Lecture Notes

PROBABILISTIC GENERATIVE MODELS

This version: 11th January 2026

Latest version: github.com/felipe-tobar/Probabilistic-Generative-Models

Felipe Tobar
Department of Mathematics
Imperial College London

f.tobar@imperial.ac.uk
https://www.ma.ic.ac.uk/~ft410/

Preface

This notes are under development for 2026.

Felipe Tobar,
London,
January 2026.

Contents

Week 1

Foundations

1.1 Introduction

A Probabilistic generative model (PGM), or simply, a GM, is a methodology for generating data. In general, the PGM is constructed and adjusted using observations with the aim to synthesise samples with the same statistical properties of the available observations. The *probabilistic* nature of the PGMs studied in this course follows from the fact that the available data will be considered to be realisations of an underlying random variable (RV), e.g., X.

In this sense, the probability distribution of X, denoted $P_X(x)$, as well as its probability density function (pdf) $p_X(x)$ will be central to the study of PGMs. In particular, targetting the pdf is one way of constructing PGMs, in which case the whole PGM paradigm becomes equivalent to the classical statistical modelling approach. However, as we will see in the course, enforcing the sought-after PGM to have an explicit parametric pdf can be rather restrictive.

Throughout the course, we will consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with 3 RVs given by the following measurable maps:

$$\begin{array}{ccc} X : \Omega \to \mathcal{X} & Y : \Omega \to \mathcal{Y} & Z : \Omega \to \mathcal{Z} \\ \text{(observed input)} & \text{(observed output)} & \text{(latent variable)} \end{array}$$

Remark 1.1.

Not all three RVs will be present in all our settings. For instance, in classification there is no justification for the latent variable Z (in general), while in clustering, there is no need for X. However, we build the general setup here for formality.

We will also consider the σ-algebra $\mathcal{F}$ to be the product Borel σ-algebra, that is, $\mathcal{F} = \mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{Y}) \otimes \mathcal{B}(\mathcal{Z})$. This is the smallest σ-algebra making the joint random variable (X, Y, Z) measurable.

Furthermore, we will assume that the joint probability of (X, Y, Z) has a density, that is, $\forall A \in \mathcal{B}(\mathcal{X}), B \in \mathcal{B}(\mathcal{Y}), C \in \mathcal{B}(\mathcal{Z})$, we have

$$\mathbb{P}\left(X \in A, Y \in B, Z \in C\right) = \int_{A\times B\times C} p(x,y,z)\mathrm{d}x\mathrm{d}y\mathrm{d}z. \tag{1.1}$$

We will also assume that all marginals and conditional have a density. This includes $p(x,y)$, $p(y|x)$, $p(z|x,y)$, etc.

1.2 Discriminative versus generative

The generative approach aims to characterise the complete generative distribution $p(x,y,z)$, whereas, in some application-specific cases, only the discriminative model, e.g., $p(y|x)$, is needed. Let us examine the following example.

Example 1.1 (Generative and discriminative views of binary classification)**.**

Consider the binary classification problem, where, given an observation $X = x$, one needs to estimate its label Y. A discriminative model would directly parametrise $\mathbb{P}(Y|X = x)$. Since this is a binary classification case, without loss of generality, we can assume $Y \in \{0,1\}$, and model $\mathbb{P}(Y = 1|X = x)$, since $\mathbb{P}(Y = 0|X = x) = 1 - \mathbb{P}(Y = 1|X = x)$. A model for this probability only needs to map $x \in \mathbb{R}^d \rightarrow \mathbb{P}(Y = 1|X = x) \in [0,1]$. For instance, a reasonable candidate for this is

$$\mathbb{P}(Y = 1|X = x) = \frac{1}{1 + e^{-\theta^\top x}}, \tag{1.2}$$

which is known as the logistic regression.

Conversely, in a generative approach, we aim to model the joint probability $p(Y = y, X = x)$. Modelling this distribution is not easy, however, observe that we can factorise it as

$$p(Y = y, X = x) = p(X = x|Y = y)p(Y = y), \tag{1.3}$$

which yields a pair of much more intuitive distributions to model:

- the class probability $p(Y = y) = (\pi, 1-\pi), \pi \in [0,1]$, and
- the class-conditional probability $p(X = x|Y = y)$, given by a two distributions over $\mathcal{X}$, denoted f_{θ_0} and f_{θ_1}.

Therefore, the classifier is

$$\begin{aligned} p(Y = 1|X = x) &= \frac{p(X = x|Y = 1)p(Y = 1)}{p(X = x)} \\ &= \frac{1}{1 + e^{-\log\left(\frac{\pi}{1-\pi}\frac{f_{\theta_1}}{f_{\theta_0}}\right)}}. \end{aligned} \tag{1.4}$$

Exercise 1.1.

Evaluate eq. (1.4) for $f_{\theta_0} = \mathcal{N}(\mu_0, \Sigma_0)$ and $f_{\theta_1} = \mathcal{N}(\mu_1, \Sigma_1)$. What happens when $\Sigma_0 = \Sigma_1$?

1.3 The pushforward measure

Despite the abundant collection of well-studied statistical models, in some scenarios we can construct a more ad hoc model by applying an appropriate transformation.

Definition 1.1.

Consider a RV $X \in \mathcal{X}$ with measure P_X, and a nonlinear map $T : \mathcal{X} \to \mathcal{X}$. The measure of the transformed RV $T(X)$ is known as the *push forward measure* of P_X through T, and it is denoted by $T_\# P_X$

Remark 1.2.

The transformations considered in the course will be such that the pushforward measure has a density. With a slight abuse of notation, we will denote this density as $T_\# p_X$.

Example 1.2 (Discrete pushforward)**.**

Let X be a discrete random variable taking values in $\{1, 2, 3\}$ with

$$\mathbb{P}(X = 1) = 0.2, \quad \mathbb{P}(X = 2) = 0.5, \quad \mathbb{P}(X = 3) = 0.3.$$

Define the map $T : \{1, 2, 3\} \to \{a, b\}$ by

$$T(1) = a, \qquad T(2) = b, \qquad T(3) = b.$$

Then the pushforward $T_\# \mathbb{P}$ satisfies

$$(T_\# \mathbb{P})(\{a\}) = 0.2, \qquad (T_\# \mathbb{P})(\{b\}) = 0.8.$$

For an illustration see Fig. 1.1.

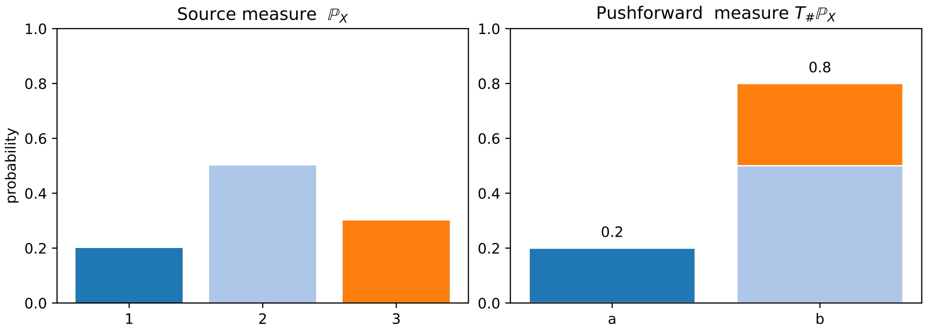

Figure 1.1: Source and pushforward distributions: discrete example.

Example 1.3 (Continuous pushforward)**.**

Let $X \sim \mathcal{N}(0,1)$ on $\mathbb{R}$ and define $T(x) = x^2$. The pushforward $T_{\#}\mathbb{P}$ is the law of $Y = T(X)$, supported on $\mathbb{R}_+$. Its density is given by

$$p_Y(y) = \frac{1}{\sqrt{2\pi y}} \exp\left(-\frac{y}{2}\right), \qquad y > 0.$$

For an illustration of this case and other related examples, see Fig. 1.2.

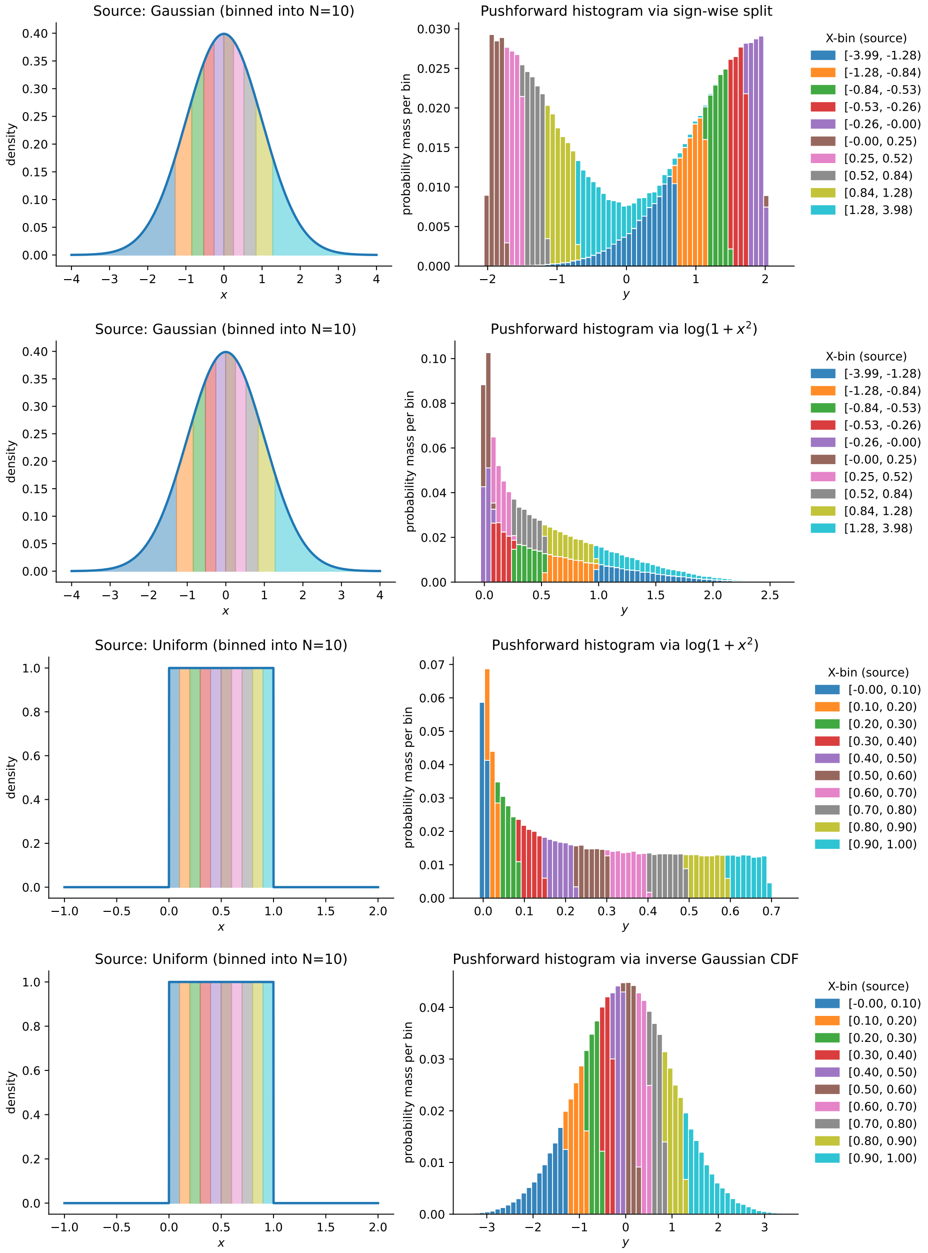

Figure 1.2: Source and target distributions: continuous examples

In general, for arbitrary source distributions and maps it is difficult to compute the target density in closed form, at least in the continuous case. For the specific case of differentiable and invertible maps T, the following theorem given a recipe to compute $p_{T(X)}$

Theorem 1.1 (Change of variable)**.**
Consider two RVs $X, Y \in \mathbb{R}^d$, such that $Y = T(X)$, where $T : \mathbb{R}^d \to \mathbb{R}^d$ is a C^1 diffeomorphism. If X and Y have densities p_X and p_Y respectively, then

$$p_Y(y) = p_X(T^{-1}(y)) \left|\det \nabla_y T^{-1}(y)\right|, \tag{1.5}$$

where $\nabla_y T^{-1}(y)$ is the Jacobian of the inverse map.

Remark 1.3.
Though the above result provides a closed-form expression for the pushforward measure only when T is a C^1 diffeomorphism (continuously differentiable with an the inverse having the same property), we can transform a source RV X into a target RV T with any measurable map. This is because

$$\mathbb{P}(T \in A) = \sum_{T(B_i)=A} \mathbb{P}(X \in B_i). \tag{1.6}$$

Though in general the pdf of T will not be available in closed form.

1.4 Likelihood-based training

Maximum likelihood (ML) is going to be the canonical methodology for training our PGMs, and, as we will see next, it will recover other forms of training criteria in particular cases.

Consider a PGM for the RV Y, with density $p_\theta(y)$, where $\theta \in \Theta$ denotes the model parameter. Also, consider the realisations of Y given by $y_1, y_2, \ldots, y_n$.

Definition 1.2 (Likelihood function)**.**
The likelihood of the parameter θ is the function $L : \Theta \to \mathbb{R}_+$ given by the probability density function of Y evaluated on the observations. That is,

$$L(\theta) = p_\theta(y_1, y_2, \ldots, y_n). \tag{1.7}$$

NB: We abused notation above stating the joint pdf for the observations.

Remark 1.4.
Very important: the likelihood function is not a probability/density function, as it is a function of the parameter. In particular, it is not true that $\int_\Theta L(\theta) \mathrm{d}\theta$ is one.

Definition 1.3 (Maximum likelihood estimator)**.**

The ML estimator is given by

$$\theta_{ML} = \arg\max L(\theta). \tag{1.8}$$

Remark 1.5.

In general (but, importantly, not always) we will consider i.i.d observations, in which case the likelihood factorises as $L(\theta) = \prod_{i=1}^{n} p_\theta(y_i)$. Furthermore, when optimising the likelihood we will consider the log-likelihood instead; in the i.i.d. case, this is

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log p(y_i). \tag{1.9}$$

Example 1.4 (Gaussian linear regression)**.**

Let us consider the PGM given by

$$Y|x \sim \mathcal{N}(ax, \sigma^2), a, x \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+. \tag{1.10}$$

This is equivalent to $Y = ax + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$. The parameters in this setting are $\theta = (a, \sigma^2)$. Now consider the observations $\{(x_i, y_i)\}_{i=1}^{n}$.

Since $p(y_i|x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2}(y_i - ax_i)^2\right)$, we can write the log-likelihood as

$$l(\theta) = \sum_{i=1}^{n} \frac{-1}{2} \log 2\pi\sigma^2 + \frac{1}{2\sigma^2}(y_i - ax_i)^2. \tag{1.11}$$

The optimal (a, σ^2) can be found in closed form using the first order optimality conditions.

Remark 1.6.

Observe that optimising eq. (1.11) recovers the least squares solution.

Example 1.5 (Binary classification)**.**

Consider observations $\{(x_i, y_i)\}_{i=1}^{n} \subset \mathbb{R}^d \times \{0, 1\}$ from a binary classification setting. Model the classifier as

$$p_\theta(y = 1|x) = \sigma(s(x)), \tag{1.12}$$

where $\sigma(s(x)) = \frac{1}{1+e^{-s(x)}}$, and $s : \mathbb{R}^d \top \mathbb{R}$ is a feature extractor (e.g., $s(x) = a^\top x + b$). Assuming that the observations are i.i.d., we have

$$L(\theta) = \prod_{i=1}^{n} p(y_i|x_i) = \prod_{i=1}^{n} \sigma(s(x_i))^{y_i}(1 - \sigma(s(x_i)))^{1-y_i}, \tag{1.13}$$

and equivalently

$$l(\theta) = \sum_{i=1}^{n} y_i \log \sigma(s(x_i)) + (1 - y_i) \log(1 - \sigma(s(x_i))). \tag{1.14}$$

Does this expression seem familiar? If not, we will find out soon what this is.

Example 1.6 (Clustering)**.**

Consider a set of observations $\{x_i\}_{i=1}^n \in \mathbb{R}^d$ and implement a clustering algorithm. We will assume that there are $K \in \mathbb{N}$ clusters, each specified by a density $p_k, k = 1, \ldots, n$; this means that the probability of a sample x coming from the k-th cluster is $\mathbb{P}(x \in C_k) = \pi_k$, where $\forall k, 0 \geq \pi_k \geq 1$ and $\sum_{k=1}^K \pi_k = 1$.

This is a mixture model, with density $p(x) = \sum_{k=1}^K \pi_k p_k(x)$, and parameters given by the cluster probabilities π_k and the parameters of the densities $p_k = p_{\theta_k}$. The log-likelihood is

$$l(\theta) = \sum_{i=1}^n \log \sum_{k=1}^K \pi_k p_k(x_i) \tag{1.15}$$

Note that there are two issues associated to optimising eq. (1.15).

- We do not recover the cluster assignments.
- The problem is ill-posed. E.g., if $p_k = \mathcal{N}(\mu_k, \Sigma_k)$, which is the usual choice, we can set $\mu_k = x_i, \Sigma_k = 0$ which gives $l = \infty$.

We can overcome this drawback by introducing a collection of latent random variables Z_{nk}, that represents the cluster assignments. That is,

$$Z_{nk} = 1 \iff x_n \in C_k. \tag{1.16}$$

This allows us to write the conditional densities $p(x_n|z_{nk}) = \prod_{k=1}^K p_k^{z_{nk}}$, and thus to express the **complete-data likelihood** given by

$$l(\theta) = \log \prod_{n=1}^N \prod_{k=1}^K p_k^{z_{nk}}(x_n) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \log p_k(x_n). \tag{1.17}$$

Good and bad news: this objective is now theoretically feasible to optimise but impractical since we do not have access to the latent cluster assignments $\{z_{nk}\}_{nk}$.

A workaround to this is to estimate the cluster assignments, via its conditional expectation wrt the observations. That is,

$$\mathbb{E}\left(z_{nk}|x_{1:N}\right) = 1 * \mathbb{P}(z_{nk} = 1|x_n) + 0 * \mathbb{P}(z_{nk} = 0|x_n) = \mathbb{P}(z_{nk} = 1|x_n), \tag{1.18}$$

which can be computed explicitly using Bayes theorem in terms of the model parameters. Then, we can perform and iterative procedure by: i) optimising $l(\theta)$ using $\mathbb{E}\left(z_{nk}|x_{1:N}\right)$, and ii) computing $\mathbb{E}\left(z_{nk}|x_{1:N}\right)$ using $\theta_{ML} = \arg\max l(\theta)$.

This means that exact ML cannot be performed in this case. Also, does this procedure seem familiar?

The maximum likelihood estimator (MLE) satisfies several important theoretical properties:

- **Consistency:** Under the assumption that the statistical model is *identifiable*—i.e., different parameter values correspond to different probability distributions—the MLE converges to the true parameter as the number of observations grows. Intuitively, maximising the likelihood asymptotically minimises the Kullback–Leibler divergence between the true distribution and the distribution induced by a candidate parameter.

- **Equivariance:** If $\hat{\theta}_{\text{MLE}}$ is the MLE of θ, then for any transformation g, the MLE of $g(\theta)$ is $g(\hat{\theta}_{\text{MLE}})$. This property allows us to compute MLEs under reparametrisations directly.

- **Asymptotic normality:** For large sample sizes, the MLE is approximately normally distributed around the true parameter with covariance matrix given by the inverse Fisher information. Formally,

$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta) \ \overset{d}{\to}\ \mathcal{N}(0, I(\theta)^{-1}),$$

  where $I(\theta)$ is the Fisher information matrix.

- **Asymptotic efficiency:** As a consequence of asymptotic normality, the MLE achieves the Cramér–Rao lower bound for the variance in the limit of large n, making it asymptotically optimal among unbiased estimators.

In practice, these properties justify the widespread use of the MLE: it not only converges to the true parameter under mild assumptions, but also allows for straightforward reparametrisations and provides an estimator with minimal asymptotic variance.

1.5 A brief intro to information theory

Motivation. Let us consider a discrete RV $X \in \{1, 2, \ldots, K\}$ with pmf p_X. Observe that $-\log p_X(x)$ represents a measure of *information* gained from obtaining the value a as a sample of X. Now consider a communication channel $A \to B$, where A is transmitting samples of X to B. When B received the samples, its *average information* can be expressed as

$$H(X) = -\sum_{x=1}^{K} p_X(x) \log p_X(x). \tag{1.19}$$

This quantity is known as *entropy* and—in connection with thermodynamics—it represent a measure of disorder or un-predictability of X.

NB: We will use $H(X)$ and $H(p_X)$ interchangeably.
NB: We will usually denote $H(X) = -\mathbb{E}\left(\log p_X\right) = \mathbb{E}\left(\log \frac{1}{p_X}\right)$.

Clearly, $H(X) \geq 0$ with equality achieved for an RV that has always the same outcome with $p_X(x) = 1$. This is the deterministic, predictable, case. To revise further properties, let us recall the following result.

Jensen's Inequality Let $\phi : \mathbb{R} \to \mathbb{R}$ be a *convex* function and let X be a random variable such that $\mathbb{E}[|X|] < \infty$. Then

$$\phi\Big(\mathbb{E}[X]\Big) \;\le\; \mathbb{E}\Big[\phi(X)\Big]. \tag{1.20}$$

since $\log(\cdots)$ is *concave*, the inequality is reversed and we have:

$$\mathbb{E}[\log X] \;\le\; \log \mathbb{E}[X]. \tag{1.21}$$

Remark 1.7.

Equality in eq. (1.20) is only achieved when either ϕ is affine, or X is constant almost surely, that is, $\mathbb{P}\left(X = c\right) = 1$. As a consequence, equality in eq. (1.21) is only achieved when X is constant almost surely (when the argument of the logarithm does not depend on x)

Keep this result in mind, as it will be used throughout the module.

Using Jensen on the definition of the entropy, we have

$$H(X) = \mathbb{E}\left(\log \frac{1}{p}\right) \le \log \mathbb{E}\left(\frac{1}{p}\right) = \log \sum \frac{1}{p} p = \log K. \tag{1.22}$$

This directly implies that the uniform distribution over $\{1, 2, \ldots, K\}$ has the largest entropy, since

$$H(U_{1:K}) = \mathbb{E}\left(\log \frac{1}{1/K}\right) = \log K. \tag{1.23}$$

Example 1.7 (Bernoulli distribution)**.**

Consider $X \sim p_X(x) = \theta^x(1-\theta)^{1-x}, \theta \in [0,1]$. The entropy is given by $H(X) = -\theta \log \theta + (1-\theta)\log(1-\theta)$. Figure 1.3 shows this function.

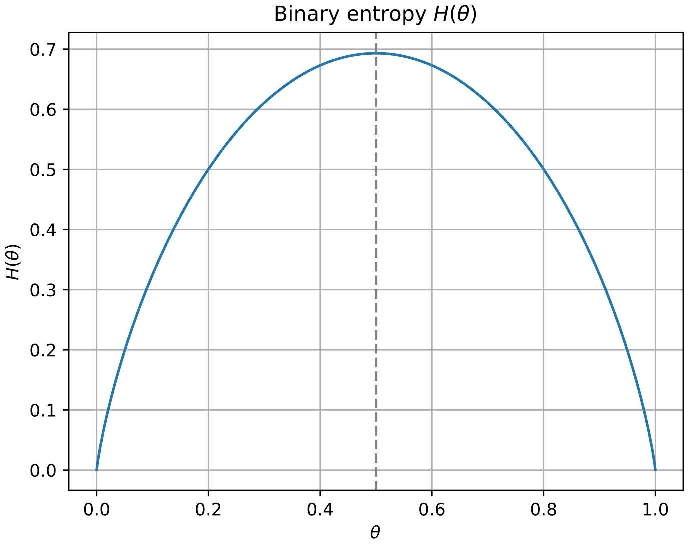

Figure 1.3: Entropy for Bernoulli.

The entropy, in addition to being a measure of disorder, can be understood as the cost of the optimal for of compression. Here, think of a compression strategy using symbols $s_1, s_2, \dots$ with increasing size (or storage cost). For instance, think of storage using logical gates, meaning that these symbols are (equivalent to)

$$s_1 = 0, s_2 = 1, s_3 = 10, s_4 = 11, \dots \tag{1.24}$$

where storing more and more symbols becomes increasingly expensive. The compression strategy is then to assign each outcome of X with a symbol s_i. Intuitively, the optimal compression would assign s_i to the i-th most frequent value in $\{1, 2, \dots, K\}$, meaning that the *cost of storage* precisely grows precisely with $\log \frac{1}{p_X}$.

As a consequence, the average message size is the entropy $H(X)$, and thus can be understood as the cost of this compression strategy.

Now let us go back to our communication channel $A \to B$, where the receiver in B now mistakenly believes that $X \sim q_X$. B's estimated entropy, or averaged information, would be

$$H_{CE}(p_X, q_X) = -\sum_{x=1}^{K} p_X(x_i) \log q_X(x_i). \tag{1.25}$$

Following the same rationale as above, this can be interpreted as the cost of compressing the sequence $x_1, x_2, x_3, \dots$ using q_x.

This quantity is known as the cross-entropy between p_X and q_X. Note that this quantity is not symmetric.

It is relevant to study how $H_{CE}(p_X, q_X)$ and $H(P) = H_{CE}(p_X, p_X)$ relate to one another, in particular if one of them is (always) larger than the other.

Let us see:

$$\begin{align}
H(p) - H_{CE}(p_X, q_X) &= \sum_x p(x) \log \frac{q(x)}{p(x)} \tag{1.26}\\
&\overset{\text{Jensen's}}{\leq} \log \sum_x \cancel{p(x)} \frac{q(x)}{\cancel{p(x)}} \tag{1.27}\\
&= \log 1 = 0. \tag{1.28}
\end{align}$$

Therefore, $H(p) \leq H(p, q)$, with equality only achieved when $p = q$, as per Remark 1.7.

Remark 1.8.

Minimising the correntropy wrt to one of its arguments is precisely an attempt to match $p = q$.

The use of the entropy/correntropy that is going to be more relevant in our case is via its application to continuous RVs. This extension is

$$H(p) = -\int_{\mathcal{X}} p(x)\log p(x)\mathrm{d}x \tag{1.29}$$

$$H(p,q) = -\int_{\mathcal{X}} p(x)\log q(x)\mathrm{d}x \tag{1.30}$$

$$\tag{1.31}$$

Remark 1.9.

Unlike its discrete formulation, $H(p)$ can be positive, negative or zero for continuous RVs.

Example 1.8 (Continuous unifom distribution)**.**

Consider a RV $X \sim U_{[a,b]}$, its entropy is

$$H(U_{[a,b]}) = -\int_a^b \frac{1}{b-a}\log\frac{1}{b-a}\mathrm{d}x = \log(b-a).$$

and can be zero (resp. negative) if $b-a=1$ (resp. $b-a\leq 1$).

The difference between the entropy and crossentropy is of critical relevance in this module (and life). We will recall a relevant definition first.

Definition 1.4 (Absolute Continuity)**.**

Let $(\Omega,\mathcal{F})$ be a measurable space and let P and Q be probability measures on it. We say that P is *absolutely continuous* with respect to Q, denoted $P \ll Q$, if for every $A \in \mathcal{F}$,

$$Q(A) = 0 \implies P(A) = 0.$$

Remark 1.10.

If $P \ll Q$, and Q admits a density q, P admits a density p satisfying $p(x) = 0$ whenever $q(x) = 0$.

Definition 1.5 (Kullback–Leibler Divergence)**.**

Let P and Q be probability measures on a measurable space $(\Omega,\mathcal{F})$ such that $P \ll Q$. If P and Q admit densities p and q with respect to a common base measure (e.g. Lebesgue measure), then

$$\mathrm{KL}(p\,\|\,q) \;=\; \mathbb{E}_p\!\left[\log\frac{p(X)}{q(X)}\right] \;=\; \int p(x)\,\log\frac{p(x)}{q(x)}\,\mathrm{d}x.$$

Definition 1.6 (Discrete KL Divergence)**.**

For discrete distributions,

$$\mathrm{KL}(p\,\|\,q) \;=\; \sum_x p(x)\,\log\frac{p(x)}{q(x)}.$$

Remark 1.11.

Notice that the KL divergence is always positive:

$$\mathrm{KL}(p \,\|\, q) = H(p,q) - H(p) \geq 0. \tag{1.32}$$

The KL is a *divergence*, i.e., a function that quantifies how far p is form q that is i) always positive, and ii) $\mathrm{KL}(p \,\|\, q) = 0 \iff p = q$ (identify of the indiscernible). However, note that the KL is not a distance, since

- is not symmetric
- does not have triangle inequality.

Critically, the $\mathrm{KL}(p \,\|\, q)$ is only defined when $P \ll Q$.

1.6 KL divergence as a metric to compare p and q

In the continuous case, we are interested in understanding what type of convergence KL gives. Let us consider, other two divergences:

- $L_1\left(p\|q\right) = \int_{\mathcal{X}} |p(x) - q(x)| \mathrm{d}x$
- $\chi^2\left(p\|q\right) = \int_{\mathcal{X}} \frac{|p(x)-q(x)|^2}{q(x)} \mathrm{d}x.$

Example 1.9 (KL versus L_1)**.**

Consider $p(x) = \mathrm{Uniform}\,(0,1)$ and $q_n(x) = \mathbb{1}_{x\in[0,1/n]} e^{-n} + \mathbb{1}_{x\in[1/n,1]} c_n$, with $c_n \geq 0$ so that q integrates 1 ($n > 1$). Note that here, we have

$$L_1\left(p\|q_n\right) = \int_0^{1/n} |e^{-n} - 1| \mathrm{d}x + \int_{1/n}^1 |c_n - 1| \mathrm{d}x = \frac{|e^{-n} - 1|}{n} + \frac{n|c_n - 1|}{n-1} \tag{1.33}$$

$$\mathrm{KL}\left(p\|q_n\right) = \int_0^{1/n} - \log e^{-n} \mathrm{d}x + \int_{1/n}^1 - \log c_n \mathrm{d}x = 1 + \frac{-n}{n-1} \log c_n. \tag{1.34}$$

Now take $n \to \infty$. For q_n to be well defined, c_n has to converge to 1. Therefore, $L_1\left(p\|q\right) \to 0$. However, note that $\mathrm{KL}\left(p\|q\right) \to 1$.

Example 1.10 (KL versus χ^2)**.**

Consider now $p_\epsilon = (1-\epsilon, \epsilon)$ and $q_\epsilon = (1-\epsilon^2, \epsilon^2)$ two Bernoulli distribution with different parameters. Again, we have:

$$\chi^2\left(p_\epsilon\|q_\epsilon\right) = \frac{||1-\epsilon-1+\epsilon^2||^2}{1-\epsilon^2} + \frac{||\epsilon - \epsilon^2||^2}{\epsilon^2} = \frac{||\epsilon^2 - \epsilon||^2}{1-\epsilon^2} + ||1-\epsilon||^2 \tag{1.35}$$

$$\mathrm{KL}\left(p_\epsilon\|q_\epsilon\right) = (1-\epsilon)\log\frac{1-\epsilon}{1-\epsilon^2} + \epsilon \log \frac{\epsilon}{\epsilon^2} = (1-\epsilon)\log\frac{1}{1+\epsilon} + \epsilon\log\frac{1}{\epsilon}. \tag{1.36}$$

This time, taking $\epsilon \to 0$, we have $\mathrm{KL}\left(p_\epsilon\|q_\epsilon\right) \to 0$ (l'Hôpital's rule), but $\chi^2\left(p_\epsilon\|q_\epsilon\right) \to 1$

Remark 1.12.

The objective of these examples is to show that under different divergences, one can have different criteria of convergence. In the first case, q_n converges to p under L_1, but not under KL. In the second case, p_ϵ converges to q_ϵ under KL but not under ξ^2. This give a sense of *hierarchy* across divergences, where some are sais to induce stronger topologies than others. The stronger the topology, the more demanding the conditions for convergence (or fewer sequences are admitted to converge). In general, we consider KL as one of the stronger divergences (but there are some that are even stronger as we just saw).

Direct versus reverse KL. Since $\text{KL}\,(p\|q)$ is not symmetric, we are interested in studying the *reverse* divergence $\text{KL}\,(q\|p)$ and understanding how it relates its *direct* counterpart.

Since $P \ll Q$ is needed for $\text{KL}\,(p\|q)$, it is required that $P \gg Q$ for $\text{KL}\,(q\|p)$. This gives intuition of $\text{KL}\,(p\|q)$ as a metric assessing how well q approximates p (and not viceversa); this is because if there is a set $A \subset \mathcal{X}$ such that $Q(A) = 0$ and $P(A) > 0$ is strongly penalised, unlike the opposite case.

Let us see a numerical example.

Example 1.11 (Assymetry of the KL between two Gaussians)**.**

The KL divergence between two Gaussians is

$$\text{KL}\,(\mathcal{N}(\mu_0, \sigma_0)\|\mathcal{N}(\mu_1, \sigma_1)) = \log\frac{\sigma_1}{\sigma_0} + \frac{\sigma_0^2 + (\mu_0 - \mu_1)^2}{2\sigma_1^2} - \frac{1}{2}. \tag{1.37}$$

Let us consider $p = \mathcal{N}(0, 1)$ and $q = \mathcal{N}(0, v^2)$, and evaluate

- $\text{KL}\,(p\|q) = \frac{1}{2}(\log v^2 + v^{-2} - 1)$
- $\text{KL}\,(q\|p) = \frac{1}{2}(-\log v^2 + v^2 - 1)$.

Fig. 1.4 shows these functions, note how the penalisation strength depends on the direction.

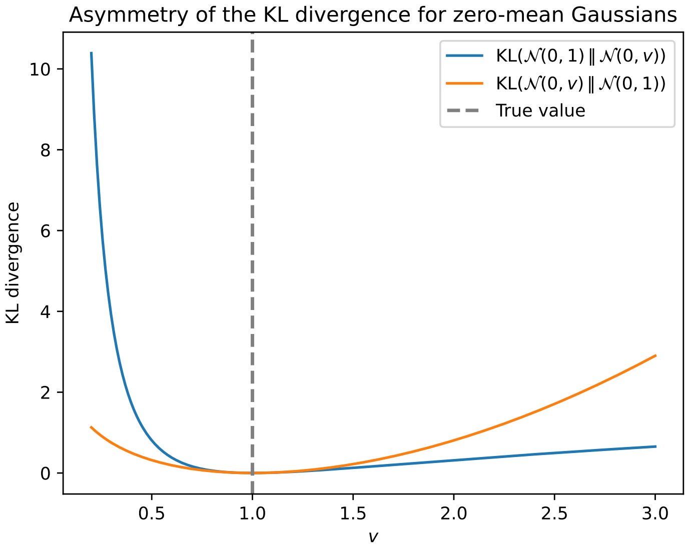

Figure 1.4: Direct and reverse KL for zero mean Gaussians as a function of the variance.

Example 1.12 (KL gradient flow)**.**

Let us now find the approximating q via optimisation for the above example. We can do this via optimisation. Differentiating eq. (1.37) wrt to μ_1 and σ_1, we have

$$\nabla_{\mu_1}\mathrm{KL}\left(\mathcal{N}(\mu_0,\sigma_0)\|\mathcal{N}(\mu_1,\sigma_1)\right) = \frac{(\mu_1-\mu_0)}{\sigma_1^2} \tag{1.38}$$

$$\nabla_{\sigma_1}\mathrm{KL}\left(\mathcal{N}(\mu_0,\sigma_0)\|\mathcal{N}(\mu_1,\sigma_1)\right) = \frac{1}{\sigma_1} - \frac{\sigma_0^2}{\sigma_1^3} = \frac{\sigma_1^2-\sigma_0^2}{\sigma_1^3} \tag{1.39}$$

Where it is clear the this is minimised for $q = p$. Additionally, we can build the gradient descent rule:

$$\mu_n \to \mu_n - \eta_\mu \frac{(\mu_n-\mu_0)}{\sigma_n^2} \tag{1.40}$$

$$\sigma_n \to \sigma_n - \eta_\sigma \frac{\sigma_n^2-\sigma_0^2}{\sigma_n^3} \tag{1.41}$$

Figure 1.5 implements these recursions.

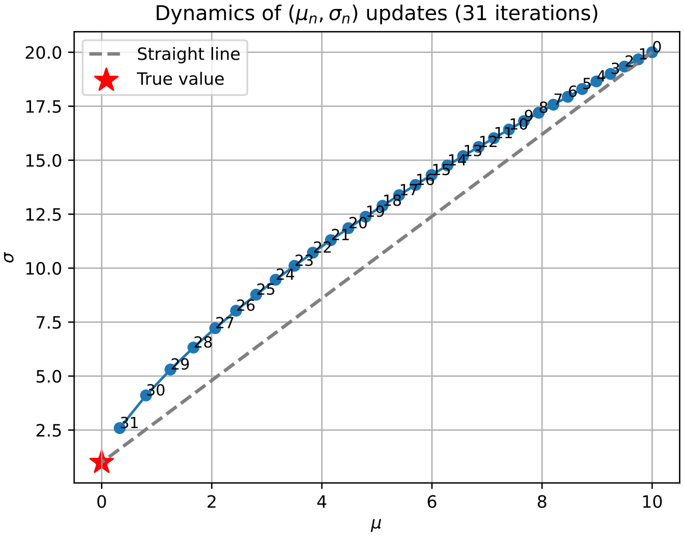

Figure 1.5: KL gradient flow between two Gaussians.

1.7 KL and maximum likelihood