

# *Machine Perception*

*Cristian Perez Jensen*

*May 12, 2024*

Note that these are not the official lecture notes of the taught course, but only notes written by me. As such, it might contain mistakes. The source code can be found at [github.com/cristianpjensen/eth-cs-notes](https://github.com/cristianpjensen/eth-cs-notes). If you find a mistake, please create an issue or open a pull request.

## Contents

1	Neural networks	1
1.1	Multi-layer perceptron	1
1.2	Loss functions	1
1.3	Backpropagation	2
1.4	Activation functions	3
1.5	Universal approximation theorem	4
2	Convolutional neural networks	5
2.1	Convolution	5
2.2	Convolutional neural network	5
3	Fully convolutional neural network	8
3.1	Upsampling methods	8
3.2	U-net	8
4	Recurrent neural networks	10
4.1	Elman RNN	10
4.2	Long-short term memory	11
4.3	Gradient clipping	13
5	Generative models	14
6	Autoencoders	15
6.1	Linear autoencoders	15
6.2	Non-linear autoencoders	15
6.3	Variational autoencoders	16
6.4	$\beta$ -VAE	17
7	Autoregressive models	19
7.1	Fully visible sigmoid belief network	19
7.2	Neural autoregressive density estimator	20
7.3	Masked autoencoder distribution estimation	20
7.4	Generating images	21
7.5	Generating audio	22
7.6	Variational RNN	22
7.7	Transformers	23
8	Normalizing flows	24
8.1	Change of variables	24
8.2	Coupling layers	25
8.3	Composing transformations	25
8.4	Training and inference	26
8.5	Architectures	26
9	Generative adversarial networks	28
9.1	Theoretical analysis	28
9.2	Training	31
10	Diffusion models	33

10.1	<i>Diffusion step</i>	33
10.2	<i>Denoising step</i>	34
10.3	<i>Training</i>	34
10.4	<i>Guidance</i>	36
10.5	<i>Latent diffusion models</i>	36
11	<i>Reinforcement learning</i>	37
11.1	<i>Monte Carlo methods</i>	40
11.2	<i>Temporal difference learning</i>	40
11.3	<i>Deep reinforcement learning</i>	41

## List of symbols

$\doteq$	Equality by definition
$\approx$	Approximate equality
$\propto$	Proportional to
$\mathbb{N}$	Set of natural numbers
$\mathbb{R}$	Set of real numbers
$i : j$	Set of natural numbers between $i$ and $j$ . I.e., $\{i, i+1, \dots, j\}$
$\mathbb{1}\{\text{predicate}\}$	Indicator function (1 if predicate is true, otherwise 0)
$\boldsymbol{v} \in \mathbb{R}^n$	$n$ -dimensional vector
$\boldsymbol{M} \in \mathbb{R}^{m \times n}$	$m \times n$ matrix
$\boldsymbol{T} \in \mathbb{R}^{d_1 \times \dots \times d_n}$	Tensor
$\boldsymbol{M}^\top$	Transpose of matrix $\boldsymbol{M}$
$\boldsymbol{M}^{-1}$	Inverse of matrix $\boldsymbol{M}$
$\det(\boldsymbol{M})$	Determinant of $\boldsymbol{M}$
$\frac{d}{dx}f(x)$	Ordinary derivative of $f(x)$ w.r.t. $x$ at point $x \in \mathbb{R}$
$\frac{\partial}{\partial x}f(x)$	Partial derivative of $f(x)$ w.r.t. $x$ at point $x \in \mathbb{R}^n$
$\nabla_x f(x) \in \mathbb{R}^n$	Gradient of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ at point $x \in \mathbb{R}^n$
$J_x f(x) \in \mathbb{R}^{n \times m}$	Jacobian of $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ at point $x \in \mathbb{R}^n$
$\nabla_x^2 f(x) \in \mathbb{R}^{n \times n}$	Hessian of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ at point $x \in \mathbb{R}^n$
$\boldsymbol{\theta} \in \Theta$	Parametrization of a model, where $\Theta$ is a compact subset of $\mathbb{R}^K$
$\mathcal{X}$	Input space
$\mathcal{Y}$	Output space
$\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$	Labeled training data

## 1 Neural networks

### 1.1 Multi-layer perceptron

The original perceptron [Rosenblatt, 1958] was a single layer perceptron with the following non-linearity,

$$\sigma(x) \doteq \mathbb{1}\{x > 0\}.$$

The classification of a single point can then be written as

$$\hat{y} = \mathbb{1}\{\mathbf{w}^\top \mathbf{x} + b > 0\}.$$

The learning algorithm then iteratively updates the weights for a data point that was classified incorrectly,

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \eta \underbrace{(y_i - \hat{y}_i)}_{\text{residual}} \mathbf{x}_i,$$

where  $\eta$  is the learning rate. This is essentially gradient descent with a Hinge loss, where if  $y < \hat{y}$ , then we decrease the weights, while if  $y > \hat{y}$ , we increase the weights. If the data is linearly separable, the perceptron converges in finite time.

The problem with the single-layer perceptron was that it could not solve the XOR problem; see Figure 2. This can be solved by introducing hidden layers,

$$\hat{y} = \sigma(\mathbf{W}_k \sigma(\mathbf{W}_{k-1} \cdots \sigma(\mathbf{W}_1 \mathbf{x}))).$$

We call this architecture a multi-layer perceptron (MLP); see Figure 3. We then want to estimate the parameters  $\boldsymbol{\theta} = \{\mathbf{W}_1, \dots, \mathbf{W}_k, \mathbf{b}_1, \dots, \mathbf{b}_k\}$ , using an optimization algorithm such as gradient descent, which we call “learning”. We can compute the gradient by backpropagation, which we will see in a later section.

### 1.2 Loss functions

We need an objective to optimize for. We typically call this objective function the loss function, which we minimize. In classification, we typically optimize the *maximum likelihood estimate* (MLE),

$$\begin{aligned} \operatorname{argmax}_{\boldsymbol{\theta}} p(\mathcal{D} \mid \boldsymbol{\theta}) &\stackrel{\text{iid}}{=} \operatorname{argmax}_{\boldsymbol{\theta}} \prod_{i=1}^n p(y_i \mid \boldsymbol{\theta}) \\ &= \operatorname{argmin}_{\boldsymbol{\theta}} -\log \prod_{i=1}^n p(y_i \mid \boldsymbol{\theta}) \\ &= \operatorname{argmin}_{\boldsymbol{\theta}} -\sum_{i=1}^n \log p(y_i \mid \boldsymbol{\theta}). \end{aligned}$$

If the model predicts the parameters of a Bernoulli distribution,<sup>1</sup> MLE is



**Figure 1.** Computation graph of a perceptron [Rosenblatt, 1958], where  $\sigma(x) = \mathbb{1}\{x > 0\}$ .



**Figure 2.** XOR problem. As can be seen, the data is not linearly separable, and thus not solvable by the perceptron.



**Figure 3.** Example multi-layer perceptron architecture.

log is monotonic.

<sup>1</sup> I.e.,  $y \in \{0, 1\}$ , but it predicts the probability of  $y = 1$  with  $\hat{y} \in [0, 1]$ .

equivalent to binary cross-entropy,

$$\begin{aligned}\mathcal{L}(\theta) &= - \sum_{i=1}^n \log \text{Ber}(y_i | \hat{y}_i \doteq f(x_i | \theta)) \\ &= - \sum_{i=1}^n \log \hat{y}_i^{y_i} (1 - \hat{y}_i)^{1-y_i} \\ &= - \sum_{i=1}^n y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i).\end{aligned}$$

This loss is minimized if  $y_i = \hat{y}_i$  for all  $i \in [n]$ ; see Figure 4. We can extend this to multi-class classification by using the softmax and a categorical distribution.

If we choose the model to be Gaussian, we end up minimizing the mean-squared error,

$$\mathcal{L}(\theta) = \sum_{i=1}^n \|y_i - f_{\theta}(x_i)\|_2^2.$$

Furthermore, the Laplacian distribution yields minimizing the  $\ell_1$  norm,

$$\mathcal{L}(\theta) = \sum_{i=1}^n \|y_i - f_{\theta}(x_i)\|_1.$$

If we have prior information about the weights  $p(\theta)$ , we could also optimize for the *maximum a posteriori* (MAP),

$$\begin{aligned}\arg\max_{\theta} p(\theta | \mathcal{D}) &= \arg\max_{\theta} p(\theta) p(\mathcal{D} | \theta) \\ &\stackrel{\text{iid}}{=} \arg\max_{\theta} p(\theta) \prod_{i=1}^n p(y_i | \theta) p(\theta) \\ &= \arg\min_{\theta} - \log \left( p(\theta) \prod_{i=1}^n p(y_i | \theta) \right) \\ &= \arg\min_{\theta} - \log p(\theta) - \sum_{i=1}^n \log p(y_i | \theta)\end{aligned}$$

Note that MAP and MLE are equivalent if  $p(\theta)$  is uniform over the domain of weights. Assuming a Gaussian prior distribution over  $\theta$ , MAP yields Ridge regression,

$$\mathcal{L}'(\theta) = \mathcal{L}(\theta) + \lambda \|\theta\|_2^2.$$

### 1.3 Backpropagation

Typically, we cannot find the optimal parameters  $\theta^*$  in closed form, so we must use an optimization algorithm. Optimization algorithms, such as gradient descent, typically require computing the gradient w.r.t. the parameters. Backpropagation is an algorithm for computing the gradient

$f$  is a model that outputs the Bernoulli parameter. Note that this parameter must be in  $[0, 1]$ , thus we use a sigmoid non-linearity,

$$\sigma(x) = \frac{1}{1 + \exp(-x)}.$$



**Figure 4.** Loss function for if  $y_i = 0$  and  $y_i = 1$  in binary cross entropy.

of any function, given that we have access to the derivatives of the primitive functions it consists of.<sup>2</sup> It then computes the gradient by making use of dynamic programming, the chain rule, and sum rule.

Let  $\hat{y}_i < y_i$ , then we want to increase  $\hat{y}_i$  to match  $y_i$ . Furthermore, consider the following loss function,

$$\ell = \frac{1}{2}(y - \hat{y})^2, \quad \hat{y} = \sigma(\mathbf{w}^\top \mathbf{h}),$$

and the MLP architecture, then intuitively, we can do two things to increase  $\hat{y}_i$ :

1. We could increase the weight connected to a node with a high activation value in the previous hidden layer. Typically, we optimize by moving the weights in the direction of the gradient,

$$\frac{\partial \mathcal{L}}{\partial w_k} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{w}^\top \mathbf{h}} \frac{\partial \mathbf{w}^\top \mathbf{h}}{\partial w_k} = (y - \hat{y})\sigma'(\mathbf{w}^\top \mathbf{h})h_k.$$

This matches our intuition, because the amount that we increase  $w_k$  by is proportional to the activation connected to that weight  $h_k$ ;

2. Or, we could increase the activation that is connected to a strong weight in the previous hidden layer. Again, we move the weights in the direction of the gradient,

$$\frac{\partial \mathcal{L}}{\partial h_k} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{w}^\top \mathbf{h}} \frac{\partial \mathbf{w}^\top \mathbf{h}}{\partial h_k} = (y - \hat{y})\sigma'(\mathbf{w}^\top \mathbf{h})w_k.$$

This also matches our intuition, since the amount that we want to increase  $h_k$  by is proportional to  $w_k$ . But, we cannot increase  $h_k$  directly, but we can update the weights connected to  $h_k$ , which brings us back to the first case. This update will be proportional to  $\frac{\partial \mathcal{L}}{\partial h_k}$  by the chain rule. In this way, we can recursively update all the weights using gradient information.

Thus, using first-order methods that iteratively move the weights in the direction of the gradient should work well. Gradient descent iteratively updates the parameters by the following,

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}),$$

until the gradient is small.

#### 1.4 Activation functions

In MLPs, the activation function should be non-linear, or the resulting MLP is just an affine mapping with extra steps. This is because the product of affine mappings are themselves affine mappings.

<sup>2</sup> For example, to compute the gradient of  $f(\mathbf{x}, \mathbf{y}) = \sigma(\mathbf{x}^\top \mathbf{y})$ , we would need access to  $\frac{d}{dx}\sigma(x)$ ,  $\frac{\partial}{\partial \mathbf{x}}\mathbf{x}^\top \mathbf{y}$ , and  $\frac{\partial}{\partial \mathbf{y}}\mathbf{x}^\top \mathbf{y}$ .

### 1.5 Universal approximation theorem

**Theorem 1** (Universal approximation theorem [Hornik et al., 1989]). Let  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  be a non-constant, bounded, and continuous activation function. Let  $I_m$  denote the  $m$ -dimensional unit hypercube  $[0, 1]^m$  and the space of real-valued function on  $I_m$  is denoted by  $\mathcal{C}(I_m)$ .

Let  $g \in \mathcal{C}(I_m)$  be any function in the hypercube. Let  $\epsilon > 0, n \in \mathbb{N}, v_i, b_i \in \mathbb{R}, w_i \in \mathbb{R}^m$  for  $i \in [n]$ , then

$$g(\mathbf{x}) \approx f_{\theta}(\mathbf{x}) = \sum_{i=1}^n v_i \sigma(w_i^{\top} \mathbf{x} + b_i),$$

where  $|f_{\theta}(\mathbf{x}) - g(\mathbf{x})| < \epsilon$  for all  $\mathbf{x} \in I_m$ .

The universal approximation theorem holds for any single hidden layer network. However, this hidden layer may need to have infinite width to approximate  $f$ . In practice, deeper networks work better than wider networks.



## 2 Convolutional neural networks

When dealing with high-dimensional data, such as images, it is not practical to work with MLPs, because the amount of parameters would explode in size.<sup>3</sup> By making use of the locality of images, we can drastically decrease the number of parameters.

### 2.1 Convolution

The *correlation* operator takes a filter  $\mathbf{K}$ , moves it along the entire image, and outputs the patch-wise multiplication, for each patch of the same size as the filter. It is defined as follows,

$$(\mathbf{K} \star \mathbf{I})[i, j] = \sum_{m=-k}^k \sum_{n=-k}^k \mathbf{K}[m, n] \mathbf{I}[i + m, j + n].$$

The *convolution* operator is very similar. The only difference is that the kernel is mirrored in a convolution,

$$(\mathbf{K} \star \mathbf{I})[i, j] = \sum_{m=-k}^k \sum_{n=-k}^k \mathbf{K}[m, n] \mathbf{I}[i - m, j - n].$$

Theoretically, the convolution operator is more useful, because it is commutative.<sup>4</sup> In practice with neural networks, it does not matter, since the weights will just be updated to be the same, except that they are mirrored. Thus, we will only be referring to the convolution from now on.

A convolution operator  $C$  is a linear, shift-equivariant transformation, *i.e.*,

$$\begin{aligned} C(\alpha \mathbf{x} + \beta) &= \alpha C(\mathbf{x}) + \beta \\ T_t(C(\mathbf{x})) &= C(T_t(\mathbf{x})). \end{aligned}$$

Since convolutions are linear, discrete convolutions can be implemented using matrix multiplication,

$$\mathbf{K} \star \mathbf{I} = \begin{bmatrix} k_1 & 0 & 0 & \cdots & 0 \\ k_2 & k_1 & 0 & \cdots & 0 \\ k_3 & k_2 & k_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & k_m \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \\ I_3 \\ \vdots \\ I_n \end{bmatrix}.$$

### 2.2 Convolutional neural network

A *convolutional neural network* (CNN) [Krizhevsky et al., 2012] is composed of a sequence of *convolutional layers* and *pooling layers*, followed by a final *dense layer* (MLP).

<sup>3</sup> Mapping a  $256 \times 256 \times 3$  input image to a 1-dimensional output would already require nearly 2 million parameters.

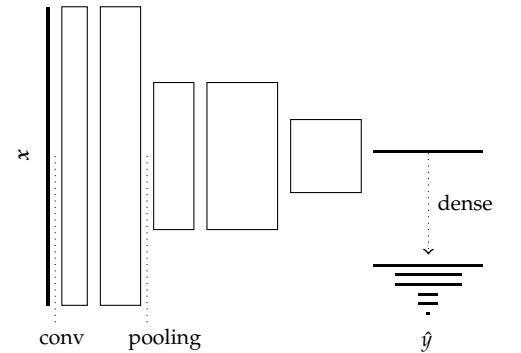


**Figure 5.** Illustration of applying a correlation to a pixel.

<sup>4</sup>  $\mathbf{I} \star \mathbf{K} = \mathbf{K} \star \mathbf{I}$ , but  $\mathbf{I} \star \mathbf{K} \neq \mathbf{K} \star \mathbf{I}$ .

Linearity.

Translation equivariant.



**Figure 6.** Example schematic of a CNN architecture.

This architecture is loosely inspired by the brain, which, at a high level, first extracts high-level features and then more and more specific features. Furthermore, the HMAX model of the brain distinguishes between simple and complex cells, which correspond to the linear layer and max-pool layer, respectively.

**Convolutional layer.** A convolutional layer applies many filters  $\mathbf{W} \in \mathbb{R}^{C_{\text{in}} \times C_{\text{out}} \times k \times k}$  to an input image  $\mathbf{Z} \in \mathbb{R}^{C_{\text{in}} \times H \times W}$  by convolution,

$$\mathbf{z}_j^{(\ell)} = \sum_{k=1}^{C_{\text{in}}} \mathbf{w}_{kj}^{(\ell)} * \mathbf{z}_k^{(\ell-1)} + b_j, \quad j \in [C_{\text{out}}].$$

This filter is learned. Hence, we need to derive its derivative. We will focus on the single filter case ( $C_{\text{in}} = C_{\text{out}} = 1$ ) for simplicity, whose forward pass is computed by

$$z^{(\ell)}[i, j] = \sum_{m=-k}^k \sum_{n=-k}^k w^{(\ell)}[m, n] z^{(\ell-1)}[i - m, j - n] + b.$$

We can express the derivative of the cost function  $\mathcal{L}$  w.r.t. the output of the  $(\ell - 1)$ -th layer as the following,

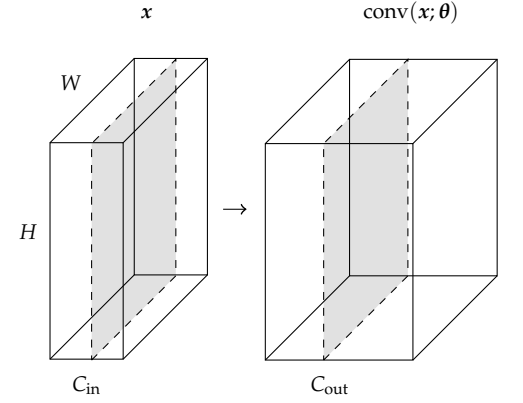
$$\begin{aligned} \delta^{(\ell-1)}[i, j] &= \frac{\partial \mathcal{L}}{\partial z^{(\ell-1)}[i, j]} \\ &= \sum_{i'} \sum_{j'} \frac{\partial \mathcal{L}}{\partial z^{(\ell)}[i', j']} \frac{\partial z^{(\ell)}[i', j']}{\partial z^{(\ell-1)}[i, j]} \\ &= \sum_{i'} \sum_{j'} \delta^{(\ell)}[i', j'] \frac{\partial}{\partial z^{(\ell-1)}[i, j]} \sum_m \sum_n w^{(\ell)}[m, n] z^{(\ell-1)}[i' - m, j' - n] + b \\ &= \sum_{i'} \sum_{j'} \delta^{(\ell)}[i', j'] w^{(\ell)}[i' - i, j' - j]. \end{aligned}$$

From this, we can see that we can compute all values of  $\delta^{(\ell-1)}$  by a single convolution,

$$\delta^{(\ell-1)} = \delta^{(\ell)} * \text{Rot}_{180}(\mathbf{W}^{(\ell)}) = \delta^{(\ell)} \star \mathbf{W}^{(\ell)}.$$

Using this value, we can compute the derivative w.r.t. the weights, which we need for the parameter update in algorithms such as gradient descent,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial w^{(\ell)}[m, n]} &= \sum_i \sum_j \frac{\partial \mathcal{L}}{\partial z^{(\ell)}[i, j]} \frac{\partial z^{(\ell)}[i, j]}{\partial w^{(\ell)}[m, n]} \\ &= \sum_i \sum_j \delta^{(\ell)}[i, j] \frac{\partial}{\partial w^{(\ell)}[m, n]} \sum_{m'} \sum_{n'} w^{(\ell)}[m', n'] z^{(\ell-1)}[i - m', j - n'] + b \\ &= \sum_i \sum_j \delta^{(\ell)}[i, j] z^{(\ell-1)}[i - m, j - n]. \end{aligned}$$



**Figure 7.** Schematic of a convolutional layer. Each input-output channel pair has its own kernel, so  $\theta$  has  $K \times K \times C_{\text{in}} \times C_{\text{out}}$  parameters.

Again, this has the form of a convolution, thus we can compute all derivatives of  $\mathbf{W}^{(\ell)}$  by convolution,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(\ell)}} = \delta^{(\ell)} * \mathbf{Z}^{(\ell-1)}.$$

As shown, we can use convolutions for both computing the forward pass, as well as the backward pass. Thus, we first do the forward pass, then compute all  $\delta^{(\ell)}$  for all layers  $\ell$  by convolution, and finally we can compute the derivative by convolution as well.

*Pooling layers.* Pooling layers makes the data more manageable. The most common pooling layer is *max pooling*, which outputs the maximum value for each patch. It can be seen as a non-linear convolutional filter, where it simply outputs the maximum value. Usually, pooling is done with a stride, such that the output becomes exponentially smaller; see Figure 8. Because of this, the *receptive field* becomes exponentially larger.

The forward pass is computed as follows,

$$z^{(\ell)}[i, j] = \max \left\{ z^{(\ell)}[i', j'] \mid i' \in [si : si + k], j' \in [sj : sj + k] \right\},$$

where  $s$  is the stride and  $k$  is the kernel size. Let  $[i^*, j^*]$  be the indices which corresponded to the maximum value in the forward pass, then we can compute the error propagation in the backward pass by

$$\frac{\partial z^{(\ell)}[i, j]}{\partial z^{(\ell-1)}[i', j']} = \mathbb{1}\{[i', j'] = [i^*, j^*]\}.$$

Note that the max pooling layer has no learnable parameters. Hence, the backward pass is only a propagation of the error, and not used for a weight update.

*Dense layer.* The dense layer is simply a linear layer that maps the final convolutional layer to the network's output. All previous convolutional and pooling layers can be seen as extracting features from the image, while the final dense layer makes the actual prediction.



Figure 8. Toy example of max pooling.

### 3 Fully convolutional neural network

*Semantic segmentation* is a computer vision task that involves assigning a semantic class to each pixel in an image. While in image classification, the model must output a single class for the entire image, semantic segmentation requires classifying a class for each pixel individually.

A naive approach would be to apply a single convolutional layer to an image, and then running a classifier on each individual pixel. However, this method is inefficient, because we have to run the classifier  $H \times W$  times. Instead, we use the output of convolutional neural networks. A naive approach of using CNNs would be to simply apply  $n$  convolutional layers with no downsampling, and then considering the last output as the predicted segmentation map. However, this method is expensive.

In practice, the most common approach is to downsample the features obtained using convolution and pooling layers and then upsample them again. By downsampling, this method is more computationally efficient, has larger receptive field, and suffers less from “The curse of dimensionality.” By upsampling, the model produces an output of the same resolution as the input.

#### 3.1 Upsampling methods

*Nearest neighbor.* Nearest neighbor upsampling copies the same value into all corresponding pixels at a higher resolution; see Figure 9.

*Bed of nails.* Bed of nails upsampling only copies each value once into the output in the top left value, and pads the rest with zero; see Figure 10.

*Max unpooling.* Max unpooling also uses zero padding, like bed of nails. However, it also remembers the original position of the maximum value before the corresponding max pooling in the downsampling phase. This information is then used to place each element back in the correct position; see Figures 8 and 11.

*Transposed convolutions.* Transposed convolution [Shelhamer et al., 2016] is a learned upsampling technique. This layer learns a kernel that is used to produce the terms whose sum will be the final output. Each term is obtained by multiplying all the element of the kernel by the same value of one single input pixel and then inserting the result in the correct position of a matrix of the same size as the output.

#### 3.2 U-net

The *U-net* [Ronneberger et al., 2015] is an FCNN architecture, whose main idea is to combine global and local feature maps by copying corresponding tensors from earlier stages in each upsampling stage; see Figure 12.



Figure 9. Nearest neighbor upsampling.



Figure 10. Bed of nails upsampling.



Figure 11. Max unpooling the output of Figure 8.

This allows the network to capture both local and global context. In each upsampling, the corresponding output from the downsampling phase is appended to the output. The copied tensor can be seen as the “global” information, while the input of the upsampling layer is the “local” information. Combining these allows for more fine-grained outputs.



**Figure 12.** U-net architecture. Down arrows are downsampling layers, up arrows are upsampling layers, and right arrows copy.

## 4 Recurrent neural networks

Recurrent neural networks (RNN) are a type of neural network that processes sequential data, such as text and video. Unlike traditional neural networks, which take fixed-length inputs, RNNs can take inputs of variable length.<sup>5</sup> Generally, RNNs have the following form,

$$\mathbf{h}^{(t)} = f_{\theta}(\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}).$$

In this way, an RNN behaves much like a dynamical system. If we unroll the timesteps, it becomes clear how  $\mathbf{h}^{(t)}$  depends on  $\mathbf{x}^{(1:t)}$ ,

$$\mathbf{h}^{(t)} = f_{\theta}(f_{\theta}(\dots(f_{\theta}(\mathbf{h}^{(0)}, \mathbf{x}^{(1)}), \dots), \mathbf{x}^{(t-1)}), \mathbf{x}^{(t)}).$$

Because of this, we can see  $\mathbf{h}^{(t)}$  as a representation of  $\mathbf{x}^{(1:t)}$ .

RNNs can have different applications, for example, *one to one*, where at each time step we have one input and one output,<sup>6</sup> *one to many*, where we have one input and we output a sequence of elements,<sup>7</sup> *many to one*, where we have a sequence of inputs and one output,<sup>8</sup> and *many to many*, where we map a sequence to another sequence of a different length.<sup>9</sup>

### 4.1 Elman RNN

The Elman RNN [Elman, 1990] is characterized by a hidden vector  $\mathbf{h}^{(t)}$ , which forms the state of the network at timestep  $t$ . The hidden state is updated at each timestep by combining the previous hidden state with the input,

$$f_{\theta}(\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}) = \tanh(\mathbf{W}_h \mathbf{h}^{(t-1)} + \mathbf{W}_x \mathbf{x}^{(t)}).$$

Since  $\mathbf{h}^{(t)}$  represents the subsequence  $\mathbf{x}^{(1:t)}$ , we can use it as input to an MLP,

$$\hat{\mathbf{y}} = \mathbf{W}_y \mathbf{h}^{(t)}.$$

Then, we can compute the loss function as the sum of each individual loss function,

$$\mathcal{L} \doteq \sum_{t=1}^T \ell^{(t)}.$$

We use *backpropagation through time* (BPTT) to compute the gradient of an RNN. This involves first unrolling the RNN; see Figure 13. Then we can compute the gradient by backpropagation on the resulting computational graph,

$$\frac{\partial \ell^{(t)}}{\partial \mathbf{W}} = \sum_{k=1}^t \frac{\partial \ell^{(t)}}{\partial \hat{\mathbf{y}}^{(t)}} \frac{\partial \hat{\mathbf{y}}^{(t)}}{\partial \mathbf{h}^{(t)}} \frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(k)}} \frac{\partial \mathbf{h}^{(k)}}{\partial \mathbf{W}},$$

where  $\frac{\partial \mathbf{h}^{(k)}}{\partial \mathbf{W}}$  is the immediate derivative that treats  $\mathbf{h}^{(k-1)}$  as constant w.r.t.  $\mathbf{W}$ .

<sup>5</sup> This is useful for data structures such as text, where the number of words in a text is not fixed.

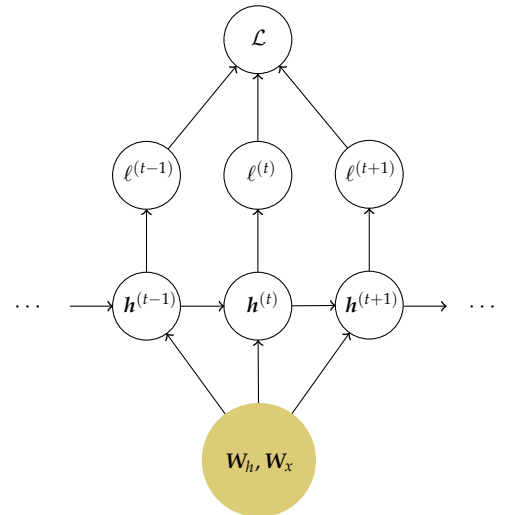
<sup>6</sup> E.g., part-of-speech tagging.

<sup>7</sup> E.g., image captioning, where the image is the one input, and the caption is a sequence of words.

<sup>8</sup> E.g., sentiment classification, where the input is a text and the output is a single output that determines how positive or negative the text is.

<sup>9</sup> E.g., machine translation, where we map a sentence in one language to a sentence of another.

We use the tanh activation function, because it is centered at 0.



**Figure 13.** The computational graph of an unrolled recurrent neural network. The inputs  $\mathbf{x}_{1:T}$  and outputs  $\mathbf{y}_{1:T}$  are omitted.

Let's only consider the following term of the product,

$$\begin{aligned}\frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(k)}} &= \prod_{i=k+1}^t \frac{\partial \mathbf{h}^{(i)}}{\partial \mathbf{h}^{(i-1)}} \\ &= \prod_{i=k+1}^t \frac{\partial}{\partial \mathbf{h}^{(i-1)}} \sigma(\mathbf{W}_h \mathbf{h}^{(i-1)} + \mathbf{W}_x \mathbf{x}^{(i)}) \\ &= \prod_{i=k+1}^t \mathbf{W}_h^\top \text{diag}\left(\sigma'(\mathbf{W}_h \mathbf{h}^{(i-1)} + \mathbf{W}_x \mathbf{x}^{(i)})\right).\end{aligned}$$

Assuming that the norm of the gradient of  $\sigma$  is upper bounded by some  $\gamma \in \mathbb{R}$ ,<sup>10</sup> i.e.,

$$\left\| \text{diag}\left(\sigma'(\mathbf{W}_h \mathbf{h}^{(t-1)} + \mathbf{W}_x \mathbf{x}^{(t)})\right) \right\| < \gamma.$$

<sup>10</sup> For example, the gradient of  $\tanh$  is bounded by 1.

Let  $\lambda_1$  be the largest eigenvalue of  $\mathbf{W}_h$ , then we have two cases,

1.  $\lambda_1 < \frac{1}{\gamma}$ . Then we have the following,

$$\left\| \frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(i-1)}} \right\| \leq \|\mathbf{W}_h\| \left\| \text{diag}\left(\sigma'(\mathbf{W}_h \mathbf{h}^{(t-1)} + \mathbf{W}_x \mathbf{x}^{(t)})\right) \right\| < \frac{1}{\gamma} \gamma = 1.$$

Triangle inequality and  $\|\mathbf{W}_h\| = \lambda_1$ .

Let  $\eta < 1$  be the upper bound of all gradients between  $\mathbf{h}^{(i)}$  and  $\mathbf{h}^{(i-1)}$ , then by induction, we have

$$\left\| \frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(k)}} \right\| = \left\| \prod_{i=k+1}^t \frac{\partial \mathbf{h}^{(i)}}{\partial \mathbf{h}^{(i-1)}} \right\| < \eta^{t-k}.$$

This converges to zero, as  $t \rightarrow \infty$ . Thus, we have a *vanishing gradient*;

2.  $\lambda_1 > \frac{1}{\gamma}$ . Using the same logic as in the other case, this yields

$$\left\| \frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(k)}} \right\| = \left\| \prod_{i=k+1}^t \frac{\partial \mathbf{h}^{(i)}}{\partial \mathbf{h}^{(i-1)}} \right\| > \eta^{t-k}.$$

for an upper bound  $\eta > 1$ . Thus, this diverges to  $\infty$  as  $t \rightarrow \infty$ . Thus, we have a *exploding gradient*.

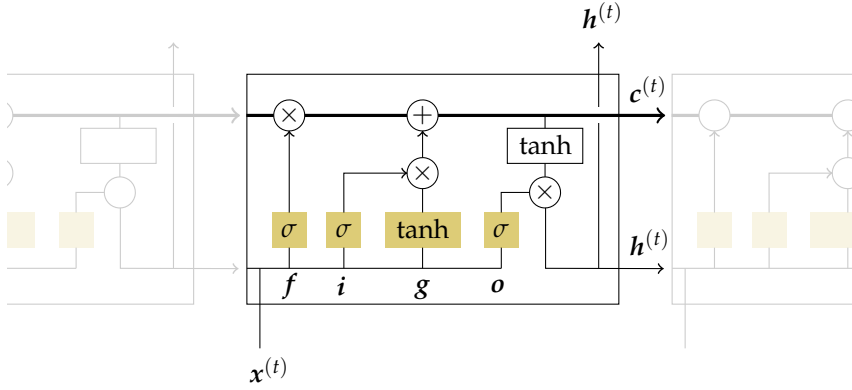
Thus, we will always have vanishing or exploding gradients when using an Elman RNN. This makes it hard for the architecture to capture long-term dependencies. A naive solution is to truncate BPTT to only go back  $m$  steps. However, then we still lose long-term signals, which we wanted to preserve.

#### 4.2 Long-short term memory

We need to make sure there is constant error flow. For this, we need a connection between timesteps that avoids exploding and vanishing gradients. A simple solution is the leaky unit,

$$\begin{aligned}\hat{\mathbf{h}}^{(t)} &= f_\theta(\mathbf{h}^{(t-1)}, \mathbf{x}^{(t)}) \\ \mathbf{h}^{(t)} &= \alpha \mathbf{h}^{(t-1)} + (1 - \alpha) \hat{\mathbf{h}}^t.\end{aligned}$$

The *long-short term memory* (LSTM) architecture [Hochreiter and Schmidhuber, 1997] takes this idea further by keeping a separate memory cell. Access to this cell is protected through gates to make sure that there is always a path between units, such that errors can propagate; see Figure 14. Furthermore, in contrast to the leaky unit, the LSTM Learns how to “remember” and “forget” information.



**Figure 14.** LSTM architecture. The yellow squares are neural networks, and the white squares are point-wise operators. As can be seen, there is an “information highway” that can easily propagate errors at the top, because of the minimal modifications made to it.

The cell of an LSTM consists of 4 gates. All gates get  $h^{(t-1)}$  and  $x^{(t)}$  as input. In particular, these gates have the following instructions,

- $f : \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, 1]^d$  is the *forget gate* and has the role of scaling the old cell state  $h^{(t-1)}$ . It “decides which information should be forgotten” from the previous cell state. It does so by outputting a vector in  $[0, 1]^d$  where 0 means forgetting completely and 1 means remembering everything;
- $g : \mathbb{R}^d \times \mathbb{R}^d \rightarrow [-1, 1]^d$  is the *gate* that decides what to write in the cell state  $h^{(t)}$ .
- $i : \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, 1]^d$  is the *input gate* and has the role of “deciding which values of the cell state  $c^{(t)}$  should be updated” at the current time step. Again, it does so by outputting a vector in  $[0, 1]^d$  to decide what of the output of the  $g$  gate should be written to the cell state  $c^{(t)}$ ;
- $o : \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, 1]^d$  is the *output gate* and has the role of “deciding which values of the current cell state  $c^{(t)}$  should be put in the output of the cell  $h^{(t)}$ ”.

We first compute all the gates,

$$\begin{aligned} f^{(t)} &= \sigma(W_{hf}h^{(t-1)} + W_{xf}x^{(t)}) \\ i^{(t)} &= \sigma(W_{hi}h^{(t-1)} + W_{xi}x^{(t)}) \\ o^{(t)} &= \sigma(W_{ho}h^{(t-1)} + W_{xo}x^{(t)}) \\ g^{(t)} &= \tanh(W_{hg}h^{(t-1)} + W_{xg}x^{(t)}). \end{aligned}$$

You can also chain multiple LSTM units one after another, which results in this computation being performed multiple times per layer. If this is the case, then we replace  $x^{(t)}$  by the hidden vector of the previous unit for all units after the first.



Then, we compute the outputs that are propagated to the next layer,

$$\begin{aligned} c^{(t)} &= f^{(t)} \odot c^{(t-1)} + i^{(t)} \odot g^{(t)} \\ h^{(t)} &= o^{(t)} \odot \tanh(c^{(t)}). \end{aligned}$$

$f^{(t)} \odot c^{(t-1)}$  computes what to keep/forget and  
 $i^{(t)} \odot g^{(t)}$  computes what to add.

The addition in the computation of  $c^{(t)}$  allows for gradients to directly propagate through  $c^{(t-1)} \odot f$ . Also, it allows the model to “select” what information should be retained. For example, at a high level in text, it might be helpful to store information such as gender and countries of origin. See [Olah, 2015] for more information about the gates.

### 4.3 Gradient clipping

While LSTMs are a great solution to the vanishing gradient problem, we still have the possibility of exploding gradients. This is what *gradient clipping* solves. The idea is to limit the maximum value of the gradient if it surpasses a predetermined threshold. In practice, the gradient descent step gets transformed into the following update rule,

$$\theta \leftarrow \begin{cases} \theta - \gamma g & \|g\| \leq k \\ \theta - \gamma \frac{k}{\|g\|} g & \text{otherwise,} \end{cases}$$

where  $g$  is the gradient,  $\gamma$  is the learning rate, and  $k$  is the gradient threshold.

## 5 *Generative models*

TODO

## 6 Autoencoders

*Autoencoders* [Kramer, 1991] are *generative models*. This means that their objective is to learn the underlying hidden structure of the data. They aim to model the distribution  $p_{\text{model}}(\mathbf{x})$  that resembles  $p_{\text{data}}(\mathbf{x})$  to generate new samples. Autoencoders are an *explicit* generative model, which means that they explicitly define the probability distribution  $p_{\text{model}}(\mathbf{x})$  and then sample from it to generate new data points.

In machine learning, we often have high-dimensional data  $\mathbf{x} \in \mathbb{R}^n$ , such as images, audio, or time-series. Hence, it is crucial to find a low-dimensional representation that can effectively compress the data while preserving its essential information.

Autoencoders offer a solution by making use of the *encoder-decoder structure*; see Figure 15. The *encoder*  $f$  projects the input space  $\mathcal{X}$  into a latent space  $\mathcal{Z}$ , while the *decoder*  $g$  maps the latent space  $\mathcal{Z}$  back to the input space  $\mathcal{X}$ . The assumption made by the autoencoder architecture is that if the decoder is capable of reconstructing the original input solely from the compressed representation, then this compressed representation must be meaningful. Consequently, the composition  $g \circ f$  aims to approximate the identity function on the data for a low reconstruction error.

Furthermore, to enable the generation of new samples from the latent space, the latent space must be well structured, characterized by *continuity* and *interpolation*. Continuity means that the entire space must be covered by the data points, while interpolation means that if we interpolate between two points, then the interpolation must also be a well behaved data point.

### 6.1 Linear autoencoders

If we restrict  $f$  and  $g$  to be linear, the encoder  $f$  becomes equivalent to the projection performed by *principal component analysis*. The advantage of such a reconstruction is that it can be found in a closed form, and it is interpretable. However, it is not very powerful.

### 6.2 Non-linear autoencoders

We can gain a lot of performance by allowing  $f$  and  $g$  to be non-linear. In this case, the encoder and decoder are implemented as neural networks. To train these networks, we optimize for the reconstruction error,

$$\phi^*, \psi^* \in \operatorname{argmin}_{\phi, \psi} \sum_{n=1}^N \|\mathbf{x}_n - g_{\psi}(f_{\phi}(\mathbf{x}_n))\|^2$$

We can distinguish between *undercomplete* and *overcomplete* latent spaces. A latent space is undercomplete if  $\dim(\mathcal{Z}) < \dim(\mathcal{X})$ , while it is over-



Figure 15. Autoencoder architecture.

complete if  $\dim(\mathcal{Z}) > \dim(\mathcal{X})$ . The idea of an undercomplete hidden representation is to enable the network to learn the important features of the data by reducing the dimensionality of the hidden space. This prevents the autoencoder from simply copying the input and forces it to extract meaningful and discriminative features. Next, overcomplete latent spaces are useful for denoising and inpainting autoencoders, where we have an imperfect input and want a perfect output. The overcompleteness allows the model to extract more features from the transformed input, leading to improved performance.

### 6.3 Variational autoencoders

While autoencoders are good at reconstruction, they struggle at generating new high quality samples. This is due to the latent space not being “well-structured”, meaning that there is no continuity or interpolation. There are large regions in the latent space where there are no observations, thus the model does not know what to output when it get an input from those regions.

*Variational autoencoders* (VAE) [Kingma and Welling, 2013] are designed to have a continuous latent space. It achieves this by making the encoder output a probability distribution over latent vectors, rather than a single latent vector. Generally, it outputs a mean vector  $\mu$  and standard deviation vector  $\sigma$  to form a Gaussian distribution over latents  $\mathcal{N}(\mu, \text{diag}(\sigma))$ . The idea is that even for the same input, the latent vector can be different, but in the same area. This means that data points cover areas in the latent space, rather than single points, ensuring continuity and interpolation.

However, since there are no limits on the values taken by  $\mu$  and  $\sigma$ , the encoder may learn to generate very different  $\mu$  for each class while minimizing  $\sigma$ . This would mean that the encoder essentially outputs points again to decrease the reconstruction error. We can avoid this by minimizing the KL-divergence between the output distribution and a standard normal distribution.<sup>11</sup> Intuitively, this encourages the encoder to distribute the encodings evenly around the center of the latent space.

To train the model, we want to maximize the likelihood of the training data,

$$p(x) = \int p_{\psi}(x | z)p(z)dz.$$

However, this is intractable, because we cannot compute it for all  $z \in \mathcal{Z}$ . Thus, we define an approximation of the posterior,  $q_{\phi}(z | x)$ , which is computed by the encoder. We can now derive the *evidence lower bound*

<sup>11</sup> The KL-divergence is defined as

$$D_{\text{KL}}(p||q) \doteq \mathbb{E} \left[ \log \left( \frac{p(x)}{q(x)} \right) \right].$$

It is not symmetric and non-negative.

$p_{\psi}(x | z)$  is induced by the decoder.

(ELBO),

$$\begin{aligned}
\log p(x) &= \mathbb{E}_{z \sim q_\phi(\cdot|x)} [\log p(x)] && x \text{ does not depend on } z. \\
&= \mathbb{E}_{z \sim q_\phi(\cdot|x)} \left[ \log \frac{p_\psi(x|z)p(z)}{p(z|x)} \right] && \text{Bayes' rule.} \\
&= \mathbb{E}_{z \sim q_\phi(\cdot|x)} \left[ \log \left( \frac{p_\psi(x|z)p(z)}{p(z|x)} \frac{q_\phi(z|x)}{q_\phi(z|x)} \right) \right] && q(z|x)/q(z|x) = 1. \\
&= \mathbb{E}_{z|x} [\log p_\psi(x|z)] - \mathbb{E}_{z|x} \left[ \log \frac{q_\phi(z|x)}{p(z)} \right] + \mathbb{E}_{z|x} \left[ \log \frac{q_\phi(z|x)}{p(z|x)} \right] \\
&= \mathbb{E}_{z|x} [\log p_\psi(x|z)] - D_{\text{KL}}(q_\phi(z|x) \| p(z)) + D_{\text{KL}}(q_\phi(z|x) \| p(z|x)) \\
&\geq \mathbb{E}_{z|x} [\log p_\psi(x|z)] - D_{\text{KL}}(q_\phi(z|x) \| p(z)). && \text{KL-divergence is non-negative.}
\end{aligned}$$

The first term of the ELBO encourages low reconstruction error, while the second term makes sure that the approximate posterior  $q_\phi$  does not deviate too far from the prior  $p$ . The second term can be computed in a closed-form, since both arguments are Gaussian,

$$\begin{aligned}
q_\phi(z|x) &= \mathcal{N}(z; \mu_\phi(x), \text{diag}(\sigma_\phi(x))) \\
p(z) &= \mathcal{N}(z; \mathbf{0}, I).
\end{aligned}$$

A minor problem is that, during training, we cannot compute the derivative of expectations w.r.t. the parameters that we wish to optimize. Thus, we must use the *reparametrization trick*, which involves treating the random sampling as a single noise term. In particular, instead of sampling  $z \sim \mathcal{N}(\mu, \text{diag}(\sigma))$ , we sample  $\epsilon \sim \mathcal{N}(\mathbf{0}, I)$  and compute  $z = \mu + \sigma \odot \epsilon$ . Using this trick, we can remove the mean and variance from the sampling operation, meaning that we can differentiate w.r.t. the model parameters.

After training, we can sample a latent vector  $z \sim \mathcal{N}(\mathbf{0}, I)$  and pass it to the decoder, which will give a good output, because the latent space is well-structured.

#### 6.4 $\beta$ -VAE

VAEs still have problems with their latent space; the representations are still *entangled*.<sup>12</sup> This means that we do not have an explicit way of controlling the output. For example, in the MNIST dataset [Deng, 2012], we have no way of explicitly sampling a specific number. The  $\beta$ -VAE [Higgins et al., 2017] solves this problem by giving more weight to the KL term with an adjustable hyperparameter  $\beta$  that balances latent channel capacity and independence constraints with reconstruction accuracy. The intuition behind this is that if factors are in practice independent from each other, the model should benefit from disentangling them.

<sup>12</sup> The latent space is disentangled if every dimension changes a single feature of the output.

In practice, we want to force the KL loss to be under a threshold  $\delta > 0$ ,

$$\begin{aligned} & \underset{\phi, \psi}{\text{maximize}} && \mathbb{E}_{x \sim \mathcal{D}} \left[ \mathbb{E}_{z \sim q_{\phi}(\cdot | x)} [\log p_{\psi}(x | z)] \right] \\ & \text{subject to} && D_{\text{KL}}(q_{\phi}(z | x) \| p(z)) < \delta. \end{aligned}$$

Rewriting this as a Lagrangian, using the Karush-Kuhn-Tucker conditions, we get

$$\begin{aligned} \mathcal{L}(\phi, \psi, \beta) &= \mathbb{E}_{z \sim q_{\phi}(\cdot | x)} [\log p_{\psi}(x | z)] - \beta (D_{\text{KL}}(q_{\phi} \| p(z)) - \delta) \\ &= \mathbb{E}_{z \sim q_{\phi}(\cdot | x)} [\log p_{\psi}(x | z)] - \beta D_{\text{KL}}(q_{\phi} \| p(z)) + \beta \delta \\ &\geq \mathbb{E}_{z \sim q_{\phi}(\cdot | x)} [\log p_{\psi}(x | z)] - \beta D_{\text{KL}}(q_{\phi} \| p(z)). \end{aligned}$$

Thus, this becomes our new objective function that we wish to maximize.

## 7 Autoregressive models

We saw that VAEs are approximate models, since they cannot exactly compute the likelihood  $p(\mathbf{x})$  to maximize it. *Autoregressive models* solve this by computing the likelihood with the chain rule,

$$p(\mathbf{x}) = \prod_{i=1}^n p(x_i \mid \mathbf{x}_{1:i-1}).$$

Autoregressive models use data from the same input variable at previous timesteps. This is where they get their name from; they perform regression of themselves. In particular, autoregressive models generate one element of the sequence at a time, conditioning on previously generated elements. Thus, it takes  $\mathbf{x}_{1:k}$  as input and outputs  $x_{k+1}$ . Because of this, they can build a probability distribution over possible sequences, using the chain rule as above.

The hard part of this approach is that we must parametrize all possible conditional distributions  $p(x_{k+1} \mid \mathbf{x}_{1:k})$ . Suppose we have a binary image consisting of  $n$  pixels. Then, we need

$$\sum_{i=1}^n 2^{i-1} \in \mathcal{O}(2^n)$$

parameters to parametrize this model. Thus, we need to make additional assumptions to find more compact representations of the distribution.<sup>13</sup> We will explore the idea of learning a function  $f_i : \{0,1\}^{i-1} \rightarrow [0,1]$ , parametrized by  $\theta_i$ , which takes as input the previous pixels and outputs the parameter for the Bernoulli distribution. The total number of parameters is

$$\sum_{i=1}^n |\theta_i|.$$

Furthermore, we need to think about in which order we generate the pixels of the image. Intuitively, it would make sense to do left-right top-down or top-down left-right. However, in practice, randomly ordering the pixels works just as well. One just needs to make sure that the ordering remains the same for all data points.

### 7.1 Fully visible sigmoid belief network

In a *fully visible sigmoid belief network* (FVSBN) [Frey, 1998], each timestep has its own function  $f_i$  that is modeled by logistic regression,

$$f_i(\mathbf{x}_{1:i-1}) = \sigma(\alpha_0^{(i)} + \alpha_1^{(i)}x_1 + \dots + \alpha_{i-1}^{(i)}x_{i-1}).$$

At the  $i$ -th timestep, we have  $i$  parameters denoted by  $\theta_i = [\alpha_0, \dots, \alpha_{i-1}]$ . Thus, the total number of parameters is

$$\sum_{i=1}^n |\theta_i| = \sum_{i=1}^n i = \frac{n^2 + n}{2} \in \mathcal{O}(n^2),$$

<sup>13</sup> A naive solution would be to assume that all points are independent, which would result in

$$p(\mathbf{x}) = \prod_{i=1}^n p(x_i).$$

Then, we only require  $n$  parameters. However, in practice, this would result in a random sampling of pixels, making the generations incoherent.

which is much better than the exponential number of parameters we had before.

### 7.2 Neural autoregressive density estimator

The problem with FVSBN is that they are likely not expressive enough for any meaningful tasks, since they only consist of a single linear layer for each timestep. The *neural autoregressive density estimator* (NADE) [Uria et al., 2016] offers an alternative parametrization based on MLPs, where we have hidden layers, increasing expressivity, and the weights are shared between timesteps, decreasing the number of parameters. Specifically, the hidden layer activations can be computed by the following,

$$\begin{aligned} \mathbf{h}_i &= \sigma(\mathbf{W}_{:,1:i-1} \mathbf{x}_{1:i-1} + \mathbf{b}) \\ \hat{x}_i &= \sigma(\mathbf{V}_i \mathbf{h}_i + c_i). \end{aligned}$$

The advantage of shared parameters is that the total number of parameters gets reduced from  $\mathcal{O}(n^2d)$  to  $\mathcal{O}(nd)$ , and the hidden unit activations can be evaluated in  $\mathcal{O}(nd)$  by using an alternative recursive definition,

$$\begin{aligned} \mathbf{a}_0 &= \mathbf{0} \\ \mathbf{a}_{i+1} &= \mathbf{a}_i + \mathbf{W}_{:,i} x_i \\ \mathbf{h}_i &= \sigma(\mathbf{a}_i + \mathbf{b}). \end{aligned}$$

Since NADE is a model for binary data,  $\hat{x} \in [0, 1]$  is the probability  $p(x_i \mid \mathbf{x}_{1:i-1})$  at each timestep  $i$ . NADE is trained by maximizing the average log-likelihood,

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \log p(\mathbf{x}^{(t)}) &= \frac{1}{T} \sum_{t=1}^T \log \prod_{i=1}^n p(x_i^{(t)} \mid \mathbf{x}_{1:i-1}^{(t)}) \\ &= \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^n \log p(x_i^{(t)} \mid \mathbf{x}_{1:i-1}^{(t)}). \end{aligned}$$

Thus, we optimize the log-likelihood exactly, which was not possible for VAEs.

During training, a *teacher forcing* approach is used, where the ground truth values of pixels are used for conditioning, rather than the predicted ones. This leads to more stable training. However, at inference time we use the predicted value, since we do not have access to the ground truth.

### 7.3 Masked autoencoder distribution estimation

The idea behind *masked autoencoder distribution estimation* (MADE) [Germain et al., 2015] is to construct an autoencoder which fulfills the autoregressive property, such that its outputs can be used as conditionals. In order to achieve this, no computational path between output unit  $\hat{x}_{k+1}$  and any of the input units  $x_{k+1}, \dots, x_n$  may exist. This has the result that  $\hat{x}_{k+1}$  is only conditioned on  $\mathbf{x}_{1:k}$ , relative to the ordering.

The way to do this is to first assign a number from 1 to  $n$  to each input unit  $x_i$ , which we call the ordering. Then, for each hidden unit, uniformly



sample an integer  $m$ , between 1 and  $n - 1$ . In the hidden layers, we then only allow units in layer  $\ell$  to propagate to units in layer  $\ell + 1$  with higher or equal value. Finally, we allow connections between the last hidden layer and the output only to units with value that is strictly greater. Let  $m^{(\ell)}(k)$  be the value assigned to the  $k$ -th element of layer  $\ell$ , then these constraints can be encoded in mask matrices,

$$\begin{aligned} \mathbf{M}_{ij}^{W^{(\ell)}} &= \mathbb{1}\{m^{(\ell-1)}(j) \leq m^{(\ell)}(i)\} \\ \mathbf{M}_{ij}^V &= \mathbb{1}\{m^{(\ell-1)}(j) < m^{(\ell)}(i)\}. \end{aligned}$$

Then, we alter the weight matrices by

$$\begin{aligned} \bar{\mathbf{W}}^{(\ell)} &= \mathbf{W}^{(\ell)} \odot \mathbf{M}^{W^{(\ell)}} \\ \bar{\mathbf{V}} &= \mathbf{V} \odot \mathbf{M}^V, \end{aligned}$$

and use those instead.

However, a problem with this approach is that it requires very large hidden layers to retain expressivity. And, while it is possible to compute  $p(\mathbf{x})$  in a single pass, sampling still requires  $n$  passes.

#### 7.4 Generating images

*Pixel-RNN.* The idea behind *Pixel-RNN* [Van Den Oord et al., 2016] is to generate image pixels starting from the corner and modeling the dependency on previous pixels using an RNN. In particular, a pixel value is dependent on its top-left neighboring pixels and the RNNs hidden state; see Figure 17. However, the problem with this approach is that the generation of new pixels depends on the hidden state, making the generation process sequential, and thus slow.

*Pixel-CNN.* We can solve the problem of *Pixel-RNN* by assuming that pixel values only depend on a context region around the pixel. This is exactly what the *Pixel-CNN* does [Van Den Oord et al., 2016]. This allows for parallelization during training, because the context region values are known during training.<sup>14</sup> Just as *Pixel-RNN*, it starts from the top-left corner, but it models the dependencies with a CNN. During training, we need to make sure that only the previously generated pixels are used for prediction of the next, thus we need a mask that masks out all unknown pixel values. However, stacking layers of masked convolutions creates a blind spot in the convolution; see Figure 18. The solution to this is to combine horizontal and vertical stacks of convolutions [Van den Oord et al., 2016]. The former conditions on the row so far, while the latter conditions on all rows above. The final output is obtained by summing the two outputs.

To enforce the autoregressive property, the model also needs to go over the color channels in an autoregressive manner. Thus the conditional

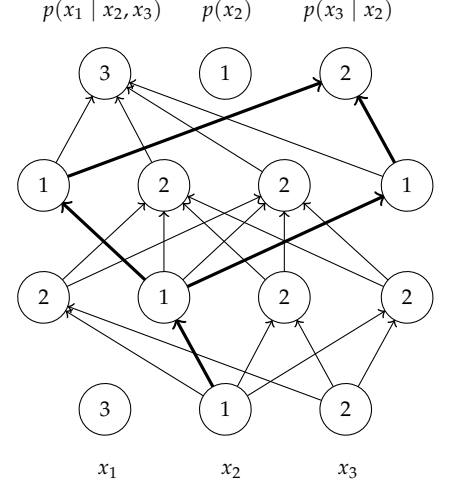


Figure 16. MADE masking with  $n = 3$ .

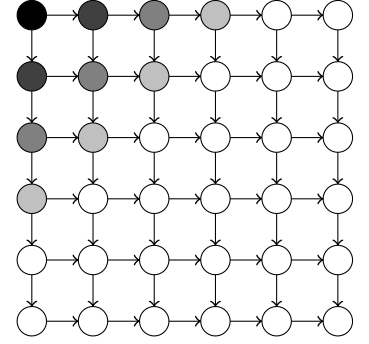


Figure 17. Pixel-RNN generation process.

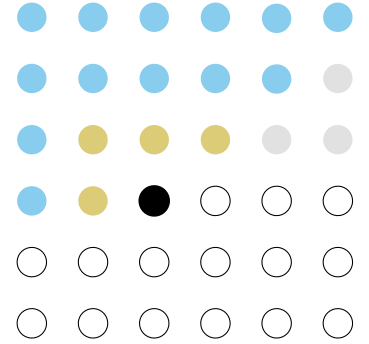


Figure 18. Pixel-CNN generation process. The black pixel depends explicitly on the yellow pixels, where the thick-lined pixels denote the masked convolutional layer. The black pixel should also depend on the gray pixels, but it does not due to the way the stacked masked convolutions work; a blind spot.

<sup>14</sup> However, we still have to sequentially predict every token during inference, but this is an issue with all autoregressive models.

probability is expressed as

$$p(\mathbf{x}_i | \mathbf{x}_{1:i-1}) = p(x_{i,R} | \mathbf{x}_{1:i-1})p(x_{i,G} | \mathbf{x}_{1:i-1}, x_{i,R})p(x_{i,B} | \mathbf{x}_{1:i-1}, x_{i,R}, x_{i,G}).$$

### 7.5 Generating audio

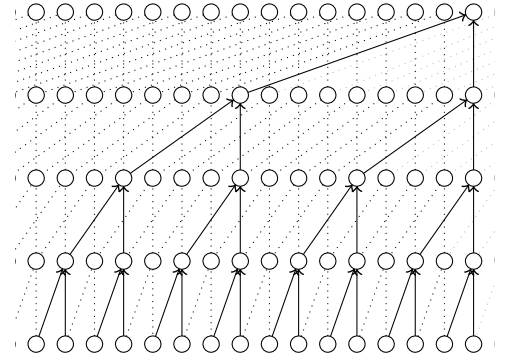
*WaveNet* [van den Oord et al., 2016] adapts the Pixel-CNN framework for audio data. However, audio typically has much longer time horizons, since they consist of 16000 samples per second. To be able to capture these long-term dependencies efficiently, WaveNet incorporates *dilated convolutions* [Yu and Koltun, 2016]. This allows for an exponential increase in the receptive field. In dilated convolutions, the filter is applied with gaps between the filter elements. WaveNet increases the dilation factor as we go up in the layers to attain an exponential receptive field; see Figure 19.

### 7.6 Variational RNN

RNNs can also be used to generate sequences by sampling  $\mathbf{h}^{(0)}$  and then sequentially predicting the next element, and updating the hidden state. However, the generation of new sequences is very slow, because of the sequential nature of the generation process. Another limitation of vanilla RNNs is that their structure is entirely deterministic and thus limited in expressive power.

The *variational VRNN* [Chung et al., 2015] (VRNN) is a recurrent version of the VAE. A different VAE is used at each timestep. However, the VAEs are not the same as we have seen. In particular, the prior distribution is no longer a standard Gaussian, but follows the distribution  $\mathcal{N}(\boldsymbol{\mu}_0^{(t)}, \text{diag}(\boldsymbol{\sigma}_0^{(t)}))$ , where  $(\boldsymbol{\mu}_0^{(t)}, \boldsymbol{\sigma}_0^{(t)})$  is predicted by the network  $\phi_{\text{prior}}(\mathbf{h}^{(t)})$  based on the hidden state of the RNN at each timestep. This introduces randomness to the generation, whereas RNNs are typically deterministic.

The *conditional VRNN* (C-VRNN) is an extension to the VRNN, where we can condition on *e.g.* style or, in the case of MNIST, which digit should be generated. During training, the model predicts the conditional variable, called the posterior, and hidden state from the input  $\mathbf{x}^{(t)}$ . From the hidden state it also predicts the conditional variable, called the prior. It then minimizes the reconstruction error from the style variable, and the KL divergence between prior and posterior. During inference, it only uses the priors for next token prediction, predicted from the hidden state, since it does not have access to  $\mathbf{x}^{(t)}$ . Then, given the predicted token, it predicts the posterior conditional variables, which are used to predict the next hidden state  $\mathbf{h}^{(t+1)}$ .



**Figure 19.** Stacked dilated convolutional layers in WaveNet.

## 7.7 Transformers

Transformers [Vaswani et al., 2017] are used in nearly all the state-of-the-art models at the moment. At the basis of transformers lies the self-attention mechanism, where we extract key, value, and query representations from the input. These are then used in a soft-lookup mechanism, where the query and key values decide how much attention is paid to the values.<sup>15</sup>

<sup>15</sup> Note the parallel with dictionaries in programming languages, which use a hard-lookup.

It computes the keys, values, and queries by

$$\mathbf{K} = \mathbf{X}\mathbf{W}_K$$

$$\mathbf{V} = \mathbf{X}\mathbf{W}_V$$

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q,$$

where  $\mathbf{X} \in \mathbb{R}^{n \times d}$  and  $\mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_Q \in \mathbb{R}^{d \times d}$ . Now, we have key, value, and query representations of the input that are all in  $\mathbb{R}^{n \times d}$ . Then, we can compute the output by

$$\mathbf{Y} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}.$$

Intuitively, the softmax computes how much “attention” should be given to each of the previous timesteps by outputting a probability distribution over timesteps.

Since  $\mathbf{Y} \in \mathbb{R}^{n \times d}$  is of the same dimensionality as  $\mathbf{X}$ , we can stack self-attention layers. This is the basis of the transformer architecture. However, this does not respect the autoregressive property, thus we need to use a mask  $\mathbf{M}$  to prevent the model from accessing future timesteps, where

$$\mathbf{M} = \begin{bmatrix} -\infty & -\infty & \cdots & -\infty \\ 0 & -\infty & \cdots & -\infty \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\infty \end{bmatrix}.$$

Then, the attention computation becomes

$$\mathbf{Y} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{M}\right)\mathbf{V}.$$

The computational complexity of self-attention is  $\mathcal{O}(n^2d)$ , which is quite high for large inputs, such as audio and images. But, we gain a maximum path length of  $\mathcal{O}(1)$  to any of the previous timesteps, making sure that no information gets “forgotten”, and it allows for easy error propagation during training, in contrast to RNNs.

## 8 Normalizing flows

As we have seen, VAEs learn meaningful representation through latent variables, but they suffer from intractable marginal likelihoods. On the other hand, autoregressive models have a tractable likelihood, but lack a latent space and direct feature learning mechanism. *Normalizing flows* [Rezende and Mohamed, 2015] try to have the best of both worlds; meaningful latent space and a tractable likelihood. They achieve this by leveraging the change of variable technique of integration.

### 8.1 Change of variables

In the one-dimensional, integration by substitution works as follows. Let  $g : [a, b] \rightarrow I$  be a differentiable function with a continuous derivative and  $f : I \rightarrow \mathbb{R}$  be a continuous function, where  $I \subset \mathbb{R}$  is an interval. Then,

$$\int_{g(a)}^{g(b)} f(x) dx = \int_a^b f(g(u)) g'(u) du.$$

Similarly, we can make the same change of variables transformation to probability distributions. Let  $z \sim p_Z$ ,  $x = f(z)$ , where  $f(\cdot)$  is a monotonic and differentiable function with an inverse  $z = f^{-1}(x)$ . Then, the probability density function of  $x$  is

$$p_X(x) = p_Z(f^{-1}(x)) \left| \frac{df^{-1}(x)}{dx} \right| = p_Z(z) \left| \frac{df^{-1}(x)}{dx} \right|.$$

We can generalize this to any dimensionality by the following,

$$p_X(x) = p_Z(f^{-1}(x)) \left| \det(J_x f^{-1}(x)) \right| = p_Z(z) |\det(J_z f(z))|^{-1},$$

where the last equality is a property of Jacobians of invertible functions. We can view the absolute determinant as computing the volume of change of  $f$ . Normalizing flow models parametrize  $f_\theta$ .

Thus, this gives us a way to compute the probability of  $x$  in an exact and tractable manner. The downside of this approach is that  $f$  must be invertible, thus we must carefully parametrize the model, which means that we cannot use any model we might want to use. Furthermore, this has the consequence that it must preserve dimensionality, thus the latent space must be as large as the data space.

From a computational perspective, we require the Jacobian of the transformation to be computed efficiently. In general, computing the Jacobian takes  $\mathcal{O}(d^3)$  to compute for a  $d \times d$  matrix. However, this is not fast enough. A way to achieve linear complexity is to design  $f$  such that its Jacobian is a triangular matrix, which takes  $\mathcal{O}(d)$  to compute. This requirement further reduces the number of modeling decisions we can make.

We need to normalize by the determinant of  $f$ , because the total probability must be 1. Recall that the determinant quantifies the volume change of an operation.

### 8.2 Coupling layers

A *coupling layer* [Dinh et al., 2014] is a type of network layer that effectively meets the above requirements of a normalizing flow function. It is invertible and offers efficient computation of the determinant. It consists of two functions;  $\beta$  and  $h$ .  $\beta$  can be any neural network and does not necessarily need to be invertible.<sup>16</sup>  $h$  is an element-wise operation that is invertible w.r.t. its first argument, given the second.  $h(x_A, \beta(x_B))$  produces the first half of the input and  $x_B$  produces the second half.

This gives the following function,

$$f : \begin{bmatrix} x_A \\ x_B \end{bmatrix} \mapsto \begin{bmatrix} h(x_A, \beta(x_B)) \\ x_B \end{bmatrix}.$$

The inverse of this function is given by

$$f^{-1} : \begin{bmatrix} y_A \\ y_B \end{bmatrix} \mapsto \begin{bmatrix} h^{-1}(y_A, \beta(y_B)) \\ y_B \end{bmatrix}.$$

The Jacobian matrix can be efficiently computed by

$$\begin{aligned} Jf(x) &= \begin{bmatrix} \frac{\partial y_A}{\partial x_A} & \frac{\partial y_A}{\partial x_B} \\ \frac{\partial y_B}{\partial x_A} & \frac{\partial y_B}{\partial x_B} \end{bmatrix} \\ &= \begin{bmatrix} h'(x_A, \beta(x_B)) \frac{\partial x_A}{\partial x_A} & h'(x_A, \beta(x_B)) \frac{\partial \beta(x_B)}{\partial x_B} \\ \frac{\partial x_B}{\partial x_A} & \frac{\partial x_B}{\partial x_B} \end{bmatrix} \\ &= \begin{bmatrix} h'(x_A, \beta(x_B)) & h'(x_A, \beta(x_B)) \beta'(x_B) \\ \mathbf{0} & \mathbf{I} \end{bmatrix}. \end{aligned}$$

When implementing this, we notice that this layer leaves part of its input unchanged. The role of the two subsets in the partition thus gets exchanged in alternating layers, so that both subsets get updates. In practice, we often randomly choose the splits to ensure proper mixing.

### 8.3 Composing transformations

Often, a single non-linear transformation is insufficient to capture complex patterns. Especially, because a single layer leaves part of the input unchanged. Thus, to achieve more complex transformations, we can compose multiple transformations together. We then get the following function

$$x = f(z) = (f_m \circ \dots \circ f_1)(z).$$

Again using the change of variables, we can then compute the likelihood by

$$p_X(x) = p_Z(f^{-1}(x)) \prod_{k=1}^m |\det(Jf_k(x))|^{-1}.$$



**Figure 20.** Diagram of a coupling layer.  $h$  is an invertible element-wise operation and  $\beta$  is can be arbitrarily complex and does not need to be invertible.

<sup>16</sup> This is very important, because requiring a neural network to be invertible would significantly reduce the number of available modeling decisions.

$$\det(AB) = \det(A)\det(B).$$

### 8.4 Training and inference

At training time, we can learn the model by maximizing the exact log likelihood over the dataset,

$$\log p_X(\mathcal{D}) = \sum_{i=1}^n \log p_Z(f^{-1}(x_i)) + \sum_{k=1}^m \log |\det(Jf_k(x_i))|^{-1}.$$

At inference time, we generate a sample  $x$  by drawing a random  $z \sim p_Z$  and transform it via  $f$ , obtaining  $x = f(z)$ . To evaluate the probability of an observation, we use the inverse transform to get its latent variable  $z = f^{-1}(x)$ , and compute its probability at  $p_Z(z)$ . Generally,  $p_Z$  is chosen to be a simple distribution, such as  $\mathcal{N}(\mathbf{0}, \mathbf{I})$ .

### 8.5 Architectures

The main difference between flow architectures is the choice of  $h$  and the way it splits the input into  $x_A$  and  $x_B$ .

**NICE.** NICE [Dinh et al., 2014] splits the data by dividing the input into two parts  $x_A = x_{1:d/2}$  and  $x_B = x_{d/2+1:d}$ , swapping the order randomly. In the coupling layer, it uses an additive coupling law and the output is computed as

$$\begin{bmatrix} y_A \\ y_B \end{bmatrix} = \begin{bmatrix} x_A + \beta(x_B) \\ x_B \end{bmatrix}.$$

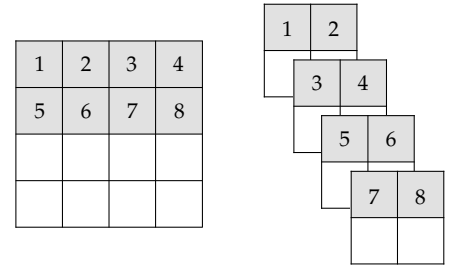
**RealNVP.** RealNVP [Dinh et al., 2016] employs a multi-scale architecture that uses two types of partitioning: checkerboard and channel-wise masking; see Figure 21. At each scale, the model alternates between checkerboard patterns. Then, it does a squeezing operation to go from  $C \times H \times W$  dimensionality to  $4C \times H/2 \times W/2$ . Lastly, it uses channel-wise masking. This sequence of steps ensures that all data can interact with each other. Instead of a simple additive  $h$  as in NICE, RealNVP implements it as an affine mapping,

$$\begin{bmatrix} y_A \\ y_B \end{bmatrix} = \begin{bmatrix} x_A \odot \exp(s(x_B)) + t(x_B) \\ x_B \end{bmatrix}.$$

Here,  $s$  and  $t$  can be arbitrarily complex.

**GLOW.** GLOW [Kingma and Dhariwal, 2018] is an architecture that utilizes invertible  $1 \times 1$  convolutions, affine coupling layers, and multi-scale architecture. It consists of  $L$  levels, each of which is composed of  $K$  steps of flow. The  $L$  levels allow for effective processing of all parts of the input, while the  $K$  steps are used to increase the flexibility of the transformation  $f$ .

A step of flow consists of applying activation norm, an invertible  $1 \times 1$  convolution, and a coupling layer, in order. The activation norm



**Figure 21.** Masking used by RealNVP [Dinh et al., 2016].

is similar to batch norm, but it normalizes each input channel. As we saw, a permutation of the input is needed in order to be able to process the entire input. The  $1 \times 1$  convolution with  $C$  filters is a generalization of a permutation in the channel dimension. This allows us to learn the required permutation. The coupling layer is implemented as in RealNVP, while the split is performed along the channel dimension only, because the convolution is  $1 \times 1$ .

## 9 Generative adversarial networks

So far, we have only seen generative models that optimize the likelihood. This has a nice interpretation and leads to nice theory, but there are cases where optimizing the likelihood will not give good results [Theis et al., 2016].

1. In the first case, we might have a good likelihood score, but poor samples. Let's say we have a model  $p$  that generates high-quality samples and a model  $q$  that generates noise. Now construct the mixture model  $0.01p + 0.99q$ .<sup>17</sup> The log-likelihood of this model is then

$$\log(0.01p(x) + 0.99q(x)) \geq \log(0.01p(x)) = \log(p(x)) - \log 100.$$

<sup>17</sup> This means that 99% of the time, the generated samples will be noise.

The term  $\log p(x)$  will be proportional to the dimensionality  $d$ , while  $\log 100$  remains constant. For high-dimensional data, this results in a high log-likelihood for the mixture model;

2. In the second case, we might have a low likelihood score with high-quality samples. This occurs when the model overfits on the training data, meaning that it only outputs data points from the training dataset. This will result in low log-likelihood on the validation dataset, despite the samples being high-quality.

*Generate adversarial networks* (GAN) [Goodfellow et al., 2014] aim to solve this by introducing a *discriminator* ( $D$ ), whose job it is to differentiate between real and fake images. The objective of the *generator* ( $G$ ) is then to maximize the discriminator's classification loss by generating images similar to the dataset. By doing so, it implicitly induces a distribution over data points  $p_{\text{model}}$ .

Specifically, the generator  $G : \mathbb{R}^d \rightarrow \mathbb{R}^n$  maps a simple  $d$ -dimensional distribution to a sample from the data distribution. The discriminator  $D : \mathbb{R}^n \rightarrow [0, 1]$  assigns a probability to samples. Its objective is to assign probability 1 to samples from the dataset and probability 0 to samples generated by the generator.

This leads us to the following value function,

$$V(D, G) = \log D(x) + \log(1 - D(G(z))), \quad x \in \mathcal{D}, z \sim \mathcal{N}(0, I).$$

The discriminator aims to maximize it, while the generator aims to minimize it, which gives the following optimization problem,

$$G^*, D^* \in \underset{G}{\operatorname{argmin}} \underset{D}{\operatorname{argmax}} V(D, G).$$

### 9.1 Theoretical analysis

*Optimal discriminator.* Given access to  $p_{\text{model}}$  and  $p_{\text{data}}$ , we can compute a closed form solution for  $D^*$  by the following theorem.<sup>18</sup>

<sup>18</sup> In practice, we do not have access to  $p_{\text{model}}$ , since we parametrize it implicitly.



**Theorem 2.** For any generator  $G$  that induces  $p_{\text{model}}$ , the optimal discriminator is

$$D^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_{\text{model}}(x)}.$$

*Proof.* Let  $G$  be a generator, then  $D$  is computed by

$$\begin{aligned} D^* &= \operatorname{argmax}_D V(G, D) \\ &= \operatorname{argmax}_D \mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim p_{\text{prior}}} [\log(1 - D(G(z)))] \\ &= \operatorname{argmax}_D \int p_{\text{data}}(x) \log D(x) dx + \int_z p_{\text{prior}}(z) \log(1 - D(G(z))) dz \\ &= \operatorname{argmax}_D \int p_{\text{data}}(x) \log D(x) dx + \int_x p_{\text{model}}(x) \log(1 - D(x)) dx && \text{Law of unconscious statistician.} \\ &= \operatorname{argmax}_D \int p_{\text{data}}(x) \log D(x) + p_{\text{model}}(x) \log(1 - D(x)) dx. \end{aligned}$$

Let  $a, b > 0$  and consider  $f(y) = a \log(y) + b \log(1 - y)$ , then  $f$ 's maximum is achieved at  $y = \frac{a}{a+b}$ , which we can easily prove by

$$\begin{aligned} f'(y) &= \frac{a}{y} - \frac{b}{1-y} = 0 \iff y = \frac{a}{a+b} && \text{Critical point.} \\ f''(y) &= -\frac{a}{y^2} - \frac{b}{(1-y)^2} < 0, \quad \forall a, b > 0. && \text{Maximum point.} \end{aligned}$$

Thus, we obtain

$$D^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_{\text{model}}(x)}, \quad \forall x \in \mathcal{X}.$$

■

*Global optimality.* Now that we have found the optimal theoretical value for the discriminator, we want to know when we have found a global optimum of the value function. *I.e.*, when is  $G^*$  obtained.

**Theorem 3.** The generator is optimal if  $p_{\text{model}} = p_{\text{data}}$  and at optimum, we have

$$V(G^*, D^*) = -\log 4.$$

*Proof.* Let  $p = p_{\text{data}}$  and  $q = p_{\text{model}}$ . We have already found the optimal  $D^*$ ,

$$D^*(x) = \frac{p(x)}{p(x) + q(x)}.$$

We substitute this into the value function,

$$\begin{aligned}
V(G, D^*) &= \mathbb{E}_{x \sim p} \left[ \log \left( \frac{p(x)}{p(x) + q(x)} \right) \right] + \mathbb{E}_{z \sim p_{\text{prior}}} \left[ \log \left( 1 - \frac{p(x)}{p(x) + q(x)} \right) \right] \\
&= \mathbb{E}_{x \sim p} \left[ \log \left( \frac{p(x)}{p(x) + q(x)} \right) \right] + \mathbb{E}_{z \sim p_{\text{prior}}} \left[ \log \left( \frac{q(x)}{p(x) + q(x)} \right) \right] \\
&= \mathbb{E}_{x \sim p} \left[ \log \left( \frac{2p(x)}{2(p(x) + q(x))} \right) \right] + \mathbb{E}_{z \sim p_{\text{prior}}} \left[ \log \left( \frac{2q(x)}{2(p(x) + q(x))} \right) \right] \\
&= \mathbb{E}_{x \sim p} \left[ \log \left( \frac{2p(x)}{p(x) + q(x)} \right) \right] - \log 2 + \mathbb{E}_{z \sim p_{\text{prior}}} \left[ \log \left( \frac{2q(x)}{p(x) + q(x)} \right) \right] - \log 2 \\
&= \mathbb{E}_{x \sim p} \left[ \log \left( \frac{2p(x)}{p(x) + q(x)} \right) \right] + \mathbb{E}_{z \sim p_{\text{prior}}} \left[ \log \left( \frac{2q(x)}{p(x) + q(x)} \right) \right] - \log 4 \\
&= D_{\text{KL}} \left( p \left\| \frac{p+q}{2} \right\| \right) + D_{\text{KL}} \left( q \left\| \frac{p+q}{2} \right\| \right) - \log 4 \\
&= 2D_{\text{JS}}(p \| q) - \log 4,
\end{aligned}$$

where the Jensen-Shannon divergence is a symmetric and smoothed version of the KL divergence, defined as

$$D_{\text{JS}}(p \| q) \doteq \frac{1}{2} D_{\text{KL}} \left( p \left\| \frac{p+q}{2} \right\| \right) + \frac{1}{2} D_{\text{KL}} \left( q \left\| \frac{p+q}{2} \right\| \right).$$

This divergence is non-negative, and equals 0 if and only if  $p = q$ . Thus,  $G^* \in \text{argmin}_G V(D^*, G)$  must satisfy  $p_{\text{data}} = p_{\text{model}}$ , and we obtain

$$V(G^*, D^*) = -\log 4.$$

■

*Convergence.* Under very strong assumptions, we can guarantee that a GAN converges.

**Theorem 4.** Assume that  $G$  and  $D$  have sufficient capacity, at each update step  $D \rightarrow D^*$ , and  $p_{\text{model}}$  is updated to improve

$$\begin{aligned}
V(p_{\text{model}}, D^*) &= \mathbb{E}_{x \sim p_{\text{data}}} [\log D^*(x)] + \mathbb{E}_{x \sim p_{\text{model}}} [\log(1 - D^*(x))] \\
&\propto \sup_D \int_x p_{\text{model}}(x) \log(1 - D(x)) dx.
\end{aligned}$$

Then,  $p_{\text{model}}$  converges to  $p_{\text{data}}$ .

Notice that  $p_{\text{model}}$  is updated directly here, rather than indirectly by optimizing  $G$ , which is actually what happens.

*Proof.* The argument of the supremum is convex in  $p_{\text{model}}$  and the supremum preserves convexity. Thus,  $V(p_{\text{model}}, D^*)$  is convex in  $p_{\text{model}}$  with global optimum as in Theorem 3. ■

However, Theorem 4 is a very weak result, because of how strong the assumptions are. In practice,  $G$  and  $D$  have finite capacity,  $D$  is optimized for only  $k$  steps and does not converge to  $D^*$ , and due to the neural network parametrization of  $G$ , the objective is no longer convex. However, despite this, GANs work well in practice, because  $D$  does stay close to  $D^*$ , providing meaningful gradients for  $G$  to optimize its generations.

## 9.2 Training

```

1: while not converged do
2:   repeat  $k$  times
3:      $\mathbf{x}_1, \dots, \mathbf{x}_n \sim p_{\text{data}}$ 
4:      $\mathbf{z}_1, \dots, \mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
5:      $\mathcal{L}_D = \frac{1}{n} \sum_{i=1}^n \log(D(\mathbf{x}_i)) + \log(1 - D(G(\mathbf{z}_i)))$ 
6:     perform a gradient ascent step on  $\mathcal{L}_D$ 
7:   end
8:    $\mathbf{z}_1, \dots, \mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
9:    $\mathcal{L}_G = \frac{1}{n} \sum_{i=1}^n \log(1 - D(G(\mathbf{z}_i)))$ 
10:  perform a gradient descent step on  $\mathcal{L}_G$ 
11: end while

```

A possible issue with this training algorithm is that, early in learning,  $G$  is poor, which means that  $D$  can easily reject samples with high confidence. In this case,  $\log(1 - D(G(z)))$  saturates, meaning that it approaches  $-\infty$  as  $D(G(z)) \rightarrow 1$ . Instead, we can train  $G$  to maximize  $\log D(G(z))$ , which does not have this problem; see Figure 22.

*Mode collapse.* We speak of *mode collapse* when the generator learns to produce high-quality samples with very low variability, covering only a fraction of the data distribution. A simple example that explains this phenomenon is a generator that generates temperature values. The generator may learn to only output cold temperatures, which the discriminator counters by predicting all cold temperatures as “fake” and all warm temperatures as “real”. Then, the generator exploits this by only generating warm temperatures. And, again the discriminator can counter this, which the generator counters, repeating cyclically.

The most common solution to mode collapse is the *unrolled GAN* [Metz et al., 2017]. The idea is to optimize the generator w.r.t. the last  $k$  discriminators. This results in the above not being able to occur, since the generator must not only fool the current discriminator, which might be unstable, but also the previous  $k$  ones.

*Training instability.* Since we optimize GANs as a two-player game, we need to find a Nash-Equilibrium, where, for both players, moving anywhere will only be worse than the equilibrium. However, this can lead to training instabilities, since making progress for one player may mean the other player being worse off.

*Optimizing Jensen-Shannon divergence.* It might be the case that the supports of  $p_{\text{data}}$  and  $p_{\text{model}}$  are disjoint. In this case, it is always possible to find a perfect discriminator with  $D(x) = 1, \forall x \in \text{supp}(p_{\text{data}})$  and  $D(x) = 0, \forall x \in \text{supp}(p_{\text{model}})$ . Then, the loss function equals zero, meaning that there will be no gradient to update the generator’s parameters.<sup>19</sup>

**Algorithm 1.** Generative adversarial network training algorithm.



**Figure 22.** Solution to that the GAN loss function saturates for the generator.

<sup>19</sup> As we saw in the proof of Theorem 3, GANs optimize the Jensen-Shannon divergence.

Nowozin et al. [2016] showed that the GAN objective can be generalized to an entire family of divergences. The Wasserstein GAN [Arjovsky et al., 2017] optimizes the Wasserstein distance between  $p_{\text{model}}$  and  $p_{\text{data}}$ . In this case, the loss does not fall to zero for disjoint supports, because it measures divergence by how different they are horizontally, rather than vertically. Intuitively, it measures how much “work” it takes to turn one distribution into the other.

*Gradient penalty.* To stabilize training, Mescheder et al. [2018] proposed adding a gradient penalty,

$$V(G, D) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} \left[ \log D(\mathbf{x}) + \lambda \|\nabla D(\mathbf{x})\|^2 \right] + \mathbb{E}_{\mathbf{x} \sim p_{\text{model}}} [\log(1 - D(\mathbf{x}))].$$

## 10 Diffusion models

Unlike VAEs and GANs, *diffusion models* [Ho et al., 2020] do not generate a sample in a single step. They do so in many small steps. They start from pure noise and iteratively denoise it. At the end of all the denoising steps, the goal is to obtain a sample from the data distribution. The diffusion chain can be traversed in two directions: going from noise to sample is called *denoising* and going from sample to noise is called *diffusion*. In general, we have

$$x_0 \sim q, \quad x_T \sim \mathcal{N}(\mathbf{0}, I),$$

where  $q$  is the data distribution and  $T$  is the total number of steps in the diffusion chain.  $x_t$  is then a noisy version of  $x_0$  at timestep  $t$ , where higher  $t$  means more noise. The goal of diffusion models is to train a model that can predict  $x_{t-1}$ , given  $x_t$  to reverse the diffusion process. It then performs these small steps  $T$  times, starting from  $x_T \sim \mathcal{N}(\mathbf{0}, I)$ .

### 10.1 Diffusion step

To generate the training data for a diffusion model, we need an efficient way of computing  $x_t$  and  $x_{t+1}$ . The diffusion chain is governed by a variance schedule  $(\beta_t \in (0, 1))_{t=1}^T$ , where  $\beta_t < \beta_{t+1}$  for all  $t$ .<sup>20</sup> A naive way of computing the diffusion steps is sequentially,

$$q(x_t | x_{t-1}) = \mathcal{N}(\sqrt{1 - \beta_t}x_{t-1}, \beta_t I).$$

However, this is very inefficient, since to compute  $x_t$ , we need to perform  $t$  diffusion steps. Luckily, there exists a closed-form solution for  $x_t$  by using the reparametrization trick,

$$x_t = \sqrt{1 - \beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, I).$$

Let  $\alpha_t = 1 - \beta_t$  and  $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$ . Using the new notation, we get

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1 - \alpha_t}\epsilon.$$

Using induction, we can find a closed-form solution,

$$\begin{aligned} x_t &= \sqrt{\alpha_t}(\sqrt{\alpha_{t-1}}x_{t-2} + \sqrt{1 - \alpha_{t-1}}\eta) + \sqrt{1 - \alpha_t}\epsilon, \quad \eta \sim \mathcal{N}(\mathbf{0}, I) \\ &= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{\alpha_t(1 - \alpha_{t-1})}\eta + \sqrt{1 - \alpha_t}\epsilon \\ &= \textcircled{*} \end{aligned}$$

Using the properties of multivariate Gaussians, we get

$$\begin{aligned} \sqrt{1 - \alpha_t}\epsilon &\sim \mathcal{N}(\mathbf{0}, (1 - \alpha_t)I) \\ \sqrt{\alpha_t(1 - \alpha_{t-1})}\eta &\sim \mathcal{N}(\mathbf{0}, \alpha_t(1 - \alpha_{t-1})I) \\ \implies \sqrt{\alpha_t(1 - \alpha_{t-1})}\eta + \sqrt{1 - \alpha_t}\epsilon &\sim \mathcal{N}(\mathbf{0}, (\alpha_t(1 - \alpha_{t-1}) + (1 - \alpha_t))I) \\ &= \mathcal{N}(\mathbf{0}, (1 - \alpha_t\alpha_{t-1})I). \end{aligned}$$

<sup>20</sup> There are two main ways of defining the scheduler; a linear schedule and a cosine schedule. It was found that linear schedulers add too much noise too quickly, making it hard for the model to learn. Thus, a cosine scheduler is usually preferred in practice.

Thus, we have

$$\sqrt{\alpha_t(1 - \alpha_{t-1})}\boldsymbol{\eta} + \sqrt{1 - \alpha_t}\boldsymbol{\epsilon} = \sqrt{1 - \alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}', \quad \boldsymbol{\epsilon}' \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Since  $\boldsymbol{\epsilon}$  and  $\boldsymbol{\epsilon}'$  are samples from the same distribution, we can continue using  $\boldsymbol{\epsilon}$ ,

$$\circledast = \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1 - \alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}$$

Continuing this pattern, we obtain

$$= \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}.$$

### 10.2 Denoising step

The reverse diffusion (denoising) process obviously has no closed-form solution. Thus, we want to parametrize a model  $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$  that performs the denoising from  $\mathbf{x}_t$  to  $\mathbf{x}_{t-1}$ . We defined  $q(\mathbf{x}_t | \mathbf{x}_{t-1})$  to be a Gaussian with known parameters. For small enough  $\beta_t$ , we can also model  $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$  as a Gaussian,

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)).$$

Similarly to VAEs, instead of predicting the full distribution (which would be very hard), we only need to predict the parameters of the Gaussian (which is not as hard). Using this model, we can compute the probability of the full process,

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t).$$

### 10.3 Training

Similarly to VAEs, we can derive an ELBO to optimize the log-likelihood,

$$\begin{aligned} \log p_\theta(\mathbf{x}_0) &= \log \int p_\theta(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T} \\ &= \log \int q(\mathbf{x}_{1:T} | \mathbf{x}_0) \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} d\mathbf{x}_{1:T} \\ &= \log \mathbb{E}_{\mathbf{x}_{1:T} \sim q(\cdot | \mathbf{x}_0)} \left[ \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \right] \\ &\geq \mathbb{E}_{\mathbf{x}_{1:T} \sim q(\cdot | \mathbf{x}_0)} \left[ \log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \right] && \text{Jensen's inequality.} \\ &\vdots \\ &= \underbrace{\mathbb{E}_{\mathbf{x}_{1:T} \sim q(\cdot | \mathbf{x}_0)} [\log p_\theta(\mathbf{x}_{0:T})]}_{\text{reconstruction term}} - \underbrace{D_{\text{KL}}(q(\mathbf{x}_T | \mathbf{x}_0) \| p(\mathbf{x}_T))}_{\text{prior matching term}} \\ &\quad - \underbrace{\sum_{t=2}^T \mathbb{E}_{\mathbf{x}_t \sim q(\cdot | \mathbf{x}_0)} [D_{\text{KL}}(q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t))]}_{\text{denoising matching term}}. \end{aligned}$$

In practice, we assume that the covariance matrix  $\Sigma_t = \sigma_t^2 \mathbf{I}$  of  $p$  and  $q$  are the same. Then, we can simplify the denoising matching term,

$$\begin{aligned}
& \underset{\theta}{\operatorname{argmin}} D_{\text{KL}}(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \parallel p_{\theta}(\mathbf{x}_{t-1} \mid \mathbf{x}_t)) \\
&= \underset{\theta}{\operatorname{argmin}} D_{\text{KL}}(\mathcal{N}(\mu_q, \Sigma_t) \parallel \mathcal{N}(\mu_{\theta}, \Sigma_t)) \\
&= \underset{\theta}{\operatorname{argmin}} \frac{1}{2} \left( \log \frac{\det(\Sigma_t)}{\det(\Sigma)} - d + \operatorname{tr}(\Sigma_t^{-1} \Sigma_t) + (\mu_{\theta} - \mu_q)^{\top} \Sigma^{-1} (\mu_{\theta} - \mu_q) \right) \\
&= \underset{\theta}{\operatorname{argmin}} \frac{1}{2} (\mu_{\theta} - \mu_q)^{\top} (\sigma_t^2 \mathbf{I})^{-1} (\mu_{\theta} - \mu_q) \\
&= \underset{\theta}{\operatorname{argmin}} \frac{1}{2\sigma_t^2} \|\mu_{\theta} - \mu_q\|^2.
\end{aligned}$$

Thus, we want to minimize the difference between the means of the two distributions. Assuming that we have a model that can predict the noise at timestep  $t$ , we can compute the two means by

$$\begin{aligned}
\mu_q(\mathbf{x}_t, t) &= \frac{1}{\sqrt{\alpha_t}} \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t} \sqrt{\alpha_t}} \epsilon \\
\mu_{\theta}(\mathbf{x}_t, t) &= \frac{1}{\sqrt{\alpha_t}} \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t} \sqrt{\alpha_t}} \hat{\epsilon}_{\theta}(\mathbf{x}_t, t).
\end{aligned}$$

A further simplification can be made by formulating the loss directly as the difference between the two actual and predicted noise,

$$\|\epsilon - \hat{\epsilon}_{\theta}(\mathbf{x}_t, t)\|^2 = \left\| \epsilon - \hat{\epsilon}_{\theta}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t) \right\|^2.$$

We then get the following algorithms for training and sampling a diffusion model.

**Require:**  $\{\beta_t\}_{t=1}^T$

```

1: while not converged do
2:    $\mathbf{x}_0 \sim q$ 
3:    $t \sim \text{Unif}([T])$ 
4:    $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
5:    $\mathcal{L} = \|\epsilon - \hat{\epsilon}_{\theta}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t)\|^2$ 
6:   perform a gradient descent step on  $\mathcal{L}$ 
7: end while

```

**Algorithm 2.** Diffusion model training algorithm.

**Require:**  $\{\beta_t\}_{t=1}^T, \hat{\epsilon}_{\theta}$

```

1:  $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
2: for  $t = T, \dots, 1$  do
3:    $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  if  $t > 1$  else  $\mathbf{z} = \mathbf{0}$ 
4:    $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \hat{\epsilon}_{\theta}(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$ 
5: end for
6: return  $\mathbf{x}_0$ 

```

**Algorithm 3.** Diffusion model sampling algorithm.

Typically,  $\epsilon_{\theta}$  is implemented as a U-net that is shared across timesteps.

## 10.4 Guidance

In use-cases such as text-to-image, we want to condition the generated image  $x$  on the input text  $y$ , meaning that we need to model a conditional distribution,

$$p_{\theta}(x | y) = p(x_T) \prod_{t=1}^T p_{\theta}(x_{t-1} | x_t, y).$$

*Contrastive language-image pretraining* (CLIP) [Radford et al., 2021] is a large image-language model that has been trained on image-caption pairs. It consists of an image encoder and a text encoder network. By using a contrastive loss, CLIP is encouraged to encode the image and caption into similar embeddings.<sup>21</sup>

The idea of *classifier guidance* is to guide denoising in a “direction” favoring images that are more reliably classified by a trained classifier. For this we need a pretrained unconditional diffusion model and a classifier trained on noisy images. Then, we guide the denoising in the “direction” of the classifier by injecting gradients of the classifier model into the sampling process. However, the problem is that this requires training a very specific classifier on noisy data, because we want to guide the diffusion model in all steps of the process. Furthermore, it is hard to interpret what the classifier guidance is doing.

*Classifier-free guidance* [Ho and Salimans, 2022] address these issues by jointly training a class-conditional and unconditional diffusion model. It then guides the generation process in the generation of conditioning by

$$\epsilon_{\theta}^*(x, y; t) = (1 + \rho)\epsilon_{\theta}(x, y; t) - \rho\epsilon_{\theta}(x; t),$$

where  $y$  is the conditioning variable and is usually obtained by encoding text using CLIP and  $\rho$  controls the strength of the guidance.<sup>22</sup> However, the problem with this approach is that generation takes twice as long when compared to classifier guidance. Overall, guidance improves the quality of the outputs, but reduces the diversity of the generated images.

## 10.5 Latent diffusion models

Diffusion models for image generation typically operate on the original, high-dimensional image size, which can result in slow training and sampling. *Latent diffusion models* [Rombach et al., 2022] address this issue by operating on the latent space of a VAE. The VAE is trained beforehand and is frozen during training of the diffusion model. The diffusion model then does the diffusion reversal on the latent space, rather than the high-dimensional space of the data. This has the advantage that it significantly improves the efficiency of sampling and training. Furthermore, in this approach, the diffusion model only needs to focus on the “semantic” aspect of the image, because the VAE has already captured the relevant information in the latent space.

<sup>21</sup> A possible application of CLIP is *zero-shot classification*, which leverages the CLIP model to predict the class of an image without any training. It achieves this by predicting the class that maximizes the cosine similarity between the image and the class name.

<sup>22</sup> In practice, we usually train a single model and just set the conditioning variable to all zero for the unconditional generation.



## 11 Reinforcement learning

**Definition 5** (Markov decision process). A Markov decision process (MDP) is a 5-tuple  $\langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$ , where

- $\mathcal{S}$  is the state space;
- $\mathcal{A}$  is the action space;
- $P : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$  is the state transition model;
- $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  is the expected reward function;
- $\gamma \in [0, 1]$  is the discount factor. If  $\gamma = 1$ , all future rewards count as much as current reward. If  $\gamma = 0$ , only immediate rewards matter. Usually, we want a value between 0 and 1, because we want a balance between the two extremes.

An MDP can be seen as a controlled Markov chain, because the next state  $s_{t+1}$  is independent of all states and actions before  $t$  if we know  $(s_t, a_t)$ ,

$$\mathbb{P}(s_{t+1} = s' \mid s_t = s, a_t = a, \dots, s_0, a_0) = P(s' \mid s, a).$$

The performance objective is maximizing the expected discounted reward.

**Definition 6** (Policy). A policy  $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$  expresses an agent's behavior by mapping states to a probability distribution over actions. A policy can be made deterministic by outputting only Dirac distributions.

With a fixed policy, we can then put the probability of the next action as

$$\mathbb{P}(a_t = a \mid s_t = s, \dots, s_0, a_0) = \pi(a \mid s).$$

At any point in the Markov chain, the objective of an agent is to maximize the (discounted) return,

$$G_t = \sum_{k=0}^{\infty} \gamma^k r(s_{t+k}, a_{t+k}).$$

This can also be written in recursive form as

$$G_t = r(s_t, a_t) + \gamma G_{t+1}.$$

**Definition 7** (Value function). The value function  $V^\pi : \mathcal{S} \rightarrow \mathbb{R}$  is defined as the expected return under policy  $\pi$ , starting from state  $s$ ,

$$V^\pi(s) \doteq \mathbb{E}_\pi[G_0 \mid s_0 = s] = \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s \right].$$



Figure 23. Diagram of an MDP.

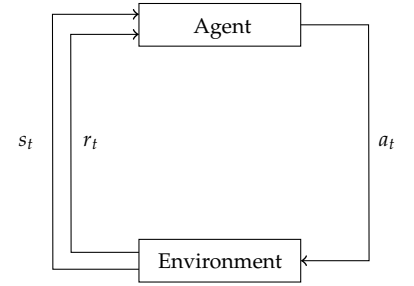


Figure 24. Reinforcement learning.

From the definition of the value function, we can derive the Bellman consistency equation,

$$\begin{aligned}
V^\pi(s) &\doteq \mathbb{E}_\pi[G_0 \mid s_0 = s] \\
&= \sum_{a \in \mathcal{A}} \pi(a \mid s) \mathbb{E}_\pi[r(s, a) + \gamma G_1 \mid s_0 = s, a_0 = a] && \text{Recursive definition of return and condition on action.} \\
&= \sum_{a \in \mathcal{A}} \pi(a \mid s) [r(s, a) + \gamma \mathbb{E}_\pi[G_1 \mid s_0 = s, a_0 = a]] && \text{Linearity of expectation.} \\
&= \sum_{a \in \mathcal{A}} \pi(a \mid s) \left[ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) \mathbb{E}_\pi[G_1 \mid s_1 = s'] \right] && \text{Condition on next state, which makes } (s_0, a_0) \text{ irrelevant for further return due to Markov property.} \\
&= \sum_{a \in \mathcal{A}} \pi(a \mid s) \left[ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) \mathbb{E}_\pi[G_0 \mid s_0 = s'] \right] && \text{Markov property.} \\
&= \sum_{a \in \mathcal{A}} \pi(a \mid s) \left[ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) V^\pi(s') \right].
\end{aligned}$$

Furthermore, we have the Bellman optimality equation,

$$V^*(s) = \max_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) V^*(s') \right\}.$$

This must always hold for the optimal policy, because the optimal policy acts greedily w.r.t. the optimal value function. We can formalize this as the Bellman optimality operator

$$(\mathcal{T}v)(s) \doteq \max_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) v(s') \right\}.$$

$v$  is the vector containing all state values, where  $v_s = V^\pi(s)$ .

This operator is a  $\gamma$ -contraction w.r.t. the  $\ell_\infty$ -norm,

$$\|\mathcal{T}v' - \mathcal{T}v\|_\infty \leq \gamma \|v' - v\|_\infty,$$

and monotonic,

$$v \leq v' \implies \mathcal{T}v \leq \mathcal{T}v'.$$

As we saw above,  $v^*$  is the fixed-point of the Bellman optimality operator,

$$v^* = \mathcal{T}v^*.$$

Thus, by applying  $\mathcal{T}$  iteratively, we converge linearly in  $\gamma$ ,

$$\|\mathcal{T}v_t - v^*\|_\infty = \|\mathcal{T}v_t - \mathcal{T}v^*\|_\infty \leq \gamma \|v_t - v^*\|.$$

This telescopes, giving

$$\|v_t - v^*\| \leq \gamma^t \|v_0 - v^*\|.$$

As  $t \rightarrow \infty$ ,  $\gamma^t \rightarrow 0$ , showing convergence. This leads us to *value iteration*, which does exactly that; see Algorithm 4. After computing the optimal value function, we can get the optimal policy by acting greedily w.r.t. the returned value function,

$$\pi^*(s) \in \operatorname{argmax}_{a \in \mathcal{A}} \left\{ r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) V^*(s') \right\}.$$

```

1:  $v_0 \leftarrow \mathbf{0}$ 
2: while  $\|v_t - v_{t-1}\|_\infty > \epsilon$  do
3:    $v_{t+1} = \mathcal{T}v_t$ 
4: end while
5: return  $v$ 

```

The runtime complexity of a single iteration is  $\mathcal{O}(|\mathcal{S}|^2|\mathcal{A}|)$ .

*Policy iteration* is another algorithm that computes the optimal policy given a known MDP. Instead of only operating in the space of the value function, it alternates between computing the value function and the greedy policy w.r.t. the value function; see Algorithm 5. This algorithm converges in fewer iterations than value iteration, but has higher iteration cost of  $\mathcal{O}(|\mathcal{S}| + |\mathcal{S}|^2|\mathcal{A}|)$ .

```

1:  $\pi_0 \leftarrow$  random policy
2: while  $\|v^{\pi_t} - v^{\pi_{t-1}}\|_\infty > \epsilon$  do
3:    $V^{\pi_t} \leftarrow \text{VALUEFUNCTION}(\pi_t)$  ▷ Policy evaluation
4:    $\pi_{t+1} \leftarrow \text{GREEDYPOLICY}(V^{\pi_t})$  ▷ Policy improvement
5: end while

```

*Reinforcement learning.* Unlike in MDPs, in reinforcement learning (RL), we do not have access to the transition model  $P$  and reward function  $r$ . Thus, we cannot make use of value iteration or policy iteration. In some way, we must learn to act optimally within an environment by interacting with it. We can distinguish between two kinds of RL algorithms: *model-free* and *model-based*. In model-based RL, we learn the underlying MDP by approximating  $P$  and  $r$ . Then, we solve this learned MDP by value or policy iteration. However, we do not *need* the underlying MDP to act optimally. In model-free RL, we learn either the value function, which induces a greedy policy, or directly learn the policy. Generally, model-free algorithms are less expensive to run, while model-based algorithms are more sample efficient.

Furthermore, we can distinguish between RL algorithms in how they learn from data; *on-policy* and *off-policy*. On-policy algorithms must learn from actions that result from its own actions, while off-policy algorithms can learn from the trajectory data of any algorithm. Generally, off-policy are preferred, since we can collect data more efficiently.

Lastly, a way that RL differs from supervised learning is that the data depends on past actions, *i.e.*, they are highly correlated, which violates the *independently and identically distributed* assumption made by supervised learning. RL algorithms learn from trajectory data,

$$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, s_2, \dots).$$

**Algorithm 4.** Value iteration, where  $\mathcal{T}$  is the Bellman optimality operator.

**Algorithm 5.** Policy iteration.



**Figure 25.** Overview of the kinds of reinforcement learning algorithms.

### 11.1 Monte Carlo methods

The most naive method is the *Monte Carlo method*, which is a value-based on-policy algorithm that estimates the value function by the empirical mean return,

$$V^\pi(s) \approx \frac{1}{n} \sum_{i=1}^n G(s)^i,$$

where  $G(s)^i$  is the return of episode  $i$ , starting from  $s$ . The problem with this method is that we need to wait for a trajectory to finish before we can use it for approximating the value function.

### 11.2 Temporal difference learning

*Temporal difference* (TD) learning makes it possible to learn from transitions  $(s, a, r, s')$ , instead of full trajectories, by approximating the Bellman equation by a single reward, action, and next-state sample,

$$V(s) \leftarrow \alpha V(s) + (1 - \alpha)(r + \gamma V(s')),$$

where  $\alpha > 0$  is the learning rate. We use *bootstrapping* here, which means that we use “old” information  $V(s')$  as labels. This can also be interpreted as updating the value by the TD error,

$$\begin{aligned} \delta &= r + \gamma V(s') - V(s) \\ V(s) &\leftarrow V(s) + \alpha \delta. \end{aligned}$$

However, to find the optimal value function, we must visit all states sufficiently often. This means that we have to find a balance between exploration and exploitation. A commonly used method is the  $\epsilon$ -greedy policy, which chooses a random action with small probability  $\epsilon$ .

**SARSA.** SARSA [Rummery and Niranjan, 1994] is an on-policy algorithm that learns the Q-values of a policy  $\pi$ . It updates Q-values given a transition  $(s, a, r, s')$  as follows,

$$Q(s, a) \leftarrow \alpha Q(s, a) + (1 - \alpha)(r + \gamma Q(s', a')), \quad a' \sim \pi(\cdot, s').$$

This is on-policy, because it requires the policy for the update.

**Q-learning.** An off-policy version of SARSA is *Q-learning* [Watkins and Dayan, 1992]. Note that this learns the optimal Q-values, rather than the Q-values of a certain policy. It updates the Q-values by

$$Q(s, a) \leftarrow \alpha Q(s, a) + (1 - \alpha)(r + \gamma \max_{a' \in \mathcal{A}} Q(s', a')), \quad a' \in \arg\max_{a' \in \mathcal{A}} Q(s', a').$$

Notice that this is an off-policy algorithm, because nowhere in the learning algorithm does it depend on the policy. Given optimal Q-values, it is easy to compute the optimal policy by always acting greedily,

$$\pi(s) \in \arg\max_{a \in \mathcal{A}} Q(s, a).$$

### 11.3 Deep reinforcement learning

Notice that the policy and the value are simply functions,

$$\pi : \mathcal{S} \rightarrow \mathcal{A}, \quad V^\pi : \mathcal{S} \rightarrow \mathbb{R}.$$

Thus, we can use neural networks to approximate them. This is especially useful for large state and action spaces, since these would not fit in memory.

*Replay buffer.* The key component to training the following models is the *replay buffer*. Since transitions are highly correlated and we want to use supervised learning techniques, we cannot always use the latest transition for training. Instead, the replay buffer stores the last  $n$  transitions and randomly samples from them, making the data points independent and identically distributed.

*Deep Q-learning.* Deep Q-networks (DQN) [Mnih et al., 2013] estimate the value function of a large, potentially infinite, state space with a finite number of parameters  $\theta$ ,

$$V(s) \approx V_\theta(s).$$

It does so in a very similar way to discrete Q-learning. However, instead of updating the Q-values directly, the parameters of the network are updated by the gradients of the following loss function given a transition  $(s, a, r, s')$ ,

$$\ell(\theta) = (Q_\theta(s, a) - (r + \gamma Q_{\bar{\theta}}(s', a')))^2, \quad a' \in \underset{a \in \mathcal{A}}{\operatorname{argmax}} Q_{\bar{\theta}}(s', a).$$

Notice that we use different parameters  $\bar{\theta}$  for the target value. This is a copy of the parameters  $\theta$  that is only updated occasionally. This helps with convergence, because the target is more stable than if we used  $\theta$ .

*Policy search methods.* The problem with DQNs is that they do not address potentially large state spaces. *Policy search methods* solve this by directly parametrizing the policy  $\pi_\theta$ . However, it is not trivial to train such a model, since there is no way of knowing what the target action should be. Furthermore, we want the policy to be a probability distribution. Thus, instead we parameterize a Gaussian as the policy,

$$\pi(a | s) = \mathcal{N}(a; \mu_\theta, \sigma_\theta^2).$$

Recall that the probability of a trajectory  $\tau$  is computed by

$$\pi_\theta(\tau) = p(s_0) \prod_{t=0}^T \pi_\theta(a_t | s_t) p(s_{t+1} | a_t, s_t).$$

And, ideally, we want our policy to make trajectories with high return more likely and trajectories with low return unlikely. Thus, we have the

following training objective that we want to maximize,

$$J(\theta) \doteq \mathbb{E}_{\tau \sim \pi_\theta} \left[ \sum_{t=0}^T \gamma^t r(s_t, a_t) \right].$$

Then, we update the parameters by gradient ascent,

$$\theta \leftarrow \theta + \eta \nabla_\theta J(\theta).$$

But, since we take the expectation w.r.t. the parameters  $\theta$ , we need to rederive it, such that we can compute its gradient. First, we can rewrite the expectation as

$$\begin{aligned} J(\theta) &= \mathbb{E}_{\tau \sim \pi_\theta} [r(\tau)] \\ &= \int \pi_\theta(\tau) r(\tau) d\tau. \end{aligned}$$

$$\text{Let } r(\tau) \doteq \sum_{t=0}^T \gamma^t r(s_t, a_t).$$

Now, we can rewrite the gradient,

$$\begin{aligned} \nabla_\theta J(\theta) &= \int \nabla_\theta \pi_\theta(\tau) r(\tau) d\tau \\ &= \int \pi_\theta(\tau) \nabla_\theta \log \pi_\theta(\tau) r(\tau) d\tau \\ &= \mathbb{E}_{\tau \sim \pi_\theta} [\nabla_\theta \log \pi_\theta(\tau) r(\tau)]. \end{aligned}$$

$$\text{By chain rule, } \nabla \log f(x) = \nabla f(x) / f(x).$$

We can evaluate  $\log \pi_\theta$  as

$$\begin{aligned} \log \pi_\theta(\tau) &= \log \left( p(s_0) \prod_{t=0}^T \pi_\theta(a_t | s_t) p(s_{t+1} | a_t, s_t) \right) \\ &= \log p(s_0) + \sum_{t=0}^T \log \pi_\theta(a_t | s_t) + \sum_{t=0}^T \log p(s_{t+1} | a_t, s_t). \end{aligned}$$

Thus, when we take the gradient, only the second term remains,

$$\nabla_\theta \log \pi_\theta(\tau) = \sum_{t=0}^T \nabla_\theta \log \pi_\theta(a_t | s_t).$$

So, we are able to update  $\theta$  without knowing the MDP. Thus, we get the following objective gradient,

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \left[ \left( \sum_{t=0}^T \nabla_\theta \log \pi_\theta(a_t | s_t) \right) \left( \sum_{t=0}^T \gamma^t r(s_t, a_t) \right) \right].$$

Despite that this value is unbiased, it exhibits high variance. To reduce the variance, we can introduce a baseline  $b(\tau)$ ,

$$\mathbb{E}_{\tau \sim \pi_\theta} \left[ \left( \sum_{t=0}^T \nabla_\theta \log \pi_\theta(a_t | s_t) \right) (r(\tau) - b(\tau)) \right],$$

which does not influence the unbiasedness.<sup>23</sup> Thus, we are allowed to shift the reward up or down to reduce the variance. REINFORCE [Sutton et al., 1999] sets this baseline to be

$$b_t(\tau) = \sum_{t'=0}^{t-1} \gamma^{t'} r(s_{t'}, a_{t'}).$$

<sup>23</sup> Critically, the baseline may not depend on the policy  $\pi_\theta$ .

Thus, we get the following expectation,

$$\mathbb{E}_{\tau \sim \pi_{\theta}} \left[ \left( \sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right) G_t \right].$$

*I.e.*, instead of total discounted reward, we only consider the reward from timestep  $t$  onward, which naturally has lower variance. Note that REINFORCE is an on-policy algorithm, because it requires training on its own trajectory data, since the expectation is w.r.t. trajectories from the policy itself.

*Actor-critic methods.* A natural next question is how we can make this off-policy. The key idea behind *actor-critic methods* is to introduce bias and reduce variance by learning a value network next to the policy network. We can then estimate  $G_t$  in the REINFORCE gradient by bootstrapping,

$$\nabla_{\theta} \log \pi_{\theta}(a_t | s_t) (V(s_t) - (r(s_t, a_t) + \gamma V(s_{t+1}))).$$

We can remove the expectation since nothing depends on  $\tau$  anymore, since we estimate  $G_t$  as  $V(s_t) - (r(s_t, a_t) + \gamma V(s_{t+1}))$ . Note that if the TD error is zero, it means that the value network has learned the optimal value function and the policy network has learned the optimal policy, leading to no update.

## References

- Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.
- Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. *Advances in neural information processing systems*, 28, 2015.
- Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.
- Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.
- Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.
- Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.
- Brendan J Frey. *Graphical models for machine learning and digital communication*. MIT press, 1998.
- Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: Masked autoencoder for distribution estimation. In *International conference on machine learning*, pages 881–889. PMLR, 2015.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017.
- Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851, 2020.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feed-forward networks are universal approximators. *Neural networks*, 2(5): 359–366, 1989.



- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible  $1\times 1$  convolutions. *Advances in neural information processing systems*, 31, 2018.
- Mark A Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2):233–243, 1991.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018.
- Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks, 2017.
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013.
- Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, 29, 2016.
- Christopher Olah. Understanding lstm networks, 2015. URL <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.
- Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.
- Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

- Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*, volume 37. University of Cambridge, Department of Engineering Cambridge, UK, 1994.
- Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation, 2016.
- Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.
- Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models, 2016.
- Benigno Uria, Marc-Alexandre Côté, Karol Gregor, Iain Murray, and Hugo Larochelle. Neural autoregressive distribution estimation. *Journal of Machine Learning Research*, 17(205):1–37, 2016.
- Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio, 2016.
- Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.
- Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8:279–292, 1992.
- Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions, 2016.