Enhancing Conformal Prediction Using E-Test Statistics

"BB-predictor"

Alexander A. Balinsky and Alexander D. Balinsky

Cardiff University and Imperial College London

Contents

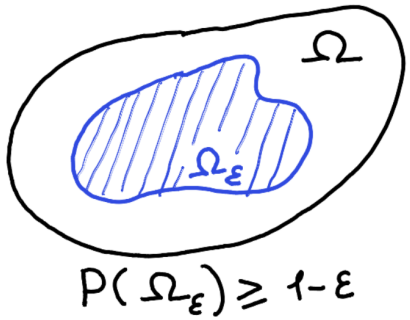
What is Conformal Prediction Really About?

Let (Ω, A, P) be a probability space.

What is Conformal Prediction Really About?

Let (Ω, A, P) be a probability space.

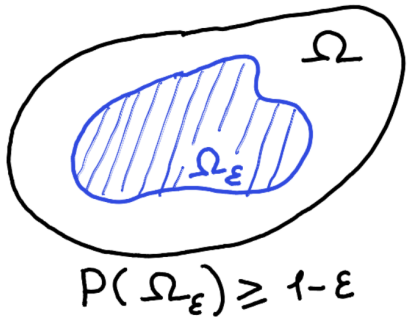
We are looking for a subspace $\Omega_\epsilon \subset \Omega$ with $P(\Omega_\epsilon) \geq 1 - \epsilon$:



What is Conformal Prediction Really About?

Let (Ω, A, P) be a probability space.

We are looking for a subspace $\Omega_\epsilon \subset \Omega$ with $P(\Omega_\epsilon) \geq 1 - \epsilon$:



Prediction:

Random sample in Ω with probability $\geq 1 - \epsilon$ will be in Ω_ϵ

Main Question:

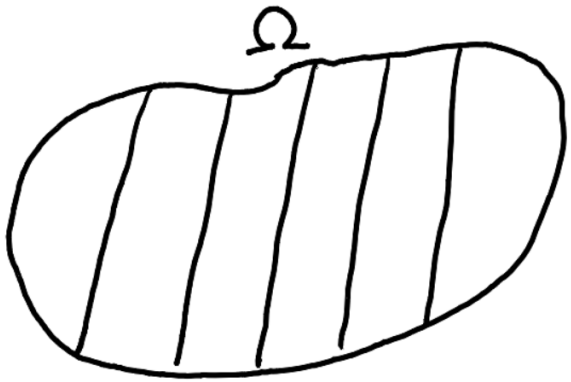
How to construct effectively Ω_ϵ ?

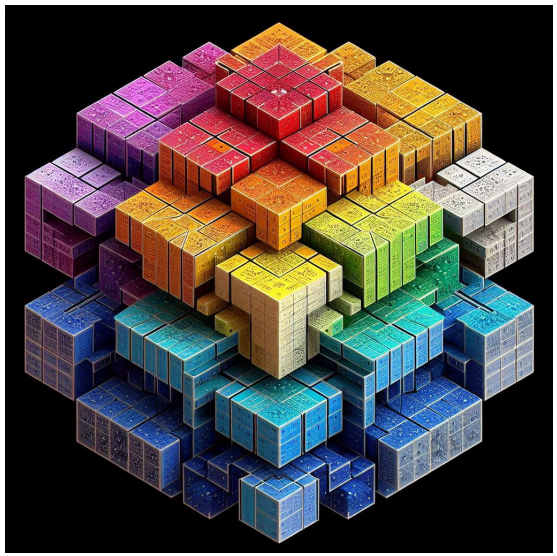
Main Question:

How to construct effectively Ω_ϵ ?

p-conformal approach:

Partition Ω into several parts of equal probabilities. Select some of them to have probability $\geq 1 - \epsilon$.

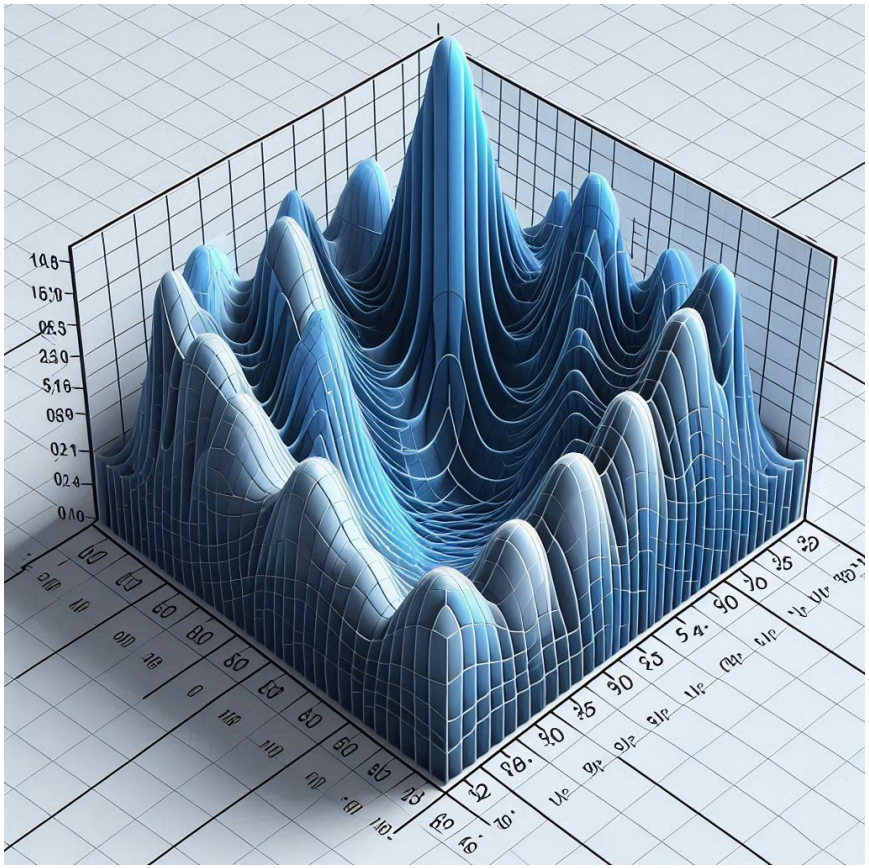




Our approach will be functional approach:

Construct a function $f : \Omega \to \mathbf{R}$ with

$$\Omega_\epsilon = \{\omega : \ f(\omega) \leq \beta(\epsilon)\}.$$



The next Lemma is the main mathematical tool of standard CP theory.

Lemma

Suppose $L_1, \ldots, L_{n+1}$ *are exchangeable random variables. Set*

$$U = \#\{i = 1, \ldots, n+1 : \ L_i \geq L_{n+1}\},$$

i.e., the number of L *that are at least as large as the last one. Then*

$$P\left\{\frac{U}{n+1} > \epsilon\right\} \geq 1 - \epsilon$$

for all $\epsilon \in [0, 1]$

Main Result: BB-predictor.

The key idea is very simple and is a straightforward application of **Markov's inequality**: if E is a non-negative random variable with the expectation $\mathbf{E}(E) \leq 1$, then

$$P(E \geq 1/\alpha) \leq \alpha,$$

for any positive α. If, for example, $\alpha = 0.05$, then the event $E \geq 20$ will have probability less than 5%.

The main theoretical result of our article is the following theorem

Theorem

Suppose $L_1, \ldots, L_{n+1}$ *are exchangeable non-negative random variables. Set*

$$F = \frac{L_{n+1}}{\left(\sum_{j=1}^{n+1} L_j\right)/(n+1)}.$$

Then the expectation $\mathbf{E}(F) = 1$ *and*

$$P\{F \geq 1/\alpha\} \leq \alpha,$$

for any positive α*.*

Corollary

Suppose $L_1, \ldots, L_{n+1}$ *are exchangeable non-negative random variables. Then for any positive* α

$$P\left\{\frac{L_{n+1}}{\frac{L_1+\ldots+L_n}{n}} \geq \frac{1}{\alpha}\left(\frac{1}{1+\frac{1-1/\alpha}{n}}\right)\right\} \leq \alpha. \tag{1}$$

So, we can now introduce our new **BB-predictor** (bounded from the below predictor)

$$P\left\{L_{n+1} \geq \frac{1}{\alpha}\left(\frac{1}{1+\frac{1-1/\alpha}{n}}\right) \times \frac{L_1 + \ldots + L_n}{n}\right\} \leq \alpha. \qquad (2)$$

Proof of the Main Theorem:

Let us introduce the following random variables

$$F_i = \frac{L_i}{\left(\sum_{j=1}^{n+1} L_j\right)/(n+1)}, \qquad i = 1, 2, \ldots, n+1.$$

Due to the exchangeability of $L_1, \ldots, L_{n+1}$, all random variables $F_i,\ i = 1, \ldots, n+1$ are identically distributed, so they all have the same expectation $\mathbf{E}(F_i)$.

$$\mathbf{E}(F_1 + F_2 + \ldots + F_{n+1}) = (n+1) \times \mathbf{E}(F_{n+1}) = (n+1) \times \mathbf{E}(F).$$

From definition of F_i we can see that $F_1 + F_2 + \ldots + F_{n+1} = n+1$. Thus, $(n+1) \times \mathbf{E}(F) = n+1$, and $\mathbf{E}(F) = 1$.

Important Remark:

In the proof of Theorem, we utilize *exchangeability* solely to demonstrate that all random variables $F_i,\ i = 1, \ldots, n+1$ are identically distributed. This property holds true even under the less stringent assumption of *cycle invariance* of the joint distribution of $L_1, \ldots, L_{n+1}$.

We can also have the same result for any collection of random variables $L_i,\ i \in I$ with invariance under transitive action of a group on the index set I.

Examples: Lattice statistical models (Ising type), images, spin-glasses, models on regular graphs, etc.

The simplest possible scenario: $L_i = LossFunction(z_i)$.

For our numerical experiments, we will employ the MNIST (Modified National Institute of Standards and Technology) database of handwritten digits. This database is a popular choice for training and testing various image processing systems and machine learning models. It comprises 60,000 training images and 10,000 testing images. Each image is a grayscale representation of a handwritten digit, sized at 28×28 pixels. Let's examine some examples:

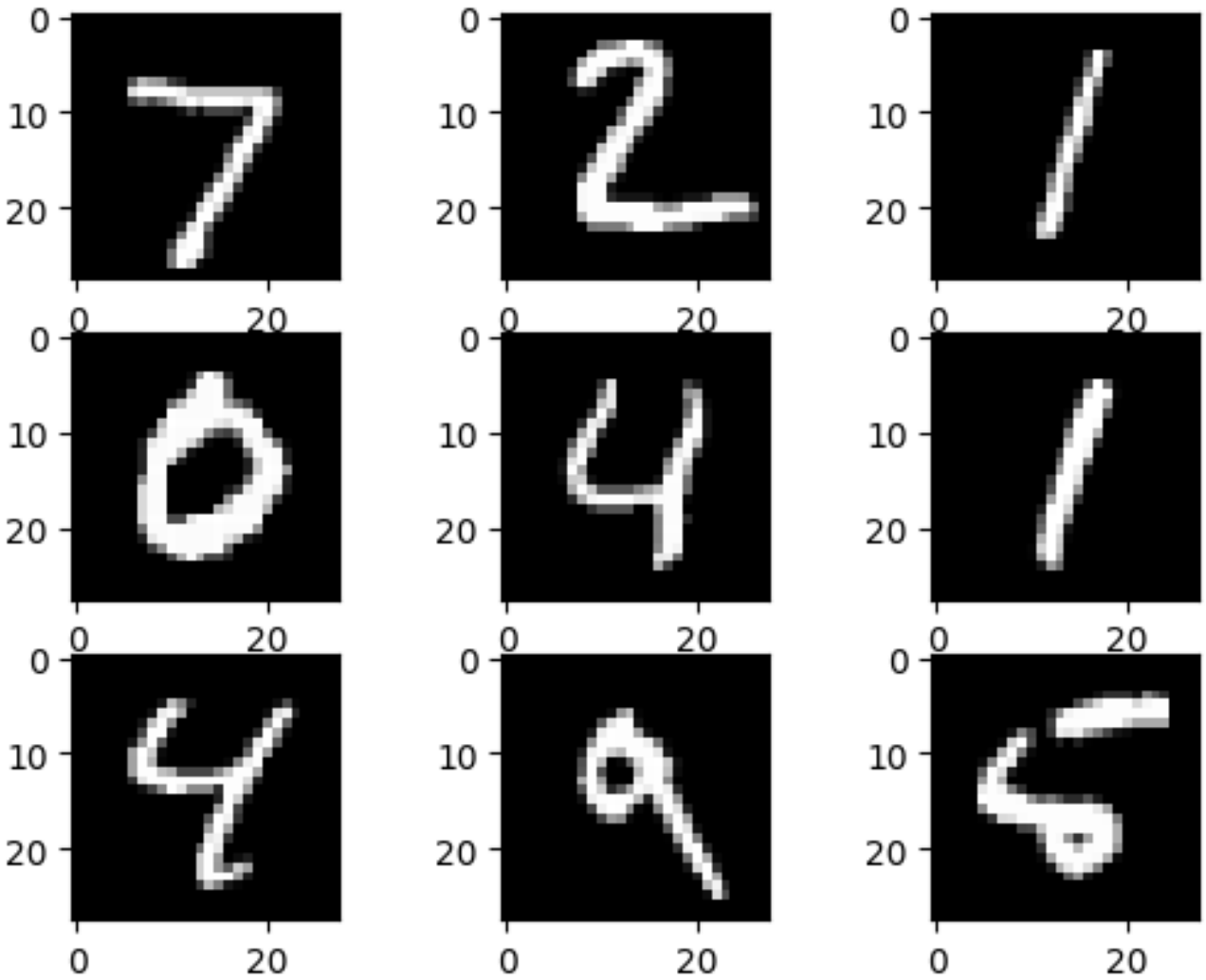


Figure: The first nine images from the MNIST database of handwritten digits.

We will train a very simple TensorFlow Keras model Figure 2

```
model = tf.keras.models.Sequential([
  tf.keras.layers.Flatten(input_shape=(28, 28)),
  tf.keras.layers.Dense(128, activation='relu'),
  tf.keras.layers.Dropout(0.2),
  tf.keras.layers.Dense(10)
])
```

Figure: Our model.

with *SparseCategoricalCrossentropy* loss function and '*adam*' optimizer.
For reproducibility we use $tf.random.set\_seed(2024)$.

Upon training the model with the training images and labels over 25 epochs, we achieved an accuracy of 99.22% on the training data and 98.17% on the test data. Thus, the model appears to be quite effective. Let's present some instances where the predictions were incorrect in Figure:

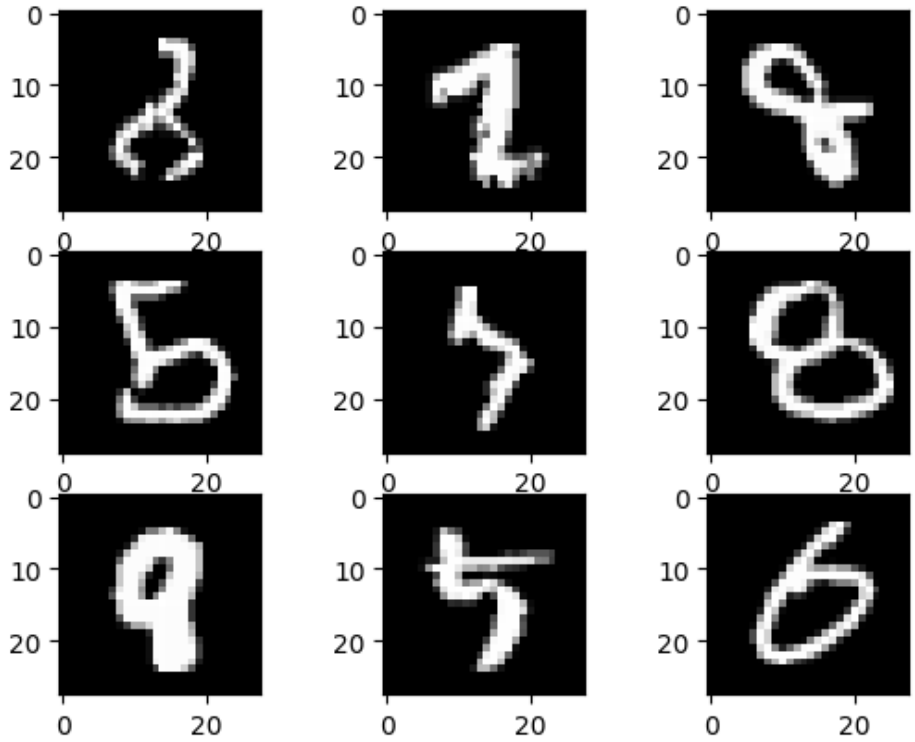
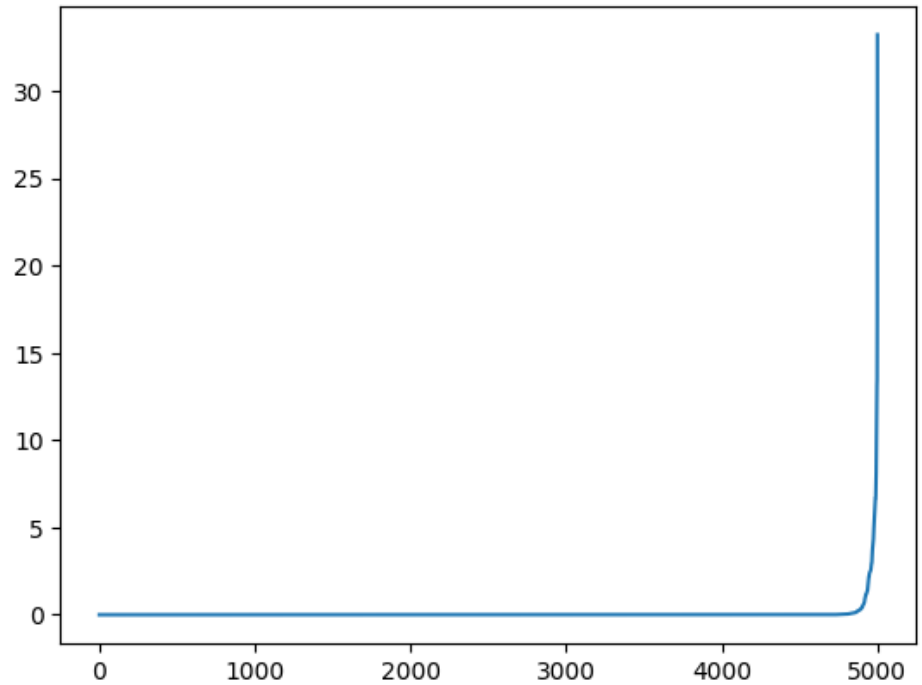


Figure: Some images that the model predicts the wrong labels.

Once the model is trained and the loss function is established, we can shift our focus away from the training data and concentrate solely on the test data. We partition the test data randomly into two subsets: the *CalibrationSet*, which comprises 50% of the test data, and the *ConformalPredictionTestSet*, which also makes up 50% of the test data. For the *CalibrationSet*, a plot of the sorted values of the loss function is depicted in Figure:



Inductive Conformal Prediction with e-test statistics

With $\alpha = 0.05$, $n = 5000$, the *BB-predictor* (2) tells us that for an (image, label) with the image from from the *ConformalPredictionTestSet*.

$$P\{LossFunction \geq 1.5002\} < 0.05.$$

Applying this criterion to the *ConformalPredictionTestSet*, we obtain that with probability more than 95%

- Examples without labels: 0
- Examples with one label: 4944
- Examples with more than one label: 56

Let us demonstrate some examples with multiple labels:

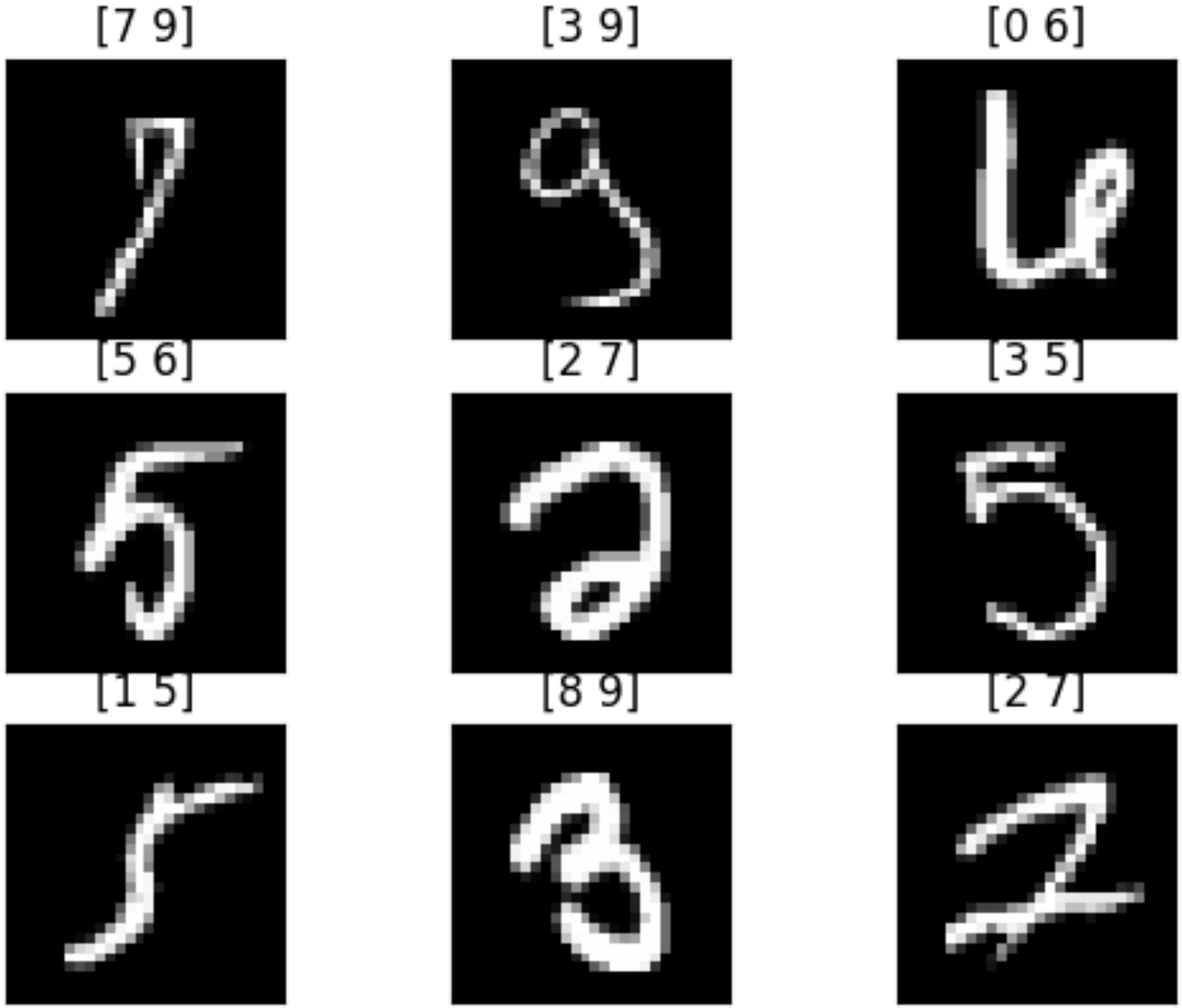


Figure: Some images that have multiple labels prediction with their labels.

Inductive Conformal Prediction with p-value statistics

Given $\epsilon = 0.05$, $n = 5000$, Lemma 1 instructs us to compute the loss of the element at index 4750 in the ordered list of losses for images from the *CalibrationSet*. This value is found to be 0.0247. When we apply this to the *ConformalPredictionTestSet*, we find that with a probability exceeding 95%,

- Examples without labels: 238
- Examples with one label: 4762
- Examples with more than one label: 0

Some examples without labels are depicted in

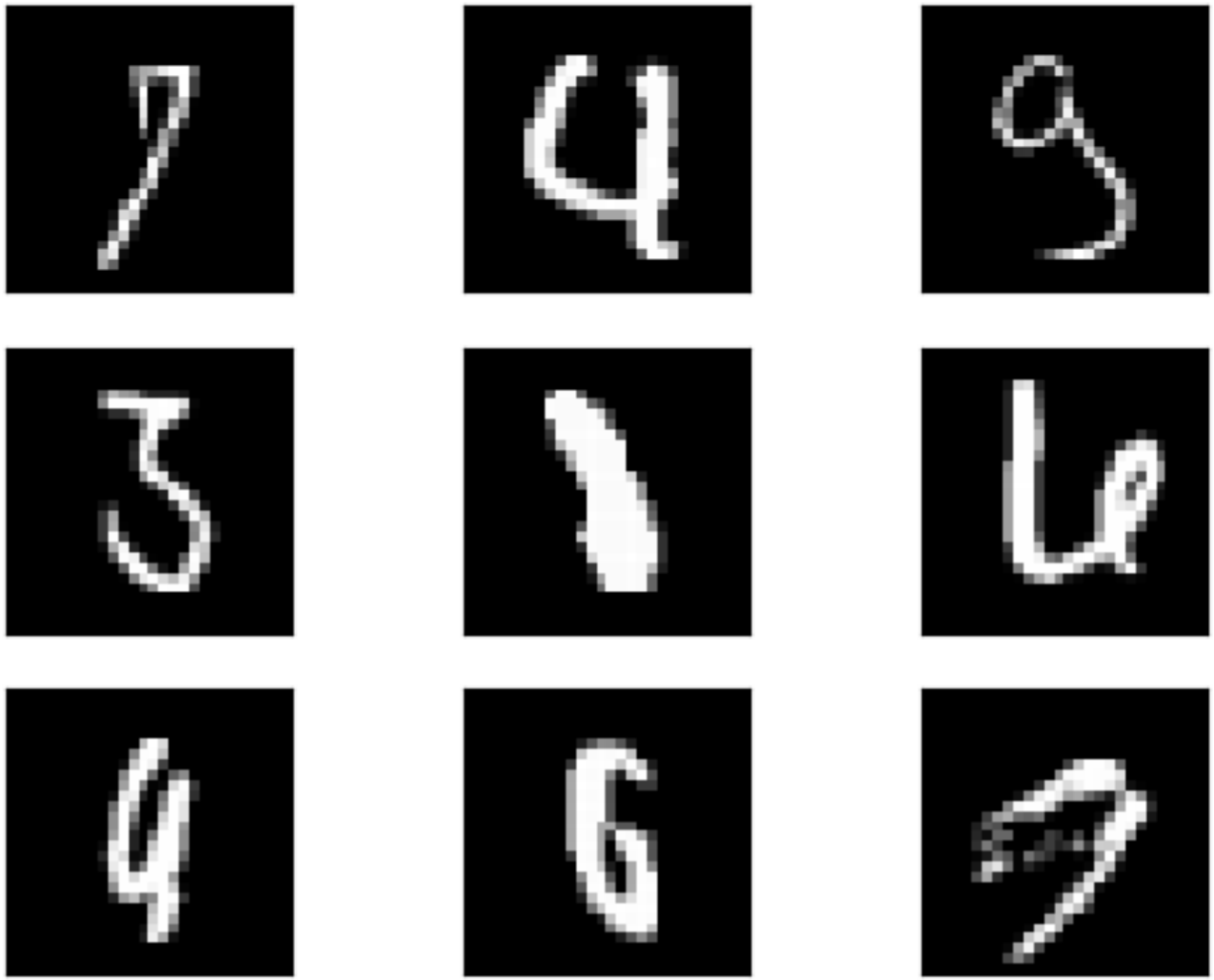


Figure: Some images that do not have labels.

Q&A