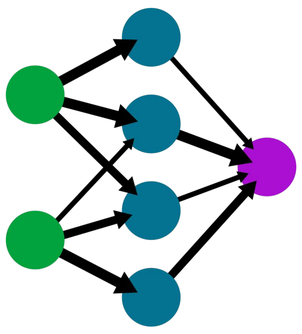
From score matching to score-based generative modeling

Eustache Le Bihan, Maxim Kondracki

Principle - Score Matching

Objective: model high-dimensional probability distributions!

$\mathbf{x} \sim p_{data}$  $f_{\boldsymbol{\theta}}(\mathbf{x}) \rightarrow \frac{e^{f_\theta}}{Z_\theta}$

- Positivity
- Sum to one

$$Z_{\boldsymbol{\theta}} = \int e^{f_{\boldsymbol{\theta}}(\mathbf{x})} \mathrm{d}\mathbf{x}$$

Intractable !

Solution: Stein score ! $s_\theta(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_{\boldsymbol{\theta}}(\mathbf{x}) = \nabla_{\mathbf{x}} f_{\boldsymbol{\theta}}(\mathbf{x})$ No Z_θ!

Explicit Score Matching (ESM) (Fisher divergence):

$$J_{ESM}(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}_{p_{\boldsymbol{data}}(\boldsymbol{x})}[\|\nabla_{\mathbf{x}} \log p_{\boldsymbol{data}}(\boldsymbol{x}) - s_\theta(\boldsymbol{x})\|^2] \quad (1)$$

Implicit Score Matching (ISM), n i.i.d. samples $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^d$:

$$\hat{J}_{ISM}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}[Tr(\nabla_{\mathbf{x}} s_\theta(\mathbf{x}_i)) + \frac{1}{2}\|s_\theta(\mathbf{x}_i)\|^2] \quad (2)$$

[1]: $\hat{\boldsymbol{\theta}}_{n,ISM} \xrightarrow{p} \hat{\boldsymbol{\theta}}^*_{ESM}$, where $p_{\hat{\boldsymbol{\theta}}^*_{ESM}} = p_{data}$, $\hat{\boldsymbol{\theta}}_{n,ISM}$ min. of $\hat{J}_{ISM}$

In practice

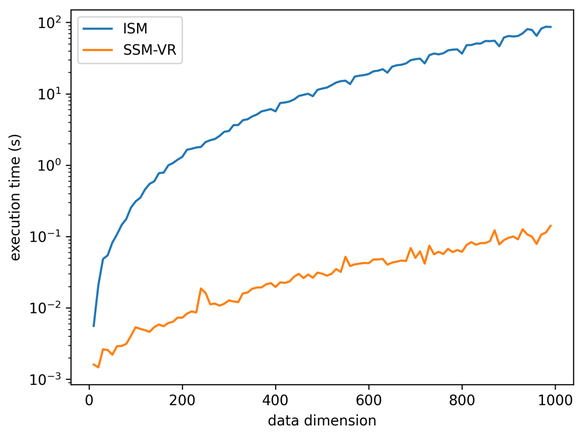
Problem: computing $Tr(\nabla_{\mathbf{x}} s_\theta(\mathbf{x}_i))$ requires d backpropagation passes!
Solution: sliced score matching (SSM)!

SSM: vector field $s_\theta(\boldsymbol{x}) \rightarrow$ projection into scalar field

$$J_{SlicedSM}(\boldsymbol{\theta}) = \mathbb{E}_{p_{\mathbf{v}}}\mathbb{E}_{p_{\boldsymbol{data}}(\boldsymbol{x})}[\mathbf{v}^T \nabla_{\mathbf{x}} s_\theta(x)\mathbf{v} + \frac{1}{2}(\mathbf{v}^T s_\theta(\boldsymbol{x}))^2] \quad (3)$$

$\mathbf{v}^T \nabla_{\mathbf{x}} s_\theta(x)\mathbf{v} = \mathbf{v}^T \nabla_{\mathbf{x}}(\mathbf{v}^T s_\theta(x))$:

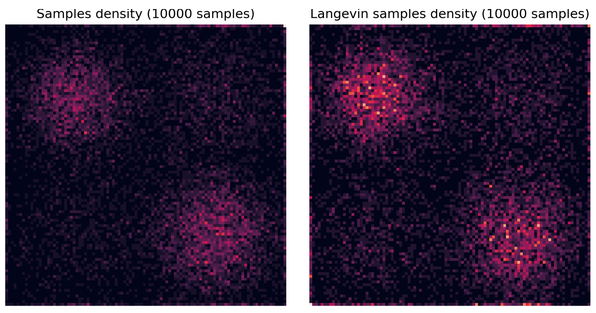
- $\mathbf{v}^T s_\theta(x)$: one forward pass (add one neuron at the end of the comp. graph)
- $\nabla_{\mathbf{x}}(\mathbf{v}^T s_\theta(x))$: one backpropagation



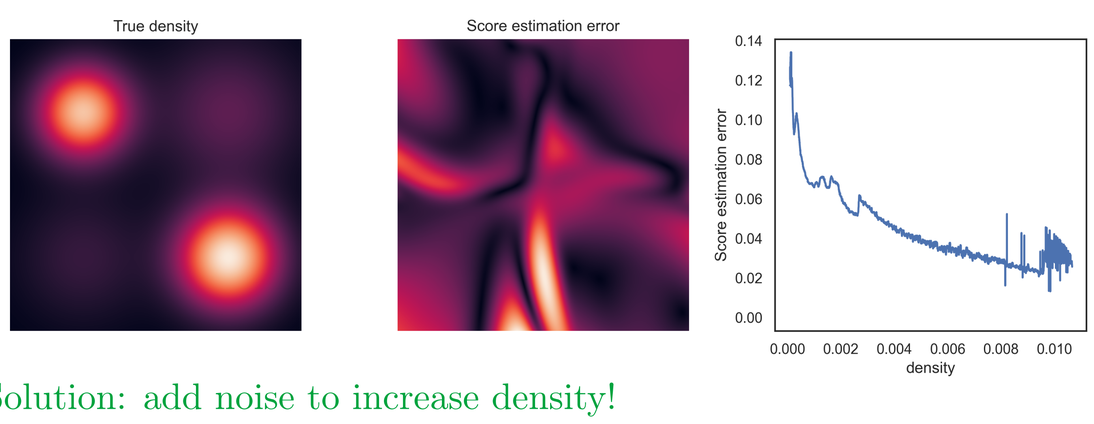
Sampling - Langevin Dynamics

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_{t-1} + \frac{\epsilon}{2} s_\theta(\tilde{\mathbf{x}}_{t-1}) + \sqrt{\epsilon}\,\mathbf{z}_t$$

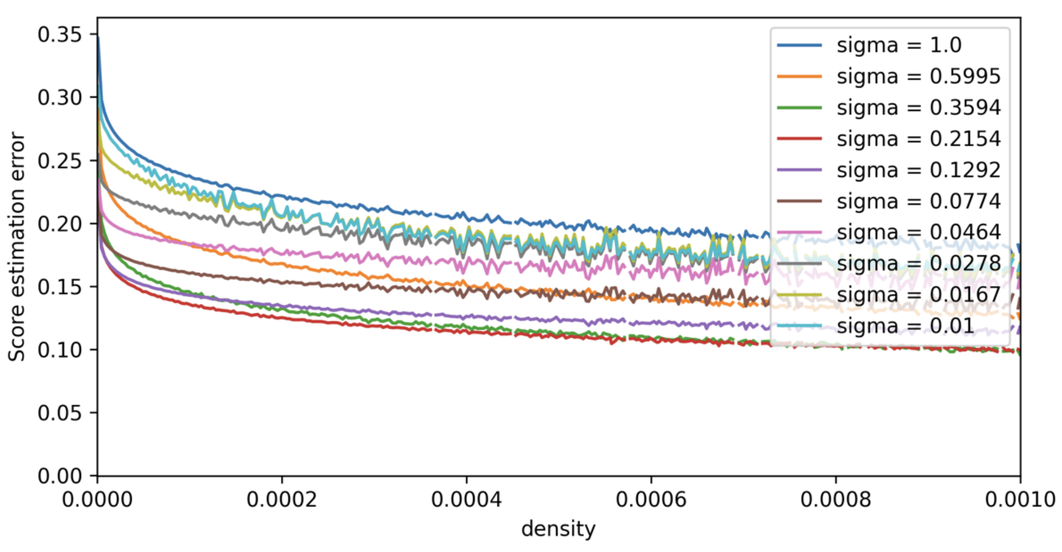
0. start from prior $\tilde{\mathbf{x}}_0$
1. follow gradient
3. add noise (avoid to collaps in low divergence regions)



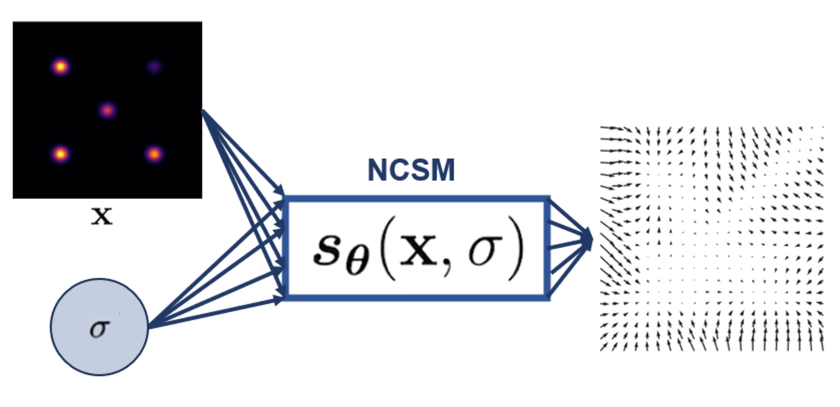
Problem: in low-density data regions, score estimation errors surge!



Solution: add noise to increase density!

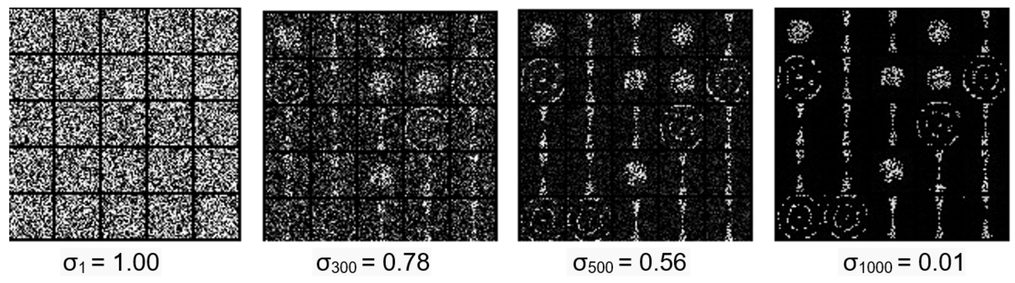


Noise Conditional Score Network (NCSM)

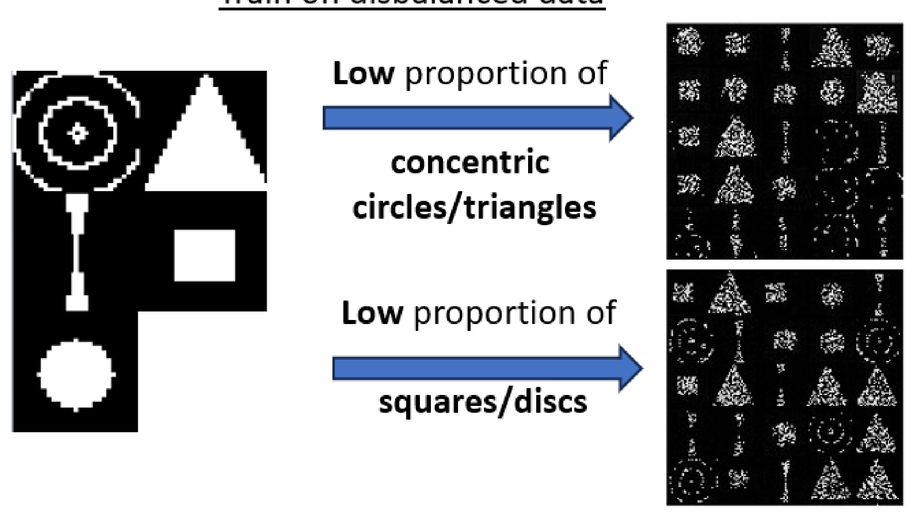


Principle: add noise to increase data density and progressively reduce noise as Langevin Dynamics reaches high-density regions.

NCSM : train a conditional score network to estimate the scores of perturbed data distributions, i.e., $\forall \sigma \in \{\sigma_i\}_{i=1} : s_\theta(x, \sigma) \approx \nabla_x \log q_\sigma(x)$ where $q_\sigma(x)$ is the perturbed data distribution with Gaussian noise $\mathcal{N}(\mu, \sigma_i)$



Train on disbalanced data



Concentric circles and triangles exhibit complex distribution that require more samples for accurate representation. Their pixels distribution is more dispersed compared to squares and circles, which have their maximum concentration centered in the middle of the image.

References

[1] Estimation of Non-Normalized Statistical Models by Score Matching, Hyvärinen, Aapo,Journal of Machine Learning Research. 2005

[2] A Connection Between Score Matching and Denoising Autoencoders; Vincent, Pascal; University of Montreal, 2010

[3] Generative Modeling by Estimating Gradients of the Data Distribution; Yang Song and Stefano Ermon; CoRR, 2019