Deep Generative Models

Chapter 2: Data Generation Problem

Ali Bereyhi

ali.bereyhi@utoronto.ca

Department of Electrical and Computer Engineering

University of Toronto

Summer 2025

Beyond Text

In Chapter 1 . . .

> *We considered only text generation which is one possible modality*

A data could though come from different natures ≡ modalities

- *Text data typically has a form of linguistic modality*
  - ↳ *We can treat it as a sequence of unit objects*
- *Images incorporate their information in their visual configuration*
  - ↳ *We represent them as arrays of pixels*
- *In audio data information is incorporated temporally as well*
- *Video encodes information in both temporal and visual form*
- . . .

Beyond Text: *Generic Datatype*

At the end of day, each sample is presented in a mathematical form

- *We modeled text as a sequence tokens*
  - ↳ *It is a sequence of integers ≡ token indices*
- *Image are multi-channel tensors of pixel values*
  - ↳ *A CIFAR-10 image is a $3 \times 32 \times 32$ tensor of 8-bit pixels*
- *Audio files are sequences of framed samples*
  - ↳ *We sample with Nyquist rate and put them in fixed-size frames*
- *Videos are sequences of tensor frames*
  - ↳ *Each frame contains multiple image tensors*

Generic Datatype

We can denote a generic data sample by a mathematical object x

Data Space

Data Space

Data sample x comes from a known data space $\mathbb{X}$, i.e., $x \in \mathbb{X}$

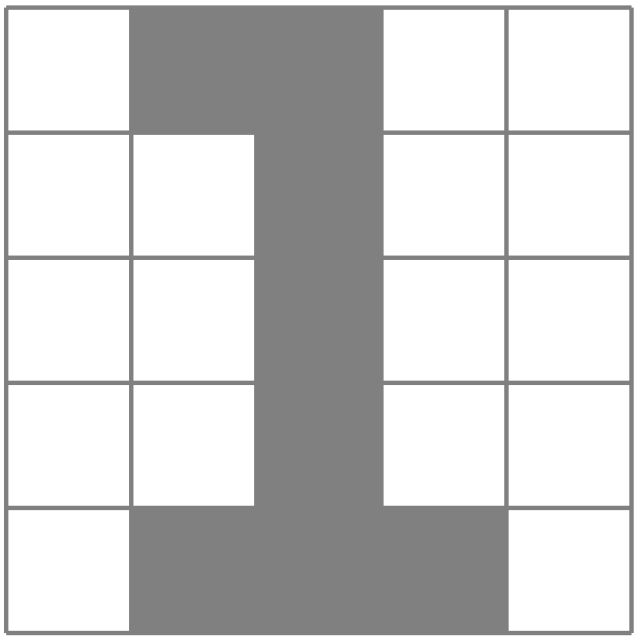
- *We have a text $x \equiv x_1, \ldots, x_T$ from a vocabulary of size I*
  - ↳ *Each token $x_t \in \{0, \ldots, I-1\}$*
  - ↳ *Data samples belong to data space $\mathbb{X} = \{0, \ldots, I-1\}^T$*
- *An $n \times n$ pixel RGB image contains three color-maps*
  - ↳ *Each color map contains $n \times n$ pixels*
  - ↳ *Each pixel is represented by 8 bits, i.e., $\{0, \ldots, 255\}$*
  - ↳ *Data sample belongs to data space $\mathbb{X} = \{0, \ldots, 255\}^{3 \times n \times n}$*

Example: 5×5 *Pixel Digits*

Our data is a 5×5 *pixel digit*

- *Each pixel is either* $0 \equiv$ *off or* $1 \equiv$ *on*

We can represent each sample by a matrix x *coming from* $\mathbb{X}\ \{0,1\}^{5\times 5}$



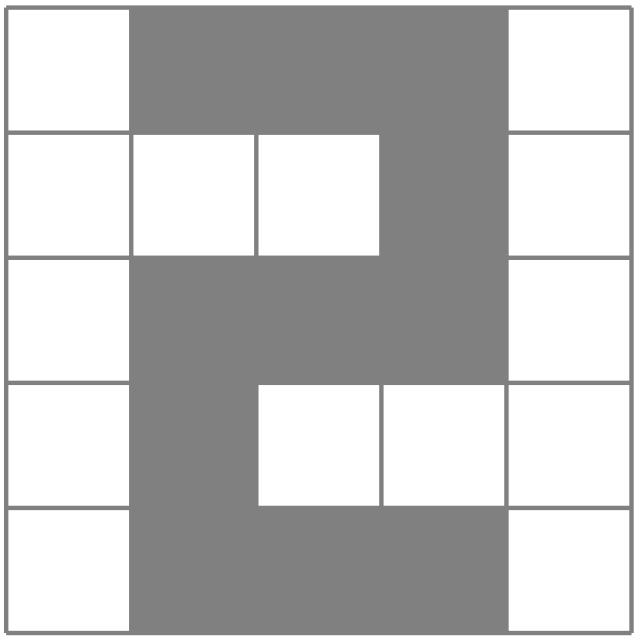
$$\rightsquigarrow \quad \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

Example: 5×5 *Pixel Digits*

Our data is a 5×5 *pixel digit*

- *Each pixel is either* $0 \equiv$ *off or* $1 \equiv$ *on*

We can represent each sample by a matrix x *coming from* $\mathbb{X}\ \{0,1\}^{5\times 5}$



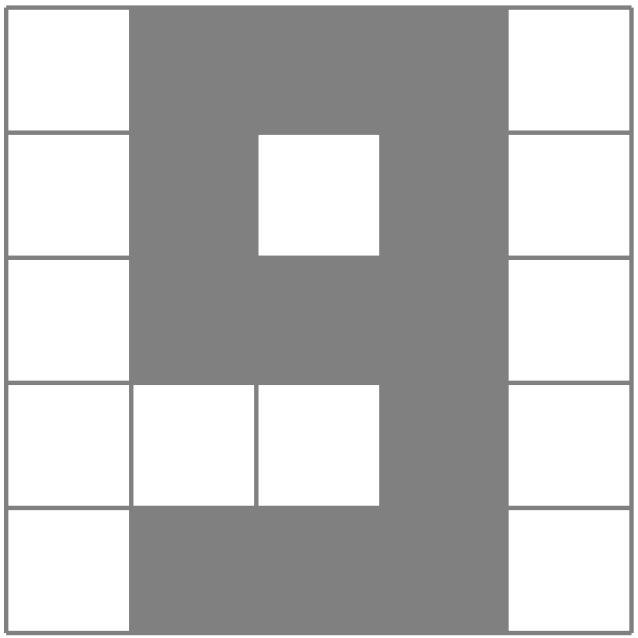
$$\rightsquigarrow \quad \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

Example: 5×5 *Pixel Digits*

Our data is a 5×5 *pixel digit*

- *Each pixel is either* $0 \equiv$ *off or* $1 \equiv$ *on*

We can represent each sample by a matrix x *coming from* $\mathbb{X}\ \{0,1\}^{5\times 5}$



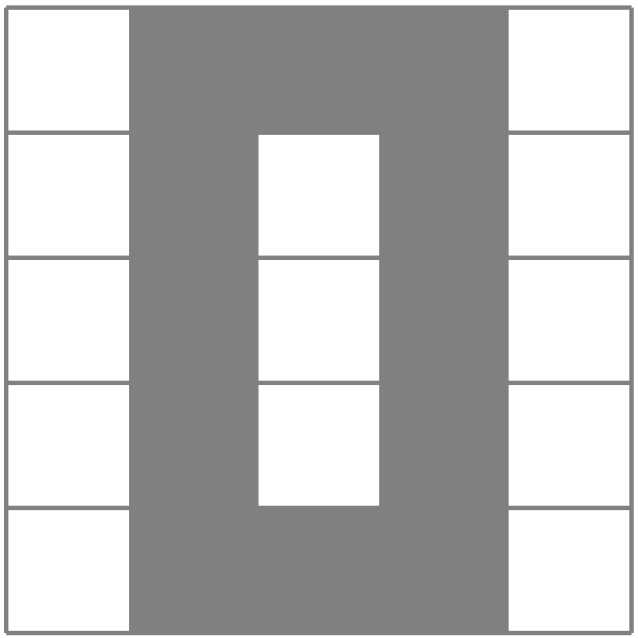
$$\rightsquigarrow \quad \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

Example: 5×5 *Pixel Digits*

Our data is a 5×5 *pixel digit*

- *Each pixel is either* $0 \equiv$ *off or* $1 \equiv$ *on*

We can represent each sample by a matrix x *coming from* $\mathbb{X}\ \{0,1\}^{5\times 5}$



$$\rightsquigarrow \quad \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

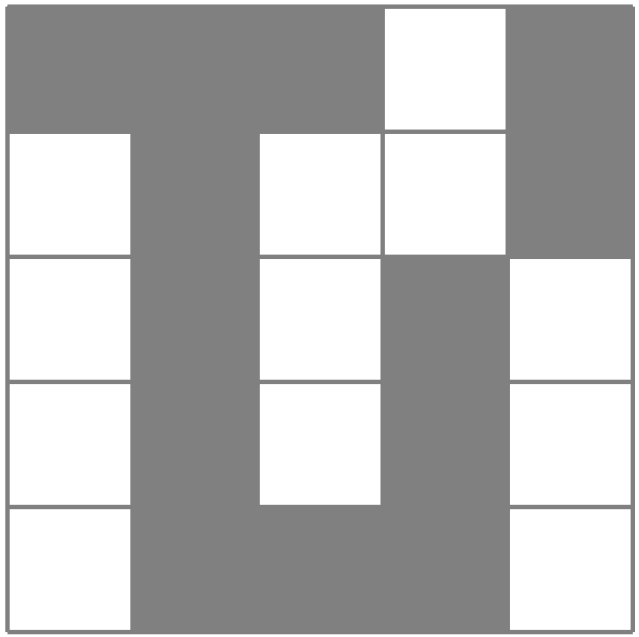
Valid Data Points

Though data space represents where data comes from . . .

not all points in the data space are valid data samples

Example: *For 5×5 pixel digits, the following sample is invalid*

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix} \rightsquigarrow$$



Valid Data Points $\equiv$ *Data Distribution*

+ *How can we then specify valid data points in the data space?*
- We can define a distribution for data

Data Distribution

Data distribution $P(x)$ is the distribution by which data samples are drawn from data space $\mathbb{X}$

+ *How does it help defining valid data points?*
– Well! The probability of each sample can define its validity
  ↳ *Invalid data points happen with probability zero*
  ↳ *Frequent data points happen with higher probability*
  ↳ *Non-frequent data points happen with less probability*

Data Distribution vs Dataset

Data distribution describes the statistical viewpoint on data

In this viewpoint: *each sample that we see*

- ↳ *is drawn randomly from data distribution*
- ↳ *this is indeed why we call it a data sample*

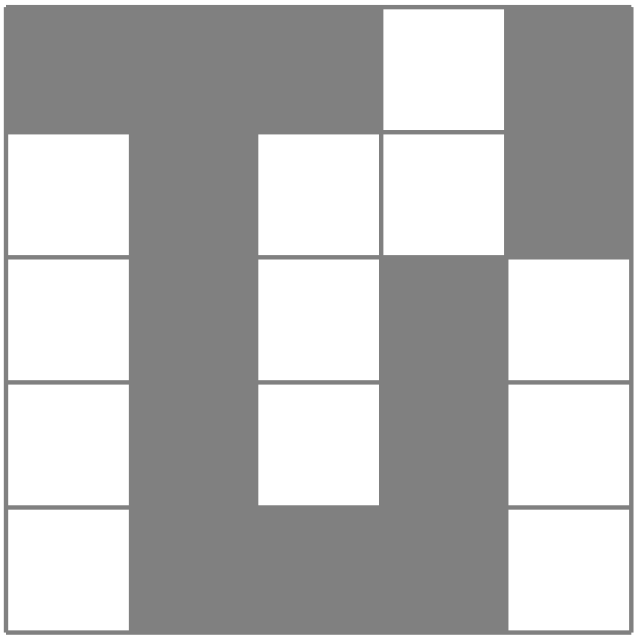
Dataset

A collection of samples drawn from data distribution $P(x)$

Example: 5×5 *Pixel Digits*

For 5×5 pixel digits

↳ *invalid matrices happen with probability zero*

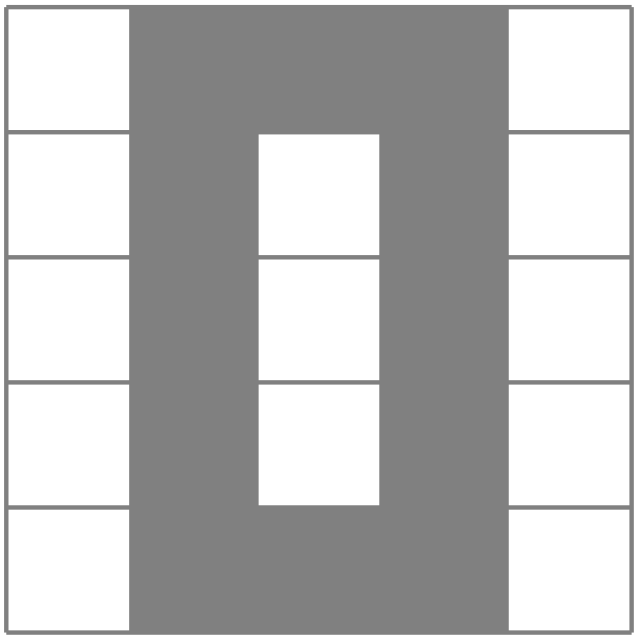


$$\rightsquigarrow \quad P(x = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix}) = 0$$

Example: 5×5 *Pixel Digits*

For 5×5 pixel digits: digits might be given at the same chances

↳ *valid matrices happen with the same probability*

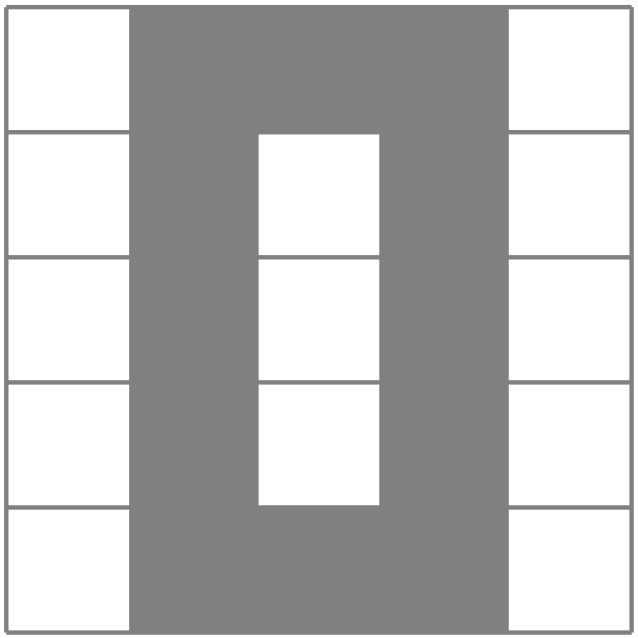


$$\rightsquigarrow \quad P(x = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix}) = 0.1$$

Example: 5×5 *Pixel Digits*

For 5×5 pixel digits: some digits might be given at higher chances

↳ *valid matrices happen with different non-zero probabilities*



$$\rightsquigarrow \quad P(x = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix}) = 0.3$$

Data Distribution: *Complexity*

+ *Wouldn't it be easier to work with the set of valid data-points instead of defining data distribution?*
- Absolutely No!

Data distribution enables us to capture

- *the pattern hidden in the data*
  - ↳ *E.g., different samples can occur wit different frequencies*
- *a feasible way to model intractable data complexity*
  - ↳ *Just think about a realistic dataset like CIFAR-10*

Example: *CIFAR-10*

In CIFAR-10, we have 32×32 RGB images with `uint8` *pixels labeled by*

$$\textit{airplanes} \equiv 0, \ldots, \textit{trucks} \equiv 9$$

A data sample is hence represented by

$$(x, y) \in \mathbb{X} = \{0, \ldots, 255\}^{3 \times 32 \times 32} \times \{0, \ldots, 9\}$$

+ *How many of such samples we have?*

$$\textit{\# possible points in } \mathbb{X} = 256^{3072} \cdot 10 = 2^{24576} \cdot 10 > 10^{2458}$$

We could have had all samples by now if we would have kept collecting

$$\textit{since big-bang with a rate more than } 10^{2440} \textit{ samples/sec!}$$

Data Distribution is *Everything!*

What we have in CIFAR-10 dataset is not even a drop of the ocean!

Defining data distribution helps us think about it rigorously, *e.g.*, *in CIFAR-10*

$P($  $, car) \neq 0$ $\qquad$ $P($  $, car) = 0$

Distribution is *All We Need*

Indeed, learning data distribution is the whole thing we do in machine learning!

Using this notion, whatever we've done is formulated in universal framework

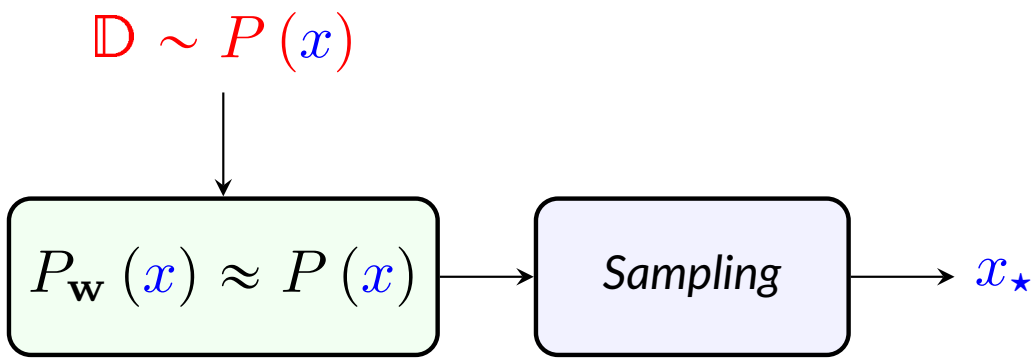
We study this universal framework in the next section

Problem of Data Generation: *Most Generic Form*

+ *Can we present the problem of data generation formally then?*
– Yes!

Data Generation

We look for a generation mechanism which can learn sampling from a data distribution after observing a limited set of samples drawn from that distribution



We will realize this mechanism with different tricks throughout the semester