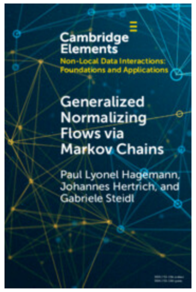


Lecture 3: Normalizing Flows

Gabriele Steidl
Applied Mathematics - Imaging Sciences
TU Berlin

Berlin Mathematics Research Center MATH+

Funded under Germany's Excellence Strategy by DFG Deutsche Forschungsgemeinschaft

Mathematisches Forschungsinstitut Oberwolfach

Member of Leibniz Leibniz Association

Oberwolfach Seminar

Variational and Information Flows in Machine Learning and Optimal Transport

Organizers: Wuchen Li, Columbia
Bernhard Schmitzer, Göttingen
Gabriele Steidl, Berlin
Francois-Xavier Vialard, Paris
Date (ID): 19 – 25 November 2023 (2347b)
Deadline: 1 September 2023

Variational and stochastic flows are now ubiquitous in machine learning and generative modeling. Indeed, many such models can be interpreted as flows from a latent distribution to the sample distribution and training corresponds to finding the right flow vector field. Optimal transport and diffeomorphic flows provide powerful frameworks to analyze such trajectories of distributions with elegant notions from differential geometry, such as geodesics, gradient and Hamiltonian flows. Recently, mean field control and mean field games offer a general optimal control variational problems on the learning problem. How do these tools lead us to a better understanding and further development of machine learning and generative models?

The Oberwolfach Seminar will address the topic from different points of view taking in particular recent developments in machine learning into account. The target audience is PhD students and post-doctoral researchers wishing to be quickly immersed in this modern, active research area. Priority will be given to young, motivated researchers.

Please see the website of the seminar for detailed information:

www.mfo.de/occasion/2347b

The seminar takes place at the Mathematisches Forschungsinstitut Oberwolfach. The Institute covers board and lodging. By the support of the Carl Friedrich von Siemens Foundation travel expenses can be reimbursed up to 150 EUR in average per person (against copies of travel receipts). The number of participants is restricted to 25.

Applications including title, ID and date of the intended seminar, together with **one pdf-file attached** containing

- full name and address, incl. e-mail address
- short CV and publication list
- present position, university
- name of supervisor of Ph.D. thesis
- a short summary of previous work and interest

should be **sent by e-mail** via **seminars@mfo.de** until 1 September 2023 to:

Prof. Dr. Matthias Hieber
Mathematisches Forschungsinstitut Oberwolfach
Schwarzwaldstr. 9 – 11
77709 Oberwolfach
Germany

www.mfo.de/scientific-program/meetings/oberwolfach-seminars

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58

- Generative Adversarial Networks (GANs) (Goodfellow et al. 2014)
- Variational Auto-Encoders (Kingma/Welling 2014)
- Diffusion Flows (Zhang/Chen 2021, ...)
- Invertible Networks
  - Residual Networks (Behrmann, Chen et al. 2019)
  - Normalizing Flows - Directly Invertible Neural Networks (Dingh et al. 2017, Aridizzone et al. 2019)

  - for continuous normalizing flows
    see, e.g. Ruthotto/Haber 2020, Hagemann/Hertrich/St. Overview paper: Generalized Normalizing Flows via Markov Chains 2021

Push Forward Measures

- $T : X \to Y$ (Borel) measureable, $\mu \in \mathcal{P}(X)$

Push forward measure

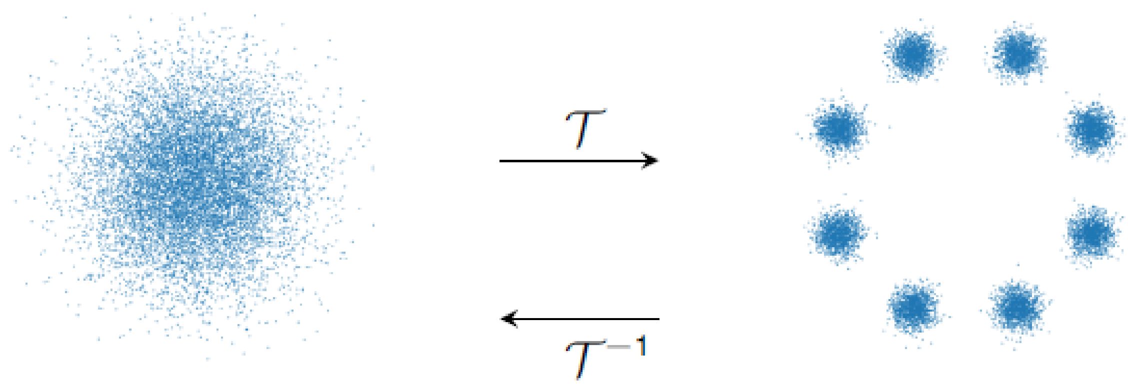
$$T_{\#}\mu := \mu \circ T^{-1}$$

Relation

$$\int_{T^{-1}(A)} h(T(x)) d\mu(x) = \int_A h(y) d\underbrace{(T_{\#}\mu)}_{\nu}(y)$$

Change of variable formula: in case of existing density p_μ and a diffeomorphism T:

$$p_{T_{\#}\mu}(y) = p_\mu\left(T^{-1}(y)\right) \left|\det \nabla T^{-1}(y)\right|$$



Latent distribution P_Z Data distribution P_X

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58

Outline

1. Residual Networks (ResNet)
   - Proximal NNs within ResNets
2. Normalizing Flows
   - Applications in Inverse Problems - The Power of Patches

Residual Networks: composition $\mathcal{T}_N \circ \ldots \circ \mathcal{T}_1$ of layers of the form

$$\mathcal{T}_k(x) = x + \Phi(x;\theta_k), \qquad \Phi(x;\theta_k) \text{ subnetworks}$$

All $\mathcal{T}_k$ are networks themselves. But they are chosen to be invertible as follows.

Invertible Residual Networks: $x^{(0)} = y(= x + \Phi(x;\theta_k))$

$$x^{(r+1)} = y - \Phi(x^{(r)};\theta_k).$$

Sequence $(x^{(r)})_r$ converges to $\mathcal{T}_k^{-1}(y)$ if $\operatorname{Lip}(\Phi(\cdot;\theta_k)) < 1$

BIG effort to ensure Lipschitz continuity in learning process!

Proposal:

$$\mathcal{T}_k(x) = x + \gamma_k \underbrace{\Phi(x;\theta_k)}_{PNN}, \quad \gamma_k > 0$$

Refs: Behrmann et al. 2019, Chen et al. 2019, J. Hertrich: Proximal Residual Flows for Bayesian Inverse Problems, 2022

Theorem Let $\Phi = (1-t)I + tR$, $t \in (0,1)$ be an averaged operator. If $\frac{1}{2} < t \leq 1$, let $0 < \gamma < \frac{1}{2t-1}$ and $\gamma > 0$ otherwise. Then $\mathcal{T}(x) := x + \gamma\Phi(x)$ is invertible and $\mathcal{T}^{-1}(y)$ is the limit of

$$x^{(r+1)} = \frac{1}{1+\gamma-\gamma t}y - \frac{\gamma t}{1+\gamma-\gamma t}R(x^{(r)}),$$

Proof: R is 1-Lipschitz. For $\frac{1}{2} < t \leq 1$ we have $\gamma < \frac{1}{2t-1} \quad \Leftrightarrow \quad \frac{\gamma t}{1+\gamma-\gamma t} < 1$. By Banach's fixed point theorem the series converges to the unique fixed point of

$$x = \frac{1}{1+\gamma-\gamma t}y - \frac{\gamma t}{1+\gamma-\gamma t}R(x)$$

Then

$$\begin{aligned} y &= (1+\gamma-\gamma t)x + \gamma t R(x) \\ &= x + \gamma\left((1-t)x + tR(x)\right) \\ &= x + \gamma\Phi(x) \end{aligned}$$

Thus, $x = \mathcal{T}^{-1}(y)$.

Minimization of log likelihood loss:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log p_{\mathcal{T}_{\theta\#}^{-1} P_Z}(x_i).$$

We have to evaluate and differentiate $\log|\det \nabla \mathcal{T}|$ which is computationally intractable in high dimensions.

Theorem Let Q be a random variable on $\mathbb{Z}_{>0}$ such that $P(Q = k) > 0$ for all $k \in \mathbb{Z}_{>0}$ and $p_k := P(Q \geq k)$. Let $\mathcal{T}(x) = x + \Phi(x)$, where $\Phi\colon \mathbb{R}^n \to \mathbb{R}^n$ is differentiable and fulfills $\mathrm{Lip}(\Phi) < 1$. Then, it holds

$$\log(|\nabla \mathcal{T}(x)|) = \mathbb{E}_{v\sim\mathcal{N}(0,I), q\sim P_Q}\left[\sum_{k=1}^{q} \frac{(-1)^{k+1}}{k} \frac{v^\mathsf{T}(\nabla\Phi(x))^k v}{p_k}\right]$$

and

$$\frac{\partial}{\partial\theta}\log(|\nabla \mathcal{T}(x)|) = \mathbb{E}_{v\sim\mathcal{N}(0,I), q\sim P_Q}\left[\left(\sum_{k=0}^{q} \frac{(-1)^k}{p_k} v^\mathsf{T}(\nabla\Phi(x))^k\right) \frac{\partial(\nabla\Phi(x))}{\partial\theta} v\right].$$

Conditional Proximal Residual Flows

Setting: $Y = F(X) + \eta$,
where $F\colon \mathbb{R}^n \to \mathbb{R}^d$ is an ill-posed/ill-conditioned forward operator and η is some noise.

Aim: NN for reconstructing all posterior distributions $P_{X|Y=y}$, $y \in \mathbb{R}^d$
$\mathcal{T}_\theta\colon \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}^n$ such that $\mathcal{T}(y,\cdot)$ is invertible for all $y \in \mathbb{R}^d$ and

$$P_{X|Y=y} = \mathcal{T}_\theta(y,\cdot)^{-1}_{\#} P_Z.$$

Learn $\mathcal{T}_\theta$ from i.i.d. samples $(x_1, y_1), ..., (x_N, y_N)$ of (X, Y) using the maximum likelihood loss

$$\mathcal{L}(\theta) = \sum_{i=1}^N p_{\mathcal{T}_\theta(y_i,\cdot)^{-1}{}_{\#} P_Z}(x_i) \approx \mathbb{E}_{y\sim P_Y}[\mathrm{KL}(P_{X|Y=y}, P_{\mathcal{T}_\theta(y,\cdot)^{-1}{}_{\#} P_Z})] + \text{const.}$$

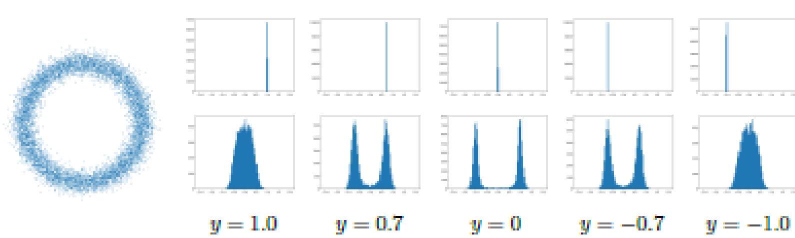


Figure 2: Left: Samples from the prior distribution of X for the circle example. Right: Histograms of samples from the reconstructed posterior distribution $P_{X|Y=y} \approx \mathcal{T}(\cdot, y)^{-1}_{\#} P_Z$ for $y \in \{1, 0.7, 0, -0.7, -1\}$ within the circle example. Top: first coordinate, Bottom: second coordinate.

$$J(\theta) = \mathbb{E}_{x\sim P}\left[\operatorname{trace}(\nabla_x^2 \log(q(x,\theta))) + \tfrac{1}{2}\|\nabla_x \log(q(x,\theta))\|^2\right].$$

Lemma:

Let $A \in \mathbb{R}^{d\times d}$ be an arbitrary matrix and let V be a random variable with $\mathbb{E}(V) = 0$ and $\operatorname{Cov}(V) = I$. Then, it holds

$$\operatorname{trace}(A) = \mathbb{E}(V^\top A V).$$

Using the lemma, J can be rewritten as

$$J(\theta) = \mathbb{E}_{x\sim P, v\sim\mathcal{N}(0,1)}\left[v^\top \nabla_x^2 \log(q(x,\theta))v + \tfrac{1}{2}\|\nabla_x \log(q(x,\theta))\|^2\right].$$

Normalizing Flows (NFs)

Parameter depending concatenation of diffeomorphisms of the form (up to some permutation matrices)

$$\mathcal{T}(\cdot;\theta) = \mathcal{T}_T \circ \cdots \circ \mathcal{T}_1$$

Invertibility by special structure of $\mathcal{T}_k$, $k = 1, \ldots, T$:

1. Real NVP (real-valued volume-preserving transformations) (Dingh et al. 2017)

$$\mathcal{T}_k(z_1, z_2) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} := \begin{pmatrix} z_1 \\ z_2\, \mathrm{e}^{s_k(z_1)} + t_k(z_1) \end{pmatrix}$$

$$\mathcal{T}_k^{-1}(x_1, x_2) = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ (x_2 - t_k(x_1))\, \mathrm{e}^{-s_k(x_1)} \end{pmatrix}$$

with neural networks s_k, t_k, $k = 1, \ldots, T$, $x_j, z_j \in \mathbb{R}^{n_j}$, $j = 1, 2$.

2. Other architecture (Aridizzone et al. 2019):

$$\mathcal{T}_k(z_1, z_2) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} := \begin{pmatrix} z_1\, \mathrm{e}^{s_{k,2}(z_2)} + t_{k,2}(z_2) \\ z_2\, \mathrm{e}^{s_{k,1}(x_1)} + t_{k,1}(x_1) \end{pmatrix}$$

$$\mathcal{T}_k^{-1}(x_1, x_2) = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} (x_1 - t_{k,2}(z_2))\, \mathrm{e}^{-s_{k,2}(z_2)}, \\ (x_2 - t_{k,1}(x_1))\, \mathrm{e}^{-s_{k,1}(x_1)} \end{pmatrix}$$

Since $T_k = T_{2,k} \circ T_{1,k}$ with

$$T_{1,k}(z_1, z_2) = \begin{pmatrix} x_1 \\ z_2 \end{pmatrix} := \begin{pmatrix} z_1 \mathrm{e}^{s_{k,2}(z_2)} + t_{k,2}(z_2) \\ z_2 \end{pmatrix},$$

$$T_{2,k}(x_1, z_2) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} := \begin{pmatrix} x_1 \\ z_2 \mathrm{e}^{s_{k,1}(x_1)} + t_{k,1}(x_1) \end{pmatrix},$$

we have

$$\nabla T_{1,k}(z_1, z_2) = \begin{pmatrix} \operatorname{diag}\left(\mathrm{e}^{s_{k,2}(z_2)}\right) & \operatorname{diag}\left(\nabla_{z_2}\left(z_1 \mathrm{e}^{s_{k,2}(z_2)} + t_{k,2}(z_2)\right)\right) \\ 0 & I_{d_2}, \end{pmatrix}$$

so that $\det \nabla T_{1,k}(z_1, z_2) = \prod_{k=1}^{d_1} \mathrm{e}^{\left(s_{k,2}(z_2)\right)_k}$ and similarly for $\nabla T_{2,k}$. Applying the chain rule and noting that $\det(AB) = \det(A)\det(B)$,

$$\log(|\det(\nabla \mathcal{T}(z))|) = \sum_{k=1}^{T} \left(\operatorname{sum}\left(s_{k,2}\left((z^k)_2\right)\right) + \operatorname{sum}\left(s_{k,1}\left((T_{1,k} z^k)_1\right)\right)\right),$$

where sum denotes the sum of the components of the respective vector, $z^1 := z$ and $z^k = T_{k-1} z^{k-1}$, $k = 2, \ldots, T$.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58

Aim: $P_X \approx \mathcal{T}_\# P_Z := P_Z \circ \mathcal{T}^{-1}$

Loss Function: Kullback-Leibler divergence (KL)

$$
\begin{aligned}
\mathcal{L}_{\mathrm{NF}}(\theta) = \mathrm{KL}(P_X, \mathcal{T}_\# P_Z) &= \int \log \left(\frac{p_X(x)}{p_{\mathcal{T}_\# P_Z}(x)} \right) p_X(x) \mathrm{d}x \\
&= \underbrace{\int \log \left(p_X(x)\right) p_X(x) \mathrm{d}x}_{\text{const}} - \int \log \left(p_{\mathcal{T}_\# P_Z}(x) \right) p_X(x) \mathrm{d}x
\end{aligned}
$$

with transformation formula $p_{\mathcal{T}_\# P_Z} = p_Z\left(\mathcal{T}^{-1}\right) |\det \nabla \mathcal{T}^{-1}|$ and Gaussian distribution P_Z

$$
\begin{aligned}
\mathcal{L}_{\mathrm{NF}}(\theta) &\sim - \int \log \left(p_Z\left(\mathcal{T}^{-1}(x)\right) \, |\det \nabla \mathcal{T}^{-1}(x)| \right) p_X(x) \mathrm{d}x \\
&= -\mathbb{E}_{x \sim P_X}[\log p_Z\left(\mathcal{T}^{-1}\right)] - \mathbb{E}_{x \sim P_X}[\log \left(|\det \nabla \mathcal{T}^{-1}|\right)] \\
&= \|\mathcal{T}^{-1}(\cdot)\|^2_{L_2(dP_X)} - \mathbb{E}_{x \sim P_X}[\left(\log |\det \nabla \mathcal{T}^{-1}|\right)]
\end{aligned}
$$

Refs: SGD (Optimizer Adam: Kingma et al. 2015), Inertial Stoch. PALM (Hertrich/St. 2022)

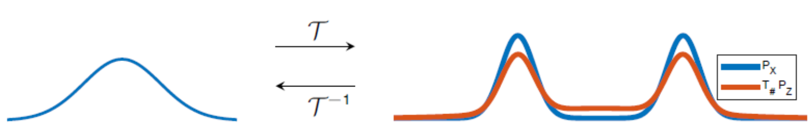
Remark on Kullback-Leibler Divergence

KL is (only) the Bregman distance of the Shannon entropy:

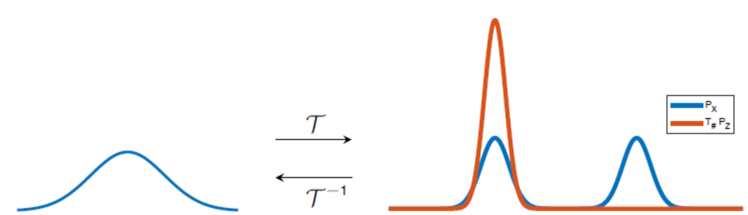
- $\mathrm{KL}(\mu,\nu) \geq 0$ for all $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$ with equality $\Leftrightarrow$ if $\nu = \mu$
- not symmetric and no triangular inequality
- finite if $\mu \ll \nu$ and $\mathrm{KL}(\mu,\nu) = +\infty$ otherwise

Different properties of $\mathrm{KL}(P_X, \mathcal{T}_\# P_Z)$ (forward KL) and $\mathrm{KL}(\mathcal{T}_\# P_Z, P_X)$ (backward KL):

- inverse problems: operator known versus operator not known
- mode covering (unrealistic samples possible) versus mode seeking (mode collapse)



$\mathrm{KL}(P_X, \mathcal{T}_\# P_Z)$



$\mathrm{KL}(\mathcal{T}_\# P_Z, P_X)$

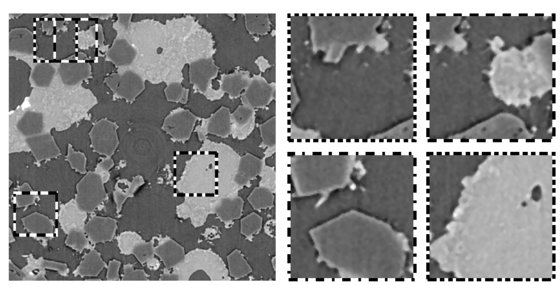
Refs: Backward KL: Kruse et al. 2020, Sun/Bouman 2021, Altekrüger/Hertrich: WPPFlows 2022, Inverse Problem for a certain class of images:

$$y = \mathrm{noisy}(F(x))$$

Find solution as minimizer of variational model:

$$\mathcal{J}(x) = \underbrace{\mathcal{D}_F(x,y)}_{\text{data term}} + \lambda \underbrace{\mathcal{R}(x)}_{\text{regularizer}}, \qquad \lambda > 0$$

Idea: Learn regularizer from many patches $P_i(x_j)$, $i = 1, \ldots, N$ of few images x_j, $j = 1, \ldots, n$



The Power of Patches

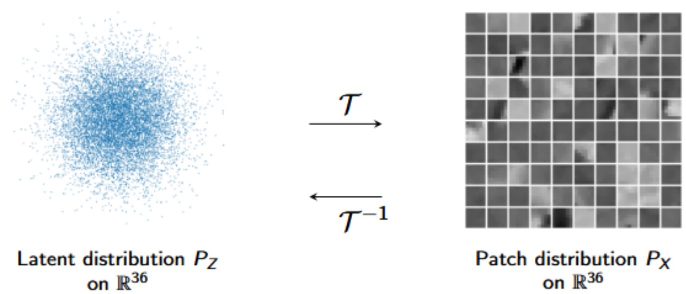
Refs: Buardes et al. 2005, Dabov et al. (BM3D) 2008; Lebrun/Morel (Denoising cuisine) 2013; Bortoli/Desolneux/Galerne/Leclaire 2019 ...,

Laus et al. 2017, Houdard et al. 2018; Hertrich et al. 2021 ...

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58

1. Learn Normalizing Flow $\mathcal{T} = \mathcal{T}(\theta, \cdot)$

$$\mathcal{L}_{\mathrm{NF}}(\theta) = \sum_{j=1}^{n} \sum_{i=1}^{N} \frac{\|\mathcal{T}^{-1}(P_i(x_j))\|^2}{2} - \log|\det \nabla \mathcal{T}^{-1}(P_i(x_j))|$$



2. Variational Model

$$\mathcal{J}(x) = \mathcal{D}_F(x, y) + \lambda\, \mathrm{patchNR}(x)$$

negative log likelihood of all patches under the probability distribution learned by the patchNR $\mathcal{T} = \mathcal{T}(\theta, \cdot)$

$$\mathrm{patchNR}(x) := \Big(\frac{1}{N} \sum_{i=1}^{N} \frac{\|\mathcal{T}^{-1}(P_i(x))\|^2}{2} - \log|\det \nabla \mathcal{T}^{-1}(P_i(x))|\Big)$$

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58

Bayesian Approach - MAP:

$$\begin{aligned}\hat{x} &\in \operatorname*{argmax}_x \log\left(p_{X|Y=y}(x)\right)\\ &= \operatorname*{argmin}_x \left\{ -\log\left(p_{Y|X=x}(y)\right) - \log\left(p_X(x)\right) \right\}\\ &= \operatorname*{argmin}_x \Big\{ \underbrace{\mathcal{D}_F(x,y)}_{\text{data-fidelity term}} + \lambda \underbrace{\mathcal{R}(x)}_{\text{regularizer}} \Big\}\end{aligned}$$

Proposition:

Let $P_Z = \mathcal{N}(0, I)$ and let $\mathcal{T}\colon \mathbb{R}^s \to \mathbb{R}^s$ be a bi-Lipschitz diffeomorphism. Then

$$\exp\left(-\lambda\, \mathrm{patchNR}(x)\right) \in L^1(\mathbb{R}^d), \qquad \lambda > 0.$$

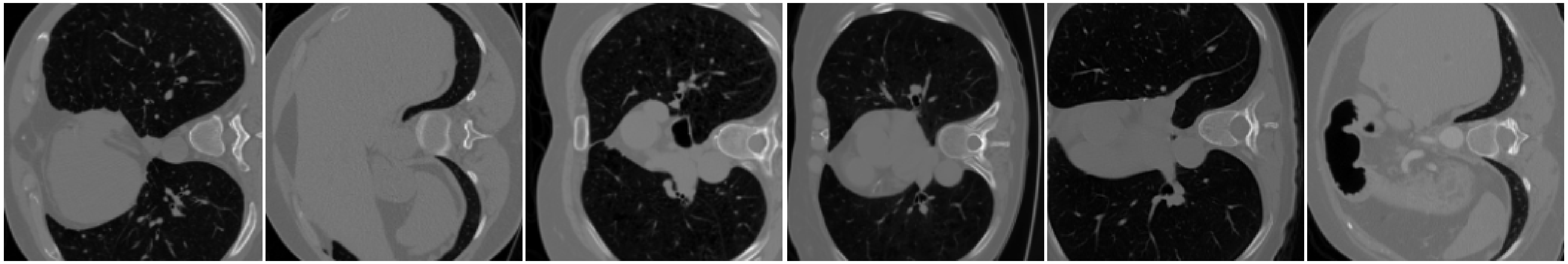
Advantages:

- Few training data: often not many training images are available - need just patches! Save electrical energy.
- Flexibility of operator and image classes: regularizer fits for every operator in the corresponding image classes

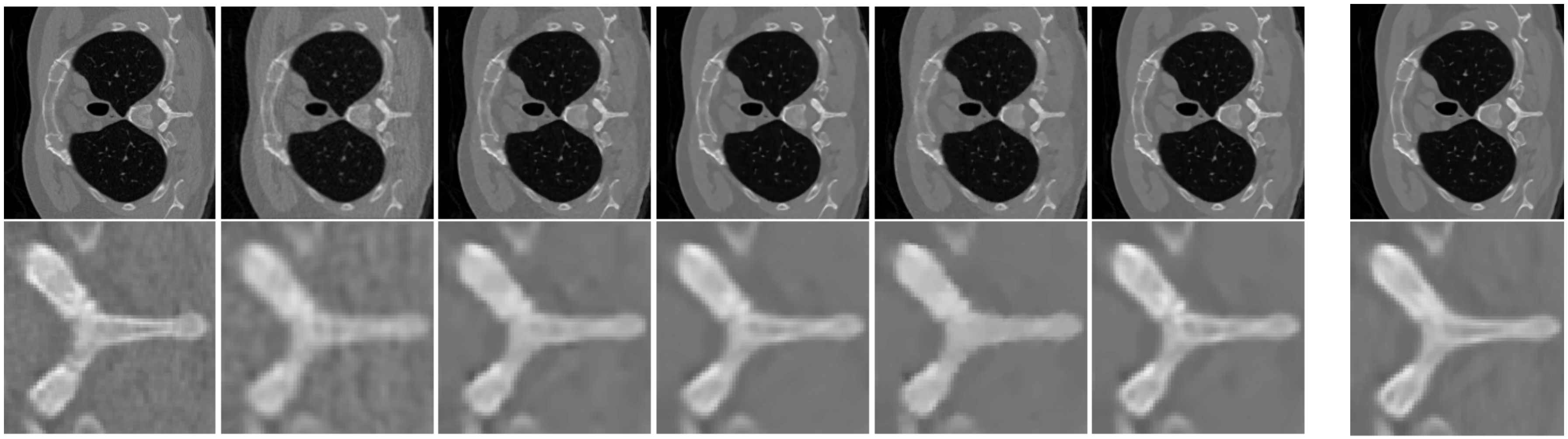
Data term: Poisson (like) noise

$$\mathcal{D}(Ax,y)=\sum_{i=1}^{d} e^{-(Ax)_i}N_0 - e^{-y_i}N_0\big(-(Ax)_i+\log(N_0)\big)$$

Regularizer: $\mathcal{R}(x)=\mathrm{patchNR}(x)$ learned from $n=6$ images



Results: Low Dose Tomography (full angle)

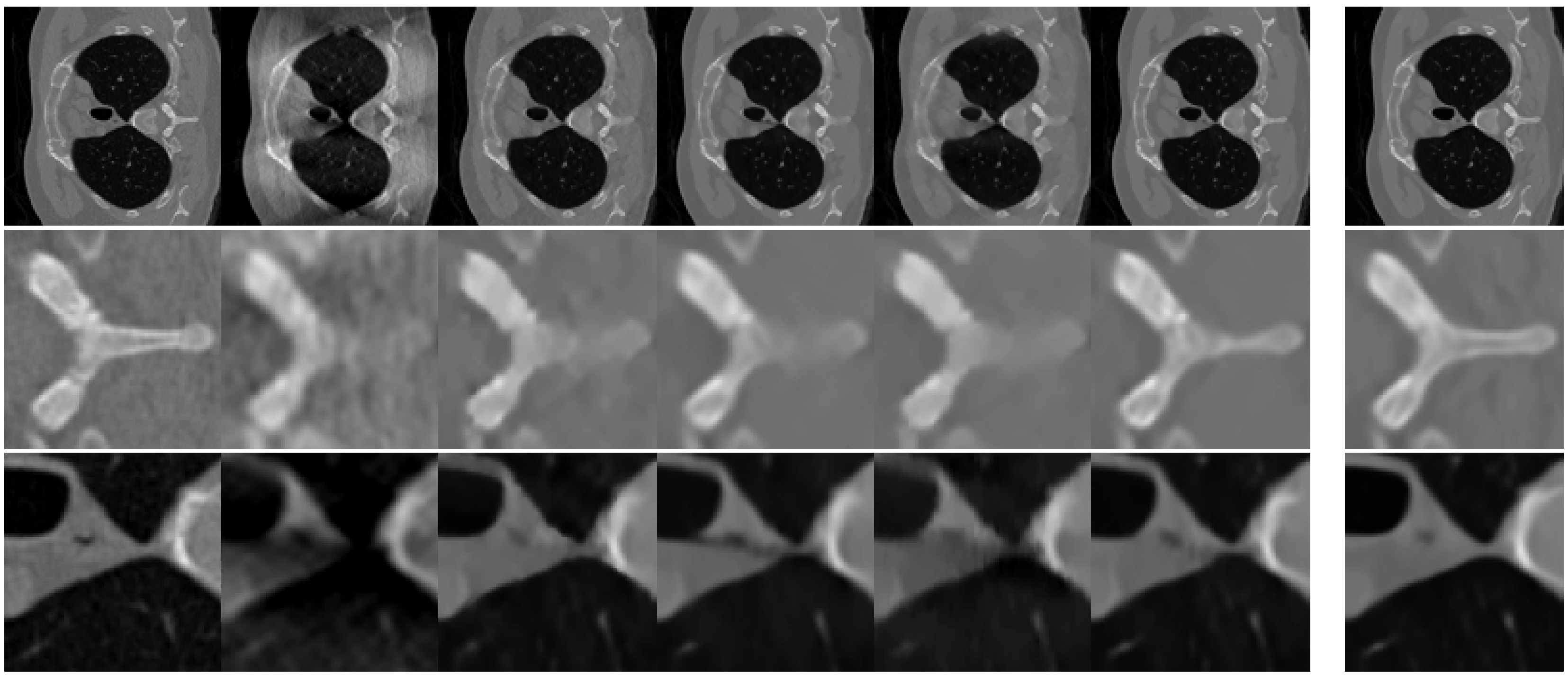


Ground truth | FBP | DIP+TV | EPLL | localAR | patchNR | FBP+UNet

	FBP	DIP + TV	EPLL	localAR	patchNR	FBP+UNet (data-based)
PSNR	30.37 ± 2.95	34.45 ± 4.20	34.89 ± 4.41	33.64 ± 3.74	**35.19** ± 4.52	35.48 ± 4.52
SSIM	0.739 ± 0.141	0.821 ± 0.147	0.821 ± 0.154	0.807 ± 0.145	**0.829** ± 0.152	0.837 ± 0.143
Runtime	0.03s	1514.33s	36.65s	30.03s	48.39s	0.46s

Refs: DIP-TV: Ulyanov et al. 2018, EPLL: Zoran et al. 2011, AR: Lunz et al. 2018, LocalAR: Prost et al. 2021, FBP+UNet: Jin et al. 2017, 35820 supervised samples

Results: Limited Angles - $36°/180°$ (same $\mathcal{R}$ for Low Dose Tomography!)



Ground truth FBP DIP+TV EPLL localAR patchNR FBP+UNet (s)

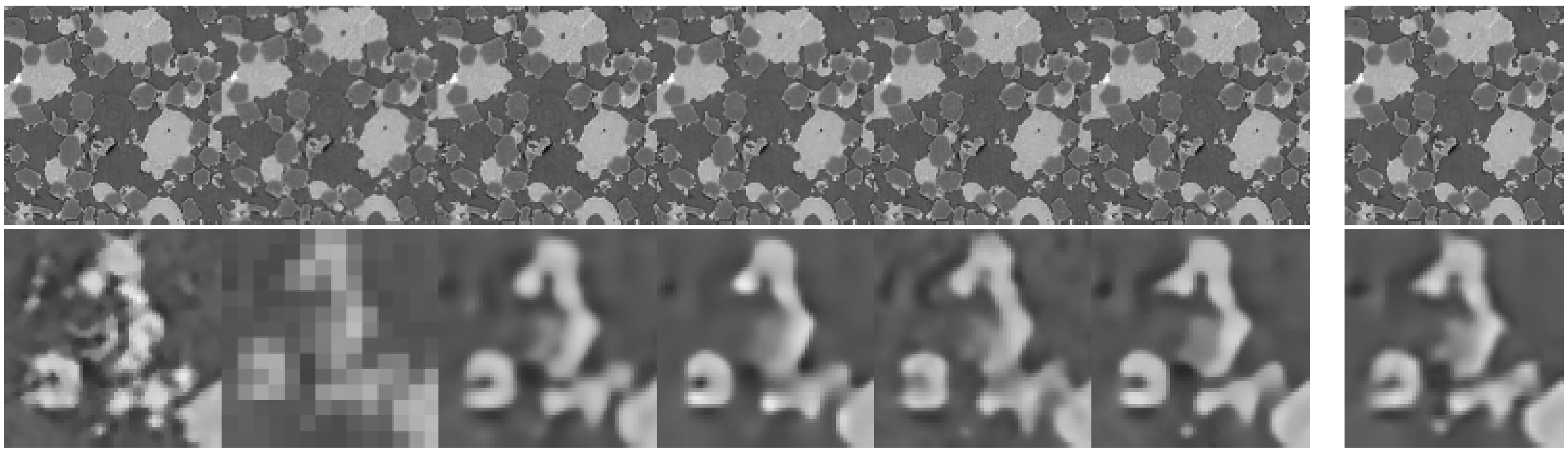
	FBP	DIP + TV	EPLL	localAR	patchNR	FBP+UNet (data-based)
PSNR	21.96 ± 2.25	32.57 ± 3.25	32.78 ± 3.46	31.06 ± 2.95	$\mathbf{33.20} \pm 3.55$	33.75 ± 3.58
SSIM	0.531 ± 0.097	0.803 ± 0.146	0.801 ± 0.151	0.779 ± 0.142	$\mathbf{0.811} \pm 0.151$	0.820 ± 0.140
Runtime	0.02s	1770.89s	127.21s	53.47s	485.93s	0.53s

Superresolution

Data term: Gaussian noise $\mathcal{D}(Ax, y) = \frac{1}{2}\|Ax - y\|^2$ stride: 4

Regularizer: learned from 1 high resolution synchotron image

Results:



	bicubic (not shown)	DPIR (not shown)	DIP+TV	EPLL	WPP	patchNR	ACNN (data-based)
PSNR	25.63 ± 0.56	27.78 ± 0.53	27.99 ± 0.54	28.11 ± 0.55	27.80 ± 0.37	**28.53** ± 0.49	28.89 ± 0.53
SSIM	0.699 ± 0.012	0.770 ± 0.011	0.764 ± 0.007	0.779 ± 0.010	0.749 ± 0.011	**0.780** ± 0.008	0.804 ± 0.010
Runtime	0.0002s	56.62s	234.00s	60.28s	387.28s	150.79s	0.03s

Data: D. Bernard (U Bordeaux), Y. Berthoumieu, JF Aujol (ANR-DFG project)

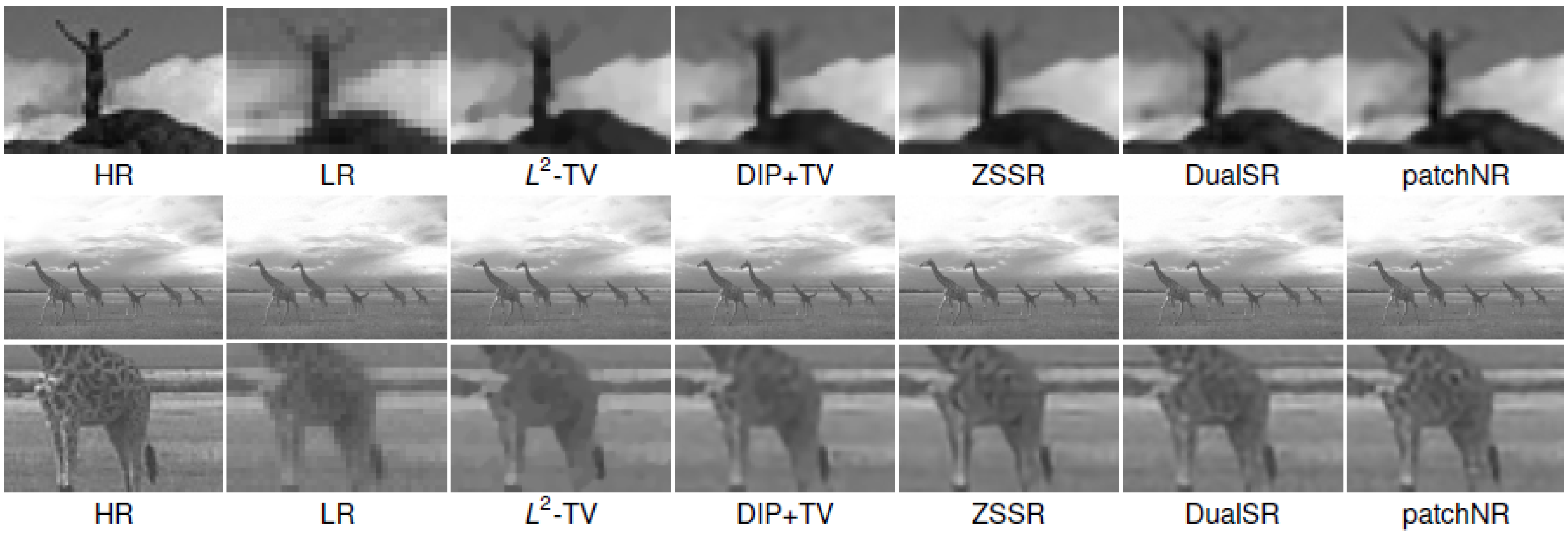
Refs: WPP: Hertrich et al. 2021, ACNN: Tian et al. 2021

Single-Shot Superresolution

Question: What to do when we do not have any high-resolution image for training the patchNR?

Assumption: Patch distribution of natural images is self-similar across the scales

Regularizer: learned from 1 low resolution image

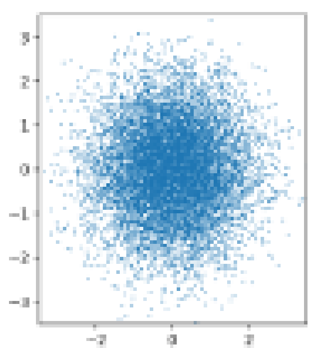


	L^2-TV	DIP+TV	ZSSR	DualSR	patchNR
PSNR	28.35 ± 3.55	28.44 ± 3.69	28.83 ± 3.57	28.64 ± 3.47	**29.08** ± 3.58
SSIM	0.820 ± 0.072	0.821 ± 0.087	0.834 ± 0.066	0.829 ± 0.061	**0.846** ± 0.061
Runtime	13.12s	171.51s	56.64s	53.47s	132.36s

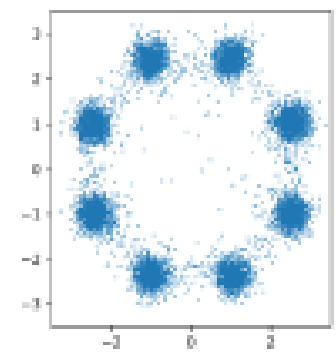
Refs: Glasner et al. 2009, ZSSR: Shocher et al. 2018 , Dual SR: Emad et al. 2021

Drawback of NF

- **Drawback: Lack in expressiveness!**: unimodal distributions are hard to map to multimodal and heavy tailed ones
- Normalizing flows mapping unimodal distributions to multimodal ones must have an exploding Lipschitz constant!
  (Refs Lipschitz: Nagarajan et al. 2018, Gulrajani et al. 2018, Hagemann/Neumayer 2021; Stéphanovitch et al. 2022; Salmona et al. 2022)



$$\xrightarrow[\mathcal{T}]{}$$



- (One) **Solution:** Application of stochastic steps within a unifying framework of Markov chains
  (Refs Stochastic NFs without Markov chains: Wu/Köhler/Noé 2020, Nielsen/Welling et al. 2021)

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58

Technische Universität Berlin

Lecture 4: Generalized Normalizing Flows

Gabriele Steidl
Applied Mathematics - Imaging Sciences
TU Berlin

Berlin Mathematics Research Center MATH+

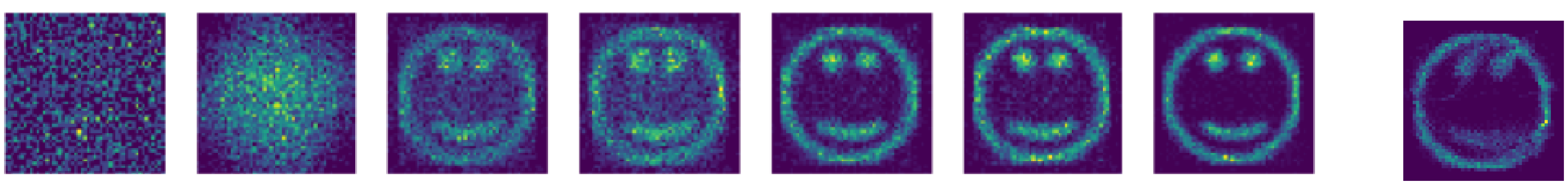
Funded under Germany's Excellence Strategy by DFG Deutsche Forschungsgemeinschaft

Generalized Normalizing Flows

Markov chain framework includes

- Normalizing flows
- Metropolis-Hastings (MH, MCMC) layer
- Langevin layer
- VAE layer
- Diffusion flow layer

Up to now the only mathematically sound way to combine these layers!



Alternation of INNs with MCMC layers, starting with the first one.

Technische Universität Berlin
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58

Outline

Markov Kernels and Markov Chains

- A Markov kernel $\mathcal{K}\colon \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0,1]$ is a mapping such that
  - i) $\mathcal{K}(\cdot, B)$ is measurable for any $B \in \mathcal{B}(\mathbb{R}^d)$, and
  - ii) $\mathcal{K}(x, \cdot)$ is a probability measure for any $x \in \mathbb{R}^d$.
- For μ on $\mathcal{P}(\mathbb{R}^n)$, the measure $\mu \times \mathcal{K}$ on $\mathbb{R}^n \times \mathbb{R}^d$ is defined by

$$(\mu \times \mathcal{K})(A \times B) := \int_A \mathcal{K}(x, B) d\mu(x).$$

  Definition captures all sets in $\mathcal{B}(\mathbb{R}^n \times \mathbb{R}^d)$ since the measurable rectangles form a $\cap$-stable generator of $\mathcal{B}(\mathbb{R}^n \times \mathbb{R}^d)$.
- For all integrable f that

$$\int_{\mathbb{R}^n \times \mathbb{R}^d} f(x,y) \mathrm{d}(\mu \times \mathcal{K})(x,y) = \int_{\mathbb{R}^n} \int_{\mathbb{R}^d} f(x,y) \mathrm{d}\mathcal{K}(x,\cdot)(y) \mathrm{d}\mu(x).$$

- Regular conditional distribution of X given Y: P_Y-a.s. unique Markov kernel $P_{Y|X=\cdot}(\cdot)$ with

$$P_X \times \underbrace{P_{Y|X=\cdot}(\cdot)}_{\mathcal{K}(\cdot,\cdot)} = P_{(X,Y)}$$

$$\int_{\mathbb{R}^n \times \mathbb{R}^d} f(x,y) \mathrm{d}P_{(X,Y)}(x,y) = \int_{\mathbb{R}^n} \int_{\mathbb{R}^d} f(x,y) \mathrm{d}P_{Y|X=x}(y) \mathrm{d}P_X(x).$$

- $(X_0, \dots, X_T)$ is called a Markov chain, if there exist Markov kernels

$$\mathcal{K}_t = P_{X_t | X_{t-1} = \cdot}(\cdot) \colon \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0,1]$$

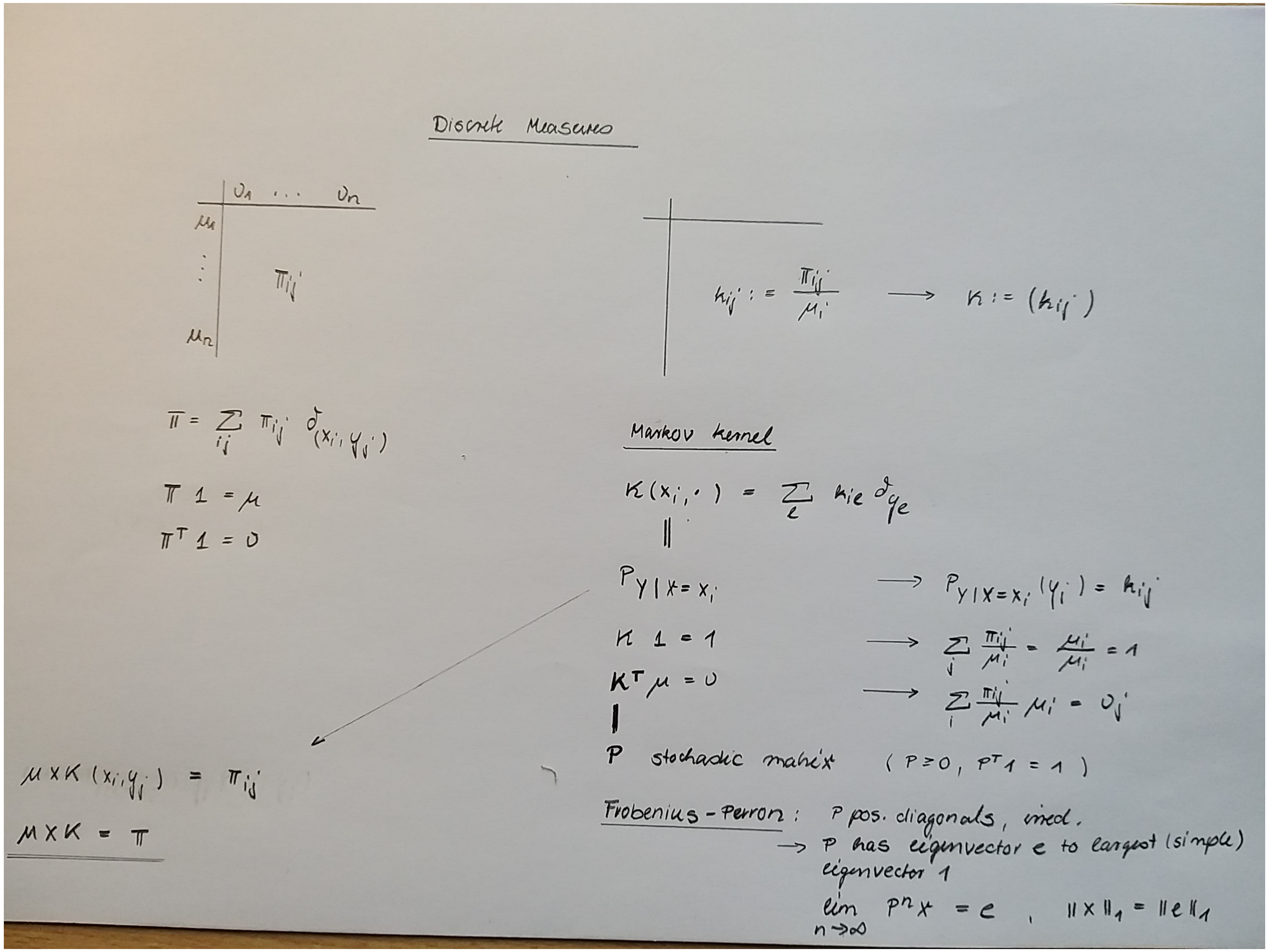
such that

$$P_{(X_0,\dots,X_T)} = P_{X_0} \times \mathcal{K}_1 \times \dots \times \mathcal{K}_T.$$

- Markov kernels $\mathcal{K}_t$ are called transition kernels
- If $\mathcal{K}_t(x,\cdot) = P_{X_t|X_{t-1}=x}$ has a density $k_t(x,y)$, and $P_{X_{t-1}}$ resp. P_{X_t} have densities $p_{X_{t-1}}$ resp. p_{X_t}, then

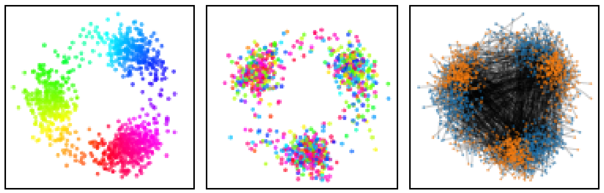
$$p_{X_t}(y) = \int_{\mathbb{R}^{d_{t-1}}} k_t(x,y) p_{X_{t-1}}(x) \mathrm{d}x.$$

Example in OT



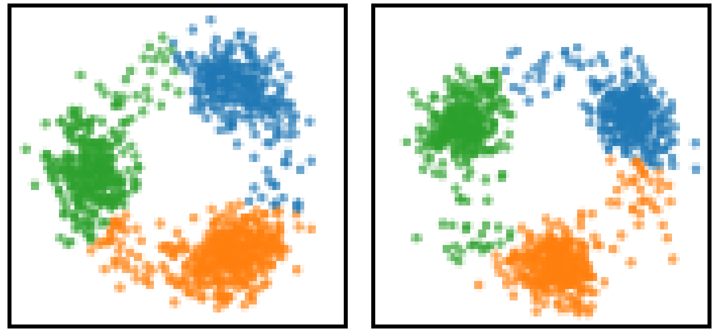
Disintegration theorem in OT (book Ambrosio et al:Thm 5.3.1): existence of a Markov kernel $\mathcal{K}$ (uniquely defined for a.e. $x \in X$) such that $\pi = \mu \times \mathcal{K}$ Usual notation $K(x, \cdot) = \pi_x$

Remark on Transfer Operators

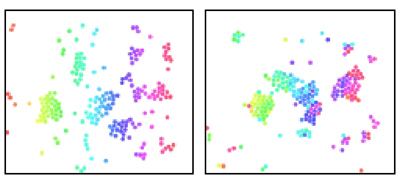


(a) Data at time $t = 0$. (b) Data at time $t = 1$. (c) Movement of particles.

Figure 7: Particles moving in a potential with circular driving force. The color scheme illustrates the particle mixing.

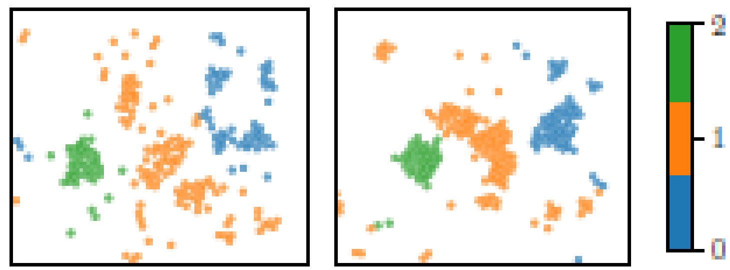


(a) Hard, $t = 0$. (b) Hard, $t = 1$.



(a) Data at time $t = 0$. (b) Data at time $t = 1$.

Figure 10: Particles moving in a potential with circular driving force. Again, the color scheme illustrates the particle mixing.



(a) Hard, $t = 0$. (b) Hard, $t = 1$.

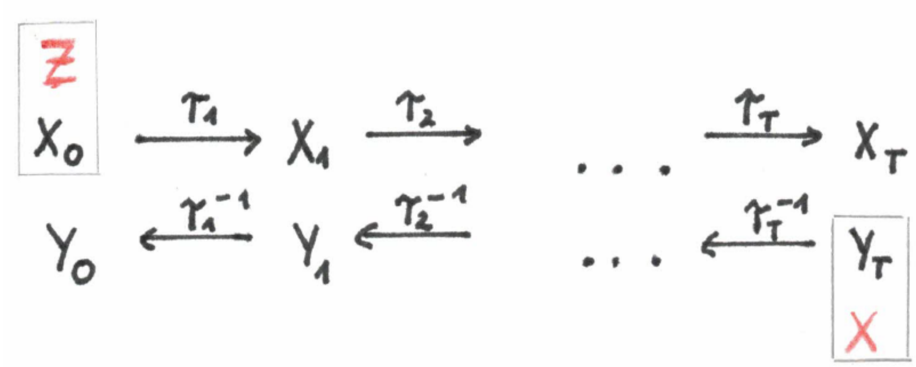
Refs: 1. P Koltai, J von Lindheim, S Neumayer, G S: Transfer operators from optimal transport plans for coherent set detection Physica D, 2021

2. F. Beier: Gromov–Wasserstein Transfer Operators

3. Junge, O., Matthes, D., Schmitzer, B.: Entropic transfer operators. 2022 (exact transfer operator is known on a finite subset of the full state space. Then, using regularized OT, a finite-dimensional approximation is constructed which limit is a regularized version of the ground truth and exhibits desirable properties, such as retention of the spectral information)

Outline

Normalizing Flow: $\mathcal{T}(\cdot;\theta) = \mathcal{T}_T \circ \cdots \circ \mathcal{T}_1$,



- $\mathcal{T}_t\colon \mathbb{R}^d \to \mathbb{R}^d$ generate a pair of Markov chains $\big((X_0, \dots, X_T), (Y_T, \dots, Y_0)\big)$

$$X_0 \sim P_Z, \quad X_t = \mathcal{T}_t(X_{t-1}) \quad \text{and}$$
$$Y_T \sim P_X, \quad Y_{t-1} = \mathcal{T}_t^{-1}(Y_t).$$

- Markov kernels $\mathcal{K}_t(x,\cdot) = P_{X_t|X_{t-1}=x} = \delta_{\mathcal{T}_t(x)}$ and $\mathcal{R}_t(x,\cdot) = \delta_{\mathcal{T}_t^{-1}(x)}$
- Minimize the ,,whole path"

$$\mathcal{L}_{\mathrm{NF}}(\theta) = \mathrm{KL}(P_X, P_{\mathcal{T}_\# Z}) = \mathrm{KL}(P_{Y_T}, P_{X_T})$$
$$= \mathrm{KL}\left(P_{(Y_0,\dots,Y_T)}, P_{(X_0,\dots,X_T)}\right) \qquad \text{in general } \leq$$

Deterministic Markov kernels:

$$\mathcal{K}_t(x,\cdot) = P_{X_t|X_{t-1}=x} = \delta_{\mathcal{T}_t(x)}$$
$$\mathcal{R}_t(y,\cdot) = P_{Y_{t-1}|Y_t=y} = \delta_{\mathcal{T}_t^{-1}(y)}$$

Proof

$$\begin{aligned}
P_{(X_{t-1},X_t)}(A\times B) &= \int_{\mathbb{R}^d\times\mathbb{R}^d} 1_{A\times B}(x_{t-1},x_t)\mathrm{d}P_{(X_{t-1},X_t)}(x_{t-1},x_t)\\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d} 1_A(x_{t-1})1_B(x_t)\mathrm{d}P_{(X_{t-1},X_t)}(x_{t-1},x_t).
\end{aligned}$$

Since $P_{(X_{t-1},X_t)}$ is by definition concentrated on the set $\{(x_{t-1},\mathcal{T}_t(x_{t-1})) : x_{t-1}\in\mathbb{R}^d\}$, this becomes

$$\begin{aligned}
P_{(X_{t-1},X_t)}(A\times B) &= \int_{\mathbb{R}^d\times\mathbb{R}^d} 1_A(x_{t-1})1_B(\mathcal{T}_t(x_{t-1}))\mathrm{d}P_{(X_{t-1},X_t)}(x_{t-1},x_t)\\
&= \int_A 1_B(\mathcal{T}_t(x_{t-1}))\mathrm{d}P_{X_{t-1}}(x_{t-1})\\
&= \int_A \delta_{\mathcal{T}_t(x_{t-1})}(B)\mathrm{d}P_{X_{t-1}}.
\end{aligned}$$

Consequently, the transition kernel $\mathcal{K}_t = P_{X_t|X_{t-1}}$ is given by $\mathcal{K}_t(x,\cdot) = \delta_{\mathcal{T}_t(x)}$. □

Outline

1. Markov Kernels and Markov Chains
2. Normalizing Flows via Markov Chains
3. Generalized Normalizing Flows (GNFs)
4. Conditional GNFs

Generalized/Stochastic normalizing flow (SNF) = pair of Markov chains

$$((X_0, \dots, X_T), (Y_T, \dots, Y_0))$$

that minimizes the loss function

$$\mathcal{L}_{\mathsf{SNF}}(\theta) = \mathrm{KL}(P_{(Y_0,\dots,Y_T)}, P_{(X_0,\dots,X_T)}) :$$

P1) P_{X_t}, P_{Y_t} have the densities $p_{X_t}, p_{Y_t} \colon \mathbb{R}^{d_t} \to \mathbb{R}_{>0}$ for any $t = 0, \dots, T$.

P2) There exist Markov kernels $\mathcal{K}_t = P_{X_t|X_{t-1}}$ and $\mathcal{R}_t = P_{Y_{t-1}|Y_t}$, $t = 1, \dots, T$:

$$P_{(X_0,\dots,X_T)} = P_{X_0} \times P_{X_1|X_0} \times \cdots \times P_{X_T|X_{T-1}},$$
$$P_{(Y_T,\dots,Y_0)} = P_{Y_T} \times P_{Y_{T-1}|Y_T} \times \cdots \times P_{Y_0|Y_1}.$$

P3) For P_{X_t}-almost every $x \in \mathbb{R}^{d_t}$, the measures $P_{Y_{t-1}|Y_t=x}$ and $P_{X_{t-1}|X_t=x}$ are absolutely continuous with respect to each other

Important: Conditional distributions $P_{X_t|X_{t-1}}$ themselves must not be absolutely continuous! Indeed not the case for NFs and MCMC layers.

Avoid: UMOs ! e.g. $\frac{\delta(x_t - T_t(x_{t-1}))}{\delta(x_{t-1} - T_t^{-1}(x_t))}$

Loss function: $\mathcal{L}_{\mathsf{SNF}}(\theta) = \mathrm{KL}(P_{(Y_0,\ldots,Y_T)}, P_{(X_0,\ldots,X_T)})$

Theorem: With Radon-Nikodym derivative $f_t(\cdot, x_t) = \frac{\mathrm{d}P_{Y_{t-1}|Y_t=x_t}}{\mathrm{d}P_{X_{t-1}|X_t=x_t}}$ we have

$$\mathcal{L}_{\mathsf{SNF}}(\theta) = \mathbb{E}_{(x_0,\ldots,x_T)\sim P_{(Y_0,\ldots,Y_T)}}\Big[\log\Big(\frac{p_X(x_T)}{p_Z(x_0)}\prod_{t=1}^T \frac{f_t(x_{t-1},x_t)p_{X_{t-1}(x_{t-1})}}{p_{X_t(x_t)}}\Big)\Big]$$

$$= \mathbb{E}_{(x_0,\ldots,x_T)\sim P_{(Y_0,\ldots,Y_T)}}\Big[\log\Big(\frac{p_X(x_T)}{p_{X_T}(x_T)}\prod_{t=1}^T f_t(x_{t-1},x_t)\Big)\Big]$$

The right-hand side can be computed for the different layers:

- NF layer: $\frac{p_{X_{t-1}}(x_{t-1})}{p_{X_t}(x_t)} = \frac{1}{|\nabla \mathcal{T}_t{}^{-1}(x_t)|}$ and $f_t(x_{t-1},x_t) = 1$
- MCMC layer: $\frac{f_t(x_{t-1},x_t)p_{X_{t-1}}(x_{t-1})}{p_{X_t}(x_t)} = \frac{p_t(x_{t-1})}{p_t(x_t)}$
- Langevin layer: $\frac{f_t(x_{t-1},x_t)p_{X_{t-1}}(x_{t-1})}{p_{X_t}(x_t)} = \exp\Big(\frac{1}{2}(\|\eta_t\|^2 - \|\tilde\eta_t\|^2)\Big)$,
  with proposal density $p_t\colon \mathbb{R}^d \to \mathbb{R}_{>0}$, $u_t = -\log(p_t)$ and
  $\eta_t := \frac{1}{a_2}(x_{t-1} - x_t - a_1\nabla u_t(x_{t-1}))$, $\tilde\eta_t := \frac{1}{a_2}(x_{t-1} - x_t + a_1\nabla u_t(x_t))$
- Diffusion layer: $\frac{f_t(x_{t-1},x_t)p_{X_{t-1}}(x_{t-1})}{p_{X_t}(x_t)} = \exp\Big(\frac{1}{2}(\|\eta_t\|^2 - \|\tilde\eta_t\|^2)\Big)$,, where
  $\eta_t := \frac{1}{\sqrt{\epsilon}h_{t-1}}(x_{t-1} - x_t + \epsilon g_{t-1}(x_{t-1}))$, $\tilde\eta_t := \frac{1}{\sqrt{\epsilon}g_t}(x_{t-1} - x_t - \epsilon(g_t(x_t) - h_t{}^2 s_t(x_t)))$
- VAEs: $\frac{f_t(x_{t-1},x_t)p_{X_{t-1}}(x_{t-1})}{p_{X_t}(x_t)} = \frac{q_\phi(x_{t-1}|x_t)}{p_\theta(x_t|x_{t-1})}$

Technische Universität Berlin
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58

Learning SNFs

• $\int_{A\times B} f(x,y)\, dP_{(X,Y)} = \int_A \int_B f(x,y)\, dP_Y(y)\, dP_{X|Y=y}(x)$

① $f := \dfrac{dP_{(Y_0,\dots,Y_T)}}{dP_{(X_0,\dots,X_T)}} \overset{!}{=} \dfrac{p_{Y_T}(x_T)}{p_{X_T}(x_T)} \prod_{t=1}^{T} \underbrace{f_t(x_{t-1}, x_t)}_{\frac{dP_{Y_{t-1}|Y_t=x_t}}{dP_{X_{t-1}|X_t=x_t}}}$

② $KL(\mu, \nu) = \int \log \underbrace{\frac{d\mu}{d\nu}}_{\text{Radon-Nikodym derivative}} d\mu(x) = \mathbb{E}_{x\sim\mu}\left[\log \frac{d\mu}{d\nu}\right]$

$KL(P_{(Y_0,\dots,Y_T)}, P_{(X_0,\dots,X_T)}) = \mathbb{E}_{(x_0,\dots,x_T)\sim P_{(Y_0,\dots,Y_T)}}\left[\log f(x_0,\dots,x_T)\right]$

• ① can be shown by induction using •

Case $T = 1$: f rhs in ①

To show: $f\, dP_{(X_0,X_1)} = dP_{(Y_0,Y_1)}$

$\int_{A\times B} \underbrace{\frac{p_{Y_1}(x_1)}{p_{X_1}(x_1)}}\ \underbrace{\frac{dP_{Y_0|Y_1=x_1}}{dP_{X_0|X_1=x_1}}}\ dP_{(X_0,X_1)}(x_0,x_1)$

$= \int_A \int_B$ — " — $\underbrace{dP_{X_1}(x_1)}\ \underbrace{dP_{X_0|X_1=x_1}(x_0)}$

$= \int_A \int_B dP_{Y_1}(x_1)\, dP_{Y_0|Y_1=x_1}(x_0)$

$= \int_{A\times B} dP_{(Y_0,Y_1)}(x_0,x_1)$

- Markov chain:

$$X_t = X_{t-1} + 1_{[U,1]}\left(\alpha_t(X_{t-1}, X_{t-1} + \xi_t)\right) \xi_t, \quad \xi_t \sim \mathcal{N}(0, \sigma^2 I)$$

  where

$$\alpha_t(x,y) := \min\left\{1, \frac{p_t(y)}{p_t(x)}\right\}$$

  with proposal density $p_t \colon \mathbb{R}^d \to \mathbb{R}_{>0}$

- Markov kernels $\mathcal{K}_t = \mathcal{R}_t \colon \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0,1]$

$$\begin{aligned}\mathcal{K}_t(x, A) := &\int_A \mathcal{N}(y; x, \sigma^2 I)\alpha_t(x,y)\mathrm{d}y \\ &+ \delta_x(A) \int_{\mathbb{R}^d} \mathcal{N}(y; x, \sigma^2 I)\left(1 - \alpha_t(x,y)\right)\mathrm{d}y\end{aligned}$$

- one sampling step of Metropolis-Hastings algorithm (known to sample from the proposal distribution p_t)

Metropolis-Hastings Algorithm

Input: $x_0 \in \mathbb{R}^d$, proposal density $p_t \colon \mathbb{R}^d \to \mathbb{R}_{>0}$
For $k = 0, 1, \ldots$ do
Draw x' from $\mathcal{N}(x_k, \sigma^2 I)$ and u uniformly in $[0, 1]$.
Compute the acceptance ratio

$$\alpha(x_k, x') := \min\left\{1, \frac{p(x')}{p(x_k)}\right\}.$$

Set

$$x_{k+1} := \begin{cases} x' & \text{if } u < \alpha(x_k, x'), \\ x_k & \text{otherwise.} \end{cases}$$

Output: $\{x_k\}_k$

VAE Layer

1. Autoencoders (AE): $\mathcal{L}_{\mathsf{AE}}(\phi,\theta) := \mathbb{E}_{x\sim P_X}\left(\|x - D_\theta(E_\phi(x))\|^2\right)$

- encoder $E = E_\phi\colon \mathbb{R}^d \to \mathbb{R}^n$, $d > n$
- decoder $D = D_\theta\colon \mathbb{R}^n \to \mathbb{R}^d$

Example: Principle Component Analysis = affine encoder/decoder

Given $x_i \in \mathbb{R}^m$, $i = 1,\ldots,N$ (e.g. realizations of a random variable)

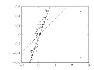
Find $A \in \mathrm{St}(m,d) \subset R^{m,d}$, $d \ll m$ i.e. $A^\mathsf{T} A = I$ and $b \in \mathbb{R}^m$ minimizing

$$\min_{A\in \mathrm{St}(m,d),b,t_i} \sum_{i=1}^N \|At_i + b - x_i\|^p, \quad p \in [1,\infty)$$

- Independent of p: $t_i = (A^\mathsf{T} A)^{-1} A^\mathsf{T}(x_i - b) = A^\mathsf{T}(x_i - b)$, $i = 1,\ldots,N$ Thus

$$\min_{A\in \mathrm{St}(m,d),b} \sum_{i=1}^N \|AA^\mathsf{T}(x_i - b) + b - x_i\|^p = \min_{A\in \mathrm{St}(m,d),b} \sum_{i=1}^N \|(I - AA^\mathsf{T})(b - x_i)\|^p$$

- $p = 1$: robust PCA (in particular computation of b not trivial ??
  Neumayer, Nimmer, Setzer, Steidl: On the robust PCA and Weiszfeld's algorithm + On the rotational invariant L1-norm PCA (2020)



$p = 2$:

$$\min_{A \in \mathrm{St}(m,d), b} \sum_{i=1}^{N} \| \underbrace{(I - AA^{\mathsf{T}})}_{P \text{ orth. proj.}} (b - x_i) \|^2$$

- Computation of b: $\sum_{i=1}^{N} P(b - x_i) = 0 \quad \Rightarrow \quad b = \sum_{i=1}^{N} \frac{1}{N} x_i + \mathcal{R}(A) = \bar{x} + \mathcal{R}(A)$
- Computation of A: $y_i = x_i - \bar{x}$

$$\min_{A \in \mathrm{St}(m,d)} \sum_{i=1}^{N} \|y_i - AA^{\mathsf{T}} y_i\|^2 = \min_{A \in \mathrm{St}(m,d)} \sum_{i=1}^{N} \|y_i\|^2 - \langle AA^{\mathsf{T}} y_i, y_i \rangle$$

Find

$$\operatorname*{argmax}_{A \in \mathrm{St}(m,d)} \sum_{i=1}^{N} y_i^{\mathsf{T}} AA^{\mathsf{T}} y_i$$

$$= \operatorname*{argmax}_{A \in \mathrm{St}(m,d)} \sum_{k=d}^{N} a_k^{\mathsf{T}} \underbrace{YY^{\mathsf{T}}}_{C = \mathsf{Cov}(X)} a_k$$

Find d orthogonal eigenvectors $a_1, \dots, a_d$ of C belonging to the largest d eigenvectors of C

Autoencoder:

$$\|x - D_\theta(E_\phi(x))\|^2 = \|x - b - AA^{\mathsf{T}}(x - b)\|^2$$
$$E_\phi(x) := A^{\mathsf{T}}(x - b), \quad D_\theta(y) = Ay + b$$

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58

2. Variational Autoencoders (VAE):

- encoder $D(z) = D_\theta(z) := (\mu_\theta(z), \Sigma_\theta(z))$

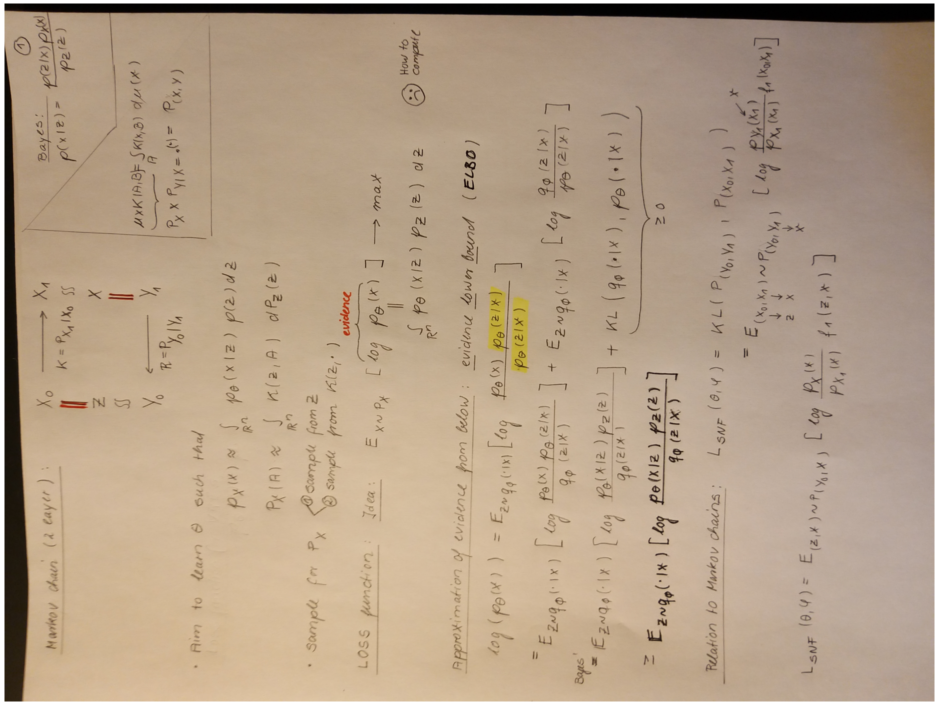
$$P_{X_1|Z=z}(\cdot) = \mathcal{K}(z,\cdot) := \mathcal{N}(\mu_\theta(z), \Sigma_\theta(z)), \quad p_\theta(x|z) = \mathcal{N}(x; \mu_\theta(z), \Sigma_\theta(z))$$

- decoder $E(x) = E_\phi(x) := (\mu_\phi(x), \Sigma_\phi(x))$

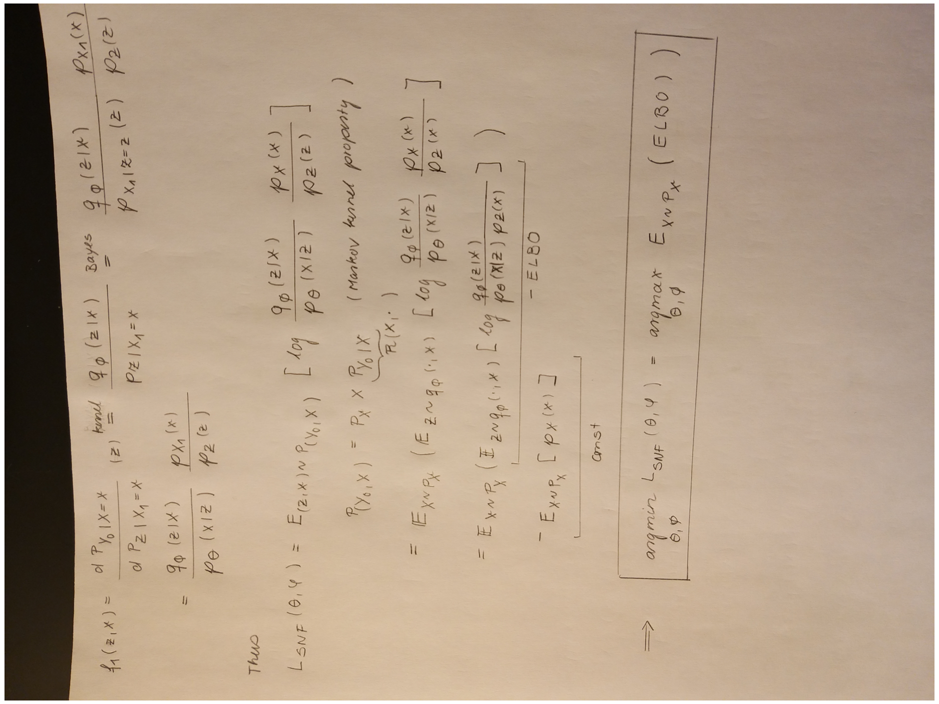
$$P_{Y_0|X=x}(\cdot) = \mathcal{R}(x,\cdot) := \mathcal{N}(\mu_\phi(x), \Sigma_\phi(x)), \quad q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \Sigma_\phi(x))$$

SNF $((X_0, X_1), (Y_1, Y_0))$ one layer Markov chain with

$$\begin{aligned} \mathcal{L}_{\mathsf{SNF}}(\theta,\phi) &= \mathbb{E}_{(z,x)\sim P_{(Y_0,Y_1)}}\left[\log\left(\frac{p_X(x) f_1(z,x)}{p_{X_1}(x)}\right)\right] \\ &= -\underbrace{\mathbb{E}_{x\sim P_X}\left[\mathcal{L}_{\theta,\phi}(x)\right]}_{\text{ELBO}} + \underbrace{\mathbb{E}_{x\sim P_X}[\log(p_X(x))]}_{\text{const}} \end{aligned}$$



1
Markov chain (2 layer):
$X_0 \xrightarrow[K = P_{X_1 \mid X_0}]{} X_1$
Z
X
$Y_0 \xleftarrow[R = P_{Y_0 \mid Y_1}]{} Y_1$
Bayes: $p(x \mid z) = \frac{p(z \mid x)\, p_X(x)}{p_Z(z)}$
$\underbrace{\mu \times K(A,B)}= \int_A K(x,B)\, d\mu(x)$
$P_X \times P_{Y \mid X = \cdot}(\cdot) = P_{(X,Y)}$
· Aim to learn θ such that
$p_X(x) \approx \int_{\mathbb{R}^n} p_\theta(x \mid z)\, p(z)\, dz$
$P_X(A) \approx \int_{\mathbb{R}^n} K(z, A)\, dP_Z(z)$
· Sample for P_X: ① sample from Z ② sample from $K(z, \cdot)$
LOSS function: Idea: $E_{X \sim P_X} [\overbrace{\log p_\theta(x)}^{\text{evidence}}] \to \max$
$= \int_{\mathbb{R}^n} p_\theta(x \mid z)\, p_Z(z)\, dz$
How to compute
Approximation of evidence from below: evidence lower bound (ELBO)
$\log(p_\theta(x)) = E_{Z \sim q_\phi(\cdot \mid x)} \left[\log \frac{p_\theta(x)\, p_\theta(z \mid x)}{p_\theta(z \mid x)} \right]$
$= E_{Z \sim q_\phi(\cdot \mid x)} \left[\log \frac{p_\theta(x)\, p_\theta(z \mid x)}{q_\phi(z \mid x)} \right] + E_{Z \sim q_\phi(\cdot \mid x)} \left[\log \frac{q_\phi(z \mid x)}{p_\theta(z \mid x)} \right]$
Bayes $= E_{Z \sim q_\phi(\cdot \mid x)} \left[\log \frac{p_\theta(x \mid z)\, p_Z(z)}{q_\phi(z \mid x)} \right] + \underbrace{KL(q_\phi(\cdot \mid x), p_\theta(\cdot \mid x))}_{\geq 0}$
$\geq E_{Z \sim q_\phi(\cdot \mid x)} \left[\log \frac{p_\theta(x \mid z)\, p_Z(z)}{q_\phi(z \mid x)} \right]$
Relation to Markov chains: $L_{SNF}(\theta, \varphi) = KL(P_{(Y_0, Y_1)}, P_{(X_0, X_1)})$
$= E_{(X_0, X_1) \sim P_{(Y_0, Y_1)}} \left[\log \frac{p_{Y_1}(x_1)}{p_{X_1}(x_1)} f_1(x_0, x_1) \right]$
$L_{SNF}(\theta, \varphi) = E_{(z,x) \sim P_{(Y_0, X)}} \left[\log \frac{p_X(x)}{p_{X_1}(x)} f_1(z, x) \right]$



$f_1(z,x) = \frac{d P_{Y_0 \mid X = x}}{d P_{Z \mid X_1 = x}}(z) \overset{\text{kernel}}{=} \frac{q_\phi(z \mid x)}{p_{Z \mid X_1 = x}} \overset{\text{Bayes}}{=} \frac{q_\phi(z \mid x)}{p_{X_1 \mid Z = z}(z)} \frac{p_{X_1}(x)}{p_Z(z)}$
$= \frac{q_\phi(z \mid x)}{p_\theta(x \mid z)} \frac{p_{X_1}(x)}{p_Z(z)}$
Thus
$L_{SNF}(\theta, \varphi) = E_{(z,x) \sim P_{(Y_0, X)}} \left[\log \frac{q_\phi(z \mid x)}{p_\theta(x \mid z)} \frac{p_X(x)}{p_Z(z)} \right]$
$P_{(Y_0, X)} = P_X \times \underbrace{P_{Y_0 \mid X}}_{R(X, \cdot)}$ (Markov kernel property)
$= E_{X \sim P_X} \left(E_{Z \sim q_\phi(\cdot, x)} \left[\log \frac{q_\phi(z \mid x)}{p_\theta(x \mid z)} \frac{p_X(x)}{p_Z(x)} \right] \right)$
$= E_{X \sim P_X} \left(\underbrace{E_{Z \sim q_\phi(\cdot, x)} \left[\log \frac{q_\phi(z \mid x)}{p_\theta(x \mid z)\, p_Z(x)} \right]}_{-\text{ELBO}} \right)$
$- \underbrace{E_{X \sim P_X} [p_X(x)]}_{\text{const}}$
$\Longrightarrow \boxed{\underset{\theta, \varphi}{\text{argmin}}\, L_{SNF}(\theta, \varphi) = \underset{\theta, \varphi}{\text{argmax}}\, E_{X \sim P_X} (\text{ELBO})}$

Markov chain:

$$X_t := X_{t-1} + a_1 \nabla \log p_t(X_{t-1}) + a_2 \xi_t,$$

where $a_1, a_2 > 0$ constants, $\xi_t \sim \mathcal{N}(0, I)$ and p_t proposal density

Markov kernels $\mathcal{K}_t = \mathcal{R}_t$: $\mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0,1]$:

$$\mathcal{K}_t(x, \cdot) = \mathcal{N}(x - a_1 \nabla u_t(x), a_2^2 I).$$

Proof Use independence of ξ_t of X_t and X_{t-1} to obtain that X_t and X_{t-1} have the common density

$$\begin{aligned} p_{(X_{t-1},X_t)}(x_{t-1}, x_t) &= p_{X_{t-1},\xi_t}(x_{t-1}, \tfrac{1}{a_2}(x_t - x_{t-1} + a_1 \nabla u_t(x_{t-1})) \\ &= p_{X_{t-1}}(x_{t-1}) p_{\xi_t}(\tfrac{1}{a_2}(x_t - x_{t-1} + a_1 \nabla u_t(x_{t-1}))) \\ &= p_{X_{t-1}}(x_{t-1}) \mathcal{N}(x_t; x_{t-1} - a_1 \nabla u_t(x_{t-1}), a_2^2 I). \end{aligned}$$

Then, for $A, B \in \mathcal{B}(\mathbb{R}^d)$, it holds

$$\begin{aligned} P_{(X_{t-1},X_t)}(A \times B) &= \int_{A \times B} p_{X_{t-1}}(x_{t-1}) \mathcal{N}(x_t; x_{t-1} - a_1 \nabla u_t(x_{t-1}), a_2^2 I) \mathrm{d}(x_{t-1}, x_t) \\ &= \int_A \int_B \mathcal{N}(x_t; x_{t-1} - a_1 \nabla u_t(x_{t-1}), a_2^2 I) \mathrm{d}x_t p_{X_{t-1}}(x_{t-1}) \mathrm{d}x_{t-1} \\ &= \int_A \mathcal{K}_t(x_{t-1}, B) p_{X_t}(x_t) \mathrm{d}P_{X_{t-1}}(x_{t-1}). \end{aligned}$$

Background: Langevin Layer (for fixed t)

Overdamped Langevin SDE

$$X(0) = X^0,$$
$$dX(\tau) = -\nabla\Psi(X(\tau))d\tau + \sqrt{2\beta^{-1}}dW(\tau)$$

Euler-Maruyama forward step:

$$X_t := X_{t-1} - \eta\nabla\Psi(X_{t-1}) + \sqrt{2\beta^{-1}\eta}\,\xi_t$$
$$= X_{t-1} + \underbrace{\eta\beta^{-1}}_{a_1}\nabla\log p(X_{t-1}) + \underbrace{\sqrt{2\beta^{-1}\eta}}_{a_2}\,\xi_t$$

where $p(x) = C^{-1}\mathrm{e}^{-\beta\Psi(x)}$, C normalizing factor

- $\nabla\log p$ is known as score

Density of random variables is described by Fokker-Planck equation

$=$ Wasserstein gradient flow of $\mathcal{F}(\mu) = \mathrm{KL}\,(\mu, p\,\mathrm{d}x)$

$$\rho(x,0) = \rho^0(x)$$
$$\frac{\partial\rho}{\partial\tau} = \mathrm{div}\,(\rho\nabla\Psi(x)) + \beta^{-1}\Delta\rho$$

As $\tau\to\infty$ stationary solution: $p(x) = C^{-1}\mathrm{e}^{-\beta\Psi(x)}$, C normalizing factor

For a proper, lsc function $\mathcal{F}\colon \mathbb{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$ and $\tau > 0$, the Wasserstein proximal mapping is the set-valued function

$$\mathrm{prox}_{\tau\mathcal{F}}(\mu) := \operatorname*{argmin}_{\nu \in \mathbb{P}_2(\mathbb{R}^d)} \left\{\tfrac{1}{2\tau} W_2^2(\mu,\nu) + \mathcal{F}(\nu)\right\}, \quad \mu \in \mathbb{P}_2(\mathbb{R}^d).$$

- existence and uniqueness of the minimizer is assured if $\mathcal{F}$ is λ-convex along generalized geodesics, where $\lambda > -1/\tau$ and $\mu \in \operatorname{dom} \mathcal{F}$

Backward scheme = Jordan-Kinderlehrer-Otto (JKO) scheme starting at $\mu_\tau^0 := \mu \in \mathbb{P}_2(\mathbb{R}^d)$ with time step size τ is the curve $\gamma_\tau : [0, +\infty) \to \mathbb{P}_2(\mathbb{R}^d)$ given by

$$\gamma_\tau|_{((n-1)\tau, n\tau]} := \mu_\tau^n := \mathrm{prox}_{\tau\mathcal{F}}(\mu_\tau^{n-1})$$

- If $\mathcal{F}\colon \mathbb{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ is coercive and λ-convex along generalized geodesics, then the JKO curves γ_τ starting at $\mu \in \overline{\operatorname{dom} \mathcal{F}}$ converge for $\tau \to 0$ locally uniformly to a locally Lipschitz curve $\gamma\colon (0, +\infty) \to \mathbb{P}_2(\mathbb{R}^d)$ which is the unique Wasserstein gradient flow of $\mathcal{F}$ with $\gamma(0+) = \mu$

Diffusion Layer

SDE: $\mathrm{d}X_t = g_t(X_t)\mathrm{d}t + h_t \mathrm{d}W_t$

Explicit Euler discretization with step size $\epsilon > 0$:

$$X_t = X_{t-1} + \epsilon g_{t-1}(X_{t-1}) + \sqrt{\epsilon} h_{t-1} \xi_{t-1}, \quad t = 1, ..., T,$$

where $\xi_{t-1} \sim \mathcal{N}(0, I)$ is independent of $X_0, ..., X_{t-1}$

Markov kernel:

$$\mathcal{K}_t(x, \cdot) = P_{X_t | X_{t-1} = x} = \mathcal{N}(x + \epsilon g_{t-1}(x), \epsilon h_{t-1}^2).$$

- Song et al. 2020 parametrized the functions $g_t \colon \mathbb{R}^d \to \mathbb{R}^d$ by some a-priori learned scoring network
- Motivated by the time-reversal process of the SDE, Zhang/Chen 2021 introduced the backward layer

$$\mathcal{R}_t(x, \cdot) = P_{Y_{t-1} | Y_t = x} = \mathcal{N}(x + \epsilon(g_t(x) - h_t^2 s_t(x)), \epsilon h_t^2)$$

Tweedie Formula

$Y = x + \eta,\ \eta \sim \mathcal{N}(0, \sigma^2 I)$

$p(y) = \int p(y|x)p(x)\,\mathrm{d}x = C \int \mathrm{e}^{-\|y-x\|^2/(2\sigma^2)} p(x)\,\mathrm{d}x$

MMSE (maximum mean square error):

$$\hat{x}(y) = \mathbb{E}[X|Y=y] = \int x p(x|y)\,\mathrm{d}x = \int x \frac{p(y|x)p(x)}{p(y)}\,\mathrm{d}x$$
$$= y + \sigma^2 \nabla_y \log p(y) \quad \text{Tweedie formula}$$

since

$$\nabla_y p(y) = C \int \frac{1}{\sigma^2} \mathrm{e}^{-\|y-x\|^2/(2\sigma^2)} (x-y) p(x)\,\mathrm{d}x$$
$$\frac{\sigma^2}{p(y)} \nabla_y p(y) = \frac{C}{p(y)} \int \mathrm{e}^{-\|y-x\|^2/(2\sigma^2)} x p(x)\,\mathrm{d}x - \frac{Cy}{p(y)} \int \mathrm{e}^{-\|y-x\|^2/(2\sigma^2)} p(x)\,\mathrm{d}x$$
$$= \underbrace{\int p(x|y) x\,\mathrm{d}x}_{\hat{x}(y)} - y \underbrace{\int p(x|y)\,\mathrm{d}x}_{1}$$
$$\hat{x}(y) - y = \frac{\sigma^2}{p(y)} \nabla_y p(y) = \sigma^2 \nabla_y \log p(y)$$

Refs: papers of Kadkhodaie, Simoncelli et al. 2020,

Laumont, Almansa, Pereyra, Delon et al.: Bayesian imaging using PnP priors: when Langevin meets Tweedie, 2022

Tweedie Formula

Special case used in MMSE denoising:

$$Y = X + \eta, \quad \eta \sim \mathcal{N}(0, \sigma^2 I), \quad X \sim \mathcal{N}(\mu_x, \Sigma_x) \quad \Rightarrow \quad \Sigma_Y = \Sigma_X + \sigma^2 I$$

Then

$$\hat{x}(y) = \mu_y + (\Sigma_y - \sigma^2 I)\Sigma_Y^{-1}(y - \mu_y)$$

Application:

- MMSE estimation for similar patches: Choose $s \times s$ neighborhood (patch) y_i centered at $i = (i_1, i_2) \in \mathcal{G}$ and interpret this and similar patches as realization of an $p = s^2$-dimensional random vector $Y_i \sim \mathcal{N}(\mu_i, \Sigma_i)$

$$\hat{y}_j = \hat{\mu}_i + (\hat{\Sigma}_i - \sigma^2 I_p)\hat{\Sigma}_i^{-1}(y_j - \hat{\mu}_i), \qquad j \in \mathcal{S}(i).$$

  where $\mathcal{S}(i)$ is the set of centers of patches similar to y_i

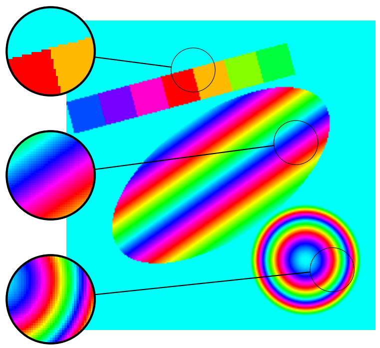
- More fun on manifolds: MMSE estimation for manifold-valued images:

$$\hat{y}_j = \exp_{\hat{\mu}_i}\left((\hat{\Sigma}_i - \sigma^2 I_{pd})\hat{\Sigma}_i^{-1}(\log_{\hat{\mu}_i} y_j)\right), \qquad j \in \mathcal{S}(i)$$

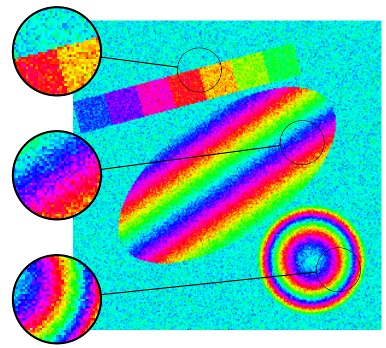
Refs: MMSE by Lebrun/Morel (Denoising cuisine) 2013,

Laus, Nikolova, Persch, Steidl: A nonlocal denoising algorithm for manifold-valued images using second order statistics 2017

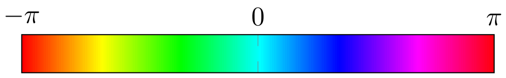
Examples

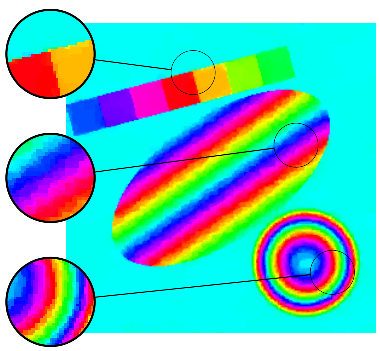


Original

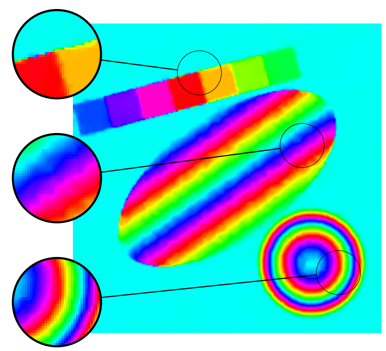


Noisy (88.5×10^{-3})

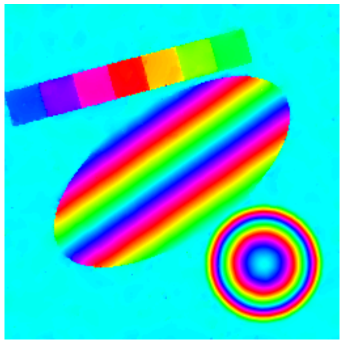




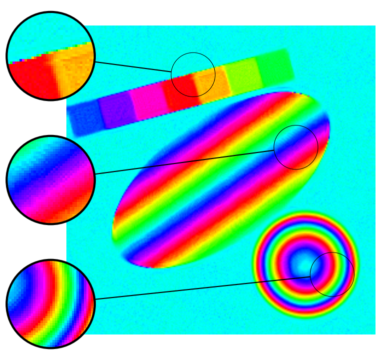
TV (7.2×10^{-3})



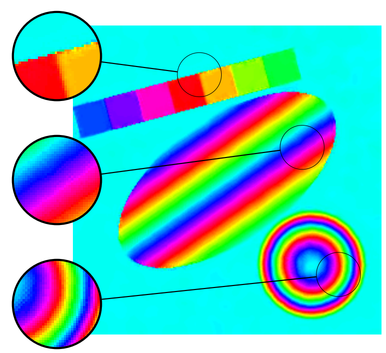
TV-TV2 (5.2×10^{-3})



TGV (2.6×10^{-3})



NL-means (8.1×10^{-3})



NL-MMSE (2.5×10^{-3})

Outline

Inverse Problems:

$$Y = F(x) + \Xi, \quad \Xi \sim \mathcal{N}(0,\sigma) \Rightarrow Y \sim \mathcal{N}(F(x),\sigma)$$
$$Y = F(X) + \Xi$$

Aim: Sample from $P_{X|Y=y}$ for some y

Idea: $P_{X|Y=y} \approx \mathcal{T}(y,\cdot)_{\#} P_Z$

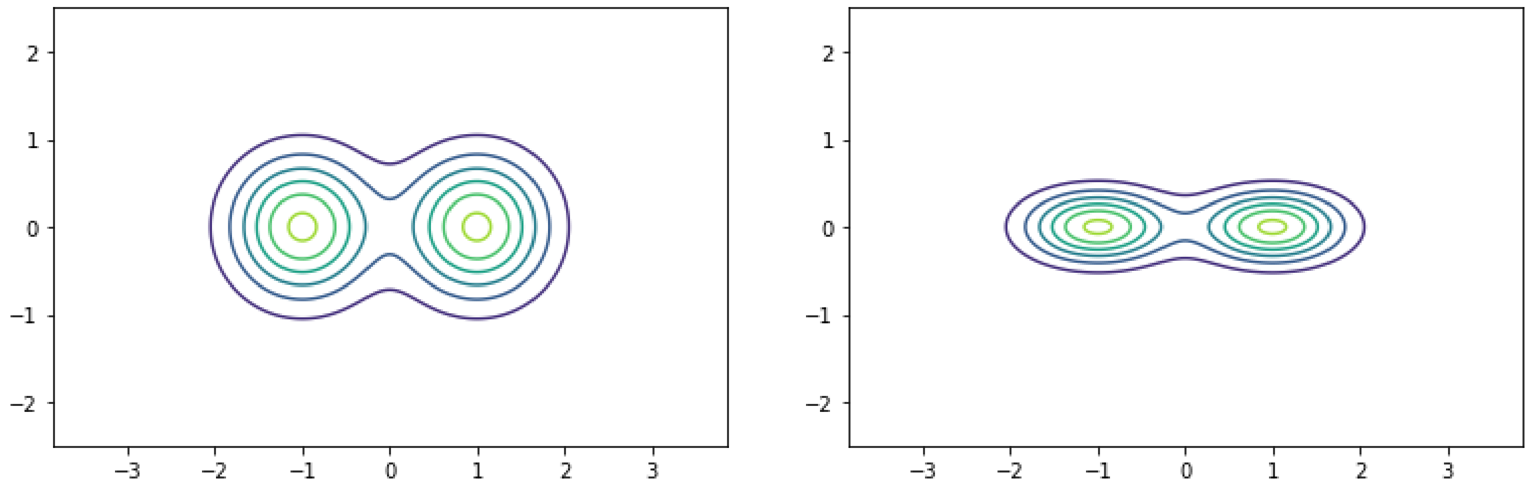


Illustration of the prior density p_X (left) and the posterior density $p_{X|Y=y}$ for $y=0$ (right) within the inverse problem with $F(x_1,x_2)=x_2$ and $\Xi \sim \mathcal{N}(0, 0.1^2)$

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58

Conditional SNF conditioned to $Y = y$ is pair of Markov chains

$$((X_0, ..., X_T), (Y_T, ..., Y_0))$$

cP1) the conditional distributions $P_{X_t|Y=y}$ and $P_{Y_t|Y=y}$ have densities

$$p_{X_t}(y, \cdot)\colon \mathbb{R}^{d_t} \to \mathbb{R}_{>0}, \quad p_{Y_t}(y, \cdot)\colon \mathbb{R}^{d_t} \to \mathbb{R}_{>0}$$

cP2) for P_Y-almost every y, there exist Markov kernels $\mathcal{K}_t\colon \mathbb{R}^{\tilde{d}} \times \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0,1]$ and $\mathcal{R}_t\colon \mathbb{R}^{\tilde{d}} \times \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0,1]$ such that

$$P_{(X_0,\dots,X_T)|Y=y} = P_{X_0} \times \mathcal{K}_1(y,\cdot,\cdot) \times \cdots \times \mathcal{K}_T(y,\cdot,\cdot),$$
$$P_{(Y_T,\dots,Y_0)|Y=y} = P_{Y_T} \times \mathcal{R}_T(y,\cdot,\cdot) \times \cdots \times \mathcal{R}_1(y,\cdot,\cdot).$$

cP3) for P_{Y,X_t}-almost every pair $(y,x) \in \mathbb{R}^{\tilde{d}} \times \mathbb{R}^d$, the measures $P_{Y_{t-1}|Y_t=x,Y=y}$ and $P_{X_{t-1}|X_t=x,Y=y}$ are absolute continuous with respect to each other.

Loss Function: $\mathcal{L}_{\mathrm{cSNF}}(\theta) = \mathrm{KL}(P_{Y,(Y_0,\dots,Y_T)}, P_{Y,(X_0,\dots,X_T)})$

- Posterior distribution is analytically known = ground truth

Lemma: Let $X \sim \sum_{k=1}^K w_k \mathcal{N}(m_k, \Sigma_k)$. Suppose that

$$Y = AX + \Xi, \quad A: \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}, \quad \Xi \sim N(0, b^2 I)$$

Then

$$P_{X|Y=y} \propto \sum_{k=1}^K \tilde{w}_k \mathcal{N}(\cdot|\tilde{m}_k, \tilde{\Sigma}_k),$$

where

$$\tilde{\Sigma}_k := (\tfrac{1}{b^2} A^\intercal A + \Sigma_k^{-1})^{-1}, \qquad \tilde{m}_k := \tilde{\Sigma}_k (\tfrac{1}{b^2} A^\intercal y + \Sigma_k^{-1} \mu_k).$$

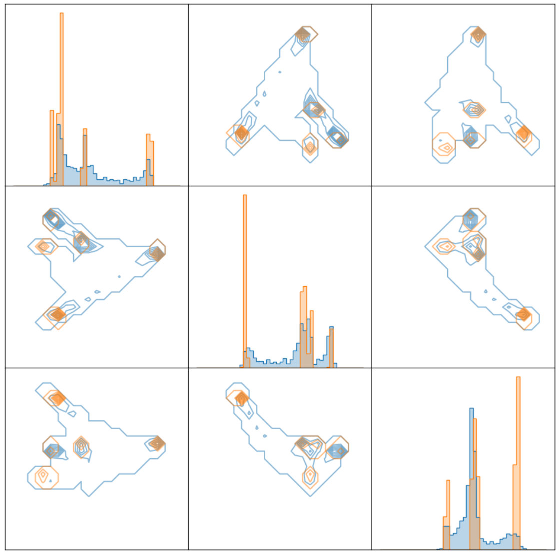
and

$$\tilde{w}_k := \frac{w_k}{|\Sigma_k|^{\frac{1}{2}}} \exp\left(\frac{1}{2}(\tilde{m}_k \tilde{\Sigma}_k^{-1} \tilde{m}_k - m_k \Sigma_k^{-1} m_k)\right).$$

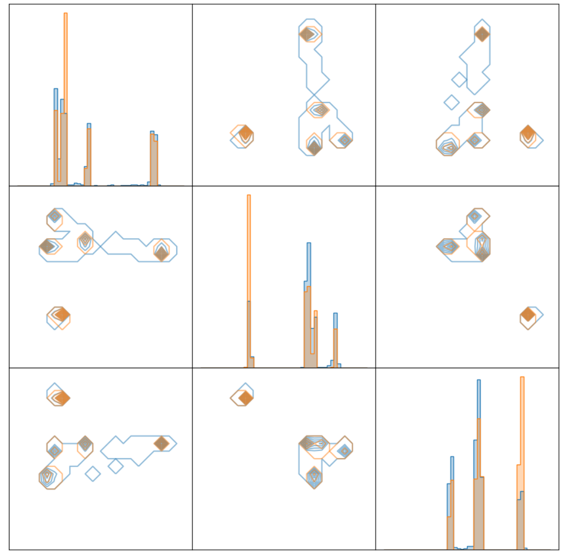
Experiment: with GMM in $\mathbb{R}^{100}$ with $K = 5$ mixture components

- A diagonal matrix
- $\Xi \sim \mathcal{N}(0, 0.05 I)$
- proposal densities: $p_t^y(x) = c_y \left(p_Z(x) p_{X|Y=y}(x)\right)^{1/2}$

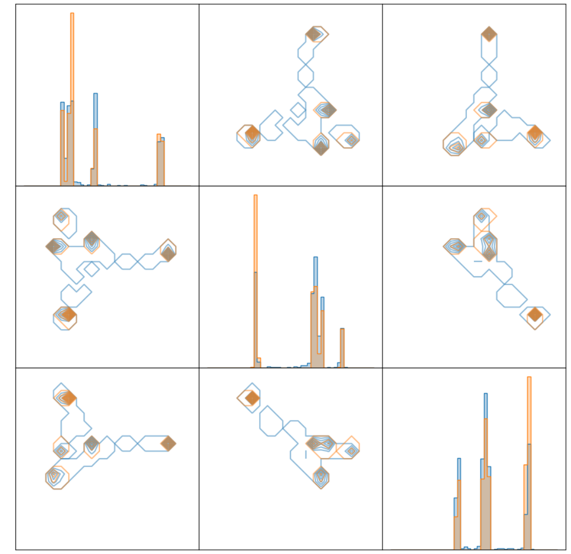
1. Approximation of the Posterior for Gaussian Mixtures



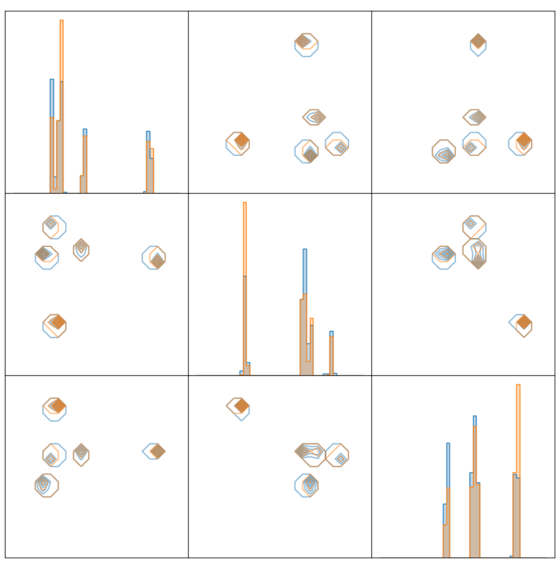
NF



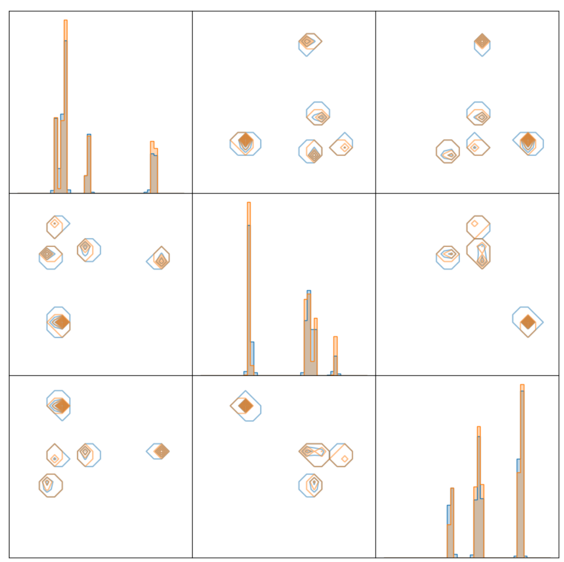
NF + MALA



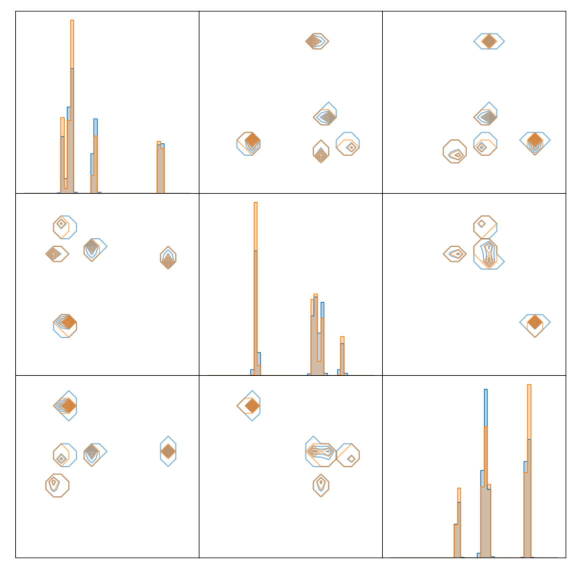
VAE



NF + VAE



VAE+MALA



NF+VAE+MALA

Method	NF	NF+MALA	VAE	VAE+NF	VAE+MALA	NF+VAE+MALA
W_1	3.55	2.92	1.22	1.18	0.8	0.82

2. Example from Scatterometry

The parameters in x-space describe the geometry of the photo masks and

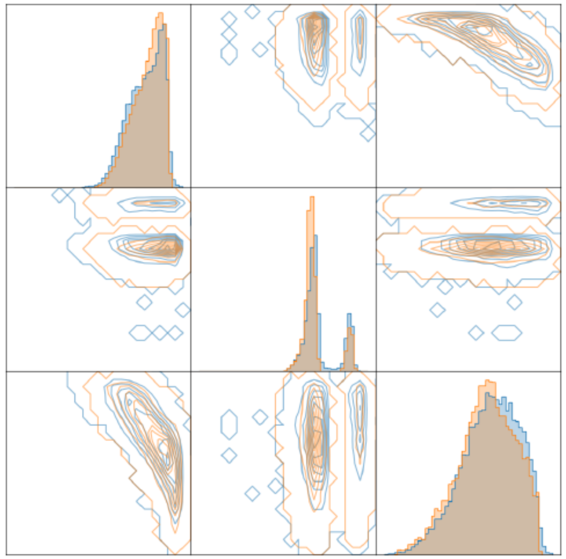
$$Y = F(X) + \eta$$

the observed diffraction pattern of light, where

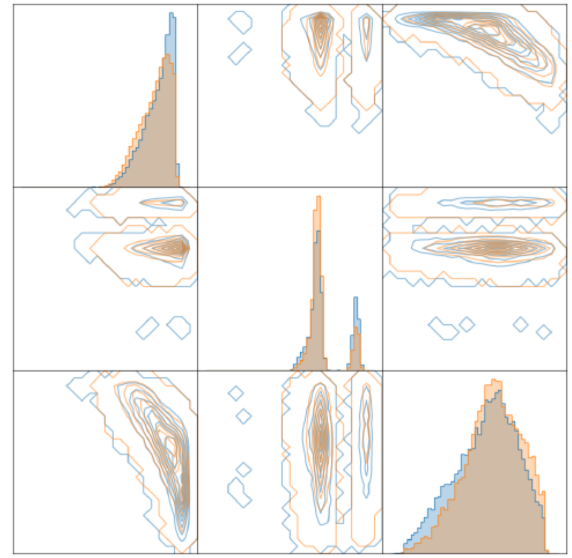
- $F\colon \mathbb{R}^3 \to \mathbb{R}^{23}$ (from a nonlinear PDE and learned by NN)
- multiplicative + additive noise model

$$\eta = aF(X)\eta_1 + b\eta_2, \quad \eta_1, \eta_2 \sim \mathcal{N}(0, I), \quad a, b > 0$$

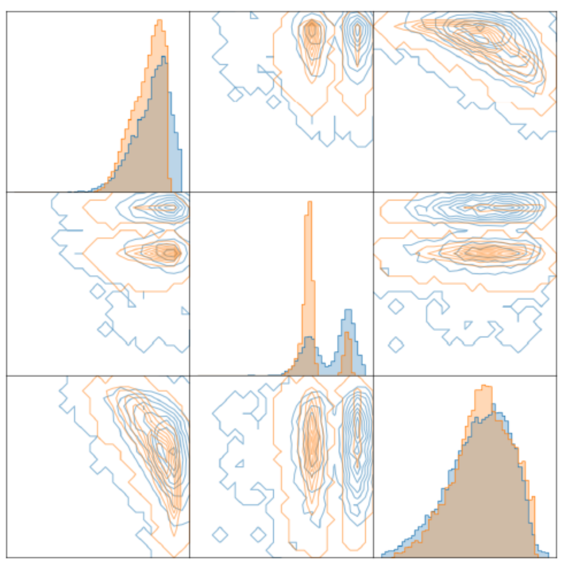
2. Example from Scatterometry



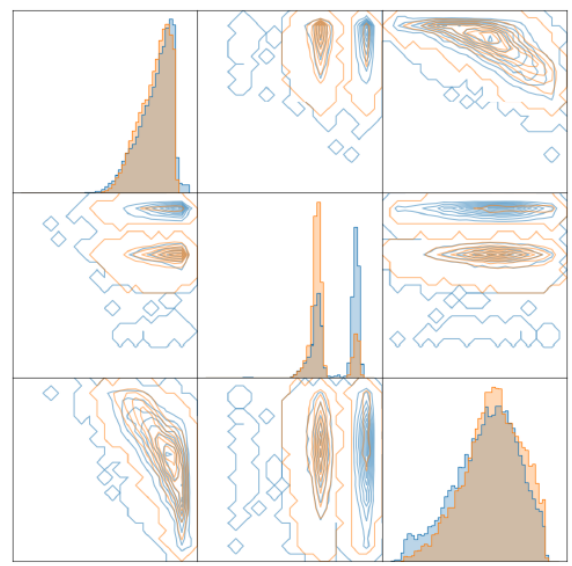
NF



NF+MALA



VAE



VAE+MALA

Method	NF	NF+MALA	VAE	VAE+MALA
KL	0.76	0.59	0.98	0.69

Cambridge Elements

Non-Local Data Interactions: Foundations and Applications

Generalized Normalizing Flows via Markov Chains

Paul Lyonel Hagemann, Johannes Hertrich, and Gabriele Steidl

Berlin Mathematics Research Center

MATH+

Funded under Germany's Excellence Strategy by
DFG Deutsche Forschungsgemeinschaft

Mathematisches
Forschungsinstitut
Oberwolfach

Member of Leibniz
Leibniz Association

Oberwolfach Seminar

Variational and Information Flows in Machine Learning and Optimal Transport

Organizers: Wuchen Li, Columbia
Bernhard Schmitzer, Göttingen
Gabriele Steidl, Berlin
Francois-Xavier Vialard, Paris
Date (ID): 19 – 25 November 2023 (2347b)
Deadline: 1 September 2023

Variational and stochastic flows are now ubiquitous in machine learning and generative modeling. Indeed, many such models can be interpreted as flows from a latent distribution to the sample distribution and training corresponds to finding the right flow vector field. Optimal transport and diffeomorphic flows provide powerful frameworks to analyze such trajectories of distributions with elegant notions from differential geometry, such as geodesics, gradient and Hamiltonian flows. Recently, mean field control and mean field games offer a general optimal control variational problems on the learning problem. How do these tools lead us to a better understanding and further development of machine learning and generative models?

The Oberwolfach Seminar will address the topic from different points of view taking in particular recent developments in machine learning into account. The target audience is PhD students and post-doctoral researchers wishing to be quickly immersed in this modern, active research area. Priority will be given to young, motivated researchers.

Please see the website of the seminar for detailed information:

www.mfo.de/occasion/2347b

The seminar takes place at the Mathematisches Forschungsinstitut Oberwolfach. The Institute covers board and lodging. By the support of the Carl Friedrich von Siemens Foundation travel expenses can be reimbursed up to 150 EUR in average per person (against copies of travel receipts). The number of participants is restricted to 25.

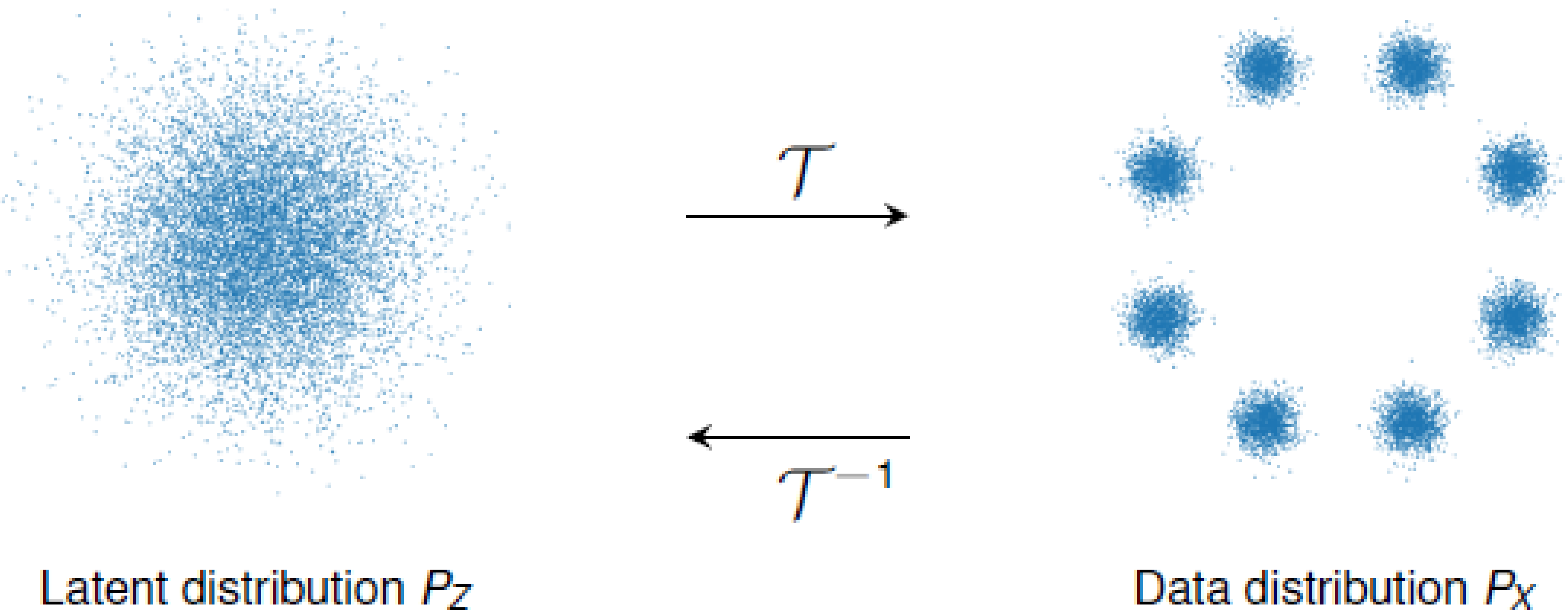
Applications including title, ID and date of the intended seminar, together with **one pdf-file attached** containing

- full name and address, incl. e-mail address
- short CV and publication list
- present position, university
- name of supervisor of Ph.D. thesis
- a short summary of previous work and interest

should be **sent by e-mail** via **seminars@mfo.de** until 1 September 2023 to:

Prof. Dr. Matthias Hieber
Mathematisches Forschungsinstitut Oberwolfach
Schwarzwaldstr. 9 – 11
77709 Oberwolfach
Germany

www.mfo.de/scientific-program/meetings/oberwolfach-seminars


$\mathcal{T}$
$\mathcal{T}^{-1}$
Latent distribution P_Z
Data distribution P_X

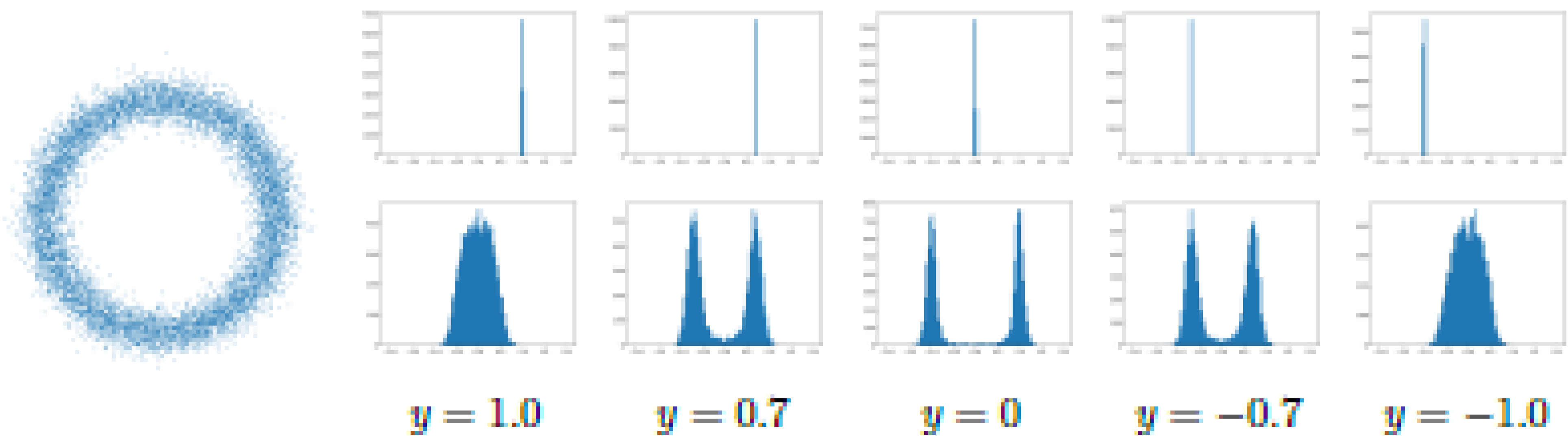
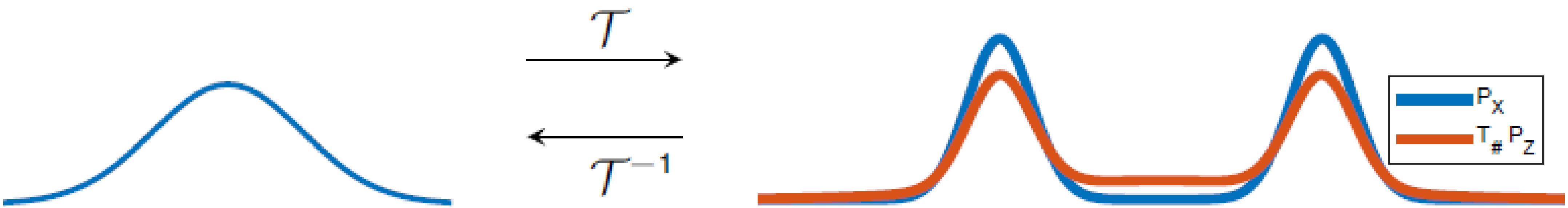
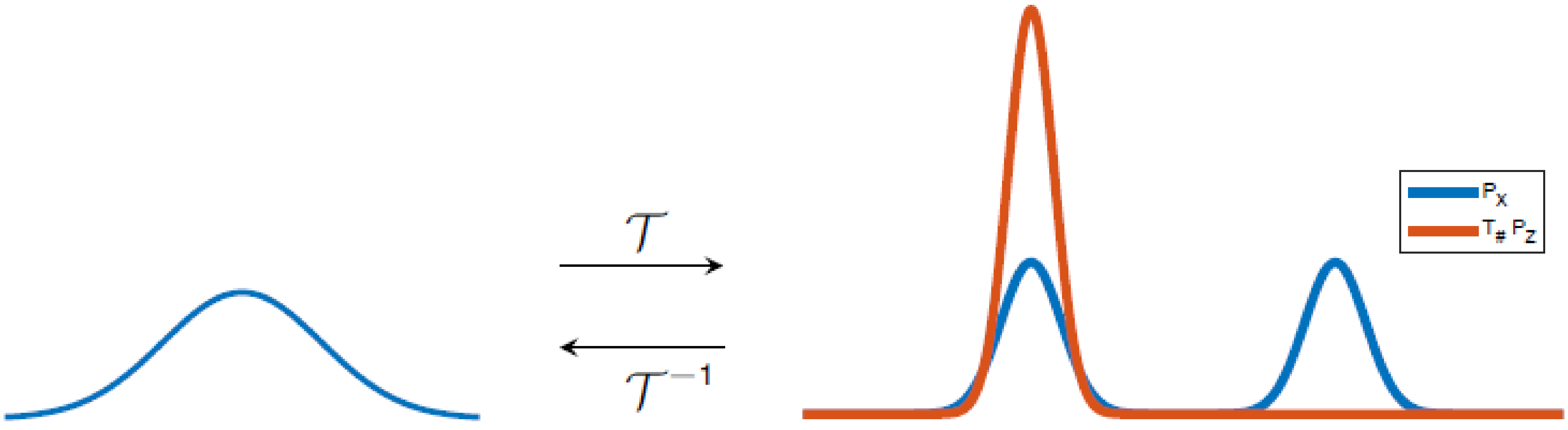
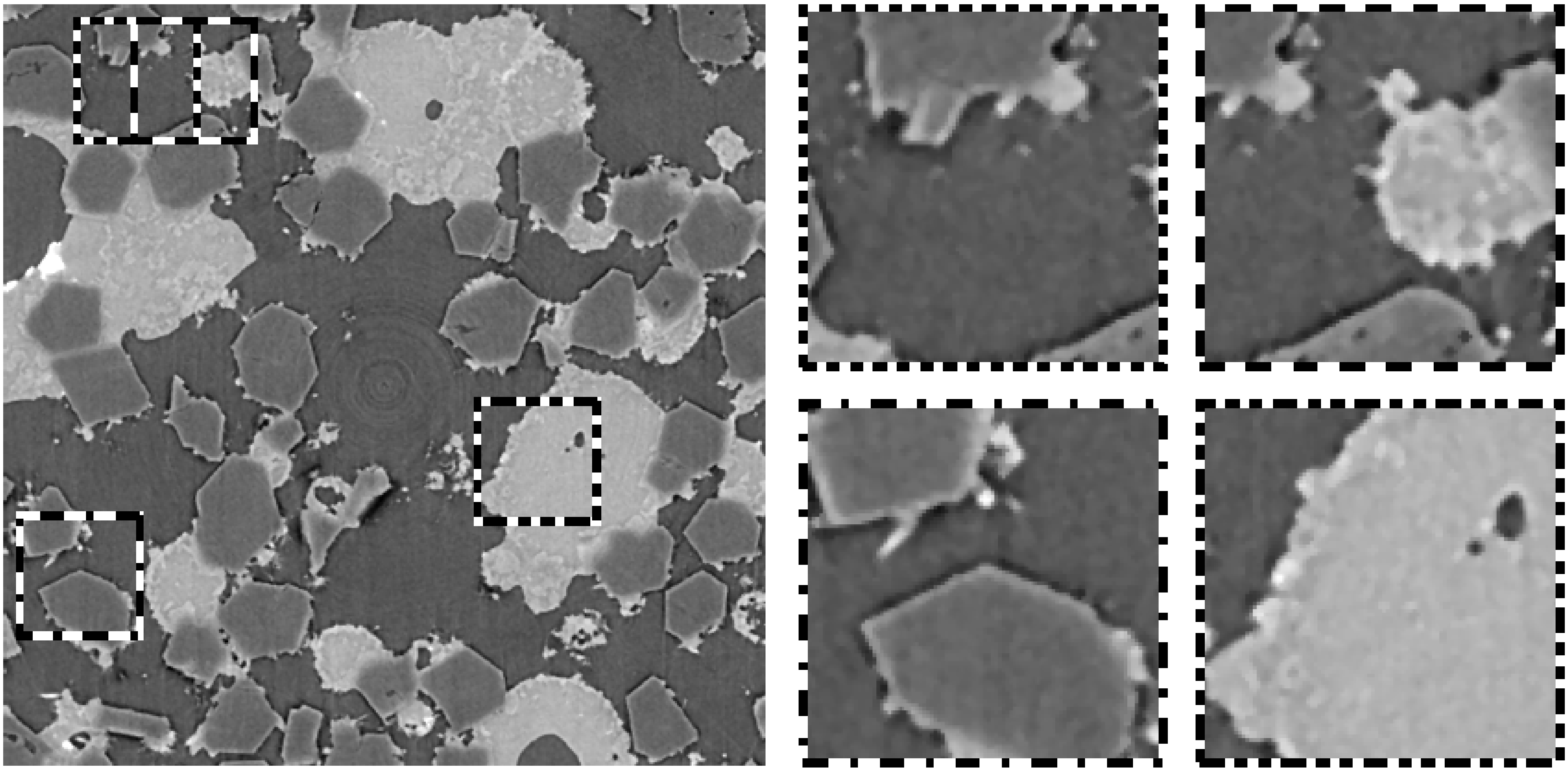


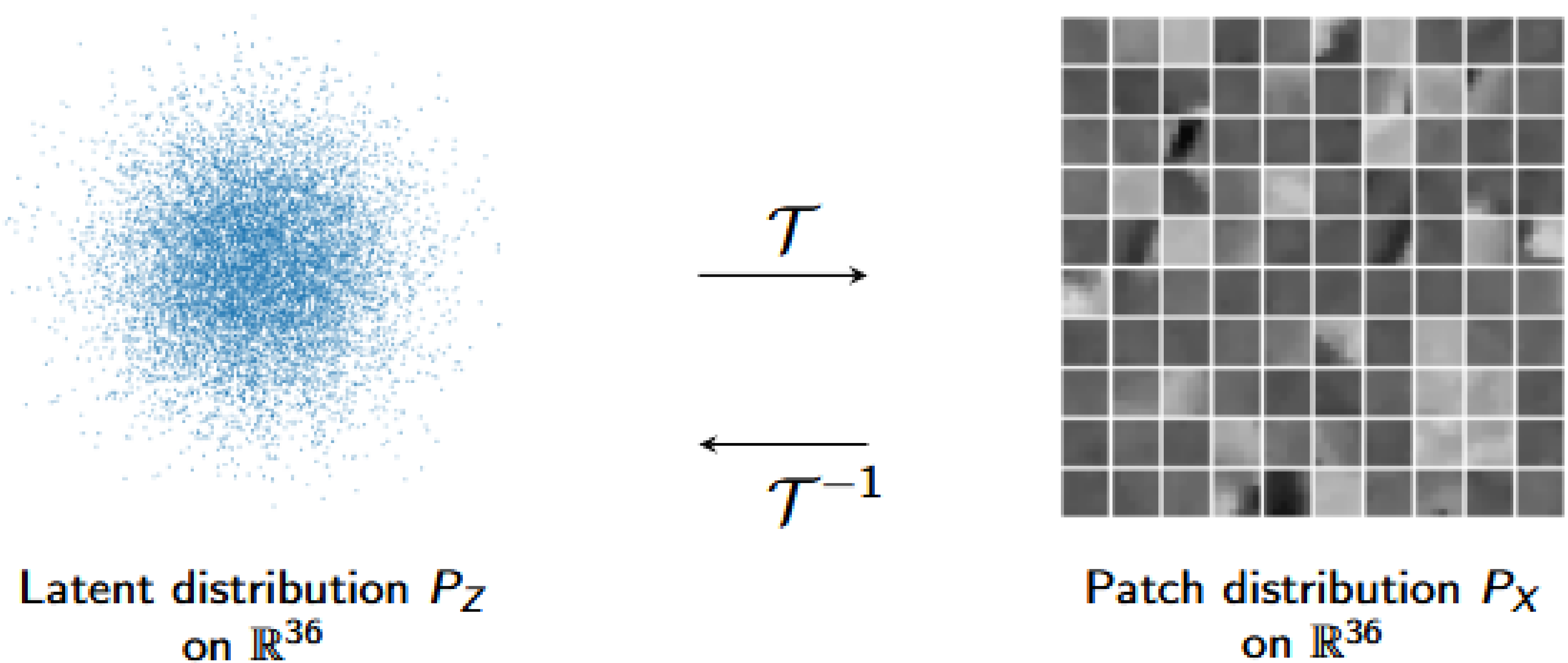
Figure 2: Left: Samples from the prior distribution of X for the circle example. Right: Histograms of samples from the reconstructed posterior distribution $P_{X|Y=y} \approx \mathcal{T}(\cdot, y)_{\#}^{-1} P_Z$ for $y \in \{1, 0.7, 0, -0.7, -1\}$ within the circle example. Top: first coordinate, Bottom: second coordinate.

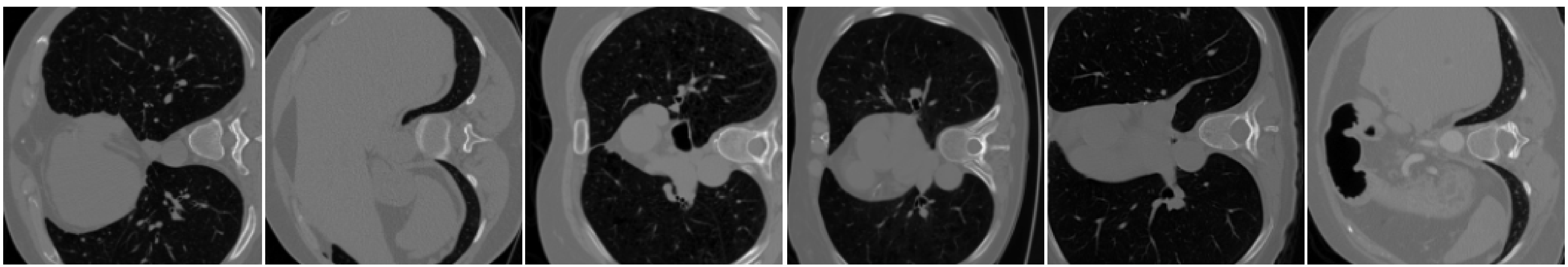


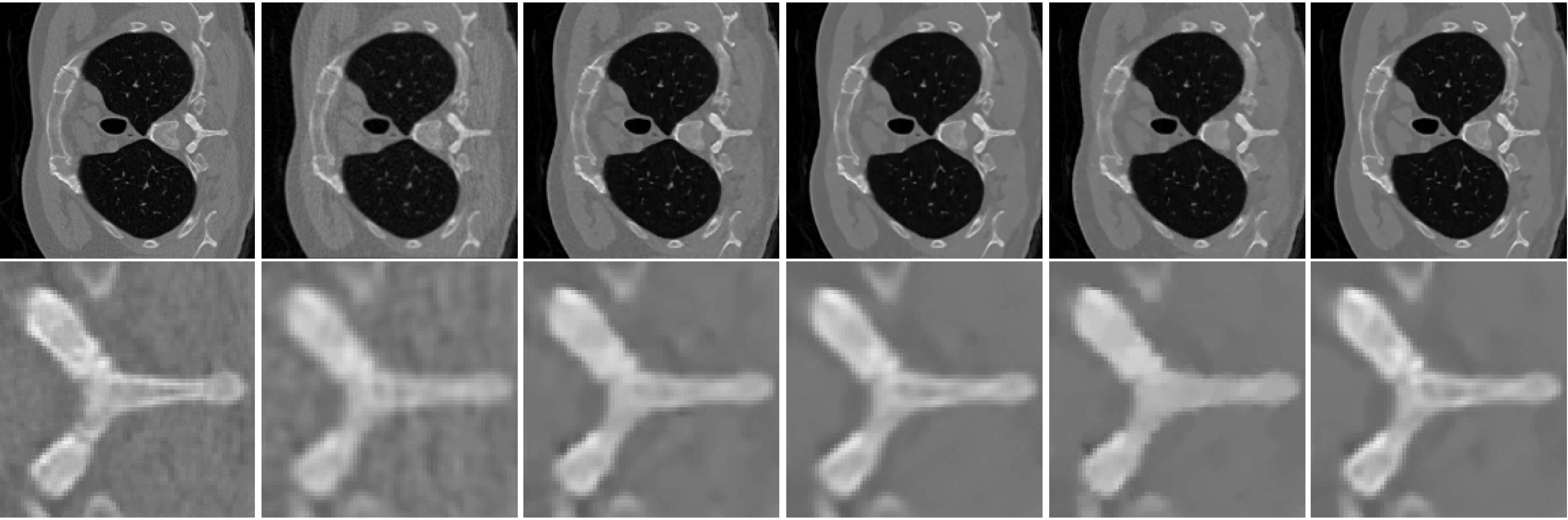
$\mathcal{T}$
$\mathcal{T}^{-1}$
P_X
$T_\# P_Z$


$\mathcal{T}$
$\mathcal{T}^{-1}$
P_X
$T_\# P_Z$

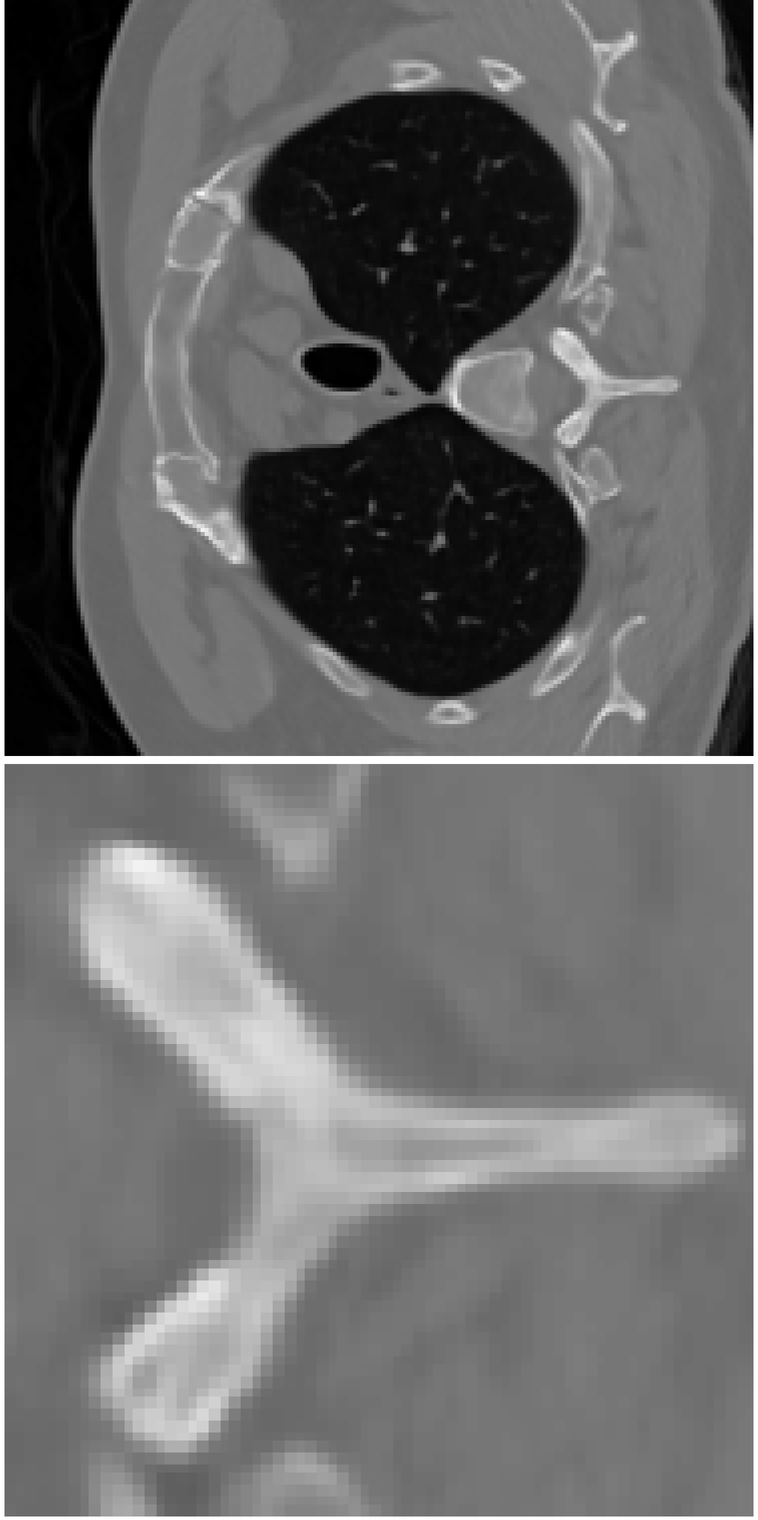



$\mathcal{T}$
$\mathcal{T}^{-1}$
Latent distribution P_Z on $\mathbb{R}^{36}$
Patch distribution P_X on $\mathbb{R}^{36}$



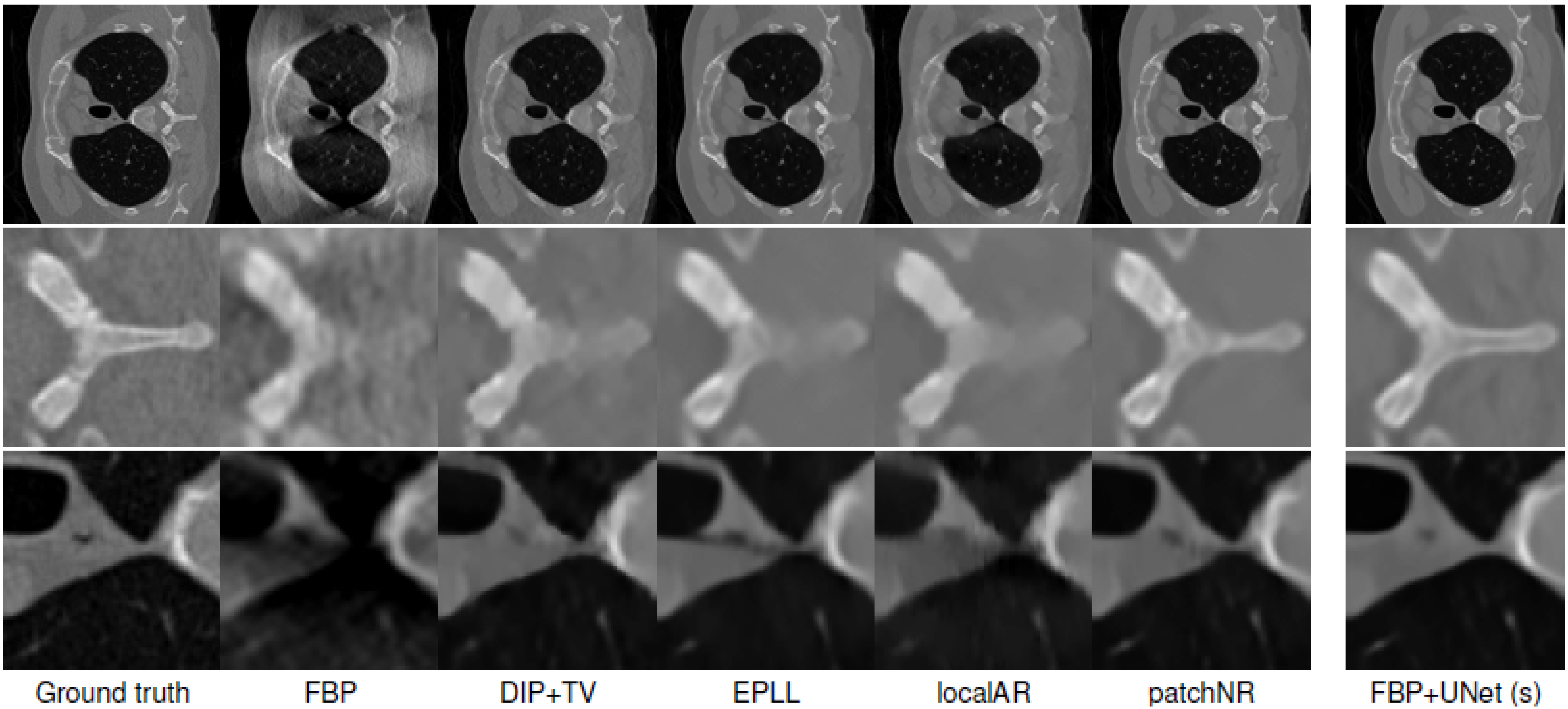


Ground truth FBP DIP+TV EPLL localAR patchNR

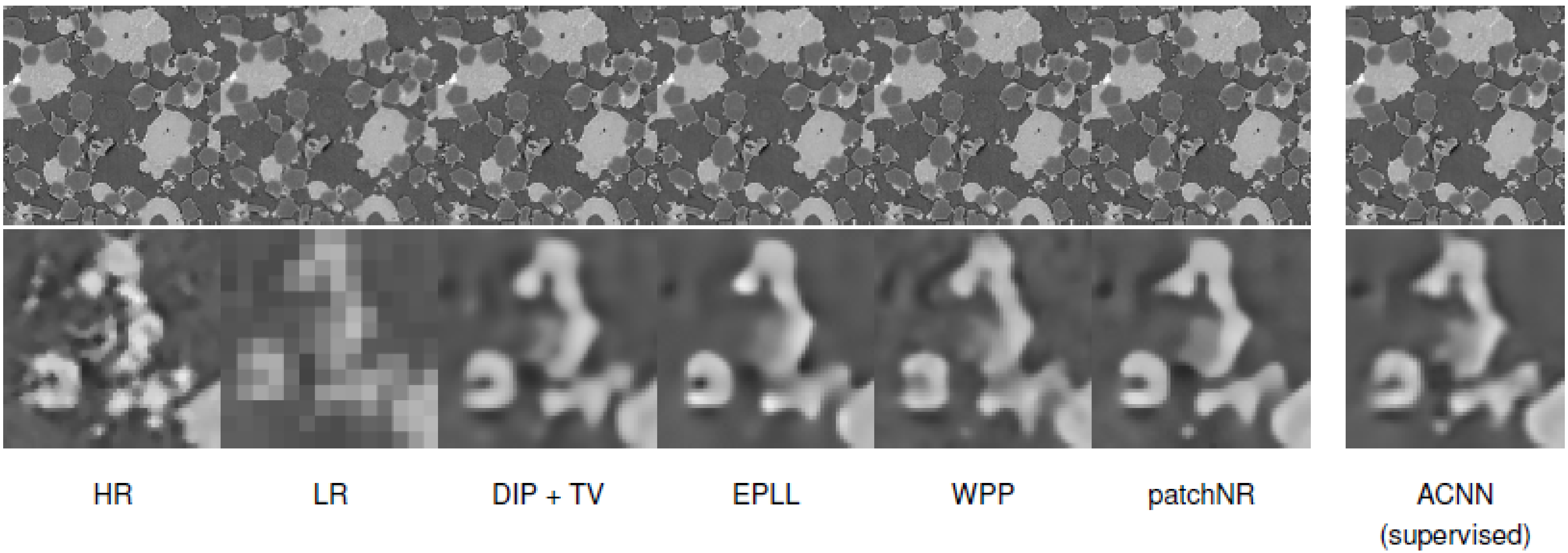


FBP+UNet

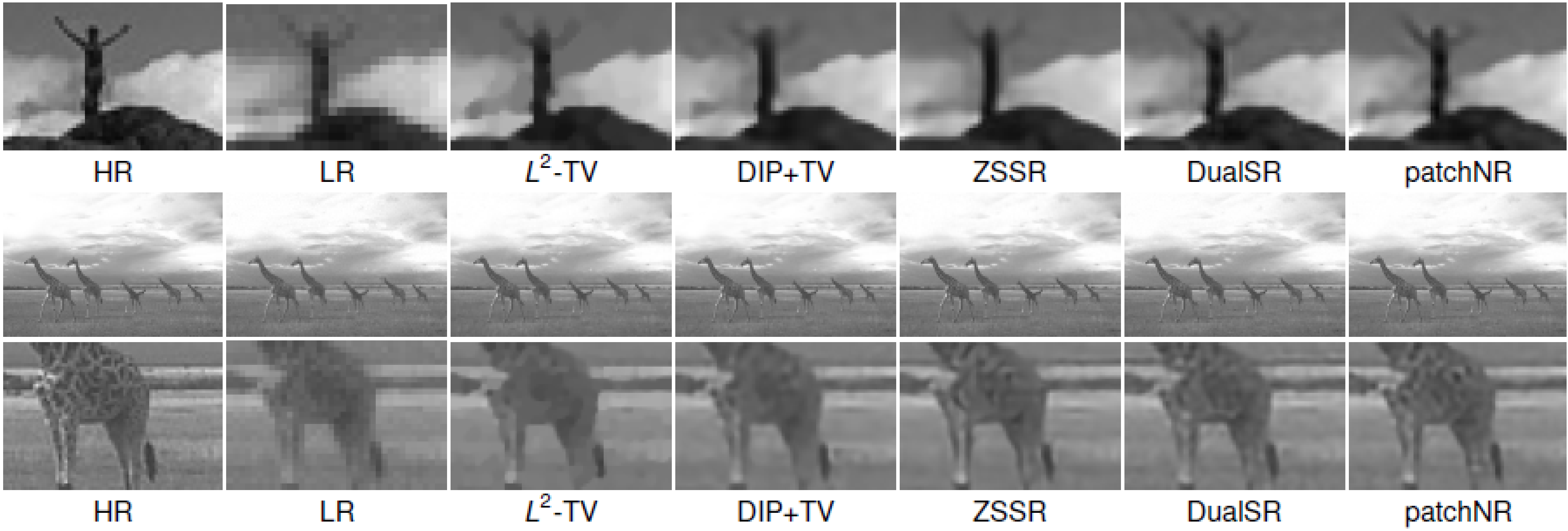
	FBP	DIP + TV	EPLL	localAR	patchNR	FBP+UNet (data-based)
PSNR	30.37 ± 2.95	34.45 ± 4.20	34.89 ± 4.41	33.64 ± 3.74	**35.19** ± 4.52	35.48 ± 4.52
SSIM	0.739 ± 0.141	0.821 ± 0.147	0.821 ± 0.154	0.807 ± 0.145	**0.829** ± 0.152	0.837 ± 0.143
Runtime	0.03s	1514.33s	36.65s	30.03s	48.39s	0.46s


Ground truth
FBP
DIP+TV
EPLL
localAR
patchNR
FBP+UNet (s)

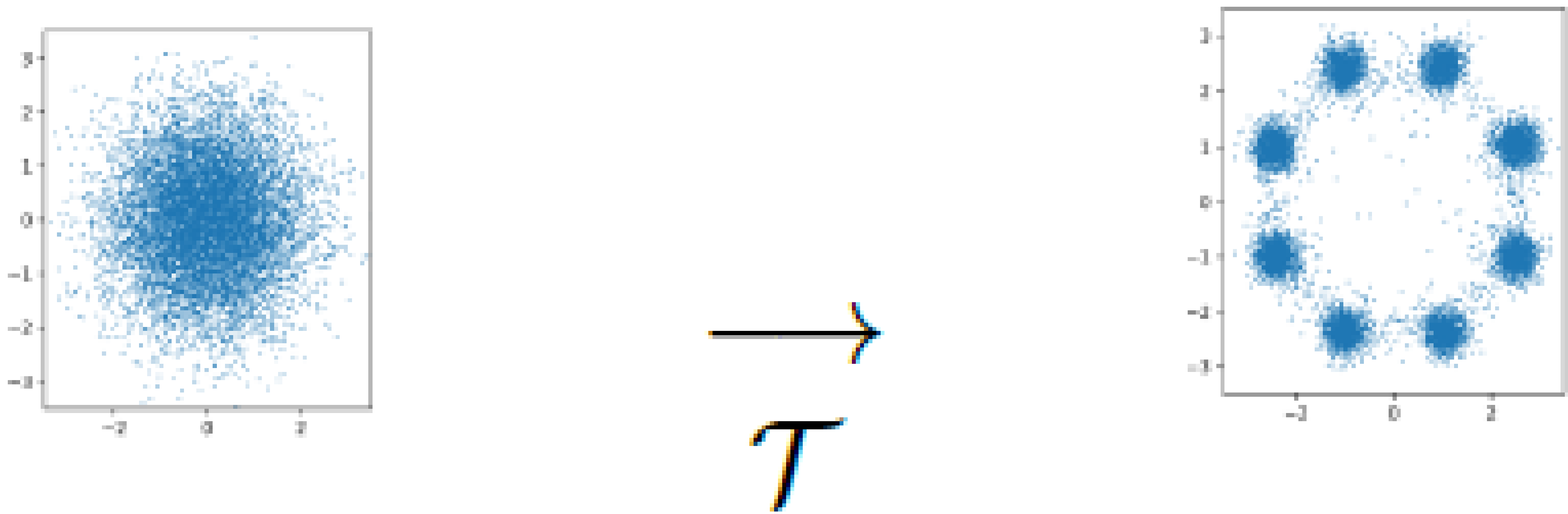
	FBP	DIP + TV	EPLL	localAR	patchNR	FBP+UNet (data-based)
PSNR	21.96 ± 2.25	32.57 ± 3.25	32.78 ± 3.46	31.06 ± 2.95	$\mathbf{33.20} \pm 3.55$	33.75 ± 3.58
SSIM	0.531 ± 0.097	0.803 ± 0.146	0.801 ± 0.151	0.779 ± 0.142	$\mathbf{0.811} \pm 0.151$	0.820 ± 0.140
Runtime	0.02s	1770.89s	127.21s	53.47s	485.93s	0.53s


HR
LR
DIP + TV
EPLL
WPP
patchNR
ACNN
(supervised)

	bicubic (not shown)	DPIR (not shown)	DIP+TV	EPLL	WPP	patchNR	ACNN (data-based)
PSNR	25.63 ± 0.56	27.78 ± 0.53	27.99 ± 0.54	28.11 ± 0.55	27.80 ± 0.37	**28.53** ± 0.49	28.89 ± 0.53
SSIM	0.699 ± 0.012	0.770 ± 0.011	0.764 ± 0.007	0.779 ± 0.010	0.749 ± 0.011	**0.780** ± 0.008	0.804 ± 0.010
Runtime	0.0002s	56.62s	234.00s	60.28s	387.28s	150.79s	0.03s


HR
LR
L^2-TV
DIP+TV
ZSSR
DualSR
patchNR
HR
LR
L^2-TV
DIP+TV
ZSSR
DualSR
patchNR

	L^2-TV	DIP+TV	ZSSR	DualSR	patchNR
PSNR	28.35 ± 3.55	28.44 ± 3.69	28.83 ± 3.57	28.64 ± 3.47	**29.08** ± 3.58
SSIM	0.820 ± 0.072	0.821 ± 0.087	0.834 ± 0.066	0.829 ± 0.061	**0.846** ± 0.061
Runtime	13.12s	171.51s	56.64s	53.47s	132.36s



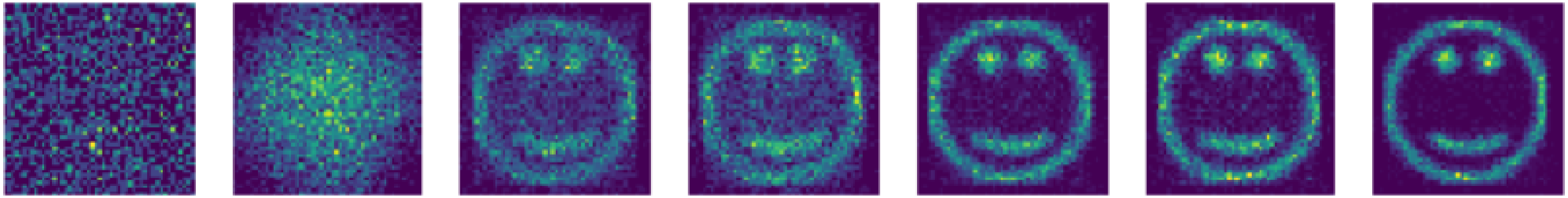
3
2
1
0
−1
−2
−3
−2
0
2
$\mathcal{T}$

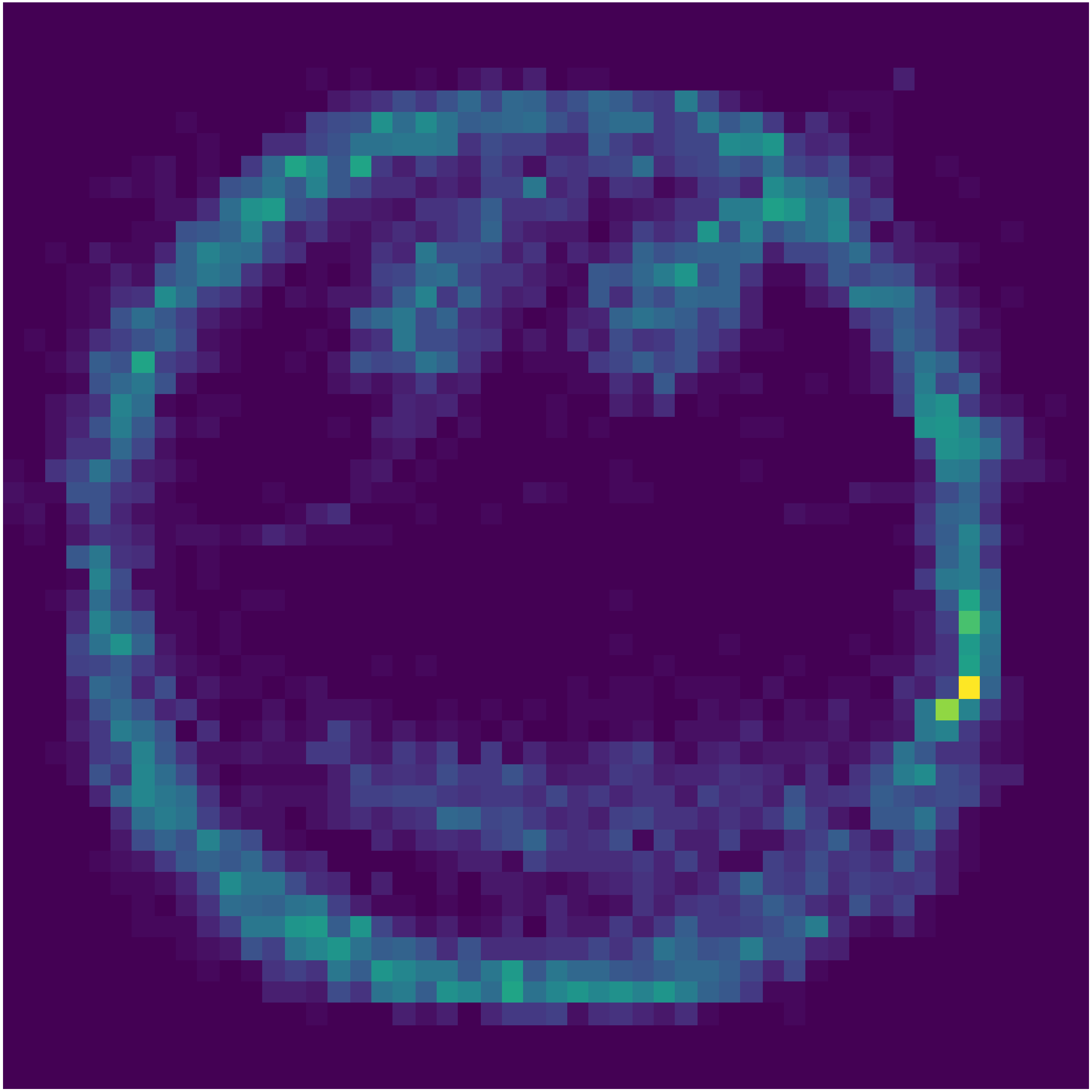
Berlin Mathematics Research Center

MATH+

Funded under Germany's Excellence Strategy by

DFG Deutsche Forschungsgemeinschaft





Discrete Measures

$$\begin{array}{c|ccc} & \nu_1 & \dots & \nu_n \\ \hline \mu_1 & & & \\ \vdots & & \pi_{ij} & \\ \mu_n & & & \end{array}$$

$$k_{ij} := \frac{\pi_{ij}}{\mu_i} \longrightarrow K := (k_{ij})$$

$$\pi = \sum_{ij} \pi_{ij} \, \delta_{(x_i, y_j)}$$

$$\pi \mathbb{1} = \mu$$

$$\pi^T \mathbb{1} = \nu$$

Markov kernel

$$K(x_i, \cdot) = \sum_{\ell} k_{i\ell} \, \delta_{y_\ell}$$

$$\|$$

$$P_{Y \mid X = x_i} \longrightarrow P_{Y \mid X = x_i}(y_j) = k_{ij}$$

$$K \mathbb{1} = \mathbb{1} \longrightarrow \sum_j \frac{\pi_{ij}}{\mu_i} = \frac{\mu_i}{\mu_i} = 1$$

$$K^T \mu = \nu \longrightarrow \sum_i \frac{\pi_{ij}}{\mu_i} \mu_i = \nu_j$$

$$\|$$

P stochastic matrix $(P \geq 0, \; P^T \mathbb{1} = \mathbb{1})$

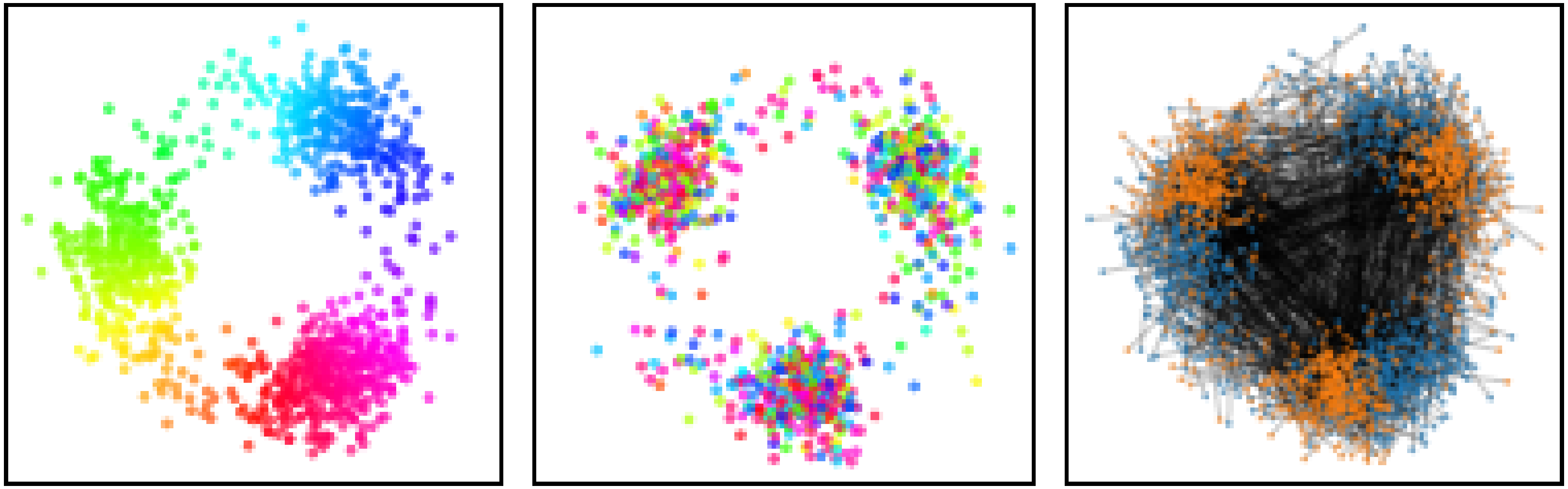
$$\mu \times K(x_i, y_j) = \pi_{ij}$$

$$\mu \times K = \pi$$

Frobenius - Perron: P pos. diagonals, irred.

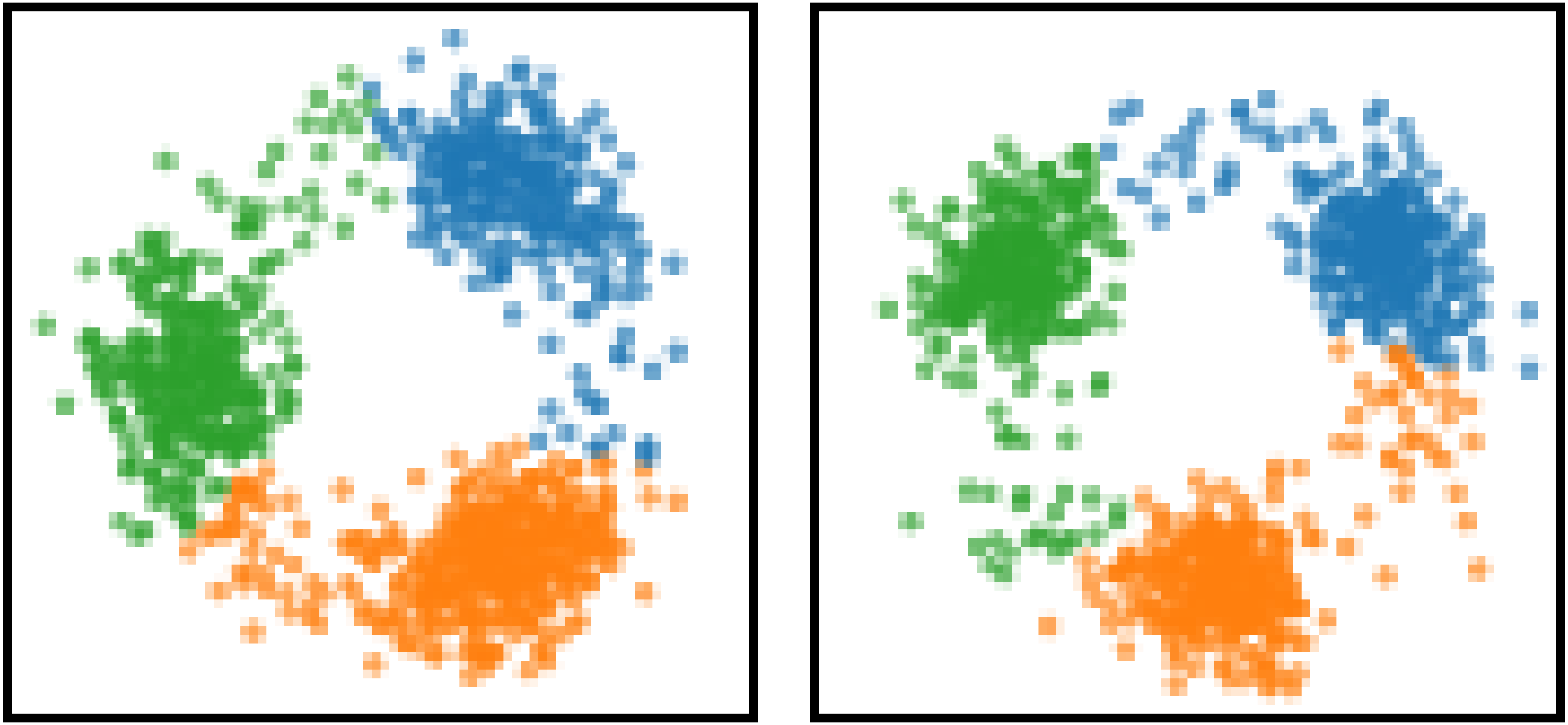
$\rightarrow$ P has eigenvector e to largest (simple) eigenvector 1

$$\lim_{n \to \infty} P^n x = e, \quad \|x\|_1 = \|e\|_1$$



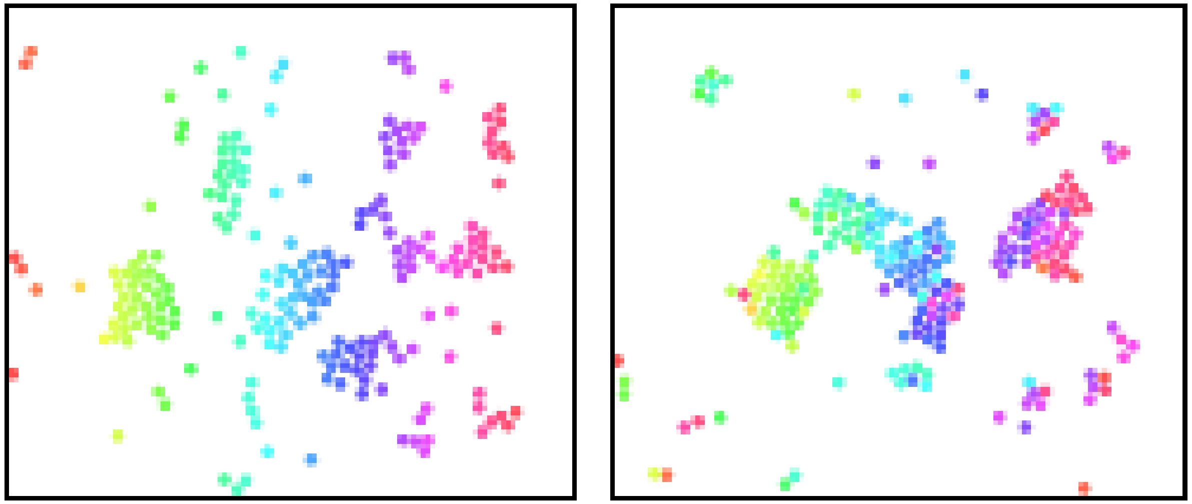
(a) Data at time $t = 0$. (b) Data at time $t = 1$. (c) Movement of particles.

Figure 7: Particles moving in a potential with circular driving force. The color scheme illustrates the particle mixing.



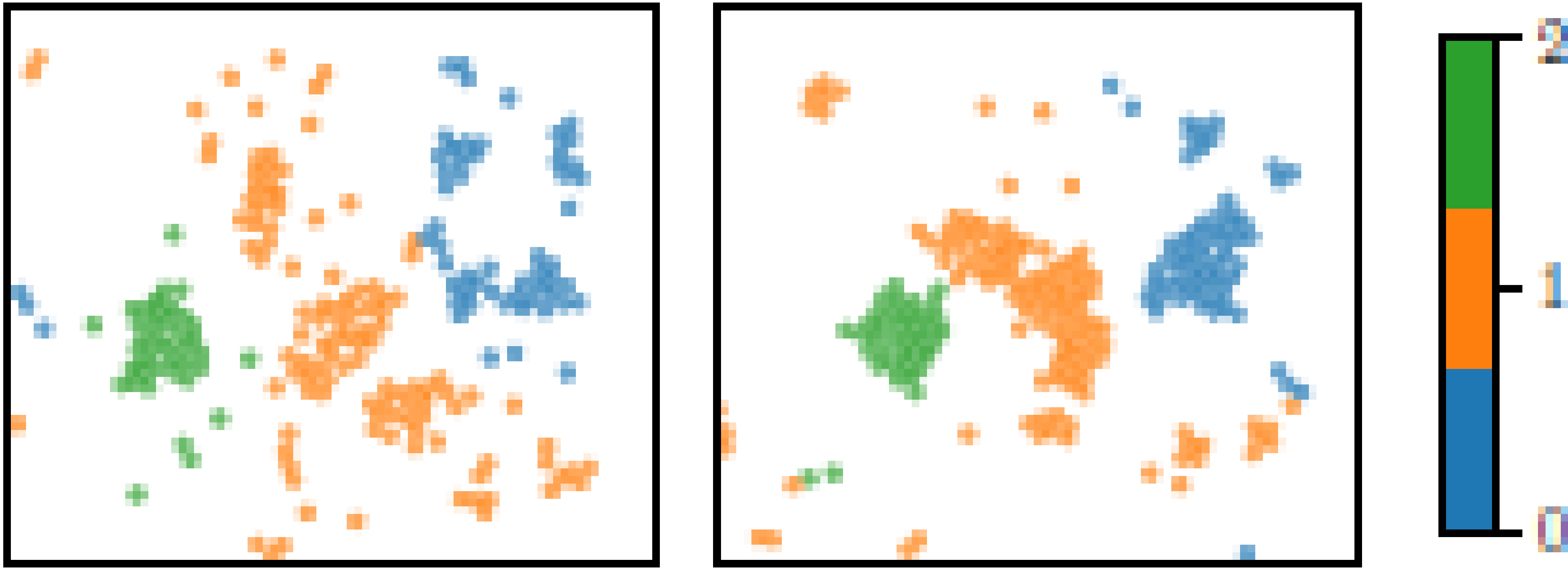
(a) Hard, $t = 0$.

(b) Hard, $t = 1$.



(a) Data at time $t = 0$. (b) Data at time $t = 1$.

Figure 10: Particles moving in a potential with circular driving force. Again, the color scheme illustrates the particle mixing.



(a) Hard, $t = 0$. (b) Hard, $t = 1$.

$$
\begin{tikzcd}
Z & & & \\
X_0 \arrow[r, "\tau_1"] & X_1 \arrow[r, "\tau_2"] & \cdots \arrow[r, "\tau_T"] & X_T \\
Y_0 & Y_1 \arrow[l, "\tau_1^{-1}"'] & \cdots \arrow[l, "\tau_2^{-1}"'] & Y_T \arrow[l, "\tau_T^{-1}"'] \\
 & & & X
\end{tikzcd}
$$

- $$\int_{A\times B} f(x,y)\, dP_{(X,Y)} = \int_A \int_B f(x,y)\, dP_Y(y)\, dP_{X|Y=y}(x)$$

(1) $$f := \frac{dP_{(Y_0,\dots,Y_T)}}{dP_{(X_0,\dots,X_T)}} \overset{!}{=} \frac{p_{Y_T}(x_T)}{p_{X_T}(x_T)} \prod_{t=1}^{T} \underbrace{f_t(x_{t-1}, x_t)}_{\frac{dP_{Y_{t-1}|Y_t=x_t}}{dP_{X_{t-1}|X_t=x_t}}}$$

(2) $$KL(\mu, \upsilon) = \int \log \underbrace{\frac{d\mu}{d\upsilon}}_{\text{Radon-Nikodym derivative}} d\mu(x) = \mathbb{E}_{x\sim\mu}\left[\log \frac{d\mu}{d\upsilon}\right]$$

$$KL(P_{(Y_0,\dots,Y_T)}, P_{(X_0,\dots,X_T)}) = \mathbb{E}_{(x_0\dots x_T)\sim P_{(Y_0,\dots,Y_T)}}\left[\log f(x_1,\dots,x_T)\right]$$

- (1) can be shown by induction using •

Case $T=1$: f rhs in (1)

To show : $f\, dP_{(X_0,X_1)} = dP_{(Y_0,Y_1)}$

$$\int_{A\times B} \underline{\frac{p_{Y_1}(x_1)}{p_{X_1}(x_1)}}\; \underline{\frac{dP_{Y_0|Y_1=x_1}}{dP_{X_0|X_1=x_1}}}\; dP_{(X_0,X_1)}(x_0,x_1)$$

$$= \int_A \int_B \quad -\text{"}- \quad \underline{dP_{X_1}(x_1)}\; \underline{dP_{X_0|X_1=x_1}(x_0)}$$

$$= \int_A \int_B dP_{Y_1}(x_1)\, dP_{Y_0|Y_1=x_1}(x_0)$$

$$= \int_{A\times B} dP_{(Y_0,Y_1)}(x_0,x_1)$$

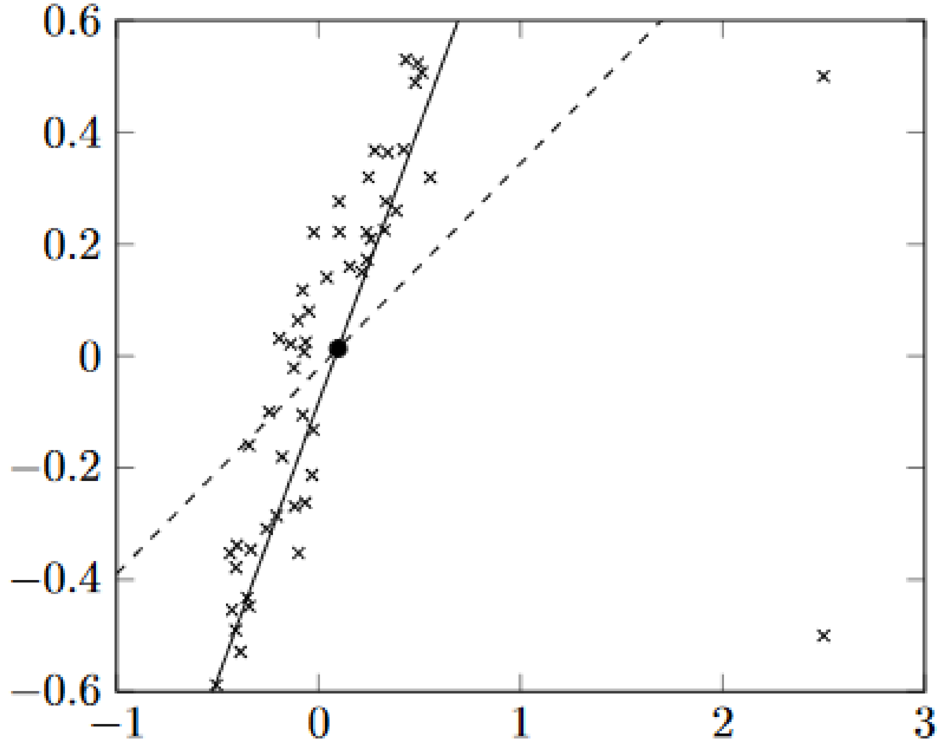


Figure 1: Demonstration of the sensitivity of standard PCA to outliers. The data set consists of 50 points close to a line through the origin and two outliers. The subspace indicated by the dashed line is the result of standard PCA (2), while the solid one corresponds to (4). In both cases the offset b was chosen as the mean (black dot).

1

Markov chain (2 layer):

$$\begin{array}{ccc} X_0 & \xrightarrow[K = P_{X_1|X_0}]{} & X_1 \\ \| & & \wr\wr \\ Z & & X \\ \wr\wr & & \| \\ Y_0 & \xleftarrow[R = P_{Y_0|Y_1}]{} & Y_1 \end{array}$$

Bayes: $p(x|z) = \dfrac{p(z|x)\, p_X(x)}{p_Z(z)}$

$\underbrace{\mu \times K(A,B)}_{} = \int_A K(x,B)\, d\mu(x)$

$P_X \times P_{Y|X=\cdot}(\cdot) = P_{(X,Y)}$

- Aim to learn θ such that

$$p_X(x) \approx \int_{\mathbb{R}^n} p_\theta(x|z)\, p(z)\, dz$$

$$P_X(A) \approx \int_{\mathbb{R}^n} K(z,A)\, dP_Z(z)$$

- sample for P_X: ① sample from Z; ② sample from $K(z,\cdot)$

LOSS function: Idea: $E_{x\sim P_X}[\overbrace{\log p_\theta(x)}^{\text{evidence}}] \longrightarrow \max$

$$p_\theta(x) = \int_{\mathbb{R}^n} p_\theta(x|z)\, p_Z(z)\, dz$$

☹ How to compute

Approximation of evidence from below: evidence lower bound (**ELBO**)

$$\log(p_\theta(x)) = E_{z\sim q_\phi(\cdot|x)}\left[\log \frac{p_\theta(x)\, p_\theta(z|x)}{p_\theta(z|x)}\right]$$

$$= E_{z\sim q_\phi(\cdot|x)}\left[\log \frac{p_\theta(x)\, p_\theta(z|x)}{q_\phi(z|x)}\right] + E_{z\sim q_\phi(\cdot|x)}\left[\log \frac{q_\phi(z|x)}{p_\theta(z|x)}\right]$$

$$\overset{\text{Bayes}}{=} E_{z\sim q_\phi(\cdot|x)}\left[\log \frac{p_\theta(x|z)\, p_Z(z)}{q_\phi(z|x)}\right] + \underbrace{KL(q_\phi(\cdot|x), p_\theta(\cdot|x))}_{\geq 0}$$

$$\geq \mathbf{E_{z\sim q_\phi(\cdot|x)}\left[\log \frac{p_\theta(x|z)\, p_Z(z)}{q_\phi(z|x)}\right]}$$

Relation to Markov chains: $L_{SNF}(\theta,\varphi) = KL(P_{(Y_0,Y_1)}, P_{(X_0,X_1)})$

$$= E_{\substack{(x_0,x_1)\sim P_{(Y_0,Y_1)} \\ \downarrow\ \downarrow\qquad\ \downarrow \\ z\ \ x\qquad\ \ x}}\left[\log \frac{p_{Y_1}(x_1)}{p_{X_1}(x_1)} f_1(x_0,x_1)\right]$$

$$L_{SNF}(\theta,\varphi) = E_{(z,x)\sim P_{(Y_0,X)}}\left[\log \frac{p_X(x)}{p_{X_1}(x)} f_1(z,x)\right]$$

$$f_1(z,x) = \frac{d\,P_{Y_0|X=x}}{d\,P_{Z|X_1=x}}(z) \overset{\text{kernel}}{=} \frac{q_\phi(z|x)}{p_{Z|X_1=x}} \overset{\text{Bayes}}{=} \frac{q_\phi(z|x)}{p_{X_1|Z=z}(z)}\,\frac{p_{X_1}(x)}{p_Z(z)}$$

$$= \frac{q_\phi(z|x)}{p_\theta(x|z)}\,\frac{p_{X_1}(x)}{p_Z(z)}$$

Thus

$$L_{SNF}(\theta,\varphi) = E_{(z,x)\sim P_{(Y_0,X)}}\left[\log \frac{q_\phi(z|x)}{p_\theta(x|z)}\,\frac{p_X(x)}{p_Z(z)}\right]$$

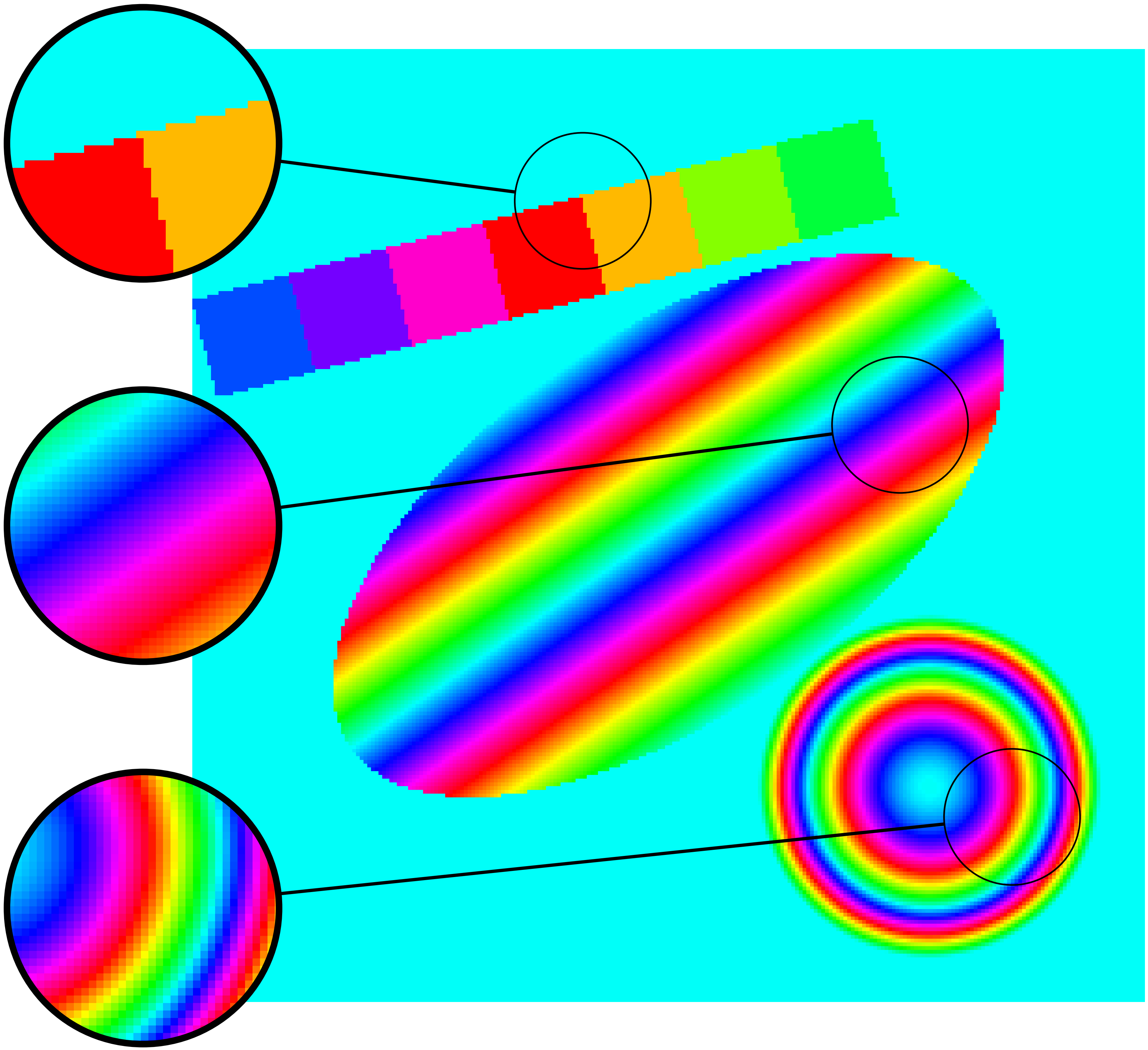
$$P_{(Y_0,X)} = P_X \times \underbrace{P_{Y_0|X}}_{R(x,\cdot)} \quad \text{(Markov kernel property)}$$

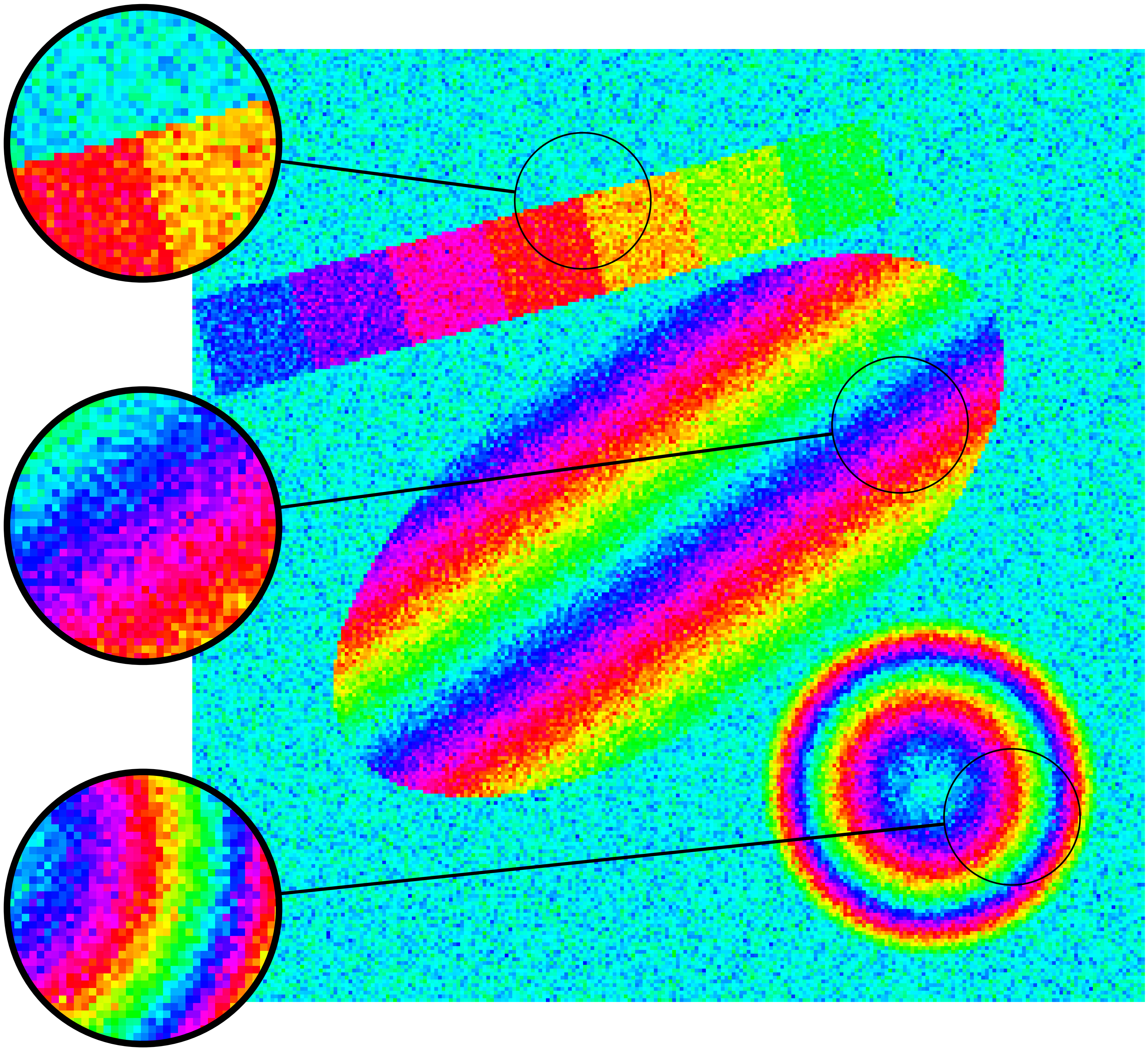
$$= \mathbb{E}_{x\sim P_X}\left(\mathbb{E}_{z\sim q_\phi(\cdot,x)}\left[\log \frac{q_\phi(z|x)}{p_\theta(x|z)}\,\frac{p_X(x)}{p_Z(x)}\right]\right.$$

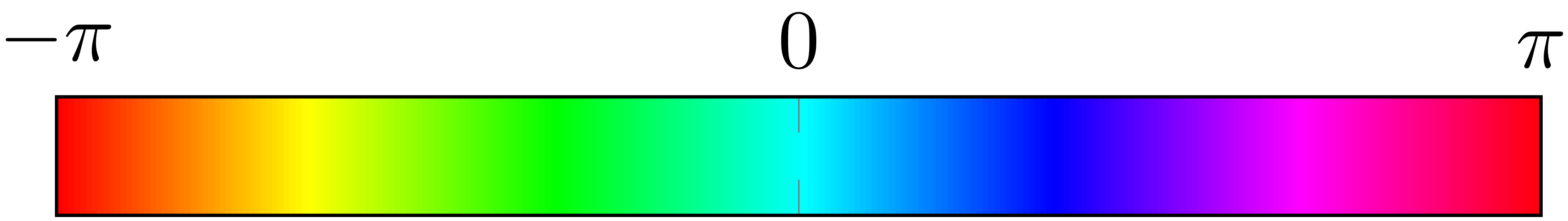
$$= \mathbb{E}_{x\sim P_X}\underbrace{\left(\mathbb{E}_{z\sim q_\phi(\cdot,x)}\left[\log \frac{q_\phi(z|x)}{p_\theta(x|z)\,p_Z(x)}\right]\right)}_{-\,ELBO}$$

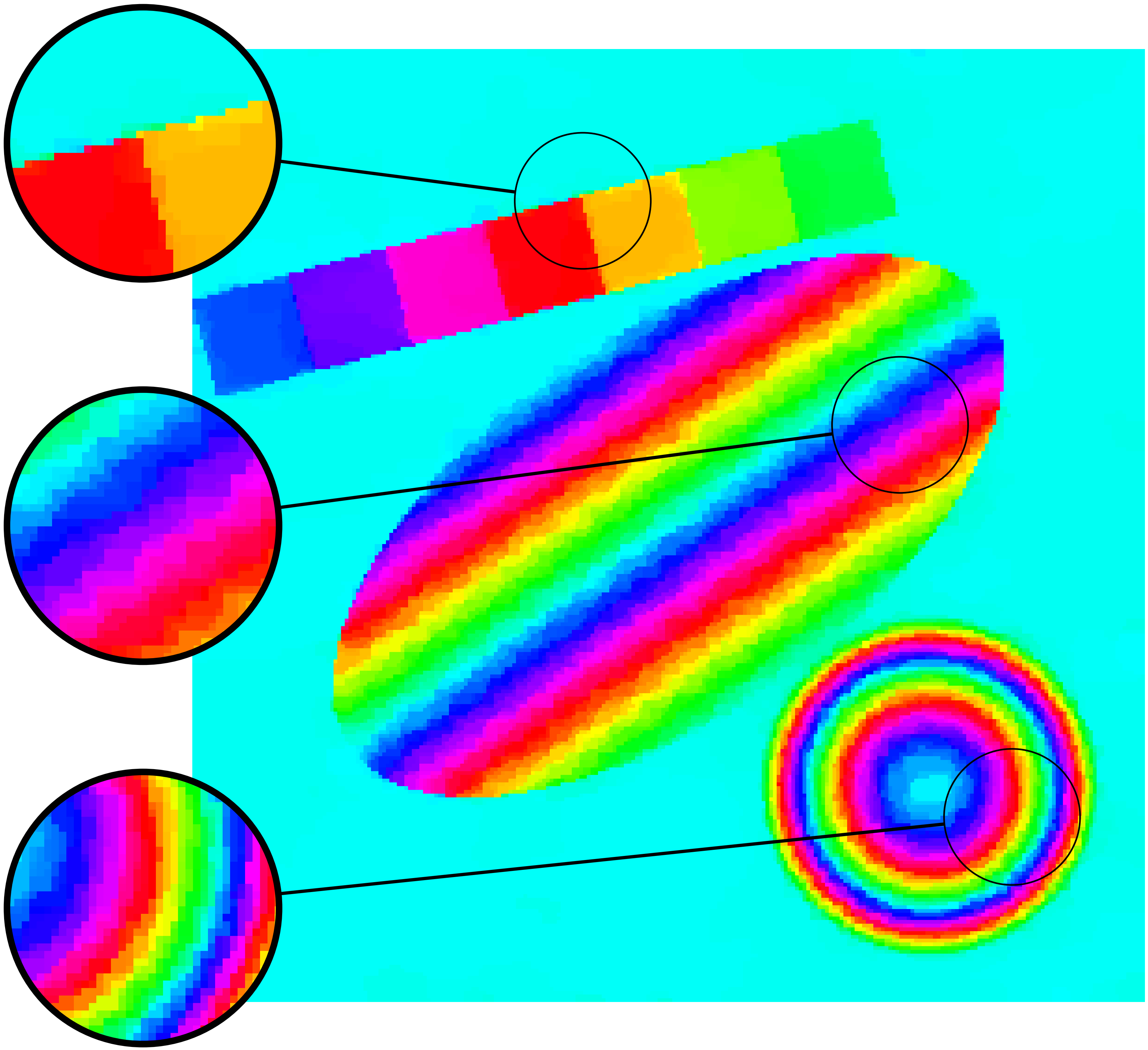
$$\underbrace{-\,E_{x\sim P_X}\left[p_X(x)\right]}_{\text{const}}$$

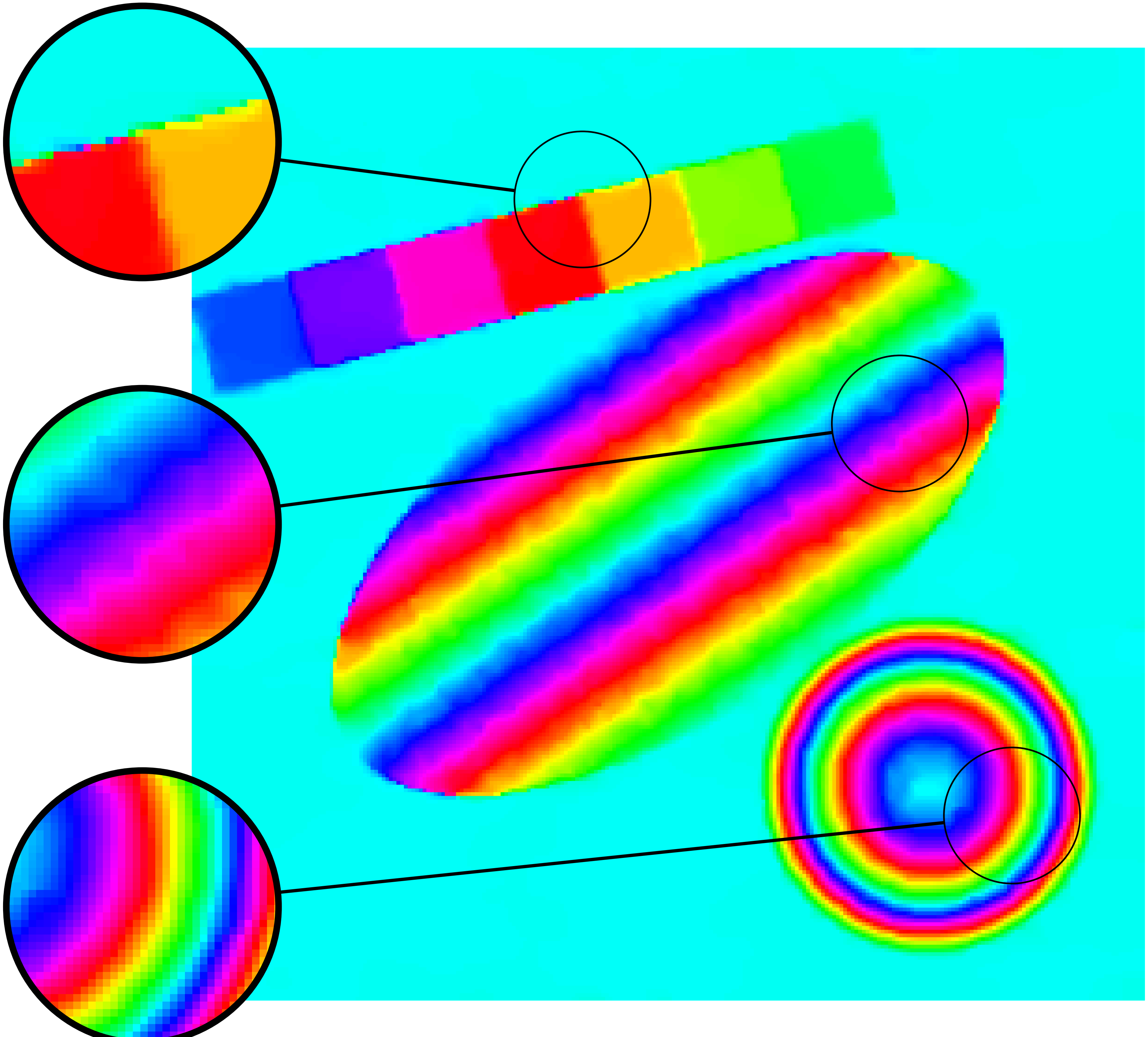
$$\Longrightarrow \quad \boxed{\underset{\theta,\phi}{\operatorname{argmin}}\, L_{SNF}(\theta,\varphi) = \underset{\theta,\phi}{\operatorname{argmax}}\, E_{x\sim P_X}(ELBO))}$$

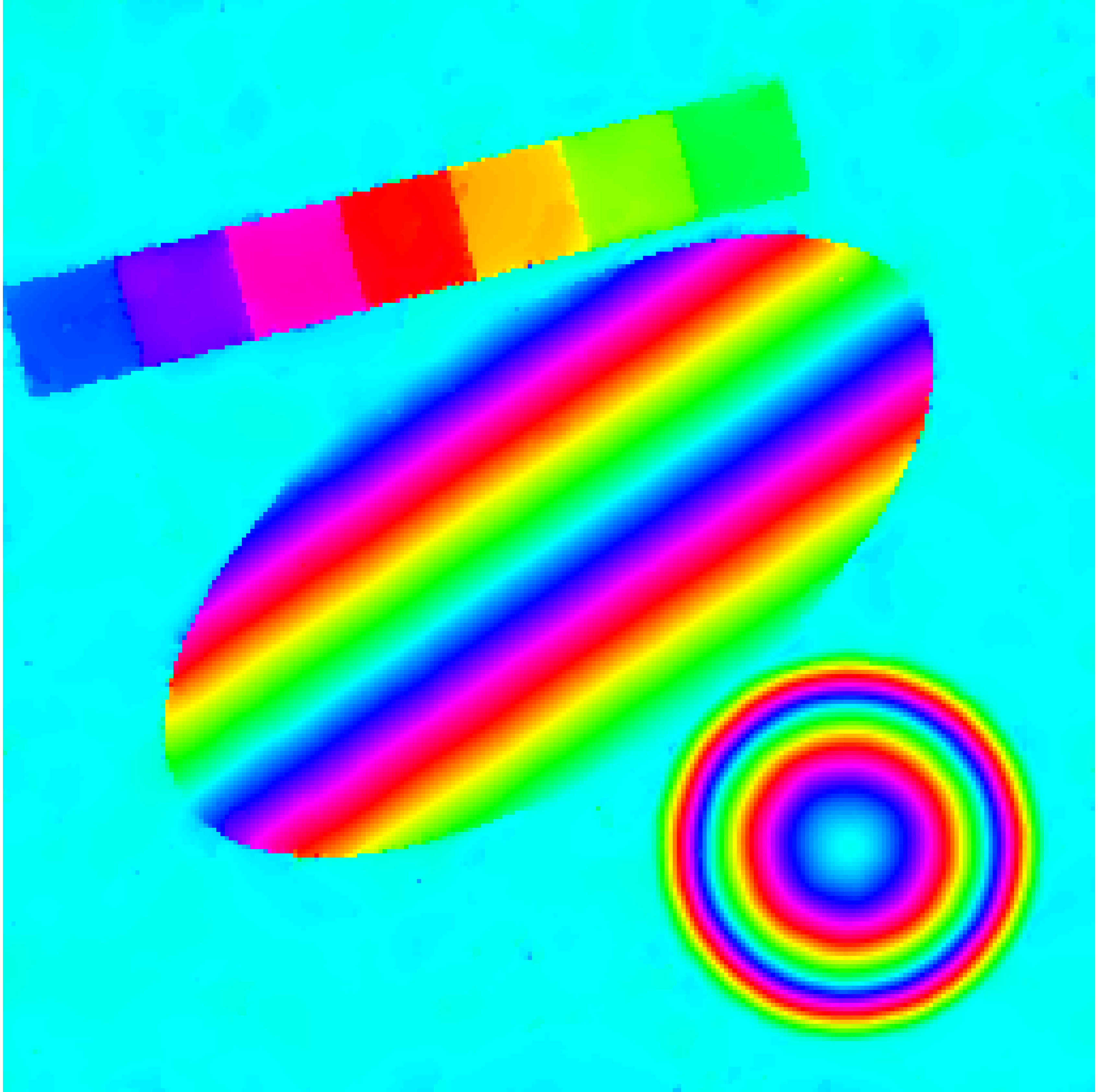


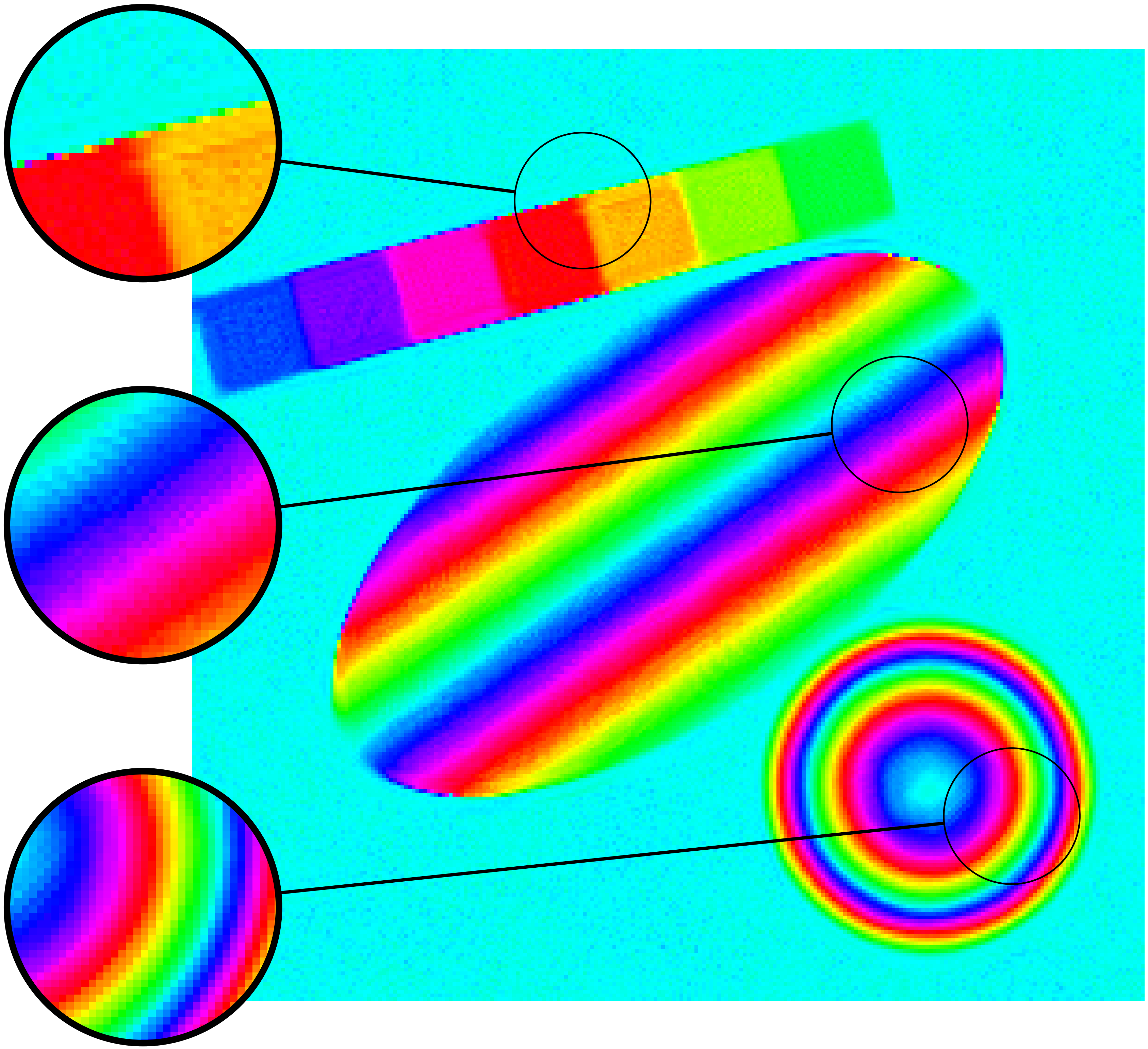


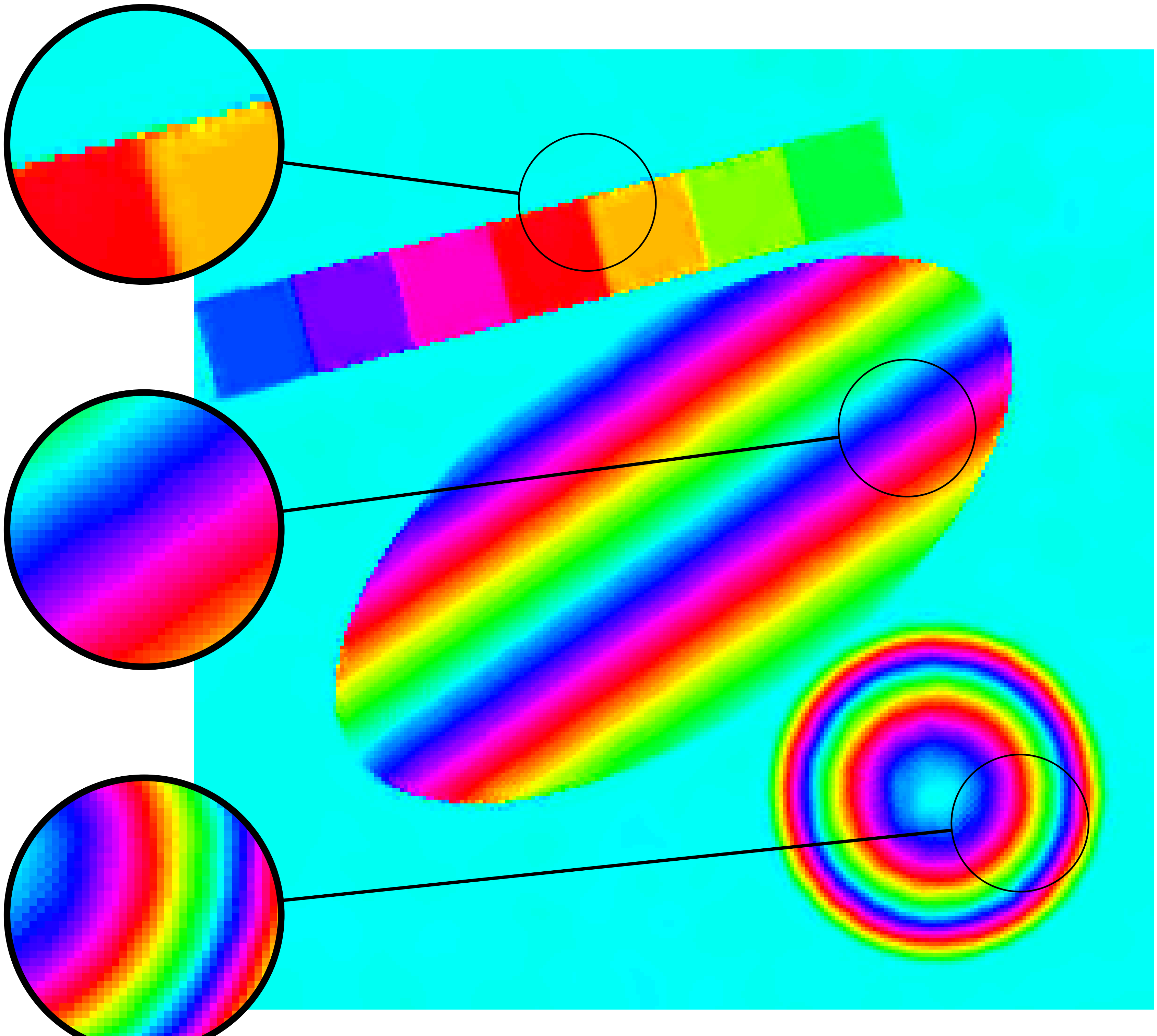

$-\pi$
0
π

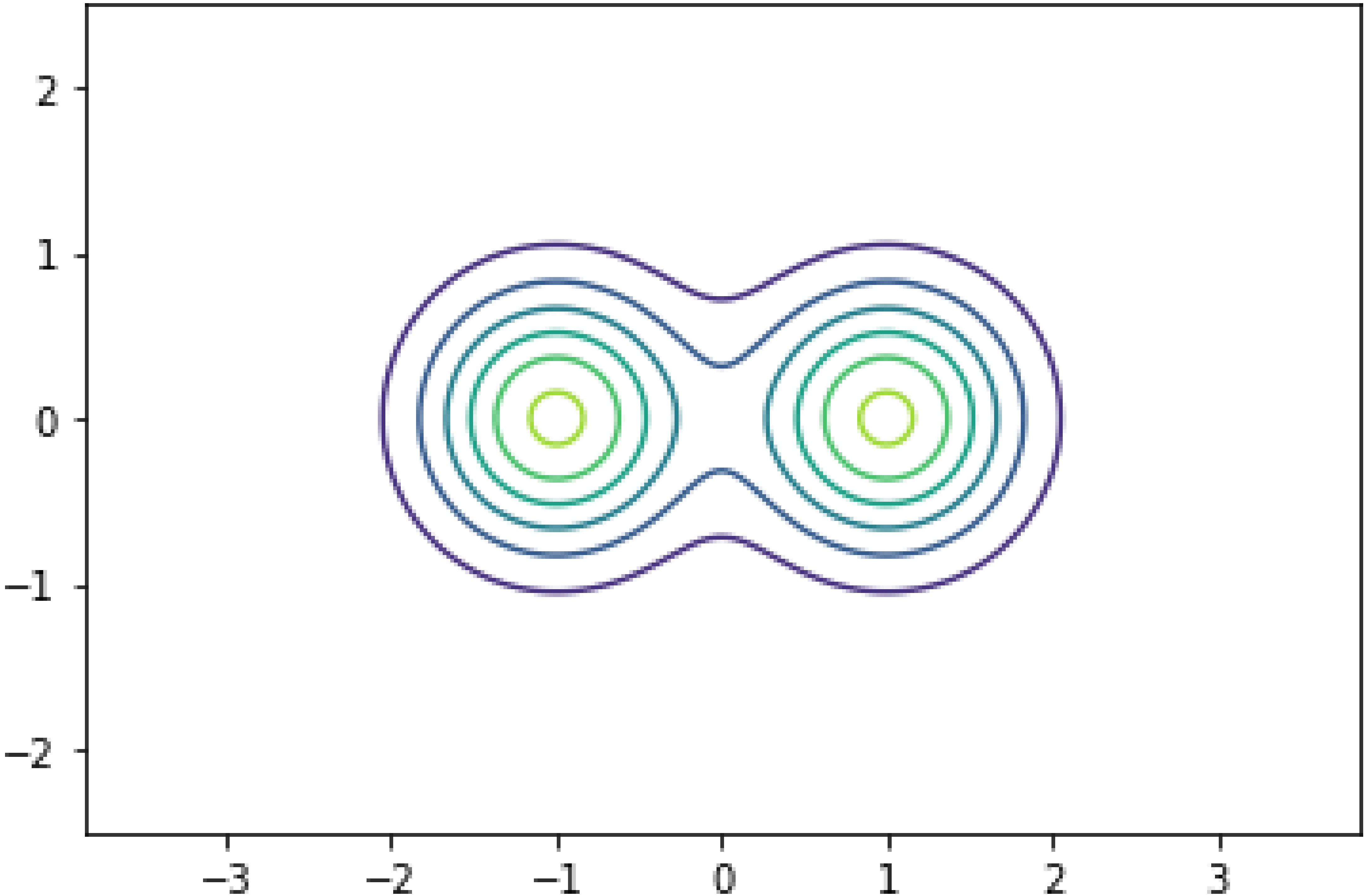




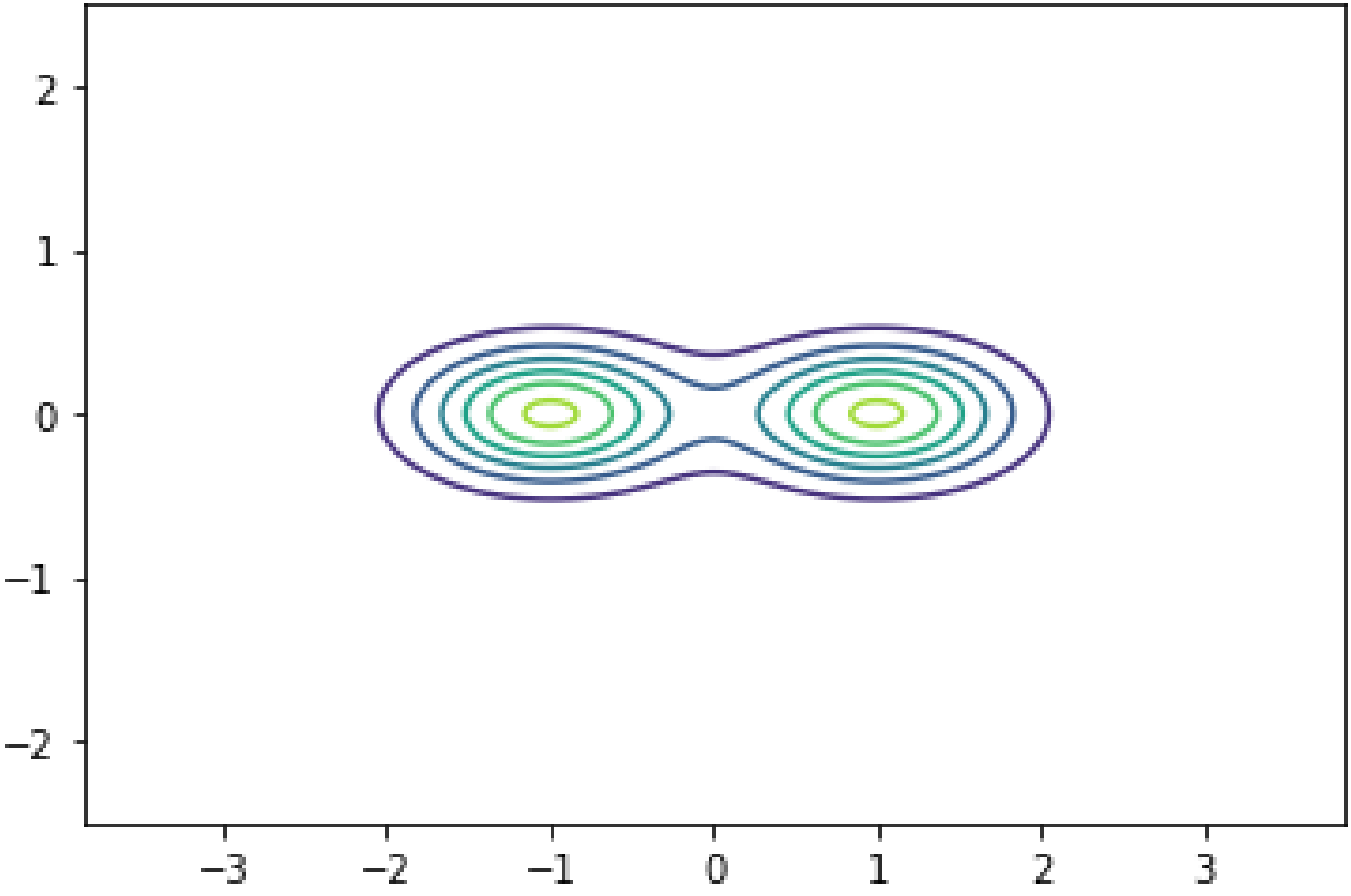








2
1
0
−1
−2
−3
−2
−1
0
1
2
3



2
1
0
−1
−2
−3
−2
−1
0
1
2
3