Normalizing Flow Models

Stefano Ermon

Stanford University

Lecture 7
Recap of likelihood-based learning so far:

- **Model families:**
  - **Autoregressive Models:** \( p_\theta(x) = \prod_{i=1}^n p_\theta(x_i|x_{<i}) \)
  - **Variational Autoencoders:** \( p_\theta(x) = \int p_\theta(x,z)dz \)

- Autoregressive models provide tractable likelihoods but no direct mechanism for learning features
- Variational autoencoders can learn feature representations (via latent variables \( z \)) but have intractable marginal likelihoods
- **Key question:** Can we design a latent variable model with tractable likelihoods? Yes!
Simple Prior to Complex Data Distributions

- Desirable properties of any model distribution $p_\theta(x)$:
  - Easy-to-evaluate, closed form density (useful for training)
  - Easy-to-sample (useful for generation)
- Many simple distributions satisfy the above properties e.g., Gaussian, uniform distributions

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

- **Key idea behind flow models**: Map simple distributions (easy to sample and evaluate densities) to complex distributions through an invertible transformation.
A flow model is similar to a variational autoencoder (VAE):

1. Start from a simple prior: \( z \sim \mathcal{N}(0, I) = p(z) \)
2. Transform via \( p(x \mid z) = \mathcal{N}(\mu_\theta(z), \Sigma_\theta(z)) \)
3. Even though \( p(z) \) is simple, the marginal \( p_\theta(x) \) is very complex/flexible. However, \( p_\theta(x) = \int p_\theta(x, z) dz \) is expensive to compute: need to enumerate all \( z \) that could have generated \( x \)
4. What if we could easily "invert" \( p(x \mid z) \) and compute \( p(z \mid x) \) by design? How? Make \( x = f_\theta(z) \) a deterministic and invertible function of \( z \), so for any \( x \) there is a unique corresponding \( z \) (no enumeration)
Continuous random variables refresher

- Let $X$ be a continuous random variable
- The cumulative density function (CDF) of $X$ is $F_X(a) = P(X \leq a)$
- The probability density function (pdf) of $X$ is $p_X(a) = F'_X(a) = \frac{dF_X(a)}{da}$
- Typically consider parameterized densities:
  - Gaussian: $X \sim \mathcal{N}(\mu, \sigma)$ if $p_X(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
  - Uniform: $X \sim \mathcal{U}(a, b)$ if $p_X(x) = \frac{1}{b-a} 1[a \leq x \leq b]$
  - Etc.
- If $X$ is a continuous random vector, we can usually represent it using its **joint probability density function**:
  - Gaussian: if $p_X(x) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)$
Let $Z$ be a uniform random variable $U[0, 2]$ with density $p_Z$. What is $p_Z(1)$?

- $\frac{1}{2}$

  As a sanity check, $\int_0^2 \frac{1}{2} = 1$

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

- $p_X(4) = p(X = 4) = p(4Z = 4) = p(Z = 1) = p_Z(1) = 1/2$ **Wrong!**

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

- To get correct result, need to use **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:

$$p_X(x) = p_Z(h(x))|h'(x)|$$

- Previous example: If $X = f(Z) = 4Z$ and $Z \sim \mathcal{U}[0, 2]$, what is $p_X(4)$?
  - Note that $h(X) = X/4$
  - $p_X(4) = p_Z(1)h'(4) = 1/2 \times |1/4| = 1/8$

- More interesting example: If $X = f(Z) = \exp(Z)$ and $Z \sim \mathcal{U}[0, 2]$, what is $p_X(x)$?
  - Note that $h(X) = \ln(X)$
  - $p_X(x) = p_Z(\ln(x))|h'(x)| = \frac{1}{2x}$ for $x \in [\exp(0), \exp(2)]$

- Note that the "shape" of $p_X(x)$ is different (more complex) from that of the prior $p_Z(z)$. 
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:

  $$p_X(x) = p_Z(h(x))|h'(x)|$$

- **Proof sketch:** Assume $f(\cdot)$ is monotonically increasing

  $$F_X(x) = p[X \leq x] = p[f(Z) \leq x] = p[Z \leq h(x)] = F_Z(h(x))$$

  Taking derivatives on both sides:

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

- 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

  $$p_X(x) = p_Z(z)\frac{1}{f'(z)}$$
Geometry: Determinants and volumes

- Let $Z$ be a uniform random vector in $[0, 1]^n$
- Let $X = AZ$ for a square invertible matrix $A$, with inverse $W = A^{-1}$. How is $X$ distributed?
- Geometrically, the matrix $A$ maps the unit hypercube $[0, 1]^n$ to a parallelotope
- Hypercube and parallelotope are generalizations of square/cube and parallelogram/parallelopiped to higher dimensions

**Figure:** The matrix $A = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$ maps a unit square to a parallelogram
The volume of the parallelepiped is equal to the absolute value of the determinant of the matrix $A$

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

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

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

$$= p_Z(Wx) |\det(W)|$$

because if $W = A^{-1}$, $\det(W) = \frac{1}{\det(A)}$. Note similarity with 1D case formula.
Generalized change of variables

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

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

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

- Note 0: generalizes the previous 1D case $p_X(x) = p_Z(h(x))|h'(x)|$
- Note 1: unlike VAEs, $x, z$ need to be continuous and have the same dimension. For example, if $x \in \mathbb{R}^n$ then $z \in \mathbb{R}^n$
- Note 2: For any invertible matrix $A$, $\det(A^{-1}) = \det(A)^{-1}$

\[
p_X(x) = p_Z(z) \left| \det \left( \frac{\partial f(z)}{\partial z} \right) \right|^{-1}
\]
Two Dimensional Example

Let $Z_1$ and $Z_2$ be continuous random variables with joint density $p_{Z_1, Z_2}$.

Let $u : \mathbb{R}^2 \to \mathbb{R}^2$ be an invertible transformation. Two inputs and two outputs, denoted $u = (u_1, u_2)$.

Let $v = (v_1, v_2)$ be its inverse transformation.

Let $X_1 = u_1(Z_1, Z_2)$ and $X_2 = u_2(Z_1, Z_2)$ Then, $Z_1 = v_1(X_1, X_2)$ and $Z_2 = v_2(X_1, X_2)$

$$p_{X_1, X_2}(x_1, x_2)$$

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

$$= p_{Z_1, Z_2}(z_1, z_2) \left| \det \begin{pmatrix} \frac{\partial u_1(z_1, z_2)}{\partial z_1} & \frac{\partial u_1(z_1, z_2)}{\partial z_2} \\ \frac{\partial u_2(z_1, z_2)}{\partial z_1} & \frac{\partial u_2(z_1, z_2)}{\partial z_2} \end{pmatrix} \right|^{-1} \text{ (forward)}$$
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 $f_\theta : \mathbb{R}^n \mapsto \mathbb{R}^n$, is deterministic and invertible such that $X = f_\theta(Z)$ and $Z = f_\theta^{-1}(X)$.

Using change of variables, the marginal likelihood $p(x)$ is given by

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

Note: $x, z$ 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

\[ z_m = f_{\theta}^m \circ \cdots \circ f_{\theta}^1(z_0) = f_{\theta}^m(f_{\theta}^{m-1}(\cdots(f_{\theta}^1(z_0)))) \triangleq f_{\theta}(z_0) \]

- Start with a simple distribution for \( z_0 \) (e.g., Gaussian)
- Apply a sequence of \( M \) invertible transformations to finally obtain \( x = z_M \)
- By change of variables

\[ p_X(x; \theta) = p_z(f_{\theta}^{-1}(x)) \prod_{m=1}^{M} \left| \det \left( \frac{\partial (f_{\theta}^m)^{-1}(z_m)}{\partial z_m} \right) \right| \]

(Note: determinant of product equals product of determinants)
Planar flows (Rezende & Mohamed, 2016)

- Base distribution: Gaussian

<table>
<thead>
<tr>
<th>$Z_0$</th>
<th>$M = 1$</th>
<th>$M = 2$</th>
<th>$M = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Unit Gaussian" /></td>
<td><img src="image2.png" alt="Unit Gaussian" /></td>
<td><img src="image3.png" alt="Unit Gaussian" /></td>
<td><img src="image4.png" alt="Unit Gaussian" /></td>
</tr>
</tbody>
</table>

- Base distribution: Uniform

<table>
<thead>
<tr>
<th>Uniform</th>
<th>Uniform</th>
<th>Uniform</th>
<th>Uniform</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image5.png" alt="Uniform" /></td>
<td><img src="image6.png" alt="Uniform" /></td>
<td><img src="image7.png" alt="Uniform" /></td>
<td><img src="image8.png" alt="Uniform" /></td>
</tr>
</tbody>
</table>

- 10 planar transformations can transform simple distributions into a more complex one
Learning and Inference

- **Learning via maximum likelihood** over the dataset $\mathcal{D}$

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

- **Exact likelihood evaluation** via inverse tranformation $x \mapsto z$ and change of variables formula

- **Sampling** via forward transformation $z \mapsto x$

  $$
  z \sim p_Z(z) \quad x = f_\theta(z)
  $$

- **Latent representations** inferred via inverse transformation (no inference network required!)

  $$
  z = f_\theta^{-1}(x)
  $$
Desiderata for flow models

- Simple prior $p_Z(z)$ that allows for efficient sampling and tractable likelihood evaluation. E.g., isotropic Gaussian
- Invertible transformations with tractable evaluation:
  - Likelihood evaluation requires efficient evaluation of $x \mapsto z$ mapping
  - Sampling requires efficient evaluation of $z \mapsto x$ mapping
- Computing likelihoods also requires the evaluation of determinants of $n \times n$ Jacobian matrices, where $n$ is the data dimensionality
  - Computing the determinant for an $n \times n$ matrix is $O(n^3)$: prohibitively expensive within a learning loop!
- **Key idea**: Choose transformations 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 $O(n)$ operation
Triangular Jacobian

\[ \mathbf{x} = (x_1, \ldots, x_n) = \mathbf{f}(\mathbf{z}) = (f_1(\mathbf{z}), \ldots, f_n(\mathbf{z})) \]

\[ J = \frac{\partial \mathbf{f}}{\partial \mathbf{z}} = \begin{pmatrix}
  \frac{\partial f_1}{\partial z_1} & \cdots & \frac{\partial f_1}{\partial z_n} \\
  \cdots & \cdots & \cdots \\
  \frac{\partial f_n}{\partial z_1} & \cdots & \frac{\partial f_n}{\partial z_n}
\end{pmatrix} \]

Suppose \( x_i = f_i(\mathbf{z}) \) only depends on \( \mathbf{z}_{\leq i} \). Then

\[ J = \frac{\partial \mathbf{f}}{\partial \mathbf{z}} = \begin{pmatrix}
  \frac{\partial f_1}{\partial z_1} & \cdots & 0 \\
  \cdots & \cdots & \cdots \\
  \frac{\partial f_n}{\partial z_1} & \cdots & \frac{\partial f_n}{\partial z_n}
\end{pmatrix} \]

has lower triangular structure. Determinant can be computed in linear time. Similarly, the Jacobian is upper triangular if \( x_i \) only depends on \( \mathbf{z}_{\geq i} \).
Planar flows (Rezende & Mohamed, 2016)

- Planar flow. Invertible transformation

\[ x = f_\theta(z) = z + uh(w^Tz + b) \]

parameterized by \( \theta = (w, u, b) \) where \( h(\cdot) \) is a non-linearity

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

\[
\left| \det \frac{\partial f_\theta(z)}{\partial z} \right| = \left| \det(I + h'(w^Tz + b)uw^T) \right|
\]

\[
= \left| 1 + h'(w^Tz + b)u^Tw \right|
\]

(matrix determinant lemma)

- Need to restrict parameters and non-linearity for the mapping to be invertible. For example, \( h = \tanh(\cdot) \) and \( h'(w^Tz + b)u^Tw \geq -1 \)