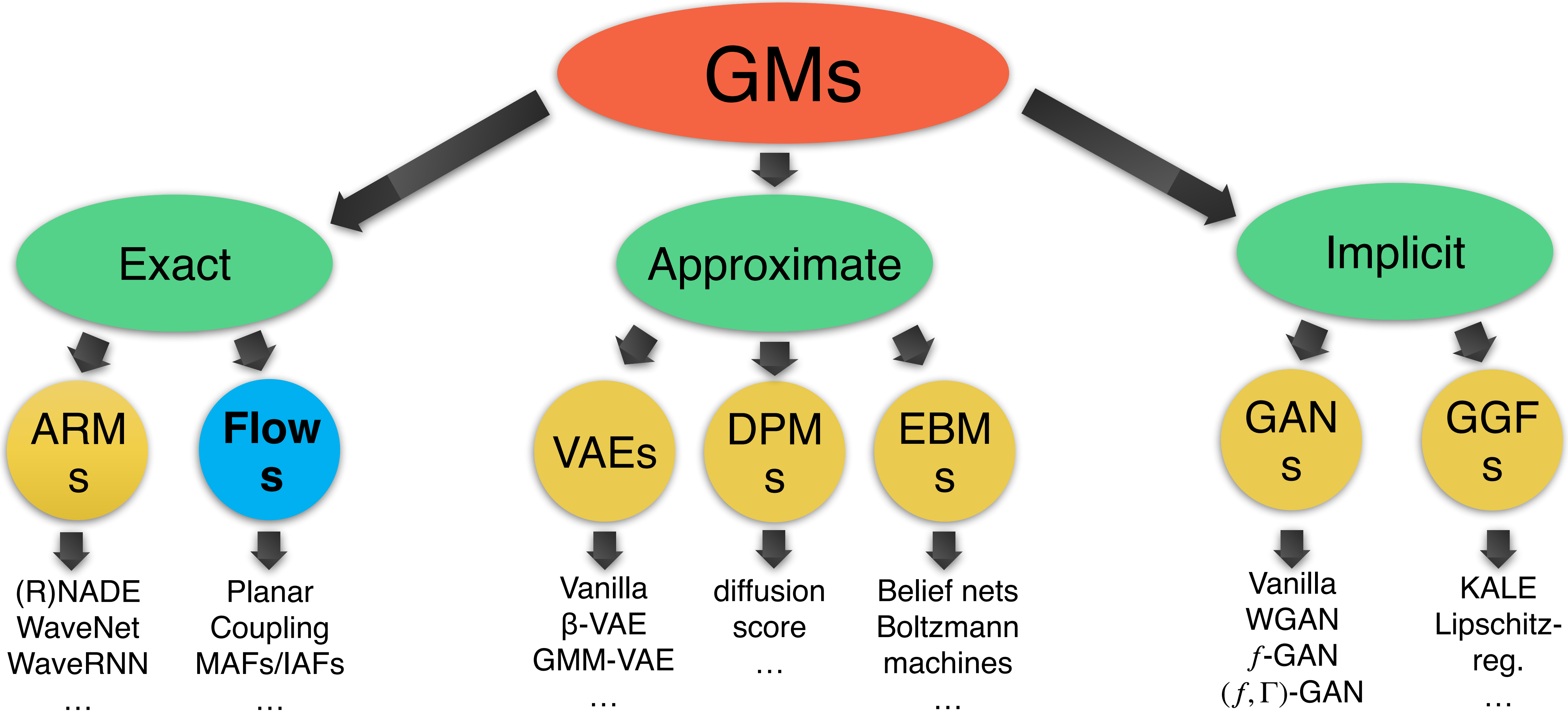
Introduction to Deep Generative Modeling

Lecture #8

HY-673 – Computer Science Dep., University of Crete
Professors: Yannis Pantazis, Yannis Stylianou
Teaching Assistant: Michail Raptakis

Taxonomy of GMs



Desiderata in Deep GMs

- **Efficient sampling** from $p_\theta(x)$:

$$z \longrightarrow \boxed{f_\theta} \longrightarrow x \sim p_\theta(x) \approx p_d(x)$$

$\hookrightarrow$ z should have simple (base/prior) distribution (e.g., isotropic Gaussian).

$\hookrightarrow$ Great advantage over previous sampling approaches (typically based on Markov Chain Monte Carlo - MCMC methods).

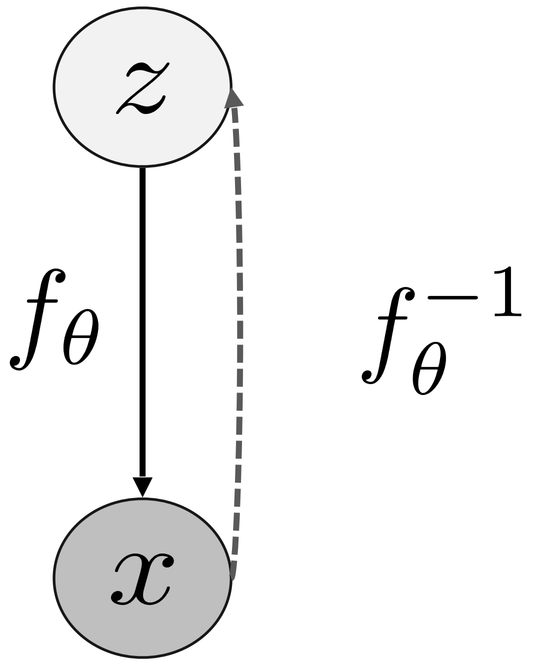
- For *exact MLE-based GMs*, **easy to compute** $p_\theta(x)$.

$\hookrightarrow$ Again, z should have simple (prior) distribution.

$\hookrightarrow$ $f_\theta(z)$ should have some structure that will result in tractable $p_\theta(x)$.

Terminology in Deep GMs

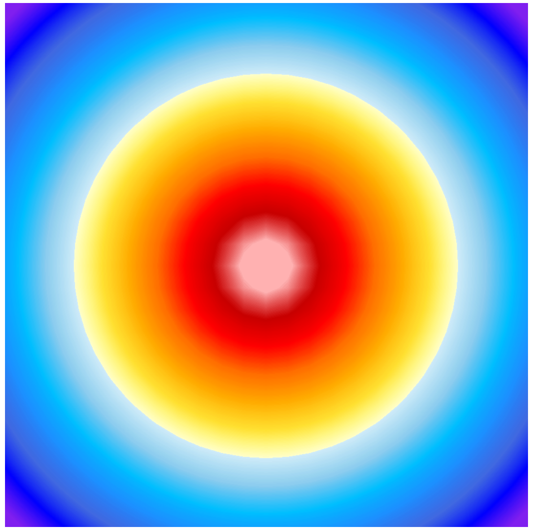
- $f_\theta(x)$: sampling or "coloring" phase — the *decoder* or the *generator*.
- $f_\theta^{-1}(x)$: inference or "normalizing" phase — the *encoder* or the *normalizer*.
- The *encoder* transforms the distribution into independent factors.



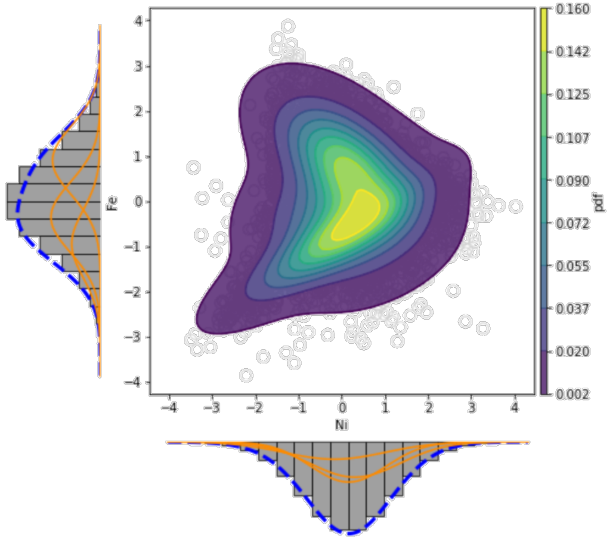
Simple Prior to Complex Data Distributions

- Desirable properties of any model distribution $p_\theta(x)$:
  1. **For Training:** Easy to evaluate, closed form density.
  2. **For Generation:** Easy to sample from.
- Many distributions satisfy these two, e.g., Gaussian, uniform, et al.

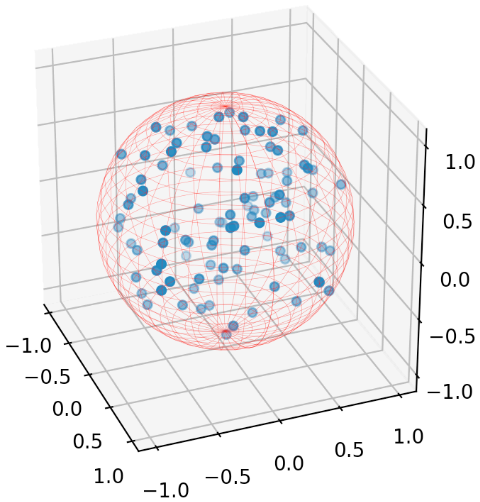
Unit Gaussian



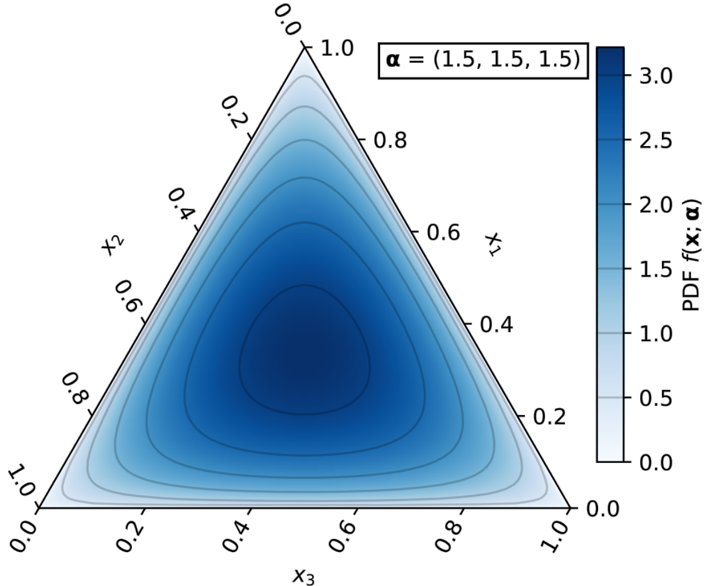
Mixture of Gaussians



Uniform in Sphere



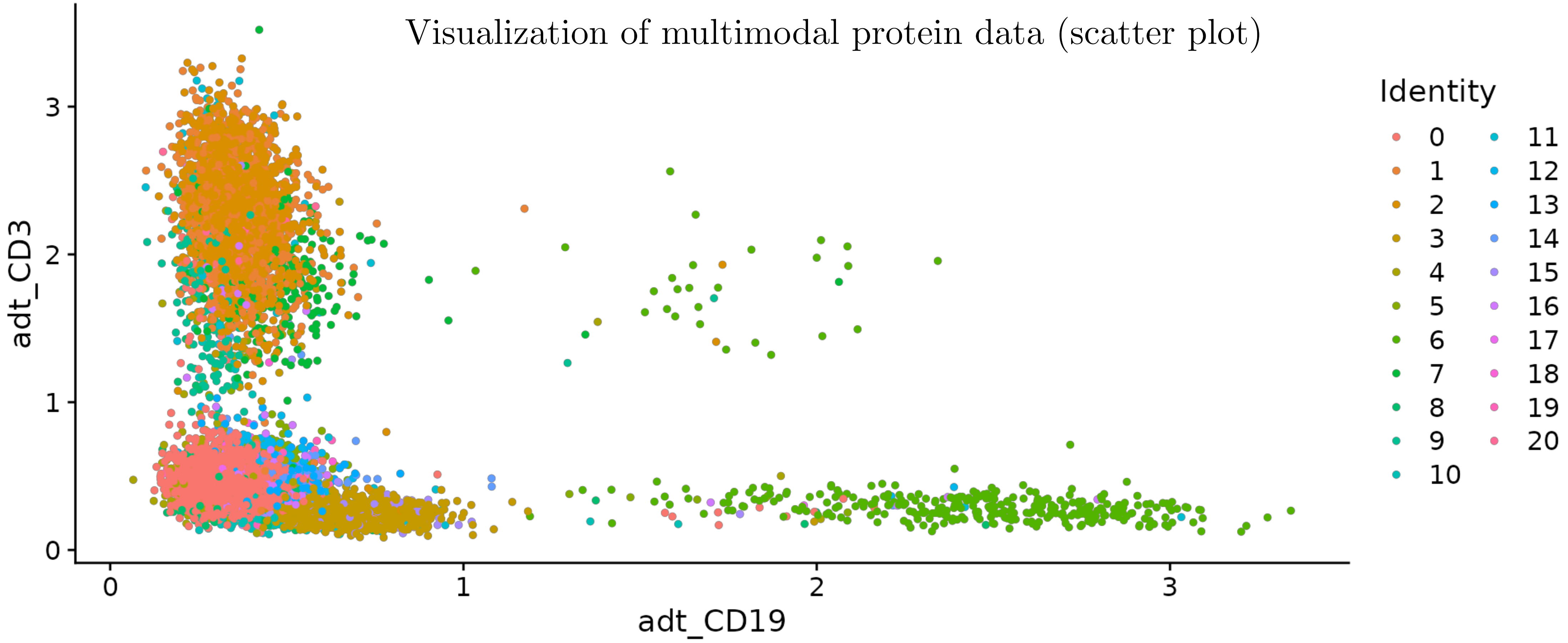
Dirichlet



Simple Prior to Complex Data Distributions

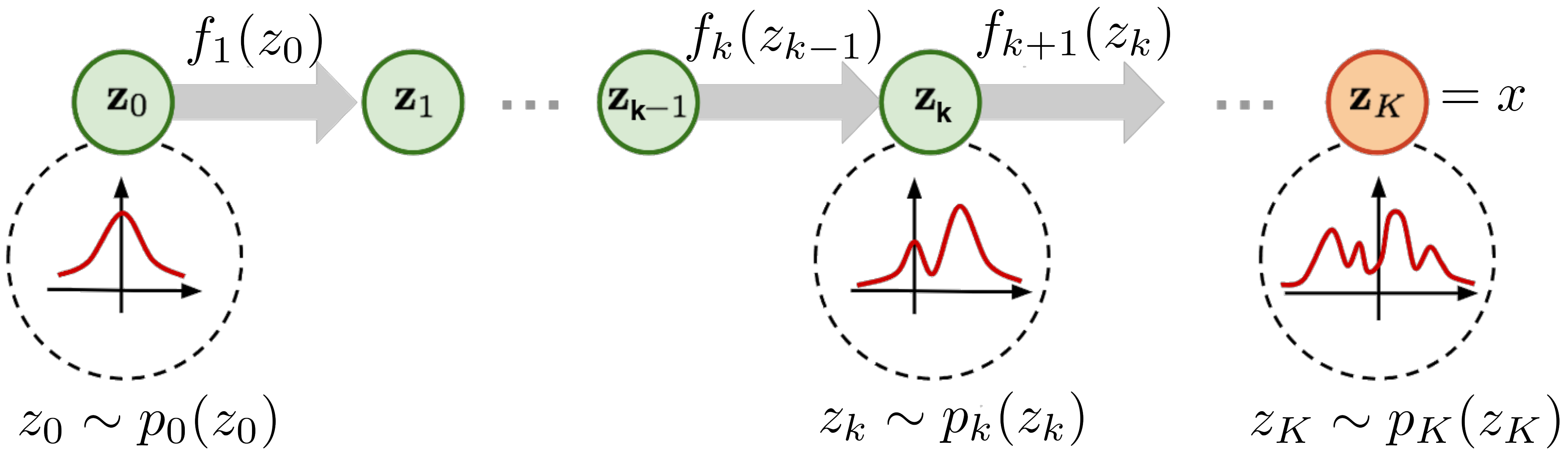
- Unfortunately, real data distributions are usually more complex (multi-modal).

Visualization of multimodal protein data (scatter plot)



What is a Flow Model?

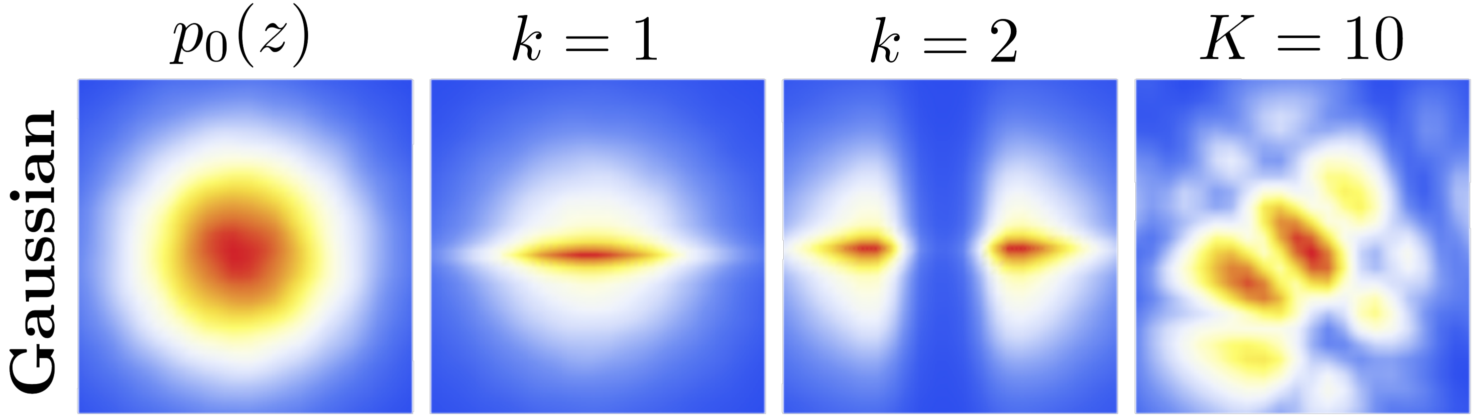
- **Key idea behind <u>flow models</u>:** Map simple distributions (i.e., easy to sample and evaluate densities) to complex distributions through a **series of invertible and differentiable transformations**.



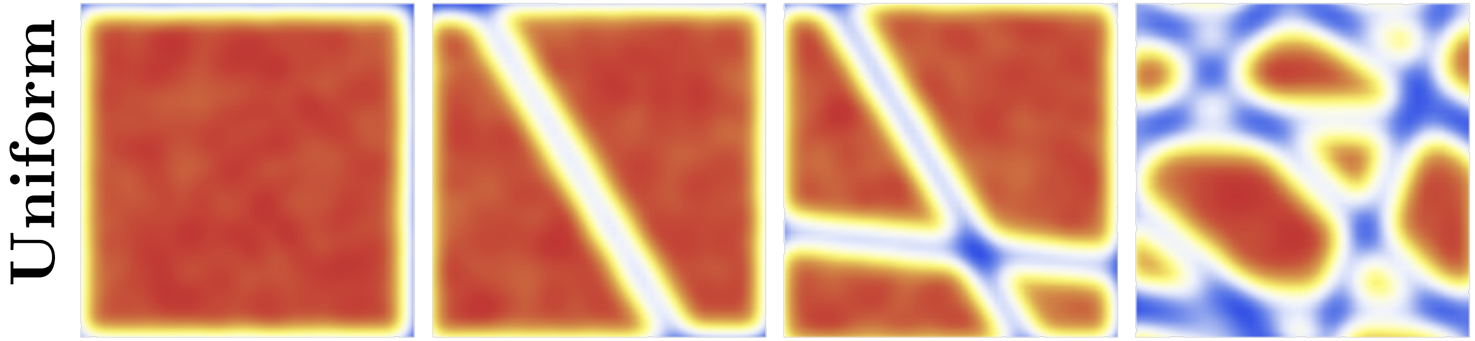
Many small steps adds up to big results.

Planar Flows (Rezende & Mohamed, 2015)

- Base distribution: Gaussian



- Base distribution: Uniform



10 planar transformations can transform a simple distribution into a far more complex one.

Flow models – Basic properties

- The decoder/generator is given by:
$$f_\theta := f_K \circ ... \circ f_k \circ ... \circ f_1.$$
- The encoder/normalizer is given by:
$$f_\theta^{-1} := f_1^{-1} \circ ... \circ f_k^{-1} \circ ... \circ f_K^{-1}.$$
- What about f_k's?
  - ↪ They have to be invertible. $\Rightarrow \dim(z) = \dim(x) = d$.
  - ↪ The output has to have tractable and fast-to-compute probability density function.
  - ↪ The *change of variables formula for random variables* imples that f_k's need to have easy-to-compute Jacobian and easy-to-compute determinant.
- How expressive/powerful is a flow model?
Answer: They are universal approximators of the density.

Example: $f_k := f_{\theta_k}$, with $\theta := \{\theta_1, ..., \theta_K\}$.

Flow models – Math Terminology

- *Bijection:* An invertible transformation.

- *Diffeomorphism:* A bijection that is differentiable.

- *Flow:* A family of diffeomorphisms f_t indexed by a real number t such that $t = 0$ indexes the identity function and $t_1 + t_2$ indexes the composition $f_{t_1} \circ f_{t_2}$.

- $p_\theta(x)$: *pushforward* of the base distribution (notation: $p_\theta = f_* p_0$).

Change of Variables Formula

- Let Z be a uniform random variable $\mathcal{U}[0, 2]$ with density $p_Z(z)$. What is $p_Z(1)$?

  Answer: $\frac{1}{2}$, sanity check: $\int_0^2 \frac{1}{2} dx = 1$.

- Let $X = 4Z$, and let $p_X(x)$ be its density. What is $p_X(4)$?

  Answer: $p_X(4) = P(X = 4) = P(4Z = 4) = P(Z = 1) = p_Z(1) = 1/2$. Wrong!

  Answer: Clearly, X is uniform in $[0, 8]$, so $p_X(4) = 1/8$.

  !!! To get the correct result, we need to use the **change of variables formula**.

Change of Variables Formula

- **Change of Variables (1D case)**: If $X = f(Z)$ and $f(\cdot)$ is monotone with inverse $Z = f^{-1}(X) = h(X)$, then:

$$p_X(x) = p_Z(h(x)) \times |\frac{d}{dx}h(x)|.$$

- More interesting example: If $X = f(Z) = \exp(Z)$ and $Z \sim \mathcal{U}[0, 2]$, what is $p_X(x)$?

  Answer: Note that $Z = h(X) = \log(X)$, thus,

  $$p_X(x) = p_Z(log(x)) \times |h'(x)| = \frac{1}{2x}, \text{ for } x \in [\exp(0), \exp(2)].$$

  – Note that the "shape" of $p_X(x)$ is different (and, essentially, more complex) from that of the base distribution $p_Z(z)$.

Change of Variables Formula

- **Change of Variables (1D case)**: If $X = f(Z)$ and $f(\cdot)$ is monotone with inverse $Z = f^{-1}(X) = h(X)$, then:

$$p_X(x) = p_Z(h(x)) \times |\tfrac{d}{dx} h(x)|.$$

- *Proof sketch:* Assuming $f(\cdot)$ is monotonic:

$$F_X(z) = P(X \leq x) = P(f(Z) \leq x) = P(Z \leq h(x)) = F_Z(h(x)).$$

Differentiating both sides:

$$p_X(x) = \frac{dF_X(x)}{dx} = \frac{dF_Z(h(x))}{dx} = p_Z(h(x)) \frac{dh(x)}{dx}.$$

Change of Variables Formula

- **Change of Variables (1D case)**: If $X = f(Z)$ and $f(\cdot)$ is monotone with inverse $Z = f^{-1}(X) = h(X)$, then:

$$p_X(x) = p_Z(h(x)) \times \left|\tfrac{d}{dx}h(x)\right|.$$

- Recall from basic calculus that $h'(x) = [f^{-1}]'(x) = \frac{1}{f'(f^{-1}(x))}$.

  So, letting $z = h(x) = f^{-1}(x)$, we can also write:

$$p_X(x) = p_Z(z) \times \left|\frac{1}{f'(z)}\right|.$$

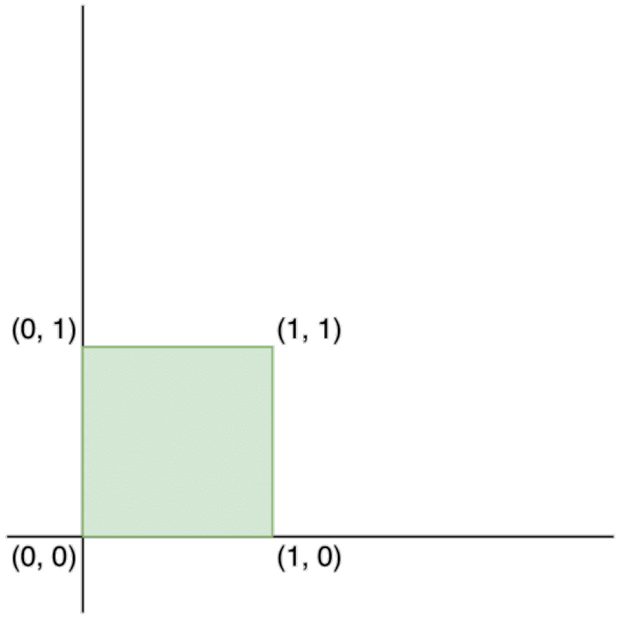
Recall:
$f'(x) \equiv \frac{d}{dx}f(x)$.

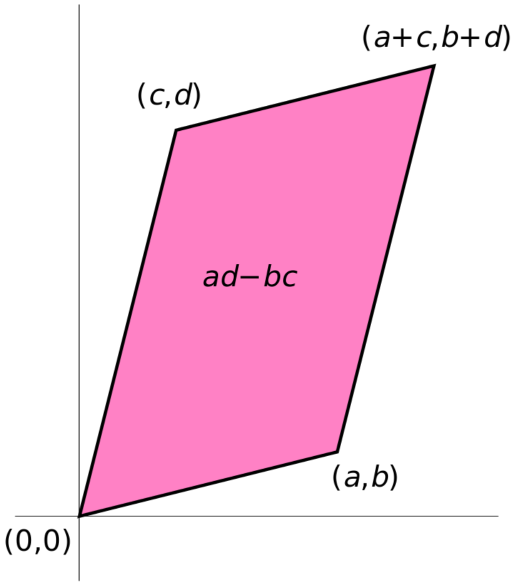
Geometry: Determinants & Volumes

- Let Z be a uniform random vector in $[0,1]^n$. Let $Z = AZ$ for a square invertible matrix A, with inverse $W = A^{-1}$. How is X distributed?

Answer: Geometrically, the matrix A maps the unit hypercube $[0,1]^n$ to a parallelotope. Hypercube and parallelotope are generalizations of square/cube and parallelogram/parallelopipes to higher dimensions.

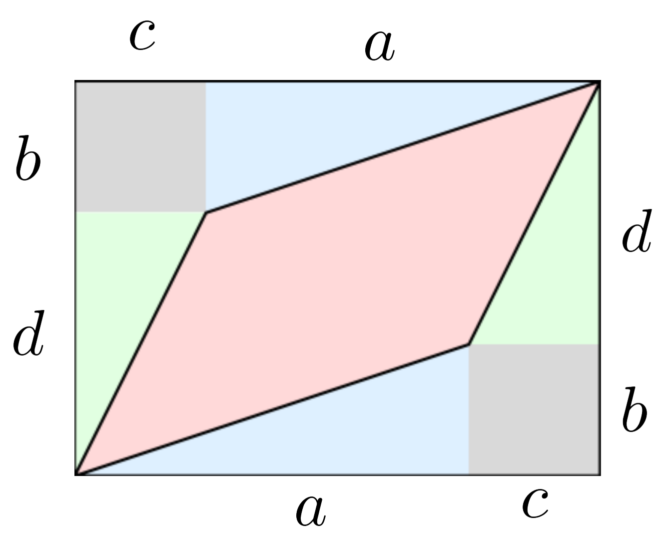
Geometry: Determinants & Volumes





The matrix $A = \begin{pmatrix} \mathrm{a} & \mathrm{c} \\ \mathrm{b} & \mathrm{d} \end{pmatrix}$

maps a unit square to a parallelogram.



$$(a+c)(b+d) - ab - 2bc - cd = ad - bc.$$

- The volume of the parallelotope is equal to the absolute value of the determinant of the matrix A:

$$\det(A) = \det \begin{pmatrix} \mathrm{a} & \mathrm{c} \\ \mathrm{b} & \mathrm{d} \end{pmatrix} = ad - bc.$$

Geometry: Determinants & Volumes

- Let $X = AZ$ for a square invertible matrix A, with inverse $W = A^{-1}$. X is uniformly distributed over the parallelotope of area $|\det(A)|$. Hence:

$$p_X(x) = p_Z(Wx)/|\det(A)| = p_Z(Wx)|\det(W)|,$$

because if $W = A^{-1}$, then $\det(W) = \frac{1}{\det(A)}$.
– Essentially, an extension of the 1D case formula.

- For linear transformations specified via A, change in volume is given by the determinant of A, and for non-linear transformations $f(\cdot)$, the *linearized* change in volume is given by the **determinant of the Jacobian** of $f(\cdot)$.

Generalized Change of Variables

- **Change of Variables (General case):** The mapping between Z and X, given by $f : \mathbb{R}^d \to \mathbb{R}^d$, is invertible such that $X = f(Z)$ and $Z = f^{-1}(X)$.

$$p_X(x) = p_Z\left(f^{-1}(x)\right)\left|\det\left(\frac{\partial f^{-1}(x)}{\partial x}\right)\right|.$$

1. Generalizes the previous 1D case: $p_X(x) = p_Z(h(x))|h'(x)|$
2. x and z need to be continuous and have the same dimension. For example, if $x \in \mathbb{R}^d$, then $z \in \mathbb{R}^d$.
3. For any invertible matrix A, $\det(A^{-1}) = \det(A)^{-1} \Rightarrow$ $p_X(x) = p_Z(z)\left|\det\left(\frac{\partial f(z)}{\partial z}\right)\right|^{-1}$.

https://web.williams.edu/Mathematics/sjmiller/public_html/probabilitylifesaver/supplementalchap_changeofvar.pdf

Two Dimensional Example

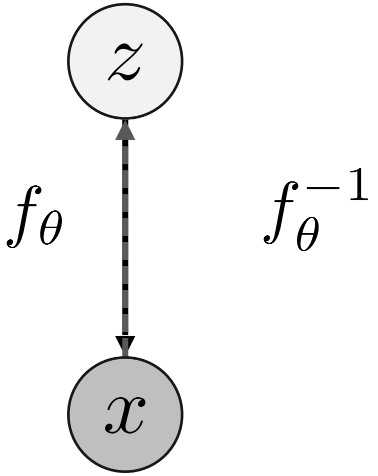
- Let Z_1 and Z_2 be continuous random variables with joint density p_{Z_1,Z_2}. Let $f = (f_1, f_2)$ be a transformation, and $h = (h_1, h_2)$ be the inverse transformation. Let $X_1 = f_1(Z_1, Z_2)$ and $X_2 = f_2(Z_1, Z_2)$. Then, $Z_1 = h_1(X_1, X_2)$ and $Z_2 = h_2(X_1, X_2)$. It follows that:

$$
\begin{aligned}
& p_{X_1,X_2}(x_1, x_2) \\
&= p_{Z_1,Z_2}(h_1(x_1, x_2), h_2(x_1, x_2)) \left| \det \begin{pmatrix} \frac{\partial h_1(x_1,x_2)}{\partial x_1} & \frac{\partial h_1(x_1,x_2)}{\partial x_2} \\ \frac{\partial h_2(x_1,x_2)}{\partial x_1} & \frac{\partial h_2(x_1,x_2)}{\partial x_2} \end{pmatrix} \right| \qquad \text{(inverse)} \\
&= p_{Z_1,Z_2}(z_1, z_2) \left| \det \begin{pmatrix} \frac{\partial f_1(z_1,z_2)}{\partial z_1} & \frac{\partial f_1(z_1,z_2)}{\partial z_2} \\ \frac{\partial f_2(z_1,z_2)}{\partial z_1} & \frac{\partial f_2(z_1,z_2)}{\partial z_2} \end{pmatrix} \right|^{-1}. \qquad \text{(forward)}
\end{aligned}
$$

Normalizing Flow Models

Lecture #8

- Consider a directed, latent-variable model over observed variables X and latent variables Z. In a **normalizing flow model**, the mapping between Z and X, given by $f_\theta : \mathbb{R}^d \to \mathbb{R}^d$, is *deterministic, invertible and differentiable*, such that $X = f_\theta(Z)$ and $Z = f_\theta^{-1}(X)$.



Using change of variables, the marginal likelihood $p_\theta(x)$ is given by:

$$p_\theta(x) = p_Z\left(f_\theta^{-1}(x)\right)\left|\det\left(\frac{\partial f_\theta^{-1}(x)}{\partial x}\right)\right|.$$

Note: x and z need to be continuous and have the same dimension.

A Flow of Transformations

- **Normalizing:** Change of variables gives a normalized density after applying an invertible transformation.

- **Flow:** Invertible transformations can be composed with each other:

$$z_K = f_\theta^K \circ \cdots \circ f_\theta^1(z_0) = f_\theta^K(f_\theta^{K-1}(\cdots f_\theta^1(z_0)\ldots)) =: f_\theta(z_0),$$

with $x = z_k$ and $z = z_0$.

A Flow of Transformations

1. Start with a simple distribution for z_0 (e.g., isotropic Gaussian).

2. Apply a sequence of K invertible transformations to finally obtain $x = z_K$.

3. By change of variables:

$$p_\theta(x) = p_Z\left(f_\theta^{-1}(x)\right) \prod_{k=1}^{K} \left|\det\left(\frac{\partial (f_\theta^k)^{-1}(z_k)}{\partial z_k}\right)\right|.$$

(*Note:* The determinant of a matrix product equals the product of the matrix determinants.)

Learning and Inference

1. Learning via **maximum likelihood** over dataset $\mathcal{D}$:

$$\max_{\theta} \log(p_\theta(\mathcal{D})) = \sum_{x \in \mathcal{D}} \left[\log p_Z(f_\theta^{-1}(x)) + \log \left| \det \left(\frac{\partial f_\theta^{-1}(x)}{\partial x} \right) \right| \right].$$

2. **Exact likelihood evaluation**: via inverse tranformation $x \to z$ and change of variables formula.

3. Sampling via forward transformation $z \to x$.

$$z \sim p_Z(z), \; x = f_\theta(z).$$

4. **Latent representations** inferred via inverse transformation (no extra inference model is required!).

$$z = f_\theta^{-1}(x).$$

Desiderata for Flow Models

1. Simple prior $p_Z(z)$ that allows for efficient sampling and tractable likelihood evaluation, e.g., isotropic Gaussian.

2. Invertible transformations with tractable evaluation:

   - Likelihood evaluation requires effcient evaluation of $x \to z$ mapping.

   - Sampling requires efficient evaluation of $z \to x$ mapping.

Desiderata for Flow Models

3. Computing likelihoods also requires the evaluation of determinants of $d \times d$ Jacobian matrices, where d is the data dimensionality:

   - Computing the determinant for an $d \times d$ matrix is $O(d^3)$: prohibitively expensive within a learning loop!

   - **Key idea**: Choose tranformations so that the resulting Jacobian matrix has special structure. For example, the determinant of a triangular matrix is the product of the diagonal entries, i.e., an $O(d)$ operation.

Triangular Jacobian

The Jacobian matrix of $x = [x_1, \cdots, x_d]^T = f(z) = [f_1(z), \cdots, f_d(z)]^T$ is given by:

$$J \equiv \frac{\partial f}{\partial z} := \begin{pmatrix} \frac{\partial f_1}{\partial z_1} & \cdots & \frac{\partial f_1}{\partial z_d} \\ \cdots & \cdots & \cdots \\ \frac{\partial f_d}{\partial z_1} & \cdots & \frac{\partial f_d}{\partial z_d} \end{pmatrix}.$$

Suppose $x_k = f_k(z)$ only depends on $z_{\leq k}$. Then, its Jacobian is given by:

$$J = \begin{pmatrix} \frac{\partial f_1}{\partial z_1} & \cdots & 0 \\ \cdots & \cdots & \cdots \\ \frac{\partial f_d}{\partial z_1} & \cdots & \frac{\partial f_d}{\partial z_d} \end{pmatrix},$$

which has lower triangular structure, thus, its determinant can be computed in **linear time**.

Planar Flows (Rezende & Mohamed, 2015)

- *Planar flow:* Invertible (residual) transformation:

$$z_k = f_{\theta_k}(z_{k-1}) = z_{k-1} + u_k h(w_k^T z_{k-1} + b_k),$$

parametrized by $\theta_k = (w_k, u_k, b_k)$ where $h(\cdot)$ is a non-linearity.

- Absolute value of the determinant of the Jacobian is given by:

$$\left|\det \frac{\partial f_{\theta_k}(z)}{\partial z}\right| = \left|\det(I + h'(w_k^T z + b_k) u_k w_k^T)\right| = \left|1 + h'(w^T z + b) u_k^T w_k\right|.$$

(via matrix determinant lemma: $\det(A + uv^T) = (1 + v^T A^{-1} u)\det(A)$)

Planar Flows (Rezende & Mohamed, 2015)

- *Planar flow:* Maximum log-likelihood:

$$\max_{\theta} \log(p_\theta(x|\mathcal{D})) = \sum_{x \in \mathcal{D}} \left[\log p_Z(z_0) - \sum_{k=1}^{K} \log \left| 1 + +h'(w^T z_k + b) u_k^T w_k \right| \right],$$

where $z_k = f_{\theta_k}^{-1}(z_{k+1})$ starting from $z_{K-1} = f_{\theta_K}^{-1}(x)$.

- Need to restrict the parameters and the non-linearity for the mapping to be invertible. For example, $h = \tanh$ and $h'(w_k^t z + b_k) u_k^T w_k > -1, \ \forall k$.

Planar Flows (Rezende & Mohamed, 2015)

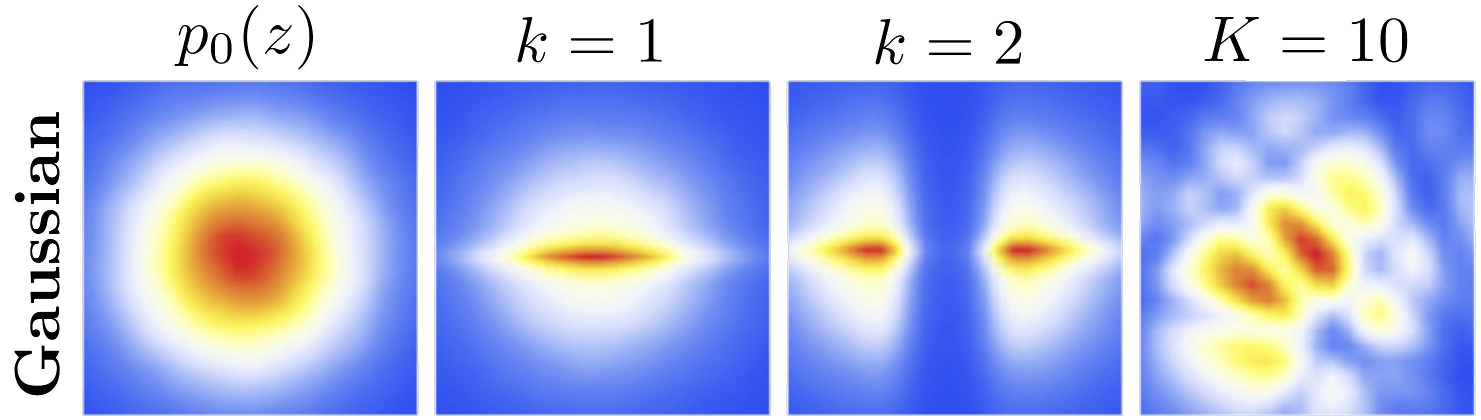
- In general, there is no analytic expression for the inverse $f_{\theta_k}^{-1}(x)$. However, it can be iteratively approximated via:

$$z_k^{(l)} = z_{k+1} - u_k h(w_k^T z_k^{(l-1)} + b_k),\ l = 1, 2, ...$$

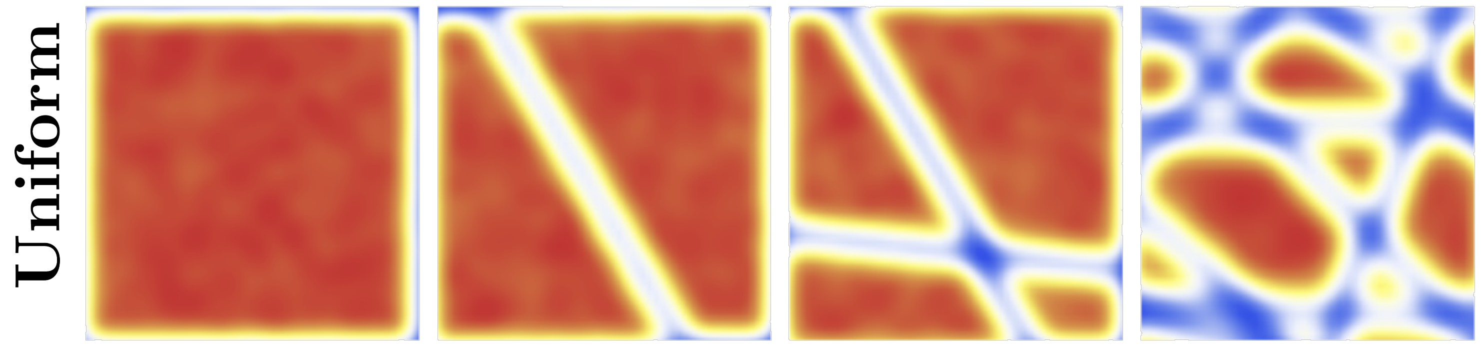
- Banach's fixed point theorem guarantees under *the contraction assumption* that the sequence $z_k^{(l)}$, $l = 1, 2, ...$ will converge to $f_{\theta_k}^{-1}(z_{k+1})$ exponentially fast.

Planar Flows (Rezende & Mohamed, 2015)

- Base distribution: Gaussian



- Base distribution: Uniform



10 planar transformations can transform a simple distribution into a far more complex one.

References

1. Probabilistic Machine Learning: Advanced Topics (*Chapter 22*)
   Kevin P Murphy, The MIT Press (2023)

2. Normalizing Flows for Probabilistic Modeling and Inference, Papamakarios et al., JMLR, 2021. (A coherent and accessible summary to normalizing flows - Highly recommended read!)

3. Variational Inference with Normalizing Flows, D. Rezende & S. Mohamed, ICML, 2015.

4. `https://github.com/janosh/awesome-normalizing-flows` (list of papers and source code).

5. `https://lilianweng.github.io/posts/2018-10-13-flow-models/`

6. `https://tech.skit.ai/normalizing-flows-part-2/`

Introduction to Deep Generative Modeling

Lecture #8

HY-673 – Computer Science Dep., University of Crete

Professors: Yannis Pantazis, Yannis Stylianou

Teaching Assistant: Michail Raptakis