The NN-QFT Correspondence

Anindita Maiti
(Northeastern University)

Based on: 2008.08601 with J. Halverson and K. Stoner

String Phenomenology Seminar Series

Neural Network “=” Euclidean Quantum Field Theory

Build up the new correspondence

Essence of modern QFT into NNs:
Wilsonian EFT and renormalization

Why Neural Networks

Neural networks (NN) are the backbones of of Deep learning:

Supervised: NN is powerful function that predicts outputs (e.g. class labels), given input.

Generative Models: NN is powerful function that maps draws from noise distribution to draws from data distribution. Learns to generate/simulate/fake data.

In physics, e.g.: Simulate GEANT4 ECAL simulator. CaloGAN, [Paganini et al, 2018].
Simulate string theory EFTs, ALP kinetic terms. used Wasserstein GAN, [Halverson, Long, 2020]

Reinforcement: NN is powerful function that, e.g., picks intelligent state-dependent actions.

In chess, eg.: Alphazero.

Train a NN k different times, different results, because it's a function from some distribution

Outline

1. Introduction to Neural Networks
2. Asymptotic Neural Networks, Gaussian Processes, and Free Field Theory
3. Finite Neural Networks, Non-Gaussian Processes, and Effective Field Theory
4. Wilsonian Renormalization in Neural Net Non-Gaussian Processes

Introduction to Neural Networks

Neural Networks : Backbone of Deep Learning

A function with continuous learnable parameters $\boldsymbol{\theta}$ and discrete hyperparameters N.

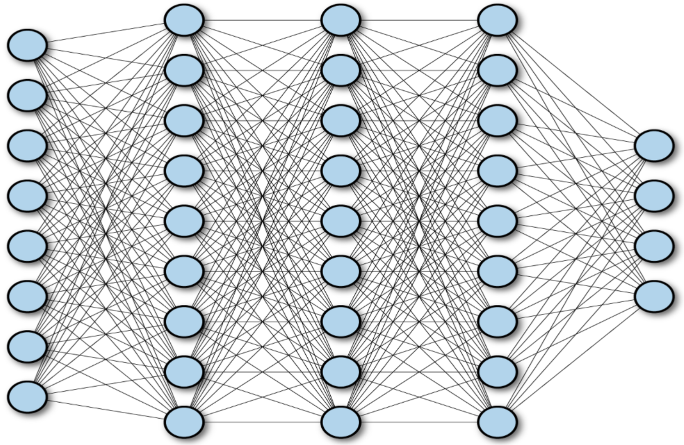
$$f_{\theta,N} : \mathbb{R}^{d_{\text{in}}} \rightarrow \mathbb{R}^{d_{out}}$$

Rough idea: neural network as computational nodes that pass information along edges.

Training mechanism updates $\boldsymbol{\theta}$ to improve performance :
Supervised learning, Generative models, Reinforcement learning

Fully Connected Networks :

→ network "depth" direction.



input → hidden → … → hidden→ output

- x : network *input*.
- σ : non-linear *activation function*
- x_j : *post-activation*
- z_j : *pre-activation*, affine transformation of post. Truncate at *output.*
- W and b : *weights* and *biases*, previously $\boldsymbol{\theta}$.

$$z_i^l(x) = b_i^l + \sum_{j=1}^{N} W_{ij}^l x_j^l(x)$$

$$x_j^l = \sigma(z_j^{l-1}(x))$$

Asymptotic Neural Networks, Gaussian Processes, and Free Field Theory

Asymptotic NN "=" GP "=" Free Field Theory

Asymptotic limit : hyperparameter N → ∞ limit .

Central Limit Theorem : Add N independently and Identically distributed random variables (iid), take N → ∞, sum is drawn from a Gaussian distribution.

NN outputs drawn from a Gaussian distribution on function space, it's a *Gaussian Process* (GP).

Infinite Width Single-Layer Feedforward Network

$$f_{\theta,N} : \mathbb{R}^{d_{\text{in}}} \xrightarrow{W_0,\, b_0} \mathbb{R}^N \xrightarrow{\sigma} \mathbb{R}^N \xrightarrow{W_1,\, b_1} \mathbb{R}^{d_{\text{out}}}$$

$$f(x) = W_1(\sigma(W_0 x + b_0)) + b_1$$

$$b_0, b_1 \sim \mathcal{N}(\mu_b, \sigma_b^2)$$

$$W_0 \sim \mathcal{N}(\mu_W, \sigma_W^2/d_{\text{in}}) \qquad W_1 \sim \mathcal{N}(\mu_W, \sigma_W^2/N)$$

Any *standard* NN architecture admits GP limit when N → ∞ .

Eg. Single-layer infinite width feedforward networks, Deep infinite width feedforward networks, Infinite channel CNNs.

[Neal], [Williams] 1990's , [Lee et al., 2017], [Matthews et al., 2018] , [Yang, 2019], [Yang, 2020]

GP property persists under appropriate training.

[Jacot et al., 2018], [Lee et al., 2019], [Yang, 2020]

Gaussian Processes and Free Field Theory

Gaussian Process:

distribution: $$P[f] \sim \exp\left[-\frac{1}{2}\int d^{d_{\rm in}}x\, d^{d_{\rm in}}x'\, f(x)\Xi(x,x')f(x')\right]$$

where: $$\int d^{d_{\rm in}}x'\, K(x,x')\,\Xi(x',x'') = \delta^{(d_{\rm in})}(x-x'')$$

K is the *kernel* of the GP.

log-likelihood: $$S = \frac{1}{2}\int d^{d_{\rm in}}x\, d^{d_{\rm in}}x'\;\; f(x)\Xi(x,x')f(x')$$

n-pt correlation functions: $$G^{(n)}(x_1,\dots,x_n) = \frac{\int df\;\; f(x_1)\dots f(x_n)\, e^{-S}}{Z}$$

GP / asymptotic NN	Free QFT
inputs $(x_1,\dots,x_k)$	external space or spacetime points
kernel $K(x_1,x_2)$	Feynman propagator
asymptotic NN $f(x)$	free field
log-likelihood	free action $S_{\rm GP}$

Free Field Theory:

“free” = non-interacting Feynman path integral:

$$Z = \int D\phi\, e^{-S[\phi]}$$

From P.I. perspective, Gaussian distributions on field space.

e.g., free scalar field theory

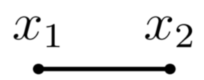
$$S[\phi] = \int d^dx\, \phi(x)(\Box + m^2)\phi(x)$$

GP Predictions for Correlation Functions

Analytic and Feynman diagram expressions for n-pt correlations of asymptotic NNs (right) :

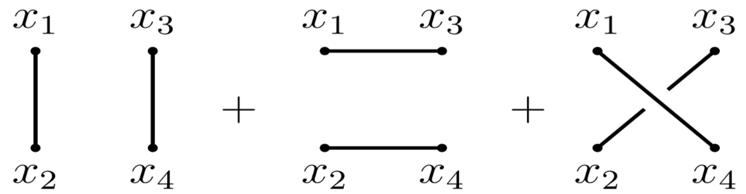
$$G^{(2)}_{\mathrm{GP}}(x_1, x_2) = K(x_1, x_2)$$

$$=$$



$$G^{(4)}_{\mathrm{GP}}(x_1, x_2, x_3, x_4) = K(x_1, x_2)K(x_3, x_4) + K(x_1, x_3)K(x_2, x_4) + K(x_1, x_4)K(x_2, x_3)$$

$$=$$



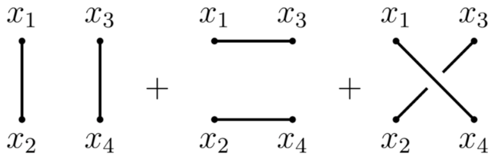
Physics analogy: mean-free GP is totally determined by 2-pt statistics, i.e. GP kernel.

kernel = propagator, so GP = a QFT where all diagrams rep. particles flying past each other.

$$\Delta G^{(n)}(x_1, \ldots, x_n) = G^{(n)}(x_1, \ldots x_n) - G^{(n)}_{\mathrm{GP}}(x_1, \ldots, x_n)$$

$$\Delta G^{(2)} = \frac{1}{n_{\mathrm{nets}}} \sum_{\alpha}^{n_{\mathrm{nets}}} f_\alpha(x_1) f_\alpha(x_2) - K(x_1, x_2)$$

$$\Delta G^{(4)} = \frac{1}{n_{\mathrm{nets}}} \sum_{\alpha}^{n_{\mathrm{nets}}} f_\alpha(x_1) f_\alpha(x_2) f_\alpha(x_3) f_\alpha(x_4) - \Big[$$



$$\Big]$$

$$\begin{aligned}
\Delta G^{(6)} &= \frac{1}{n_{\mathrm{nets}}} \sum_{\alpha}^{n_{\mathrm{nets}}} f_\alpha(x_1) f_\alpha(x_2) f_\alpha(x_3) f_\alpha(x_4) f_\alpha(x_5) f_\alpha(x_6) \\
&- \Big[K_{12}K_{34}K_{56} + K_{12}K_{35}K_{46} + K_{12}K_{36}K_{45} + K_{13}K_{24}K_{56} + K_{13}K_{25}K_{46} + K_{13}K_{26}K_{45} \\
&+ K_{14}K_{23}K_{56} + K_{14}K_{25}K_{36} + K_{14}K_{26}K_{35} + K_{15}K_{23}K_{46} + K_{15}K_{24}K_{36} + K_{15}K_{26}K_{34} \\
&+ K_{16}K_{23}K_{45} + K_{16}K_{24}K_{35} + K_{16}K_{25}K_{34} \Big]. \qquad (2.31)
\end{aligned}$$

Experiments with Single-Layer Networks

Erf-net: $\sigma(z) = \mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z dt\, e^{-t^2}$

$$K_{\mathrm{Erf}}(x, x') = \sigma_b^2 + \sigma_W^2 \frac{2}{\pi} \arcsin\left[\frac{2(\sigma_b^2 + \frac{\sigma_W^2}{d_{\mathrm{in}}} xx')}{\sqrt{\left(1 + 2(\sigma_b^2 + \frac{\sigma_W^2}{d_{\mathrm{in}}} x^2)\right)\left(1 + 2(\sigma_b^2 + \frac{\sigma_W^2}{d_{\mathrm{in}}} x'^2)\right)}}\right]$$

Gauss-net: $\sigma(x) = \frac{\exp(W x + b)}{\sqrt{\exp\left[2(\sigma_b^2 + \frac{\sigma_W^2}{d_{\mathrm{in}}} x^2)\right]}}$

$$K_{\mathrm{Gauss}}(x, x') = \sigma_b^2 + \sigma_W^2 \exp\left[-\frac{\sigma_W^2 |x - x'|^2}{2 d_{\mathrm{in}}}\right]$$

ReLU-net: $\sigma(z) = \begin{cases} 0 & z < 0 \\ z & z \geq 0 \end{cases}$

$$\begin{aligned} K_{\mathrm{ReLU}}(x, x') &= \sigma_b^2 + \sigma_W^2 \frac{1}{2\pi} \sqrt{(\sigma_b^2 + \frac{\sigma_W^2}{d_{\mathrm{in}}} x \cdot x)(\sigma_b^2 + \frac{\sigma_W^2}{d_{\mathrm{in}}} x' \cdot x')}(\sin\theta + (\pi - \theta)\cos\theta), \\ \theta &= \arccos\left[\frac{\sigma_b^2 + \frac{\sigma_W^2}{d_{\mathrm{in}}} x \cdot x'}{\sqrt{(\sigma_b^2 + \frac{\sigma_W^2}{d_{\mathrm{in}}} x \cdot x)(\sigma_b^2 + \frac{\sigma_W^2}{d_{\mathrm{in}}} x' \cdot x')}}\right], \end{aligned}$$

Q . Measure experimental falloff to GP correlation functions (theoretical predictions) as N → ∞ ?

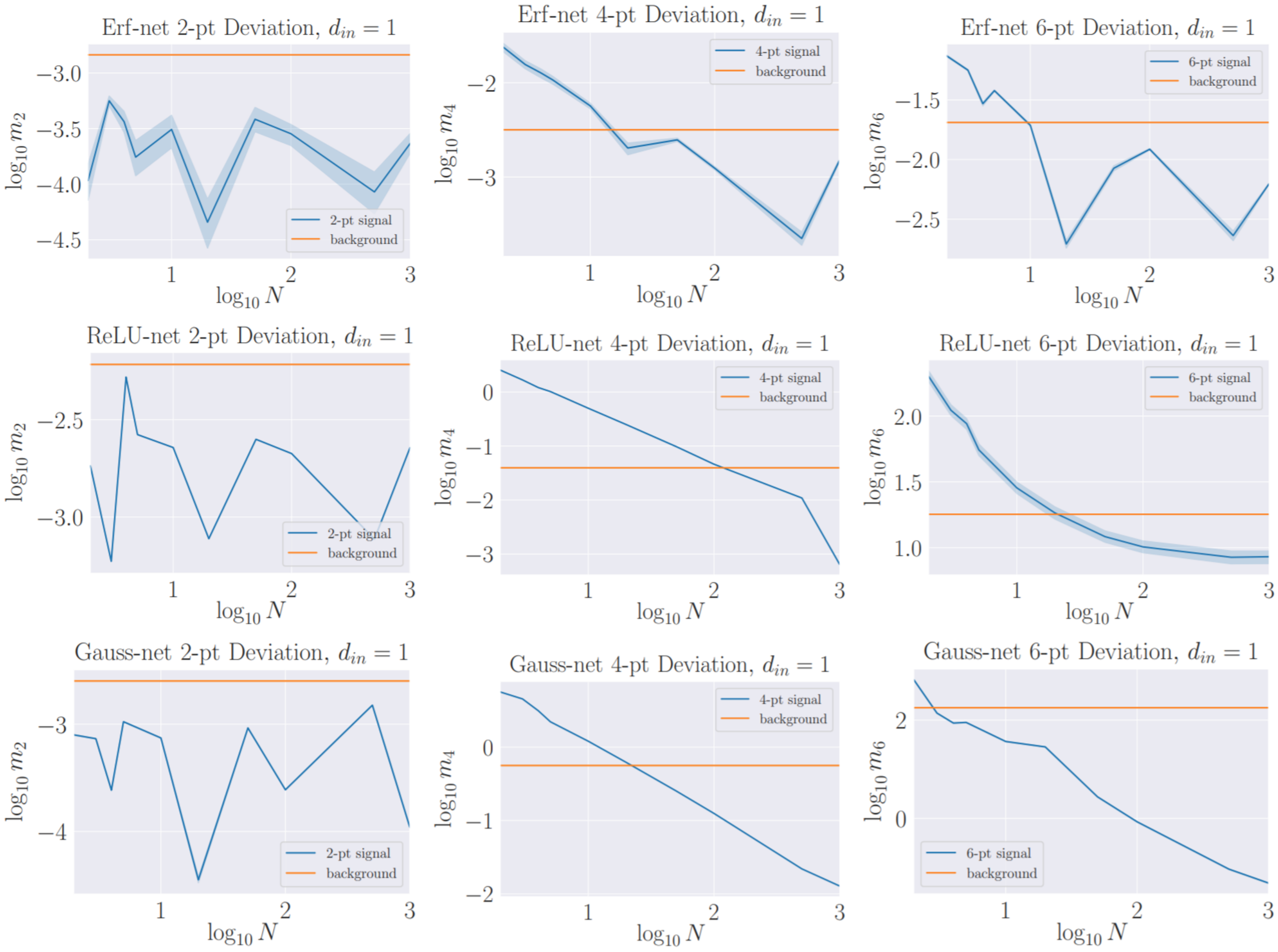
	inputs $\{x_i\}$	(σ_W^2, σ_b^2)
Gauss-net	$\{-0.01, -0.006, -0.002, +0.002, +0.006, +0.01\}$	$(1, 1)$
Erf-net	$\{-1, -0.6, -0.2, +0.2, +0.6, +1\}$	$(1, 1)$
ReLU-net	$\{+0.2, +0.4, +0.6, +0.8, +1.0, +1.2\}$	$(1, 0)$

Specifications of experiments :

$N \in \{2, 3, 4, 5, 10, 20, 50, 100, 500, 1000\}$

10 experiments of 10^6 neural nets each.

Experimental Falloff to GP Predictions



$$m_n = \Delta G^{(n)} / G^{(n)}_{\mathrm{GP}}$$

Correlation functions = Ensembled average across 10 expts

Background := average of standard deviation of m_n across 10 expts

GP kernel is exact 2-pt function at all widths

For n>2 , experimentally determined scaling $\Delta G^{(n)} \propto N^{-1}$

GP for asymptotic NNs different than Free Field theory GP

Neural Networks, Non-Gaussian Processes, and Effective Field Theory

At finite N, the NN distribution must receive 1/N suppressed non-Gaussian corrections.

Essence of perturbative Field Theory, “turning on interactions.”

Non-Gaussian Process “=” Effective Field theory

Finite N networks that admit a GP limit should be drawn from non-Gaussian process. (NGP)

$$S = S_{\text{GP}} + \Delta S$$

in general

$$\Delta S = \int d^{d_{\text{in}}}x \; [g\, f(x)^3 + \lambda\, f(x)^4 + \alpha\, f(x)^5 + \kappa\, f(x)^6 + \dots]$$

such non-Gaussian terms are interactions in QFT, with coefficients = “couplings.”

Wilsonian EFT Rules for NGPs

- Determine the symmetries (or desired symmetries) respected by the system of interest.
- Fix an upper bound k on the dimension of any operator appearing in ΔS.
- Define ΔS to contain all operators of dimension $\leq k$ that respect the symmetries.

NGP / finite NN	Interacting QFT
inputs $(x_1, \dots, x_k)$	external space or spacetime points
kernel $K(x_1, x_2)$	free or exact propagator
network output $f(x)$	interacting field
log probability	effective action S

Single-layer finite width networks : Odd-pt functions vanish (experimentally) → odd couplings vanish.

$$S = S_{\text{GP}} + \int d^{d_{\text{in}}}x \; [\lambda\, f(x)^4 + \kappa\, f(x)^6]$$

In Wilsonian sense, κ more irrelevant than λ , can be ignored in expts. *even simpler NGP distribution*.

More parameters in NN means fewer in EFT, due to “irrelevance” of operators, in Wilsonian sense.

NGP Correlation Functions by Feynman Diagrams

Compute correlation functions of NN outputs using Feynman diagrams by EFT.

$$G^{(n)}(x_1,\dots,x_n)=\frac{\int df\ f(x_1)\dots f(x_n)\,e^{-S}}{Z_0}$$

$$=\frac{\int df\ f(x_1)\dots f(x_n)\left[1-\int d^{d_{\rm in}}x\,g_k f(x)^k+O(g_k^2)\right]\ e^{-S_{\rm GP}}/Z_{\rm GP,0}}{\int df\ \left[1-\int d^{d_{\rm in}}x\,g_k f(x)^k+O(g_k^2)\right]\ e^{-S_{\rm GP}}/Z_{\rm GP,0}}$$

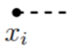
Note : exact 2-pt correlations of NGP indicate a different GP than usual free field theory.

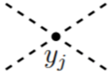
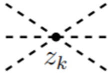
Couplings : constants or functions?

Use technical naturalness by ‘t Hooft.

Conjecture: couplings in NGPs associated to neural network architectures are constants (or nearly constants) if the kernel $K(x,y)$ associated with their GP limit is translationally invariant.

Feynman Rules:

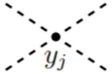
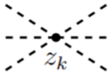
1) For each of the n external points x_i, draw 
 .

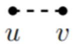
2) For each y_j, draw 
 . For each z_k, draw 
 .

3) Determine all ways to pair up the loose ends associated to x_i's, y_j's, and z_k's. This will yield some number of topologically distinct diagrams. Draw them with dashed lines.

4) Write a sum over the diagrams with an appropriate combinatoric factor out front, which is the number of ways to form that diagram. Each diagram corresponds to an analytic term in the sum.

4.5) Throw away any diagram that has a component with a λ- or κ correction to the 2-pt function.

5) For each diagram, write $-\int d^{d_{\rm in}}y_j\,\lambda$ for each 
 , and $-\int d^{d_{\rm in}}z_k\,\kappa$ for each 
 .

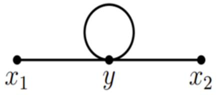
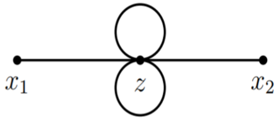
6) Write $K(u,v)$ for each 
 .

7) Throw away any terms containing vacuum bubbles.

$$\lambda(x)=\bar{\lambda}+\delta\lambda(x)$$

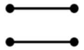
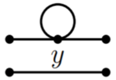
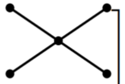
In our cases, GP kernel of Gauss-net is the only T-invt one, and only example with coupling *constants.*

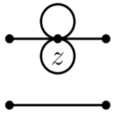
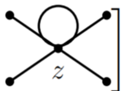
2-pt, 4-pt, and 6-pt Correlation Functions

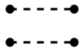
$G^{(2)}(x_1,x_2) =$  $- \lambda \Big[12$  $\Big] - \kappa \Big[90$  $\Big]$

$=$ 

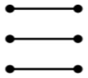
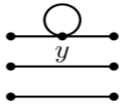
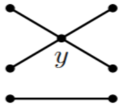
$$= K(x_1,x_2), \qquad (3.17)$$

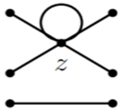
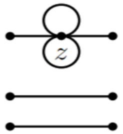
$G^{(4)}(x_1,x_2,x_3,x_4) = 3$  $- \lambda \Big[72$  $+ 24$  $\Big]$

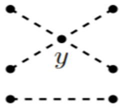
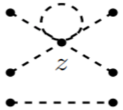
$- \kappa \Big[540$  $+ 360$  $\Big]$

$= 3$  $- 24\,\lambda$  $- 360\,\kappa$ 

$$\begin{aligned}
&= K(x_1,x_2)K(x_3,x_4) + K(x_1,x_3)K(x_2,x_4) + K(x_1,x_4)K(x_2,x_3) \\
&- 24 \int d^{d_{\text{in}}}y\, \lambda\, K(x_1,y)K(x_2,y)K(x_3,y)K(x_4,y) \\
&- 360 \int d^{d_{\text{in}}}z\, \kappa\, K(x_1,z)K(x_2,z)K(x_3,z)K(x_4,z)K(z,z) \qquad (3.18)
\end{aligned}$$

$G^{(6)}(x_1,x_2,x_3,x_4,x_5,x_6) = 15$  $- \lambda \Big[540$  $+ 360$  $\Big]$

$- \kappa \Big[720$  $+ 5400$  $+ 4050$  $\Big]$

$= 15$  $- 360\,\lambda$  $- \kappa \Big[720$  $+ 5400$  $\Big]$

$$\begin{aligned}
= &\Big[K_{12}K_{34}K_{56} + K_{12}K_{35}K_{46} + K_{12}K_{36}K_{45} + K_{13}K_{24}K_{56} + K_{13}K_{25}K_{46} + K_{13}K_{26}K_{45} + K_{14}K_{23}K_{56} \\
+ &K_{14}K_{25}K_{36} + K_{14}K_{26}K_{35} + K_{15}K_{23}K_{46} + K_{15}K_{24}K_{36} + K_{15}K_{26}K_{34} + K_{16}K_{23}K_{45} + K_{16}K_{24}K_{35} \\
+ &K_{16}K_{25}K_{34} \Big] - 24 \int d^{d_{\text{in}}}y\, \lambda \Big[K_{1y}K_{2y}K_{3y}K_{4y}K_{56} + K_{1y}K_{2y}K_{3y}K_{5y}K_{46} + K_{1y}K_{2y}K_{4y}K_{5y}K_{36} \\
+ &K_{1y}K_{3y}K_{4y}K_{5y}K_{26} + K_{2y}K_{3y}K_{4y}K_{5y}K_{16} + K_{1y}K_{2y}K_{3y}K_{6y}K_{45} + K_{1y}K_{2y}K_{4y}K_{6y}K_{35} \\
+ &K_{1y}K_{3y}K_{4y}K_{6y}K_{25} + K_{2y}K_{3y}K_{4y}K_{6y}K_{15} + K_{1y}K_{2y}K_{5y}K_{6y}K_{34} + K_{1y}K_{3y}K_{5y}K_{6y}K_{24} \\
+ &K_{2y}K_{3y}K_{5y}K_{6y}K_{14} + K_{1y}K_{4y}K_{5y}K_{6y}K_{23} + K_{2y}K_{4y}K_{5y}K_{6y}K_{13} + K_{3y}K_{4y}K_{5y}K_{6y}K_{12} \Big] \\
- &720 \int d^{d_{\text{in}}}z\, \kappa\, K_{1z}K_{2z}K_{3z}K_{4z}K_{5z}K_{6z} - 360 \int d^{d_{\text{in}}}z\, \kappa \Big[K_{zz}K_{1z}K_{2z}K_{3z}K_{4z}K_{56} \\
+ &K_{zz}K_{1z}K_{2z}K_{3z}K_{5z}K_{46} + K_{zz}K_{1z}K_{2z}K_{4z}K_{5z}K_{36} + K_{zz}K_{1z}K_{3z}K_{4z}K_{5z}K_{26} \\
+ &K_{zz}K_{2z}K_{3z}K_{4z}K_{5z}K_{16} + K_{zz}K_{1z}K_{2z}K_{3z}K_{6z}K_{45} + K_{zz}K_{1z}K_{2z}K_{4z}K_{6z}K_{35} \\
+ &K_{zz}K_{1z}K_{3z}K_{4z}K_{6z}K_{25} + K_{zz}K_{2z}K_{3z}K_{4z}K_{6z}K_{15} + K_{zz}K_{1z}K_{2z}K_{5z}K_{6z}K_{34} \\
+ &K_{zz}K_{1z}K_{3z}K_{5z}K_{6z}K_{24} + K_{zz}K_{2z}K_{3z}K_{5z}K_{6z}K_{14} + K_{zz}K_{1z}K_{4z}K_{5z}K_{6z}K_{23} \\
+ &K_{zz}K_{2z}K_{4z}K_{5z}K_{6z}K_{13} + K_{zz}K_{3z}K_{4z}K_{5z}K_{6z}K_{12} \Big], \qquad (3.19)
\end{aligned}$$

EFT is Effective: Measure Couplings, Verify Predictions

EFT: *Effective* at describing experimental system

- Give a candidate ΔS for the NGP.
- Fix coefficients of operators in ΔS with experiments.
- Once fixed, make predictions for other experiments and verify them.

Case, λ constant: measure from 4-pt function expts

$$\lambda = \frac{K(x_1,x_2)K(x_3,x_4) + K(x_1,x_3)K(x_2,x_4) + K(x_1,x_4)K(x_2,x_3) - G^{(4)}(x_1,x_2,x_3,x_4)}{24\int d^{d_{\rm in}}y\, K(x_1,y)K(x_2,y)K(x_3,y)K(x_4,y)}$$

call denominator integrand $\boldsymbol{\Delta}_{1234y}$.

Case, λ function: write as constant + space varying $\lambda(y) = \bar\lambda + \delta\lambda(y)$ then we have

$$\bar\lambda = \frac{K_{12}K_{34} + K_{13}K_{24} + K_{14}K_{23} - G^{(4)}(x_1,x_2,x_3,x_4)}{24\int d^{d_{\rm in}}y\,\Delta_{1234y}} - \frac{\int d^{d_{\rm in}}y\,\delta\lambda(y)\,\Delta_{1234y}}{\int d^{d_{\rm in}}y\,\Delta_{1234y}}$$

and expression from before not constant.

When variance in

$$\lambda_m(x_1,x_2,x_3,x_4) := \frac{K_{12}K_{34} + K_{13}K_{24} + K_{14}K_{23} - G^{(4)}(x_1,x_2,x_3,x_4)}{24\int d^{d_{\rm in}}y\,\Delta_{1234y}}$$

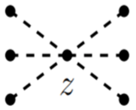
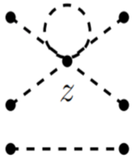
is small relative to mean, our definition of "measuring λ":

$$\lambda \simeq \bar\lambda \simeq \mathrm{mean}(\lambda_m(x_1,x_2,x_3,x_4))$$

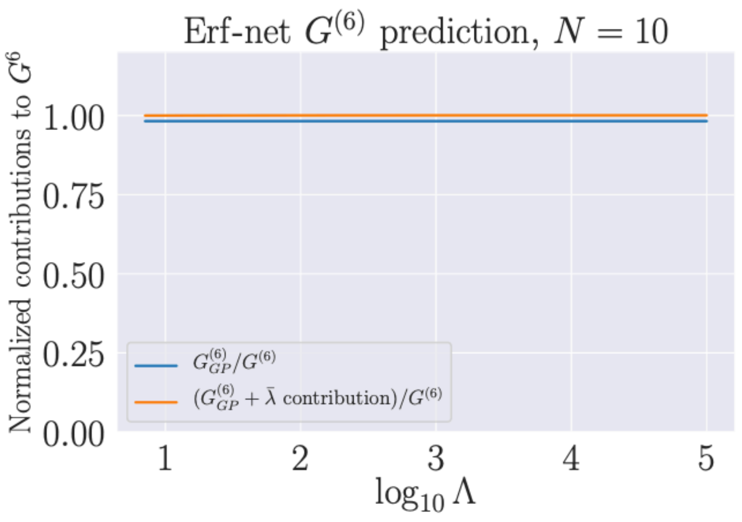
Effectiveness: Expt 6-pt - GP prediction = NGP correction

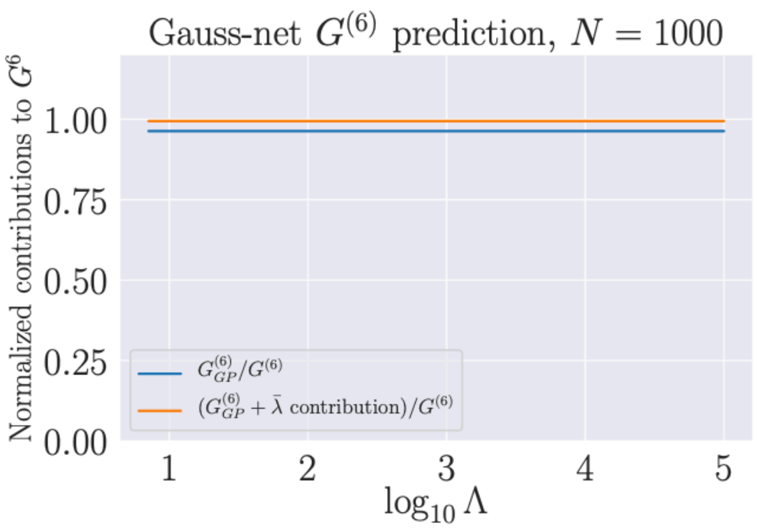
$$\begin{aligned}\delta'(x_1,\dots,x_6) \;&:=\; G^{(6)}(x_1,\dots,x_6) - \sum_{\text{15 combinations}}\Big[K(x_i,x_j)K(x_k,x_l)K(x_m,x_n) \\ &\quad - 24\int d^{d_{\rm in}}y\;\lambda\;K(x_i,y)K(x_j,y)K(x_k,y)K(x_l,y)K(x_m,x_n)\Big]\end{aligned}$$

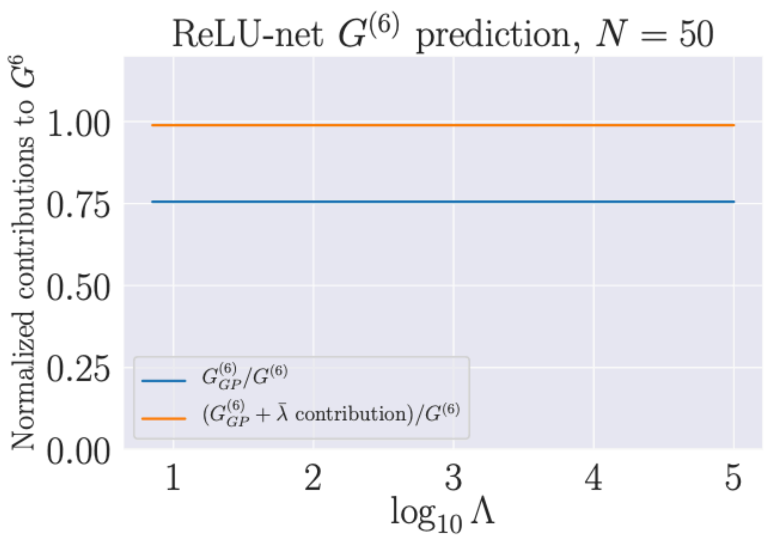
$$= -\kappa\Big[720 \;\text{(diagram)}\; + 5400\;\text{(diagram)}\Big]$$

$= -\kappa\,[\,720$  $+ 5400$  $]$

Experimental Verification of NN “=” EFT







$(G^{(6)}_{GP} + \bar\lambda \text{ contribution})/G^{(6)}$ ≈ 1 indicates EFT is effective!

Implicit $S \to S_\Lambda$, replacing one effective action with a continuous family parameterized by Λ.

Experimental NN correlation functions,

$$G^{(n)}(x_1, \ldots, x_n) = \frac{1}{n_{\text{nets}}} \sum_{\alpha \in \text{nets}}^{n_{\text{nets}}} f_\alpha(x_1) \, \ldots \, f_\alpha(x_n)$$

depend on outputs evaluated at set of inputs

$$S_{\text{in}} = \{x_1, \ldots, x_{N_{\text{in}}}\} \qquad |x_i| \ll \Lambda$$

Kernels introduce *tree-level divergence* in n-pt functions.

Regulate by sufficiently large cutoffs in effective action

$$\Delta S_\Lambda = \int_{-\Lambda}^{\Lambda} d^{d_{\text{in}}}x \sum_{l \le k} g_{\mathcal{O}_l}(\Lambda)\, \mathcal{O}_l$$

Q. One set of experiments match with infinite number of S_Λ ?

Wilsonian Renormalization

$$\frac{dG^{(n)}(x_1, \ldots, x_n)}{d\Lambda} = 0$$

Extracting β-functions from theory

NN effective actions (distributions) with different Λ may make the same predictions by absorbing the difference into couplings, ***"running couplings"***,

$$\beta(g_{\mathcal{O}_l}) := \frac{d\, g_{\mathcal{O}_l}}{d\, log\Lambda}$$

encoded in the β-functions, which capture how the couplings vary with the cutoff.

Induces a "flow" in coupling space as Λ varies, ***Wilsonian renormalization group flow***. (RG)

Extract from hitting n-pt functions, expressed using kernel functions, with derivatives.

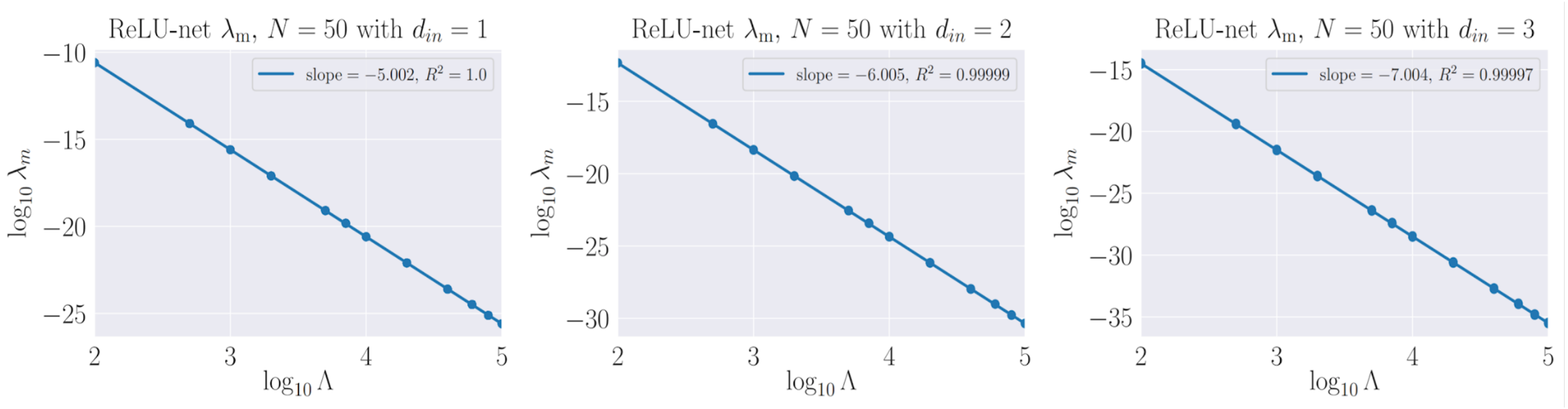
$$\frac{\partial G^{(4)}(x_1,x_2,x_3,x_4)}{\partial \log\Lambda} = 0 = \frac{\partial\lambda}{\partial\log\Lambda}\int_{-\Lambda}^{\Lambda} d^{d_{\rm in}}x\,(\gamma_{4,\lambda}+\varrho_{4,\lambda}) + \lambda\frac{\partial(\int_{-\Lambda}^{\Lambda} d^{d_{\rm in}}x\,(\gamma_{4,\lambda}+\varrho_{4,\lambda}))}{\partial\log\Lambda}$$
$$+ \frac{\partial\kappa}{\partial\log\Lambda}\int_{-\Lambda}^{\Lambda} d^{d_{\rm in}}x\,(\gamma_{4,\kappa}+\varrho_{4,\kappa}) + \kappa\frac{\partial(\int_{-\Lambda}^{\Lambda} d^{d_{\rm in}}x\,(\gamma_{4,\kappa}+\varrho_{4,\kappa}))}{\partial\log\Lambda}, \quad (4.13)$$

$$\frac{\partial G^{(6)}(x_1,x_2,x_3,x_4,x_5,x_6)}{\partial \log\Lambda} = 0 = \frac{\partial\lambda}{\partial\log\Lambda}\int_{-\Lambda}^{\Lambda} d^{d_{\rm in}}x\,(\gamma_{6,\lambda}+\varrho_{6,\lambda}) + \lambda\frac{\partial(\int_{-\Lambda}^{\Lambda} d^{d_{\rm in}}x\,(\gamma_{6,\lambda}+\varrho_{6,\lambda}))}{\partial\log\Lambda}$$
$$+ \frac{\partial\kappa}{\partial\log\Lambda}\int_{-\Lambda}^{\Lambda} d^{d_{\rm in}}x\,(\gamma_{6,\kappa}+\varrho_{6,\kappa}) + \kappa\frac{\partial(\int_{-\Lambda}^{\Lambda} d^{d_{\rm in}}x\,(\gamma_{6,\kappa}+\varrho_{6,\kappa}))}{\partial\log\Lambda} \quad (4.14)$$

Our examples: κ more irrelevant than λ, in sense of Wilson.

Extract β-function for λ from deriv. of 4-pt.

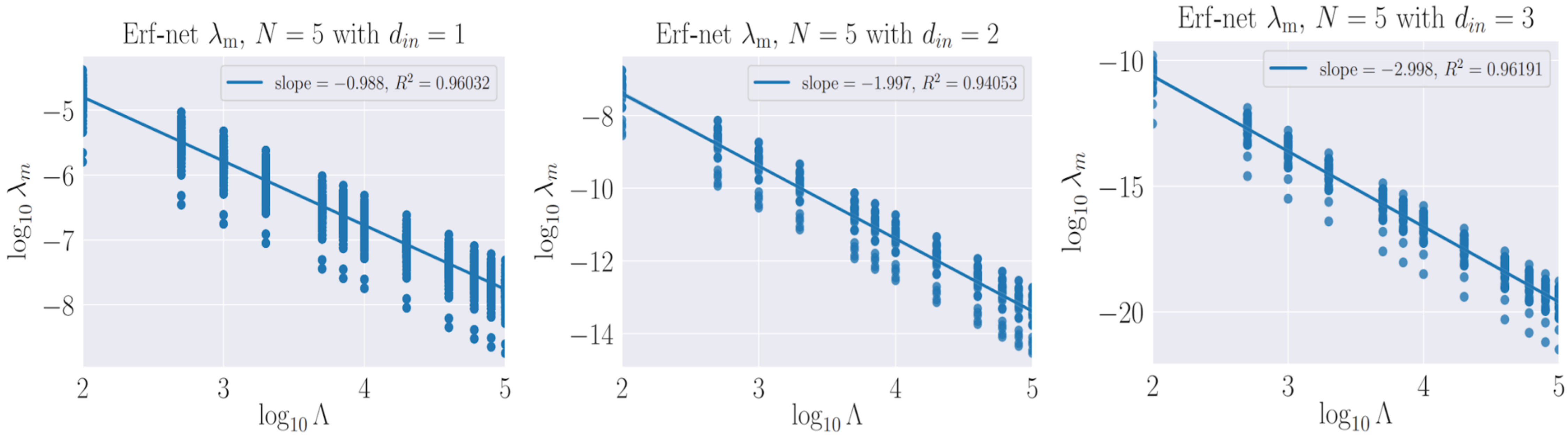
Theory vs. Experiment: ReLU-net



$$\beta(\lambda) := \frac{\partial\lambda}{\partial\log(\Lambda)} = -(d_{\rm in}+4)\lambda$$

experimentally measured d_{in}-dependent slope matches theory predictions from Wilsonian RG

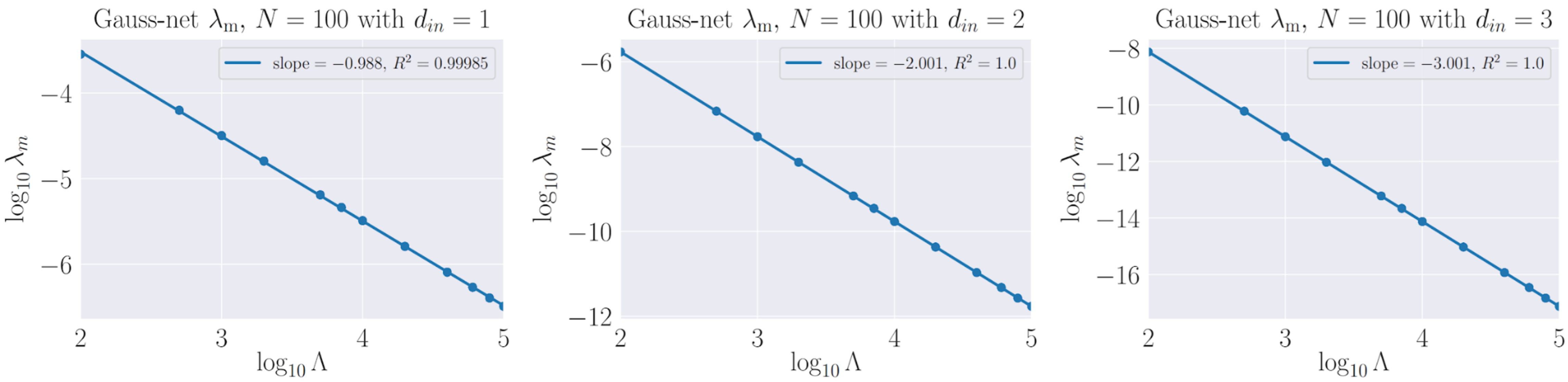
Theory vs. Experiment: Erf-net



$$\beta(\lambda) := \frac{\partial \lambda}{\partial \log \Lambda} = -\lambda\, d_{\mathrm{in}}$$

experimentally measured d_{in}-dependent slope matches theory predictions from Wilsonian RG

Theory vs. Experiment: Gauss-net



$$\beta(\lambda) := \frac{\partial \lambda}{\partial \log \Lambda} = -\lambda\, d_{\mathrm{in}}$$

experimentally measured d_{in}-dependent slope matches theory predictions from Wilsonian RG

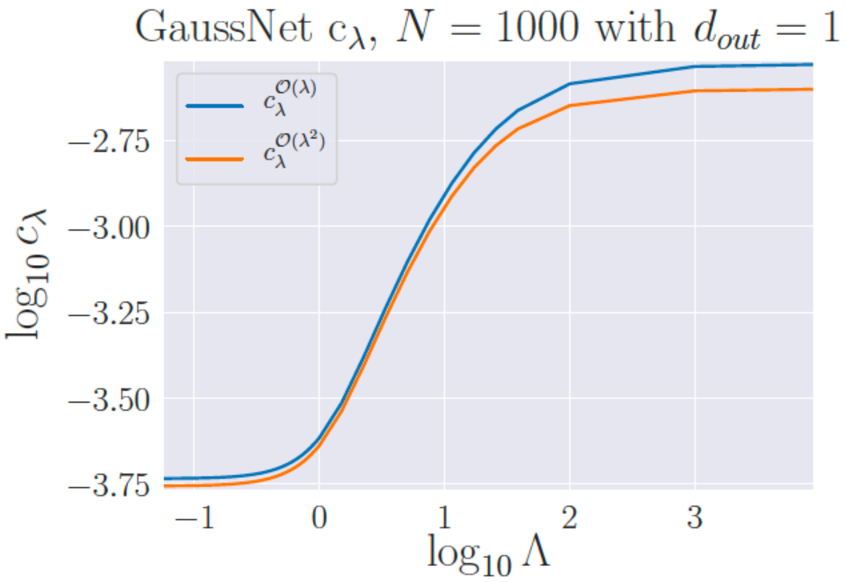
SO(d_{out}) Symmetry and Fixed Points

At arbitrary d_{out}, NN output ⟷ interacting field with d_{out} species, in EFT description.

Correlation functions SO(d_{out}) symmetric in GP limit.

$$G^{(4)}_{ijkl}(x_1,x_2,x_3,x_4) = K_{12}K_{34}\,\delta_{ij}\delta_{kl} + K_{13}K_{24}\,\delta_{ik}\delta_{jl} + K_{14}K_{23}\,\delta_{il}\delta_{jk}$$
$$- 8\int d^{d_{\rm in}}x\,\lambda\,K_{1x}K_{2x}K_{3x}K_{4x}(\delta_{ij}\delta_{kl}+\delta_{ik}\delta_{jl}+\delta_{il}\delta_{jk})$$
$$+ 32\int d^{d_{\rm in}}x d^{d_{\rm in}}y\,\frac{\lambda^2}{2}\,K^2_{xy}$$
$$\times\left[(K_{1x}K_{2x}K_{3y}K_{4y}+x\leftrightarrow y)\left((d_{\rm out}+4)\delta_{ij}\delta_{kl}+2(\delta_{ik}\delta_{jl}+\delta_{il}\delta_{jk})\right)\right.$$
$$+ (K_{1x}K_{3x}K_{2y}K_{4y}+x\leftrightarrow y)\left((d_{\rm out}+4)\delta_{ik}\delta_{jl}+2(\delta_{ij}\delta_{kl}+\delta_{il}\delta_{jk})\right)$$
$$\left.+ (K_{1x}K_{4x}K_{2y}K_{3y}+x\leftrightarrow y)\left((d_{\rm out}+4)\delta_{il}\delta_{jk}+2(\delta_{ij}\delta_{kl}+\delta_{ik}\delta_{jl})\right)\right],$$

Many architectures have universal UV fixed point for dimension-less coupling $c_\lambda = \lambda\Lambda^{d_{\rm in}}$



λ extracted at quadratic order receives $1/d_{out}$ corrections

Higher d_{out} suppresses leading non-Gaussian coefficients

Additionally, Gauss-net c_λ also approaches a fixed point at IR

Conclusion

- NN-QFT Correspondence works!

- As NN gets more and more parameters towards GP limit, fewer and fewer important non-Gaussian coefficients in Field Theory distribution.

- Wilsonian RG: limiting Λ, can ignore even more coefficients; even fewer important coefficients in distribution of highly parameterized NNs.

- Increasing d_{out} decreases magnitude of leading non-Gaussian coefficients.

- **Particularly acute in our experiments:** a *single number* can correct GP to NGP correlation functions, although in moving away from the GP limit an infinite # of NN parameters are lost.

- "Supervised learning" is just learning the 1-pt function ≈ symmetry breaking.

Thank You