Learning normalized image densities via dual score matching

FLATIRON INSTITUTE

Florentin Guth
NYU & Flatiron Institute

Zahra Kadkhodaie
Flatiron Institute

Eero Simoncelli
NYU & Flatiron Institute

NYU

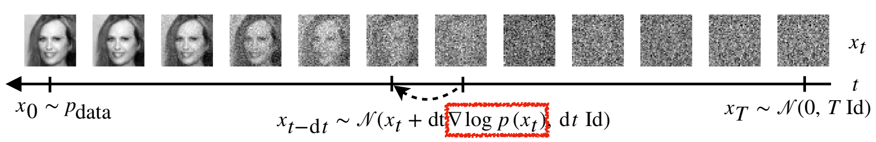
Dual score matching

We want to learn an energy function (negative log probability) that takes small values on the data and high values **everywhere else.**

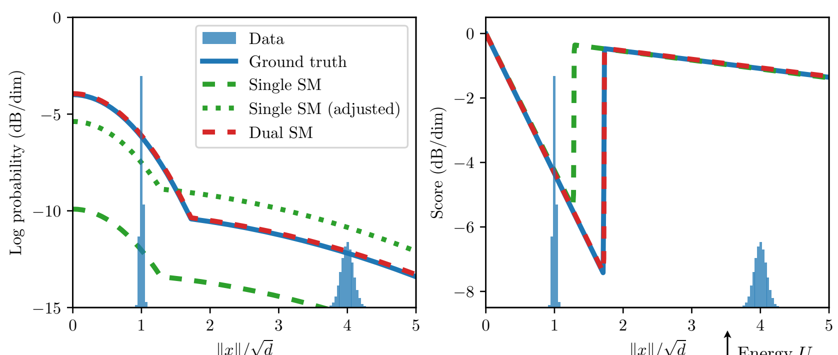
$$p_\theta(x) = \frac{1}{Z_\theta} \mathrm{e}^{-U_\theta(x)}$$

This is difficult because normalization is intractable.

Diffusion generative models sidestep this issue: images are generated by a gradient ascent algorithm on log p, which amounts to denoising!



Integrating the score is not trivial: we need to learn an energy function that is consistent with all scores.



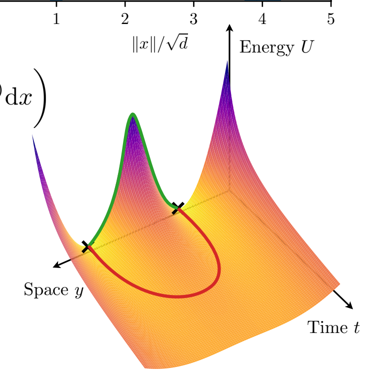
$$U(y,t) = -\log\left(\int p(x) \mathrm{e}^{-\frac{1}{2t}\|y-x\|^2 - \frac{d}{2}\log(2\pi t)} \mathrm{d}x\right)$$

$$\nabla_y U(y,t) = \mathbb{E}_x\left[\frac{y-x}{t} \,|\, y\right]$$

$$\ell_{\mathrm{DSM}}(\theta,t) = \mathbb{E}_{x,y}\left[\|\nabla_y U_\theta(y,t) - \frac{y-x}{t}\|^2\right]$$

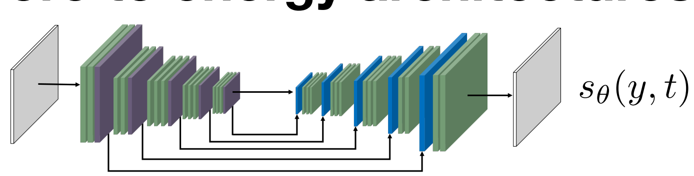
$$\partial_t U(y,t) = \mathbb{E}_x\left[\frac{d}{2t} - \frac{\|y-x\|^2}{2t^2} \,|\, y\right]$$

$$\ell_{\mathrm{TSM}}(\theta,t) = \mathbb{E}_{x,y}\left[\left(\partial_t U_\theta(y,t) - \frac{d}{2t} + \frac{\|y-x\|^2}{2t^2}\right)^2\right]$$



From score to energy architectures

Score network: UNet



We want to define U_θ such that $\nabla_y U_\theta(y,t) = s_\theta(y,t)$

Define $U_\theta(y,t) = \frac{1}{2}\langle y, s_\theta(y,t)\rangle$ with a homogeneous s_θ

Summary

How to learn a prior probability model from data? And what are the properties of the learned prior? We propose a denoising-based objective to learn a prior model. We find that probability and local dimensionality vary greatly based on image content, bringing precisions to the manifold hypothesis.

Do the probabilities generalize?

We trained two models on disjoint sets of N samples. For large N, **they learn the same probability model!**

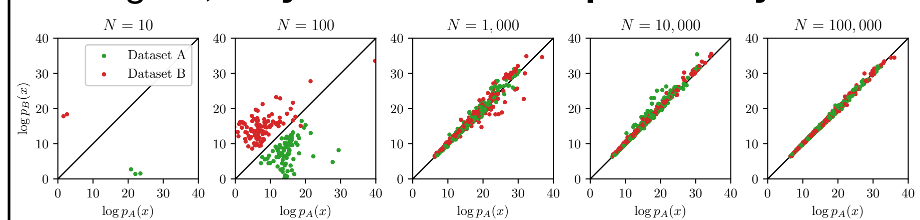
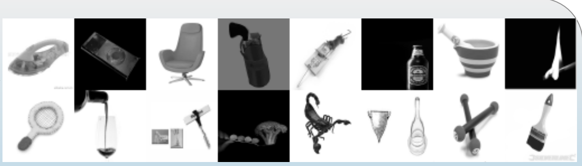
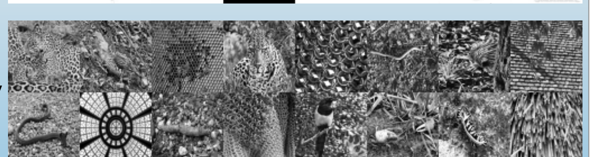


Image content and probability

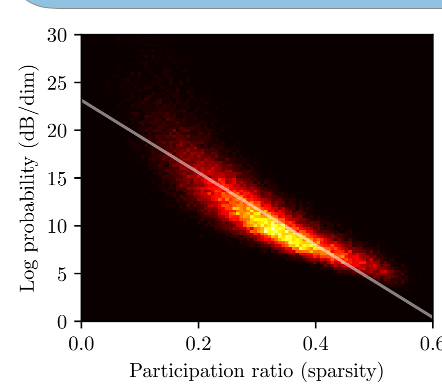
Highest probability

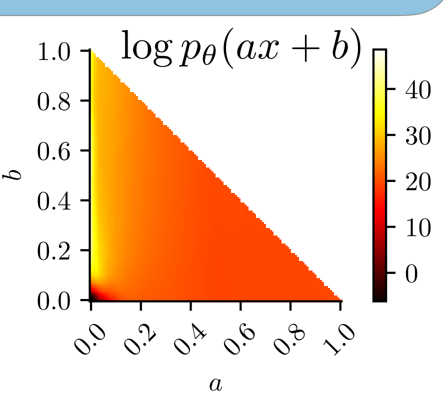


Lowest probability



The probability ratio between these is $10^{14,000}$!





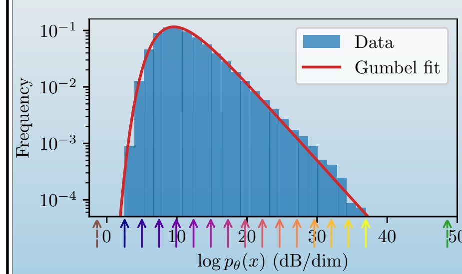
- Sparser and lower contrast images have a higher probability (while brightness has no effect)
- The support of the distribution is connected

Distribution of probabilities

Differential entropy of ImageNet: -11.4 dB/dimension (roughly volume of $[0, 0.1]^d$ compared to $[0, 1]^d$).

After quantization with 256 grayscale levels, there are $10^{5,180}$ natural images out of $10^{9,860}$ total images.

Distribution of probabilities on the ImageNet test set:

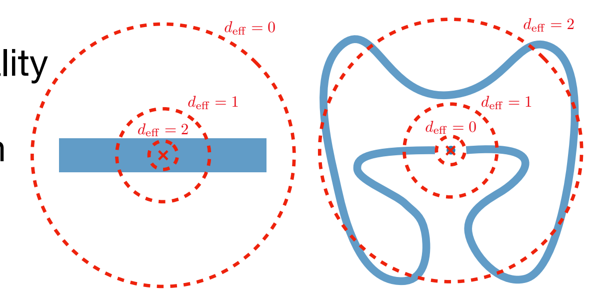


- Very wide and asymmetrical distribution of log p
- Probability is different from typicality
- Most extreme images: constant image and noise

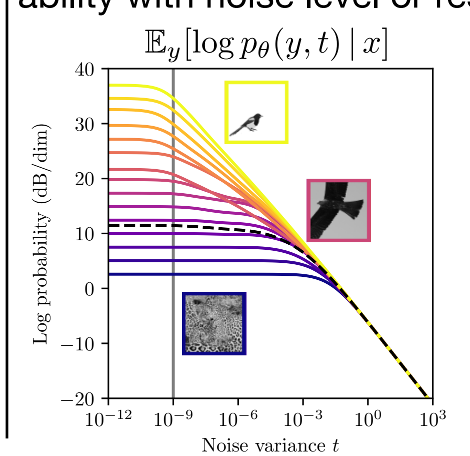
Local dimensionality

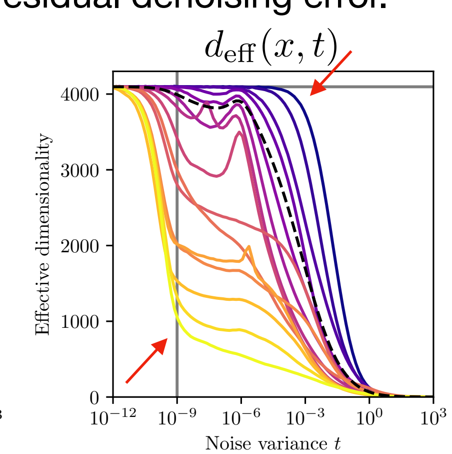
Locally around an image, the probability is supported in some subspace. What is its dimension?

Local dimensionality depends on the image but also on the **size** of the neighborhood.



Dimensionality is given by rate of change of probability with noise level or residual denoising error.





florentin.guth@nyu.edu

https://arxiv.org/abs/2506.05310