Normalizing Flow Models

Stefano Ermon, Aditya Grover

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!
Desirable properties of any model distribution:
- Analytic density
- Easy-to-sample

Many simple distributions satisfy the above properties e.g., Gaussian, uniform distributions

Unfortunately, data distributions could be much more complex (multi-modal)

**Key idea:** Map simple distributions (easy to sample and evaluate densities) to complex distributions (learned via data) using *change of variables*.
Let $Z$ be a uniform random variable $\mathcal{U}[0, 2]$ with density $p_Z$. What is $p_Z(1)$? $\frac{1}{2}$

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

$p_X(4) = p(X = 4) = p(4Z = 4) = p(Z = 1) = p_Z(1) = 1/2$ No

Clearly, $X$ is uniform in $[0, 8]$, so $p_X(4) = 1/8$
**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 $X = 4Z$ and $Z \sim \mathcal{U}[0, 2]$, what is $p_X(4)$?

- Note that $h(X) = X/4$
- $p_X(4) = p_Z(1)h'(4) = 1/2 \times 1/4 = 1/8$
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 parallelotope is equal to the determinant of the transformation $A$

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

$X$ is uniformly distributed over the parallelotope. Hence, we have

$$p_X(x) = p_Z(Wx) |\det(W)|$$
$$= p_Z(Wx) / |\det(A)|$$
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 linearized change in volume is given by the determinant of the Jacobian of $f(\cdot)$.

**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 1: $x, z$ 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 $A$, $\text{det}(A^{-1}) = \text{det}(A)^{-1}$

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

- Let \( Z_1 \) and \( Z_2 \) be continuous random variables with joint density \( p_{Z_1,Z_2} \).
- Let \( u = (u_1, u_2) \) be a transformation
- Let \( v = (v_1, v_2) \) be the 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 \( Z_2 = v_2(X_1, X_2) \)

\[
\begin{align*}
  p_{X_1,X_2}(x_1, x_2) &= p_{Z_1,Z_2}(v_1(x_1, x_2), v_2(x_1, x_2)) \\
  &= p_{Z_1,Z_2}(z_1, z_2) \left| \frac{\partial v_1(x_1, x_2)}{\partial x_1} \frac{\partial v_2(x_1, x_2)}{\partial x_2} \right| (\text{inverse})
\end{align*}
\]

\[
\begin{align*}
  &= p_{Z_1,Z_2}(z_1, z_2) \left| \frac{\partial u_1(z_1,z_2)}{\partial z_1} \frac{\partial u_2(z_1,z_2)}{\partial z_2} \right|^{-1} (\text{forward})
\end{align*}
\]
Normalizing flow models

- 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 $Z = f_\theta^{-1}(X)$

![Diagram of normalizing flow model]

- 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 \( x \triangleq z_M \)
- By change of variables

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

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

- Planar flow. Invertible transformation

\[ x = f_\theta(z) = z + u h(w^T z + b) \]

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

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

\[
\left| \det \frac{\partial f_\theta(z)}{\partial z} \right| = \left| \det(I + h'(w^T z + b)uw^T) \right| \\
= \left| 1 + h'(w^T z + b)u^T w \right|
\]

(matrix determinant lemma)

- Need to restrict parameters and non-linearity for the mapping to be invertible. For example, \( h = tanh() \) and \( h'(w^T z + b)u^T w \geq -1 \)
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>
</table>

- Base distribution: Uniform

- 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 transformation $x \mapsto z$ 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

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

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

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

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

has lower triangular structure. Determinant can be computed in **linear time**. Similarly, the Jacobian is upper triangular if \( x_i \) only depends on \( z_{\geq i} \).

**Next lecture**: Designing invertible transformations!