Machine Learning Explainers

Last updated: September 26, 2025

Abstract

Summaries of some machine/deep learning-related topics. My main motive in writing these summaries is as a reminder for my future self.

Contents

1 What is Machine Learning?

I think of machine learning as the broad discipline of studying methods of fitting models to data — like statistics, as a discipline, if it demanded far less rigour. For a description of machine learning which helps to clarify why it's difficult to define as a term, consider the following excerpt from Herbert Jaeger's lecture notes for his Machine Learning course at the University of Groningen during the academic year 2023/24:

> ML as a field, which perceives itself as a field under this name, is relatively young, say, about 40 years (related research was called "pattern recognition" earlier). It is interdisciplinary and has historical and methodological connections to neuroscience, cognitive science, linguistics, mathematical statistics, AI, signal processing and control; it uses mathematical methods from statistics (of course), information theory, signal processing and control, dynamical systems theory, mathematical logic and numerical mathematics; and it has a very wide span of applications. This diversity in traditions, methods and applications makes it difficult to study "Machine Learning". Any given textbook, even if it is very thick, will reflect the author's individual view and knowledge of the field and will be partially blind to other perspectives. This is quite different from other areas in computer science, say for example formal languages/theory of computation/computational complexity where a widely shared repertoire of standard themes and methods cleanly define the field.

My interests in machine learning lie in its rapid development since the era of deep learning began in 2012 and the mysteries of why its methods are so effective. At their core, many methods in machine learning are statistically-principled, corresponding to maximum likelihood estimation. The fact that something as simple as maximum likelihood estimation can be used to fit very complex distributions to at least a small extent isn't too surprising. What fascinates me is the notable extent to which it does so in practice. It performs so well that we are able to produce models which take natural language as input and output entirely realistic corresponding images/videos. What right does maximum likelihood estimation have to facilitate such effective fitting of complex distributions in practice? Why are deep learning architectures able to encode such rich function classes? We only have partial answers to the many natural questions like these and so our understanding is far from complete. What an interesting time to be alive. :)

2 Supervised Learning

Supervised learning methods fit functions $f_\theta : \Omega_{\mathbf{X}} \to \Omega_{\mathbf{Y}}$ from model variables $\mathbf{X} = (X_1, \ldots, X_q)$ to output variables $\mathbf{Y} = (Y_1, \ldots, Y_r)$ (often $r = 1$ in which case we write $\mathbf{Y} = Y$) by tweaking θ in line with concrete examples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ of how the function may behave in practice. The two most prominent examples of supervised learning tasks are regression (continuous output variables) and classification (discrete output variables).

2.1 Linear Regression

Perhaps the simplest example of supervised learning is linear regression. Suppose we are given a dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{2n}$ where

$$\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,q}) \in \Omega_{X_1} \times \cdots \times \Omega_{X_q} =: \Omega_{\mathbf{X}} \subseteq \mathbb{R}^q$$

are the feature values of the i^{th} sample and $y_i \in \Omega_Y \subseteq \mathbb{R}$ is its corresponding output. The reason I chose $|D| = 2n$ is that it yields a convenient partition of the data into D_{train} and D_{test} each of n samples. A linear regression models fits a model of the form

$$\begin{aligned} f_\theta : \Omega_{\mathbf{X}} &\to \Omega_Y \\ \mathbf{x} &\mapsto \theta^\top \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix} = \theta_0 + \theta_1 x_1 + \cdots + \theta_q x_q \end{aligned}$$

where $\theta = (\theta_0, \theta_1 \ldots, \theta_q) \in \mathbb{R}^{q+1}$ are the model parameters, i.e. the values we can tweak to our heart's content until testing error is sufficiently low. Some people refer to the paramter θ_0, which dictates the elevation of the hyperplane corresponding to $f_\theta(\mathbf{x}) = 0$, as the bias of the model which I find very confusing as there are a bunch of other intended meanings of the term 'bias' in statistics and machine learning. I prefer to refer to it as the elevation. Anyway, once such a linear function has been fit, given feature values $\mathbf{x} \in \Omega_{\mathbf{X}}$ our model predict the corresponding output as $y = f_\theta(\mathbf{x})$.

What does the 'linear' in linear regression actually refer to?

Casella Berger, as well as other pieces of literature, define linear regression models as being linear in their parameters. By such a definition, linear regression, as a term, encompasses polynomial regression models and other regression models with non-linear basis functions. This confuses me since, in my experience, 'linear regression' is most often intended to mean models that fit a hyperplane to $\Omega_{\mathbf{X}} \times \Omega_Y$.

The first paragraph of the Wikipedia page[a] for polynomial regression reiterates this potential confusion.

[a] `https://en.wikipedia.org/wiki/Polynomial_regression`

An intuitive approach to finding the 'optimal' parameters for a linear regression model, which we denote by θ^*, is to split D into training and testing datasets D_{train} and D_{test} (each consisting of n sample in our case) and minimising some pre-determined loss function of said parameters over D_{train}. Essentially, minimising something like

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} d(f_\theta(\mathbf{x}_i), y_i)$$

where $f_\theta(\mathbf{x}_i)$ is the model's prediction for feature values $\mathbf{x}_i$ and $d : \Omega_Y \times \Omega_Y \to \mathbb{R}_{\geq 0}$ is some goodness-of-prediction metric or loss function. Said loss function gives one an idea of how well θ fits the true underlying relationship which we wish to model. A common choice for the loss function is the mean square error

$$\text{MSE}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_\theta(\mathbf{x}_i))^2$$

which pertains to taking $d(f_\theta(\mathbf{x}_i), y_i) = (y_i - f_\theta(\mathbf{x}_i))^2$. Note that the factor of $1/n$ is often left out when discussing MSE as minimising the expression in θ is invariant to the inclusion of the factor. So in our case, we seek to compute

$$\theta^* = \underset{\theta \in \mathbb{R}^{q+1}}{\arg\min} \left[\sum_{i=1}^{n}(y_i - f_\theta(\mathbf{x}_i))^2\right].$$

We know that in the context of linear regression, the optimal paramters θ^* are typically taken to be those which minimise the mean square error over D_{train} but how do we actually compute θ^*? This could be done through numerical methods, which is often the case in machine learning, e.g. using

gradient descent in computing the optimal parameters of a logistic regression model, but linear regression has a closed form solution. This is pretty cool since it's not so common for such closed form solutions to exist in machine learning-related contexts (though stats people might not like linear regression being referred to as machine learning-related). The informal method which I use to remember the closed form solution for the optimal paramters of a linear regression model is

$$X\theta^* = y \implies X^\top X\theta^* = X^\top y \implies \theta^* = (X^\top X)^{-1}X^\top y.$$

In practice, if this matrix $X^\top X$ is singular then just add some small values to its diagonal. That is, instead compute

$$\theta^* = (X^\top X + \delta I)^{-1}X^\top y$$

for some small $\delta \in \mathbb{R}$. That said, what's given above is not rigorous as it assumes the existance of θ^* and does not make use of MSE, so let's derive it. Note that

$$\begin{aligned}
\text{MSE}(\theta) &= \sum_{i=1}^{n}(y_i - f_\theta(\mathbf{x}_i))^2 \\
&= \begin{bmatrix} y_1 - f_\theta(\mathbf{x}_1) & \cdots & y_n - f_\theta(\mathbf{x}_n) \end{bmatrix} \begin{bmatrix} y_1 - f_\theta(\mathbf{x}_1) \\ \vdots \\ y_n - f_\theta(\mathbf{x}_n) \end{bmatrix} \\
&= (y - X\theta)^\top (y - X\theta) \\
&= ||y - X\theta||^2
\end{aligned}$$

where

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n, X = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,q} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,q} \end{bmatrix} \in \mathbb{R}^{n\times(q+1)}, \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_q \end{bmatrix} \in \mathbb{R}^{q+1}$$

and $x_{i,j}$ denotes the jth element of the i^{th} sample. To find the minimiser(s) of MSE(θ), i.e. the optimal parameters θ^*, we compute its gradient with respect to θ and find its root(s). This of course only works when our function is both differentiable and convex, which is the case here. To see convexity, simply compute the Hessian of MSE(θ) and see that it is semi-positive

definite. On that note, we have

$$\begin{aligned}\nabla \text{MSE}(\theta) &= \nabla ||y - X\theta||^2 \\ &= \nabla (y - X\theta)^\top (y - X\theta) \\ &= \nabla \left[\theta^\top X^\top X\theta - \theta^\top X^\top y - y^\top X\theta + y^\top y\right] \\ &= \nabla \left[\theta^\top X^\top X\theta - 2y^\top X\theta + y^\top y\right] \\ &= 2X^\top X\theta - 2X^\top y\end{aligned}$$

and so the optimiser θ^* is given by

$$\theta^* = (X^\top X)^{-1} X^\top y.$$

Funny misuse of linear regression: Momentous sprint at the 2156 Olympics?

Read this Quora answer[a] based on a paper[b] published in Nature. The top comment is worth reading too. Related xkcd comic[c].

[a] `https://qr.ae/pslbEN`
[b] `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3173856/`
[c] `https://xkcd.com/1007/`

2.1.1 Statistical Motivation

The idea of minimising the mean square error of the model over D_{train} has a rigorous statistical motivation, corresponding to maximum likelihood estimation. Assume that the residuals corresponding of the model's output over D_{train} are independent and identically normally distributed with mean 0. That is, assume

$$E_i = Y_i - f_\theta(\mathbf{X}_i) \sim \mathcal{N}(0, \sigma^2)$$

for $i = 1, \ldots, n$ are i.i.d. This assumption is reasonable in practice due to the central limit theorem (CLT). We have $Y_i | \mathbf{X}_i \sim \mathcal{N}(f_\theta(\mathbf{x}_i), \sigma^2)$ and so

$$p_\theta(y_i | \mathbf{x}_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - f_\theta(\mathbf{x}_i))^2}{2\sigma^2}\right).$$

The maximum likelihood estimate θ_{MLE} of the parameters of our linear regression model is that which maximise the relevant log-likelihood. That

is,

$$\begin{aligned}
\theta_{\text{MLE}} &= \arg\max_{\theta \in \mathbb{R}^{q+1}} \left[\log \left(\prod_{i=1}^{n} p_\theta(y_i | \mathbf{x}_i) \right) \right] \\
&= \arg\max_{\theta \in \mathbb{R}^{q+1}} \left[\sum_{i=1}^{n} \log(p_\theta(y_i | \mathbf{x}_i)) \right] \\
&= \arg\max_{\theta \in \mathbb{R}^{q+1}} \left[n \log \left(\frac{1}{\sqrt{2\pi}\sigma} \right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - f_\theta(\mathbf{x}_i))^2 \right] \\
&= \arg\min_{\theta \in \mathbb{R}^{q+1}} \left[\sum_{i=1}^{n} (y_i - f_\theta(\mathbf{x}_i))^2 \right].
\end{aligned}$$

As such, the maximum likelihood estimate of the parameters of our linear regression model are precisely those which minimise the mean square error of the model over D_{train}.

Are residuals really normally distributed?

[a]Central limit theorem to the rescue: "Regression analysis, and in particular ordinary least squares, specifies that a dependent variable depends according to some function upon one or more independent variables, with an additive error term. Various types of statistical inference on the regression assume that the error term is normally distributed. This assumption can be justified by assuming that the error term is actually the sum of many independent error terms; even if the individual error terms are not normally distributed, by the central limit theorem their sum can be well approximated by a normal distribution."

[a]`https://en.wikipedia.org/wiki/Central_limit_theorem#Regression`

2.1.2 Goodness of fit: R^2

If we'd like a way to measure the goodness-of-fit of a linear regresison model beyond test MSE, a natural avenue is to assess what portion of the sample variance is explained by the model. As usual, we do this over some sample $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \subset \Omega_{\mathbf{X}} \times \Omega_Y$. That is, computing

$$\frac{\text{Var}(f_\theta(\mathbf{X}))}{\text{Var}(Y)} \approx \frac{\frac{1}{n}\sum_{i=1}^{n} \left(f_\theta(\mathbf{x}_i) - \bar{y}\right)^2}{\frac{1}{n}\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2} =: R^2$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. The reason that this quantity is denoted by R^2 is that it can be shown that it is equal to the square of the Pearson correlation between Y and $f_\theta(\mathbf{X})$. For one predictor, i.e. $q = 1$, said Pearson correlation coefficient is given by

$$R = \frac{\sum_{i=1}^{n} \left(f_\theta(\mathbf{x}_i) - \bar{f}_\theta\right)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} \left(f_\theta(\mathbf{x}_i) - \bar{f}_\theta\right)^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}.$$

The derivation of this statement is boring, so I'll leave it out. It's worth noting that R^2 is typically expressed as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - f_\theta(\mathbf{x}_i))^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{s_e^2}{s_Y^2}$$

where s_e^2 and s_Y^2 denote the sample variances of the residuals and Y respectively. I prefer the ratio of explained variance to underlying model variance. I find it more intuitive, especially its motviation as it doesn't feel like it just comes out of nowhere.

This notion of the portion of explained variance can be extended to logistic regression but the details are a bit much for this document.

How can we get an idea of the extent to which our features and output are linearly related without plotting?

To be honest, I thought a streamline test for this existed. As per this Stack Exchange answer[a], a good method for this is to simply fit a linear and a non-linear model (e.g. a cubic spline smoother model) and see which explains a larger amount of the variance of the output via ANOVA tests.

[a] `https://stats.stackexchange.com/a/239142`

2.1.3 Why call it 'regression'?

Nowadays, regression models are those which predict a value belonging to some continuous space but where does the term come from? In 1886, Francis Galton authored "Regression Towards Mediocrity in Hereditary Stature" which is where the 'regression towards the mean' term comes from. Loosely speaking, Galton noticed that tall fathers tend to have sons which are taller than average but shorter than them, short fathers tend to have sons which are shorter than average but taller than them and average height fathers

tend to have average height sons. Taken from a Stack Exchange comment: "Galton derived a linear approximation to estimate a son's height from the father's height in that paper. His equation was fitted so an average height father would have an average height son, but a taller than average father would have a son that is taller than average by 2/3 the amount his father is. Same with shorter than average. This could be argued to be a simple linear regression."

Put mathematically, suppose random variables X and Y are related via $Y = \alpha + \beta X + \epsilon$ where $\alpha, \beta \in \mathbb{R}$ are regression coefficients and ϵ is noise with mean 0. Then $\mathbb{E}[Y|X] = \alpha + \beta X$, so if X and Y are centred then $\mathbb{E}[Y|X] = \rho X$ where ρ is the correlation coefficient between X and Y. Since $|\rho| \leq 1$ we know that the expected value of Y is closer to the mean than X unless $\rho = 1$. So extreme values of X tend to correspond to values of Y that are closer to the mean, i.e. regression towards the mean is observed.

2.2 Logistic Regression

Logistic regression is to binary classification what linear regression is to regression. That is, both are the first method learned when introduced to regression and binary classification. The name might seem strange at first as regression models predict continuously distributed things and binary classification models predict either 0 or 1. The reason regression appears in its name is that it classifies samples by transforming them to $(0, 1)$ and imposing a threshold-based decision rule to yield 0 or 1.

Before detailing precisely how a logistic regression model classifies samples, how might one classify a sample at all? A natural idea is to fit a hyperplane to feature space which separates samples of the two classes. A hyperplane is an $(n-1)$-dimensional plane-like object embedded in n-dimensional space. For example, a hyperplane in $\mathbb{R}^2$ is a line and in $\mathbb{R}^3$ it is a plane. As an object, a hyperplane in $(q+1)-$dimensional space is the set of points $(x_1, \ldots, x_q) \in \mathbb{R}^q$ which satisfy

$$\theta_0 + \theta_1 x_1 + \cdots + \theta_q x_q = 0$$

where $\theta = (\theta_0, \ldots, \theta_q) \in \mathbb{R}^{q+1}$ are parameters which characterise the hyperplane. For brevity, we denote the function whose set of roots is the hyperplane by

$$\begin{aligned} f_\theta : \{1\} \times \mathbb{R}^q &\to \mathbb{R} \\ \mathbf{x} &\mapsto \theta_0 + \theta_1 x_1 + \cdots + \theta_q x_q =: \theta^\top \mathbf{x}. \end{aligned}$$

The purpose of the 1 in the first index of $\mathbf{x}$ is similar to its purpose in linear regression: it accounts for the elevation term θ_0 and makes notation a lot cleaner. After learning the parameters of a hyperplane, we may classify a sample $\mathbf{x}$ according to which side of the hyperplane $f_\theta(\mathbf{x}) = 0$ it lies. For example, samples 'below' the hyperplane, i.e. $f_\theta(\mathbf{x}) \leq 0$, could be classified as 0 and samples 'above', i.e. $f_\theta(\mathbf{x}) > 0$, could be classified as 1. That is, classify samples according to $C_\theta(\mathbf{x}) = \mathbb{1}(f_\theta(\mathbf{x}) > 0)$.

To obtain a notion of the probability that the class label of a sample is 1, consider how far it deviates from the plane. To make this idea concrete, apply a logistic (sigmoid) transformation to the output $f_\theta(\mathbf{x})$ as in

$$h_\theta(\mathbf{x}) = \sigma(f_\theta(\mathbf{x})) = \frac{1}{1 + \exp(-f_\theta(\mathbf{x}))} \in (0, 1).$$

From here, the class label of a sample $\mathbf{x}$ is 1, according to the model, if $h_\theta(\mathbf{x}) > \delta$ for some threshold $\delta \in (0, 1)$. Note that $\delta = 0.5$ corresponds precisely to the 'above/below the hyperplane' approach above. The advantage of this broad decision rule is that we can select δ in a way that improves precision at the expense of recall and vice versa. For example, for higher precision and lower recall, $\delta > 0.5$ and classify samples according to $C^\delta_\theta(\mathbf{x}) = \mathbb{1}(h_\theta(\mathbf{x}) > \delta)$.

See that for the model to match the underlying probability that the class label of a given sample is 1, we require that

$$\begin{aligned}
& h_\theta(\mathbf{x}) = p(Y = 1|\mathbf{X} = \mathbf{x}) \\
\iff & \frac{1}{1 + \exp(-f_\theta(\mathbf{x}))} = p(Y = 1|\mathbf{X} = \mathbf{x}) \\
\iff & f_\theta(\mathbf{x}) = \log\left(\frac{p(Y = 1|\mathbf{X} = \mathbf{x})}{1 - p(Y = 1|\mathbf{X} = \mathbf{x})}\right) \\
\iff & f_\theta(\mathbf{x}) = \log\left(\frac{p(Y = 1|\mathbf{X} = \mathbf{x})}{p(Y = 0|\mathbf{X} = \mathbf{x})}\right)
\end{aligned}$$

and so logistic regression models can be seen as fitting a linear regression model to the log-odds of each sample being of class 1.

While their motivation is intuitive, how might we learn suitable parameters of logistic regression models from data? Fortunately, as with linear regression, a statistically-grounded method exists.

2.2.1 Statistical Motivation

Like in linear regression, for a statistical motivation we'll make some assumptions regarding the class labels to derive a way of finding the optimal

parameters θ^*. Suppose that $Y|(\mathbf{X} = \mathbf{x}) \sim \text{Bernoulli}(h_\theta(\mathbf{x}))$. In this case

$$p(y|\mathbf{x};\theta) = (h_\theta(\mathbf{x}))^y(1 - h_\theta(\mathbf{x}))^{1-y}.$$

We construct the log-likelihood over $D_{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, assuming its samples are i.i.d., as

$$\begin{aligned}
l(\theta) &= \sum_{i=1}^n y \log(h_\theta(\mathbf{x}_i)) + (1 - y_i) \log(1 - h_\theta(\mathbf{x}_i)) \\
&= \sum_{i=1}^n \left[y_i \log\left(\frac{1}{1 + e^{-\theta^\top \mathbf{x}_i}}\right) + (1 - y_i) \log\left(1 - \frac{1}{1 + e^{-\theta^\top \mathbf{x}_i}}\right)\right] \\
&= \sum_{i=1}^n \left[y_i \log\left(\frac{1}{1 + e^{-\theta^\top \mathbf{x}_i}}\right) + (1 - y_i) \left(-\theta^\top \mathbf{x}_i + \log\left(\frac{1}{1 + e^{-\theta^\top \mathbf{x}_i}}\right)\right)\right] \\
&= \sum_{i=1}^n \left[- \log\left(1 + e^{-\theta^\top \mathbf{x}_i}\right) - (1 - y_i)\theta^\top \mathbf{x}_i\right]
\end{aligned}$$

which we maximise numerically in θ, e.g. via gradient descent, to obtain the optimal paramateres

$$\theta^* = \underset{\theta \in \mathbb{R}^{q+1}}{\arg\max} \left[\sum_{i=1}^n \left[- \log\left(1 + e^{-\theta^\top \mathbf{x}_i}\right) - (1 - y_i)\theta^\top \mathbf{x}_i\right]\right].$$

Note that, in practice, it's unlikely that any samples will lie on the hyperplane $f_{\theta^*}(\mathbf{x}) = 0$ itself.

2.3 Support Vector Machines (SVMs)

Logistic regression and SVMs both fit a hyperplane to feature space. The distinction is the metric of goodness of said hyperplanes. In logistic regression, the metric of goodness is how much the parameters of said hyperplane maximise the relevant log-likelihood. SVMs, however, look to maximise the distance between the hyperplane and the nearest samples. So in some sense, logistic regression is a probabilistic approach while SVMs take a raw constraint-based approach.

The authors' insight came from structural risk minimization in which instead of focusing on minimising training error (as neural networks and decision trees were doing at the time), one focuses on minimising an upper bound on the generalisation error. The inclusion of 'support vector' becomes clear from their construction but 'machine' stood out to me as a bit odd. It

turns out that at time of their development, around 1960, it was common to use 'machine' when referring to algorithms that learned from data, i.e. algorithms belonging to statistical learning theory. This reflects Arthur Samuel's coining of machine learning as a term in 1959 while working at IBM.

Inconsistent terminology surrounding SVMs

Some pieces of literature describe the decision boundary learned by an SVM as strictly linear, others allow for a non-linear decision boundary. It'd be nice if authors stuck to the former and explicitly stated 'non-linear SVM' when describing the latter. In this section, the term refers to those which correspond to linear decision boundaries. I'll state explicitly when considering non-linear SVMs.

2.3.1 Hard-margin SVMs

Since we aim to fit a hyperplane to feature space, i.e. $\mathbb{R}^{q+1}$, we use the same notation for the function $f_\theta(\mathbf{x})$ corresponding to the linear decision boundary (i.e. hyperplane) used to motivate logistic regression. To make notation a bit easier, let $\theta_+ = (\theta_1, \ldots, \theta_q)$ so that θ is the ordered concatentation of θ_0 and θ_+. Further, let $\mathbf{x}_i$ denote the i^{th} sample's features and $y_i \in \{-1, 1\}$ its class such that y_i is 1 if $\theta_+^\top \mathbf{x}_i + \theta_0 > 0$ and -1 otherwise.

Let $\mathbf{p}_i$ denote the projection of $\mathbf{x}_i$ onto the hyperplane and let d_i denote the distance between $\mathbf{p}_i$ and $\mathbf{x}_i$. Noting that the normal to the hyperplane is θ_+, we have $\mathbf{p}_i = \mathbf{x}_i - \frac{\theta_+}{||\theta_+||} d_i$ and so

$$\theta_+^\top \left(\mathbf{x}_i - \frac{\theta_+}{||\theta_+||} d_i \right) + \theta_0 = 0$$

which, after some rearranging, yields

$$d_i = \frac{\theta_+^\top \mathbf{x}_i + \theta_0}{||\theta_+||}.$$

Similarly, taking a point $\mathbf{x}_i$ whose class is -1 yields $d_i = -\frac{\theta_+^\top \mathbf{x}_i + \theta_0}{||\theta_+||}$ and so a neater way of writing this distance for an arbitrary sample $\mathbf{x}_i$ is $d_i = \frac{y_i(\theta_+^\top \mathbf{x}_i + \theta_0)}{||\theta_+||}$. Hard-margin SVMs are only applicable to linearly sparable data and their construction, from data, effectively boils down to finding which

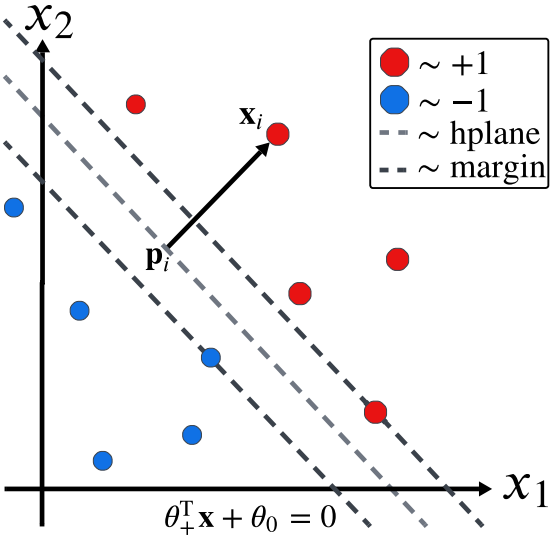


Figure 1: A hard-margin SVM in two dimensions. Samples lying on the margin are referred to as support vectors.

parameters maximise $\min_{i=1,\ldots,n} d_i$. That is, we seek to compute

$$\begin{aligned}\theta^* &= \arg\max_{(\theta_0,\theta_+)} \left[\min_{i=1,\ldots,n} d_i\right] \\ &= \arg\max_{(\theta_0,\theta_+)} \left[\min_{i=1,\ldots,n} \frac{y_i(\theta_+^\top \mathbf{x}_i + \theta_0)}{||\theta_+||}\right]\end{aligned}$$

which we'd like to translate into a convex optimisation problem. First, notice that computing θ^* is equivalent to solving

$$\max_{(\theta_0,\theta_+)} \frac{r}{||\theta_+||} \text{ s.t. } y_i\left(\theta_+^\top \mathbf{x}_i + \theta_0\right) \geq r \ (i = 1, \ldots, n).$$

in which r may be scaled arbitrarily by positives, so it is equivalent to

$$\max_{(\theta_0,\theta_+)} \frac{1}{||\theta_+||} \text{ s.t. } y_i\left(\theta_+^\top \mathbf{x}_i + \theta_0\right) \geq 1 \ (i = 1, \ldots, n).$$

This problem is still non-convex so we make a convenient switcheroo in realising that it is equivalent to solving

$$\min_{(\theta_0,\theta_+)} ||\theta_+||^2 \text{ s.t. } y_i\left(\theta_+^\top \mathbf{x}_i + \theta_0\right) \geq 1 \ (i = 1, \ldots, n)$$

which is solved in the usual convex problem solving ways.

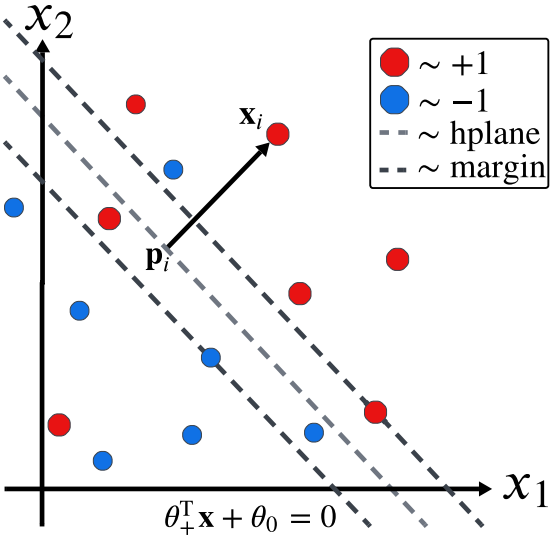


Figure 2: A soft-margin SVM in two dimensions. Samples are labelled according to their true class.

2.3.2 Soft-margin SVMs

Hard-margin SVMs are rarely applicable. In practice, samples are not entirely linearly separable and so allowing for some misclassification is pragmatic. Going from hard-margin to soft-margin is pretty straightforward, just include some slack variables $\xi = (\xi_1, \ldots, \xi_n)$ that ultimately allow the model to violate the constraints while penalising said violations. More precisely, it involves solving

$$\min_{(\theta_0, \theta_+, \xi)} \left[||\theta_+||^2 + \lambda \sum_{i=1}^{n} \xi_i \right] \text{ s.t. } y_i \left(\theta_+^\top \mathbf{x}_i + \theta_0 \right) \geq 1 - \xi_i, \ \xi_i \geq 0$$

for $i = 1, \ldots, n$ where $\lambda \geq 0$ is a regularisation parameter that influences the tradeoff of margin size and misclassification rate. Larger λ corresponds to prioritising a larger margin while smaller λ corresponds to prioritising the minimisation of misclassification.

As illustrated in Figure 2, it allows for samples to be closer to the decision boundary than the margin as well as outright misclassifications. Again, it is typically solved in the usual convex problem solving ways. Going a step further, it turns out that this can be reduced to computing

$$\arg\min_{(\theta_0, \theta_+)} \left[||\theta_+||^2 + \lambda \sum_{i=1}^{n} \max \left(0, 1 - y_i \left(\theta_+^\top \mathbf{x}_i + \theta_0 \right) \right) \right]$$

which can be done using gradient descent.

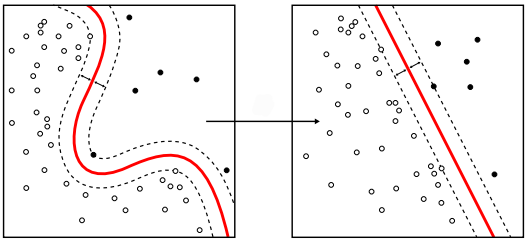


Figure 3: Non-linearly separable data being transformed into linearly separable data.

2.3.3 Non-linear SVMs

Motivating non-linear SVMs is straightforward: we'd sometimes like to separate samples by a non-linear decision boundary. With this in mind, a super intuitive approach is to find a transformation ϕ which maps samples to a space in which they are linearly separable. In said space, employ a linear SVM.

With this idea in mind, we seek to solve

$$\min_{(\theta_0,\theta_+,\xi)} \left[\frac{1}{2}||\theta_+||^2 + \lambda\sum_{i=1}^{n}\xi_i\right] \text{ s.t. } y_i\left(\phi(\theta_+)^\top \mathbf{x}_i + \theta_0\right) \geq 1-\xi_i,\ \xi_i \geq 0$$

which is difficult if the dimension of the image of ϕ is large. To make things easier, the dual of problem is optimised instead. My understanding of primal problems and their dual problems isn't great, so I'll just give the dual outright without deriving it:

$$\max_{(\alpha_1,\ldots,\alpha_n)\in[0,\lambda]^n} \left[\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j)\right] \text{ s.t. } \sum_{i=1}^{n}\alpha_i y_i = 0$$

where $K(\mathbf{x},\mathbf{z}) = \langle\phi(\mathbf{x}),\phi(\mathbf{z})\rangle$ is a pre-chosen kernel. See how it never requires an explicit computation involving ϕ as long as a closed form of $K(\mathbf{x},\mathbf{z})$ not involving ϕ is available. This is referred to as the kernel trick and reduces having to solve an optimisation problem in a space whose dimension is the number of features to a space whose dimension is the number of training samples.

At inference time, given some $\mathbf{x}$, we seek to compute

$$\begin{aligned} y &= \text{sign}\left(\theta_+^\top\phi(\mathbf{x}) + \theta_0\right) \\ &= \text{sign}\left(\sum_{i=1}^{n}\alpha_i y_i K(\mathbf{x}_i,\mathbf{x}) + b_k\right) \end{aligned}$$

where $b_k = y_k - \sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}_k)$ for some $(\mathbf{x}_k, y_k)$ that satisfies $\alpha_k \in (0, \lambda)$. Note that the second line in the equation above is derived using to argument made in deriving the dual.

The two simplest kernels are polynomial kernels and the radial basis function (RBF) kernel, the latter of which is far more popular. That said, polynomial kernels have had their place in applying SVMs to natural language processing tasks. The polynomial kernels

$$K(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^\top \mathbf{z} + c)^d$$

capture feature interactions up to degree d, which is useful when the separation of classes depends on combinations of input variables, e.g. quadratic boundaries in $\mathbb{R}^2$. Therein lies their weakness: polynomial kernels can be limited if said decision boundary is particuarly irregular. The RBF kernel

$$K(\mathbf{x}, \mathbf{z}) = \exp\left(-\frac{||\mathbf{x} - \mathbf{z}||^2}{2\sigma^2}\right)$$

is more flexible as a result of it measuring similarity based on distance, effectively producing localised bumps in feature space. From this, the model is able to form very non-linear decision boundaries. So, all in all, polynomial kernels provide interpretable higher-order interactions, while RBF kernels offer universal approximation.

2.4 Decision Trees and Random Forests

I find decision trees and random forests so boring that I'm not going to write about them and will instead point to Figure 4 which illustrates random forests well. This section mostly exists so that the list of subsections in this section form a well-rounded list of supervised learning methods.

Despite my disinterest in them, learning about decision trees encourages you to understand entropy which is good. There are other topics which encourage the same thing though, e.g. Variational Autoencoders, through the use of KL-divergences. Also, random forests are a nice introduction to ensemble methods which nicely demonstrate how to prevent overfitting by reducing variance — directly illustrating the importance of the bias-variance tradeoff! Oh, and bootstrapping. One day I'll write this subsection properly.

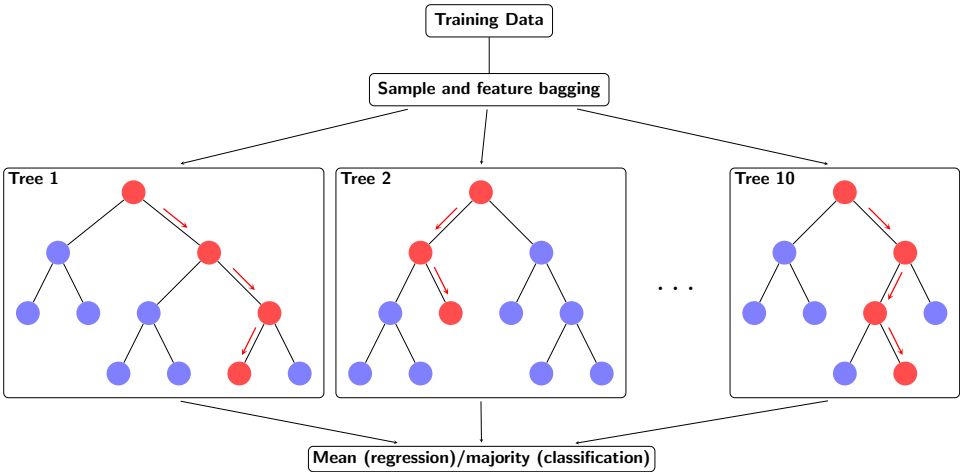


Figure 4: A random forest.

Logistic regression > SVM/RF when?

This is a pretty natural question once shown more sophisticated methods which deal with linearly separable and non-linearly separable data. The largest benefit of logistic regression is that it allows for statistical significance tests on parameters. It also helps that its implementation and interpretation are straightforward.

2.5 The Bias-Variance Tradeoff

The bias-variance tradeoff is a statement pertaining to the goodness of estimators (learning algorithms) according to mean square error (MSE) in terms of their variance — with respect to training data — and the square of their bias — due to architecture-related assumptions. It turns out that the optimal estimator, according to mean square error, is a careful tradeoff of both. Before deriving the tradeoff, let's consider the consequences of high squared bias and high variance: underfitting and overfitting.

2.5.1 Underfitting and Overfitting

Crudely put, if the architecture pertaining to an estimator is too simple to represent the point estimate then the estimator's bias (and thus its square) will be large in magnitude. Such an estimator underfits that which it is intended to model. I like to think of such an estimator as being almost

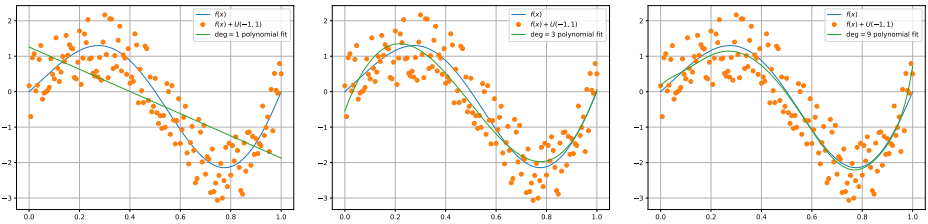


Figure 5: Lots of data, little noise.

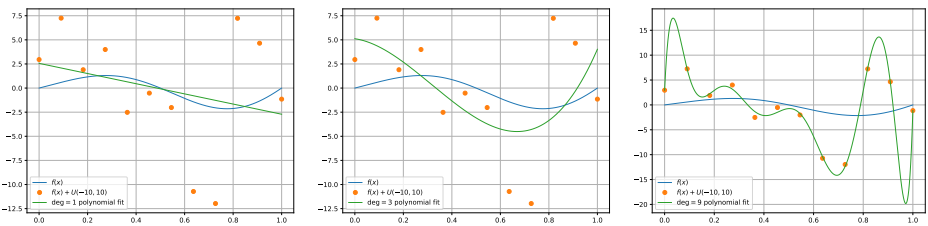


Figure 6: Little data, lots of noise.

invariant to the training data. As a practical example, think of fitting a line to a very non-linear function as in the leftmost plots of Figure 5 and Figure 6. Underfitting is reflected during training by both high training error and high validation error.

At the other extreme, if the architecture pertaining to an estimator is more complex than is necessary (e.g. overparameterised) to represent the point estimate then the excess parameters often[1] cause the model to fit the noise in the training data. In this case, the estimator is highly variant with respect to the training data. Such an estimator overfits the training data instead of fitting the point estimate. As a practical example, think of fitting a polynomial from small and heavily-noised training data, as in the rightmost plot of Figure 6. Overfitting is reflected during training by low training error and high validation error (relative to the training error).

2.5.2 Derivation with MSE loss

Recall that target values are often effected by noise, i.e. $Y|(\mathbf{X} = \mathbf{x}) = f(\mathbf{x}) + \epsilon(\mathbf{x})$ where $\epsilon(\mathbf{x}) \sim \mathcal{N}(0, \sigma^2(\mathbf{x}))$ and so

$$Y|(\mathbf{X} = \mathbf{x}) \sim \mathcal{N}\left(f(\mathbf{x}), \sigma^2(\mathbf{x})\right).$$

[1] See the double-descent phenomenon in Figure 8.

Given a fixed architecture, the estimate $\hat{f}_D$ of f is a purely function of the training dataset $D \subset (\Omega_{\mathbf{X}} \times \Omega_Y)^n$. Such a dataset is a realisation of the random variable $\mathcal{D}$ distributed according to $p(\mathbf{x}, y)^{\otimes n}$. In line with this, the bias and variance derived are that of the estimator $f_{\mathcal{D}}$.

As is common in regression contexts, the goodness of an estimate $\hat{f}_D$ is given by its expected risk

$$R\left(\hat{f}_D\right) = \mathbb{E}_{(\mathbf{X},Y)}\left[\left(Y - \hat{f}_D(\mathbf{X})\right)^2\right]$$

in which mean square error (MSE) is used as loss. Thus, the quality of the corresponding estimator (or learning algorithm) used to determine $\hat{f}_D$ from $D \in \Omega_{\mathcal{D}}$ is given by

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[R\left(\hat{f}_{\mathcal{D}}\right)\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{(\mathbf{X},Y)}\left[\left(Y - \hat{f}_{\mathcal{D}}(\mathbf{X})\right)^2\right]\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{(\mathbf{X},Y)}\left[\left(Y - f(\mathbf{X}) + f(\mathbf{X}) - \hat{f}_{\mathcal{D}}(\mathbf{X})\right)^2\right]\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{(\mathbf{X},Y)}\left[(Y - f(X))^2\right]\right] + \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{X}}\left[\left(f(\mathbf{X}) - \hat{f}_{\mathcal{D}}(\mathbf{X})\right)^2\right]\right] \\
&+ \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{(\mathbf{X},Y)}\left[2(Y - f(\mathbf{X}))\left(f(\mathbf{X}) - \hat{f}_{\mathcal{D}}(\mathbf{X})\right)\right]\right].
\end{aligned}$$

Let's address this term-by-term beginning with the term in red. See that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{(\mathbf{X},Y)}\left[2(Y - f(\mathbf{X}))\left(f(\mathbf{X}) - \hat{f}_{\mathcal{D}}(\mathbf{X})\right)\right]\right] \\
&= 2\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[(Y - f(\mathbf{X}))\left(f(\mathbf{X}) - \hat{f}_{\mathcal{D}}(\mathbf{X})\right)\right]\right]\right] \\
&= 2\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{X}}\left[\left(f(\mathbf{X}) - \hat{f}_{\mathcal{D}}(\mathbf{X})\right)\mathbb{E}_{Y|\mathbf{X}}\left[Y - f(\mathbf{X})\right]\right]\right] \\
&= 2\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{X}}\left[\left(f(\mathbf{X}) - \hat{f}_{\mathcal{D}}(\mathbf{X})\right)\cdot 0\right]\right] \\
&= 0
\end{aligned}$$

in which the tower property of conditional expectations

$$\begin{aligned}
\mathbb{E}_{(\mathbf{X},Y)}[g(\mathbf{X}, Y)] &= \iint g(\mathbf{x}, y) f_{(\mathbf{X},Y)}(\mathbf{x}, y) dx dy \\
&= \int \left(\int g(\mathbf{x}, y) f_{Y|\mathbf{X}}(y|\mathbf{x}) dy\right) f_{\mathbf{X}}(\mathbf{x}) dx \\
&= \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[g(\mathbf{X}, Y)\right]\right]
\end{aligned}$$

is applied. As such, the quantity of interest reduces to only two terms as in

$$\mathbb{E}_{\mathcal{D}}\left[R\left(\hat{f}_{\mathcal{D}}\right)\right] = \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{(\mathbf{X},Y)}\left[(Y - f(\mathbf{X}))^2\right]\right] + \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{X}}\left[\left(f(\mathbf{X}) - \hat{f}_{\mathcal{D}}(\mathbf{X})\right)^2\right]\right]$$

of which we will now address the first in orange. Noting that

$$\mathbb{E}_{Y|(\mathbf{X}=\mathbf{x})}\left[Y - f(\mathbf{x})\right] = 0$$

for all $\mathbf{x} \in \Omega_{\mathbf{X}}$ and that what is inside the expectation over $\mathcal{D}$ is independent of $\mathcal{D}$, we see that the first term reduces to

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{(\mathbf{X},Y)}\left[(Y - f(\mathbf{X}))^2\right]\right] &= \mathbb{E}_{(\mathbf{X},Y)}\left[(Y - f(\mathbf{X}))^2\right] \\
&= \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[(Y - f(\mathbf{X}))^2\right]\right] \\
&= \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{Y|\mathbf{X}}\left[\left(Y - \mathbb{E}_{Y|\mathbf{X}}[Y]\right)^2\right]\right] \\
&= \mathbb{E}_{\mathbf{X}}\left[\mathrm{Var}\left(Y|\mathbf{X}\right)\right] \\
&= \mathbb{E}_{\mathbf{X}}[\sigma^2(\mathbf{X})]
\end{aligned}$$

This is simply the expected noise, e.g. due to imperfect calibration in the instruments used to obtain samples.

To address the term in green, let $\bar{f}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}}\left[\hat{f}_{\mathcal{D}}(\mathbf{x})\right]$ for all $\mathbf{x} \in \Omega_{\mathbf{X}}$ and see that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{X}}\left[\left(f(\mathbf{X}) - \hat{f}_{\mathcal{D}}(\mathbf{X})\right)^2\right]\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{X}}\left[\left(f(\mathbf{X}) - \bar{f}(\mathbf{X}) + \bar{f}(\mathbf{X}) - \hat{f}_{\mathcal{D}}(\mathbf{X})\right)^2\right]\right] \\
&= \mathbb{E}_{\mathbf{X}}\left[\left(f(\mathbf{X}) - \bar{f}(\mathbf{X})\right)^2\right] + \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{X}}\left[\left(\bar{f}(\mathbf{X}) - \hat{f}_{\mathcal{D}}(\mathbf{X})\right)^2\right]\right] \\
&+ 2\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{X}}\left[\left(f(\mathbf{X}) - \bar{f}(\mathbf{X})\right)\left(\bar{f}(\mathbf{X}) - \hat{f}_{\mathcal{D}}(\mathbf{X})\right)\right]\right].
\end{aligned}$$

See that the final term vanishes as $\mathbb{E}_{\mathcal{D}}\left[\bar{f}(\mathbf{x}) - \hat{f}_{\mathcal{D}}(\mathbf{x})\right] = 0$ for all $\mathbf{x} \in \Omega_{\mathbf{X}}$

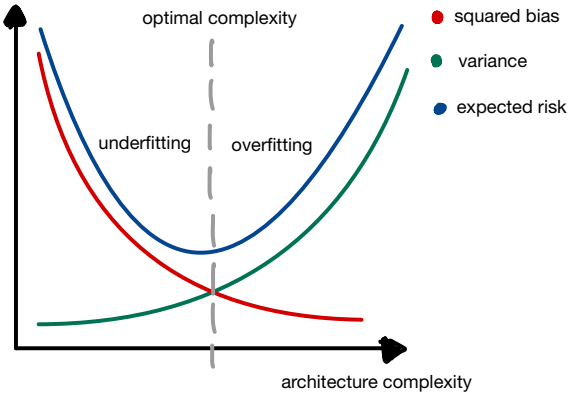


Figure 7: Expected risk, squared bias and variance against architecture complexity.

and so

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{X}}\left[\left(f(X)-\hat{f}_{\mathcal{D}}(\mathbf{X})\right)^2\right]\right] \\
&= \mathbb{E}_{\mathbf{X}}\left[\left(f(\mathbf{X})-\bar{f}(\mathbf{X})\right)^2\right] + \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{X}}\left[\left(\bar{f}(\mathbf{X})-\hat{f}_{\mathcal{D}}(\mathbf{X})\right)^2\right]\right] \\
&= \mathbb{E}_{\mathbf{X}}\left[\left(f(\mathbf{X})-\bar{f}(\mathbf{X})\right)^2 + \mathbb{E}_{\mathcal{D}}\left[\left(\bar{f}(\mathbf{X})-\hat{f}_{\mathcal{D}}(\mathbf{X})\right)^2\right]\right] \\
&= \mathbb{E}_{\mathbf{X}}\left[\left(f(\mathbf{X})-\mathbb{E}_{\mathcal{D}}\left[\hat{f}_{\mathcal{D}}(X)\right]\right)^2 + \mathbb{E}_{\mathcal{D}}\left[\left(\hat{f}_{\mathcal{D}}(\mathbf{X})-\mathbb{E}_{\mathcal{D}}\left[\hat{f}_{\mathcal{D}}(\mathbf{X})\right]\right)^2\right]\right].
\end{aligned}$$

Put together, we obtain

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[R\left(\hat{f}_{\mathcal{D}}\right)\right] \\
&= \mathbb{E}_{\mathbf{X}}\left[\left(f(\mathbf{X})-\mathbb{E}_{\mathcal{D}}\left[\hat{f}_{\mathcal{D}}(\mathbf{X})\right]\right)^2 + \mathbb{E}_{\mathcal{D}}\left[\left(\hat{f}_{\mathcal{D}}(\mathbf{X})-\mathbb{E}_{\mathcal{D}}\left[\hat{f}_{\mathcal{D}}(\mathbf{X})\right]\right)^2\right]\right] + \sigma^2 \\
&= \mathbb{E}_{\mathbf{X}}\left[\mathrm{Bias}\left(\hat{f}_{\mathcal{D}}(\mathbf{X})\right)^2 + \mathrm{Var}\left(\hat{f}_{\mathcal{D}}(\mathbf{X})\right)\right] + \sigma^2
\end{aligned}$$

where $\sigma^2 = \mathbb{E}_{\mathbf{X}}\left[\sigma^2(\mathbf{X})\right]$ (ignore the poor choice of notation). With this in mind, it's clear that we do not seek the estimator $\hat{f}_{\mathcal{D}}$ which minimises expected square bias or expected variance individually. Instead, we seek the estimator which hits the sweet spot in minimising their sum. An illustration of this statement, as the estimator's architecture is made more complex, is given in Figure 7.

Beyond MSE

Natural question: MSE is a common choice of loss when dealing with regression but not classification, in which cross-entropies are typically used as loss, so why do derivations of the bias-variance tradeoff use MSE? The quick answer is that it's the only choice which yields a relatively simple derivation of this elegant decomposition of estimator error into the square of its bias and its variance.

I've read online that some other choices of loss also yield decompositions into squared bias and variance but that their derivation is less elegant. There seems to be a good amount of work in this area but it's too in-depth for this document.

2.5.3 Double descent

Double descent refers to a behaviour observed in select set ups in which overparameterised models seemingly fly in the face of the bias-variance tradeoff. As the number of parameters (i.e. the model complexity) increases, the test error increases in line with the usual bias-variance tradeoff until reaching the number of training samples n. At this point, the test error spikes and training error hits 0. That is, the model is able to entirely interpolate the training data. The phenomenon begins as the number of parameters exceeds n. Not only does training error remain 0 (expected) but the test error begins to decrease and, in many set ups, reduces to an amount below the minimum test error observed during the traditional bias-variance region.

The phenomenon is illustrated nicely in Figure 8 and motivates the massively overparameterised contemporary models used these days. AlexNet is an example of an overparameterised model lying in the second descent region with around 60 million parameters but only 1.2 million training samples.

I like to imagine fitting a third degree polynomial by polynomials of higher and higher degree. If double descent were observed in this set up, one might expect wild polynomial fits until the polynomial's degree matches the number of training samples. At which point, the model has interoplated the training data perfectly and begins to regulate how it fits what's inbetween.

There has been a good amount of work investigating the phenomenon and some of the mystery has reduced as a result. The most convincing hand-wavy explanation of the behaviour is that parameter exploration via gradient descent is inherently regulatory. Not only that but the use of explcit regularisation, like SGD and Adam, on top of the self-regularisation bias toward parameters which offer 'simpler' representations.

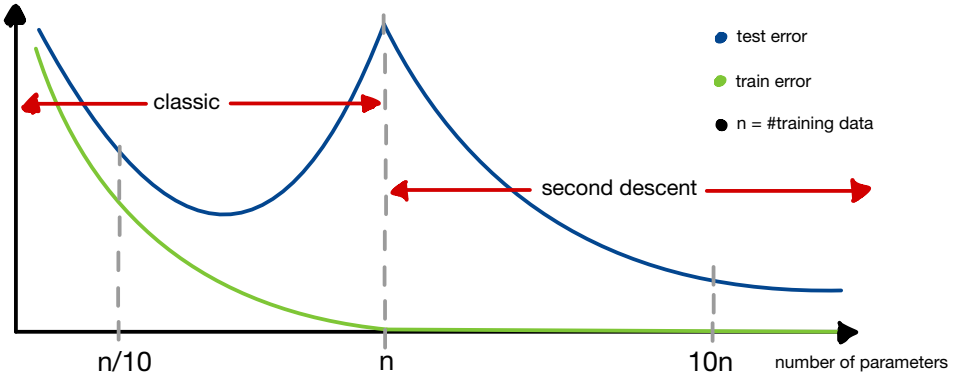


Figure 8: The double descent phenomenon, $n = |D_{\text{train}}|$.

3 Parameter Exploration and its Optimisation

In fitting a parametric model $f_\theta : \Omega_{\mathbf{X}} \to \Omega_Y$, the loss induced by a choice of model parameters θ for a sample $(\mathbf{x}, y)$ is given by $\mathcal{L}\left(y, f_\theta(\mathbf{x})\right)$ for some loss function $\mathcal{L} : \Omega_Y \times \Omega_Y \to \mathbb{R}_{\geq 0}$. In line with this, let the empirical risk R_D over a given dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be given by

$$\begin{aligned} R_D : \Theta &\to \mathbb{R}_{\geq 0} \\ \theta &\mapsto \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, f_\theta(\mathbf{x}_i)) \end{aligned}$$

In fitting f_θ, we seek to compute

$$\theta^* = \arg\min_{\theta \in \Theta} R_D(\theta),$$

i.e. to minimise the empirical risk over the dataset in the parameters. As such, methods of function minimisation are of interest.

Many function minimisation methods exist: Newton's method, genetic algorithms, gradient descent, etc. but most are horribly slow/unstable in high dimensions; not gradient descent though! As a result, gradient descent is *the* method for computing good parameters numerically in deep learning, hence the inclusion of this section.

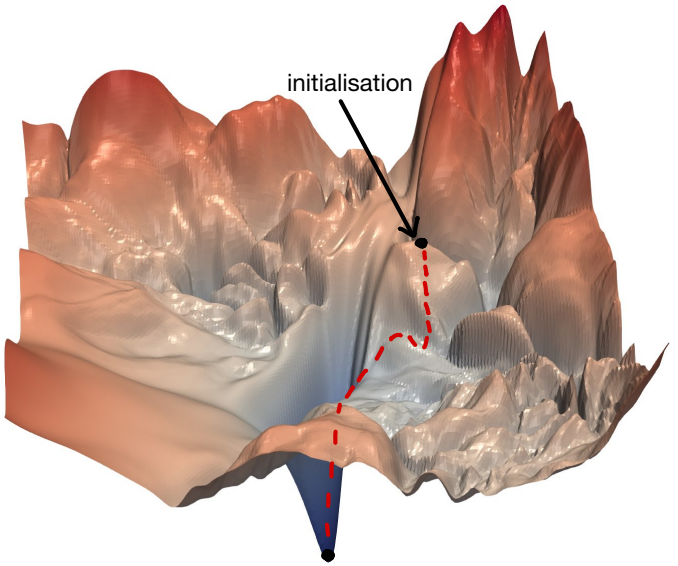


Figure 9: The graph $\{(\theta, R_D(\theta))|\theta \in \mathbb{R}^2\}$ of a complex loss landscape and a visualisation of how me might hope gradient descent looks.

Illustrations of loss landscapes are dataset-dependent!

When loss landscapes are illustrated, they are really plots of the graph $\{(\theta, R_D(\theta))|\theta \in \Theta\}$. That is, they are shaped by the fixed dataset D. As such, computations of the form $R_{D'}(\theta)$ for $D' \subsetneq D$, e.g. computations related to mini-batch or stochastic gradient descent, don't necessarily coincide with $R_D(\theta)$.

3.1 Gradient Descent

Since gradient descent is a general function minimisation method and not specific to machine learning, I'll use more general mathsy notation than θ, R_D, etc.

Say we're interested in finding minima of a function f whose co-domain is $\mathbb{R}_{\geq 0}$. A natural idea is to start with a guess $\mathbf{x}^{(0)}$ and send it in the direction in which f most decsends locally from $\mathbf{x}^{(0)}$, or equivalently the opposite of the direction in which f most ascends locally from $\mathbf{x}^{(0)}$, i.e. computing

$$\underset{||\mathbf{v}^{(0)}||=1}{\arg\max} f\left(\mathbf{x}^{(0)} + \eta \mathbf{v}^{(0)}\right),$$

for some small $\eta > 0$, and updating our guess to $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \eta \mathbf{v}^{(0)}$. Iteratively applying this idea yields the usual gradient descent rule

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta \mathbf{v}^{(t)}.$$

Eventually, after taking a ton of these small steps, we hope to reach a minimum of f.

3.1.1 In which direction does f most ascend locally from $\mathbf{x}^{(t)}$?

In the following physicist-like argument, we consider only unit vectors $\mathbf{v}$. This is because we care only about the direction of the vector in question and an equation later on simplifies quite nicely due to its unit length.

Multivariate Taylor expansion

Recall that if $f : \mathbb{R}^q \to \mathbb{R}$ then its second order Taylor expansion about $\mathbf{a} \in \mathbb{R}^q$ is given by

$$f(\mathbf{x}) = f(\mathbf{a}) + \nabla f(\mathbf{a}) \cdot (\mathbf{x} - \mathbf{a}) + \frac{1}{2}(\mathbf{x} - \mathbf{a}) \cdot \nabla^2 f(\mathbf{a})(\mathbf{x} - \mathbf{a}).$$

In gradient ascent, we are looking for the unit vector $\mathbf{v}$ whose direction f increases most locally from $\mathbf{x}$, i.e. we seek

$$\arg\max_{||\mathbf{v}||=1} \Big[f(\mathbf{x} + \eta \mathbf{v}) - f(\mathbf{x}) \Big]$$

where $\eta > 0$ is small. Approximating $f(\mathbf{x} + \eta \mathbf{v})$ using its Taylor expansion about $\mathbf{x}$ yields

$$f(\mathbf{x} + \eta \mathbf{v}) - f(\mathbf{x}) \approx \nabla f(\mathbf{x}) \cdot \eta \mathbf{v}.$$

Thus, the problem translates to finding the unit vector $\mathbf{v}$ that maximises

$$\nabla f(\mathbf{x}) \cdot \eta \mathbf{v} = ||\nabla f(\mathbf{x})|| \cdot ||\eta \mathbf{v}|| \cdot \cos(\theta) = \eta ||\nabla f(\mathbf{x})|| \cdot \cos(\theta)$$

where $\theta \in [0, \pi)$ denotes the angle between $\nabla f(\mathbf{x})$ and $\mathbf{v}$. This expression is maximised when $\cos(\theta) = 1$, i.e. when $\theta = 0$, which necessitates $\mathbf{v}$ having the same direction as $\nabla f(\mathbf{x})$. So $\mathbf{v}$ is a unit vector in the direction of $\nabla f(\mathbf{x})$, i.e. $\mathbf{v} = \frac{\nabla f(\mathbf{x})}{||\nabla f(\mathbf{x})||}$. To see that f ascends after being sent in the direction of $\mathbf{v} = \frac{\nabla f(\mathbf{x})}{||\nabla f(\mathbf{x})||}$ from $\mathbf{x}$, note that

$$f(\mathbf{x} + \eta \mathbf{v}) - f(\mathbf{x}) \approx \nabla f(\mathbf{x}) \cdot \eta \mathbf{v} = \eta \frac{\nabla f(\mathbf{x}) \cdot \nabla f(\mathbf{x})}{||\nabla f(\mathbf{x})||} = \eta ||\nabla f(\mathbf{x})|| > 0.$$

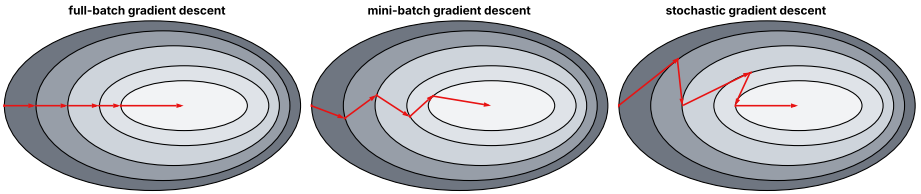


Figure 10: How the convergence of each method might look.

The edge case in which $\nabla f(\mathbf{x}) = 0$ corresponds to $f(\mathbf{x})$ being a local maximum.

3.1.2 Sensitivity to $\mathbf{x}^{(0)}$

I don't know much about initialisation. I know that it's a bad idea to start at 0 (if certain symmetries hold then parameters may change identically), points of large magnitude (exploding gradients) and points of small magnitude (vanshing gradients). Simon Price speaks of some interesting quirks of parameter initialisation around 01:08:00 in a YouTube interview.

One day when my interest in initialisation is sparked, I'll write this part properly.

3.1.3 Batch learning

If your training dataset is huge and you don't have enough VRAM to store gradient information then do not fear! It turns out that we can approximate full gradient updates reasonably well using many smaller updates over subsets of the dataset. How well mini-batch learning and stochastic gradient descent perform (and even just *that* they perform) surprised me when I first learned about them.

Epoch terminology

During training, the model tweaks its parameters based on training data. In practice, the model does so for the same sample many times. Each full cycle of the model having learned from the training data is called an epoch.

1) Full-batch

Full-batch gradient descent involves one parameter update per epoch as in

$$\theta^{(t+1)} = \theta^{(t)} - \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \mathcal{L}\left(y_i, f_{\theta^{(t)}}(\mathbf{x}_i)\right).$$

2) Mini-batch

The information stored in order to compute gradients during training might be too much for your PC's VRAM. If so, instead perform mini-batch gradient descent in which the dataset D is partitioned into m disjoint subsets $B_1, \ldots, B_m$, referred to as batches, and m parameter updates are made according to

$$\theta^{(t+j)} = \theta^{(t+j-1)} - \frac{1}{|B_j|}\sum_{(\mathbf{x},y)\in B_j} \nabla_\theta \mathcal{L}\left(y, f_{\theta^{(t+j-1)}}(\mathbf{x})\right).$$

We see that if we split into m batches then a full epoch during training corresponds to m parameter updates (or optimisation steps).

3) Stochastic gradient descent (SGD)

It turns out that updating based on a single sample $(\mathbf{x}, y)$ (uniformly randomly sampled), as in

$$\theta^{(t+1)} = \theta^{(t)} - \nabla_\theta \mathcal{L}\left(y, f_{\theta^{(t)}}(\mathbf{x})\right)$$

is viable. This surprised me a ton when I first read about it: I would have thought that it'd produce nonsense parameters. Unsurprisingly, SGD yields relatively unstable convergence, a bit like the steps that a drunk man would take in walking down a hill.

Terminology warning

What we call mini-batch gradient here is referred to by some as stochastic gradient descent. Confusing.

3.1.4 Batch + Layer normalisation

I've only seen these forms of normalisation apply to training neural networks. I'm unaware of them being grounded in theory. They're very widley used so I've included them.

1) Batch norm

For a batch $B = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$, batch normalisation is the act of normalising the post-activation values of some layer of d neurons in a neural network. Denote the d activation values for the sample $\mathbf{x}_i \in B$ as $(a_{i,1}, \dots, a_{i,d})$. The entire batch's activation values for said layer can be written as a matrix as in

$$\begin{bmatrix} a_{1,1} & \dots & a_{1,d} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \dots & a_{m,d} \end{bmatrix}.$$

Batch norm first normalises the columns of A as in

$$\begin{bmatrix} \hat{a}_{1,1} & \dots & \hat{a}_{1,d} \\ \vdots & \ddots & \vdots \\ \hat{a}_{m,1} & \dots & \hat{a}_{m,d} \end{bmatrix} = \begin{bmatrix} (a_{1,1} - \mu_1)/\sqrt{\sigma_1^2 + \epsilon} & \dots & (a_{1,d} - \mu_d)/\sqrt{\sigma_d^2 + \epsilon} \\ \vdots & \ddots & \vdots \\ (a_{m,1} - \mu_1)/\sqrt{\sigma_1^2 + \epsilon} & \dots & (a_{m,d} - \mu_d)/\sqrt{\sigma_d^2 + \epsilon} \end{bmatrix}$$

where $\mu_j = \frac{1}{m}\sum_{i=1}^m a_{i,j}$, $\sigma_j^2 = \frac{1}{m}\sum_{i=1}^m (a_{i,j} - \mu_j)^2$ and $\epsilon > 0$ is small, appearing for the sake of computational stability. At this point, the columns have mean 0 and variance 1. The output of batch norm for sample $i \in \{1, \dots, m\}$ is given by

$$(\gamma_1 \hat{a}_{i,1} + \beta_1, \dots, \gamma_d \hat{a}_{i,d} + \beta_d)$$

where $\gamma_1, \dots, \gamma_d$ and $\beta_1, \dots, \beta_d$ are learnable parameters.

I don't think that the effect of batch norm is immediately clear and the informal explanations given in the paper which introduced it have been largely discarded since. Perhaps it helps avoid exploding/vanishing gradients, I honestly don't know lol.

2) Layer norm

If batch statistics aren't particularly meaningful or batch size is small then layer norm may be preferable. Layer norm applies to individual samples, as opposed to entire batches, and so normalises over the features instead. Given activation values $a_1, \dots, a_d$, layer norm first normalises the activation values as in

$$(\hat{a}_1, \dots, \hat{a}_d) = \left(\frac{a_1 - \mu}{\sqrt{\sigma^2 + \epsilon}}, \dots, \frac{a_d - \mu}{\sqrt{\sigma^2 + \epsilon}} \right)$$

where $\mu = \frac{1}{d}\sum_{j=1}^d a_j$ and $\sigma^2 = \frac{1}{d}\sum_{j=1}^d (a_j - \mu)^2$. Then, as in batch norm, we scale and shift to obtain

$$(\gamma_1 \hat{a}_1 + \beta_1, \dots, \gamma_d \hat{a}_d + \beta_d)$$

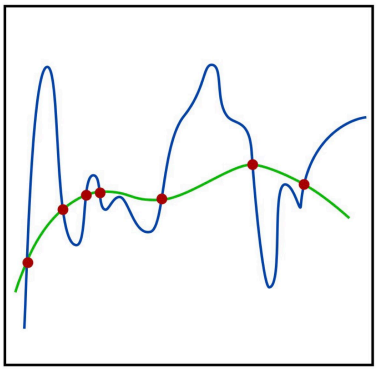


Figure 11: Motivating regularisation. Two fits, both with 0 test error.

where $\gamma_1, \ldots, \gamma_d$ and $\beta_1, \ldots, \beta_d$ are learnable parameters. As with batch norm, I honestly don't know the influence layer norm has on training.

3.2 Regularisation

Overfitting is a pain. How might we regulate the learning procedure in a way that (hopefully) prevents overfitting? Regularisation!

The most well known methods of regularisation are L1 and L2 regularisation. Both involve adding a penalty term to the loss function used during training with the intention of encouraging some behaviour in the learning process. There are tons of ways of motivating both L1 and L2 but the most intuitive to me (by a long shot) is a Bayesian approach: assume independence of parameters and impose a prior distribution on them according to whatever bias we hope to bake into the learning process. Then, instead of maximum likelihood estimation, compute

$$\begin{aligned}\arg\max_{\theta\in\mathbb{R}^{q+1}} \log(p(\theta|D)) &= \arg\max_{\theta\in\mathbb{R}^{q+1}} \left[\log(p(D|\theta)) + \log(p(\theta))\right] \\ &= \arg\min_{\theta\in\mathbb{R}^{q+1}} \left[-\sum_{i=1}^{n} \log(p(y_i|\mathbf{x}_i,\theta)) - \sum_{j=1}^{q} \log(p(\theta_j))\right].\end{aligned}$$

1) L1 regularisation (LASSO)

If we would like our to-be-learned parameters to be sparse, then it makes sense to choose a prior which has a lot of mass around 0 and tapers off rather harshly at the tails. A great choice for this is a Laplace prior with mean 0 and variance $2/\lambda$ which yields L1 regularisation. The corresponding density function is given by $p(\theta_j) = \lambda \exp(-2\lambda|\theta_j|)$ and explicit examples are illustrated in Figure 12.

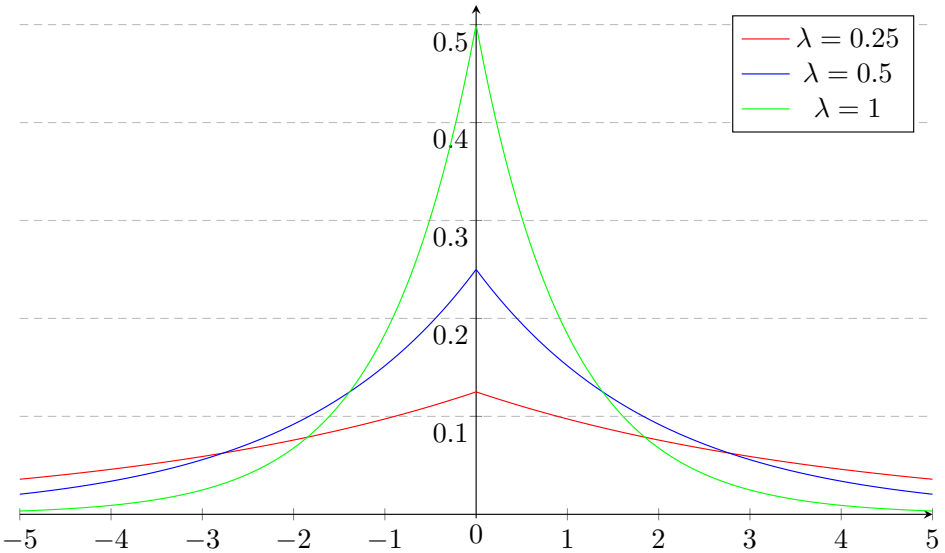


Figure 12: Laplace pdf with means 0 and decreasing variances according to $\lambda \in \{0.5, 1, 2, 3\}$.

See how as λ increases, the variance of the corresponding Laplace distribution decreases and more of the mass becomes centred around 0. The result is that the to-be-learned parameters are further encouraged to flatten out around 0. In line with this, assuming $\theta_1, \ldots, \theta_q \overset{\text{i.i.d.}}{\sim} \text{Lap}(0, 2/\lambda)$, the quantity we seek to compute is given by

$$\arg\max_{\theta \in \mathbb{R}^{q+1}} \log(p(\theta|D)) = \arg\min_{\theta \in \mathbb{R}^{q+1}} \left[-\sum_{i=1}^{n} \log(p(y_i|\mathbf{x}_i, \theta)) + \lambda \sum_{j=1}^{q} |\theta_j| \right].$$

2) L2 regularisation (Ridge)

If we would like no given parameter to be too large then imposing some sort of MSE penalty on the parameters makes sense. From the Bayesian point of view, this is achieved by imposing a normal prior on the parameters and yields L2 regularisation. That is, if $\theta_1, \ldots, \theta_q \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ then the quantity we seek is

$$\arg\max_{\theta \in \mathbb{R}^{q+1}} \log(p(\theta|D)) = \arg\min_{\theta \in \mathbb{R}^{q+1}} \left[-\sum_{i=1}^{n} \log(p(y_i|\mathbf{x}_i, \theta)) + \lambda \sum_{j=1}^{q} \theta_j^2 \right]$$

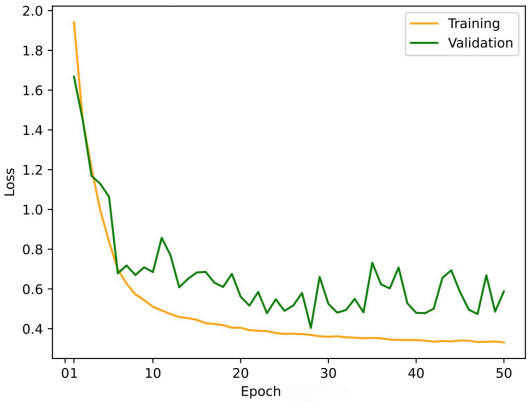


Figure 13: Training and validation losses over 50 epochs for some model. Lowest validation obtained at epoch 28.

where $\lambda = 1/2\sigma^2$.

There's a particularly elegant way to visualise the effects of L1 and L2 reglularisation in terms of how they change the shape of the corresponding loss landscape. It's as if each point in the landscape is sunk according to its distance from the origin and the hyperparameter λ. The sharpness or curvature of said sinking is prior-dependent: for L1 regularisation it is far sharper, hence the origin acts as more of a sinkhole than in L2 regularisation, inducing sparser minima.

3) Early stopping

To motivating early stopping, consider the familiar scenario of training via gradient descent. A natural thought when logging the training loss is whether or not a decrease in training loss from one epoch to the next genuinely pertains to a model that better generalises. How can we be confident that our model isn't simply overfitting the training data if all we see is decrease in training loss?

Early stopping partially alleviates this fear by splitting the training dataset into a smaller training set and a validation set (perhaps 80%/20%). After each epoch (or every few), compute both training and validation losses. Said validation loss acts as some measure of the model's ability to generalise. If training loss is decreasing while validation loss is not then the model may be overfitting. As illustration, consider Figure 13 in which the training loss continues to decrease in the number of epochs while the validation loss seems to obtain its lowest value around epoch 28.

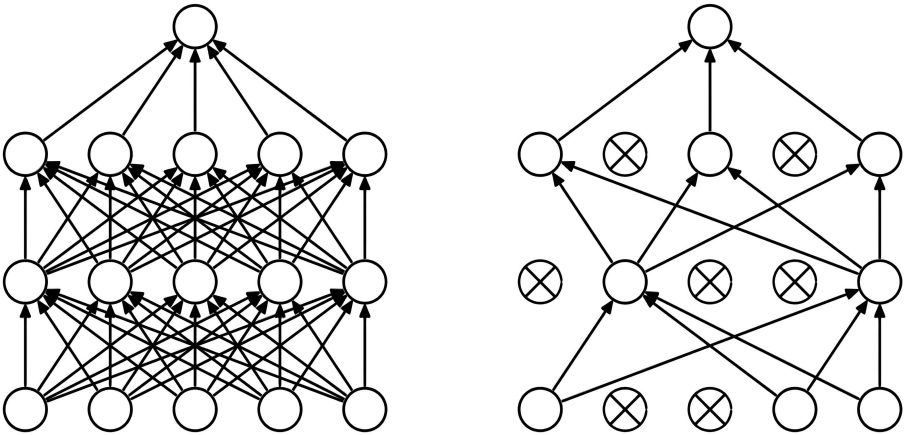


Figure 14: Dropout.

4) Dropout

Dropout is a form of regularisation specific to neural network architectures (I think). Its rough motivation is that we'd prefer not to be overly-reliant on any given subset of neurons at inference time. To prevent such an over-reliance, at the beginning of each epoch, independently set the activation of each neuron to 0 with some probability p. This way some portion of neurons are silenced during training for said epoch. As a result, its corresponding parameters are not updated during backpropagation. This idea is illustrated in Figure 14. Make sure to not apply dropout at ineference time!

3.3 Momentum + Adaptive Learning Rates

There are some intuitive ideas which aim to accelerate the gradient descent process. One is momentum, which increases steps made when gradients consistently pointed in the same direction over previous steps. Another is an adaptive learning rate, where each parameter has its own step size that adapts over time to its past gradients.

1) Momentum

Think of a ball rolling downhill: momentum lets it keep moving in the same direction which smoothes out noisy gradients. If gradients keep pointing

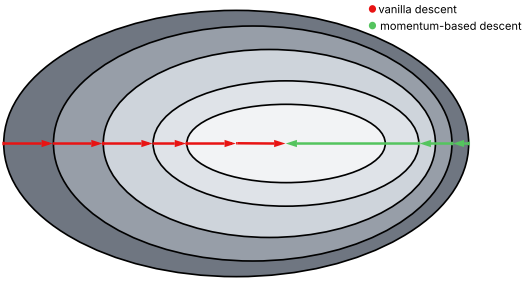


Figure 15: How we might hope convergence would look using momentum.

in the same direction then momentum builds. If gradients oscillate, e.g. traversing like a ravine, then the motion is dampened and momentum is lost.

We impose a momentum-like idea by introducing a velocity vector

$$\mathbf{v}^{(t+1)} = \mu\mathbf{v}^{(t)} - \eta\nabla R_D\left(\theta^{(t)}\right)$$

with $\mathbf{v}^{(0)} = 0$ and change the gradient descent update rule to

$$\theta^{(t+1)} = \theta^{(t)} - \eta\nabla R_D\left(\theta^{(t)}\right) + \mu\mathbf{v}^{(t)}$$

where $\mu \in [0, 1)$ is the momentum coefficient. Typical values include, $\mu \in \{0.9, 0.95, 0.99\}$. To get a sense of how this refelcts a momentum-like idea, see that

$$\mathbf{v}^{(1)} = -\eta\nabla R_D\left(\theta^{(0)}\right)$$

and so

$$\mathbf{v}^{(2)} = -\eta\left(\mu\nabla R_D\left(\theta^{(0)}\right) + \nabla R_D\left(\theta^{(1)}\right)\right).$$

If $\mu \approx 1$ and the gradients $\nabla R_D\left(\theta^{(0)}\right) \approx \nabla R_D\left(\theta^{(1)}\right)$ then $\mathbf{v}^{(2)}$ will confidently boost $\theta^{(3)}$ in the direction of the first two gradients. If $\nabla R_D\left(\theta^{(0)}\right) \approx -\nabla R_D\left(\theta^{(1)}\right)$ then, as intuition would suggest, $\mathbf{v}^{(2)} = 0$.

2) Adaptive learning rates

Why have a fixed learning rate? Why have the same learning rate for each parameter? Adaptive learning rates aim to alleviate the oversimplications these questions pertain to. For some intuition as to how learning rates might change: if up to the t^{th} optimisation step the gradients for a given parameter have been large then it has presumably converged decently so its

learning rate would ideally get smaller. AdaGrad does this is by storing a per parameter running total of gradients, as in

$$G_i^{(t)} = \sum_{\tau=1}^{t} \left(\nabla_{\theta_i} R_D(\theta_i^{(\tau)}) \right)^2$$

and switching the optimisation step to

$$\theta_i^{(t+1)} = \theta_i^{(t)} - \eta_i \nabla_{\theta_i}^{(t)} R_D(\theta_i^{(t)})$$

where $\eta_i^{(t)} = \frac{\eta}{\sqrt{G_i^{(t)} + \epsilon}}$ is the i^{th} parameter's learning rate at step t. There are more elaborate adaptive learning rate methods than this, e.g. RMSProp and Adam, but AdaGrad conveys the idea sufficiently well.

3.4 IMPROVE: Hyperparameter Tuning

Hyperparameters are the tunable parts of an architecture which are not learned during training. Bad hyperparameters will yield training issues like divergence, overfitting or underfitting. A simple example of a hyperparameter I can think of is the learning rate in vanilla gradient descent. Some choices of learning rate will result in better generalisation than other choices. For a more sophisticated example, consider architecture-related hyperparameters, e.g. the number of hidden layers in a multi-layer perceptron.

1) Random search

Begin by designating some portion of the training data for validation (maybe 80% training, 20% validation).

Each hyperparameter may take on a set of values, i.e. the m hyperparameters $h_1, \ldots, h_m$ induce the samples spaces $\Omega_{h_1}, \ldots, \Omega_{h_m}$ and the grid of all possible hyperparameter configurations is thus $\Omega_1 \times \cdots \times \Omega_m$. In a perfect world, you'd be able to train a model for each hyperparameter configuration and choose whichever yields the lowest validation loss. We don't live in a perfect world and the cardinality of said grid is often too large. In line with this, instead of traversing it entirely, uniformly randomly sample from it a number of times and train with said hyperparameter configuration. Choose whichever configuration yielded the lowest validation loss.

2) $k-$fold cross-validation

For a better idea of how a configuration of hyperparameters will help generalise post-training, consider partitioning the training data into k disjoint subsets (or folds) $D_1, \ldots, D_k$. Then, for $i = 1, \ldots, k$, train on $D \backslash D_i$ and validate on D_i. Take the mean of the k validation losses. Choose whichever hyperparameter configuration yields the lowest mean validation loss.

Benchmark overfitting

Consider the following scenario: research groups A and B independently publish results regarding almost-identical single hidden layer MLP architectures applied to the same task. The only distinction in their architectures is the number of neurons in their hidden layer. Group A decide on 99 neurons and group B decide on 100. Every aspect of their training is identical: their training data, how they split it in order to perform validation, etc. Group A's MLP obtains a lower test error than group B and so those producing commercial products that may make use of an MLP choose group A's architecture. How is this scenario distinct from a single group selecting a single hidden layer MLP architcture with 99 neurons over 100 on the basis of the test loss of both? More broadly, how is this distinct to tuning hyperparameter based on test losses?

The post-training test error becomes a biased (in this case optimistic) estimate of the generalisation of the model. That said, if this scenario were repeated, the community would be effectivevly tuning the model to the test set. This is why benchmark overfitting is talked about. It emphasises the need for benchmark evolution and the shuffling of data in the training-inference pipeline.

4 Neural Networks

It's often the case that someone's intended meaning when stating 'neural networks' is more precisely multi-layer perceptrons (MLPs), perhaps the simplest class of neural networks beyond single-layer perceptrons. Rigorously formulating MLPs is one of those things that's useful to do at least once; consistent notation is 90% of the effort.

4.1 Multi-Layer Perceptrons (MLPs)

Multi-layer perceptrons (MLPs) are fully-connected feed-forward networks consisting of an input layer (where data is input), hidden layers and an

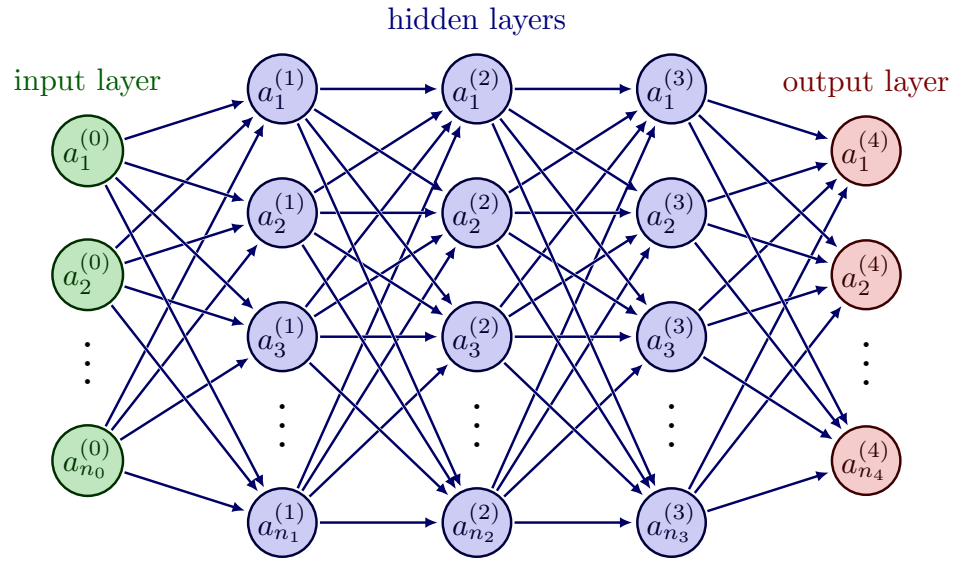


Figure 16: A multi-layer perceptron (MLP) with $k = 3$ hidden layers.

output layer. An MLP with three hidden layers is illustrated in Figure 16. Each layer is made up of a number of neurons, each of which has a real-valued activation value. For any neuron in a non-input layer, its activation value is the output of a non-linear activation function given, as input, the neuron's real-valued bias plus a weighted sum of the activation values of the neurons in its preceding layer. Regarding their inspiration from the human brain, consider this footnote[2].

The application of MLPs to learning functions from data is due to their expressivity, expressive-efficiency and tractability. Their expressivity is known due to a proof of their universal function approximation by Cybenko. To add to this, their high expressive-efficiency has been demonstrated empirically and ensures that the number of model components (hidden layers and neurons) needed to represent arbitrary functions is far lower than competing model classes. As for their tractability, MLPs allow for an evaluation of the approximation of the function of interest with complexity quadratic in the number of neurons of each layer of the MLP (after the heavy simplifcation of assuming a roughly equal number of neurons in each layer).

For brevity, given $m, n \in \mathbb{N}$ with $m \leq n$ let $[n]_m = \{m, \ldots, n\}$ and let $[n] = [n]_1$. For further ease of notation, given an MLP with k hidden layers,

[2] `https://stats.stackexchange.com/a/159172`

denote the index of the input layer by 0, the output layer by $k+1$ and by extension the j^{th} layer by $j \in [k+1]_0$. Additionally, consider the following denotations

- $n_j \in \mathbb{Z}_{\geq 1}$: the number of neurons in layer j.
- $a_i^{(j)} \in \mathbb{R}$: the activation value of neuron i in layer j.
- $w_{i,l}^{(j)} \in \mathbb{R}$: the weight associated with the edge to neuron i in layer $j \in [k+1]$ from neuron l in the previous layer.
- $b_i^{(j)} \in \mathbb{R}$: the bias of neuron i in layer $j \in [k+1]$.
- $\sigma_j : \mathbb{R} \to \mathbb{R}$: the non-linear activation function of layer $j \in [k+1]$.

From here we can express the activation value of any neuron in a non-input layer in terms of the activation values of the neurons in the layer which precedes it as

$$\begin{aligned} a_i^{(j+1)} &= \sigma_{j+1}\left(\sum_{l=1}^{n_j} w_{i,l}^{(j+1)} a_l^{(j)} + b_i^{(j+1)}\right) \\ &= \sigma_{j+1}\left(\begin{bmatrix} w_{i,1}^{(j+1)} & \dots & w_{i,n_j}^{(j+1)} \end{bmatrix} \begin{bmatrix} a_1^{(j)} \\ \vdots \\ a_{n_j}^{(j)} \end{bmatrix} + b_i^{(j+1)}\right) \end{aligned}$$

for $j \in [k]_0$. The reason for writing the second equality above, involving the dot product of two vectors, is that it helps us to see how using matrix-vector notation allows us to write an elegant and compact expression for the activation values of all neurons in a non-input layer in terms of the activation values of the neurons belonging to its preceding layer as

$$\begin{bmatrix} a_1^{(j+1)} \\ \vdots \\ a_{n_{j+1}}^{(j+1)} \end{bmatrix} = \sigma_{j+1}\left(\begin{bmatrix} w_{1,1}^{(j+1)} & \dots & w_{1,n_j}^{(j+1)} \\ \vdots & \ddots & \vdots \\ w_{n_{j+1},1}^{(j+1)} & \dots & w_{n_{j+1},n_j}^{(j+1)} \end{bmatrix} \begin{bmatrix} a_1^{(j)} \\ \vdots \\ a_{n_j}^{(j)} \end{bmatrix} + \begin{bmatrix} b_1^{(j+1)} \\ \vdots \\ b_{n_{j+1}}^{(j+1)} \end{bmatrix}\right)$$

which we abbreviate to

$$\mathbf{a}^{(j+1)} = \sigma_{j+1}\left(\mathbf{W}^{(j+1)}\mathbf{a}^{(j)} + \mathbf{b}^{(j+1)}\right)$$

where the activation function σ_{j+1} is applied element-wise.

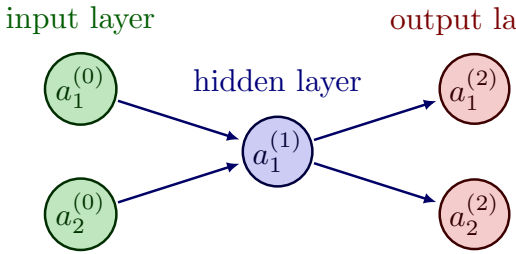


Figure 17: A neural network with one hidden layer ($k = 1$).

If we fix the structure and choice of activation functions of an MLP then all that is left to learn are its weights and biases. This is often done using gradient descent to minimise some loss function in which gradients are computed via back-propagation. Such a loss function can be a principled measure, e.g. corresponding to maximum likelihood, but this is not always necessary: ad-hoc loss functions are sometimes employed.

Advocating for depth over width

The complexity of a forward pass of an MLP can be expressed in terms of the number of neurons in each layer as

$$\mathcal{O}\left(\sum_{j=1}^{k+1} n_{j-1} \cdot n_j\right).$$

Very crudely supposing $n_0 = \cdots = n_{k+1} = n$ yields a complexity of

$$\mathcal{O}((k+1)n^2),$$

i.e. linear growth in the number of layers and quadratic growth in the number of neurons per layer. A super crude argument for increasing depth over width.

4.1.1 Example

Consider the neural network in Figure 17 whose input, hidden and output layers consist of two, one and two neurons respectively. With such a simple neural network, we may explicitly express the output neurons $a_1^{(2)}$ and $a_2^{(2)}$ in terms of the input neurons $a_1^{(0)}$ and $a_2^{(0)}$. We see that $n_0 = 2$, $n_1 = 1$ and $n_2 = 2$ and so $\mathbf{W}^{(1)} = \begin{bmatrix} w_{1,1}^{(1)} & w_{1,2}^{(1)} \end{bmatrix}$ and $\mathbf{W}^{(2)} = \begin{bmatrix} w_{1,1}^{(2)} \\ w_{2,1}^{(2)} \end{bmatrix}$. Noting additionally

that $\mathbf{b}^{(1)} = b_1^{(1)}$ and $\mathbf{b}^{(2)} = \begin{bmatrix} b_1^{(2)} \\ b_2^{(2)} \end{bmatrix}$ we can explicitly express the activation of the single neuron in the hidden layer as

$$\begin{aligned} \mathbf{a}^{(1)} &= \sigma_1\left(\mathbf{W}^{(1)}\mathbf{a}^{(0)} + \mathbf{b}^{(1)}\right) \\ &= \sigma_1\left(w_{1,1}^{(1)}a_1^{(0)} + w_{1,2}^{(1)}a_2^{(0)} + b_1^{(1)}\right) \end{aligned}$$

which is just the scalar value $a_1^{(1)}$. For the output layer, we have

$$\begin{aligned} \mathbf{a}^{(2)} &= \sigma_2\left(\mathbf{W}^{(2)}\mathbf{a}^{(1)} + \mathbf{b}^{(2)}\right) \\ &= \sigma_2\left(\begin{bmatrix} w_{1,1}^{(2)}a_1^{(1)} + b_1^{(2)} \\ w_{2,1}^{(2)}a_1^{(1)} + b_2^{(2)} \end{bmatrix}\right) \\ &= \sigma_2\left(\begin{bmatrix} w_{1,1}^{(2)}\sigma_1\left(w_{1,1}^{(1)}a_1^{(0)} + w_{1,2}^{(1)}a_2^{(0)} + b_1^{(1)}\right) + b_1^{(2)} \\ w_{2,1}^{(2)}\sigma_1\left(w_{1,1}^{(1)}a_1^{(0)} + w_{1,2}^{(1)}a_2^{(0)} + b_1^{(1)}\right) + b_2^{(2)} \end{bmatrix}\right) \\ &= \begin{bmatrix} \sigma_2\left(w_{1,1}^{(2)}\sigma_1\left(w_{1,1}^{(1)}a_1^{(0)} + w_{1,2}^{(1)}a_2^{(0)} + b_1^{(1)}\right) + b_1^{(2)}\right) \\ \sigma_2\left(w_{2,1}^{(2)}\sigma_1\left(w_{1,1}^{(1)}a_1^{(0)} + w_{1,2}^{(1)}a_2^{(0)} + b_1^{(1)}\right) + b_2^{(2)}\right) \end{bmatrix}. \end{aligned}$$

As such, the activations of the output neurons are given by

$$\begin{aligned} a_1^{(2)} &= \sigma_2\left(w_{1,1}^{(2)}\sigma_1\left(w_{1,1}^{(1)}a_1^{(0)} + w_{1,2}^{(1)}a_2^{(0)} + b_1^{(1)}\right) + b_1^{(2)}\right) \\ a_2^{(2)} &= \sigma_2\left(w_{2,1}^{(2)}\sigma_1\left(w_{1,1}^{(1)}a_1^{(0)} + w_{1,2}^{(1)}a_2^{(0)} + b_1^{(1)}\right) + b_2^{(2)}\right). \end{aligned}$$

Further advocating for depth over width

A perhaps more important result is how, geometrically, MLPs split the input space into a bunch of regions. It turns out that the number of regions by which a single-hidden layer MLP with two input neurons and n neurons in its hidden layer splits the space is bounded above by

$$\frac{1}{2}(n^2 + n + 2)$$

which is quadratic in n. For an MLP with k hidden layers of n neurons each, the number of regions by which the space is split is bounded above by

$$\frac{1}{2}(n^2 + n + 2)\left(\frac{n}{n_0} + 1\right)^{n_0(k-1)}$$

which is exponential in k. Again, we argue for increasing depth k over width n. A natural quesiton is to what extent these bounds are tight in practice.

4.1.2 Universal Function Approximation with MLPs

A class of functions, or function class, $\mathcal{F}_S$ is capable of universal function approximation on $S \subset \mathbb{R}^q$ compact (closed and bounded) if for all continuous $f : S \to \mathbb{R}$ and $\epsilon > 0$ there exists some $\hat{f} \in \mathcal{F}_S$ such that

$$\max_{x \in S} \left| f(x) - \hat{f}(x) \right| < \epsilon.$$

Note that such a function family can be used to approximate functions whose codomain is $\mathbb{R}^m$ by performing coordinate-wise approximation.

The family of polynomials is capable of universal function approximation as shown by Weierstrass in 1885. A directly equivalent statement to a function class satisfying the universal function approximation is the function ckass being dense in $C(S)$ with respect to the supremum norm. Here, $C(S)$ denotes the set of all continuous real-valued functions with compact domain $S \subset \mathbb{R}^q$. Quick note: usually the domain of the to-be-approximated function is denoted by K but I denote the number of layers in an MLP using a k so I'd like to avoid the confusion by using S instead. Instead of offering a rigorous proof, I'll motivate the overall idea in three steps starting with a simpler case: functions of the form $f : S \to \mathbb{R}$ where $S \subset \mathbb{N}$ with $\#S < \infty$.

Step 1: Motivating the idea for S finite

Without loss of generality, let $S = [\#S]$. We seek to show that for all $\epsilon > 0$ there exists an MLP with one hidden layer and one node in its output layer, whose output is denoted by $\hat{f}(x)$, which satisfies

$$\max_{x \in [\#S]} \left| f(x) - \hat{f}(x) \right| < \epsilon.$$

First note that for all $x \in [\#S]$,

$$\begin{aligned} f(x) &= \sum_{s=1}^{\#S} f(x) \mathbb{1}(x = s) \\ &= \sum_{s=1}^{\#S} f(x)(\mathbb{1}(x \geq s) - \mathbb{1}(x \geq s+1)) \end{aligned}$$

so we already see that a single neuron in the ouput layer of a single hidden layer MLP ($2 \cdot \#S$ neurons in hidden layer) with weights

$$\mathbf{W}^{(2)} = \begin{bmatrix} f(1) & -f(1) & \dots & f(\#S) & -f(\#S) \end{bmatrix},$$

i.e. $w_{1,s}^{(2)} = (-1)^{s+1} f\left(\left\lfloor \frac{s+1}{2} \right\rfloor\right)$, and bias $\mathbf{b}^{(2)} = 0$ is sufficient if the $2 \cdot \#S$ neurons in the hidden layer approximate

$$\mathbb{1}(x \geq 1), \mathbb{1}(x \geq 2), \mathbb{1}(x \geq 2), \mathbb{1}(x \geq 3), \dots, \mathbb{1}(x \geq \#S), \mathbb{1}(x \geq \#S + 1)$$

with the absolute error of each approximation bounded by ϵ. Note that the final indicator $\mathbb{1}(x \geq \#S + 1)$ is redundant so you only really need $2 \cdot \#S - 1$ neurons but I'll keep it for the nicer number. Also note that these are just renditions of the Heaviside function but I prefer sticking with the indicator function notation. This exact idea can be realised elegantly using the sigmoid as activation for the hidden layer. To see this, note that

$$\begin{aligned} \sigma\left(10^{\#S}(x-n) + 10^{\#S-9}\right) &= \frac{1}{1 + \exp\left(-10^{\#S}(x-n) + 10^{\#S-9}\right)} \\ &= \frac{1}{1 + \exp\left(n - x + 10^{-9}\right)^{10^{\#S}}} \end{aligned}$$

approximates the Heaviside function $H(x-n)$ with some small leakage. To hit certain levels of precision, i.e. attaining the desired ϵ bound, simply decrease the exponent in which -9 currently lies. You can probably express

the exponent in terms of ϵ explicitly (maybe ϵ^{-1}) but the idea is the point here. So what we want is for the $\left\lceil \frac{s+1}{2} \right\rceil^{\text{th}}$ neuron in the hidden layer to approximate $H\left(x - \left\lceil \frac{s+1}{2} \right\rceil\right)$ for $s \in [\#S]$. This is straightforward using a sigmoid activation function in the hidden layer given the approximation above. Simply set the weights and biases accordingly: $w_{s,1}^{(1)} = 10^{\#S}$ and $b_s^{(1)} = -10^{\#S}\left(\left\lceil \frac{s+1}{2} \right\rceil - 10^{-9}\right)$ and we're done.

Step 2: Extending the idea to $S \subset \mathbb{R}$ compact

If $f \in C(S)$ for $S \subset \mathbb{R}$ compact then f is uniformly continuous and so for all $\epsilon > 0$ there exists $\delta > 0$ such that

$$|x - y| < \delta \implies |f(x) - f(y)| < \epsilon.$$

Note that if we have a Lipschitz constant L for a given f then $\delta = \epsilon/L$ suffices. The reason this $\epsilon - \delta$ statement is useful is that if we split S into N subintervals whose width is bounded by δ then for all c, x within the same subinterval we have

$$|f(x) - f(c)| < \epsilon.$$

To realise this idea, if $S = [a, b]$ then split S according to

$$a = x_0 < x_1 < \cdots < x_N = b$$

such that $x_{i+1} - x_i < \delta$ and take $c = (x_i + x_{i+1})/2$, the midpoint of $[x_i, x_{i+1}]$. The number of subintervals N needed depends on a, b and δ, e.g. $N = \left\lceil \frac{b-a}{\delta/2} \right\rceil$ works but the denominator just needs to be less than δ so $\delta/1.1$ would give a tighter N. From this we see that

$$N \approx \frac{b-a}{\delta} = \frac{L(b-a)}{\epsilon}.$$

The higher we take N the more accurate of an approximation we get. This idea of splitting S into N intervals is nice because it requires us to approximate only one point $c_i = (x_i + x_{i+1})/2$ on each subinterval, i.e. we are back to the simpler case which started our motivation of approximating at a fixed number of points.

With this idea in mind, the number of required neurons to realise our idea

$$2N = 2\frac{L(b-a)}{\epsilon}$$

grows linearly in L, the interval length $b-a$ and the reciprocal of the desired level of precision $1/\epsilon$. In practice, you only have control of the desired level of precision ϵ.

Step 3: Extending the idea to $S \subset \mathbb{R}^q$ compact

The idea is really not very different. Imagine $S \subset \mathbb{R}^2$ compact. The compactness is very hepful for visualisation. As long as we can create indicator-approximate functions on separate regions of S then use exactly the idea described above: split S into separate regions and approximate f on a single point in each. To see that we can approximate the indicator on regions of $S \subset \mathbb{R}^2$ simply see that we can create wall-like objects using the same idea of creating a 'steep' plane and sigmoiding it. As such, to isolate some region, produce three walls which enclose a region and you're done. If you want a more accurate enclosing of the region then use more walls. Without boundedness, which is given by compactness, this idea wouldn't work. Extend this idea to $\mathbb{R}^q$ and voilà.

Is this idea practical?

What the idea conveyed shows is *that* (and even *how*) an MLP can approximate any function in $C(S)$ for $S \in \mathbb{R}^q$ compact. It is not at all an indicator of how MLPs learned via gradient descent, in practice, go about approximating functions from data.

An interesting addition to this note is how this formulation gives an idea of how the weights from the hidden layer to the output explicitly encode the evaluation of the to-be-approximated function. Again, this is not what we'd expect MLPs to do in practice, instead they find a much more efficient representation, sharing information between neurons. Reminds me a lot of the research 3Blue1Brown discussed in one of his videos about transfromers and LLMs in which we spoke about how the LLM 'knows' to assign a high probability to the token `Jordan` when given `My favourite basketballer is Michael` as input. How does it 'know' to do that? It isn't at all implied by the grammar/semantics/whatever of the sentence. Said research concluded things about such pieces of knowledge being stored in the weights of the MLPs of the transformer blocks.

4.1.3 Posing classification as a regression problem

Which space does one transform their samples into in order to linearly separate them? In the case of neural net approaches, for $K-$classification, it's the $(K-1)$-probability simplex given by

$$\left\{(x_1, \ldots, x_K) \in [0,1]^K | x_1 + \cdots + x_K = 1\right\}.$$

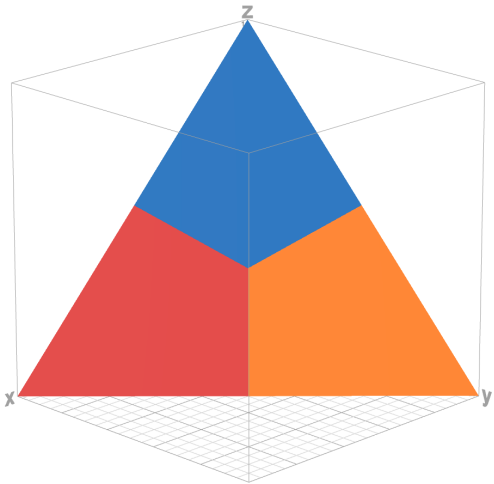


Figure 18: The 2−probability simplex coloured according to argmax in the three coordinates.

The 2−probability simpelx coloured according to argmax is illustrated in Figure 18. From the figure, it is clear how argmax can be computed by splitting the 2-probability simplex via two planes, e.g. the intersection of $x > y$ and $x > z$ on the simplex yield the red region. As such, neural nets used for classification can be seen as performing regression on the 2−probability simplex.

4.2 Backpropagation for MLPs

Backpropagation is a method of computing the gradient of a loss function pertaining to an MLP (or more generally some neural network) with respect to its weights and biases. Said gradients are used to learn suitable MLP parameters via gradient descent. We'll soon see why it's called backpropagation: it computes the gradients of interest in a way that requires going forwards through the MLP and then backwards. As with formulating MLPs, it's good to rigorously derive backpropagation at least once. It's sort of strange that I haven't done that yet (August 2025 is time of writing) and have just trusted that it holds up.

Onto the derivation, which makes use of the notation introduced earlier. Recall that a loss function is of the form

$$\mathcal{L} : \Omega_Y \times \Omega_Y \to \mathbb{R}_{\geq 0}$$

and that the loss induced by a choice of model parameters θ for a sample $(\mathbf{x}, y)$, which we will henceforth crudely denote by $\mathcal{L}$, is given by

$$\mathcal{L}\left(y, f_\theta(\mathbf{x})\right),$$

where f_θ is the model itself. Note that taking the gradient of $\mathcal{L}(y, f_\theta(\mathbf{x}))$ with respect to θ is perfectly valid even though the function $\mathcal{L}$ itself is not a function of θ. For an MLP,

$$\theta = \left(\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(k+1)}, \mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(k+1)}\right).$$

Let $\mathbf{z}^{(j)} = \mathbf{W}^{(j)}\mathbf{a}^{(j-1)} + \mathbf{b}^{(j)}$ for $j \in [k+1]$ from which we have $\mathbf{a}^{(j)} = \sigma_j\left(\mathbf{z}^{(j)}\right)$. Further, let

$$\delta_i^{(j)} = \frac{\partial \mathcal{L}}{\partial z_i^{(j)}}$$

for $j \in [k+1]$ and $i \in [n_j]$ where $\left(z_1^{(j)}, \ldots, z_{n_j}^{(j)}\right) = \mathbf{z}^{(j)}$. We seek to derive four statements:

1. $\delta_i^{(k+1)} = \sigma'_{k+1}\left(z_i^{(k+1)}\right) \frac{\partial \mathcal{L}}{\partial a_i^{(k+1)}}$ for $i \in [k+1]$
2. $\delta_i^{(j)} = \sigma'_j\left(z_i^{(j)}\right)\left(\left(\mathbf{W}^{(j+1)}\right)^\top \delta^{(j+1)}\right)_i$ for $j \in [k]$ and $i \in [n_j]$
3. $\frac{\partial \mathcal{L}}{\partial b_i^{(j)}} = \delta_i^{(j)}$ for $j \in [k+1]$ and $i \in [n_j]$
4. $\frac{\partial \mathcal{L}}{\partial w_{l,i}^{(j)}} = \delta_l^{(j)} a_i^{(j-1)}$ for $j \in [k+1]$, $i \in [n_{j-1}]$ and $l \in [n_j]$

Before deriving each statement in order, we can see already what was meant earlier by propagating backwards through the MLP: to compute the gradients relevant to performing gradient descent, backpropagation requires the computation of $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(k+1)}$ from which $\delta^{(k+1)}$ can be computed. Note that the $\frac{\partial \mathcal{L}}{\partial a_i^{(k+1)}}$ term in the first statement is often easily computed, e.g. for MSE loss we have

$$\frac{\partial \mathcal{L}}{\partial a_i^{(k+1)}} = a_i^{(k+1)} - y_i.$$

From there, one can compute $\delta^{(k)}, \ldots, \delta^{(1)}$ in that order, effectively propagating through the MLP backwards. With $\delta^{(1)}, \ldots, \delta^{(k+1)}$ and $\mathbf{a}^{(0)}, \mathbf{a}^{(1)} = \sigma_1\left(\mathbf{z}^{(1)}\right), \ldots, \mathbf{a}^{(k+1)} = \sigma_{k+1}\left(\mathbf{z}^{(k+1)}\right)$, computing the relevant gradients is straightforward.

Statement 1. Using the chain rule, see that

$$\begin{aligned}\delta_i^{(k+1)} &= \frac{\partial \mathcal{L}}{\partial z_i^{(k+1)}} \\ &= \frac{\partial \mathcal{L}}{\partial a_i^{(k+1)}} \frac{a_i^{(k+1)}}{\partial z_i^{(k+1)}} \\ &= \sigma'_{k+1}\left(z_i^{(k+1)}\right) \frac{\partial \mathcal{L}}{\partial a_i^{(k+1)}}.\end{aligned}$$

Statement 2. First, note that

$$\begin{aligned}\left(\mathbf{W}^{(j+1)}\right)^\top \delta^{(j+1)} &= \begin{bmatrix} w_{1,1}^{(j+1)} & \dots & w_{n_{j+1},1}^{(j+1)} \\ \vdots & \ddots & \vdots \\ w_{1,n_j}^{(j+1)} & \dots & w_{n_{j+1},n_j}^{(j+1)} \end{bmatrix} \begin{bmatrix} \delta_1^{(j+1)} \\ \vdots \\ \delta_{n_{j+1}}^{(j+1)} \end{bmatrix} \\ &= \sum_{l=1}^{n_{j+1}} \delta_l^{(j+1)} \begin{bmatrix} w_{l,1}^{(j+1)} \\ \vdots \\ w_{l,n_j}^{(j+1)} \end{bmatrix}\end{aligned}$$

and so

$$\left(\left(\mathbf{W}^{(j+1)}\right)^\top \delta^{(j+1)}\right)_i = \sum_{l=1}^{n_{j+1}} \delta_l^{(j+1)} w_{l,i}^{(j+1)}$$

which reduces the to-be-derived statement to

$$\delta_i^{(j)} = \sigma'_j(z_i^{(j)}) \sum_{l=1}^{n_{j+1}} \delta_l^{(j+1)} w_{l,i}^{(j+1)}.$$

Using the chain rule, we deduce that

$$\delta_i^{(j)} = \frac{\partial \mathcal{L}}{\partial z_i^{(j)}} = \sum_{l=1}^{n_{j+1}} \frac{\partial \mathcal{L}}{\partial z_l^{(j+1)}} \frac{\partial z_l^{(j+1)}}{\partial z_i^{(j)}} = \sum_{l=1}^{n_{j+1}} \delta_l^{(j+1)} \frac{\partial z_l^{(j+1)}}{\partial z_i^{(j)}}.$$

To complete our deduction, see that

$$z_l^{(j+1)} = \sum_{i=1}^{n_j} w_{l,i}^{(j+1)} \sigma_{j+1}\left(z_i^{(j)}\right) + b_l^{(j+1)}$$

from which we obtain

$$\frac{\partial z_l^{(j+1)}}{\partial z_i^{(j)}} = \sigma'_{j+1}(z_i^{(j)}) w_{l,i}^{(j+1)}$$

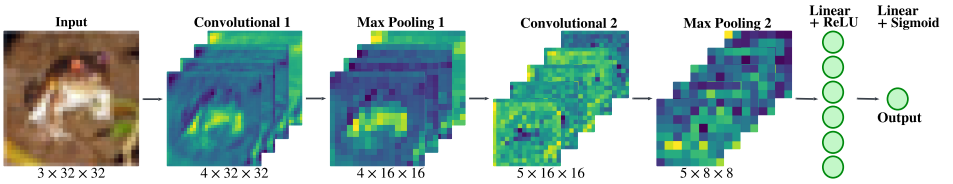


Figure 19: A binary classification CNN processing an image of a frog.

completing the deduction.

Statement 3. This one's a simple application of the chain rule, as in

$$\frac{\partial \mathcal{L}}{\partial b_i^{(j)}} = \sum_{l=1}^{n_{j+1}} \frac{\partial \mathcal{L}}{\partial z_l^{(j)}} \frac{\partial z_l^{(j)}}{\partial b_i^{(j)}} = \frac{\partial \mathcal{L}}{\partial z_i^{(j)}} = \delta_i^{(j)}.$$

Statement 4. As before, see that

$$z_l^{(j)} = \sum_{i=1}^{n_j} w_{l,i}^{(j)} a_i^{(j-1)} + b_l^{(j)}$$

which yields

$$\frac{\partial z_l^{(j)}}{\partial w_{l,i}^{(j)}} = a_i^{(j-1)}.$$

To complete the derivation, consider

$$\frac{\partial \mathcal{L}}{\partial w_{l,i}^{(j)}} = \frac{\partial \mathcal{L}}{\partial z_l^{(j)}} \frac{\partial z_l^{(j)}}{\partial w_{l,i}^{(j)}} = \delta_l^{(j)} a_i^{(j-1)}.$$

4.3 Convolutional Neural Networks (CNNs)

Convolutional neural networks (CNNs) were initially designed to process image and, by extension, video data. Over time, they have been applied to processing far richer forms of data. A forward pass through a CNN designed for image processing is illustrated in Figure 19, in which we see that their architecture can be described on a high level as consisting of two parts. The first part performs feature extraction through convolution and pooling operations. The second part employs an MLP to produce the overall model's output given the earlier-extracted features, e.g. the probability that an input image is of a given class.

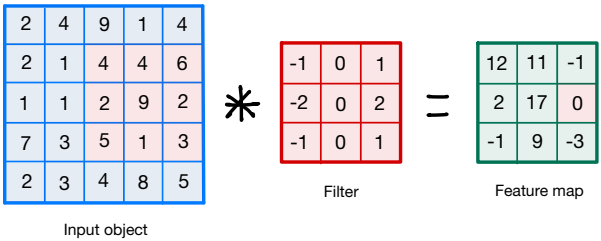


Figure 20: Convolution with a vertical edge detection filter with 0 bias.

Before deep learning architectures became mainstream, researchers used handcrafted features (SIFT, HOG, SURF, etc.) for feature extraction and fed extracted features to a given model, e.g. an SVM for classification. That is, feature extraction wasn't learned, nor was it entwined with the inference component of their models. For image processing tasks, humans turn out to be far worse than deep learning architectures at knowing which features are most informative of what is being predicted. Hence, CNNs are now the go-to architecture for image and video processing. There's something to be said here about vision transformers and the extent to which they are competitive with CNNs but this can wait. Relevant interactive visualisation: `adamharley.com/nn_vis/cnn/3d.html`.

4.3.1 Convolutions

Convolution operations involve using a (typically square) matrix to, in some sense, summarise the information embedded in parts of the object over which it traverses. This matrix is referred to as a filter or a kernel — I prefer using filter as the term kernel has so many meanings elsewhere in maths. You can think of convolution operations as a sort of dot product between the given filter and the grid of pixel values over which lies during its traversal. For a simple illustration of this idea, consider Figure 20. Since CNNs consists of multiple convolution layers, the output of a convolution layer is often referred to as a feature map.

More formally, given a feature map $I \in \mathbb{R}^{C\times W\times H}$ and a filter $K \in \mathbb{R}^{C\times k\times k}$ with $k < W$ and $k < H$, the feature map $F = K * I \in \mathbb{R}^{W'\times H'}$ is the convolution of K over I is as in

$$F_{i,j} = \sum_{c=1}^{C}\sum_{x=1}^{k}\sum_{y=1}^{k} K_{c,x,y} \cdot I_{c,x+i,y+j} + b_c$$

where b_c is the filter's bias in channel c. Nowadays, $k = 3$ is a very popular choice. Note that this notation is intended to illustrate what a convolution

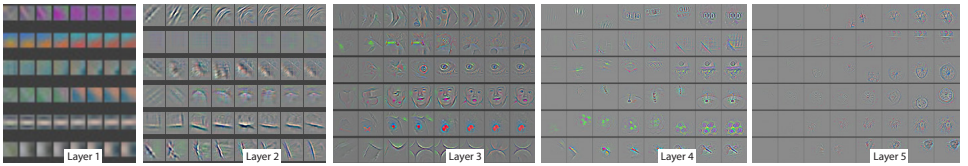


Figure 21: The level of abstraction of a CNN's learned filters. Earlier layers pertain to low-level features like edges or textures. Later layers pertain to higher-level features like faces or wheels.

looks like algebraically. As it stands, what's written above does not account for padding, stride, etc.

Generally, earlier filters in a CNN architecture extract lower-level features, like edges and textures, while later convolutions extract higher-level features like entire components of object, e.g. wheels. This is illustrated in Figure 21.

Post-convolution spatial dimensions

Given a feature map $F \in \mathbb{R}^{W \times H}$, suppose we compute the convolution $O \in \mathbb{R}^{W' \times H'}$ of the filter $K \in \mathbb{R}^{k \times k}$ over F with uniform padding P and stride S. The precise dimensions W' and H' of the new feature map O are given by

$$W' = \left\lfloor \frac{W + 2P - k}{S} \right\rfloor + 1$$

and

$$H' = \left\lfloor \frac{H + 2P - k}{S} \right\rfloor + 1.$$

It's not too tricky to derive these. It's the kinda thing you do once and never again.

4.3.2 Pooling

Pooling operations reduce the spatial dimensions of a given feature map by applying relatively straightforward transformations: max pooling, min pooling, mean pooling, etc. There's not much point to expressing them rigorously. Instead, consider Figure 22. Benefits of pooling are improved translation-invariance and reduced required compute.

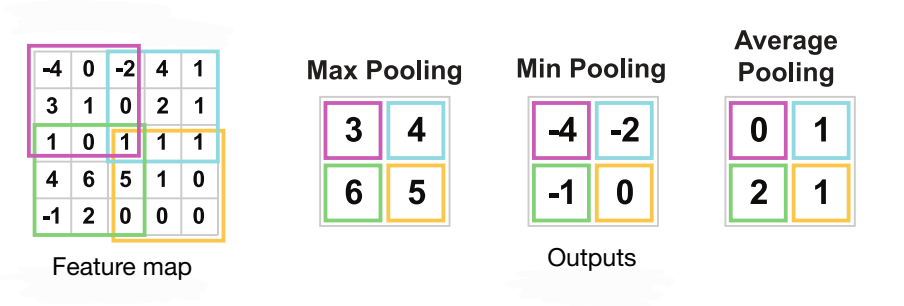


Figure 22: Common flavours of pooling.

Why use a CNN over an MLP?

Massively improved parameter-efficiency, translation invariance and learning efficiency[a] (fewer samples needed to learn distributions to similar degrees of precision). These are in part due to the inductive bias present of their architecture.

[a] ai.stackexchange.com/questions/27407

4.4 TODO: Recurrent Neural Networks (RNNs)

4.4.1 Long Short-Term Memory (LSTMs)

Schmidhuber

4.4.2 Gated Recurrent Units (GRUs)

4.5 TODO: Transformers

Just another (very effective) architecture for correlation extraction.

4.5.1 Tokens and Embedding

Tokenisation is the process of slicing a sample up into tokens. Think of splitting a 512×512 image into its four 256×256 quadrants or the sentence 'The chef cooked a meal for the critic.' into

The purpose of tokenisation is self-explanatory: we'd like an idea of dependence between different parts of the input as to aid our output prediction.

An embedding function $E : \{1, \ldots, |S|\} \to \mathbb{R}^d$, where $d \in \mathbb{N}$ is the embedding dimension, offers high dimensional numeric representations of tokens. What makes token embedding functions interesting is how well they can capture underlying semantics, e.g. of natural language. For example, an embedding function learned on pieces of text might satisfy

$$E(\text{Germany}) + E(\text{fascist}) - E(\text{Mussolini}) \approx E(\text{Hitler})$$

for which the tokenisations occur.

4.5.2 Positional encoding

Use sinusoids to form a position vector P_k for the kth token in an input sequence of length L. Add P_k to the word embedding E_k for $k = 0, \ldots, L-1$ to obtain the input $[E_0 + P_0, \ldots, E_{L-1} + P_{L-1}]$ fed to the model.

4.5.3 Attention

Has been around as an idea for a long time in less clear terms.

4.5.4 Transformer blocks

...

4.6 Misc. questions

Q1: What is the purpose of activation functions?

Without the application of at least one activation function, the values of the output neurons would simply be the result of matrix-vector multiplication and vector addition. That is to say, the output of the neural network, without an any activation function, would be a purely linear transformation of the input. The issue with this is that most problems require some degree of non-linearity. The staple example of this is the classification of two-dimensional samples which are not linearly separable. An example is illustrated in the left of Figure 23.

The point is that there is no line that would separate the classes in either case: non-linearity is needed! In the latter case, it is desirable to be able to somehow punch the interior of the unit disc to form a blue mountain with the surrounding ground covered in red. This can be achieved by the non-linear transformation $\phi(x, y) = \max(0, 1-(x^2+y^2))$ which is illustrated in the right of Figure 23.

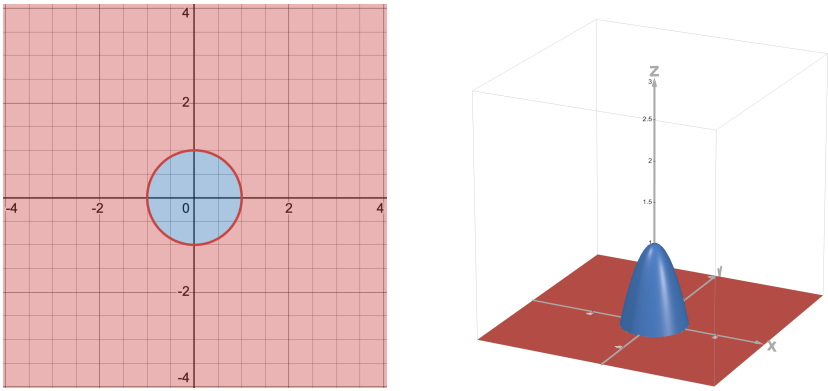


Figure 23: Non-linearly separable data (left) and its image under a non-linear transformation (right).

From here, a natural linear decision boundary is the plane $z = 0$ (i.e. the red ground surrounding the blue mountain). Any sample above the plane, i.e. any sample whose $z-$coordinate is positive, is classified as blue. It is otherwise classified as red. We can be more clever than this though. This transformation required us to map samples to 3D — how about mapping to just 1D? To do this, a complete decision function is given by

$$D(x, y) = \left\lceil \frac{\lfloor x^2 + y^2 \rfloor}{x^2 + y^2} \right\rceil$$

and define $D(0, 0) = 0$ to account for the removable singularity.

Q2: Why not just one activation function in the output layer?

If there were a single activation function towards the end of the architecture then everything that came before would be a series of linear transformations of the input. The composition of linear transformations is just a linear transformation. As such, said model would be equivalent to a single layer perceptron, i.e. linear transformation + non-linear activation function, which would yield linear decision boundaries only.

To be super pedantic, single-layer perceptrons yielding only linear decision boundaries is the case if the activation function used is monotonic, which is often the case. If you allow for weird, non-monotonic activations then this linear decision boundary result doesn't necessary hold but, again, this is a pretty pedantic detail.

Q3: Must activation functions be monotonic?

Nope, it just makes things related to optimisation easier while having a sufficiently-small effect on performance in practice. Some results surrounding neural networks necessitate monotonic activation functions though.

Q4: Which activation functions should be used and when?

Roughly speaking, I wouldn't worry too much about this in practice. The usual choices are all nice and differentiable (with small caveats, like with ReLU at 0) facilitating backpropagation. You should sometimes care about the output given the problem at hand. For example, if you need outputs in $[-1, 1]$ then tanh is a natural choice.

Note that some choices, like Leaky ReLU, introduce hyperparameters to the model which is sometimes undesirable, e.g. if one seeks to minimise the number of hyperparameters for the strength of their result (if there's no hyperparameter to cherry pick then your result is more trustworthy).

Q5: How does one prevent vanishing/exploding gradients?

Skip connections help. Architectures which use them are sometimes called residual networks.

- Most intuitive benefit: it helps prevent some layer in the architecture from degrading the input entirely. It's like a nice reminder to the model what it was working from before it got to this point
- Reduces vanishing or exploding gradients: opens the door to far deeper architectures
- Ensures that the model has to learn the residual as opposed to the full underlying function: faster convergence usually
- Takes pressure off parameter initialisation: if you set parameters to zero then initially model is just the identity and learning from there is feasible

It's worth mentioning that said vanishing behaviour can be observed during forward passes too.

5 Generative Models

Modelling data-generating processes is the heart of science: the laws which dictate graviation, ... From a probabilistic point of view, the analogous methodology is to model the joint distribution of model variables $\mathbf{X} = (X_1, \ldots, X_m)$. For example, if we're interested in understanding what gave rise to the distribution of 16×16 images of cats (to whatever extent of detail) then one might seek to represent their joint probability function

$$p(\mathbf{x}) = p(x_1, \ldots, x_{256}).$$

If the representation allows, we may realise (or generate) new samples pertaining to said distribution. Alternatively, we may use the representation to answer natural probabilistic queries about given samples, e.g. the likelihood that a given image is in-distribution, the class label of an image, the marginal likelihood of an incomplete image and so on. In the case of LLMs, we're typically interested in modelling the distribution of the next token conditioned on the tokens that came before. We simply sample from said distribution (almost feels like classification) to extend the text and repeate in an autoregressive manner.

This section will be about methods of representing the joint distributions of distributions pertaining to images. The discussed methods are often applicable to generative modelling in general.

Small note: to avoid having to write *probability mass/density function* in full throughout this section, *probability function* will be written.

5.1 IMPROVE: Bayesian networks (BNs)

A natural approach to representing the joint probability function of model variables is to decompose it into smaller probability functions and model those. For example, you might decompose the probability function according to the chain rule of probability as in

$$\begin{aligned} p(x_1, \ldots, x_m) &= p(x_1)p(x_2|x_1)\cdots p(x_m|x_1, \ldots, x_{q-1}) \\ &= \prod_{i=1}^{q} p(x_i|x_1, \ldots, x_{i-1}). \end{aligned}$$

From here, sampling may be done autoregressively: sample x_1 according to $p(x_1)$, x_2 according to $p(x_2|x_1)$ and so on until an entire realisation of $\mathbf{X} = (X_1, \ldots, X_m)$ is obtained. Why stop at decomposition though? In an image, nearby pixel values may correlate but far away pixels may not, so

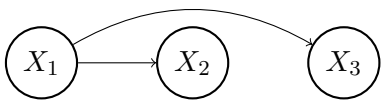
why incoroporate them in the conditional? Why not learn the dependencies between the model variables? Boom, Bayesian networks.

5.1.1 Formulating BNs

With their motivation out of the way, we can move on to their formulation. A Bayesian network consists of a directed acyclic graph (DAG) $\mathcal{G}$ and a parameter set θ. The graph $\mathcal{G}$ has a single node for each model variable $X_1, \ldots, X_m$ and directed edges represent conditional dependencies. For example, for model variables X_1, X_2 and X_3, if $X_3|X_1$ is independent of X_2 then their joint probability function may be decomposed as

$$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1)$$

whose corresponding DAG is given by



Notice that the choice of X_1 as the model variable upon which nothing is conditioned is entirely arbitrary. As a result, for a given decomposition, DAGs are not unique.

As a result, if the treewidth $(\ldots)$ of the Bayesian network is sufficiently small then it answering probabilistic queries is efficient.

The parameter set θ pertains to the parameters of the conditional distributions shown in $\mathcal{G}$.

5.1.2 Learning BNs from data

Learn their DAG via scoring. Learn their parameters via ?

5.2 Variational Autoencoders (VAEs)

In describing variational autoencoders (VAEs), it's natural to first motivate autoencoders and then spice things up with variational inference do generative things.

5.2.1 Autoencoders: no variation yet!

An autoencoder $(\Omega_{\mathbf{X}}, d, \phi, \theta)$ consists of a sample space $\Omega_{\mathbf{X}} \subset \mathbb{R}^m$, a latent dimension $d \in \mathbb{Z}_{\geq 1}$, an encoder $\phi : \mathbb{R}^m \to \mathbb{R}^d$ and a decoder $\theta : \mathbb{R}^d \to \mathbb{R}^m$.

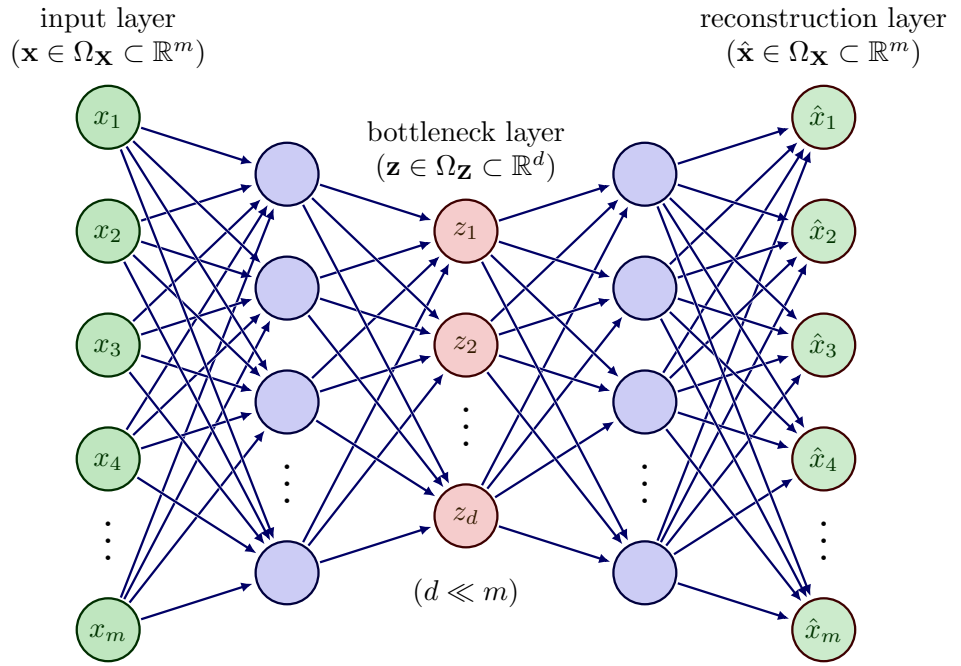


Figure 24: An autoencoder in which ϕ and θ are fit using multi-layer perceptrons (MLPs).

Using a subset of $\Omega_{\mathbf{X}}$, one typically learns the encoder ϕ and the decoder θ with the goal of approximating the identity on $\Omega_{\mathbf{X}}$ via $\theta \circ \phi$. That is, roughly put, one seeks to obtain an encoder/decoder pair such that $\theta(\phi(\mathbf{x})) \approx \mathbf{x}$ for all $\mathbf{x} \in \Omega_{\mathbf{X}}$.

In intuitive terms, one can see the encoder ϕ as a compressor of samples in $\Omega_{\mathbf{X}}$ to their compressed (or latent) representation in $\mathbb{R}^d$ and so one might refer to the image of ϕ as the latent space $\Omega_{\mathbf{Z}}$. Similarly, the decoder θ can be seen as a decompressor of compressed representations $\mathbf{z} \in \Omega_{\mathbf{Z}}$ yielding the original sample $\mathbf{x}$. In line with this notion of an autoencoder as a compressor/decompressor, the latent dimension d is typically taken to be far smaller than the dimension of the distribution of interest, i.e. $d \ll q$. Of course, when $d \ll q$, learning such mappings ϕ and θ typically involves some loss of information if $\Omega_{\mathbf{X}}$ is a manifold whose intrinsic dimension is greater than d. That is, autoencoders are typically lossy compressors.

Autoencoder example

Suppose $\Omega_{\mathbf{X}} = \{(a, a, b) \in \mathbb{R}^3 : ||(a, a, b)|| \leq 1\}$ and $d = 2$. One immediately notices that $\Omega_{\mathbf{X}}$ is a two-dimensional surface embedded in $\mathbb{R}^3$ as it is the intersection of the unit ball and the plane $\{(a, a, b) : a, b \in \mathbb{R}\}$. To produce representations of $\mathbf{x} = (a, a, b) \in \Omega_{\mathbf{X}}$ in $\mathbb{R}^2$ (i.e. to compress a sample) one might employ the encoder $\phi_2(x, y, z) = (x, z)$. To reconstruct samples from their latent representation (i.e. to decompress) one might employ the decoder $\theta_2(x, z) = (x, x, z)$. The autoencoder $(\Omega_{\mathbf{X}}, 2, \phi_2, \theta_2)$ offers lossless compression on the sample space as $(\theta_2 \circ \phi_2)(\mathbf{x}) = \mathbf{x}$ for all $\mathbf{x} \in \Omega_{\mathbf{X}}$.

If we instead desire latent representations of $\mathbf{x} = (a, a, b) \in \Omega_{\mathbf{X}}$ in $\mathbb{R}$, i.e. if $d = 1$, one might employ the encoder $\phi_1(x, y, z) = x$ and the decoder $\theta_1(x) = (x, x, x)$. The autoencoder $(\Omega_{\mathbf{X}}, 1, \phi_1, \theta_1)$ offers lossy compression, i.e. some information pertaining to a sample is lost during its compression and decompression as $(\theta_1 \circ \phi_1)(a, a, b) = (a, a, a)$ for all $a, b \in \mathbb{R}$.

One's tolerance for the loss incurred by a given encoder/decoder pair is task-dependent. Given a dataset $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \Omega_{\mathbf{X}}$ and a latent dimension d, one might compute the pair $(\phi^*, \theta^*) \in \Phi \times \Theta$ which minimises the empirical risk over D where Φ and Θ are chosen function families. That is, computing

$$(\phi^*, \theta^*) = \underset{(\phi,\theta)\in\Phi\times\Theta}{\arg\min} \sum_{i=1}^{n} ||\mathbf{x}_i - \theta(\phi(\mathbf{x}_i))||^2.$$

If the function families permit the computation of gradients of $\phi \in \Phi$ and $\theta \in \Theta$ then this computation may be done using gradient descent. For example, one could take Φ and Θ to be the family of MLPs of appropriate input and output dimensions, as illustrated in Figure 24.

5.2.2 Motivating VAEs

Autoencoders are great for compression, denoising, disecting LLMs (sparse autoencoders), etc. but we're yet to see how we might apply them to generative tasks, e.g. sampling from the complex underlying distribution $p(\mathbf{x})$ from which their training dataset $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ was sampled. An intuitive first idea is to train an autoencoder, randomly sample from its latent space and feed the sampled latent through the learned decoder, as in Figure 25.

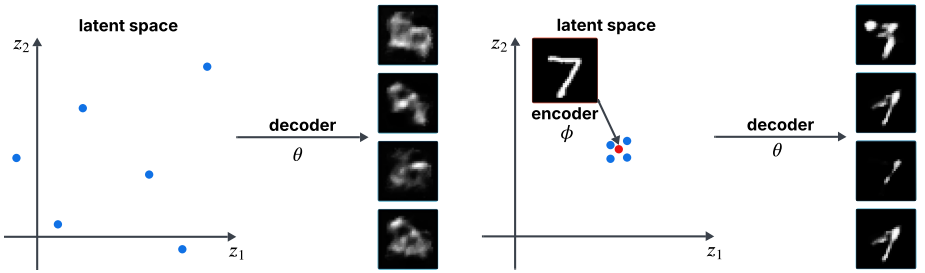


Figure 25: Failed intuitive ideas for using an autoencoder as a generative model.

The outputs of the decoder should be similar to some subet of the training data, right? No.

Disaster strikes and we begin to see just how unstructured the latent space of an autoencoder is: randomly sampling from it and decoding yields nonsense outputs from the decoder because most of the latent space itself is meaningless (nothing is encoded to most of it, so in some sense a lot of it is 'un-utilised'). This is perhaps unsurprising as at no point during training an autoencoder do we encourage it to interpolate nicely between encodings.

An intuitive second idea is to feed the decoder points in the latent space which are close to the encoding of a given training sample, as in Figure 25. The outputs of the decoder should be a sample of the class of said encoding's sample, right? No. Distaster strikes again. For meaningful decodings using this idea, one must take points in the latent space which are ridiculously close to this chosen sample's latent embedding. So close that you'd be practically reconstructing the original training sample each time — certainly not what we mean when we say that we'd like to generate new samples. So how do we do autoencoder-like things in a way that yields well-structured latent spaces? Variational autoencoders (VAEs) to the rescue!

5.2.3 Formulating VAEs

Leaving aside, for now, how to learn one from data, VAEs can informally be seen as an extension of autoencoders in which the encoder and decoder each output parameters of a distribution belonging to some pre-chosen distribution family. Extending the notation used to define autoencoders, a variational autoencoder $(\Omega_{\mathbf{X}}, d, \mathcal{Q}_d, \mathcal{P}_m, \phi, \theta)$ consists of the sample space of the distribution of interest $\Omega_{\mathbf{X}} \subset \mathbb{R}^m$, a latent dimension $d \in \mathbb{Z}_{\geq 1}$, the

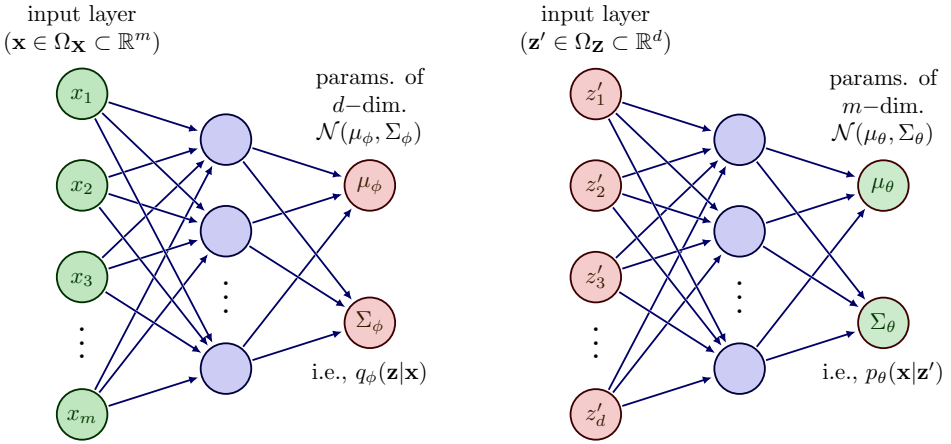


Figure 26: Left: an encoder which outputs parameters of the latent distribution $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu_\phi, \Sigma_\phi)$. Right: a decoder which outputs parameters of the reconstruction distribution $p_\theta(\mathbf{x}|\mathbf{z}') = \mathcal{N}(\mu_\theta, \Sigma_\theta)$ in which $\mathbf{z}' \sim q_\phi(\mathbf{z}|\mathbf{x})$.

parameter space $\mathcal{Q}_d$ of a family of $d-$dimensional conditional distributions denoted $q_\phi(\mathbf{z}|\mathbf{x})$, the parameter space $\mathcal{P}_m$ of a family of $m-$dimensional conditional distributions denoted $p_\theta(\mathbf{x}|\mathbf{z})$, an encoder $\phi : \mathbb{R}^m \to \mathcal{Q}_d$ and a decoder $\theta : \mathbb{R}^d \to \mathcal{P}_m$.

To illustrate the intended meaning of the newly-introduced parameter spaces $\mathcal{Q}_d$ and $\mathcal{P}_m$, one's encoder might yield the expectation vector and covariance matrix of a $d-$dimensional Gaussian $\mathcal{N}(\mu, \Sigma)$, i.e. $\mathcal{Q}_d$ could be the parameter space of the family of $d-$dimensional Gaussian distributions yielding

$$\mathcal{Q}_d = \{(\mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \mathcal{S}^d_{++}\} = \mathbb{R}^d \times \mathcal{S}^d_{++}$$

where $\mathcal{S}^d_{++}$ denotes the set of all positive-definite matrices in $\mathbb{R}^{d \times d}$.

As the encoder of a VAE yields a $d-$dimensional distribution $q_\phi(\mathbf{z}|\mathbf{x})$ given the sample $\mathbf{x} \in \Omega_{\mathbf{X}}$, to obtain a latent $d-$dimensional representation $\mathbf{z}' \in \mathbb{R}^d$ of $\mathbf{x}$, one computes the parameters $\phi(\mathbf{x}) \in \mathcal{Q}_d$ of $q_\phi(\mathbf{z}|\mathbf{x})$, e.g. a $d-$dimensional Gaussian, via the encoder and samples $\mathbf{z}' \sim q_\phi(\mathbf{z}|\mathbf{x})$. To reconstruct $\mathbf{x}$ from its latent representation $\mathbf{z}'$, one computes the parameters $\theta(\mathbf{z}') \in \mathcal{P}_m$ of the $q-$dimensional reconstruction distribution $p_\theta(\mathbf{x}|\mathbf{z}')$ via the decoder. Sampling $\mathbf{x}' \sim p_\theta(\mathbf{x}|\mathbf{z}')$ yields (ideally) a sufficiently-accurate reconstruction of the original sample $\mathbf{x}$. Achieving accurate reconstructions is done in a similar manner to autoencoders: by including a penalty term pertaining to reconstruction quality in the loss function used to learn the encoder and decoder.

Note that one's choice of $\mathcal{Q}_d$ and $\mathcal{P}_m$ should take into account the need to sample efficiently and so they should correspond to families of distributions which offer efficient means of sampling, e.g. Gaussians, as in Figure 26.

At this point, a natural question arises: for which tasks is learning a VAE more appropriate than learning an autoencoder? The answer lies in the purpose of VAEs which is two-fold: 1) to perform sufficiently-accurate compression/decompression and 2) to produce a sufficiently regularised approximation of the latent space $\Omega_{\mathbf{Z}}$. The latter is ensured by how one learns a VAE from data, which we soon consider. In brief, when learning an autoencoder one never imposes restrictions on the latent space beyond encouraging the model to yield sufficiently-accurate reconstructions. As a result, the latent space of an autoencoder is not well-structured. For example, for latent samples $\mathbf{z}_1, \mathbf{z}_2 \in \Omega_{\mathbf{Z}}$ which are 'close' in the latent space, their reconstructions are not necessarily 'close' in $\mathbb{R}^m$. VAEs seek to remedy this.

To learn a VAE from data, we look to maximise the evidence lower bound (ELBO) over some dataset $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \Omega_{\mathbf{X}}$. Over a single sample $\mathbf{x} \in \Omega_{\mathbf{X}}$, the ELBO is a tight lower bound of $\log(p(\mathbf{x}))$. As such, maximising the ELBO over D can be seen as performing approximate maximum-likelihood estimation over D (sometimes referred to as evidence maximisation). To derive the ELBO, first note that given a decoder θ (which parameterises the reconstruction distribution $p_\theta(\mathbf{x}|\mathbf{z})$), we may express the marginal distribution $p(\mathbf{x})$ using $p_\theta(\mathbf{x}|\mathbf{z})$ as

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z} = \int p_\theta(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z}. \tag{1}$$

Using Equation 1, along with the encoder ϕ (which parameterises the latent

distribution $q_\phi(\mathbf{z}|\mathbf{x})$) we derive the ELBO over a single sample $\mathbf{x} \in \Omega_{\mathbf{X}}$:

$$\begin{aligned}
\log(p(\mathbf{x})) &= \log\left(\int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}\right) \\
&= \log\left(\int q_\phi(\mathbf{z}|\mathbf{x})\frac{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}d\mathbf{z}\right) \\
&= \log\left(\mathbb{E}_{\mathbf{Z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\frac{p_\theta(\mathbf{x}|\mathbf{Z})p(\mathbf{Z})}{q_\phi(\mathbf{Z}|\mathbf{x})}\right]\right) \\
&\geq \mathbb{E}_{\mathbf{Z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left(\frac{p_\theta(\mathbf{x}|\mathbf{Z})p(\mathbf{Z})}{q_\phi(\mathbf{Z}|\mathbf{x})}\right)\right] \\
&= \mathbb{E}_{\mathbf{Z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log(p_\theta(\mathbf{x}|\mathbf{Z}))\right] - \mathbb{E}_{\mathbf{Z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left(\frac{q_\phi(\mathbf{Z}|\mathbf{x})}{p(\mathbf{Z})}\right)\right] \\
&= \mathbb{E}_{\mathbf{Z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log(p_\theta(\mathbf{x}|\mathbf{Z}))\right] - D_{\mathrm{KL}}(q_\phi(\mathbf{Z}|\mathbf{x})||p(\mathbf{Z})) \\
&=: \text{ELBO}
\end{aligned}$$

in which the inequality arises due to Jensen's inequality, as in $\mathbb{E}[\log(f(\mathbf{Z}))] \leq \log(\mathbb{E}[f(\mathbf{Z})])$, and $D_{\mathrm{KL}}(Q||P)$ denotes the KL-divergence between distributions Q and P, which is detailed in the appendix. The effect of maximising

$$\text{ELBO} = \mathbb{E}_{\mathbf{Z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log(p_\theta(\mathbf{x}|\mathbf{Z}))\right] - D_{\mathrm{KL}}(q_\phi(\mathbf{Z}|\mathbf{x})||p(\mathbf{Z}))$$

over a dataset is often explained term-by-term. The first term is a principled measure of the VAE's ability to reconstruct latent representations, as with autoencoders. The second term is a principled measure of the similarity of the latent distribution $q_\phi(\mathbf{z}|\mathbf{x})$ and the prior $p(\mathbf{z})$. The prior is chosen before training and the most common choice is a $d-$dimensional standard Gaussian, i.e. $p(\mathbf{z}) = \mathcal{N}(0, I_d)$. As such, minimising the negative KL-divergence between the latent distribution and said prior is often interpreted as encouraging the learning of the encoder such that latent representations are distributed according to the prior. This is particularly useful in the case that one is learning a VAE to generate new samples from $\Omega_{\mathbf{X}}$: post-training, sample $\mathbf{z}' \sim p(\mathbf{z})$, compute the parameters $\theta(\mathbf{z}')$ of the reconstruction distribution $p_\theta(\mathbf{x}|\mathbf{z}')$ via the decoder and sample from it. Note that using a VAE for generative purposes does not invoke the use of the encoder, only the decoder is required post-training.

Given a dataset $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \Omega_{\mathbf{X}}$, we learn a VAE by choosing function classes Φ and Θ (e.g. MLPs as in Figure 26) and computing

$$\underset{(\phi,\theta)\in\Phi\times\Theta}{\arg\max}\left[\sum_{i=1}^{n}\mathbb{E}_{\mathbf{Z}\sim q_\phi(\mathbf{z}|\mathbf{x}_i)}\left[\log(p_\theta(\mathbf{x}_i|\mathbf{Z}))\right] - D_{\mathrm{KL}}(q_\phi(\mathbf{Z}|\mathbf{x}_i)||p(\mathbf{Z}))\right].$$

In practice, after choosing a latent dimension d, we often take $p(\mathbf{z}) = \mathcal{N}(0, I_d)$, $\mathcal{Q}_d = \mathbb{R}^d \times \mathcal{S}^d_{++}$ and $\mathcal{P}_m = \mathbb{R}^m \times \mathcal{S}^m_{++}$, i.e. Gaussians for the latent and reconstruction distribution families and the standard $d-$dimensional Gaussian for the prior. A benefit of these choices is that it yields a differentiable and easy-to-implement closed form for the KL-divergence term in the ELBO. Additionally, the expectation pertaining to the reconstruction term is typically approximated via a single sample, boiling down to a mean square error term[3].

5.2.4 REVIEW: Backpropagation for VAEs

To find suitable parameters of a given VAE architecture using gradient descent, we require that the ELBO is differentiable. With respect to θ, the gradient is approximated according to

$$\begin{aligned}\nabla_\theta \text{ELBO} &= \mathbb{E}_{\mathbf{Z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log(p_\theta(\mathbf{x}|\mathbf{Z}))\right] \\ &= \frac{1}{k}\sum_{j=1}^{k} \nabla_\theta \log(p_\theta(\mathbf{x}|\mathbf{z}_j))\end{aligned}$$

which is no issue when the decoder's architecture is backpro-compatible. However, with respect to ϕ, the gradient of the ELBO is less friendly.

5.2.5 TODO: Quirks of VAEs

Blur outside of central object.

5.3 IMPROVE: Generative Adversarial Networks (GANs)

To motivate generative adversarial networks (GANs), first consider what we want a generative model to do: produce convincing samples. As such, if there is some method of distinguishing real samples from those produced by a generative model, a perfect model would be able to fool the method. That is, the decisions made by the discriminative model should be akin to random guessing if the generative model is particularly good. This idea was realised by Ian Goodfellow while at a bar (the story is detailed in Genius Makers). From what I recall, there are disputes over whether or not the idea should be attributed to someone else. I'll look into it at some point.

This is precisely the idea that motivates GANs with inspiration taken from game theory. A GAN consists of a generative component $G : \Omega_\mathbf{Z} \to \Omega_\mathbf{X}$

[3] `https://n8python.github.io/mnistLatentSpace/`

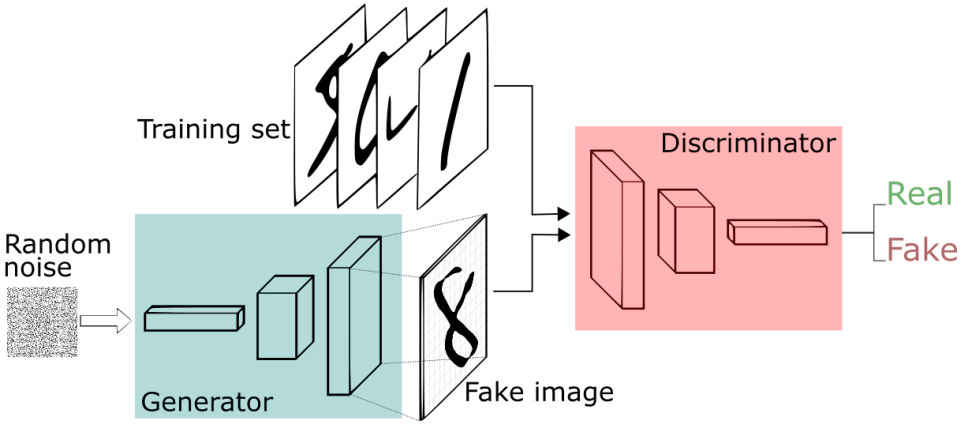


Figure 27: Example GAN architecture.

and a discriminative component $D : \Omega_{\mathbf{X}} \to \{0, 1\}$, which are trained in parallel. That is, G produces a new sample given a latent $\mathbf{z} \in \Omega_{\mathbf{Z}}$ while D learns to distinguish between training samples and outputs of G. Essentially, the models compete. In terms of loss functions, the discriminator employs binary cross entropy, as in

$$\mathcal{L}_D = -\mathbb{E}_{\mathbf{X}} \left[\log(D(\mathbf{X}))\right] - \mathbb{E}_{\mathbf{Z}} \left[\log\left(1 - D(G(\mathbf{Z}))\right)\right]$$

and the generator makes use of

$$\mathcal{L}_G = -\mathbb{E}_{\mathbf{Z}} \left[\log(D(G(\mathbf{Z})))\right].$$

As for architectures, the usual choices are natural: the disrciminator is fit using a CNN and the generator is fit using a deconvolutional CNN. So, as with other image generative models, to generate a new sample, we sample Gaussian noise and run it through the deconvolution process.

5.3.1 Quirks of GANs

- Training them is a fucking headache. Constant model collapse.
- GANs for non-image generation are hard to train.
- Bluriness

5.3.2 Flavours of GANs

Perhaps the most important contribution of GAN-related things is not the vanilla generative model itself but the many flavours in which it comes: conditional GANs, cycle GANs, multiple player GANs, style GANs, etc.

5.4 TODO: Normalising Flows

Learn a bunch of bijections between distributions starting with the complex distribution from which we would like to sample and an isotropic Gaussian. Sample from the isotropic Gaussian and run it through the composition of said bijections to obtain a sample from the distribution of interest.

5.5 Diffusion Models

The purpose of diffusion models is to facilitate the generation of samples from complex distributions from which sampling is typically intractable. While this overarching motive is not unique to diffusion models, how a diffusion model learns and how it generates new samples is quite distinct from other well-known generative models like VAEs and GANs (though some ideas are certainly comparable). A diffusion model can be described by its forward (noising) and backward (denoising) processes. Its forward process iteratively noises a sample $\mathbf{x}_0 \in \mathcal{X}$, belonging to the distribution of interest, T-many times to obtain its noised equivalent $\mathbf{x}_T$. Formally, this is done via the Markov process

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t \middle| \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t I\right)$$

where $\beta_1, \ldots, \beta_T \in [0,1]$ are hyperparameters satisfying $\beta_i < \beta_{i+1}$, often referred to as the noise schedule of the model. With an appropriately chosen final time T, these noised equivalents $\mathbf{x}_T$ are akin to random noise sampled from $\mathcal{N}(0, I)$.

Letting $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, we can describe the backward process of a diffusion model again as a Markov process in which we sample some random noise $\mathbf{x}_T$ from $\mathcal{N}(0, I)$ and obtain the $(t-1)$st denoised sample from $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ via

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}$$

where $\sigma_1, \ldots, \sigma_T$ are hyperparameters, ϵ_θ is the diffusion model's denoiser and $\mathbf{z}$ is sampled from $\mathcal{N}(0, I)$. The purpose of the denoiser ϵ_θ, where θ is

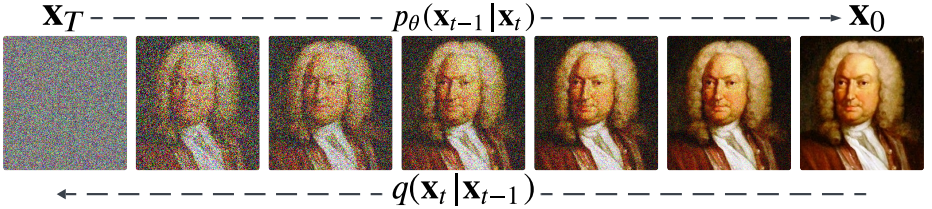


Figure 28: Johann Bernoulli being denoised in line with the backward process of a diffusion model.

its tuple of parameters, is akin to its name: it is used to iteratively turn sampled noise $\mathbf{x}_T \in \mathcal{N}(0, I)$ into something resembling a sample $\mathbf{x}_0 \in \mathcal{X}$ from the distribution of interest. If the architecture of its denoiser can be backpropagated through then training a diffusion model can be done in the usual manner of choosing an appropriate loss function and performing gradient descent in which gradients are computed via backpropagation through the entire model. We leave further details of training a diffusion model out for the sake of brevity but these can easily be found in literature.

So, what is an appropriate choice for the architecture of a diffusion model's denoiser? Before considering transformers for this task, we consider a more often-used choice, U-Net: a class of convolutional neural networks (CNNs).

5.5.1 U-Net denoisers

Despite not being the choice of denoiser made in the seminal paper introducing diffusion models, U-Net became the go to choice of denoiser architecture for contemporary diffusion models. To understand U-Net's popularity in this regard, it is worth understanding its architecture which consists of a contracting branch, terminating at its bottleneck, followed by an expansion branch, illustrated in Figure 29. This architecture can be thought of as forming a 'U'-like shape, with the bottleneck lying at the bottom, hence its name.

In the context of denoising, the contraction branch takes a noised image as input and iteratively completes a series of convolution operations followed by max pooling until reaching the bottleneck layer. At the bottleneck, the model has heavily reduced the spatial dimension of the input image while extracting varying levels of feature abstractions. For example, after the first iteration of convolution and max pooling, the spatial dimension of the input

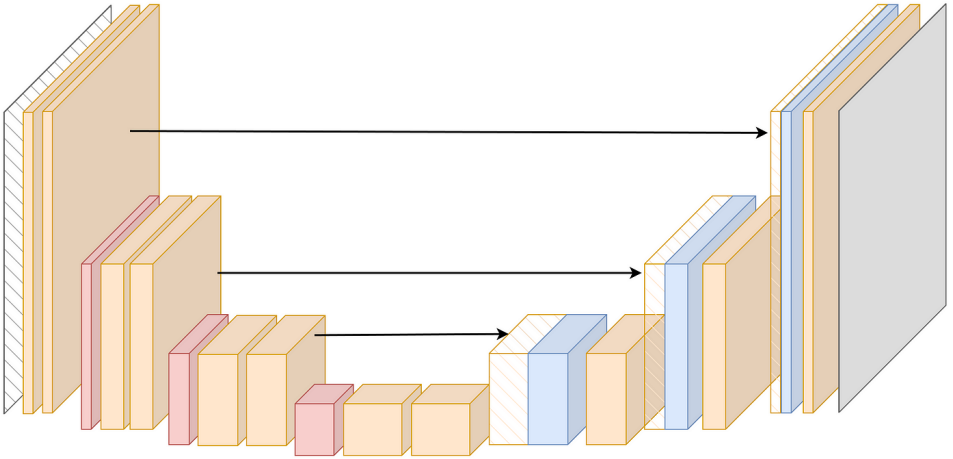


Figure 29: Example U-Net architecture.

image may be reduced from 1024×1024 to 512×512 but edges and textures within the original image may be encoded in the feature maps at this stage. Then, during expansion, the model looks to upscale from the bottleneck in a way that retains the underlying image while removing noise. This is done by iteratively completing a series of deconvolutions, which correspond to upscaling, and using skip connections from its corresponding component in the contraction branch, seen in Figure 29, in order to retain the underlying image. With this in mind, it is clear why U-Net has been the go to choice of denoiser when developing a diffusion model.

5.6 TODO: Evaluating Generative Models

Since samples are not labeled, how does one evaluate a generative model? What does under/overfitting mean for generative models?

Appendices

A Probability Theory Things

Here are some probability theory things, e.g. brief statements, long derivations, and so on, which belong in an appendix. I wish my understanding of probability theory, measure theory, stochastic processes, etc. were better. In another life.

I advise anyone who is entirely new to probability theory or statistics to consider playing around with interative visual demonstrations as in

`seeing-theory.brown.edu/index.html`.

P.S. Sometimes, I'm unsure in which section a topic belongs (probability theory or statistics), e.g. Pearson correlation, so please forgive any disagreeable placements, future me.

A.1 Derivations Related to Common Distributions

TODO: Find a new symbol for the probability p since it's used as a probability function throughout the document.

1) From Bernoulli to Binomial

Let's try to extend the Bernoulli distribution to the binomial distribution. If $X \sim \text{Ber}(p)$ then $\Omega_X = \{0, 1\}$ and $\mathbb{P}(X = x) = p^x(1-p)^{1-x}$.

Taking n independent Bernoulli random variables and summing them yields the random variable $\mathbf{X} = \sum_{i=1}^{n} X_i \sim \text{Bin}(n, p)$ with $\Omega_{\mathbf{X}} = \{0, \dots, n\}$. Let k denote a realisation of $\mathbf{X}$, i.e. $k = x_1 + \cdots + x_n$. The corresponding mass function is given by

$$\begin{aligned}
\mathbb{P}(\mathbf{X} = k) &= Z \cdot \prod_{i=1}^{n} \mathbb{P}(X_i = x_i) \\
&= Z \cdot \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\
&= Z \cdot p^{x_1+\cdots+x_n}(1-p)^{n-(x_1+\cdots+x_n)} \\
&= Z \cdot p^k(1-p)^{n-k}
\end{aligned}$$

where Z is a normalising constant such that $\sum_{k=0}^{n} Zp^k(1-p)^{n-k} = 1$. It's more intuitively seen as the number of ways of placing k-many 1s among a

series of $n \geq k$ bits, i.e. $Z = \binom{n}{k}$, so

$$\mathbb{P}(\mathbf{X} = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

2) From Categorical to Multinomial

The categorical distribution is an extension of the Bernoulli distribution and the multinomial distribution is an extension of the binomial distribution. In line with this, we'd hope to be able to extend the categorical distribution to the multinomial distribution.

If $X \sim \text{Cat}(p_1, \ldots, p_C)$ then realisations of X can be represented by single integers in $\{1, \ldots, C\}$ but I prefer to represent them in terms of their $C-$long one-hot encodings. So I denote the presence of the c^{th} class in a realisation as the unit row vector $\mathbf{e}_c^\top$, e.g. $(0, 1, 0 \ldots, 0)$ denotes a sample in which only the second class is present. The corresponding mass function is then given by

$$\mathbb{P}(X = (x_1, \ldots, x_C)) = \prod_{j=1}^{C} p_j^{x_j}.$$

Note that, in this notation, all but one of these x_j terms are zero, so it really boils down to just a single probability value, e.g. $\mathbb{P}(X = (0, 1, \ldots, 0)) = p_2$. The mass function can also be written using indicator functions but the form offered above is the one I find easiest to use.

Applying the same idea used to extend Bernoulli to binomial, consider the random variable $\mathbf{X} = \sum_{i=1}^{n} X_i$ pertaining to n independent categorical trials. Each realisation of $X_1, \ldots, X_n$ can be represented by a $C-$long one-hot encoded vector and so n realisations of $X \sim \text{Cat}(p)$ can be thought of as a matrix $\mathbf{x} \in \{0, 1\}^{n \times C}$ whose rows are unit vectors in $\mathbb{Z}_{\geq 0}^{C}$. As such, letting $k_1, \ldots, k_C$ denote the number of 1s in the columns of $\mathbf{x}$ (so $k_1 + \cdots + k_C = n$) we see immediately that $(k_1, \ldots, k_C)$ is a realisation of $\mathbf{X}$. So what is the corresponding mass function? Let x_{ij} denote the element in the i^{th} row and j^{th} column of $\mathbf{x}$ and let $k_j = x_{1j} + \cdots + x_{nj}$ denote the sum of all elements in the j^{th} column of $\mathbf{x}$. Note that k_j is the number of realisations of $X_1, \ldots, X_n$

in which the j^{th} class is present. We have

$$\begin{aligned}
\mathbb{P}(\mathbf{X} = (k_1, \ldots, k_C)) &= Z \cdot \prod_{i=1}^{n} \mathbb{P}(X_i = (x_{i1}, \ldots, x_{iC})) \\
&= Z \cdot \prod_{i=1}^{n} \prod_{j=1}^{C} p_j^{x_{ij}} \\
&= Z \cdot \prod_{j=1}^{C} p_j^{x_{1j} + \cdots + x_{nj}} \\
&= Z \cdot \prod_{j=1}^{C} p_j^{k_j} \\
&= Z \cdot p_1^{k_1} \cdots p_C^{k_C}.
\end{aligned}$$

So all that's left to do is derive the normalisation constant Z. Note that for each $j \in \{1, \ldots, C\}$ we know that there are k_j-many 1s in column j so k_1 of these n rows must pertain to the first class for which there are $\binom{n}{k_1}$ possible orderings. From here, see that k_2 of the $n - k_1$ remaining rows must pertain to the second class for which there are $\binom{n-k_1}{k_2}$-many orderings. Continuing this line of reasoning, we obtain

$$\begin{aligned}
Z &= \binom{n}{k_1}\binom{n-k_1}{k_2}\binom{n-(k_1+k_2)}{k_3} \cdots \binom{n-(k_1+\cdots+k_{C-1})}{k_C} \\
&= \frac{n!}{k_1!} \cdot \frac{(n-k_1)!}{k_2!(n-(k_1+k_2))!} \cdot \frac{(n-(k_1+k_2))!}{k_3!(n-(k_1+k_2+k_3))!} \cdots \frac{k_C!}{k_C!0!} \\
&= \frac{n!}{k_1! \cdots k_C!}
\end{aligned}$$

from which we obtain

$$\mathbb{P}(\mathbf{X} = (k_1, \ldots, k_C)) = \frac{n!}{k_1! \cdots k_C!} p_1^{k_1} \cdots p_C^{k_C}.$$

3) Negative Binomial and Geometric

We keep flipping our (maybe biased) coin until we observe r successes (r is a parameter) where individual trials are $\text{Ber}(p)$ distribured. So if $X \sim \text{NB}(r, p)$ then

$$\mathbb{P}(X = k) = Z \cdot p^r (1-p)^k.$$

A trial must be of the form $(x_1, \ldots, x_{k+r-1}, 1)$ which means that $r-1$ of the first $k+r-1$ elements must be a success of which there are $\binom{k+r-1}{r-1}$ possibilities, thus

$$\mathbb{P}(X = k) = \binom{k+r-1}{r-1} p^r (1-p)^k.$$

The case $r = 1$ yields the shifted geometric distribution, whose pmf is given by

$$\mathbb{P}(X = k) = p(1-p)^k.$$

4) From Binomial to Poisson

Let $X_n \sim \text{Bin}(n, p)$ and $\lambda = np$. How might X_n look as $n \to \infty$? Why might we be interested in this at all? It can be thought of as ...

To see what X_n converges to in distribution, note that for all $x \in \{0, \ldots, n\}$

$$\begin{aligned} p(X_n = x) &= \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\ &\approx \frac{n!}{x!(n-x)!} \frac{\lambda^x}{n^x} e^{-\lambda} e^{\lambda x/n} \\ &= e^{\lambda x/n} \frac{n(n-1)\cdots(n-x+1)}{n \cdots n} e^{-\lambda} \frac{\lambda^x}{x!} \\ &\xrightarrow{n\to\infty} e^{-\lambda} \frac{\lambda^x}{x!}. \end{aligned}$$

Showing this statement more rigorously might make use of Stirling's approximation of $n!$. Since X_n is discrete for all $n \in \mathbb{N}$ we conclude that

$$X_n \xrightarrow[d]{n\to\infty} \text{Poi}(\lambda).$$

5) From Poisson to Exponential

When working with a Poisson process, a natural question is the distribution of time between the occurence of events. Let T denote the time at which the first event occurs, so $\Omega_T = [0, \infty)$. See that if $N(t) \sim \text{Poi}(\lambda t)$ is a Poisson process then $p(N(t) = x)$ is the probability of x events occuring in $[0, t]$. We

have

$$\begin{aligned} p(T < t) &= \sum_{x=1}^{\infty} p(N(t) = x) \\ &= 1 - p(N(t) = 0) \\ &= 1 - e^{-\lambda t} \end{aligned}$$

and so $p(t) = \lambda e^{-\lambda t}$.

A.2 Motivating Variance

You want some idea of expected deviation from the mean. So something like $\mathbb{E}[|X - \mathbb{E}[X]|^a]$ where $a \geq 1$ seems natural. Which exponent is most suitable? $a = 2$ turns out to be convenient as it yields additivity for independently distributed random variables. That is, with $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$ we obtain

$$\begin{aligned} &\mathrm{Var}(X + Y) \\ &= \mathbb{E}[(X + Y)^2] - \mathbb{E}[X + Y]^2 \\ &= \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2] - \left(\mathbb{E}[X]^2 + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[Y]^2\right) \\ &= \left(\mathbb{E}[X^2] - \mathbb{E}[X]^2\right) + \left(\mathbb{E}[Y^2] - \mathbb{E}[Y]^2\right) + 2\left(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\right) \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y). \end{aligned}$$

I can't be bothered to write the analogous statement for non-scalar random variables but such a statement holds. Anyyway, if X and Y are independent then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, i.e. $\mathrm{Cov}(X, Y) = 0$, and so

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

It turns out that you do have additivity with cubes of the deviations but not cubes of the absolute values of the deviations.

It's worth noting that there are plenty of contexts in statistics in which different measure of distribution spread are more convenient, e.g. the mean absolute deviation about the median $\mathbb{E}[|X - \mathrm{median}(X)|]$ in the context of robust statistics.

A.3 Flavours of Convergence of Random Variables

Stochastic processes and asmymptotic statistics people love knowing how their sequences of random variables converge, if at all. That is, if $X_1, \ldots, X_n$ are random variables then what can be said about things related to taking the limit of the sequence to infinity?

Convergence in Distribution (Weak)

Given the sequence $X_1, \ldots, X_n$ of random variables with CDFs $F_1, \ldots, F_n$, the sequence of RVs converges in distribution to a random variable X with CDF F if for all $x \in \Omega_X$ at which F is continuous

$$\lim_{n\to\infty} F_n(x) = F(x).$$

The reason that this convergence is often described as weak is that the other two flavours of convergence necessitate convergence in distribution. A nice illustration of convergence in distribution, stolen from Wikipedia, if offered in ?.

It illustrates the convergence

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow{n\to\infty} \mathcal{N}(0, 1/3)$$

due to the central limit theorem where $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} U(-1, 1)$.

A.4 The Law of Large Numbers

If $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. and $\mathbb{E}[\mathbf{X}] < \infty$ then

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \xrightarrow{n\to\infty} \mathbb{E}[\mathbf{X}]$$

with respect to some type of convergence. The law comes in a weak flavour and a stronger one. The weak flavour pertains to convergence in probability and the latter to asmymptotic convergence.

Frankly, the details aren't important for my purposes, so I simply think of it as meaning "The sample mean of a buncha samples is the population mean.". It's useful for justifying certain approaches, e.g. Monte Carlo integration.

A.5 The (univariate) Central Limit Theorem (CLT)

If $X_1, \ldots, X_n$ are i.i.d. and $\mathrm{Var}(X) < \infty$ then

$$\sqrt{n}\left(\frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}[X]\right) \xrightarrow[d]{n\to\infty} \mathcal{N}(0, \mathrm{Var}(X)).$$

It's one of the most important results in probability theory and statistics: it helps explain why the normal distribution appears so often in real world data. I think of it as meaning "The scaled sum of i.i.d. samples will look normal upon repeatedly sampling.".

Its clearest use-case to me is in justifying the modelling of residuals in linear regression as normally distributed.

A.6 Jensen's Inequality

If $f : \mathbb{R} \to \mathbb{R}$ is convex and $\mathbb{E}[\mathbf{X}] < \infty$ then

$$f(\mathbb{E}[\mathbf{X}]) \leq \mathbb{E}[f(\mathbf{X})].$$

If f is concave then

$$f(\mathbb{E}[\mathbf{X}]) \geq \mathbb{E}[f(\mathbf{X})].$$

I've never come across a particularly inciteful proof of it, so I won't include one.

A.7 Entropy and its Friend KL-divergence

The entropy of a distribution can be motivated by the notion of the surprise of (or information learned from) observing samples drawn from it. Given a discrete random variable $\mathbf{X}$, an event $\mathbf{x} \in \Omega_{\mathbf{X}}$ and a surprise function $S : \Omega_{\mathbf{X}} \to [0, \infty)$, the surprise of observing $\mathbf{x}$ is $S(\mathbf{x})$. Before continuing, it's useful to lay out what we want out of our surprise function. Following the use of 'surprising' in day-to-day communication, we want events with low probability to be highly surprising and events with high probability to be less surprising, with some extra conditions. So if $p(\mathbf{x}) = 0.01$ then we want $S(\mathbf{x})$ to be relatively high (strictly speaking it doesn't need to be bounded above) and if $p(\mathbf{x}) = 0.99$ then we want $S(\mathbf{x})$ to be close to 0.

An easy way to achieve this is to take $S(\mathbf{x}) = -\log(p(\mathbf{x}))$ where log denotes the natural logarithm unless stated otherwise. Quickly see that $\log(0.01) = 4.61$ and $\log(0.99) = 0.01$. From here, we define the entropy of the distribution p as the expected surprise

$$H(p) = \mathbb{E}_{\mathbf{X} \sim p}[-\log(p(\mathbf{X}))] = -\sum_{\mathbf{x} \in \Omega_{\mathbf{X}}} p(\mathbf{x}) \log(p(\mathbf{x})).$$

More precisely, Claude Shannon wanted such a surprise function to satisfy three intuitive properties. Firstly, the surprise of an event with probability 1 should be 0. Secondly, the surprise of two independent events occurring

should be the sum of the surprises of the events individually. Thirdly, the surprise of a given event should be higher than the surprise of any less probable event. So, for all $\mathbf{x}_1, \mathbf{x}_2 \in \Omega_{\mathbf{X}}$, we desire

1. $p(\mathbf{x}) = 1 \implies S(\mathbf{x}) = 0$
2. $p(\mathbf{x}_1, \mathbf{x}_2) = p(\mathbf{x}_1) \cdot p(\mathbf{x}_2) \implies S(\mathbf{x}_1, \mathbf{x}_2) = S(\mathbf{x}_1) + S(\mathbf{x}_2)$
3. $p(\mathbf{x}_1) > p(\mathbf{x}_2) \implies S(\mathbf{x}_1) < S(\mathbf{x}_2)$

It's straightforward to see that $S(\mathbf{x}) = -\log(p(\mathbf{x}))$ satisfies these three properties but it turns out that it is unique in satisfying these properties, up to its base.

In literature, entropies are measured in nats (natural units of information) if the natural logarithm is used for their computation and bits if $\log_2$ is used.

Technical note regarding $\mathrm{dom}(S)$

It's clear from the second condition that S is actually a function whose domain is the powerset of $\Omega_{\mathbf{X}}$ but accommodating this detail isn't worth it — the idea being conveyed is clear without.

A.7.1 Kullback-Leibler Divergence (KL-divergence)

Given probability density/mass functions p and q on the same space $\Omega_{\mathbf{X}}$, their Kullback-Leibler divergence (KL-divergence) is given by

$$D_{\mathrm{KL}}(p||q) = \mathbb{E}_{\mathbf{X}\sim p}\left[\log\left(\frac{p(\mathbf{X})}{q(\mathbf{X})}\right)\right].$$

It is often used as a measure of similarity between two distributions: it follows from Gibbs' inequality that it is non-negative and it is often proposed as a loss function when assessing how well a model q encodes an underlying distribution p. In many cases, a loss function is chosen whose minimisation is equivalent to minimising the relevant KL-divergence.

The KL-divergence of two distributions can be expressed in terms of a self-entropy and a cross-entropy as in

$$\begin{aligned} D_{\mathrm{KL}}(p||q) &= \mathbb{E}_{\mathbf{X}\sim p}\left[\log\left(\frac{p(\mathbf{X})}{q(\mathbf{X})}\right)\right] \\ &= \mathbb{E}_{\mathbf{X}\sim p}\left[-\log(q(\mathbf{X}))\right] - \mathbb{E}_{\mathbf{X}\sim p}\left[-\log(p(\mathbf{X}))\right] \\ &= H(p,q) - H(p) \end{aligned}$$

where $H(p,q)$ denotes the cross-entropy between p and q and $H(p)$ denotes the self-entropy of p. It follows that minimising the KL-divergence $D_{\mathrm{KL}}(p,q)$ in q corresponds to minimising the cross-entropy $H(p,q)$.

Example: fitting via cross-entropy/KL-divergence

Suppose we have the geometric distribution with $p = 1/2$ and would like to fit it using a Poisson distribution. That is, we would like to best approximate

$$\begin{aligned} p : \mathbb{N} &\to [0,1] \\ k &\mapsto 2^{-(k+1)} \end{aligned}$$

by finding a nice λ for

$$\begin{aligned} q : \mathbb{N} &\to [0,1] \\ k &\mapsto \frac{e^{-\lambda}\lambda^k}{k!}. \end{aligned}$$

To find a good λ, it makes sense to first come up with a metric for how good a given value is. For this, we can use the cross-entropy of the distributions p and q given by

$$H(p,q) = -\mathbb{E}_{K\sim p}\left[\log(q(K))\right] = -\sum_{k=1}^{\infty} p(k)\log(q(k))$$

which can be seen as the incurred cost of using the model q in place of the true underlying model p. There are a bunch of different ways of describing/interpreting this quantity. The most interesting is perhaps the one related to encodings. Anyway, in our case, if we can find a closed form of this expression in terms of λ then we can look to compute which value of λ it's minimised by. In line with this, we compute

$$\begin{aligned} H(p,q) &= -\mathbb{E}_{K\sim p}\left[\log(q(K))\right] \\ &= -\sum_{k=1}^{\infty} p(k)\log(q(k)) \\ &= -\sum_{k=1}^{\infty} 2^{-k}\left(-\lambda + k\log(\lambda) - \log(k!)\right) \\ &= \lambda\sum_{k=1}^{\infty}\frac{1}{2^k} - \log(\lambda)\sum_{k=1}^{\infty}\frac{k}{2^k} + \sum_{k=1}^{\infty}\frac{\log(k!)}{2^k} \\ &= \lambda - \log(\lambda) + C \end{aligned}$$

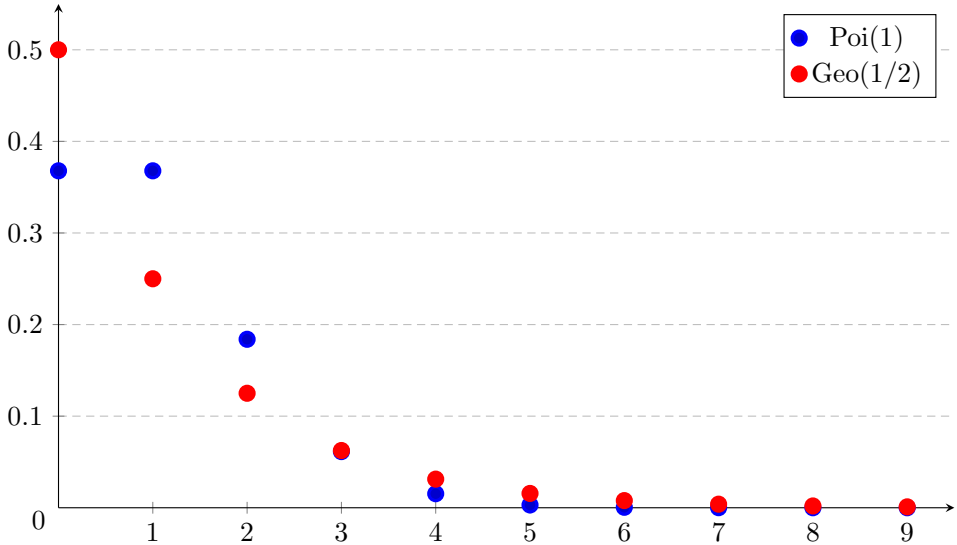


Figure 30: Our geometric distribution p and Poisson fit q.

where $C = \sum_{k=1}^{\infty} \frac{\log(k!)}{2^k}$ is independent of λ. In computing the minmum of $H(p,q)$ in λ we obtain

$$\frac{\partial}{\partial \lambda} H(p,q) = 1 - \frac{1}{\lambda}$$

which yields $\lambda = 1$. So according cross-entropy, the best-fitting Poisson distribution to our geometric distribution is Poi(1).

For $\lambda = 1$ we compute a cross-entropy of

$$H(p,q;1) = 1 + \sum_{k=1}^{\infty} \frac{\log(k!)}{2^k} \approx 2.01567.$$

On its own, this quantity isn't too useful. To get a sense of goodness-of-fit we desire the KL-divergence, i.e. the cross-entropy minus the self-entropy

of p. Let's compute said self-entropy:

$$\begin{aligned} H(p) &= -\sum_{k=1}^{\infty} p(k)\log(p(k)) \\ &= -\sum_{k=1}^{\infty} 2^{-k}\log(2^{-k}) \\ &= \log(2)\sum_{k=1}^{\infty}\frac{k}{2^k} \\ &= 2\log(2) \\ &\approx 1.386 \end{aligned}$$

from which we know that the KL-divergence between the underlying geometric distribution p and our Poisson fit q is given by

$$\mathrm{KL}(p||q) \approx 2.016 - 1.386 = 0.63.$$

A.8 Pearson Correlation and Mutual Information

If X and Y are scalar random variables then their Pearson correlation is given by

$$\rho_{X,Y} = \frac{\mathbb{E}\left[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\right]}{\sqrt{\mathbb{E}\left[(X-\mathbb{E}[X])^2\right]\mathbb{E}\left[(Y-\mathbb{E}[Y])^2\right]}} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y} \in [-1,1]$$

where the expectation in the numerator is over their joint $p_{(X,Y)}$. The overall quantity yields the linear dependence between X and Y and the denominator is for normalisation.

What if we want an idea of the non-linear dependence between random variables? This desire motivates mutual information. The mutual information of the joint over $\mathbf{X}$ and $\mathbf{Y}$ is often spoken of as a property of two random variables $\mathbf{X}$ and $\mathbf{Y}$ and is denoted as such as in

$$I(\mathbf{X};\mathbf{Y}) = D_{\mathrm{KL}}(p_{\mathbf{X},\mathbf{Y}}||p_{\mathbf{X}}\otimes p_{\mathbf{Y}}).$$

I'm massively in favour of writing it as a KL-divergence as its meaning is immediately clear: its a measure of divergence between the joint and the product of the marginals. Informally, I see it as asking "How well does the product of marginals model the joint?" which is precisely the question we seek to answer when utilising things like Pearson correlation and mutual information.

A.9 Quality of fit ≈ Encoding quality?

TODO: Perhaps worthy of its own section, unsure

B Statistics Things

Here are some statistics things.

B.1 Sample Independence and Terminology

The intended meaning of the term 'sample' in statistics is a set of independently obtained realisations of a random variable, i.e. $\{x_1, \ldots, x_n\} \subset \Omega_X$. You could also define a sample as a set of i.i.d. random variables $\{X_1, \ldots, X_n\}$. There are times where I reason with the latter but most of the time I think of realisations of random variables when talking about a sample. As a term, 'sample' can be especially confusing when context switching since in machine learning contexts, it most often refers to a single data point, which is not its intended meaning in statistics contexts. This was a huge cause of confusion during my first few meetings with my master's thesis supervisor who is a Bayesian statistician.

As for sample independence, I was think of the following simple example. Suppose we'd like to estimate the population mean of the heights of men and women. If we randomly sample 100 men and 100 women separately then the sample mean serves as a fine estimate. This could computed by independently sampling, without replacement, from the categorical distribution which corresponds to the ID numbers of all individuals. If we instead randomly sampled 100 couples then the obtained samples would not be independent as couples' heights correlate.

Sample independence is an important assumption for estimation, e.g. for maximum likelihood estimation. It's also important for other statistics-related things like t-tests, ANOVA, etc. but my classical statistics is bad so I won't pretend to know much about it.

B.2 Estimators and their Bias

Suppose you have a bunch of i.i.d. random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$, an estimator is a function $\hat{\theta} : \Omega_{\mathbf{X}}^n \to \mathbb{R}^k$. Given the sample $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \Omega_{\mathbf{X}}$, the value $\hat{\theta}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is typically denote by $\hat{\theta}$ for shorthand and is referred to as an estimate. The reason that this lack of rigour in notation is allowable is

that the contexts in which it is used often make it clear when an estimator is being considered as opposed to an estimate, and vice versa.

The simplest example of a often-used estimator is the estimator of the mean as in

$$\hat{\mu}(\mathbf{X}_1, \ldots, \mathbf{X}_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i.$$

In this case, the dimension of the codomain of $\hat{\mu}$ is simply the dimension of the sample space of $\mathbf{X}$. Given a sample $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, an estimate of the mean of the corresponding distribution is given by $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$.

Given that estimators are simply statistics used to approximate some quantity of the corresponding distribution, a natural question is how to measure the goodness of an estimator. The most well known such measure is the bias of an estimator in estimating a quantity θ given by

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

where the expectation is with respect to $p_{\mathbf{X}}^{\otimes n}$. The need for the precise terminology 'in estimating a quantity θ' is because one could, for example, use the sample mean as an estimator of both the population mean and the population variance (as ridiculous as this choice may be for the latter). As an estimator of the population mean, the sample mean is unbiased. Not so much for the latter. That said, an estimator being unbiased does not necessitate that it is 'good'. For example, if you had $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$ then X_1 is an unbiased estimator of p. To follow this, there are strange special cases in which every unbiased estimator of a quantity is worse than many biased estimators.

As an example of computing the bias of an estimator, see that the sample mean is an unbiased estimator of the population mean μ as

$$\begin{aligned}
\text{Bias}(\hat{\mu}) &= \mathbb{E}[\hat{\mu}] - \mu \\
&= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i\right] - \mu \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\mathbf{X}_i] - \mu \\
&= 0.
\end{aligned}$$

It turns out that the most natural estimator of population variance σ^2, given by

$$\frac{1}{n} \sum_{i=1}^{n} (\mathbf{X}_i - \hat{\mu})^2,$$

is biased. This can be seen after a bunch of tedious computations which yield a bias of $-\sigma^2/n$. To account for this, an unbiased estimator of the population variance is given by

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \hat{\mu})^2 .$$

Most of the time, 'sample variance' refers to the unbiased estimator.

B.2.1 Sampling bias

For a given estimator, sampling bias is not its bias as an estimator but the bias induced by samples being drawn from a distribution $q_{\mathbf{X}}^{\otimes n}$ which is distinct to the true distribution $p_{\mathbf{X}}^{\otimes n}$ attributed to $\mathbf{X}$ as in

$$\mathbb{E}_{q_{\mathbf{X}}^{\otimes n}}[\hat{\theta}] - \theta.$$

B.3 TODO: Hypothesis Testing

B.4 TODO: Markov Chains

B.5 TODO: Bayes' Theorem

B.6 The Expectation-Maximisation (EM) Algorithm

Maximum likelihood estimation (MLE) is great but what do we do when the data generating process is not determined entirely by observed model variables $\mathbf{X} = (X_1, \ldots, X_m)$. That is, what if $p(\mathbf{x})$ is simply the marginal of the true joint $p(\mathbf{x}, \mathbf{z})$ where $\mathbf{Z} = (Z_1, \ldots, Z_{m'})$ are hidden variables?

It turns out that there's a clever way of perfmorming MLE while accounting for hidden variables called the expectation-maximisation (EM) algorithm. It involves obtaining a lower bound for the log-likelihood of a given sample of the observed variables and instead maximising the lower bound in the model's parameters. If certain conditions are met for said lower bound to in fact reach equality with the sample's log-likelihood then the application of EM is MLE itself. If the bound is not exact then we refer to the process as variational EM.

In obtaining a lower bound, let θ denote the parameters of the true joint

$p_\theta(\mathbf{x}, \mathbf{z})$ and see that if q is a probability mass function with domain $\Omega_{\mathbf{Z}}$ then

$$\begin{aligned}
\log(p_\theta(\mathbf{x})) &= \log\left(\sum_{\mathbf{z}\in\Omega_{\mathbf{Z}}} p_\theta(\mathbf{x}, \mathbf{z})\right) \\
&= \log\left(\sum_{\mathbf{z}\in\Omega_{\mathbf{Z}}} q(\mathbf{z})\frac{p_\theta(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})}\right) \\
&= \log\left(\mathbb{E}_{\mathbf{Z}\sim q}\left[\frac{p_\theta(\mathbf{x}, \mathbf{Z})}{q(\mathbf{Z})}\right]\right) \\
&\geq \mathbb{E}_{\mathbf{Z}\sim q}\left[\log\left(\frac{p_\theta(\mathbf{x}, \mathbf{Z})}{q(\mathbf{Z})}\right)\right] \\
&= \mathbb{E}_{\mathbf{Z}\sim q}\left[\log\left(p_\theta(\mathbf{x}, \mathbf{Z})\right)\right] + H(q) \\
&=: \mathrm{ELBO}(q, \theta)
\end{aligned}$$

in which 'ELBO' stands for evidence lower bound (swap the sums for integrals if you'd like continuously distributed hidden variables). Note that this is not the same ELBO dervied in motivating variational autoencoders (VAEs).

The EM algorithm performs coordinate ascent (component-wise equivalent of gradient ascent) on the ELBO: iteratively maximising $\mathrm{ELBO}(q, \theta)$ in q (with θ fixed) and then in θ (with q fixed) until some convergence crtieria is met. With q fixed, maximising the ELBO in θ amounts to computing

$$\begin{aligned}
\arg\max_\theta \mathrm{ELBO}(q, \theta) &= \arg\max_\theta \mathbb{E}_{\mathbf{Z}\sim q}\left[\log\left(p_\theta(\mathbf{x}, \mathbf{Z})\right)\right] \\
&= \arg\max_\theta \left(\mathbb{E}_{\mathbf{Z}\sim q}\left[\log\left(p_\theta(\mathbf{x}|\mathbf{Z})\right)\right] + \mathbb{E}_{\mathbf{Z}\sim q}\left[\log\left(p_\theta(\mathbf{Z})\right)\right]\right) \\
&= \arg\max_\theta Q(\theta; q)
\end{aligned}$$

where $Q(\theta; q) = \mathbb{E}_{\mathbf{Z}\sim q}\left[\log\left(p_\theta(\mathbf{x}|\mathbf{Z})\right)\right] + \mathbb{E}_{\mathbf{Z}\sim q}\left[\log\left(p_\theta(\mathbf{Z})\right)\right]$. Note that equality is offered by the bound precisely when $q(\mathbf{z}) = p_\theta(\mathbf{z}|\mathbf{x})$ and so, with θ fixed, maximising the ELBO in q amounts to computing

$$\arg\max_q \mathrm{ELBO}(q, \theta) = p_\theta(\mathbf{z}|\mathbf{x}),$$

i.e. computing the posterior.

With both statements in mind, the EM algorithm amounts to initialising the parameters $\theta^{(0)}$ and repeating the following two steps from $t = 0$ until stopping criteria is met:

1. (E step: $\theta^{(t)} \to q_t$) Obtain a closed form for $q_t(\mathbf{z}) = p_{\theta^{(t)}}(\mathbf{z}|\mathbf{x})$ and in turn a closed form for $Q(\theta; q_t)$
2. (M step: $q_t \to \theta^{(t+1)}$) Compute $\theta^{(t+1)} = \arg\max_\theta Q(\theta; q_t)$

Variational EM

Computing the posterior exactly is often infeasible in practice. Varational EM seeks to alleviate this by instead restricting q to a family of functions $\mathcal{F}$, e.g. some family of MLPs, and replacing the E-step with the computation

$$q_t = \arg\max_{q \in \mathcal{F}} \text{ELBO}\left(q, \theta^{(t)}\right).$$

Of course, if the posterior $p_{\theta^{(t)}}(\mathbf{z}|\mathbf{x})$ belongs to $\mathcal{F}$ for all relevant $t \in \mathbb{N}$ then variational EM is simply EM.

Regarding the correctness of the EM algorithm, we seek to show that

$$\log(p_{\theta^{(t+1)}}(\mathbf{x})) \geq \log(p_{\theta^{(t)}}(\mathbf{x}))$$

for all $t \in \mathbb{N}$, i.e. individual steps in the loss landscape are neglible at worst and otherwise increase the likelihood of the observed data. Showing this turns out to be elegant, as in

$$\log(p_{\theta^{(t+1)}}(\mathbf{x})) \geq \text{ELBO}\left(q_t, \theta^{(t+1)}\right) \geq \text{ELBO}\left(q_t, \theta^{(t)}\right) = \log(p_{\theta^{(t)}}(\mathbf{x})).$$

The leftmost inequality is due to how the ELBO is defined. The second inequality is seen from

$$\theta^{(t+1)} = \arg\max_\theta \text{ELBO}\left(q_t, \theta\right).$$

Finally, the equality follows from the 'matching the true posterior with q_t' argument offered earlier, i.e.

$$q_t(\mathbf{z}) = p_{\theta^{(t)}}(\mathbf{z}|\mathbf{x}) \implies \text{ELBO}\left(q_t, \theta^{(t)}\right) = \log(p_{\theta^{(t)}}(\mathbf{x})).$$