

Modelos Generativos Profundos para Imágenes

Normalizing Flows

Pablo Musé

pmuse@fing.edu.uy

Instituto de Ingeniería Eléctrica
Facultad de Ingeniería



UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Agenda

① Motivación

¿Por qué Normalizing Flows?

② Transformaciones de densidades de probabilidad

Determinantes y volúmenes

Cambio de variables y transformación de densidades

③ Normalizing flows

El modelo

Entrenamiento e inferencia

Ejemplo: Planar Flows

① Motivación

¿Por qué Normalizing Flows?

② Transformaciones de densidades de probabilidad

Determinantes y volúmenes

Cambio de variables y transformación de densidades

③ Normalizing flows

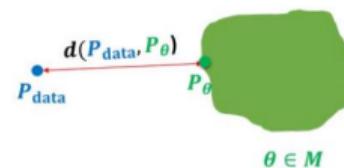
El modelo

Entrenamiento e inferencia

Ejemplo: Planar Flows

Modelado generativo

Recordemos el problema que queremos resolver:



$$\mathbf{x}_n \sim p_{\text{data}}(\mathbf{x}), n = 1, \dots, N \quad \text{Familia de modelos}$$

- **Objetivo:** aprender una densidad $p(\mathbf{x})$ a partir de $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ que nos permita:
 - Generar muestras realistas
 - Tener una estimación fiable de la densidad (e.g. para detección de anomalías)
 - Aprender buenas representaciones (**vectores de features**) de los datos
- **¿Cómo representar $p(\mathbf{x})$?**
- **¿Cómo aprender o estimar $p(\mathbf{x})$?**

Repaso: modelos autorregresivos

Modelos autorregresivos: $p(\mathbf{x}) = \prod_{i=1}^d p_\theta(x_i | \mathbf{x}_{<i}).$

- Las distribuciones de probabilidad se factorizan en un producto de factores
- $p(\mathbf{x})$ se puede representar de forma eficiente mediante independencia condicional y/o parametrizaciones neuronales

Ventajas:

- La evaluación de las densidades de probabilidad es tratable
- El entrenamiento de $p(\mathbf{x})$ por máxima verosimilitud y ascenso por gradiente es manejable

Desventajas:

- Requieren elegir un orden sobre las variables.
- La generación es secuencial (en general lenta)
- No pueden aprender características de forma no supervisada.

① Motivación

¿Por qué Normalizing Flows?

② Transformaciones de densidades de probabilidad

Determinantes y volúmenes

Cambio de variables y transformación de densidades

③ Normalizing flows

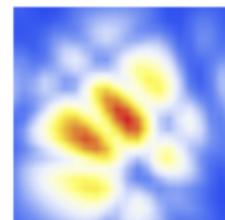
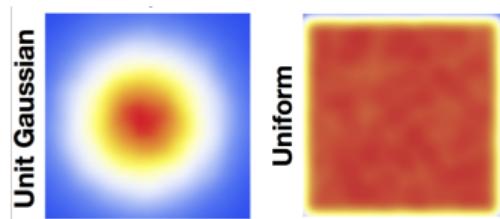
El modelo

Entrenamiento e inferencia

Ejemplo: Planar Flows

De priors simples a distribuciones de datos complejas

- **Propiedades deseables** de cualquier distribución de modelo $p_\theta(\mathbf{x})$:
 - Densidad en forma cerrada, **fácil de evaluar** (útil para el entrenamiento)
 - **Fácil de muestrear** (útil para la generación)
- Muchas distribuciones simples satisfacen las propiedades anteriores, e.g., distribuciones gaussianas, uniformes.
- Limitante: las distribuciones de datos reales son más complejas (multi-modales)
- **Idea clave detrás de los modelos basados en flujo:** mapear distribuciones simples (fáciles de muestrear y evaluar densidades) a distribuciones complejas a través de una transformación invertible.



Repaso de variables aleatorias continuas

- Sea X una variable aleatoria continua.
- Función de distribución o densidad acumulada (cdf): $F_X(x) = \Pr(X \leq x)$.
- Función de densidad de probabilidad (pdf): $p_X(x) = F'_X(x) = \frac{dF_X(x)}{dx}$.
- Ejemplos de densidades paramétricas comunes:

Gaussiana: $X \sim \mathcal{N}(\mu, \sigma^2) = p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$

Uniforme: $X \sim \mathcal{U}(a, b) = p_X(x) = \frac{1}{b-a} \mathbb{1}[a \leq x \leq b]$.

- Sea $\mathbf{X} = (X_1, X_2, \dots, X_d)$ un vector aleatorio continuo de dimensión \Rightarrow Densidad conjunta $p_{\mathbf{X}}(\mathbf{x}) = p_{\mathbf{X}}(x_1, x_2, \dots, x_d)$.

Ejemplo: Gaussiana: $p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$.

① Motivación

¿Por qué Normalizing Flows?

② Transformaciones de densidades de probabilidad

Determinantes y volúmenes

Cambio de variables y transformación de densidades

③ Normalizing flows

El modelo

Entrenamiento e inferencia

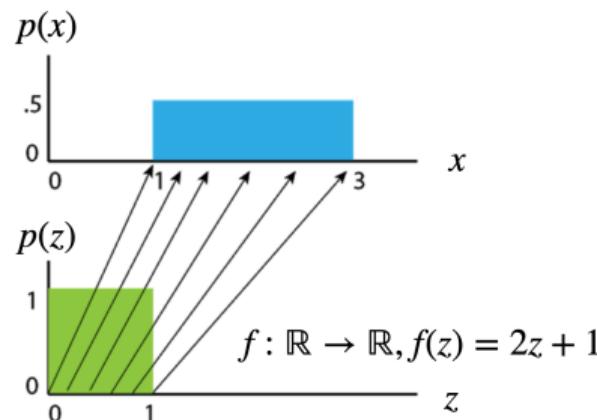
Ejemplo: Planar Flows

Fórmula de Cambio de Variables

Fórmula de cambio de variables (caso 1D): Si $X = f(Z)$ y $f(\cdot)$ es monótona con inversa $Z = f^{-1}(X) = h(X)$, entonces

$$p_X(x) = p_Z(h(x))|h'(x)|, \quad p_X(x)dx = p_Z(z)\frac{1}{f'(z)}dz.$$

- **Ejemplo 1:** $Z \sim \mathcal{U}[0, 1]$, $X = f(Z) = 2Z + 1$



$$h(x) = \frac{x}{2} - 1,$$
$$h'(x) = \frac{1}{2},$$

$$p_X(x) = p_Z\left(\frac{x}{2} - 1\right) \left|\frac{1}{2}\right| = \frac{1}{2} \mathbb{1}_{\{1 \leq x \leq 3\}}.$$

Fórmula de Cambio de Variables

Fórmula de cambio de variables (caso 1D): Si $X = f(Z)$ y $f(\cdot)$ es monótona con inversa $Z = f^{-1}(X) = h(X)$, entonces

$$p_X(x) = p_Z(h(x))|h'(x)|, \quad p_X(x)dx = p_Z(z)\frac{1}{f'(z)}dz.$$

- **Ejemplo 2:** Si $X = f(Z) = \exp(Z)$ y $Z \sim \mathcal{U}[0, 2]$, ¿cómo es $p_X(x)$?

$$h(x) = \ln x,$$

$$h'(x) = \frac{1}{x},$$

$$p_X(x) = p_Z(\ln x) \left| \frac{1}{x} \right| = \frac{1}{2x} \mathbb{1}_{\{e^0 \leq x \leq e^2\}}.$$

Fórmula de Cambio de Variables

Fórmula de cambio de variables (caso 1D): Si $X = f(Z)$ y $f(\cdot)$ es monótona con inversa $Z = f^{-1}(X) = h(X)$, entonces

$$p_X(x) = p_Z(h(x))|h'(x)|, \quad p_X(x)dx = p_Z(z)\frac{1}{f'(z)}dz.$$

Prueba (esbozo): Asuma que $f(\cdot)$ es monótonamente creciente:

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

Tomando derivadas en ambos lados:

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

Como $h'(x) = [f^{-1}]'(x) = \frac{1}{f'(f^{-1}(x))}$, usando $z = h(x) = f^{-1}(x)$, también podemos escribir $p_X(x)dx = p_Z(z)\frac{1}{f'(z)}dz$.

① Motivación

¿Por qué Normalizing Flows?

② Transformaciones de densidades de probabilidad

Determinantes y volúmenes

Cambio de variables y transformación de densidades

③ Normalizing flows

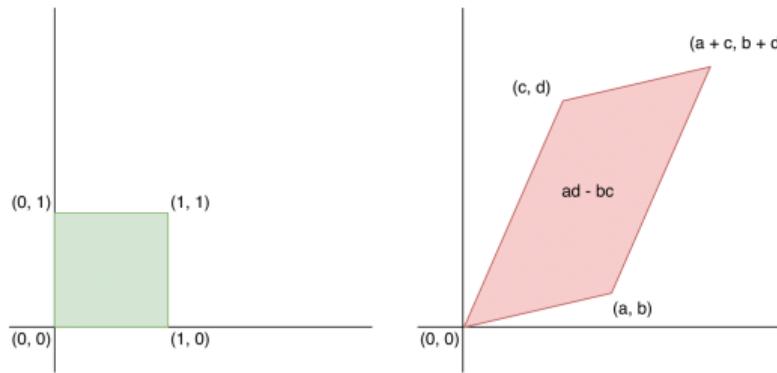
El modelo

Entrenamiento e inferencia

Ejemplo: Planar Flows

Geometría: Determinantes y volúmenes

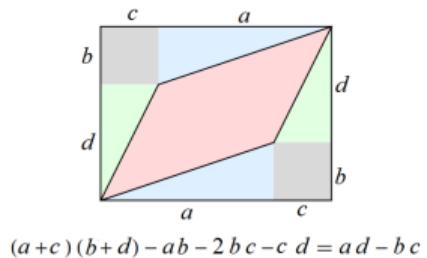
- Sea $\mathbf{Z} \sim \mathcal{U}[0, 1]^d$.
- Sea $\mathbf{X} = A\mathbf{Z}$, A matriz cuadrada invertible, con inversa $W = A^{-1}$.
- ¿Cómo se distribuye \mathbf{X} ?
- Geométricamente, la matriz A mapea el hipercubo unitario $[0, 1]^d$ a un paralelepípedo



Geometría: Determinantes y volúmenes

- El volumen del paralelepípedo es igual al valor absoluto del determinante de la matriz A :

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



- $\mathbf{X} = A\mathbf{Z}$ se distribuye uniformemente sobre el paralelepípedo de área $|\det(A)|$.
- Por lo tanto, tenemos

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{|\det(A)|} p_{\mathbf{Z}}(W\mathbf{x}) = p_{\mathbf{Z}}(W\mathbf{x}) |\det(W)|. \quad (\det A)^{-1} = \det(A^{-1}) = \det W$$

① Motivación

¿Por qué Normalizing Flows?

② Transformaciones de densidades de probabilidad

Determinantes y volúmenes

Cambio de variables y transformación de densidades

③ Normalizing flows

El modelo

Entrenamiento e inferencia

Ejemplo: Planar Flows

Cambio de variables: caso general

- Para transformaciones lineales $\mathbf{z} \mapsto A\mathbf{z}$, el cambio en el volumen está dado por $|\det A|$.
- Para transformaciones no lineales $\mathbf{z} = f(\mathbf{x})$, el cambio linealizado en el volumen está dado por el determinante del Jacobiano de f .
- **Cambio de variables (Caso general):**

Sea $f : \mathbb{R}^d \mapsto \mathbb{R}^d$ invertible, tal que $X = f(Z)$ y $Z = f^{-1}(X)$. Entonces

$$p_{\mathbf{X}}(\mathbf{x}) = p_{\mathbf{Z}}(f^{-1}(\mathbf{x})) \left| \det \left(\frac{\partial f^{-1}(\mathbf{x})}{\partial \mathbf{x}} \right) \right|, \quad p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = p_{\mathbf{Z}}(\mathbf{z}) \left| \det \left(\frac{\partial f(\mathbf{z})}{\partial \mathbf{z}} \right) \right|^{-1} d\mathbf{z}.$$

Ejemplo bidimensional

- Sean Z_1 y Z_2 V.A.s reales con densidad conjunta p_{Z_1, Z_2} .
- Sea $\mathbf{u} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ una transformación invertible.
- Dos entradas y dos salidas, denotadas $\mathbf{u} = (u_1, u_2)$.
- Sea $\mathbf{v} = (v_1, v_2)$ su transformación inversa.
- Sean $X_1 = u_1(Z_1, Z_2)$, $X_2 = u_2(Z_1, Z_2)$, $Z_1 = v_1(X_1, X_2)$, y $Z_2 = v_2(X_1, X_2)$.

$$p_{X_1, X_2}(x_1, x_2) = p_{Z_1, Z_2}(v_1(x_1, x_2), v_2(x_1, x_2)) \left| \det \begin{pmatrix} \frac{\partial v_1(x_1, x_2)}{\partial x_1} & \frac{\partial v_1(x_1, x_2)}{\partial x_2} \\ \frac{\partial v_2(x_1, x_2)}{\partial x_1} & \frac{\partial v_2(x_1, x_2)}{\partial x_2} \end{pmatrix} \right|$$

$$p_{X_1, X_2}(x_1, x_2) dx_1 dx_2 = p_{Z_1, Z_2}(z_1, z_2) \left| \det \begin{pmatrix} \frac{\partial u_1(z_1, z_2)}{\partial z_1} & \frac{\partial u_1(z_1, z_2)}{\partial z_2} \\ \frac{\partial u_2(z_1, z_2)}{\partial z_1} & \frac{\partial u_2(z_1, z_2)}{\partial z_2} \end{pmatrix} \right|^{-1} dz_1 dz_2.$$

① Motivación

¿Por qué Normalizing Flows?

② Transformaciones de densidades de probabilidad

Determinantes y volúmenes

Cambio de variables y transformación de densidades

③ Normalizing flows

El modelo

Entrenamiento e inferencia

Ejemplo: Planar Flows

① Motivación

¿Por qué Normalizing Flows?

② Transformaciones de densidades de probabilidad

Determinantes y volúmenes

Cambio de variables y transformación de densidades

③ Normalizing flows

El modelo

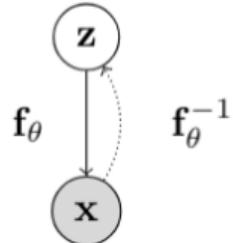
Entrenamiento e inferencia

Ejemplo: Planar Flows

Modelos de Normalizing Flows

- Consideramos un modelo de variable latente dirigido sobre variables observadas \mathbf{X} y variables latentes \mathbf{Z} , **ambas continuas y de misma dimensión**.
- En un modelo de *Normalizing Flow*, el mapeo entre \mathbf{Z} y \mathbf{X} , dado por $f_\theta : \mathbb{R}^d \mapsto \mathbb{R}^d$, es determinista e invertible:

$$\mathbf{X} = f_\theta(\mathbf{Z}), \quad \mathbf{Z} = f_\theta^{-1}(\mathbf{X}).$$



- La verosimilitud de los datos $p(\mathbf{x})$ está dada por

$$p_{\mathbf{X}}(\mathbf{x}; \theta) = p_{\mathbf{Z}}(f_\theta^{-1}(\mathbf{x})) \left| \det \left(\frac{\partial f_\theta^{-1}(\mathbf{x})}{\partial \mathbf{x}} \right) \right|, \quad p_{\mathbf{X}}(\mathbf{x}; \theta) = p_{\mathbf{Z}}(\mathbf{z}) \left| \det \left(\frac{\partial f_\theta(\mathbf{z})}{\partial \mathbf{z}} \right) \right|^{-1}.$$

Un flujo de transformaciones

Queremos construir una transformación entre densidades f_θ que sea:

- **Invertible**: para poder aplicar el cambio de variable.
- **Expresiva**: para poder aprender distribuciones complejas.
- **Computacionalmente sencilla**: para poder optimizarla y evaluarla.

Idea: componer transformaciones simples e invertibles, que den expresividad a la composición:

- Comenzar por $\mathbf{Z} = \mathbf{Z}_0$ con densidad sencilla (e.g. gaussiana).
- Aplicar una serie de K transformaciones invertibles para finalmente obtener $\mathbf{X} = \mathbf{Z}_K$

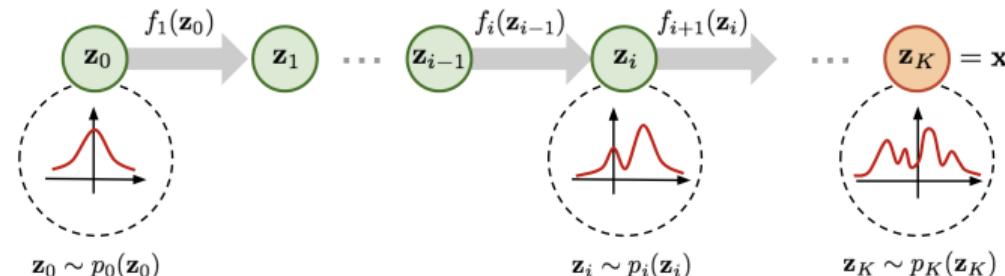


Figura tomada de *Flow-based Deep Generative Models*, Lilian Weng, 2018

Un flujo de transformaciones

$$\mathbf{x} = \mathbf{z}_K = f_K \circ f_{K-1} \circ \cdots \circ f_1(\mathbf{z}_0) \triangleq f_\theta(\mathbf{z}_0), \quad p_{\mathbf{z}_i}(\mathbf{z}_i) = p_{\mathbf{z}_{i-1}}(\mathbf{z}_{i-1}) \left| \det \left(\frac{\partial f_i(\mathbf{z}_{i-1})}{\partial \mathbf{z}_{i-1}} \right) \right|^{-1}.$$

¿Cómo queda la log-verosimilitud de \mathbf{X} ?

$$\begin{aligned}\log p_{\mathbf{X}}(\mathbf{x}) &= \log p_{\mathbf{z}_K}(\mathbf{z}_K) = \log p_{\mathbf{z}_{K-1}}(\mathbf{z}_{K-1}) + \log \left[\left| \det \left(\frac{\partial f_K(\mathbf{z}_{K-1})}{\partial \mathbf{z}_{K-1}} \right) \right|^{-1} \right] \\ &= \log p_{\mathbf{z}_{K-1}}(\mathbf{z}_{K-1}) - \log \left| \det \left(\frac{\partial f_K}{\partial \mathbf{z}_{K-1}} \right) \right| \\ &= \log p_{\mathbf{z}_{K-2}}(\mathbf{z}_{K-2}) - \log \left| \det \left(\frac{\partial f_{K-1}}{\partial \mathbf{z}_{K-2}} \right) \right| - \log \left| \det \left(\frac{\partial f_K}{\partial \mathbf{z}_{K-1}} \right) \right| \\ &= \dots \\ &= \log p_{\mathbf{z}_0}(\mathbf{z}_0) - \sum_{i=1}^K \log \left| \det \left(\frac{\partial f_i}{\partial \mathbf{z}_{i-1}} \right) \right|.\end{aligned}$$

① Motivación

¿Por qué Normalizing Flows?

② Transformaciones de densidades de probabilidad

Determinantes y volúmenes

Cambio de variables y transformación de densidades

③ Normalizing flows

El modelo

Entrenamiento e inferencia

Ejemplo: Planar Flows

Aprendizaje e Inferencia

- Tenemos

$$\log p_{\mathbf{x}}(\mathbf{x}) = \log p_{\mathbf{z}_0}(\mathbf{z}_0) - \sum_{i=1}^K \log \left| \det \left(\frac{\partial f_i}{\partial \mathbf{z}_{i-1}} \right) \right|$$

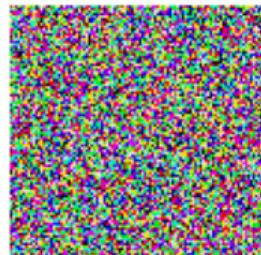
pdf simple conocida por diseño

- Aprendemos los parámetros maximizando la verosimilitud sobre un conjunto de datos \mathcal{D} :

$$\log p_{\theta}(\mathcal{D}) = \sum_{\mathbf{x} \in \mathcal{D}} \log p_{\mathbf{z}}(f_{\theta}^{-1}(\mathbf{x})) + \log \left| \det \left(\frac{\partial f_{\theta}^{-1}(\mathbf{x})}{\partial \mathbf{x}} \right) \right|$$

- **Evaluación exacta de la verosimilitud** a través de la transformación inversa $\mathbf{x} \mapsto \mathbf{z}$ y la fórmula de cambio de variables.
- **Muestreo** a través de la transformación directa $\mathbf{z} \mapsto \mathbf{x}$: $\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z}), \mathbf{x} = f_{\theta}(\mathbf{z})$
- **Representaciones latentes** inferidas a través de la transformación inversa: $\mathbf{z} = f_{\theta}^{-1}(\mathbf{x})$.

Inferencia



$$\mathbf{z} \sim p_{simple}(\mathbf{z})$$

$$\mathbf{x} \sim p_{model}(\mathbf{x})$$



$$p_{model}(\mathbf{x}) = p_{simple}(f_\theta^{-1}(\mathbf{x}))$$

$$\mathbf{x} \sim p_{data}(\mathbf{x})$$

① Motivación

¿Por qué Normalizing Flows?

② Transformaciones de densidades de probabilidad

Determinantes y volúmenes

Cambio de variables y transformación de densidades

③ Normalizing flows

El modelo

Entrenamiento e inferencia

Ejemplo: Planar Flows

Ejemplo: Planar Flows

- Transformación invertible

$$\mathbf{x} = f_{\theta}(\mathbf{z}) = \mathbf{z} + \mathbf{u}h(\mathbf{w}^T \mathbf{z} + b), \quad (h(\cdot) \text{ una no-linealidad})$$

parametrizada por $\theta = (\mathbf{w}, \mathbf{u}, b)$.

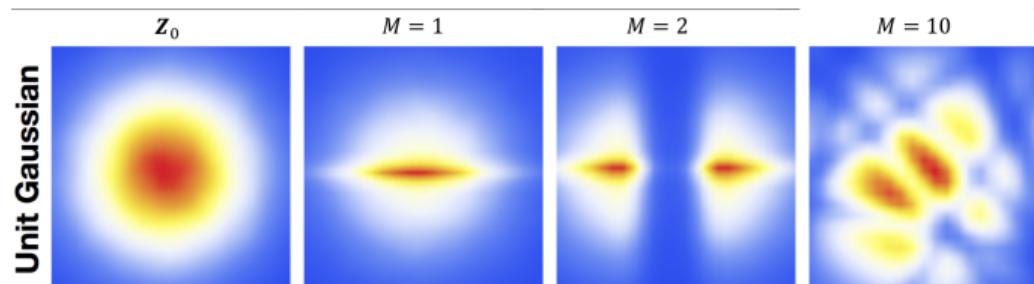
- Valor absoluto del determinante del Jacobiano:

$$\left| \det \left(\frac{\partial f_{\theta}(z)}{\partial z} \right) \right| = \left| \det \left(I + h'(\mathbf{w}^T \mathbf{z} + b) \mathbf{u} \mathbf{w}^T \right) \right| = \left| 1 + h'(\mathbf{w}^T \mathbf{z} + b) \mathbf{u}^T \mathbf{w} \right|. \quad (\text{lema del determinante})$$

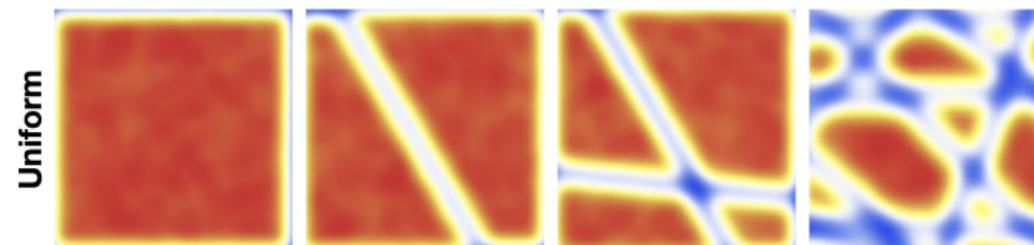
- Es necesario restringir los parámetros y la no-linealidad para que la transformación sea invertible. E.g., $h = \tanh()$ y $h'(\mathbf{w}^T \mathbf{z} + b) \mathbf{u}^T \mathbf{w} \geq -1$.

Ejemplo: Planar Flows

- Distribución base: gaussiana



- Distribución base: uniforme



- 10 transformaciones planas pueden transformar distribuciones simples en compleja.

Características deseables en los modelos de flujo

- Prior simple $p_z(z)$ que permita un muestreo eficiente y una evaluación de verosimilitud tratable. E.g., gaussiana isotrópica.
- Transformaciones invertibles con evaluación calculable:
 - La evaluación de la verosimilitud requiere una evaluación eficiente del mapeo $\mathbf{x} \mapsto \mathbf{z}$
 - El muestreo requiere una evaluación eficiente del mapeo $\mathbf{z} \mapsto \mathbf{x}$.
- Calcular las verosimilitudes también requiere evaluar determinantes de matrices Jacobianas de $d \times d$.
 - Determinante de matriz de $d \times d$ es $O(d^3) \Rightarrow$ prohibitivo en bucle de aprendizaje
 - **Idea:** elegir transformaciones para que la Jacobiana tenga una estructura especial.
 - Ejemplo: si la matriz Jacobiana es triangular su determinante cuesta $O(d)$

Jacobiano Triangular

- $\mathbf{x} = (x_1, \dots, x_d) = f(\mathbf{z}) = (f_1(\mathbf{z}), \dots, f_d(\mathbf{z}))$

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

- Si $x_i = f_i(\mathbf{z})$ solo depende de $\mathbf{z}_{\leq i}$, entonces:

$$J = \frac{\partial f}{\partial \mathbf{z}} = \begin{pmatrix} \frac{\partial f_1}{\partial z_1} & \cdots & 0 \\ \cdots & \cdots & \cdots \\ \frac{\partial f_d}{\partial z_1} & \cdots & \frac{\partial f_d}{\partial z_d} \end{pmatrix}.$$

⇒ Estructura triangular inferior. El determinante se puede calcular en tiempo lineal.

- De la misma forma el Jacobiano es triangular superior si x_i solo depende de $\mathbf{z}_{\geq i}$.

Referencias

-  C. M. Bishop, *Pattern Recognition and Machine Learning*.
Springer, 2006.
-  Stanford, “CS236 Deep Generative Models.” <https://deepgenerativemodels.github.ioLecture>, 2024.
-  J. M. Tomczak, *Deep Generative Modeling*.
Springer Cham, 2024.