UNIVERSITY OF AMSTERDAM

w.i.p. with Miranda Cheng
& Max Welling

Diffusion models & RG

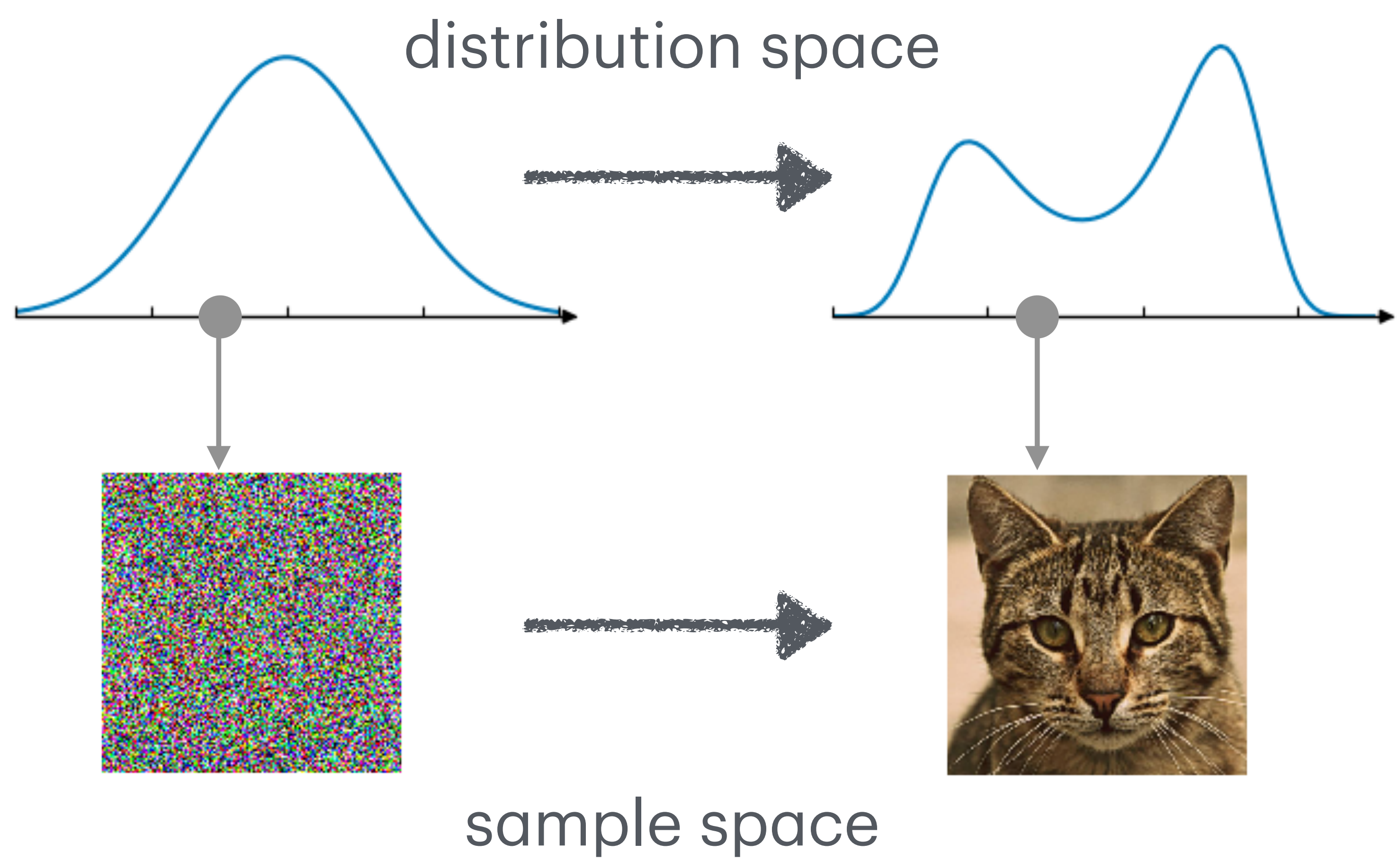
RG inspired perspective on diffusion models

Mathis Gerdes — ECT* 2024 — Machine Learning and the Renormalization Group

Generative Models

Distribution to distribution

- Map between distributions
- Map between samples
- Probabilistic, deterministic



Diffusion models

Autoregressive models

Normalizing flows

VAEs

Energy based models

...

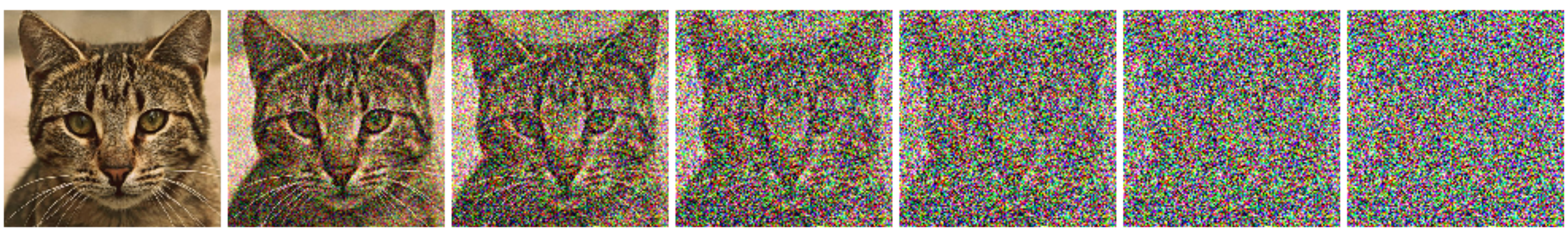
Diffusion Models

Inverting Brownian Motion

Forward: iterative diffusion process $\phi_{t+\delta t} = \sqrt{1-\beta}\,\phi_t + \beta\,\epsilon$

noising

$p_0(\phi)$ $p_1(\phi)$



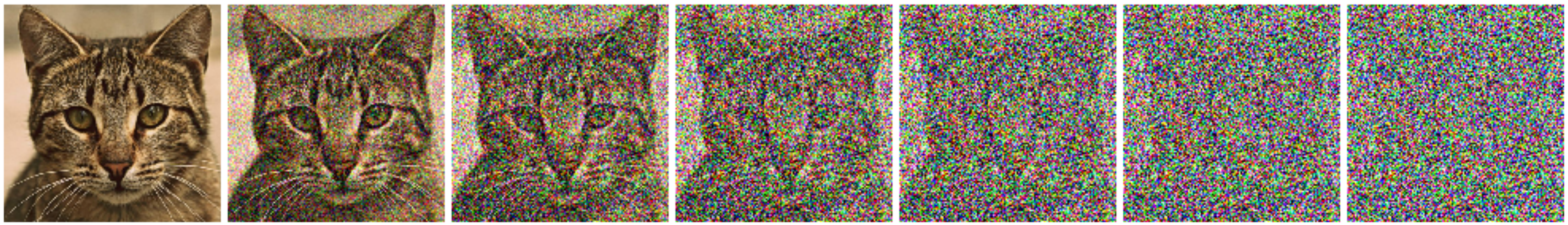
ML inverse noising

We want to learn the inverse "generative" process.

Diffusion Models

Continuum limit

$p_0(\phi)$ $\xrightarrow{t}$ $p_1(\phi)$



In the continuum limit we get a Brownian motion SDE:

$$d\phi = -\frac{1}{2}\beta\,\phi\,dt + \beta\,dw$$

Solving the SDE starting at $p_0(\phi)$ leads to a path in distributions $p_t(\phi)$.

All information about the flow is encoded in the Stein score:

Want to learn $s_\theta(\phi, t) \approx -\nabla_\phi \log p_t(\phi)$ $\longrightarrow$ Know inverse SDE!

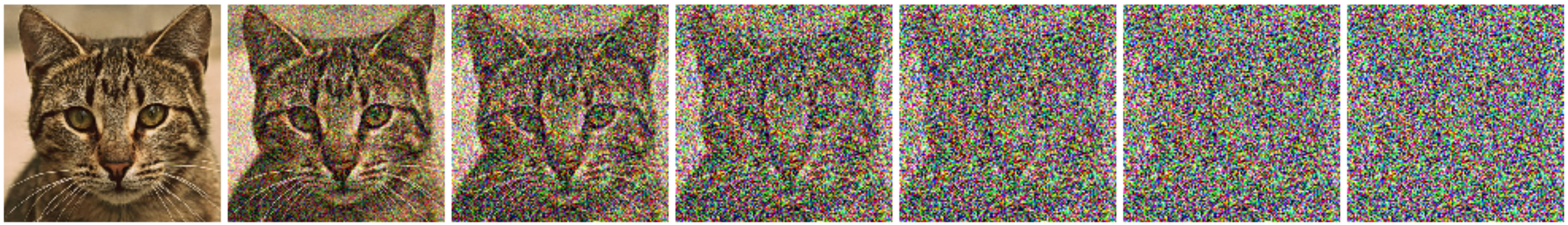
see e.g. [2011.13456] Song et al

Diffusion Models

What makes them work?

$$d\phi = -\frac{1}{2}\beta\,\phi\,dt + \beta\,dw$$

$p_0(\phi)$ $\quad t \longrightarrow \quad$ $p_1(\phi)$



- We can solve the forward SDE exactly: $\phi(t) = \sigma_t \phi(0) + \alpha_t \epsilon$.

Ornstein-Uhlenbeck process

- Given $p_t\left(\phi(t) \mid \phi(0)\right)$ we know a score loss function.

Denoising score-matching

- Can train the score at each "noise level" t independently.

score matching + linear diffusion process + multi scale

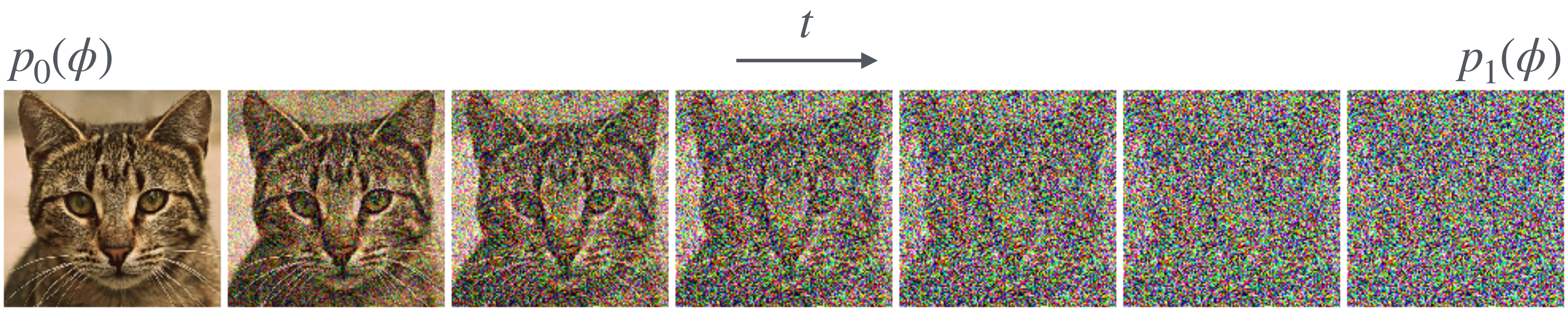
Diffusion model design space

Design degrees of freedom

- Form of diffusion process (prior, noising scheme, scaling)
- Score network architecture
- Conditional training
- Score matching loss function, training scheme
- Combination with other methods and extensions (e.g. latent space diffusion)

Diffusion Models

What makes them work?



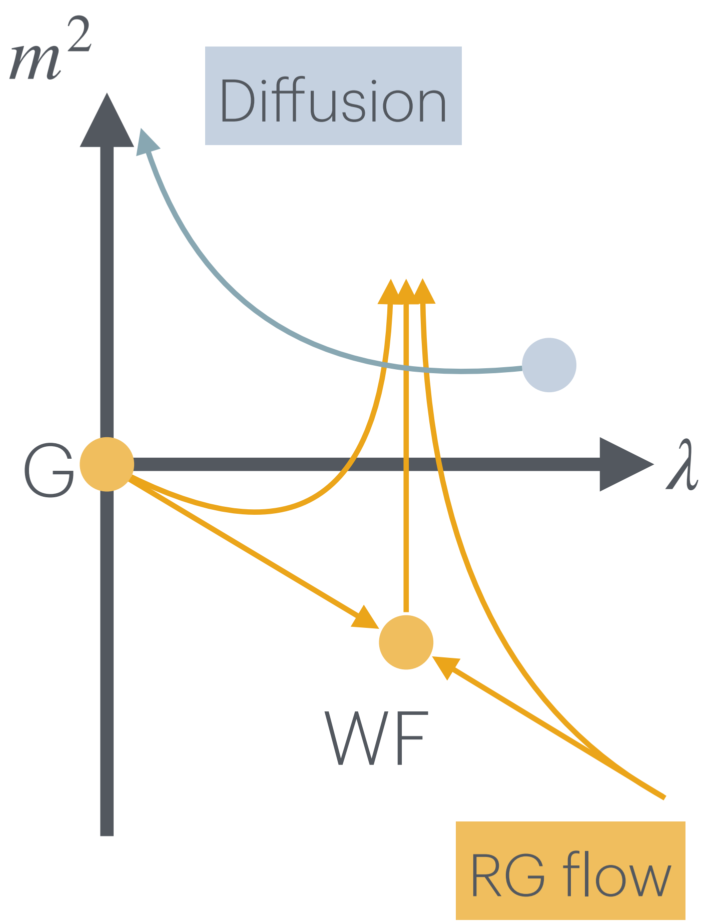
score matching + **linear diffusion process** + **multi scale**

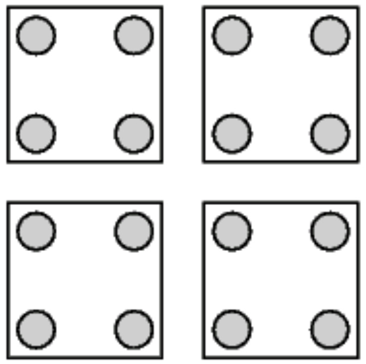
- Can we improve on the forward diffusion process?
- Can we understand and improve on the multi-scale structure?

What can we learn from an RG perspective?

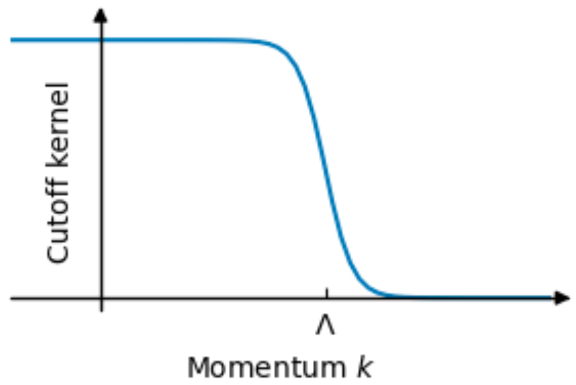
RG Perspectives

Universality & information erasure





Blockspin RG



Momentum space RG

- Flows in distribution space
- Diffusion models: always trivial f.p. given by noise distribution

- Erase/suppress momentum amplitudes from high-k to low-k
- Diffusion models: just add white noise to each pixel!

We want more control!

Can make connection to ERG & gradient flows more rigorous: [2308.12355]

Universality /
Power Spectra

Information Erasure

Forward process

Component-wise schedules

General SDE: $d\phi = F(\phi, t)\, dt + \mathbf{G}(\phi, t)\, dw$

Need simple, solvable process to get $p_t\big(\phi(t) \mid \phi(0)\big)$

Simultaneously diagonalizable: $d\phi = UAU^\dagger \phi\, dt + UBU^\dagger\, dw$

Ornstein-Uhlenbeck process

Variance preserving: $d\phi = \dfrac{1}{2} U\boldsymbol{\beta} U^\dagger \phi\, dt + U\sqrt{\boldsymbol{\beta}} S U^\dagger\, dw$

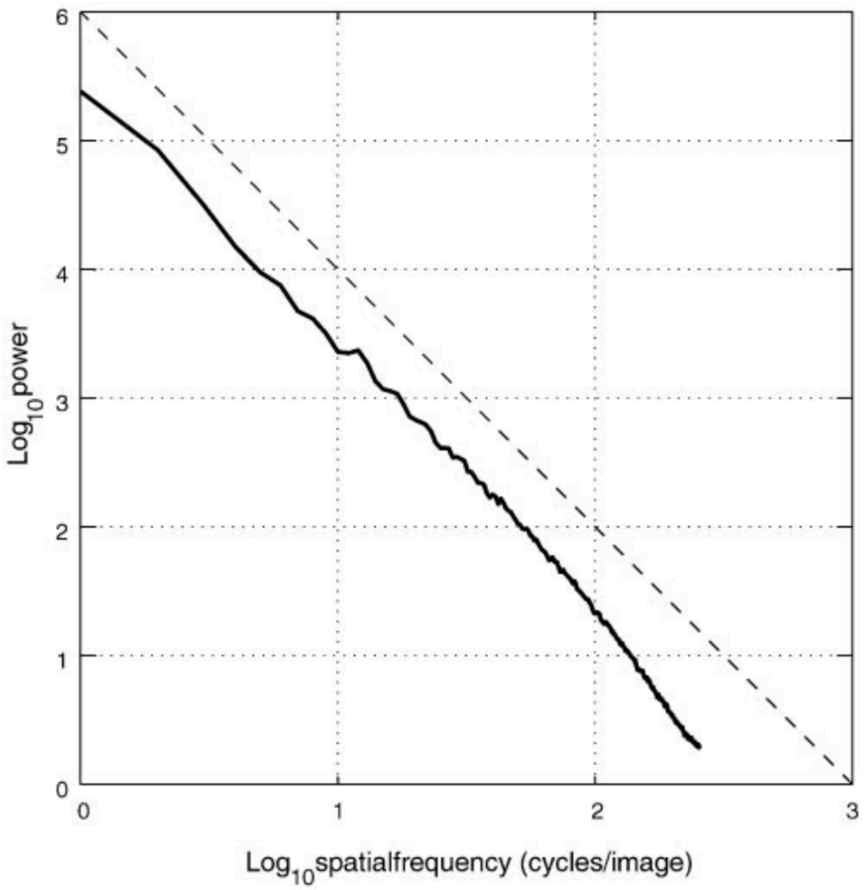
S

Noise "color" and fixed point

$\boldsymbol{\beta}(t)$

Multi-scale information erasure

Noise Spectrum

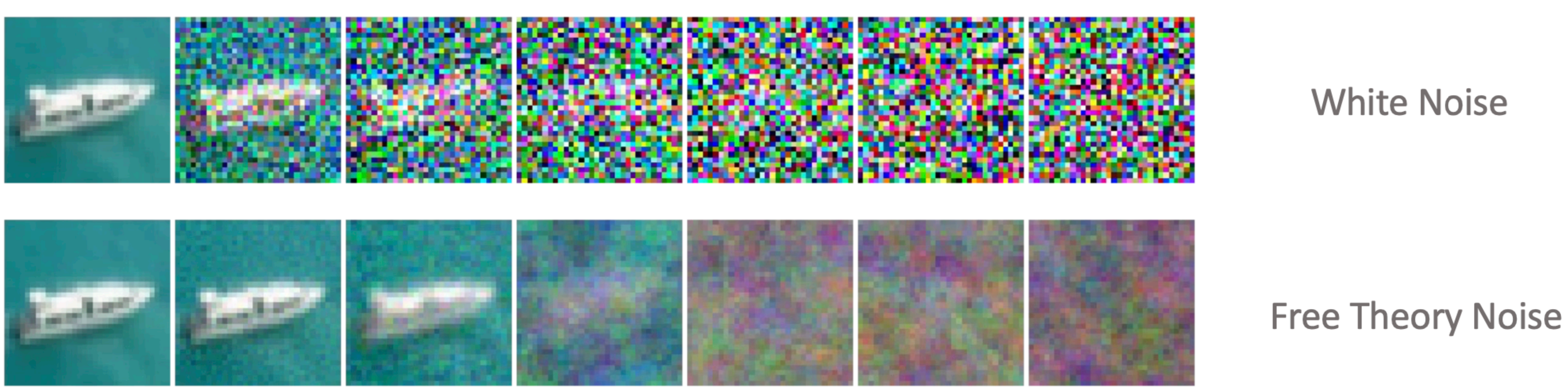


[Simoncelli-Olshausen 2001]

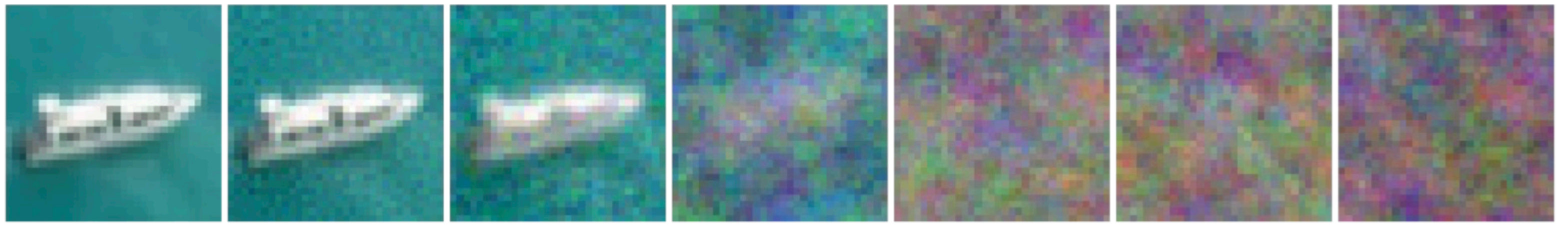
Empirically: Natural images often have power-law spectra.

$$\langle |\phi_k|^2 \rangle = \Sigma_{kk} \sim \frac{1}{k^2}$$

White noise diffusion: transition from data spectrum to white noise.



Noise Spectrum



Free Theory Noise

Initialize with (colored) Gaussian score: $\nabla_\phi \log p_{\text{norm}}(\phi) = -\Sigma^{-1}\phi$

Automatically match second order statistics!

Now network only has to learn **higher order correction**:

$$s_\theta(\phi, t) = \Sigma^{-1}\phi + \text{NN}(\phi, t)$$

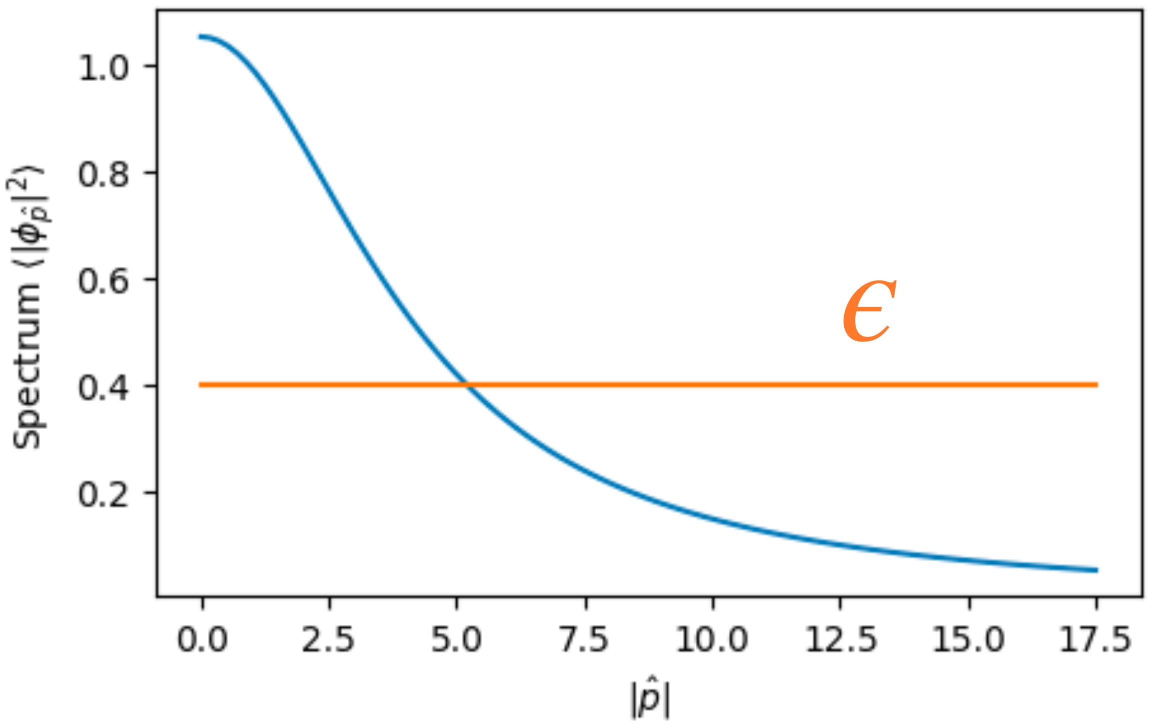
New “fixed point” is a free theory. Matches the data distribution.

Information Erasure

In the usual diffusion models

Diffusion models already *implicitly* destroy information by-scale.

Forward OU: $\phi(t) = \alpha_t \phi(0) + \sigma_t \epsilon$



- Multi-scale information erasure implicit, depending on data magnitude.
- No explicit control over this.

Information Erasure

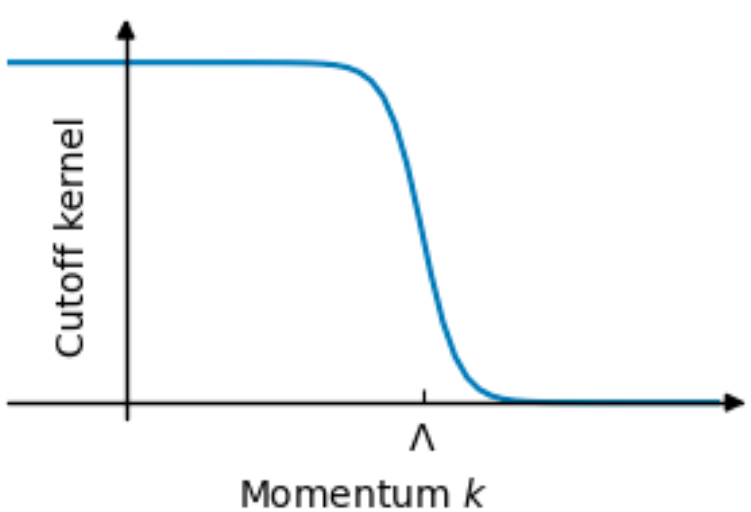
Component-wise schedules

RG intuition: erase information scale-by-scale.

Recall Polchinski RG:

$$S_\Lambda[\phi] = \int \frac{d^D k}{(2\pi)^{D/2}} \tilde{\phi}(k)\tilde{\phi}(-k)\frac{k^2+m^2}{K_\Lambda(k)} + S_{\mathrm{int},\Lambda}[\phi]$$

$K_\Lambda(k)$: cutoff kernel, $K_\Lambda(k) \to 0$ as $|k| \gg \Lambda$



E.g. sigmoid cutoff:

$$K_\Lambda(k) = \sigma(\Lambda - |k|)$$

We can translate this directly into a component-wise $\beta_k(t)$!

Forward process

Component-wise schedules

$$d\phi = \frac{1}{2} U\boldsymbol{\beta} U^{\dagger}\phi\, dt + U\sqrt{\boldsymbol{\beta}} S U^{\dagger}\, dw$$

S	$\boldsymbol{\beta}(t)$
Noise "color" and fixed point	Multi-scale information erasure

RG:	Free theory	Change theory cutoff
ML:	Good initialization matching 2nd order stats	Set how "autoregressive" generative process is

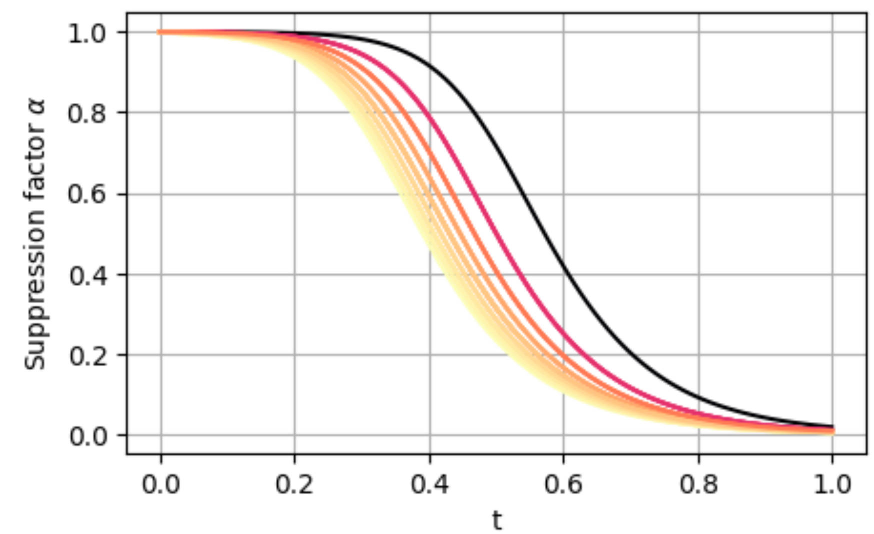
Soft conditioning to auto-regressive

Multi-scale information erasure

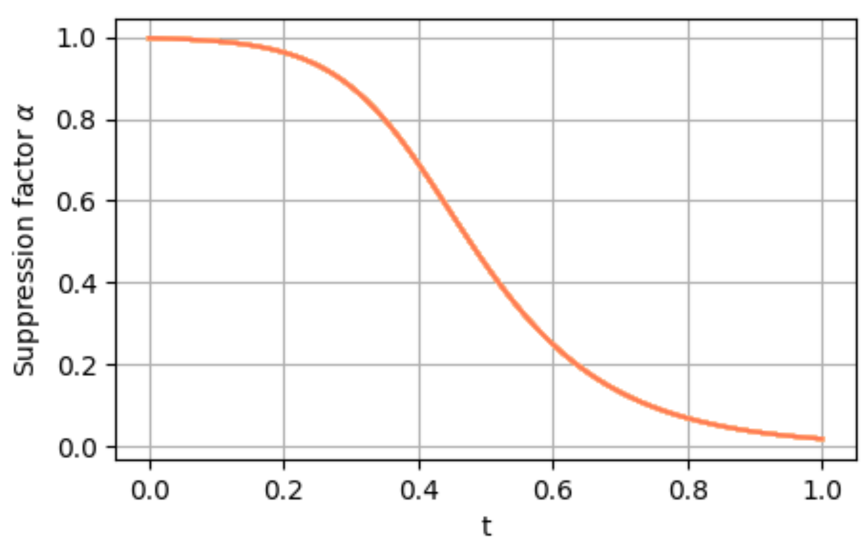
Forward OU: $\phi(t) = \alpha_t \phi(0) + \sigma_t \epsilon$

New hyper-parameter space
to optimize!

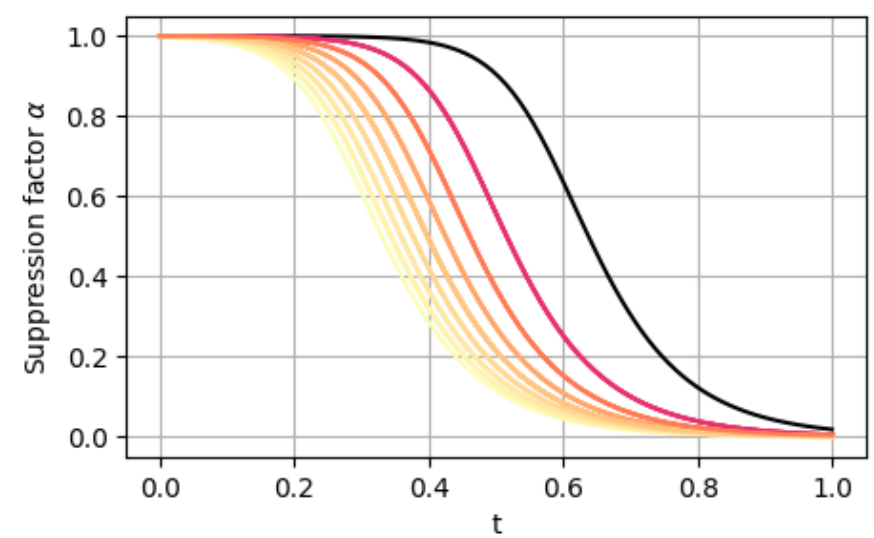
standard diffusion



noise "all at once"



toward "autoregressive"



Noising CIFAR-10

Matching power spectrum, component-specific noising

Suppression of components

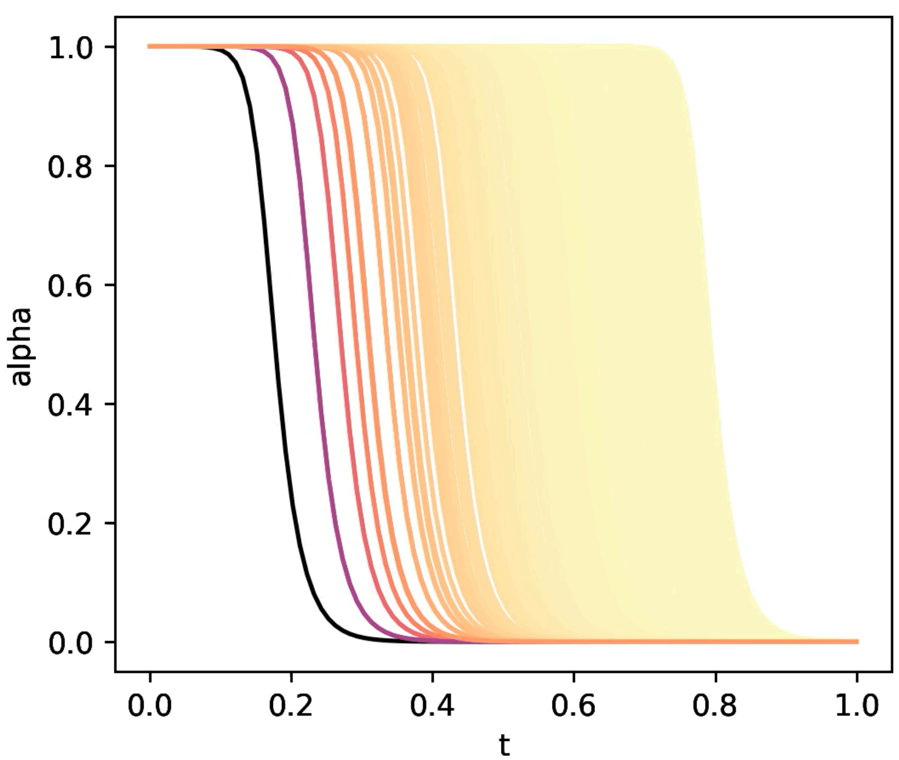
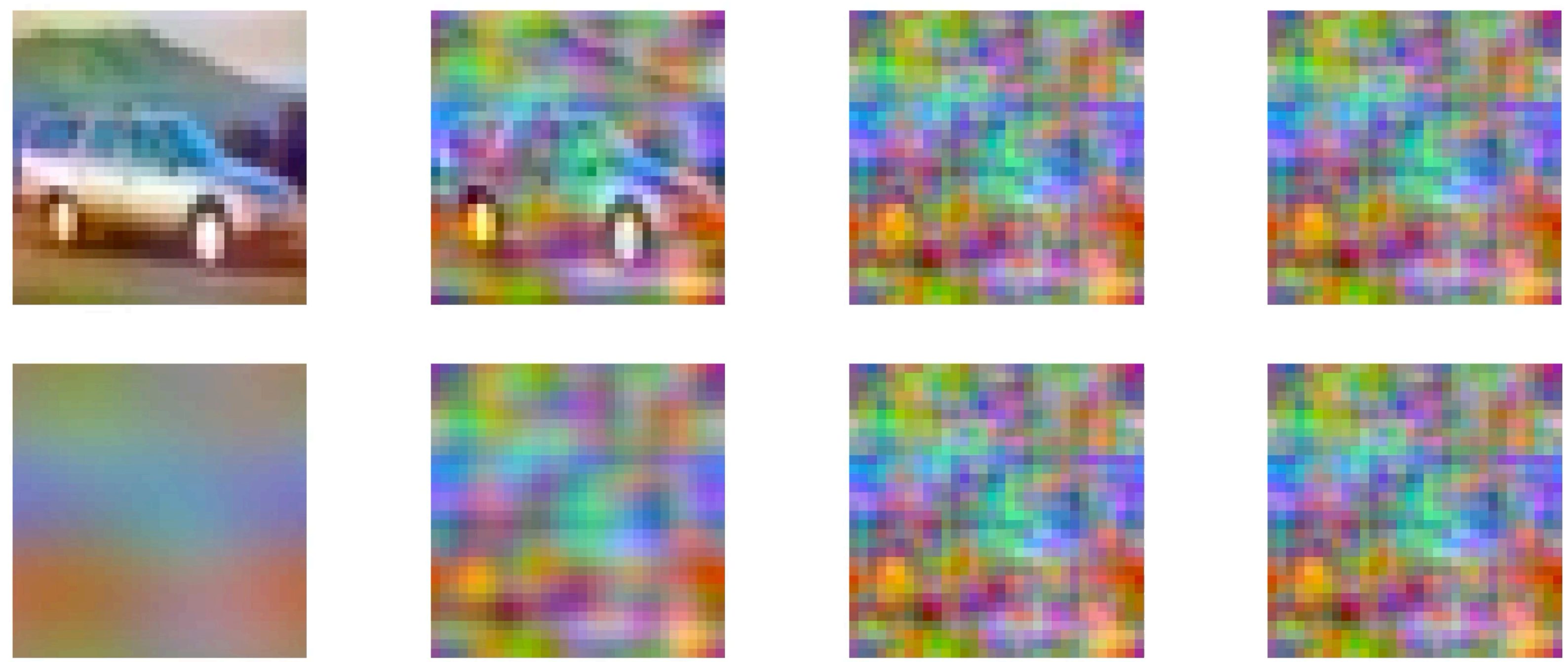


Image noising $\xrightarrow{t}$



Noise component that is added to data

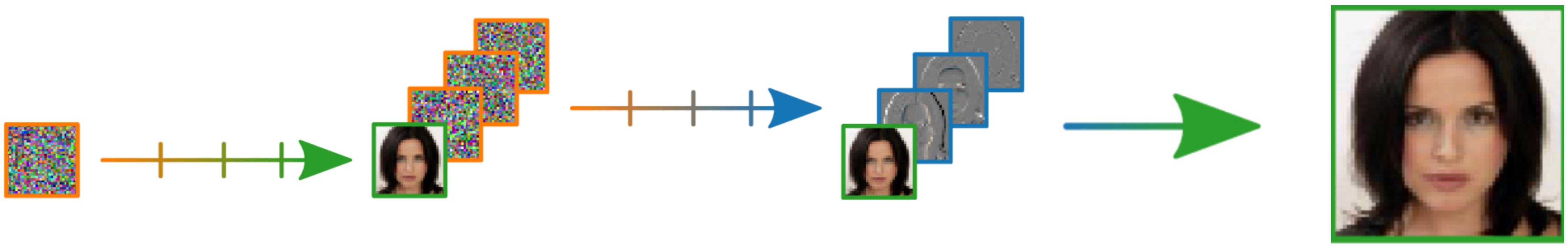
Component spaces

$$d\phi = UAU^{\dagger}\phi\, dt + UBU^{\dagger}\, dw$$

Fourier space	Principle components (PCA)
Momentum components	Whitened PCA components
Physics inspired noise/match PS	Could reinterpret as momenta/ match 2nd order statistics
RG inspired schedule/optimize	— // —

Component spaces

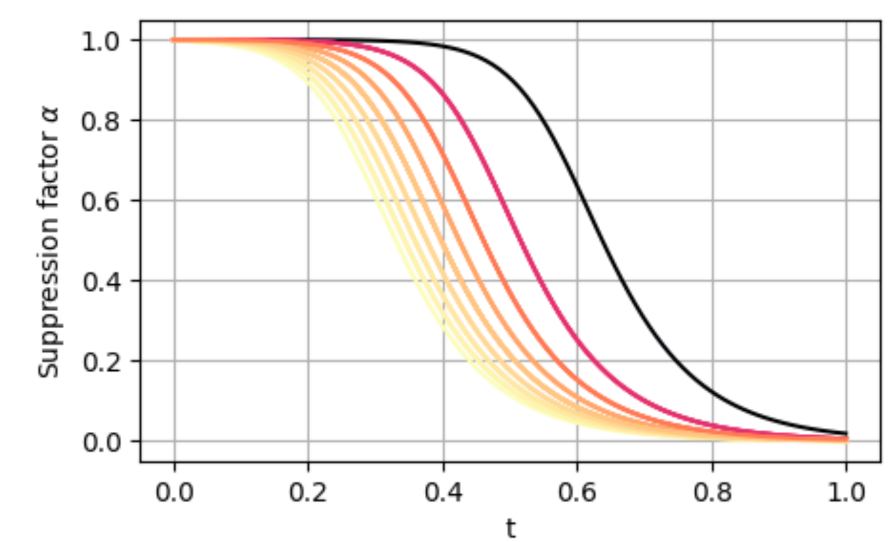
Wavelet components



modified, from [2208.05003]

Special case:

- Linear change of basis U given by wavelets
- "Hard" conditioning: generate each higher-frequency wavelet components given fixed lower-frequency data



Wavelet Score-Based Generative Modeling [2208.05003]

Summary

Tried on 6x6 phi4 samples and experiments on CIFAR-10 (w.i.p.)

- ✅ Matching noise + component-wise schedule is improvement!
- ✅ Choice of local schedule has significant impact.
- ⌛ Exploring various families of schedules, loss functions, datasets.