

- The process is isotropic.
- The process wide-sense self-similar with Hurst exponent $H = 3/2$ (i.e., s and $a^H s(\cdot/a)$, $a > 0$ have the same second-order statistics).

Asymptotics of ReLU Processes

$$s_\lambda \sim \mathcal{RP}(\lambda; \mathbf{P}_V)$$

λ controls the width of the random ReLU neural network

$$s_\lambda|_\Omega(\boldsymbol{x}) = \boldsymbol{w}_0^\mathsf{T}\boldsymbol{x} + b_0 + \sum_{k=1}^{N_{\lambda,\Omega}} v_k(\boldsymbol{w}_k^\mathsf{T}\boldsymbol{x} - b_k)_+, \quad N_{\lambda,\Omega} \text{ Poisson with mean } \propto \lambda|\Omega|$$

As $\lambda \to \infty$ the network s_λ is an **infinite-width** network.

$\mathbf{P}_V$	fixed λ	$\lambda \to \infty$
Gaussian	s_λ is non-Gaussian	s_∞ is Gaussian
Laplacian	s_λ is non-Gaussian	s_∞ is Laplacian
α-stable, $\alpha < 2$	s_λ is non-Gaussian	s_∞ is α-stable

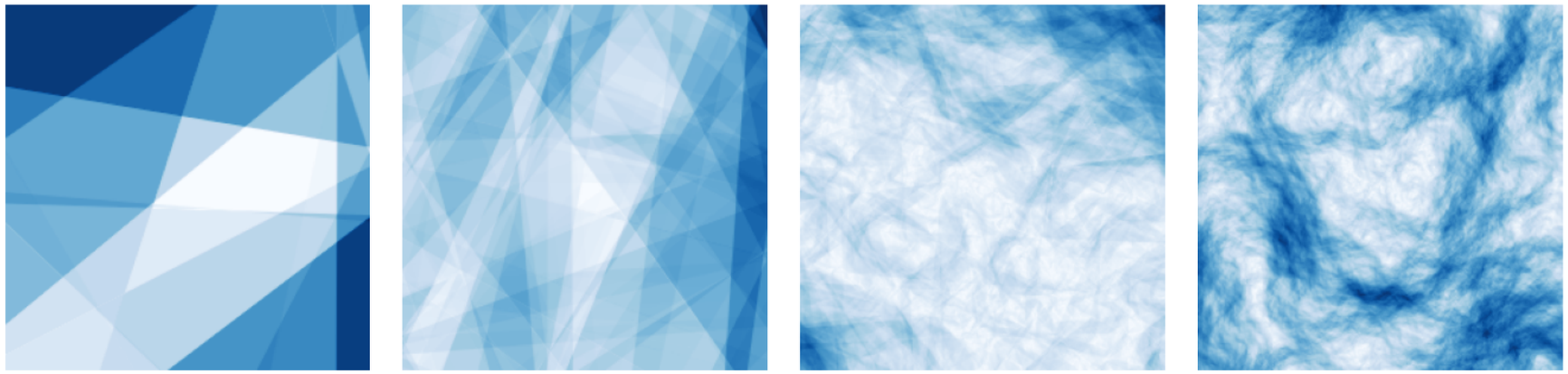
Random neural networks are non-Gaussian processes **even in wide limits**.

short proof via characteristic functional + Lévy-Fernique theorem

can check for non-Gaussianity by inspection

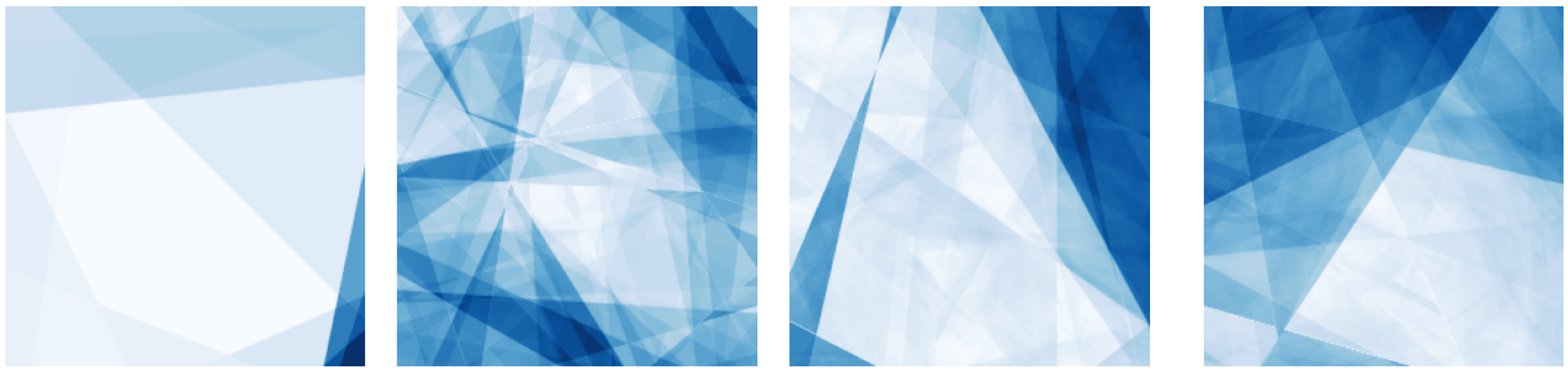
Pretty Pictures

$\mathbf{P}_V$ is Gaussian



$\lambda = 1$ $\lambda = 10$ $\lambda = 100$ $\lambda = 1000$

$\mathbf{P}_V$ is 1.25-stable



$\lambda = 1$ $\lambda = 10$ $\lambda = 100$ $\lambda = 1000$

Conclusion

Random shallow ReLU neural networks can be viewed as the deterministic "mixing" of a point process on the Radon domain

- Generic procedure to derive the characteristic functional of these random neural networks $\Rightarrow$ law of the stochastic process
- The characteristic functional streamlines the derivation of properties (asymptotic and non-asymptotic) of stochastic processes
- We can prove things that "contradict" machine learning folklore

Next steps:

- Explore other Radon-domain point processes (and general Radon -domain innovation processes)
- Deep neural networks? Compositions of stochastic processes?

arXiv > stat > arXiv:2405.10229

Statistics > Machine Learning

[Submitted on 16 May 2024]

Random ReLU Neural Networks as Non-Gaussian Processes

Rahul Parhi, Pakshal Bohra, Ayoub El Biari, Mehrsa Pourya, Michael Unser