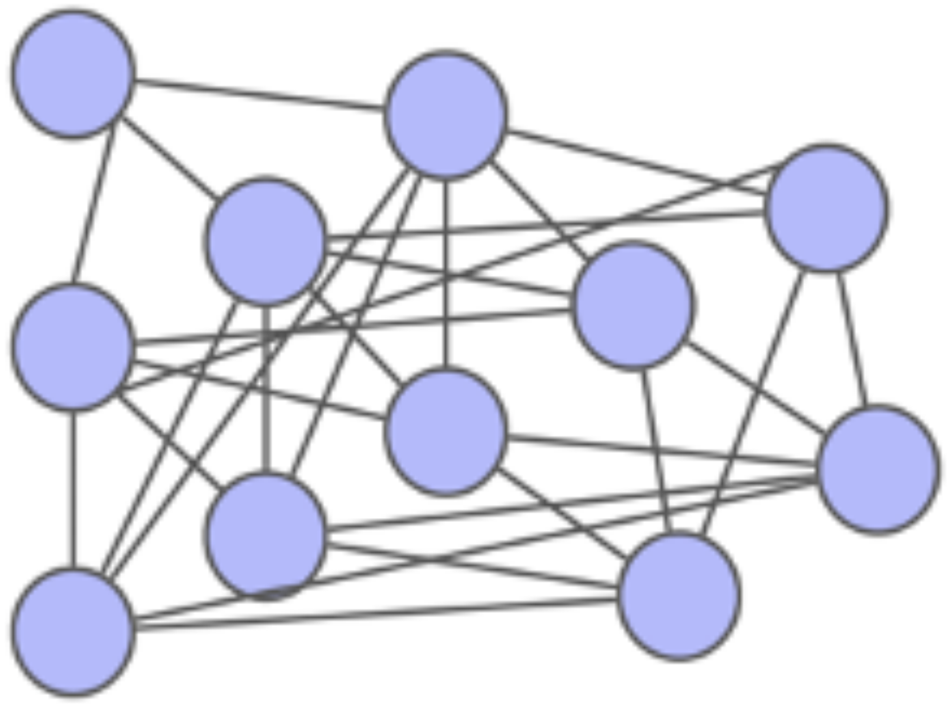
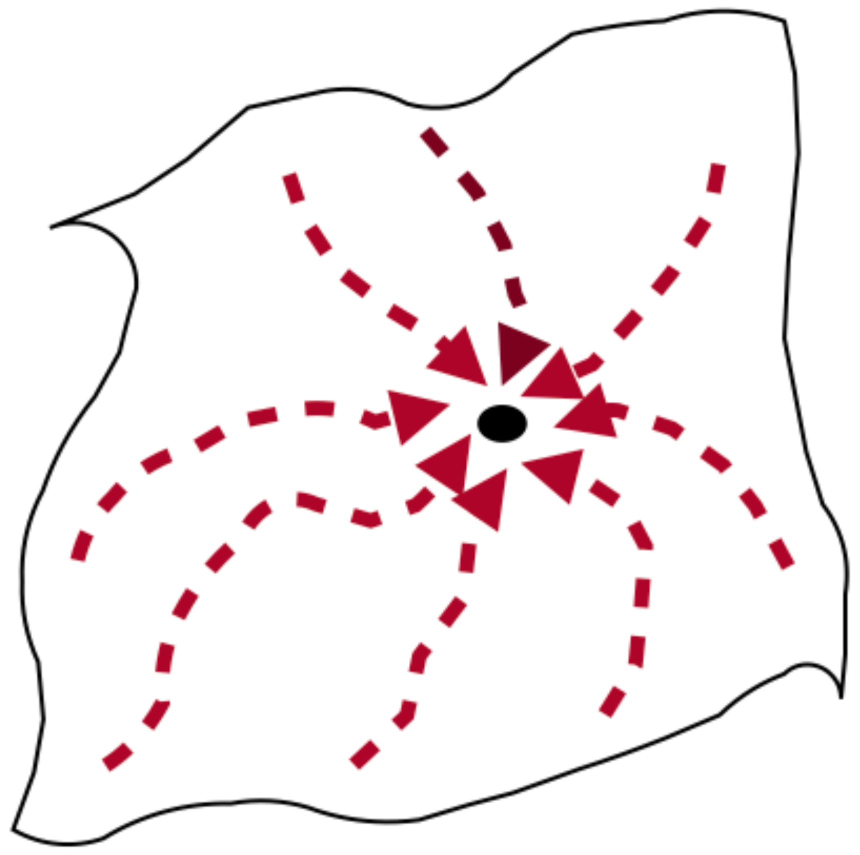
A Wilsonian RG framework for Regression Tasks in Machine Learning



Anindita Maiti

Email: amaiti@perimeterinstitute.ca

Probing the Frontiers of Nuclear Physics with AI at the EIC (II) 2025



PERIMETER INSTITUTE

Postdoctoral Fellow

QUANTUM INTELLIGENCE LAB

Based on arxiv 2405:06008v2 with Z. Ringel, R. Jefferson, J. N. Howard

20 Mar 2025

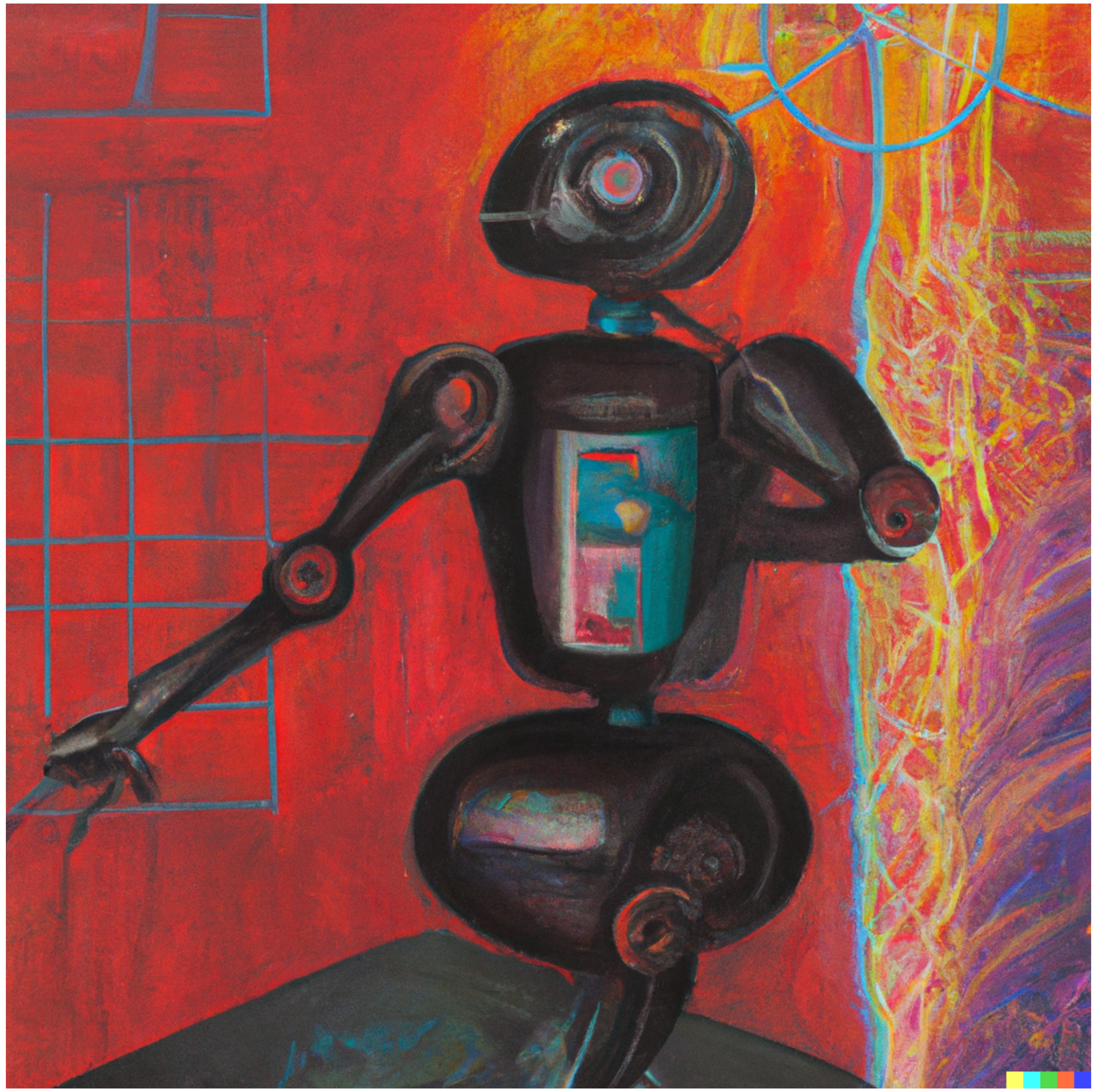
CFNS, Stony Brook University

Q. Are standard ML / AI packages effective for theoretical physics (particle / nuclear / quantum etc)?

Short answer: if those ML / AI models are trustworthy for fundamental physics, then yes.

What makes ML / AI trustworthy?

- ➡ Precision
- ➡ Interpretability
- ➡ Robustness
- ➡ Reliability (including fairness, ethics) etc.



What makes ML precise, interpretable & robust?

Precision: depends on an ability to recognize order of relevance among data features.

❖ If some data is coarse grained, trustworthy ML / AI can auto tune results according to data relevance.

Interpretability: depends on an ability to learn about the target following some rule(s) / pattern(s) / algorithm(s), rather than ad hoc data matching.

❖ Trustworthy ML / AI is much more than a glorified data fitting tool.

Robustness: depends on an ability to autofill for missing or noisy information.

❖ Trustworthy ML / AI can autofill or predict missing data following complex, hidden patterns.

Q. Shouldn't ensuring AI /ML trust be the job of AI industry, or computer scientists?

Short answer: modern state-of-the-art ML / AI models are stories of empirical success, with very little support to their reliability, across various application domains.

- Research from AI industry or computer scientists are hard to translate into language of theoretical physics.
- Still a nascent field, needs more theoretical tools for foundation.
- Performance can break down with drastic changes in data length scale!

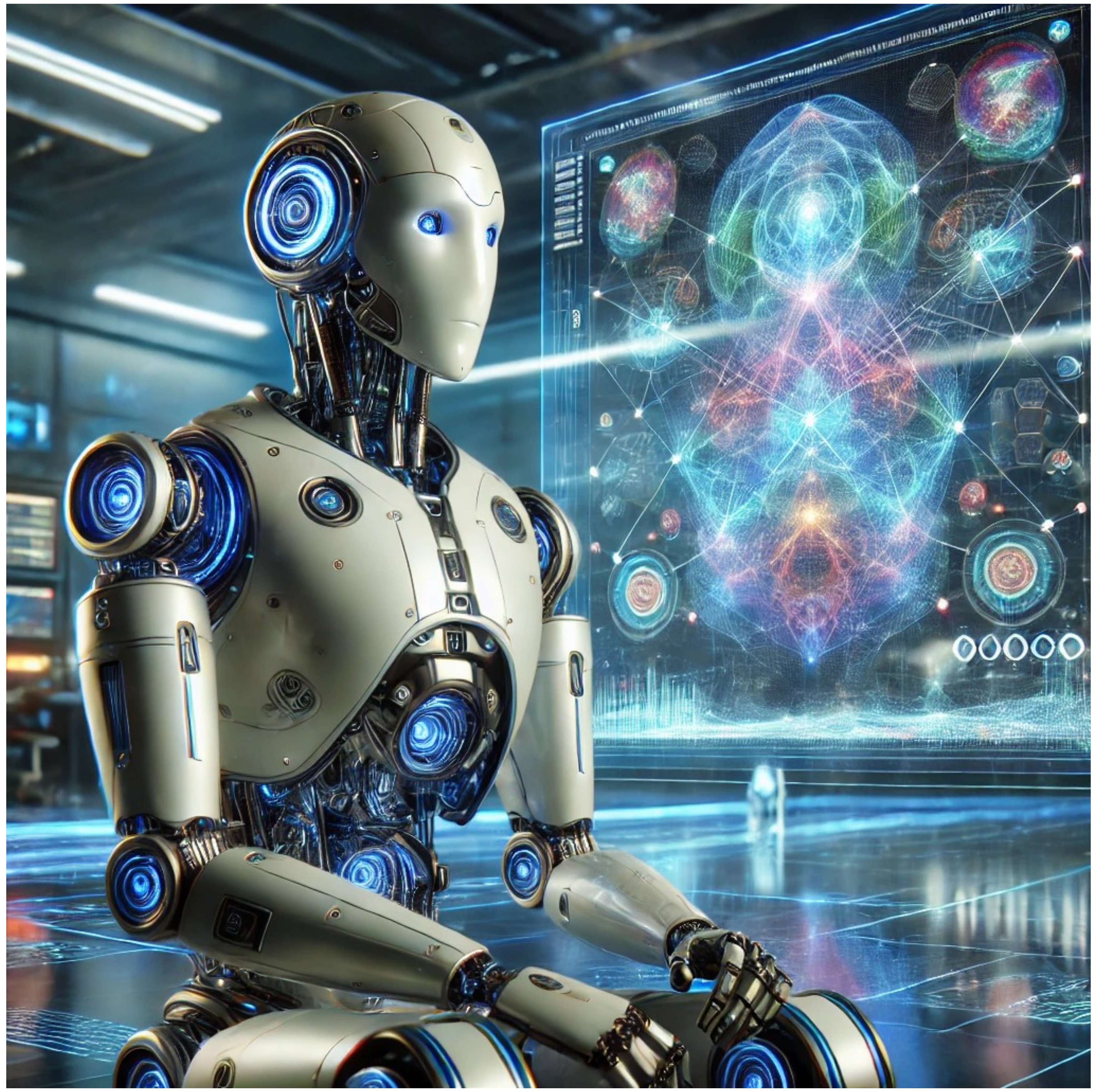
Why Wilsonian RG for trustworthy ML?

Can we track how precision of ML outputs depends on data attributes?

- Locate a separatrix in data feature space

- Study dynamics of ML output noise, as a means to precision, as a function of this separatrix (scale).

But those are same as steps in RG!



Related works: RG meets ML

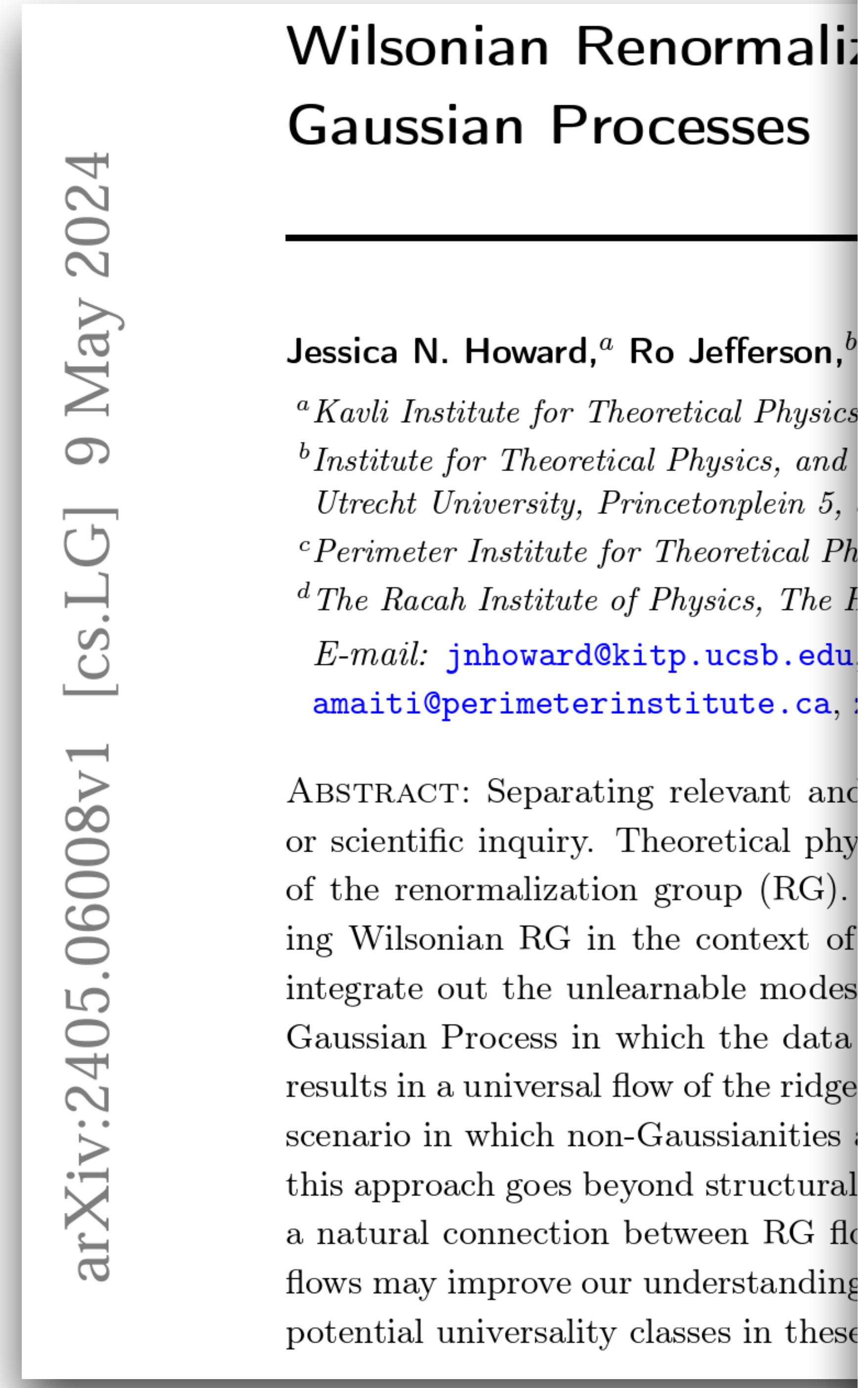
Bayesian inference, optimal transport, and diffusion processes modeled after RG.

[Cotler, Rezchikov 2022], [Cotler, Rezchikov 2023], [Berman, Heckman, Klinger 2022], [Berman, Klinger 2022], [Berman, Klinger, Stapleton 2023], [Berman, Klinger, Stapleton 2024], [Cheng, Gerdes, ... 202x]

Finite width / depth effects in initialized DNN ensembles in terms of RG.

[Halverson, AM, Stoner 2020], [Erbin, Lahoche, O. Samary 2021], [Erbin, Lahoche, O. Samary 2022], [Erbin, Finotello, Kprera, Lahoche, O. Samary 2023], [Grosvenor, Jefferson 2021], [Roberts, Yaida, Hanin 2021], [Erdmenger, Grosvenor, Jefferson 2021], [Banta, Cai, Craig, Zhang 2023]

RG to explain ML output quality (precision and noise)


arXiv:2405.06008v1 [cs.LG] 9 May 2024

Wilsonian Renormali
Gaussian Processes

Jessica N. Howard,[a] Ro Jefferson,[b]

[a] Kavli Institute for Theoretical Physics
[b] Institute for Theoretical Physics, and
Utrecht University, Princetonplein 5,
[c] Perimeter Institute for Theoretical Ph
[d] The Racah Institute of Physics, The H

E-mail: jnhoward@kitp.ucsb.edu
amaiti@perimeterinstitute.ca,

Abstract: Separating relevant and
or scientific inquiry. Theoretical phy
of the renormalization group (RG).
ing Wilsonian RG in the context of
integrate out the unlearnable modes
Gaussian Process in which the data
results in a universal flow of the ridge
scenario in which non-Gaussianities
this approach goes beyond structural
a natural connection between RG flo
flows may improve our understanding
potential universality classes in these

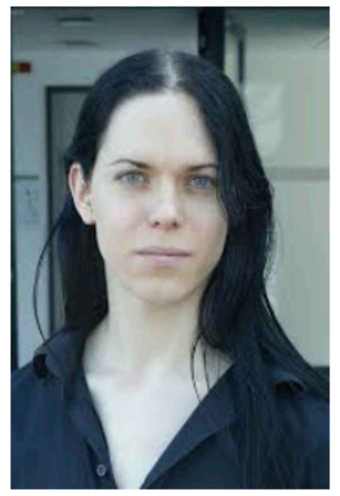
Acknowledgement

Zohar Ringel

Hebrew University of Jerusalem

Ro Jefferson

Utretch University



Jessica N. Howard

Kavli Institute for Theoretical Physics (UCBS)

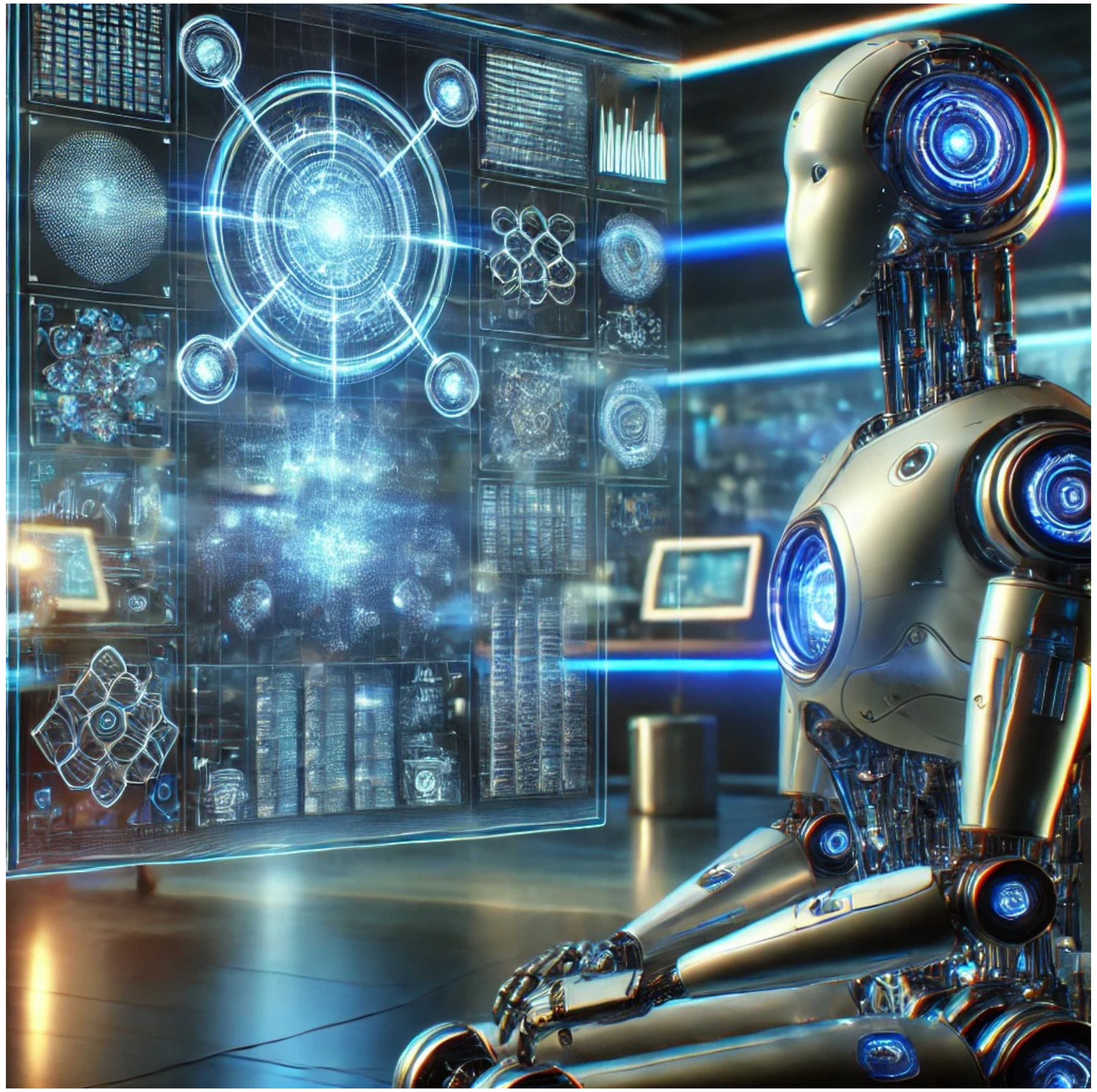
Talk outline

I. Neural Network Gaussian Process Regression

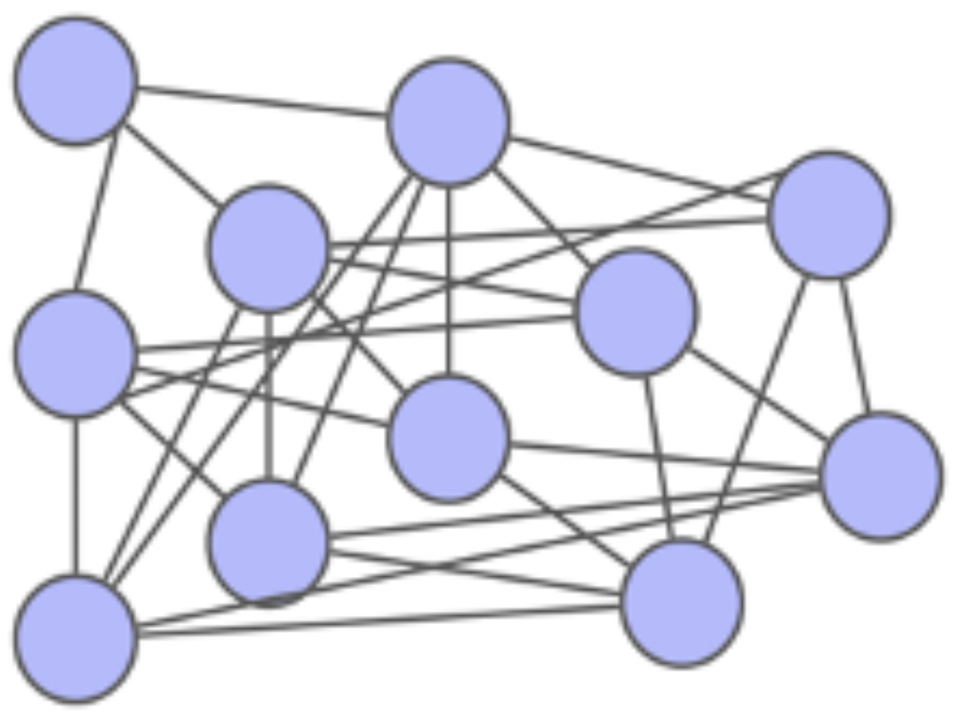
II. Wilsonian RG framework

III. Universal RG flows

IV. Functional RG flows



1. Neural Network Gaussian Process (NNGP) Regression



Neural Network Gaussian Process Regression

Overparameterized NNs produce noisy outputs, even when nicely trained.

$$f(x) \sim \boxed{y(x) + \epsilon}$$

Target

Noise in output

Covariance of noise is called 'ridge'

$\epsilon \sim \mathcal{N}(0,\sigma^2)$

$$\underset{x_*}{\boxed{\text{Test data}}} \longrightarrow \underset{y(x)}{\boxed{\text{Target}}}$$

Noise evolves if data features / information is removed.

Neural Network Gaussian Process Regression

If data features get coarse grained, average prediction of trained NN becomes more noisy, i.e. precision reduces.

- Wilsonian RG framework can track evolution of noise.
- From noise dominated learning regime, one can go bottom-up to find regimes where NN predictions are more precise and trustable.

Average predictor obtained using replica partition function, then setting $\lim M \to 0$.

$$\langle Z^M \rangle_\eta = e^{-\eta} \int \prod_{m=1}^{M} \mathcal{D} f_m e^{-S}$$

Neural Network Gaussian Process Regression

Relevant & irrelevant data features interact in replica action

$$S = \sum_{m=1}^{M} \frac{1}{2} \int \mathrm{d}\mu_x \mathrm{d}\mu_{x'}\, f_m(x) K^{-1}(x,x') f_m(x') - \eta \int \mathrm{d}\mu_x\, e^{-\sum_{m=1}^{M} \frac{(f_m(x)-y(x))^2}{2\sigma^2}}$$

MSE loss term

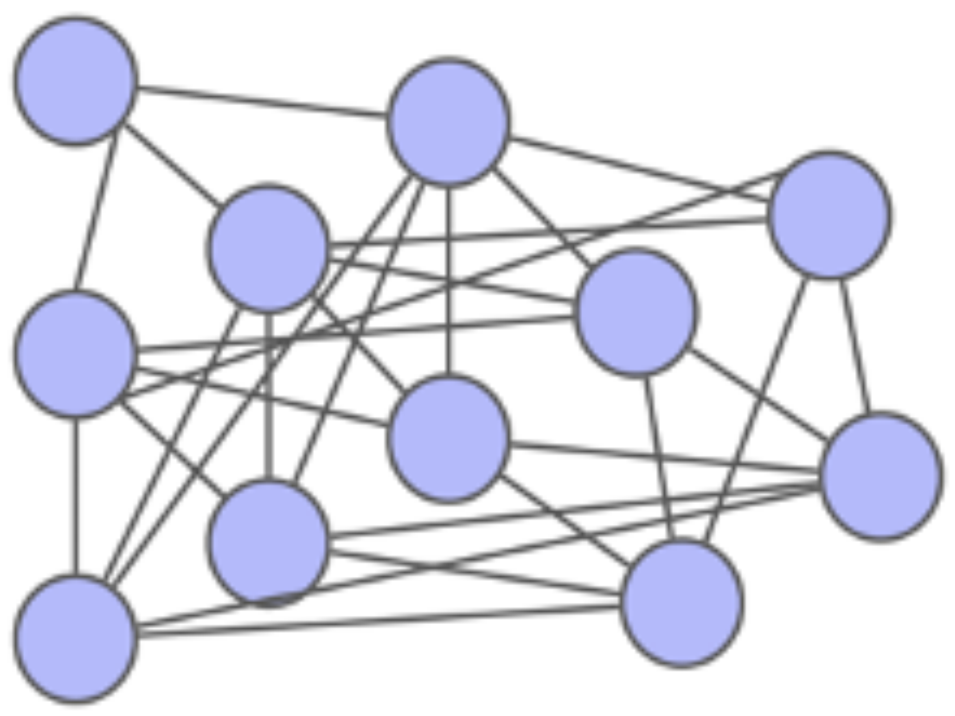
When $\eta/\sigma^2 \ll 1$ + Spectral decomposition in NNGP kernel eigenspace.

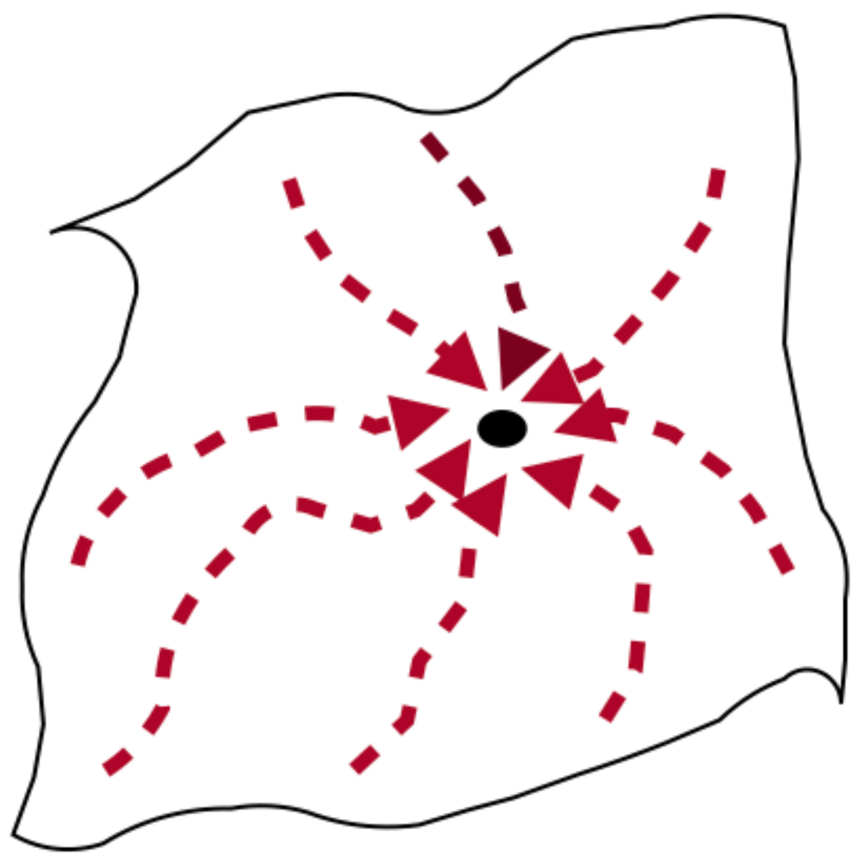
$$f_m(x) = \sum_{k=1}^{\infty} f_{mk} \phi_k(x) \qquad y(x) = \sum_{k=1}^{\infty} y_k \phi_k(x)$$

NNGP kernel eigenfunctions / **feature modes**

GP modes

II. Wilsonian RG framework



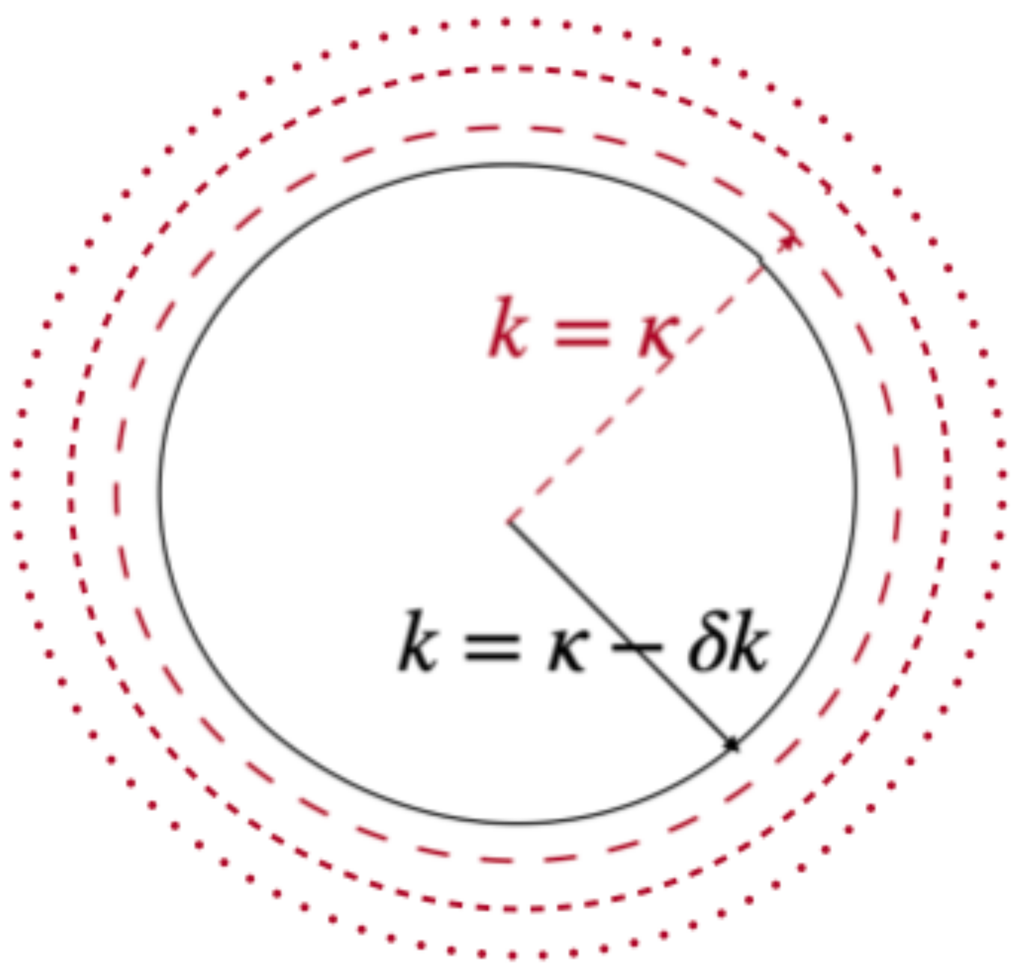


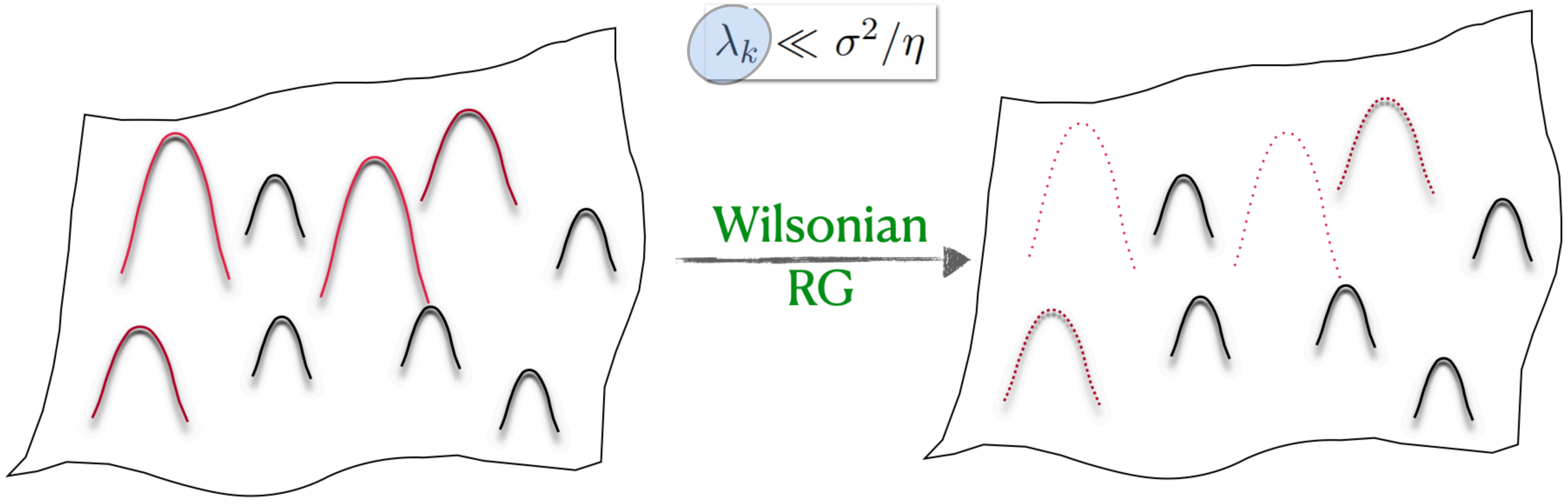
Wilsonian RG framework

Momentum shell RG to coarse grain irrelevant features from interacting replica action

Kernel eigenvalues

- Data sets IR cutoff κ.
- ϕ_k with low λ_k correspond to high momentum modes.





Feature modes $k = 1, \cdots, \kappa$

Feature modes
$k = 1, \cdots, \kappa - \delta k$

Wilsonian RG framework

Coarse graining of features renormalize noise $\epsilon \sim \mathcal{N}(0,\sigma^2)$ in predictor.

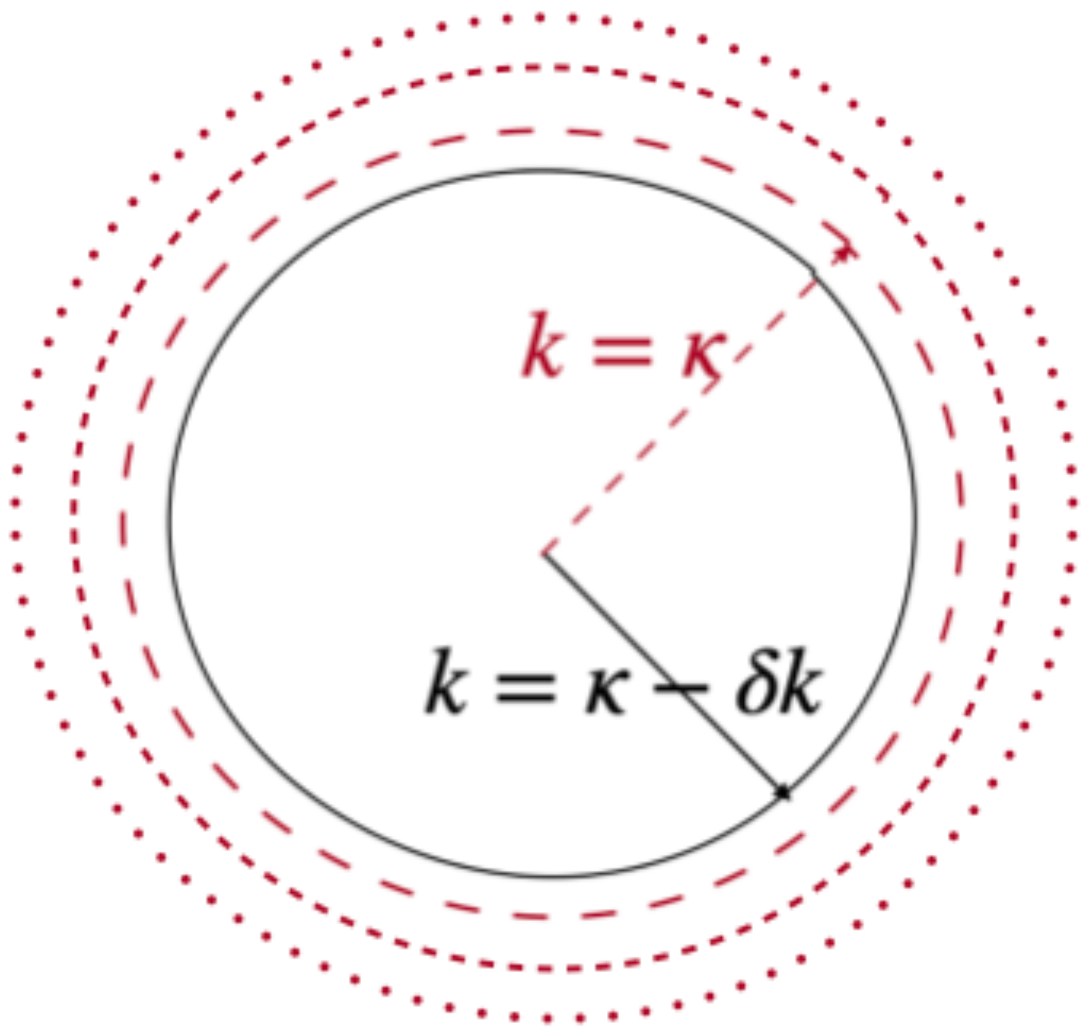
$$\langle Z^M \rangle_\eta = e^{-\eta} \int \prod_m \mathcal{D} f_{m<}\, e^{-S_0[f_{m<}]} \int \prod_m \mathcal{D} f_{m>}\, e^{-S_0[f_{m>}] - S_{\text{int}}[f_{m<}, f_{m>}]}$$

RG flow of ridge parameter σ^2.

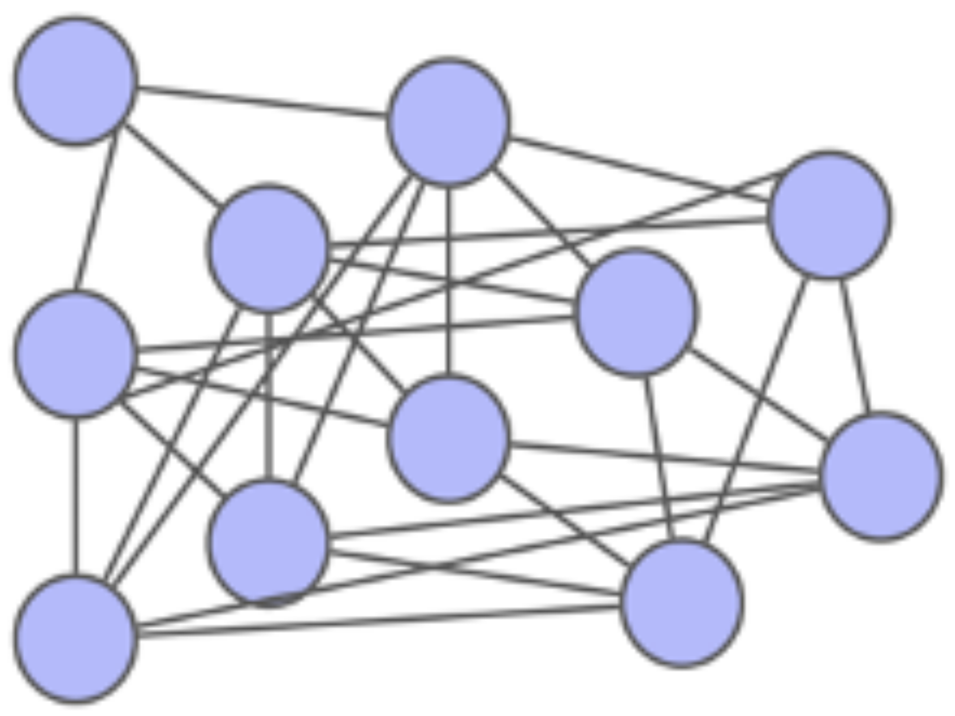
$$\langle Z^M \rangle_\eta = e^{-\eta} \int \prod_m \mathcal{D} f_{m<}\, e^{-S_{\text{eff}}[f_{m<}]}$$

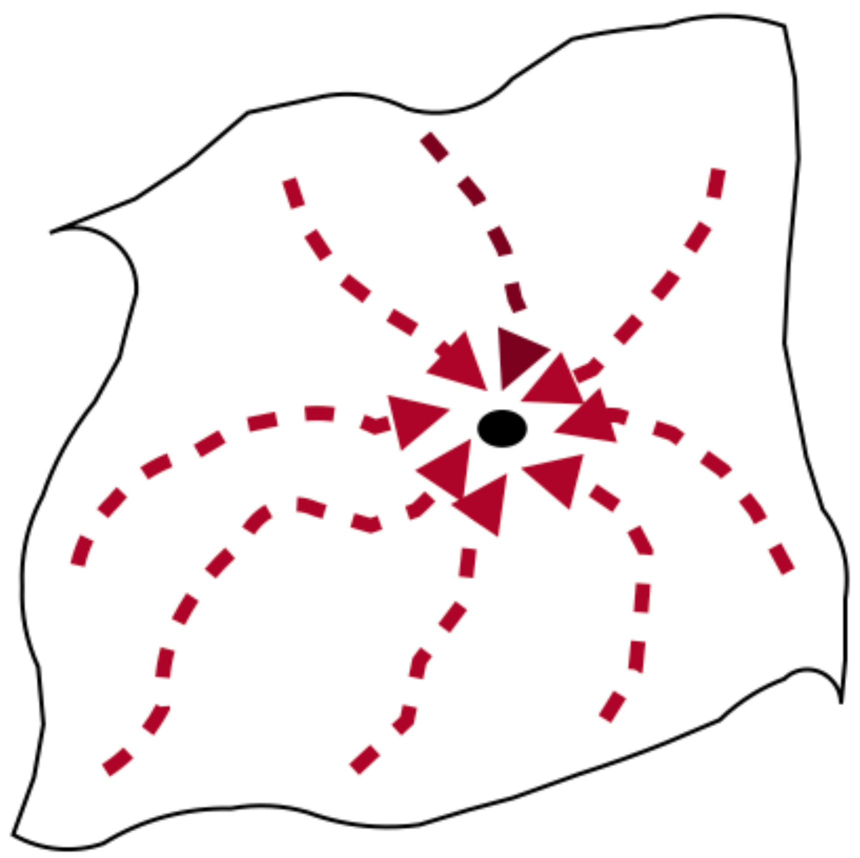
Induces flow of ML predictor & its precision.

$m <$ denotes all GP modes with $m, k < \kappa$.



III. Universal RG Flows





Gaussian irrelevant features

Step 1. Integrate Gaussian higher feature modes $\phi_{k>\kappa}$.

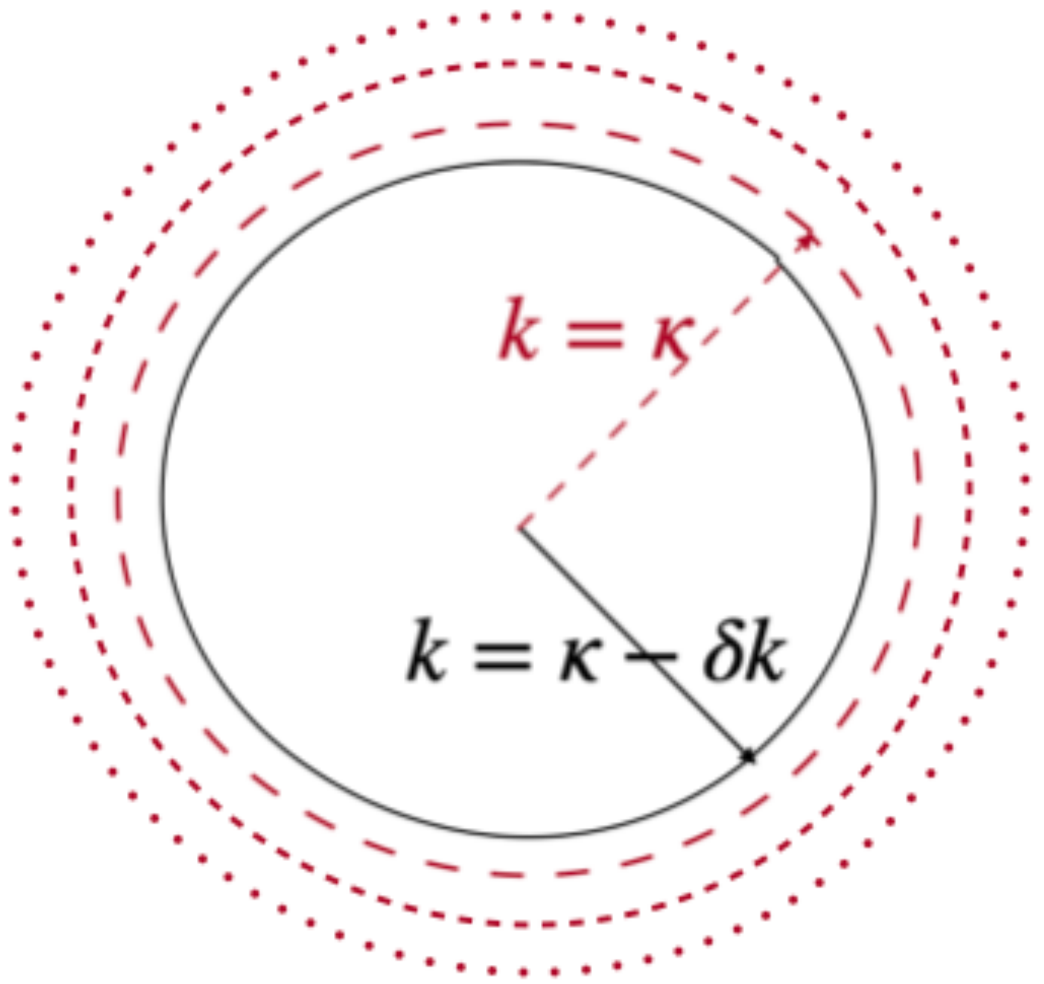
Step 2. Integrate higher GP modes $f_{mk>\kappa}$ (always Gaussian).

$$P[\varphi_>|\varphi_<] = P[\varphi_>] = \mathcal{N}[0, \mathbb{1}; \varphi_>]$$

$$\langle Z^M \rangle_\eta = e^{-\eta} \int \prod_m \mathcal{D} f_{m<} \, e^{-S_{\text{eff}}[f_{m<}]}$$

Universal RG flows

All $\phi_{k>\kappa}(x)$ and $\phi_{k<\kappa}(x)$ are Gaussian



$$\sum_{k=\kappa-\delta\kappa}^{\kappa} \lambda_k =: \delta c \ll \sigma^2$$

Each momentum shell

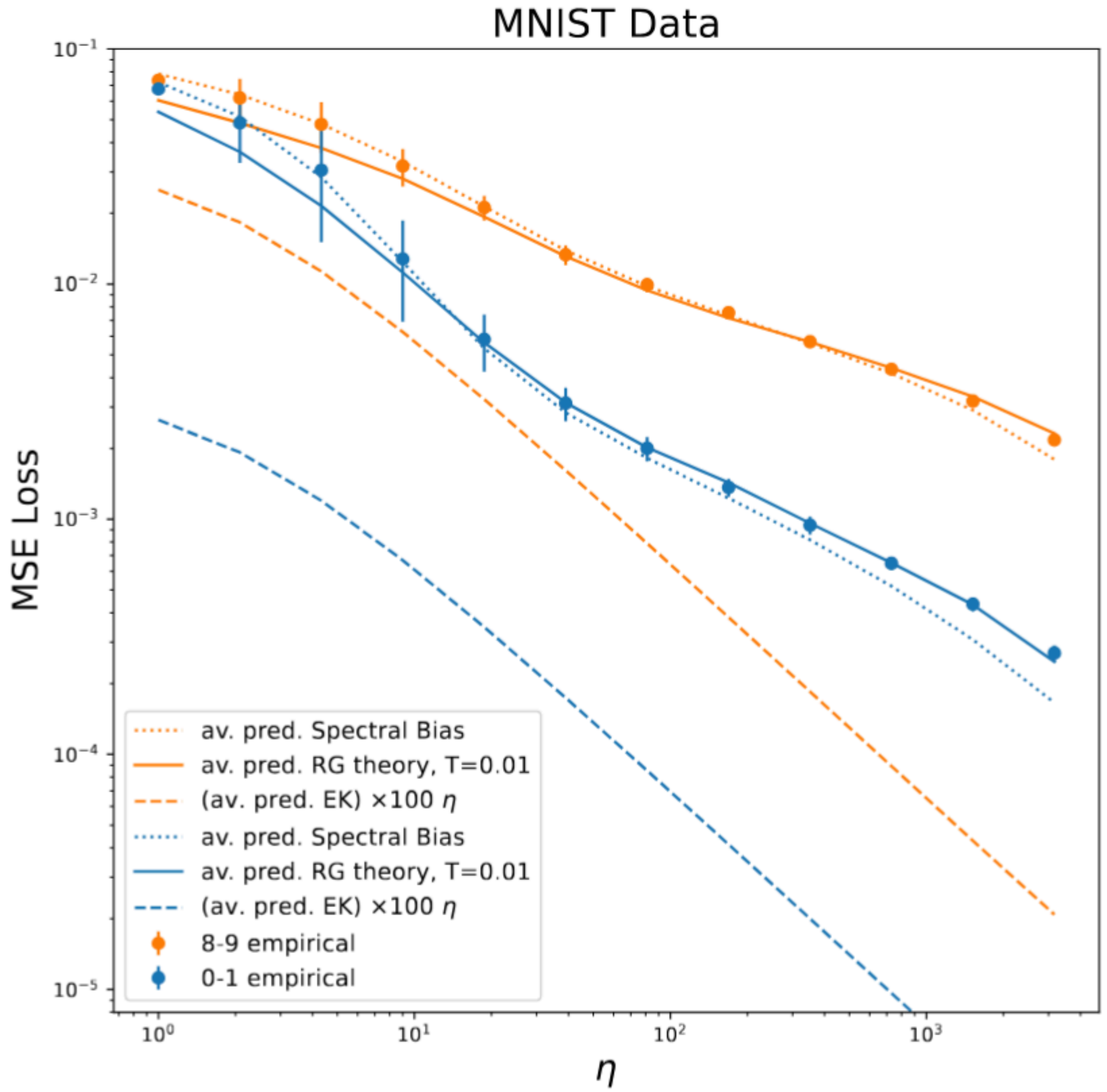
$$\sigma'^2 = \sigma^2 + \delta c$$

Universal RG flow of ridge

$$\sigma_c^2 = \sigma^2 + c$$

Note: the universal ridge renormalization result isn't entirely new.
Our framework provides an RG interpretation to [Canatar, Bordelon, Pehlevan 2021]

Wilsonian RG for Neural Scaling Laws

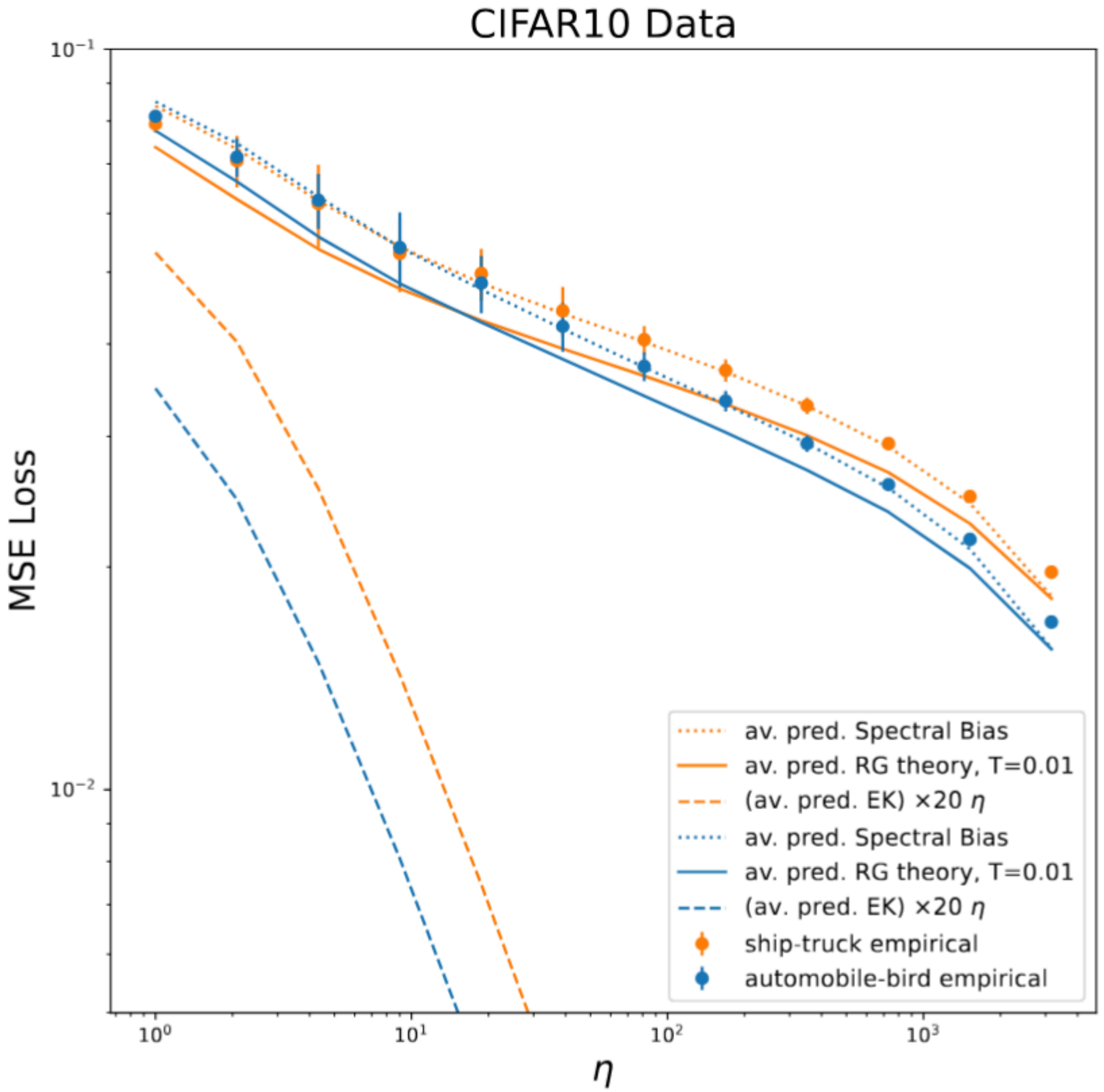


Learnability threshold

$T \in (0,1)$

learnable modes κ

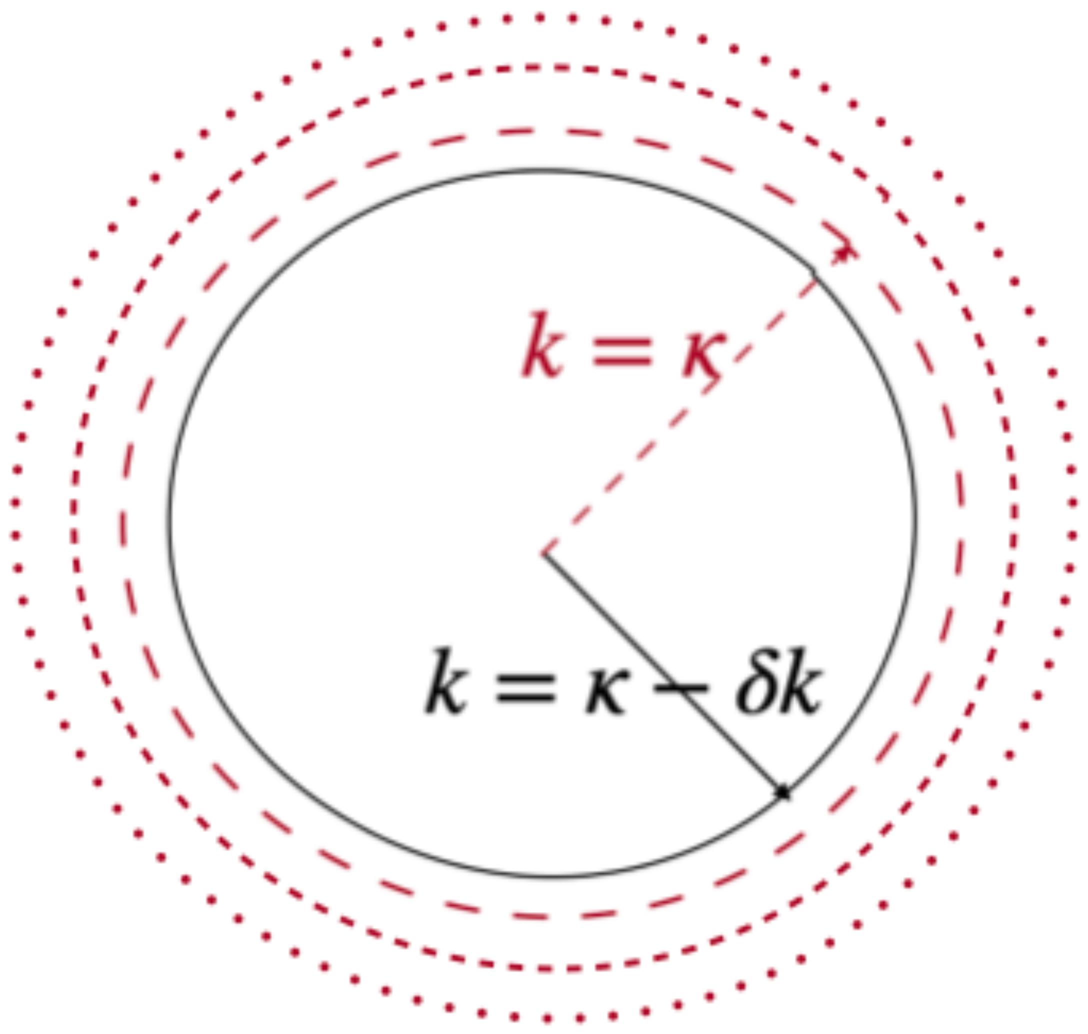
$L_\kappa \approx T$



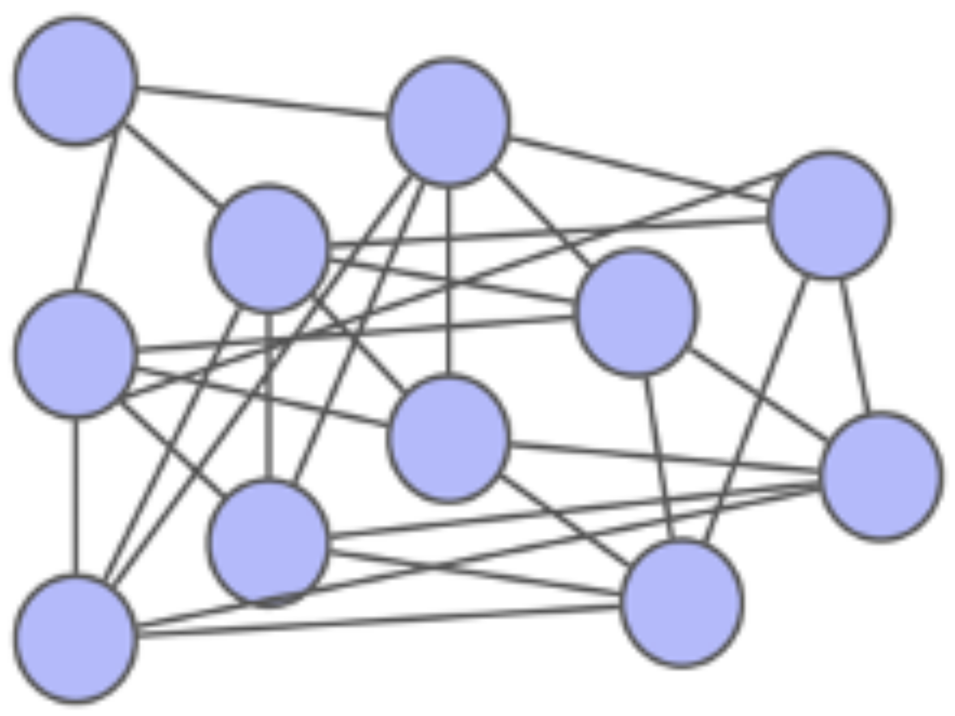
$\sigma^2 = 10^{-8}$

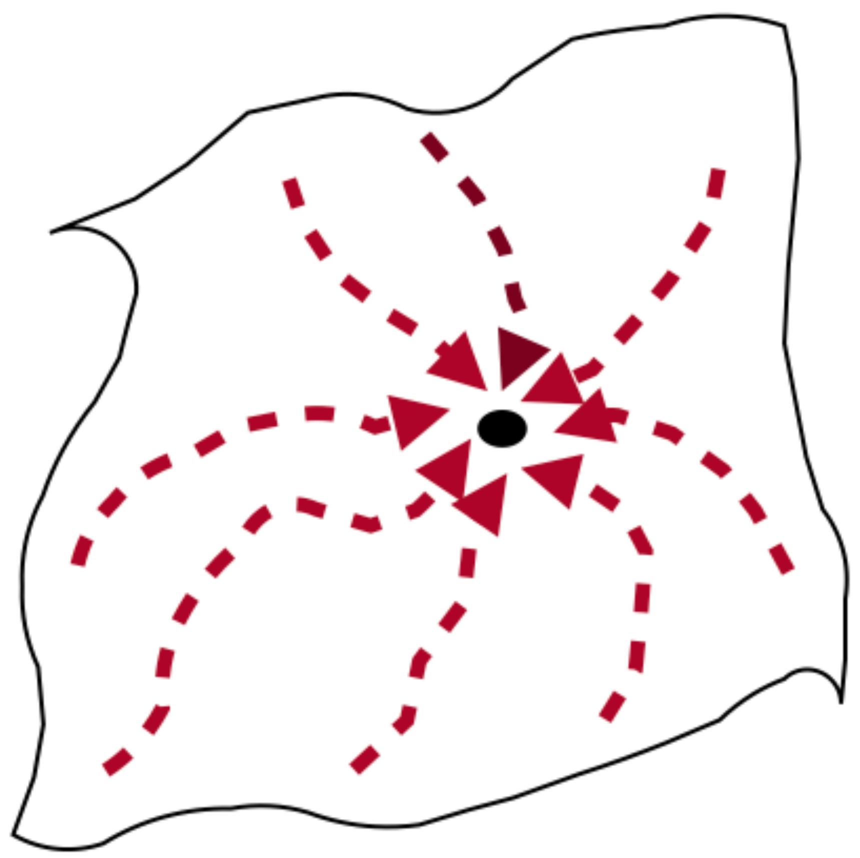
$$\mathrm{MSE}\,(y,\hat{y}) = \frac{1}{N}\sum_{k=1}^{N} |\ \bar{f}_k - y_k|^2 = \frac{1}{N}\sum_{k=1}^{N} L_k^2\ y_k^2$$

$$L_k := \frac{\sigma_{\mathrm{eff}}^2}{\eta\lambda_k + \sigma_{\mathrm{eff}}^2}$$



IV. Functional RG flows

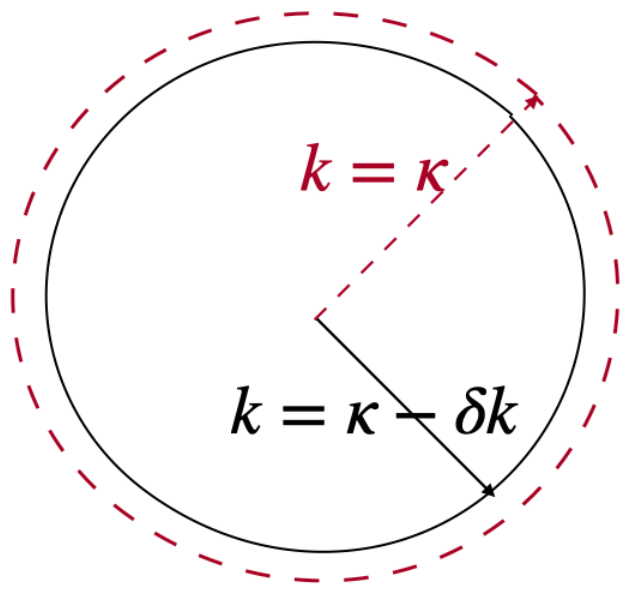




Functional RG flows

Consider Gaussian $\phi_{k<\kappa}$, while $\phi_{k>\kappa}$ are perturbatively non-Gaussian.

$$\langle Z^M\rangle_\eta = e^{-\eta}\int \mathcal{D}\boldsymbol{f}\, e^{-S_0[\boldsymbol{f}]+\eta\int\mathcal{D}\varphi P[\varphi]\,\exp\left[-\frac{1}{2\sigma^2}\left(\Phi_<^\top\Phi_<+2\Phi_<^\top\Phi_>+\Phi_>^\top\Phi_>\right)\right]}$$



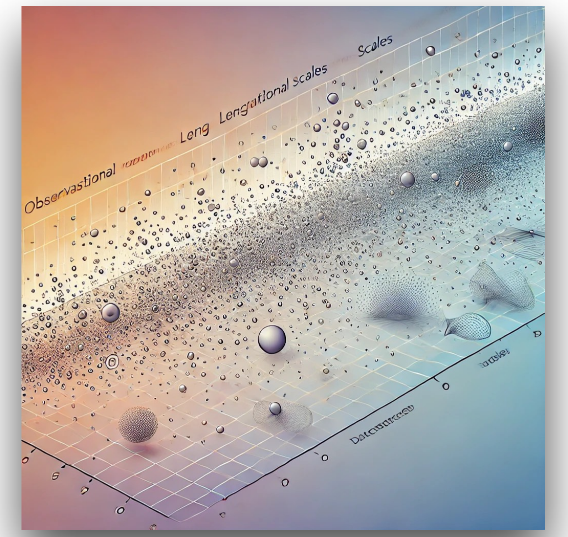
Integrate over shell $\kappa \equiv q$

$$\langle Z^M\rangle_\eta = e^{-\eta}\int\mathcal{D}\boldsymbol{f}_<\, e^{-S_0[\boldsymbol{f}_<]}\exp\left\{\eta\int\mathcal{D}\varphi_<\, P[\varphi_<]\, e^{-\frac{\Phi_<^\top\Phi_<}{2\sigma^2}+\lambda_q(1+2B)\frac{\Phi_<^\top\Phi_<}{2\sigma^4}}\right\}$$

Covariance of GP modes over infinitesimal shell

Spatial RG of ridge

$$W_{\delta c}(x) = W_0(x) - \frac{2\,\delta c}{\sigma_0^2}B_0(x) + O\left(\delta c^2/\sigma_0^4\right)$$



A Solvable Toy Model

Rank-2 kernel

$$K(x,y) = \lambda_1 xy + \lambda_2 (x^2 - 1)(y^2 - 1)$$
$$= \lambda_1 He_1(x) He_1(y) + \lambda_2 He_2(x) He_2(y)$$

Target function with no apparent overlap with kernel eigenmodes

$$y(x) = x^5 - 10x^3 + 15x = He_5(x)$$

Yet, nonzero avg predictor due to nonzero coarse-grained λ_2. **New result.**

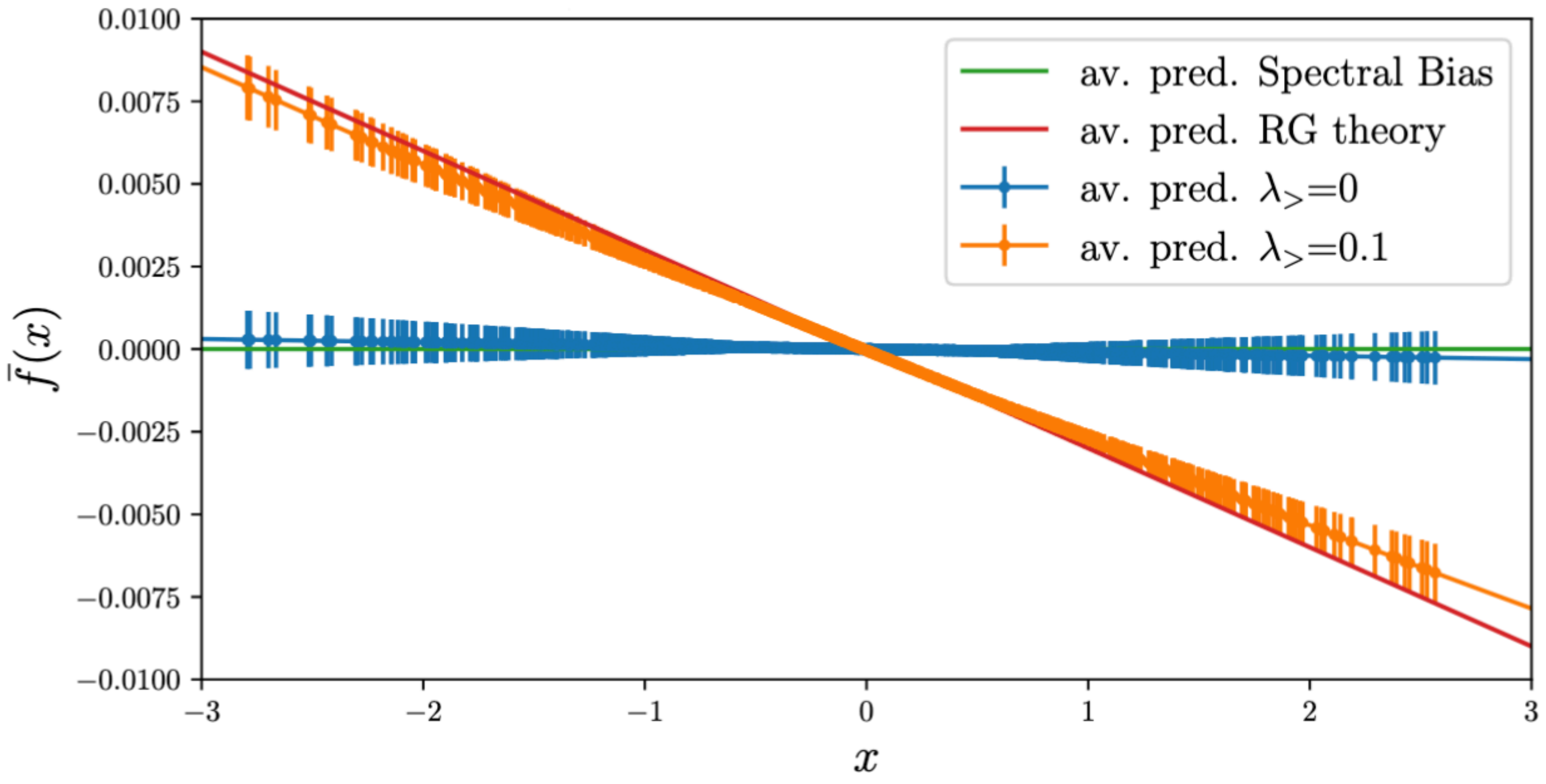


Figure 3. **Non-Gaussian features and spatial re-weighting effects.** Theory versus experiment for the model of sec. V B with $n = 100$ datapoints $\sigma^2 = 400$ and $\lambda_> := \lambda_2 = 0.1$ (unless stated otherwise). Learning a 5th Hermite polynomial, using a kernel capable of expressing only 1st and 2nd Hermite polynomials should give a zero average predictor (green line) based on the standard theory [5, 33]. However, due to spatial re-weighting, a coupling between 1st and 5th Hermite polynomial arises leading to a non-zero result. For both $\lambda_> = 0$ and $\lambda_> = 0.1$, $m = 5$ million trials are performed. The average and standard error (i.e. standard deviation/$\sqrt{m}$) are reported.

Conclusion

- A first principle Wilsonian RG approach for NNGP regression.

- When irrelevant features are Gaussian, universal RG flow; for non-Gaussian irrelevant features, spatial dependences in RG.

- Wilsonian RG shows nonzero correction to average predictions, even when feature modes and target do not overlap.

- Average predictor receives corrections through ridge renormalization in a perturbative manner. We stopped at first order corrections.

- Scaling laws of MSE loss as a function of size of data sets correctly predicted.

Thank You!

Questions?

https://aninditamaiti.github.io/

Email: amaiti@perimeterinstitute.ca

Back-up Slides

NNGP Regression: Some Details

Average predictors: equivalence kernel limit

- Different feature modes ϕ_k do not interact in replica action.

$$\bar{f}_k = \frac{\lambda_k}{\lambda_k + \sigma^2/\eta}\, y_k$$

Average per GP mode

$$\mathrm{Var}[f_k] = \frac{1}{\lambda_k^{-1} + \eta\,\sigma^{-2}}$$

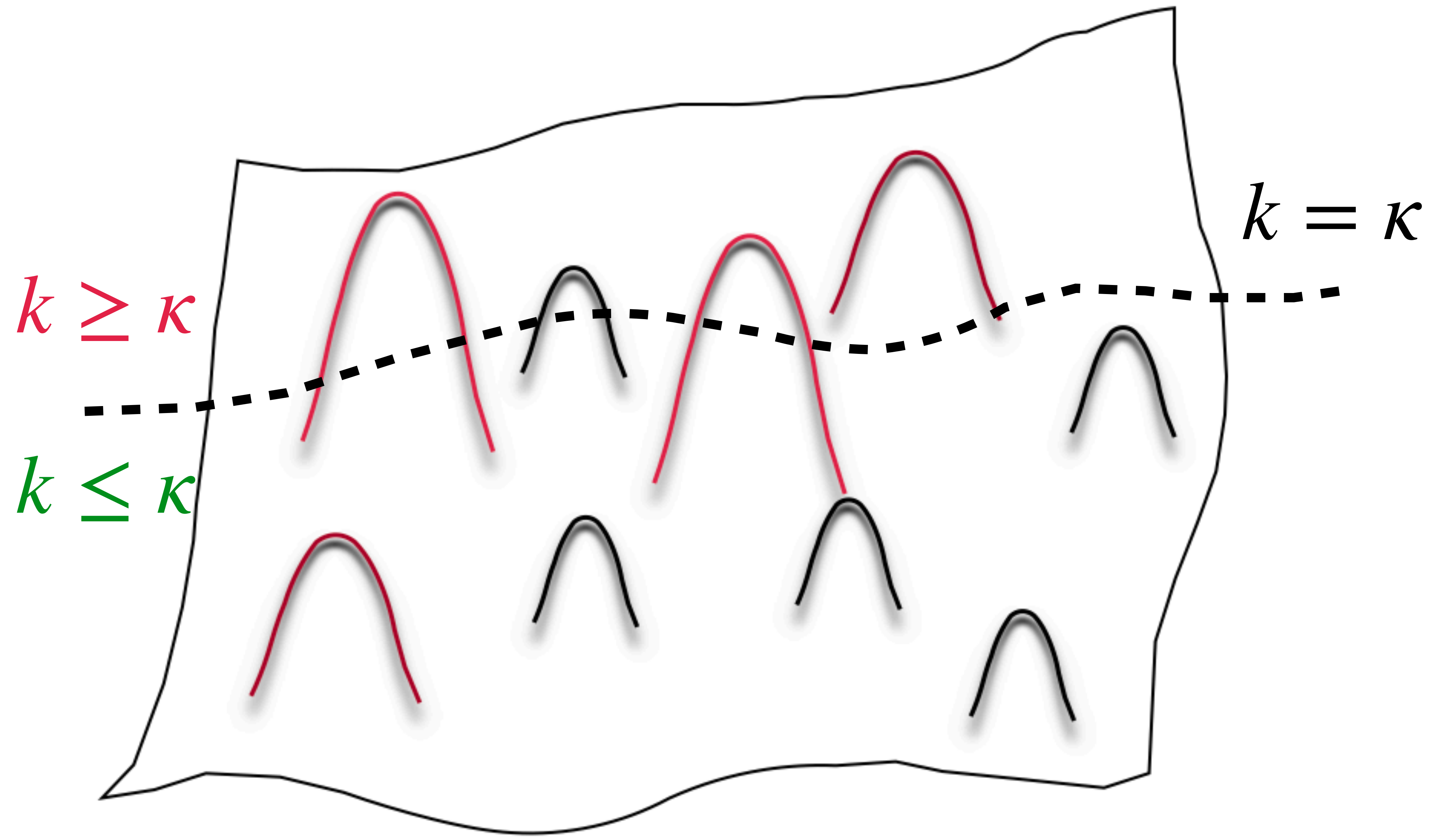
Variance per GP mode

NNGP Regression: Some Details

Irrelevant feature modes (for inference)

- Modes get decoupled from the inference problem, if

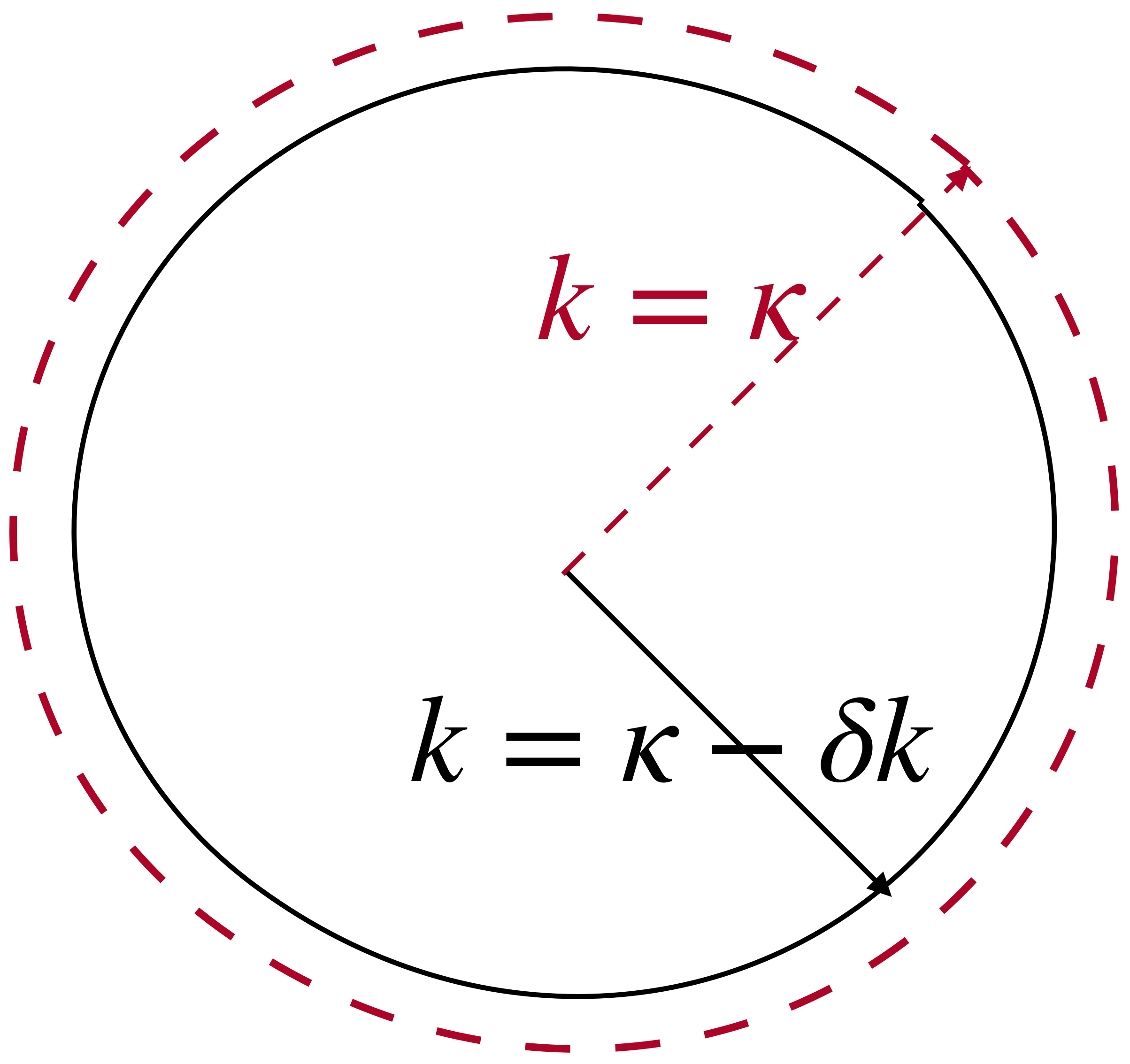
$$\lambda_k \ll \sigma^2/\eta$$



Wilsonian RG framework: Some Details

Do we actually need Wilsonian RG to coarse grain over irrelevant modes?!

Ans. Not necessarily in equivalence kernel limit, where higher modes ($k \geq \kappa$) and lower modes ($k < \kappa$) decouple in replica action.



$\lambda_k \ll \sigma^2/\eta$

Wilsonian RG framework: Some Details

Effects of higher modes show up in S_{eff}

$$\langle Z^M \rangle_\eta = e^{-\eta} \int \prod_m \mathcal{D} f_{m<} \, e^{-S_0[f_{m<}]} \int \prod_m \mathcal{D} f_{m>} \, e^{-S_0[f_{m>}] - S_{\text{int}}[f_{m<}, f_{m>}]}$$

Covariance of higher GP modes

$$[C]_{mn} = (f_{mk} - y_k)(f_{nk} - y_k)$$

$$\langle Z^M \rangle_\eta = e^{-\eta} \int \prod_m \mathcal{D} f_{m<} \, e^{-S_{\text{eff}}[f_{m<}]}$$

Universal RG flows: Some Details

Step 1. Integrate higher GP modes $f_{mk>\kappa}$ (always Gaussian).

Step 2. Integrate Gaussian higher feature modes $\phi_{k>\kappa}$.

$$P[\varphi_>|\varphi_<] = P[\varphi_>] = \mathcal{N}[0, \mathbb{1}; \varphi_>]$$

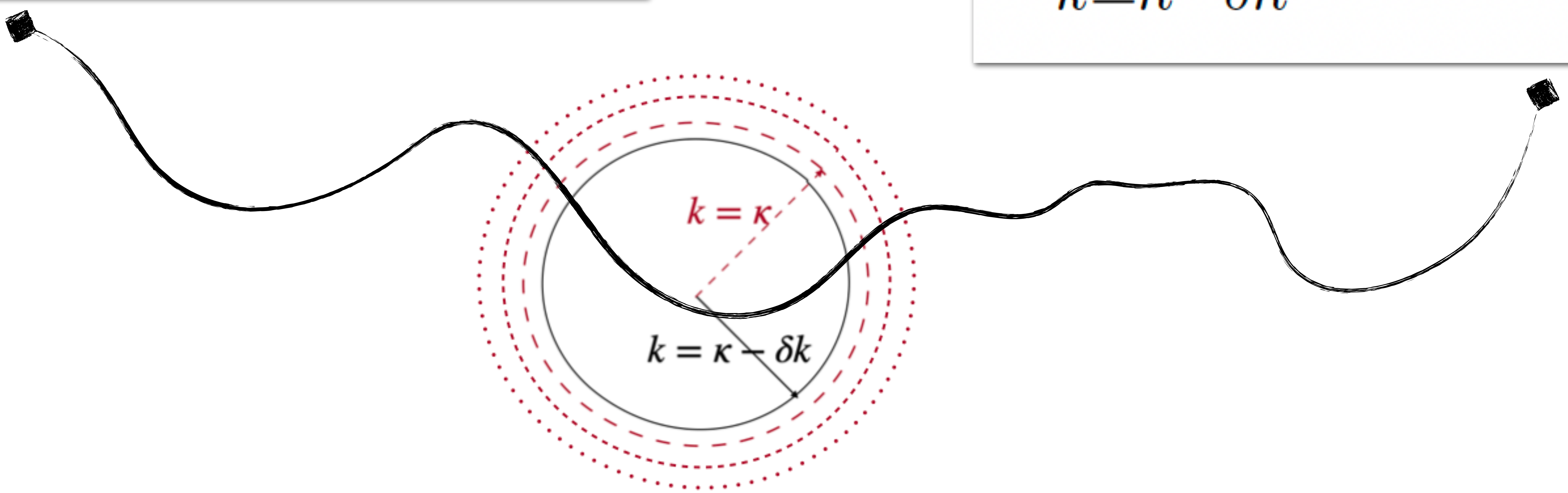
$$\langle Z^M \rangle_\eta = e^{-\eta} \int \prod_m \mathcal{D} f_{m<} \, e^{-S_{\text{eff}}[f_{m<}]}$$

Universal RG flows: Some Details

Assumption over expectation value of GP covariance matrix.

$$\langle C_{y_>=0} \rangle_{S_{0>}} = \mathbb{1}_{M\times M} \sum_{k>\kappa} \lambda_k$$

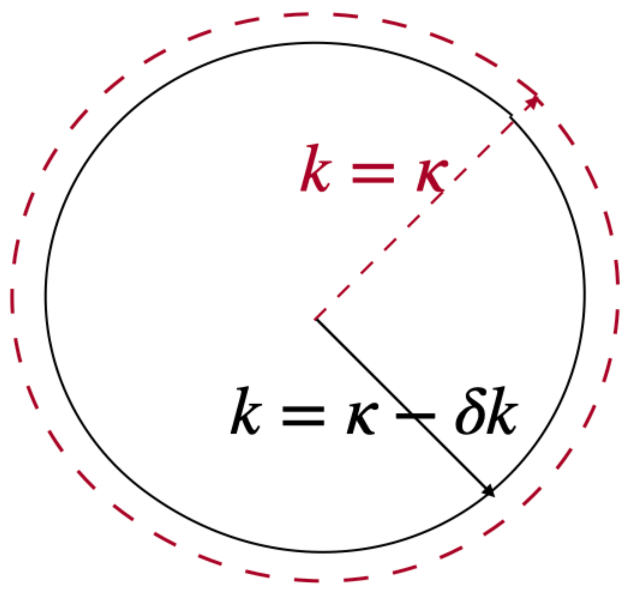
$$\sum_{k=\kappa-\delta\kappa}^{\kappa} \lambda_k =: \delta c \ll \sigma^2$$



Non-Gaussian irrelevant features: Some Details

Spatial re-weighting of MSE loss

$$\langle Z^M \rangle_\eta = e^{-\eta} \int \mathcal{D}\boldsymbol{f}\, e^{-S_0[\boldsymbol{f}]+\eta \int \mathcal{D}\varphi P[\varphi] \exp\left[-\frac{1}{2\sigma^2}\left(\Phi_<^\top\Phi_< + 2\Phi_<^\top\Phi_> + \Phi_>^\top\Phi_>\right)\right]}$$



Integrate over shell $\kappa \equiv q$

$$\Phi_{m<} := \sum_{k \le \kappa} (f_{mk} - y_k)\, \varphi_k$$

$$\langle Z^M \rangle_\eta = e^{-\eta} \int \mathcal{D}\boldsymbol{f}_<\, e^{-S_0[\boldsymbol{f}_<]} \exp\left\{\eta \int \mathcal{D}\varphi_<\, P[\varphi_<]\, e^{-\frac{\Phi_<^\top\Phi_<}{2\sigma^2} + \lambda_q (1+2B) \frac{\Phi_<^\top\Phi_<}{2\sigma^4}}\right\}$$