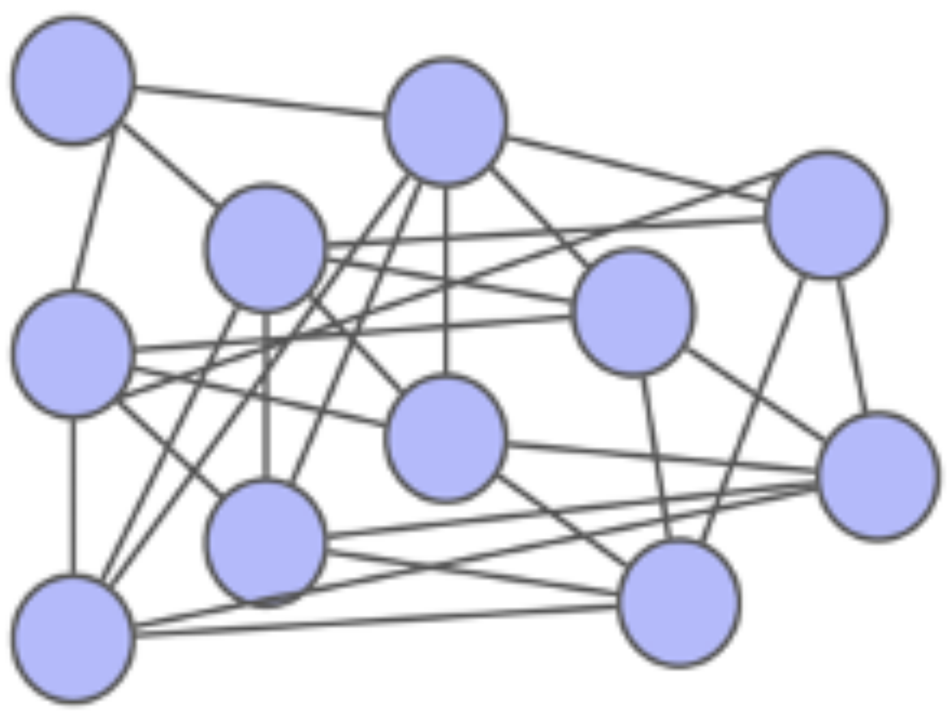
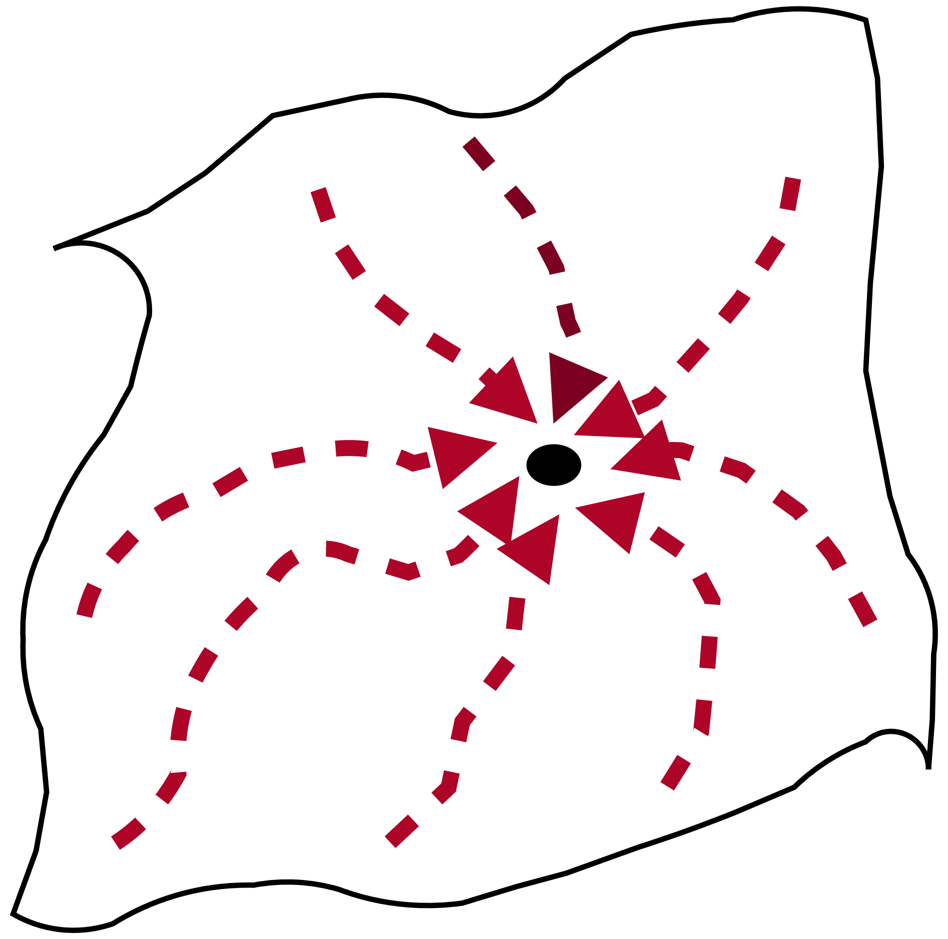
ML through the lens of Renormalization Group





Anindita Maiti
Email: amaiti@perimeterinstitute.ca

Machine Learning & the Renormalization Group

PERIMETER INSTITUTE
Postdoctoral Fellow

QUANTUM INTELLIGENCE LAB

27 May 2024

ECT* Trento,
Italy

Parallels between ML & RG

(a) Training dynamics: coarse graining over Neural Network parameters as per RG schemes.

[Cotler, Rezchikov 2022], [Cotler, Rezchikov 2023], [Berman, Heckman, Klinger 2022], [Berman, Klinger 2022], [Berman, Klinger, Stapleton 2023], [Berman, Klinger, Stapleton 2024]

(b) Initialization: RG of statistical field theories associated with NN ensembles.

[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],

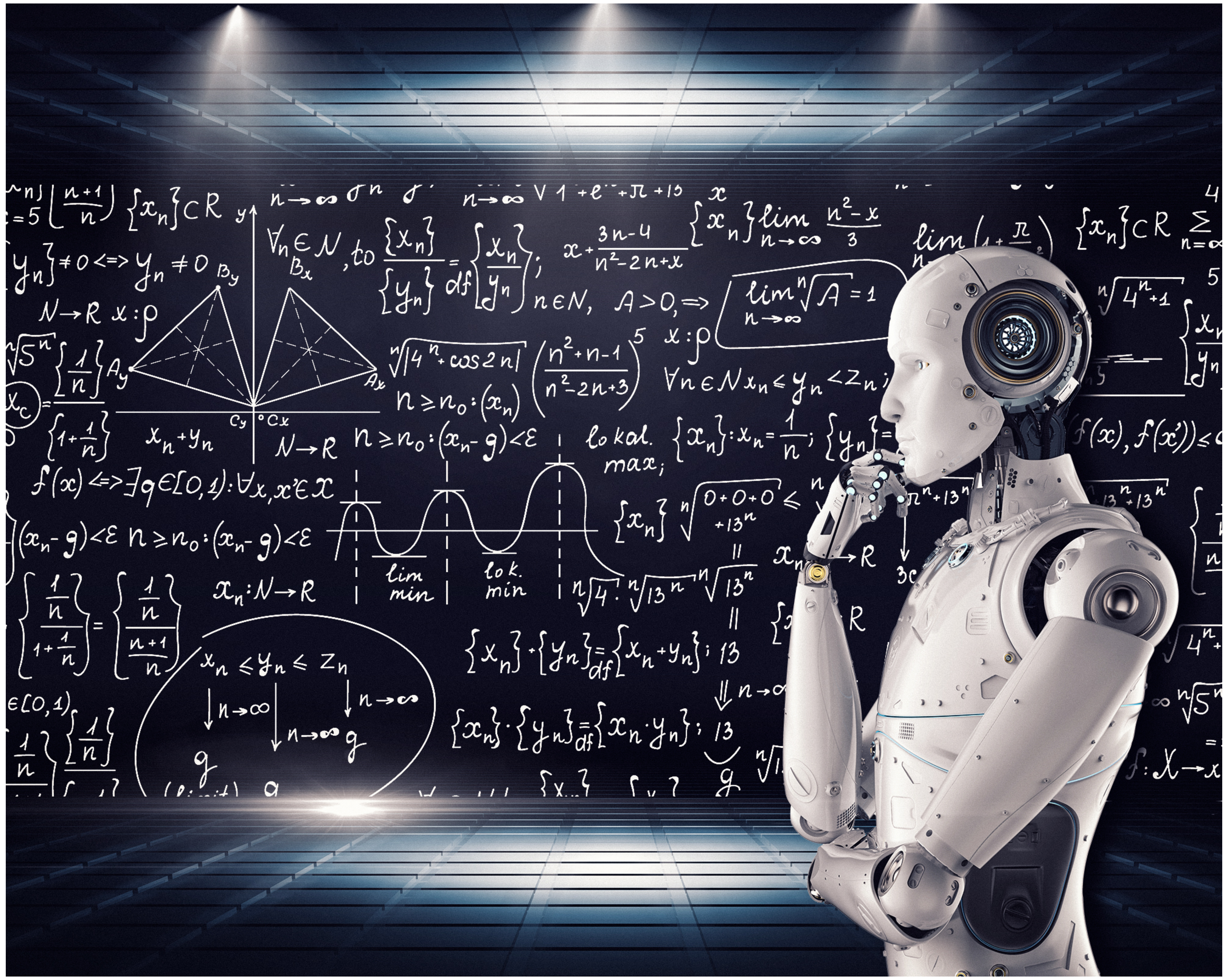


Image courtesy: google images

Parallels between ML & RG

Q. Can RG + field theory improve existing theoretical frameworks in ML?

Let's find out.

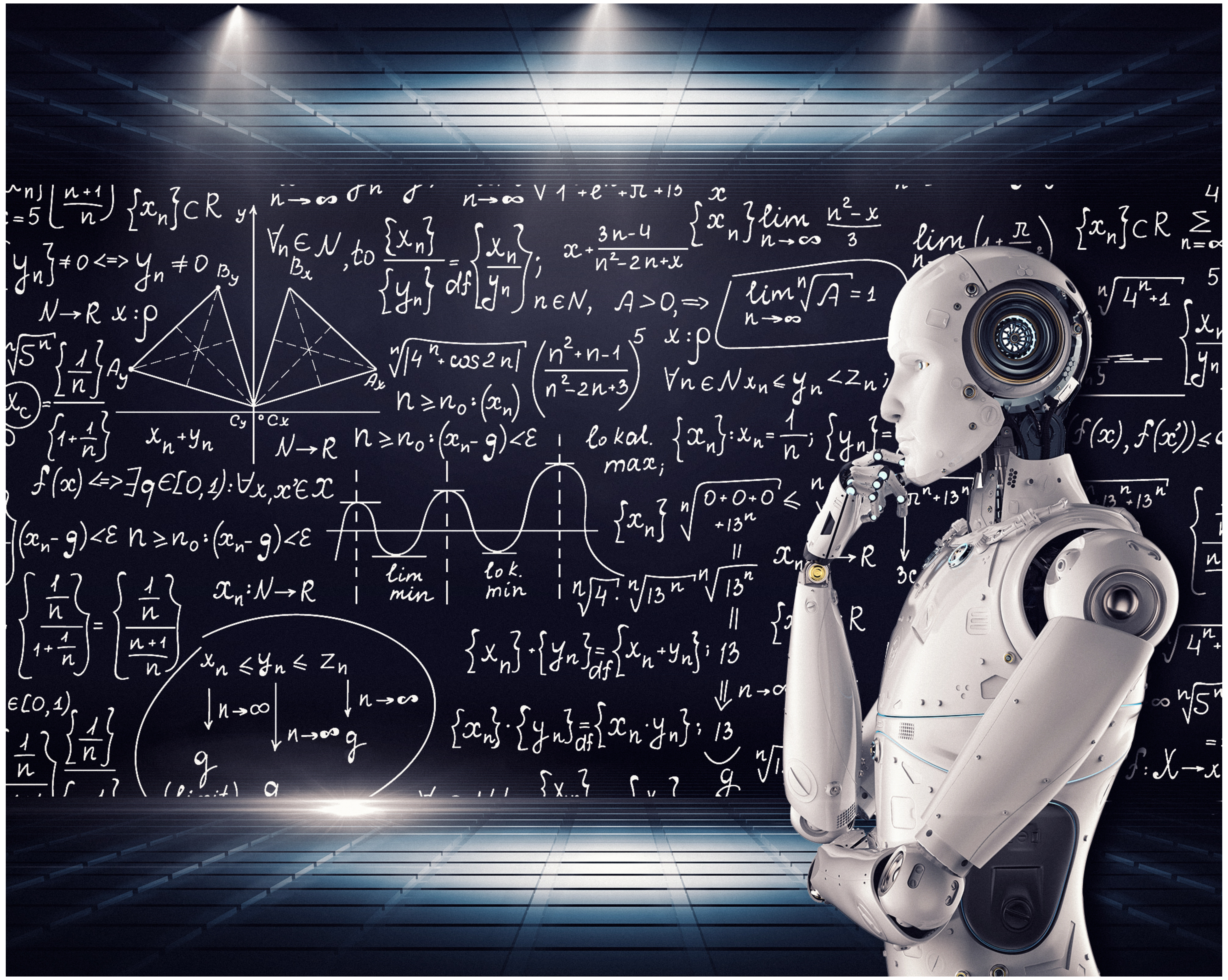


Image courtesy: google images

Parallels between ML & RG

Take a well known problem in ML:

- Neural Network Gaussian Process (NNGP) regression.

- What can RG tell us here, that we don't already know?

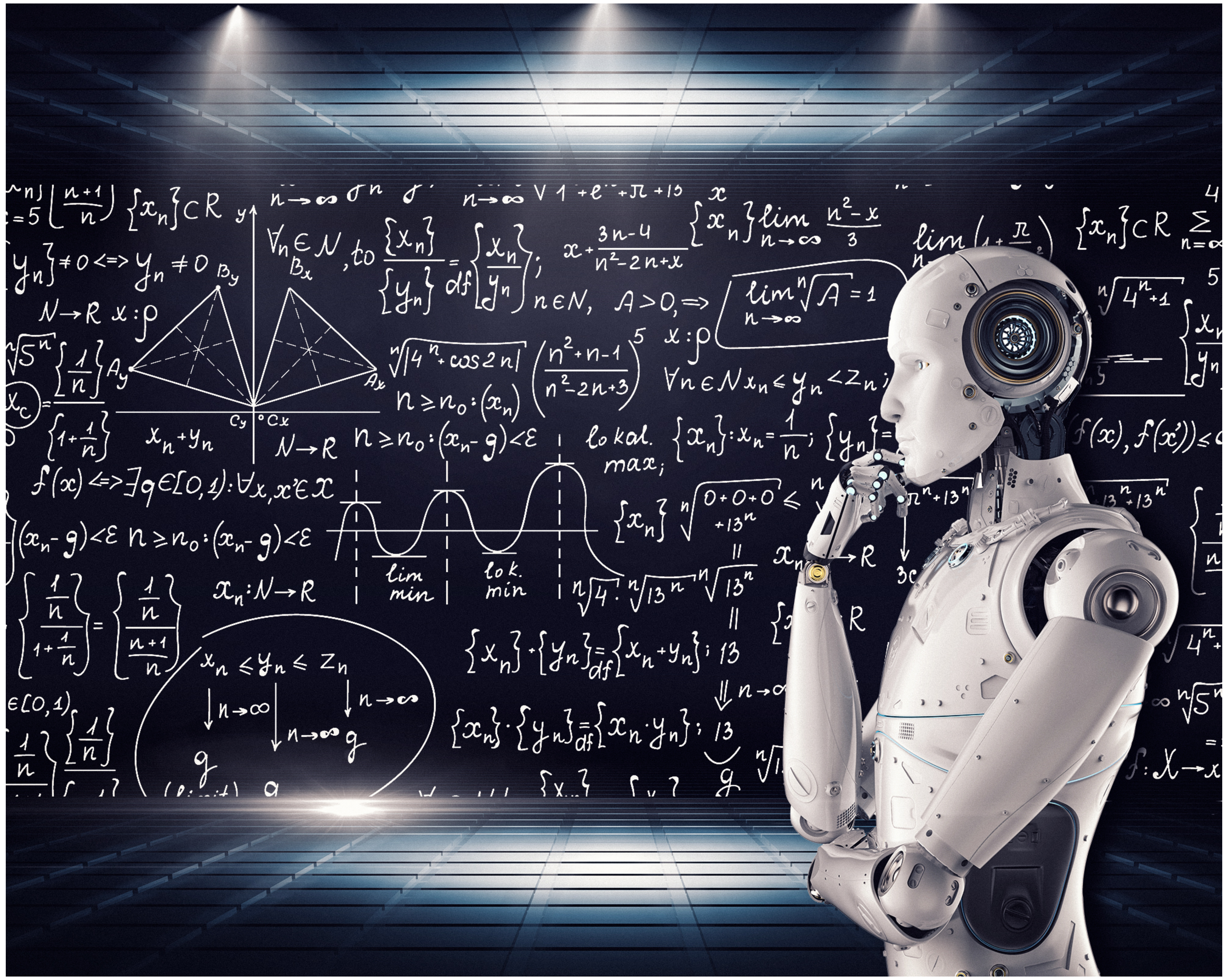


Image courtesy: google images

Wilsonian Renormalization of Neural Network Gaussian Processes

Jessica N. Howard,[a] **Ro Jefferson,**[b] **Anindita Maiti,**[c] **and Zohar Ringel**[d*]

[a] *Kavli Institute for Theoretical Physics, Santa Barbara, CA USA*
[b] *Institute for Theoretical Physics, and Department of Information and Computing Sciences Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands*
[c] *Perimeter Institute for Theoretical Physics, Waterloo, Ontario, N2L 2Y5, Canada*
[d] *The Racah Institute of Physics, The Hebrew University of Jerusalem*

E-mail: jnhoward@kitp.ucsb.edu, r.jefferson@uu.nl, amaiti@perimeterinstitute.ca, zohar.ringel@mail.huji.ac.il

Abstract: Separating relevant and irrelevant information is key to any modeling process or scientific inquiry. Theoretical physics offers a powerful tool for achieving this in the form of the renormalization group (RG). Here we demonstrate a practical approach to performing Wilsonian RG in the context of Gaussian Process (GP) Regression. We systematically integrate out the unlearnable modes of the GP kernel, thereby obtaining an RG flow of the Gaussian Process in which the data plays the role of the energy scale. In simple cases, this results in a universal flow of the ridge parameter, which becomes input-dependent in the richer scenario in which non-Gaussianities are included. In addition to being analytically tractable, this approach goes beyond structural analogies between RG and neural networks by providing a natural connection between RG flow and learnable vs. unlearnable modes. Studying such flows may improve our understanding of feature learning in deep neural networks, and identify potential universality classes in these models.

Acknowledgement

Zohar Ringel

Hebrew University of Jerusalem

Ro Jefferson

Utretch University

Jessica N. Howard

Kavli Institute for Theoretical Physics (UCBS)

Talk Outline

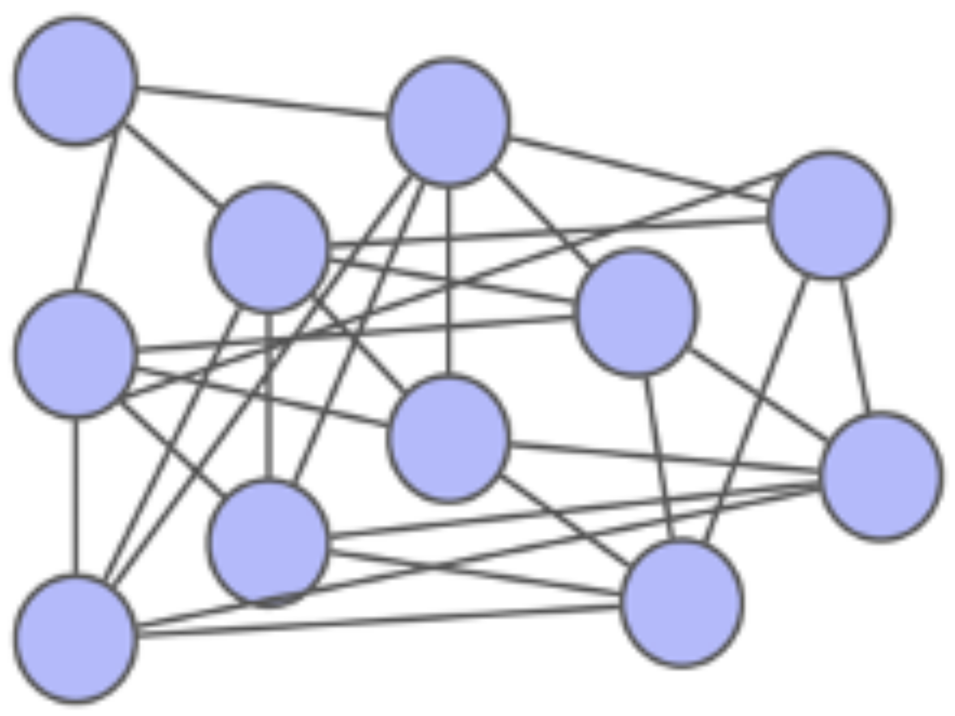
I. Neural Network Gaussian Process (NNGP) regression.

II. Wilsonian RG for NNGP regression.

III. When irrelevant modes are Gaussian.

IV. When irrelevant modes are non-Gaussian.

I. Neural Network Gaussian Process (NNGP) Regression



NNGP Regression: Part I

Consider Neural Network Gaussian Process limit.

➡ Mean 0

➡ Covariance $K(x, x')$

Train data

$x, x' \in \mathbb{R}^d$

$X_n \sim p_{\text{data}}(x)$

$n \times d$ matrix X_n

Test data $\longrightarrow$ Target

x_*

$y(x)$

NNGP Regression: Part II

Average predictor on test data

$$\tilde{f}(x_*|X_n) = K(x_*, X_n)[K(X_n, X_n) + \sigma^2 \mathbb{1}_{n\times n}]^{-1} y(X_n)$$

Train data

$x, x' \in \mathbb{R}^d$

$X_n \sim p_{\text{data}}(x)$

$n \times d$ matrix X_n

$\sigma^2 \longrightarrow \boxed{y(x) + \epsilon}$

Ridge parameter

Test data $\longrightarrow$ Target

x_* $\quad$ $y(x)$

In real life, noisy versions of target are available.

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

NNGP Regression: Part III

Equivalence kernel limit

$$\tilde{f}(x_*|X_n) = K(x_*, X_n)[K(X_n, X_n) + \sigma^2 \mathbb{1}_{n\times n}]^{-1} y(X_n)$$

Obtained using replica partition function over M copies of the same system, then setting $\lim M \to 0$.

η: average #(train data points).

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

NNGP Regression: Part IV

Equivalence kernel limit

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

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

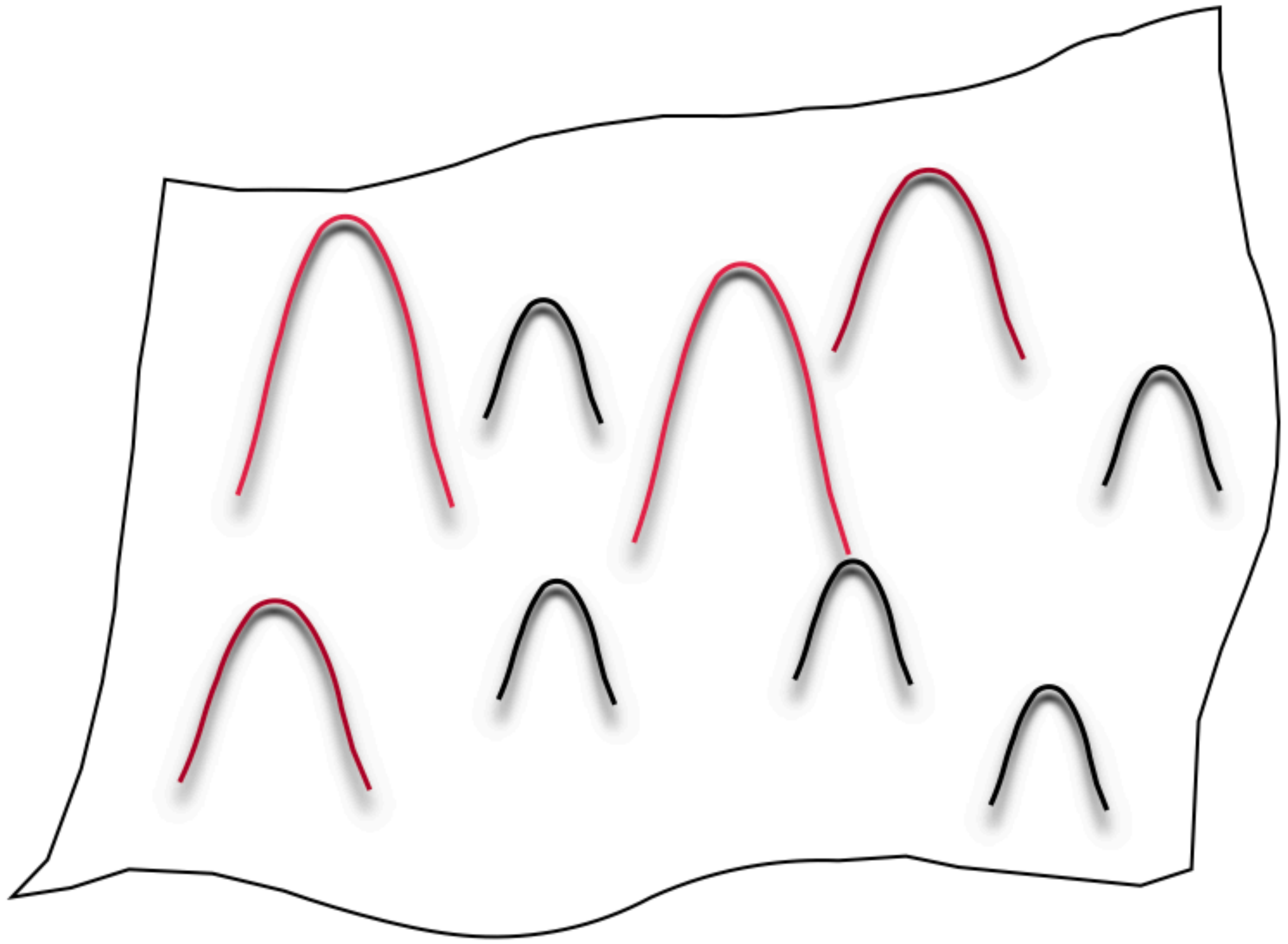
$\eta, \sigma^2 \to \infty$ with η/σ^2 fixed

Spectral decomposition: in NNGP kernel eignespace.

NNGP Regression: Part V

Introduce NNGP kernel eigenspace.

- Eigenfunctions / feature modes: ϕ_k .
- Eigenvalues: λ_k .
- GP modes: f_{mk} ; m ~ replica index.



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

NNGP Regression: Part VI

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

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

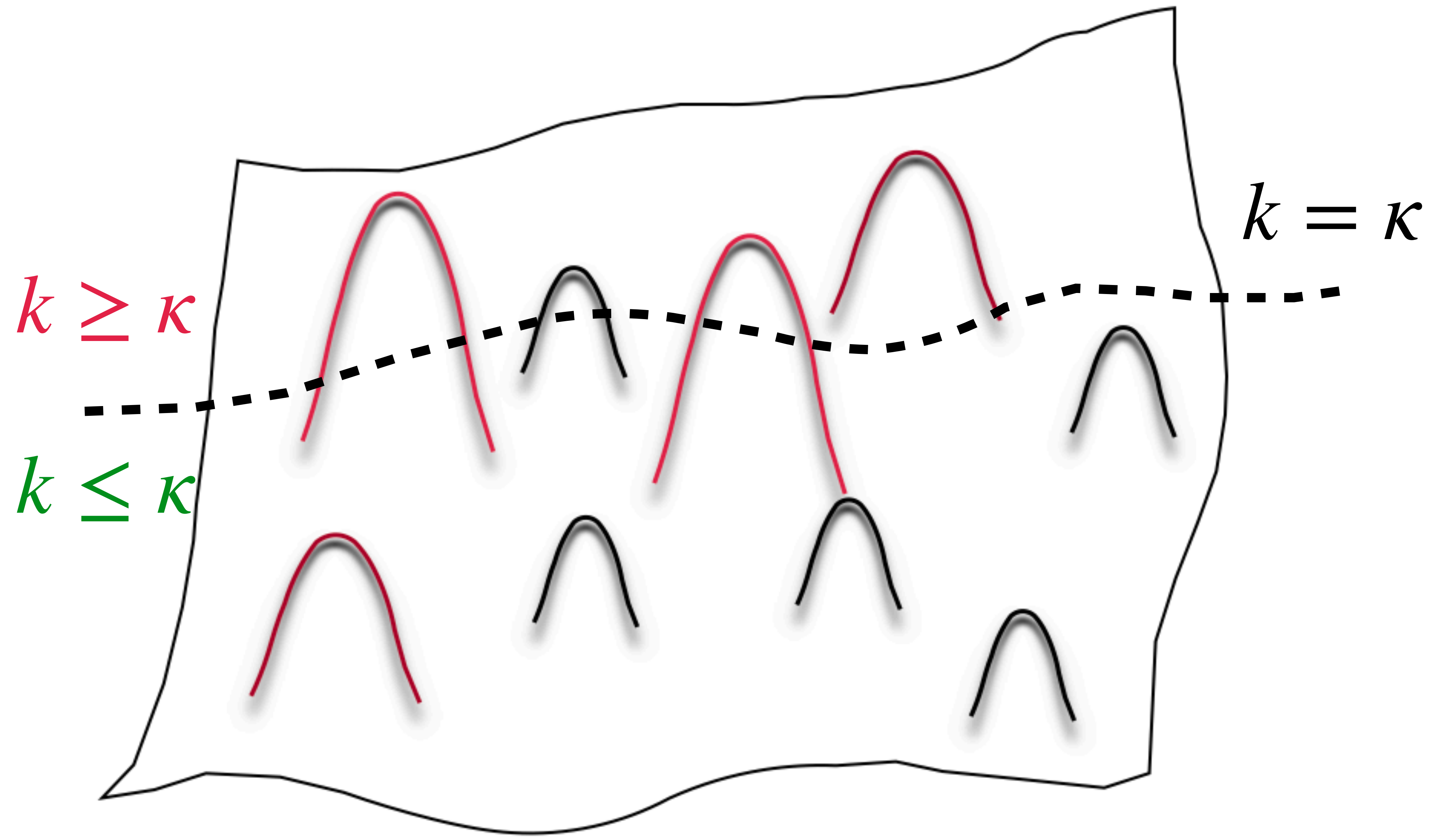
Variance per GP mode

NNGP Regression: Part VII

Irrelevant feature modes (for inference)

- Modes get decoupled from the inference problem, if

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

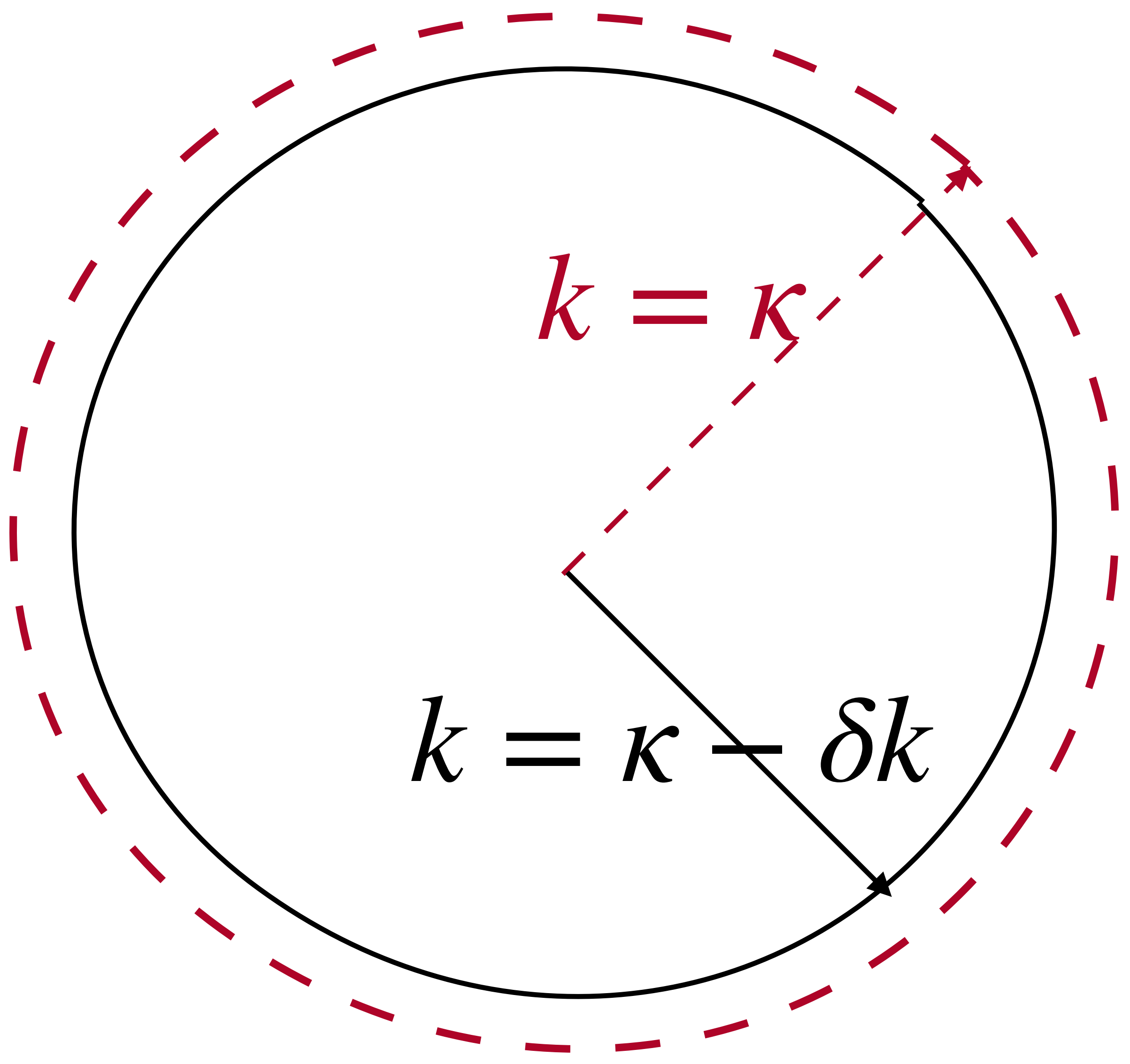


NNGP Regression: Part VIII

Feature space/ kernel eigen-space ≈ inverse of momentum space

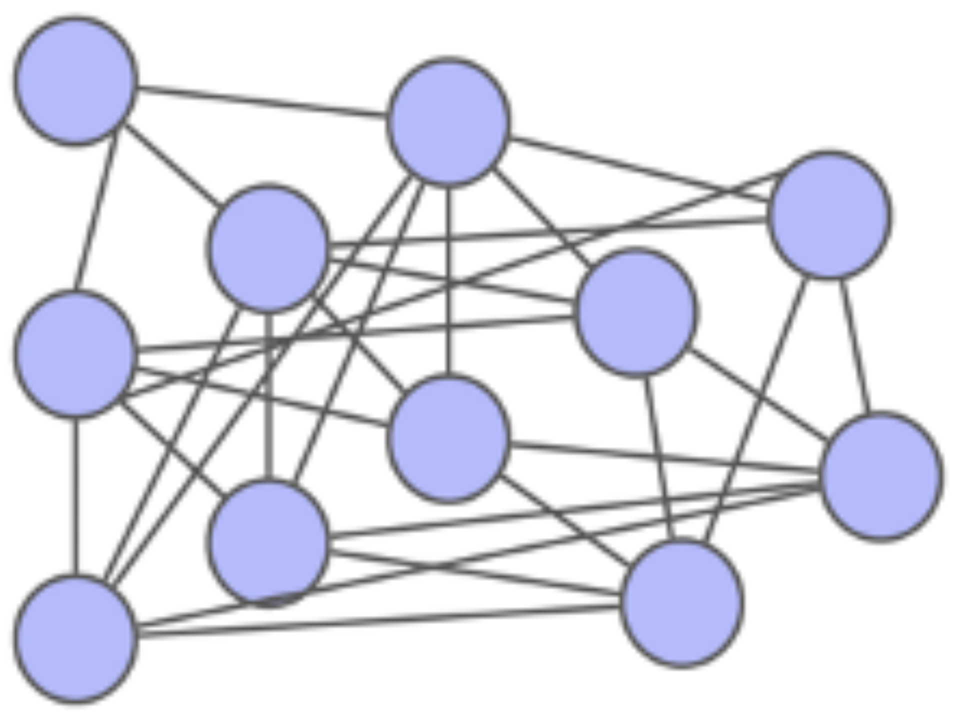
Q1. Can we use momentum shell RG to integrate out irrelevant feature modes?

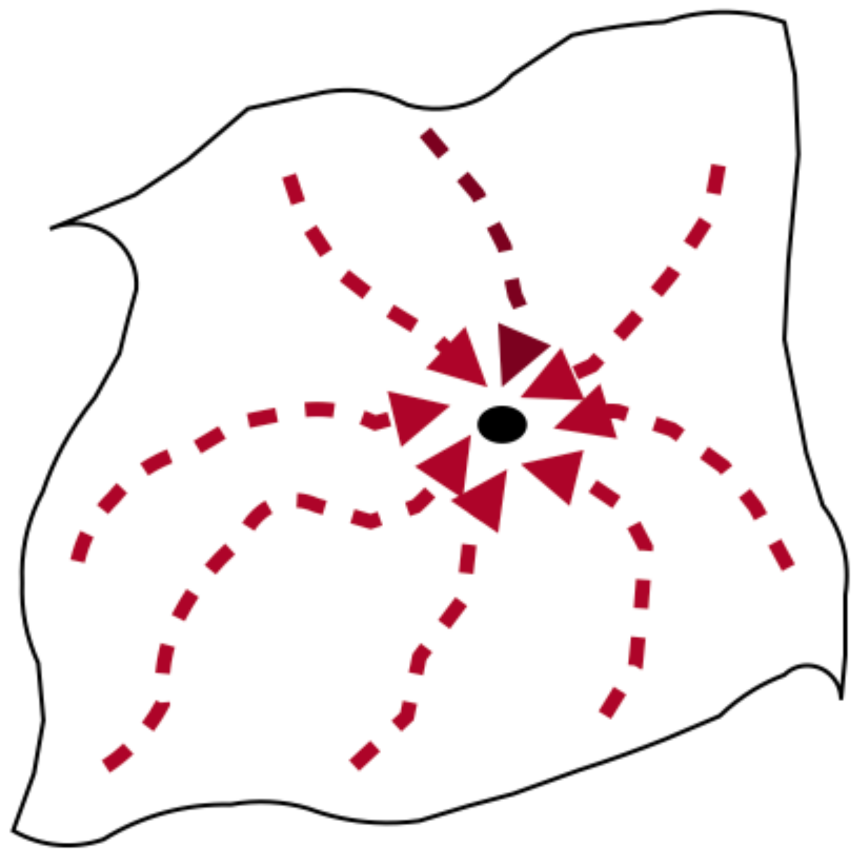
Q2. How does that impact noise σ^2 renormalization for average predictor?



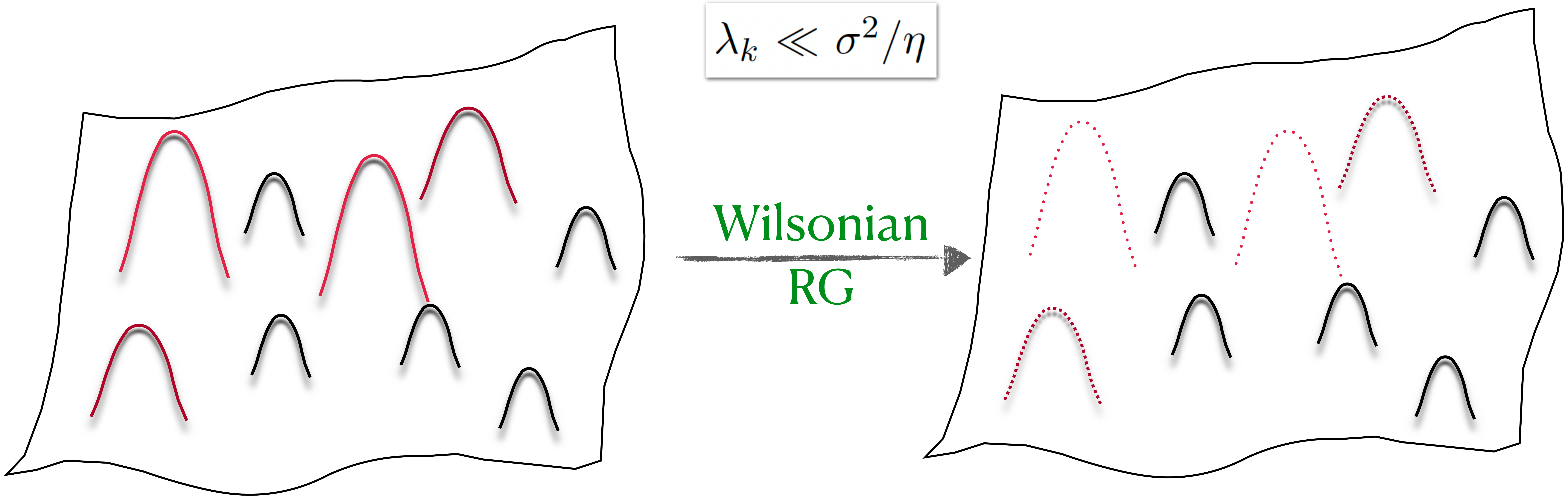
$\lambda_k \ll \sigma^2/\eta$

II. Wilsonian RG for NNGP regression





Wilsonian RG for NNGP regression (I)



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

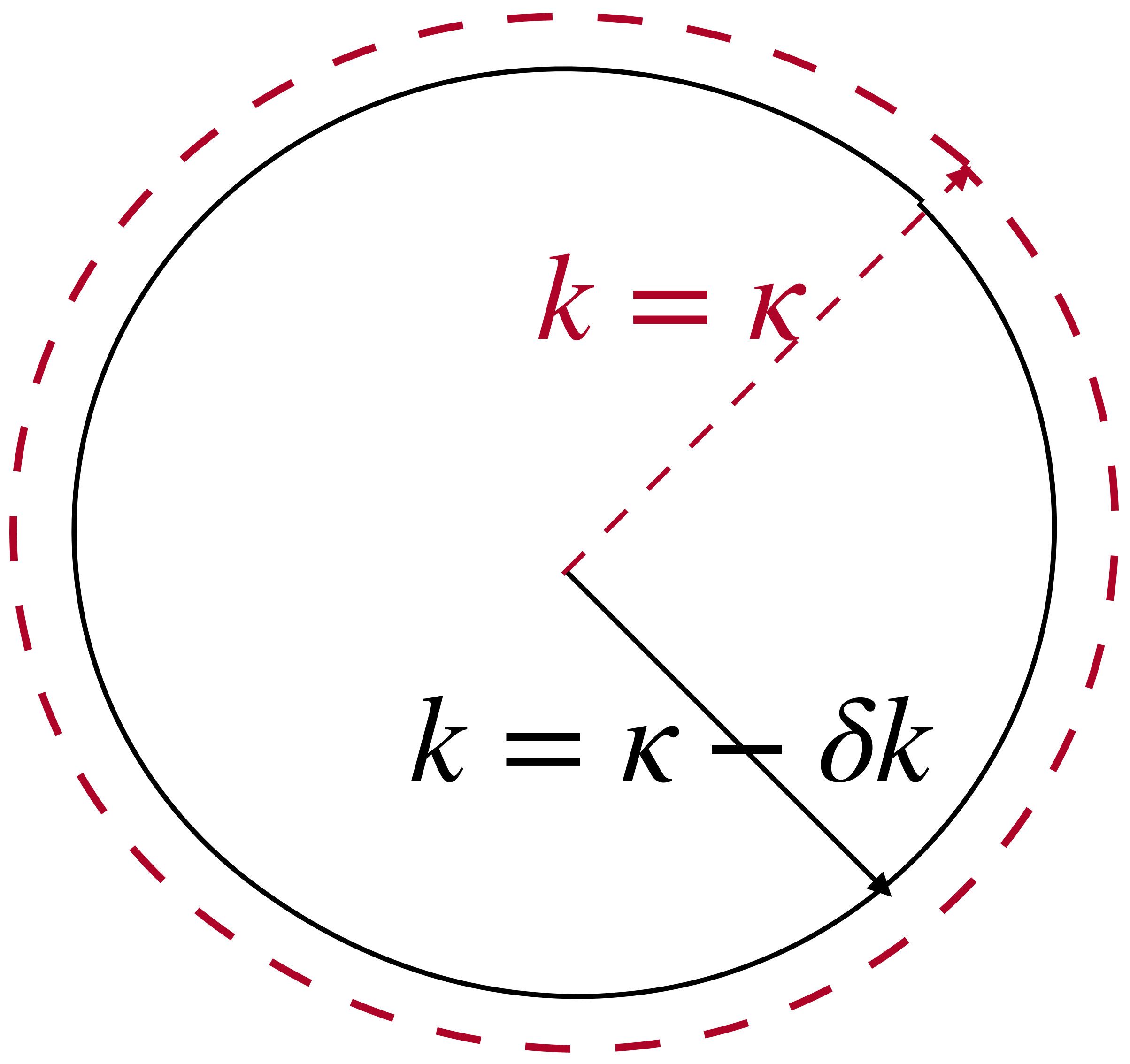
Feature modes
$k = 1, \cdots, \kappa - \delta k$

Wilsonian RG for NNGP regression (II)

Integrate out irrelevant modes from replica action.

Step 1. Integrate out $\phi_{k\geq\kappa}$.

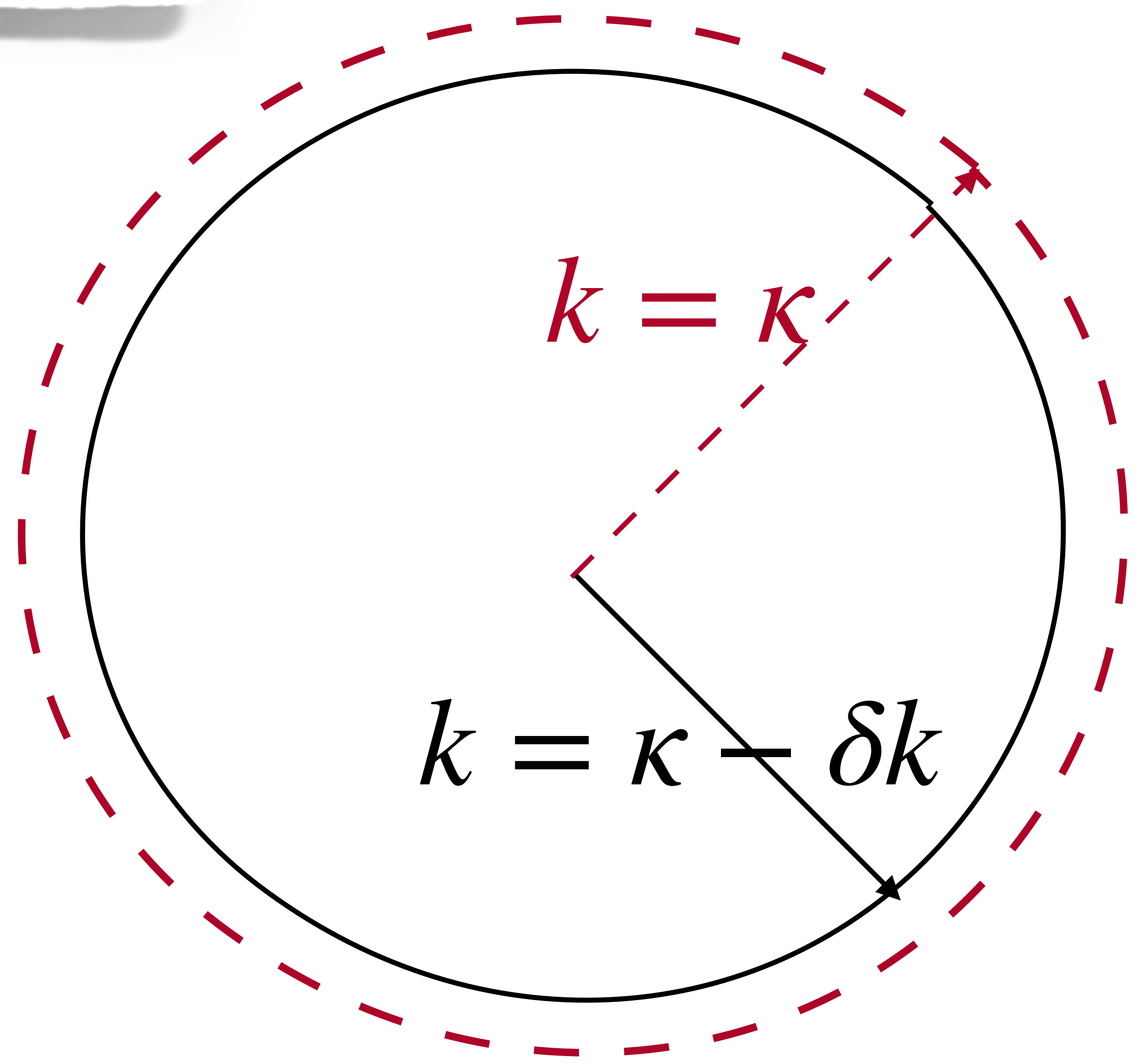
$\lambda_k \ll \sigma^2/\eta$



Step 2. Integrate out $f_{mk\geq\kappa}$.

Wilsonian RG for NNGP regression (III)

But, do we actually need Wilsonian RG to coarse grain over irrelevant modes?!



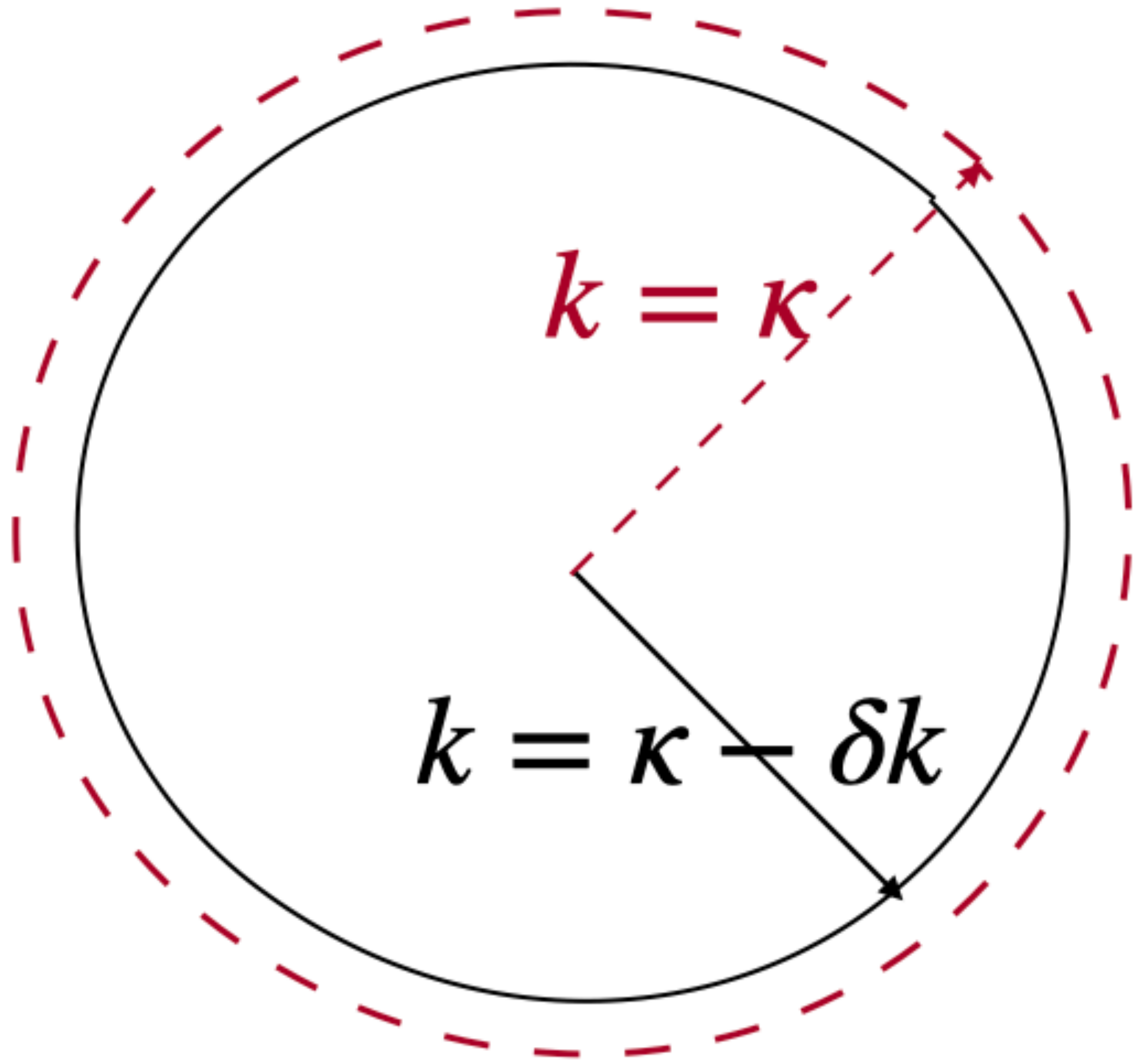
Ans. Not necessarily in equivalence kernel limit, where higher modes ($k \geq \kappa$) and lower modes ($k < \kappa$) decouple in replica action.

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

Wilsonian RG for NNGP regression (IV)

Q. Then when do we require Wilsonian RG?

Ans. Finite η, σ^2.



Higher modes $(k \geq \kappa)$ and lower modes $(k < \kappa)$ interact in replica action.

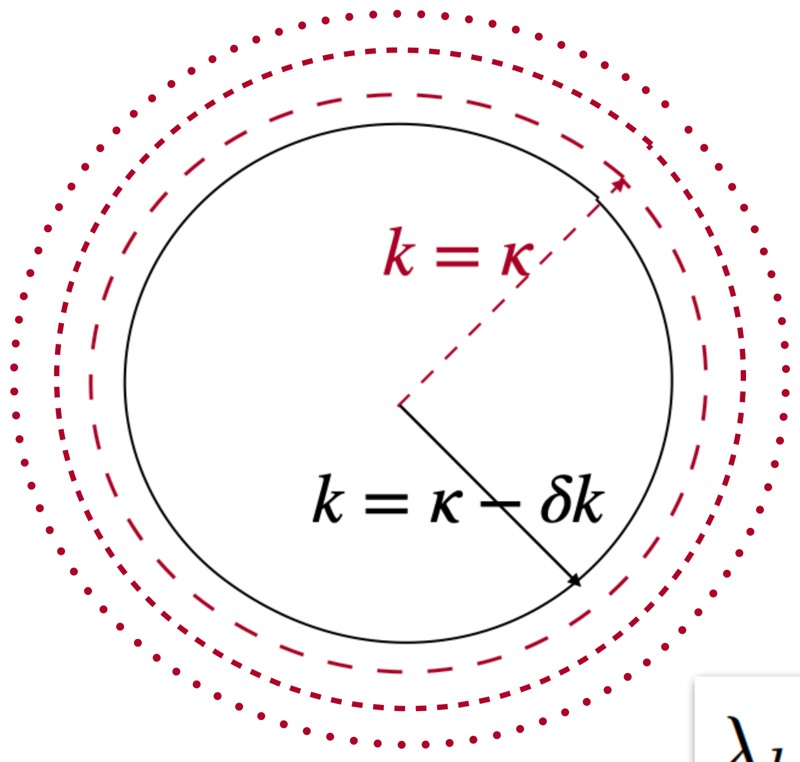
$$\langle Z^M \rangle_\eta = e^{-\eta} \int \prod_m \mathcal{D} f_{m<} \, e^{-S_0[f_{m<}]} \int \prod_m \mathcal{D} f_{m>} \, e^{-S_0[f_{m>}] - S_{\text{int}}[f_{m<}, f_{m>}]}$$

Wilsonian RG for NNGP regression (V)

Higher modes ($k \geq \kappa$) and lower modes ($k < \kappa$) interact in replica action.

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

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



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

Wilsonian RG for NNGP regression (VI)

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

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

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

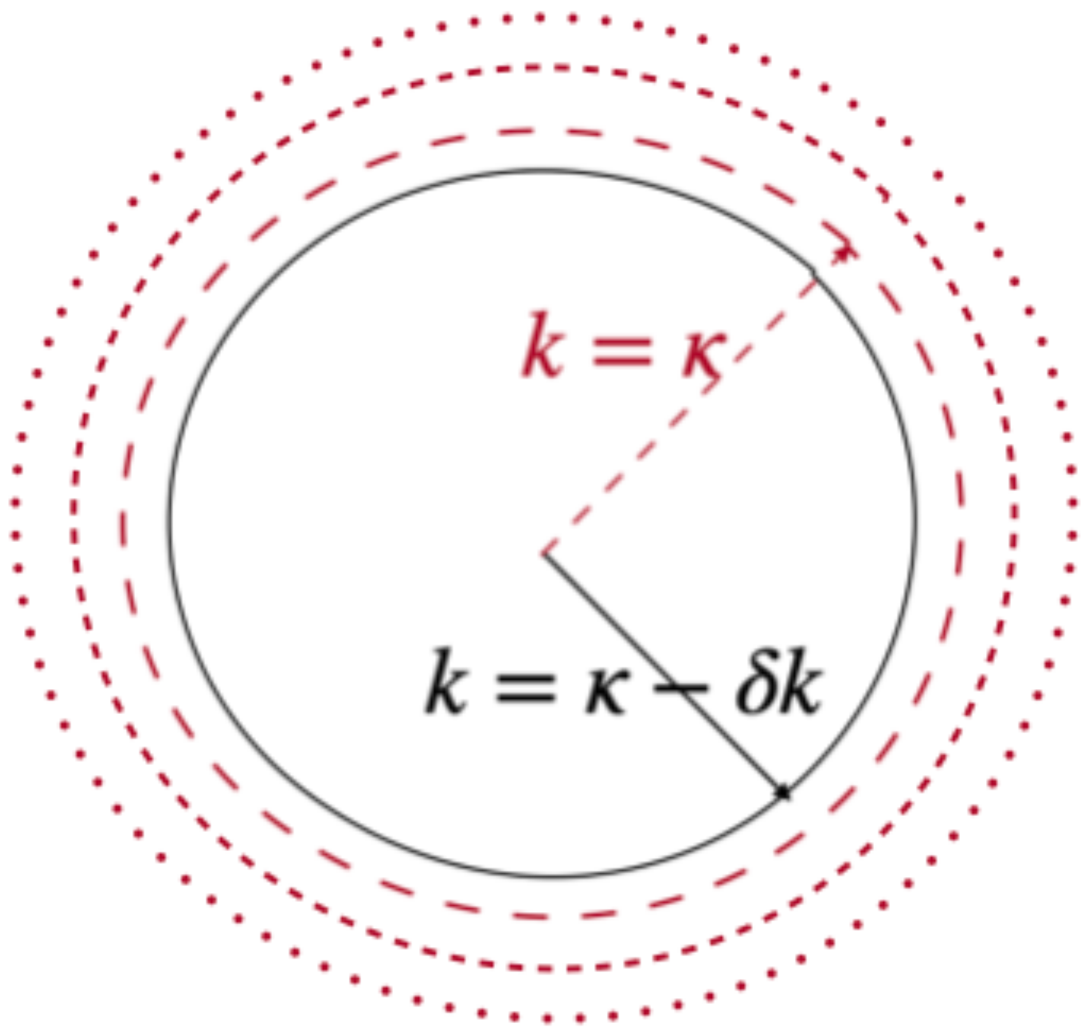
Wilsonian RG for NNGP regression (VI)

Q. How does Wilsonian RG renormalize noise $\epsilon \sim \mathcal{N}(0,\sigma^2)$?

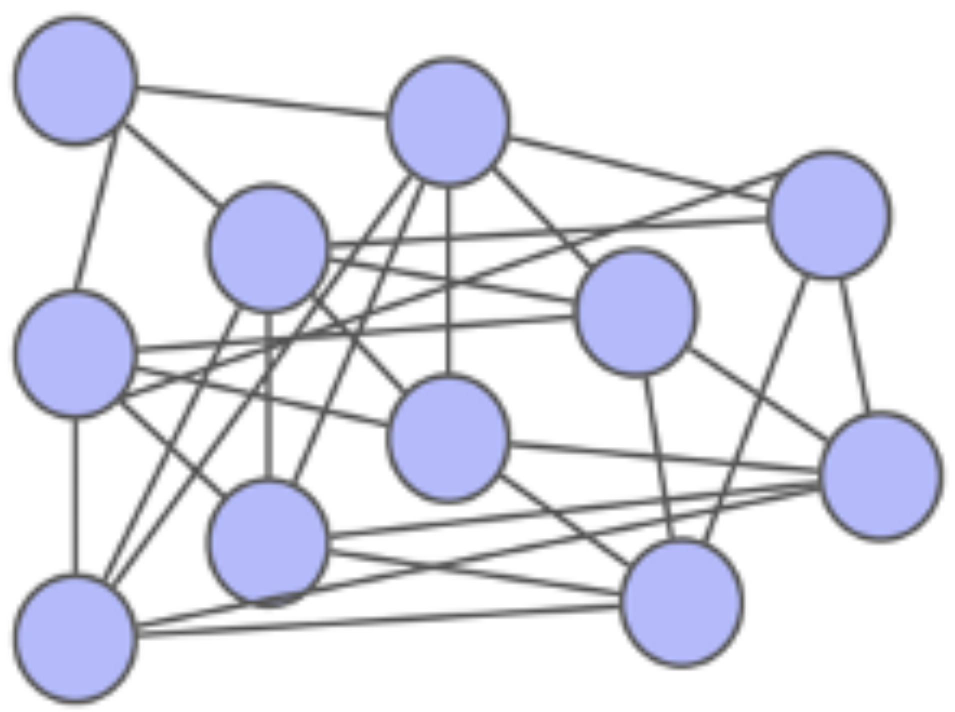
$$\langle Z^M \rangle_\eta = e^{-\eta} \int \prod_m \mathcal{D} f_{m<} \, e^{-S_0[f_{m<}]} \int \prod_m \mathcal{D} f_{m>} \, e^{-S_0[f_{m>}] - S_{\text{int}}[f_{m<}, f_{m>}]}$$

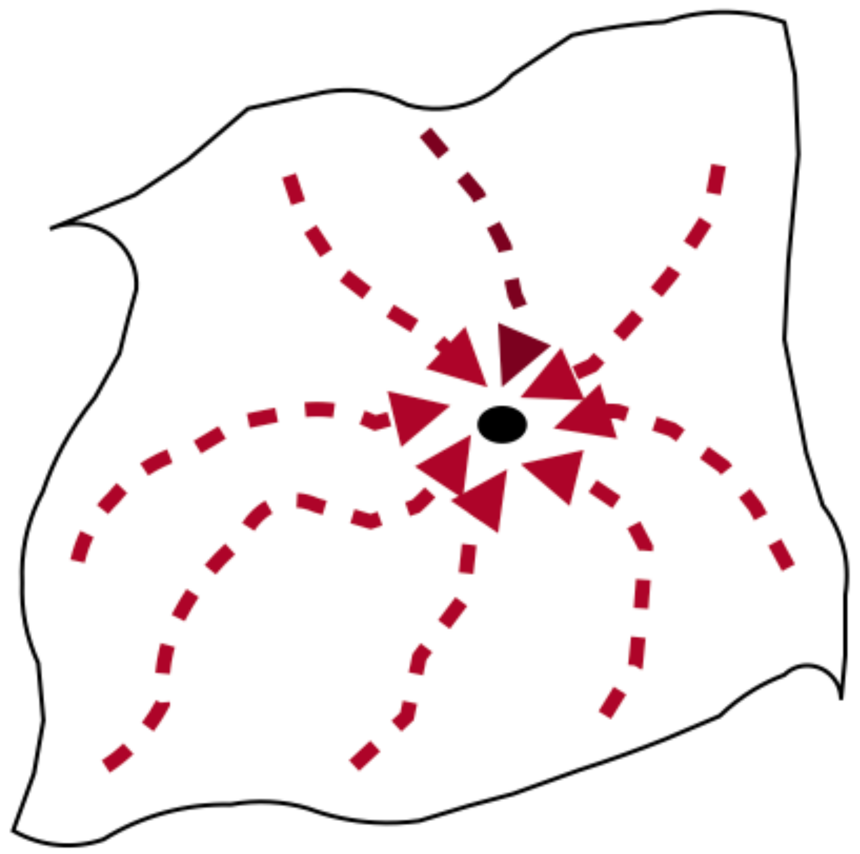
RG flow of ridge parameter σ^2.

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



III. When irrelevant modes are Gaussian





Gaussian irrelevant features: Part I

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

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

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

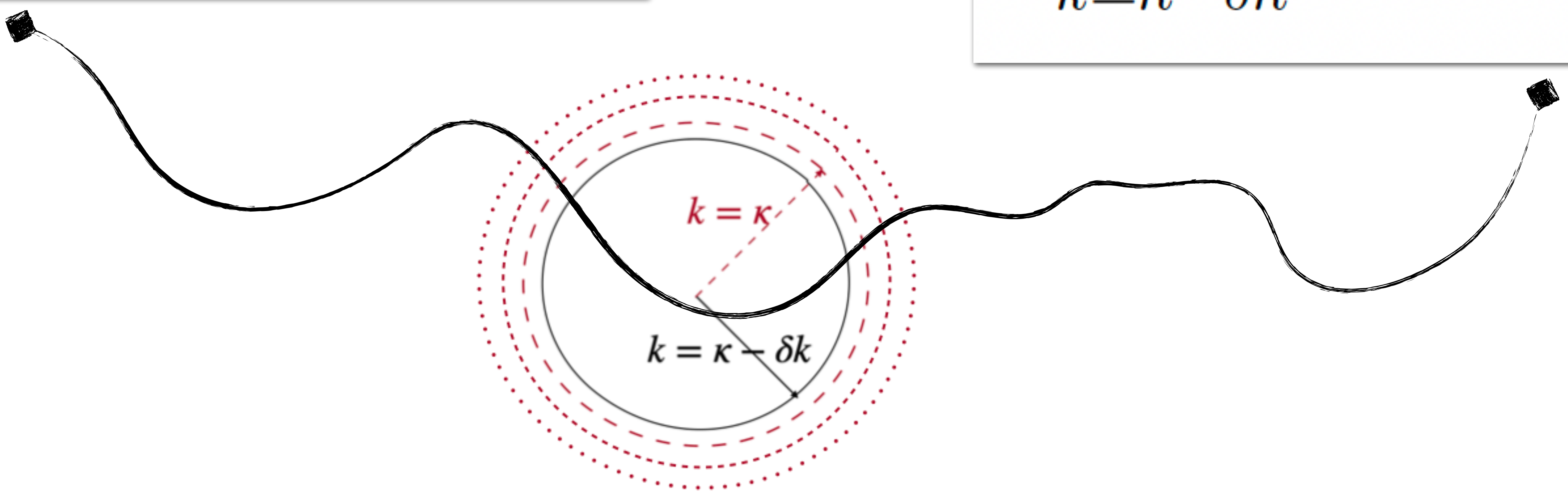
$$\langle Z^M \rangle_\eta = e^{-\eta} \int \prod_m \mathcal{D} f_{m<} \, e^{-S_{\text{eff}}[f_{m<}]}$$

Gaussian irrelevant features: Part II

Assumption over expectation value of GP covariance matrix.

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

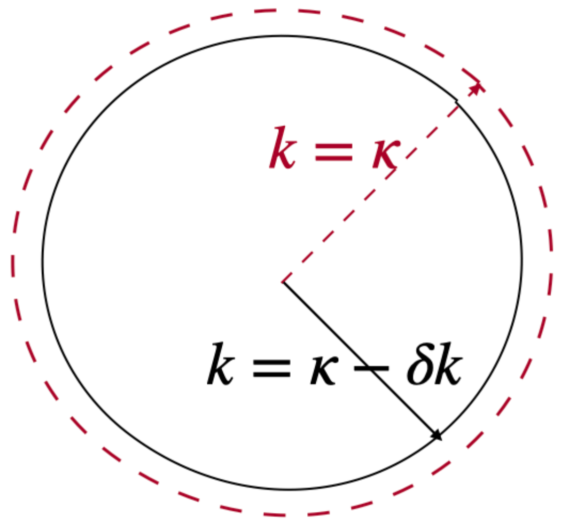
$$\sum_{k=\kappa-\delta\kappa}^{\kappa}\lambda_k =: \delta c \ll \sigma^2$$



Gaussian irrelevant features: Part III

Wilsonian RG over higher feature + GP modes

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

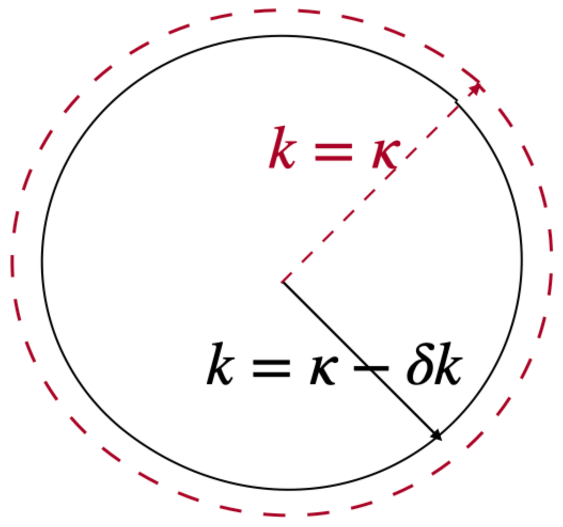


$$S_{\text{int}} = -\eta \int \mathrm{d}\mu_x \, e^{-\frac{1}{2}(\boldsymbol{f}_{m<}(x) - \boldsymbol{y}_<(x))^\top \left(\sigma^2 + C_{y_> = 0}\right)^{-1} (\boldsymbol{f}_{m<}(x) - \boldsymbol{y}_<(x))}$$

Gaussian irrelevant features: Part IV

RG flow of ridge parameter

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



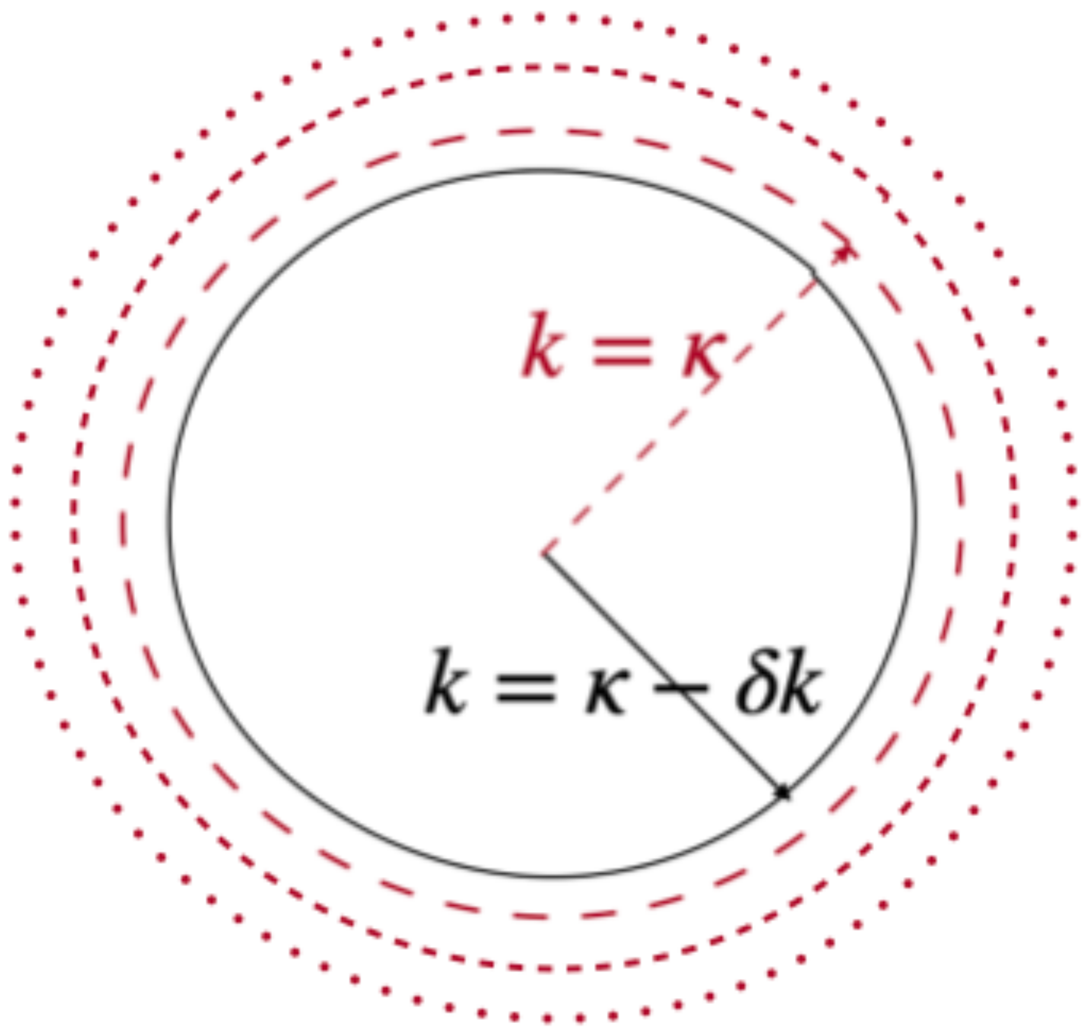
$$S_{\text{int}} = -\eta \int \mathrm{d}\mu_x \, e^{-\frac{1}{2}(\boldsymbol{f}_{m<}(x)-\boldsymbol{y}_<(x))^\top \left(\sigma^2 + C_{y_>=0}\right)^{-1} (\boldsymbol{f}_{m<}(x)-\boldsymbol{y}_<(x))}$$

Each shell

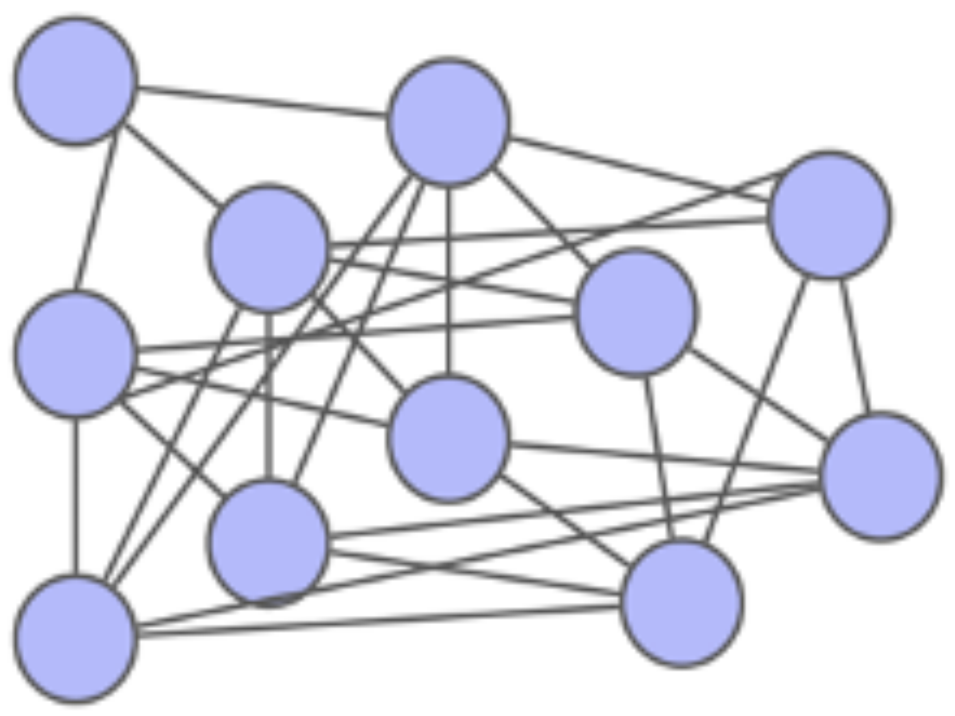
$$\sigma'^2 = \sigma^2 + \delta c$$

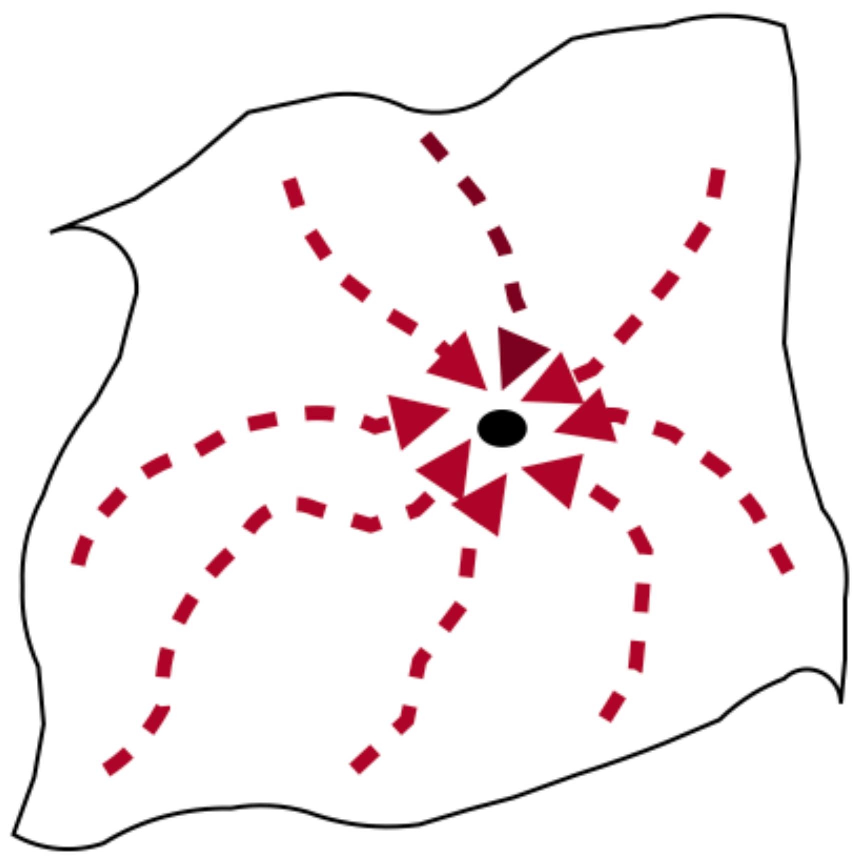
Renormalization of ridge

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



IV. When irrelevant modes are non-Gaussian





Non-Gaussian irrelevant features (I)

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

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

At leading order in δc and $1/d$

$$P[\varphi_q|\varphi_<] \approx \mathcal{N}[0, \mathbb{1}; \varphi_q] \left[1 + A\varphi_q + B\left(\varphi_q^2 - 1\right)\right]$$

$$B := \frac{1}{2!} U_{k_1 k_2 q q} He_2(\varphi_{k_1}, \varphi_{k_2})$$

$$U_{k_1 k_2 k_3 k_4} := \langle \varphi_{k_1} \varphi_{k_2} \varphi_{k_3} \varphi_{k_4} \rangle_{P[\varphi],\text{connected}}$$

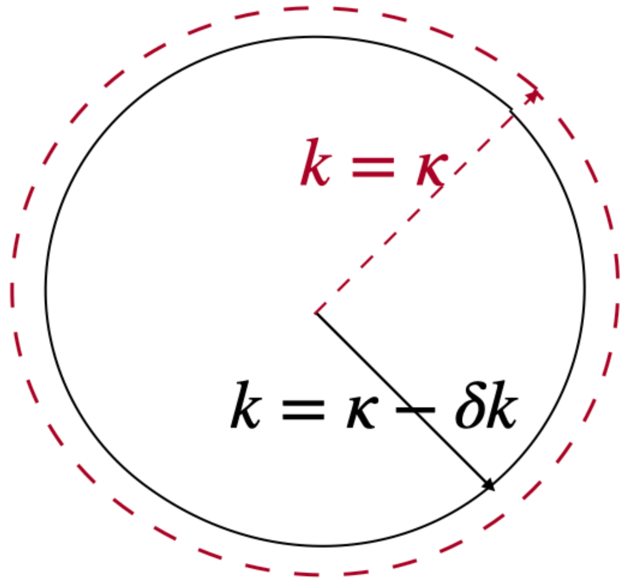
$$A := \frac{1}{3!} U_{k_1 k_2 k_3 q} He_3(\varphi_{k_1}, \varphi_{k_2}, \varphi_{k_3})$$

Hermite polynomials

Non-Gaussian irrelevant features (II)

Spatial re-weighting of MSE loss

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



Integrate over shell $\kappa \equiv q$

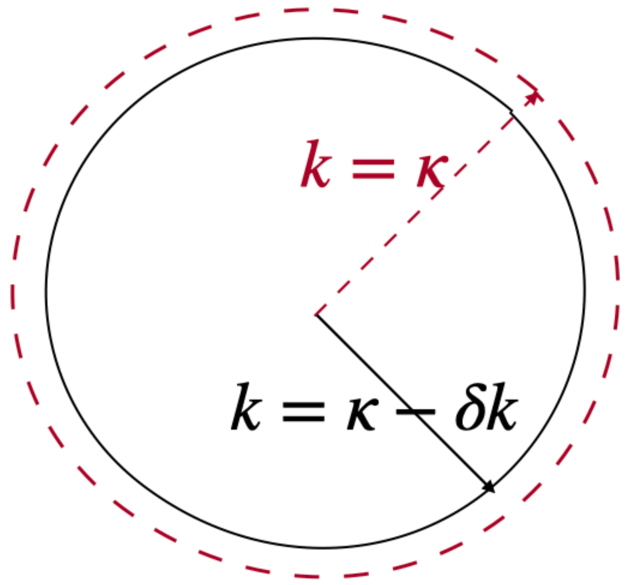
$$\Phi_{m<} := \sum_{k \le \kappa} (f_{mk} - y_k)\, \varphi_k$$

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

Non-Gaussian irrelevant features (III)

Input dependent RG flow of ridge parameter

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



Integrate over shell $\kappa \equiv q$

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

Covariance of GP modes over infinitesimal shell

Each shell

$$\frac{W_{\delta c}(x)}{\sigma_{\delta c}^2} = \frac{W_0(x)}{\sigma_0^2} - \frac{\delta c(1 + 2B_0(x))}{\sigma_0^4}$$

Spatial RG of ridge

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

Conclusion

- We introduce a Wilsonian RG framework for NNGP regression.

- Integrate out modes irrelevant to the inference problem ~ high momentum modes.

- Results in RG flow of the ridge parameter (noise covariance) appearing in average predictor.

- Gaussian and non-Gaussian irrelevant feature modes lead to simple RG and spatial (input-dependent) RG flows of the ridge.

Thank You!

Questions?

https://aninditamaiti.github.io/