Lecture Notes on Bayesian statistics

Kirill Neklyudov

September 22, 2025

Abstract

This document contains the notes that I've been writing *for myself* to prepare for teaching of the course "Methods of Bayesian statistics" at UdeM. This is not the truth chiseled in stone, this hasn't been proof-read and is very subjective especially when deviates to other subjects.

Contents

1 Basics of probability

1.1 Random Variable, Density

Definition 1 (Random Variable)**.** *Random variable is defined by the Kolmogorov triplet $(\Omega, \mathcal{F}, P)$, where*

1. *Ω is the set of all possible outcomes [Kirill:set of all values of our random variable]*
2. *$\mathcal{F}$ is the sigma-algebra (it's a specific set of subsets) defined on Ω [Kirill:we would like to ask some questions like "what's the probability that our random variable is less than 5?"; hence we need to reason about subsets of Ω]*
   - *(a) $\Omega \in \mathcal{F}$ [Kirill:indeed, we have to reason about the probability of the entire set]*
   - *(b) if $A \in \mathcal{F}$ then $\Omega \backslash A \in \mathcal{F}$ [Kirill:because we want to reason about event A happening and NOT happening]*
   - *(c) if $A_i \in \mathcal{F}\,,\ \forall i = 1, 2, \ldots$ then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$ [Kirill:because of the additivity of the measure below]*
3. *$P : \mathcal{F} \to [0, 1]$ the probability measure [Kirill:it's a map from subsets to numbers]*
   - *(a) $P(\Omega) = 1$ [Kirill:that's quite clear]*
   - *(b) $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ for $A_i \in \mathcal{F}\,,\ \forall i$ and $A_i \cap A_j = \varnothing\,,\ \forall i, j$. [Kirill:indeed, if we consider a union of several non-overlapping events, their probabilities have to add up]*

The sigma-algebra in this definition carries a lot of the work, because it designs the set of subsets which probability you can measure (input into P). To understand this, the following two

Exercise 1. *Prove that $\mathcal{F}$ also contains the intersections, i.e. if $A_i \in \mathcal{F}\,,\ \forall i = 1, 2, \ldots$ then $\cap_{i=1}^{\infty} A_i \in \mathcal{F}$.*

Exercise 2. *Construct the Borel sigma-algebra (minimal sigma-algebra containing all the open sets $(a, b), a < b$) on the unit interval $[0, 1]$.*

The definition of the random variable is the central concept of probability theory (formulated by Kolmogorov) and you can't escape it. This definition is not how people reason about probabilities in practice, though. It is a distilled concept that serves as a solid foundation to the probability theory and everything that uses it. If you would like to know more about it you should read books on Measure theory, Probability theory, Stochastic processes. Note that this will lead you downwards to the studies of the essence and the atomic ideas (at "nanoscales") of probability theory and pure mathematics. In this course we will go upwards and will construct concepts on top of the probability theory. For that purpose we will need more visual and intuitive definitions of random variables.

Definition 2 (Discrete random variable)**.** *We can define discrete random variables in a straightforward way following the definition of the random variable*

1. *Ω is a countable set of all outcomes [Kirill:think about categorical distribution, or a random walk on integers]*
2. *$\mathcal{F} = 2^{\Omega}$ [Kirill:we can simply choose the set of all subsets, and, in that sense the sigma-algebra definition is not working at its fullest, but, this is not the only possible choice of a sigma-algebra on a finite set]*
3. *$P : 2^{\Omega} \to [0, 1]$. For instance, $\forall A \in \mathcal{F}$ it's going to be a simple enumeration $P(A) = \sum_{\omega \in A} P(\omega)$. [Kirill:hence, we just need to define $P : \Omega \to [0, 1]$, and the rest is an accounting job.]*

You can see that in Definition 2 we barely used the definition of a random variable. Indeed, for every outcome ω in the (countable) set of all outcomes Ω, we can just define a non-negative number $P(\omega) \geq 0$ (we can also exclude all the outcomes ω for which $P(\omega) = 0$) and then make sure that the total probability sums up to 1, i.e.

$$P(\Omega) = \sum_{\omega \in \Omega} P(w) = 1\,. \tag{1}$$

From the perspective of the probability theory, this is great because our definition describes one of the most important use-cases in a meaningful way. From the practical perspective, the theory doesn't tell us anything new, just gives us a thumbs up. [Kirill:note that the probability theory for discrete variables in this form existed long before Hilbert and Kolmogorov (mathematical foundation of probability theory was one of the problems proposed by Hilbert in 1900, which was solved by Kolmogorov in 1933)].

Definition 3 (Continuous random variable)**.** *The definition of the random variable shines when we have to meticulously define a new random variable which is a subject of our studies [Kirill:which is not what usually happens in practice].*

1. *$\Omega = \mathbb{R}^n$ [Kirill:in the course we will consider only the Euclidian space or some simple subset of it.]*
2. *$\mathcal{F}$ is the Borel sigma-algebra, i.e. the minimal sigma-algebra containing all the open subsets (i.e. start constructing from $(a, b), a < b$ for $\mathbb{R}$).*
3. *$P : \mathcal{F} \to [0, 1]$ probability measure is what defines our random variable [Kirill:in the sense that there are a lot of different random variables with the same $\Omega, \mathcal{F}$ and different P].*

Continuous random variables is the main reason why people still learn/teach/discuss Definition 1. It is the rock bottom (of probability theory and everything on top of it) that you hit if you keep asking the question "why?". So far nobody has found cracks in it (doesn't mean there is no any) and we keep on using it and everything works. However, keeping in mind the sigma-algebras and all the corner cases is not necessary in most of the practical scenarios. It is simply not the right tool that allows you to solve practical problems. Much more useful tool in practice is the concept of the probability density.

Definition 4 (Probability Density Function)**.** *Consider the random variable X defined as the triplet $(\Omega, \mathcal{F}, P)$, then for any $A \in \mathcal{F}$ we have*

$$\int_A dx\; p(x) = P(X \in A)\,, \tag{2}$$

where the integrated (Lebesgue integral) function $p(x)$ is called Probability Density Function (PDF) or simply "density".

The probability density function $p(x)$ is not a probability and doesn't tell you anything about any probability until you perform the integration! [Kirill:you can prove that points also belong to sigma-algebras but their measure is zero, i.e., for $A = \{y\}$ the probability $P(X \in A) = 0$]

The useful way to think about the density is geometric, i.e. think that you can draw any area in the Euclidian space [Kirill:but not something egregious, which is taken care of by the sigma-algebra] then the density is a magic black box that allows you to calculate the probability that your random variable is going to be in this area. The density can be visualized in the same way as any function (see Figure 1).

Clearly, the following properties hold

$$\forall x \in \Omega\,,\; p(x) \geq 0 \;\text{ and }\; \int_\Omega dx\; p(x) = P(\Omega) = 1\,. \tag{3}$$

Note that this follows from the definition of the random variable. However, quite often people see some non-negative function $f(x) \geq 0$, and they prove that $\int_\Omega dx\; f(x) < \infty$, i.e. the function is normalizable. Then they jump to defining the corresponding probability density function as

$$p(x) = \frac{f(x)}{\int_\Omega dx\; f(x)}\,, \tag{4}$$

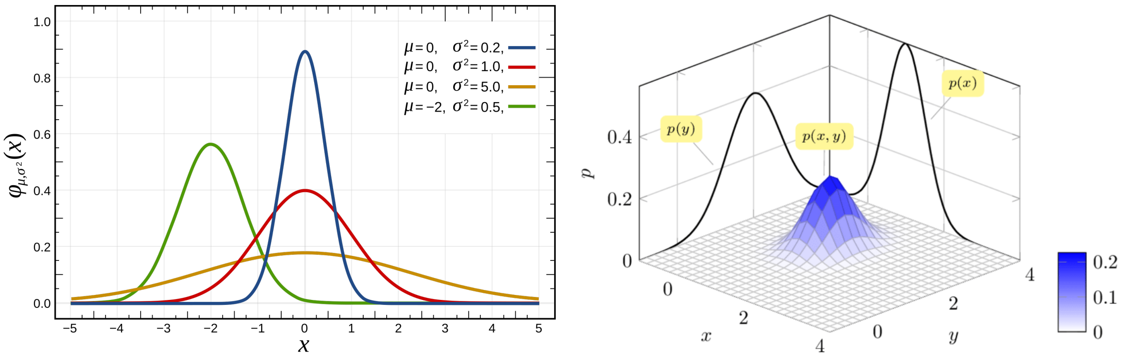

Figure 1: Visualization of 1d and 2d Gaussian densities

because if the function $f(x)$ is good enough it is true, and because it is useful in practice [Kirill:which is the whole point of the Monte Carlo methods].

Example 1. *Consider* $f(x) = \exp(-\frac{1}{2}x^2)$*, find the corresponding density by finding the normalization constant* $\int_{\mathbb{R}} dx\ f(x)$*.*

Proof. Let's denote

$$Z := \int_{\mathbb{R}} dx\ \exp\left(-\frac{1}{2}x^2\right), \tag{5}$$

then we have

$$Z^2 = \int_{\mathbb{R}^2} dxdy\ \exp\left(-\frac{1}{2}(x^2+y^2)\right). \tag{6}$$

Taking $x = r\cos\phi\,, y = r\sin\phi$, we have

$$Z^2 = \int_0^{2\pi} d\phi\ \int_0^{\infty} dr\ \exp\left(-\frac{1}{2}r^2\right)r = 2\pi\int_0^{\infty} dy\ \exp(-y) = 2\pi\,. \tag{7}$$

Thus, we have

$$p(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right). \tag{8}$$

□

Definition 5 (Normal distribution)**.** *The density of the Normal distribution or Gaussian distribution is given by the following formula*

$$\mathcal{N}(x\,|\,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{9}$$

For $\mu = 0, \sigma = 1$*, we call it the standard normal distribution.*

1.2 Joint, Marginal, Conditional distributions

Consider two random variables X, Y that we observe together. The joint distribution of these random variables is another random variable, which space of outcomes and the corresponding sigma-algebra can be formally constructed using Definition 1.

Definition 6 (Joint distribution)**.** *Consider two random variables X with $(\Omega_x, \mathcal{F}_x, P_x)$ and Y with $(\Omega_y, \mathcal{F}_y, P_y)$, then we can define*

1. *$\Omega = \Omega_x \times \Omega_y$, i.e. the joint outcome space is the direct product of the outcome spaces.*
2. *$\mathcal{F} = \mathcal{F}_x \times \mathcal{F}_y$, the product of sigma-algebras can be constructed as the minimal sigma-algebra that contains the set $\{(A_1, A_2) : A_1 \in \mathcal{F}_x\,,\ A_2 \in \mathcal{F}_y\}$.*
3. *$P : \mathcal{F} \to [0, 1]$, and, importantly, it cannot be defined simply through the knowledge of X and Y. We have to define this probability measure from other principles, e.g. the observations that we collected by observing X and Y simultaneously.*

Once again, this definition is something that tells us that we have the right to reason about these things, but doesn't give us the tools to reason.

In practice, we usually start from the joint probability or the joint density because it contains the full information about all the random variables involved. That is, for the discrete variables $\{X_i\}_{i=1}^n$ taking values x_i in integer numbers $\mathbb{Z}$, the joint probability is usually defined as a function that maps integers to probability values, i.e.

$$P : \mathbb{Z}^n \to [0, 1]\,, \text{ and for the value at given point we write } P(X_1 = x_1, \ldots, X_n = x_n) = P(x_1, \ldots, x_n)\,. \tag{10}$$

For the continuous random variables, we define the joint distribution through the joint density. That is, for the random variables $\{X_i\}_{i=1}^n$ taking values x_i in the Euclidean space $\mathbb{R}^m$ (not necessarily of the same dimension), we have the density function p that evaluates probabilities

$$\int_A dx\; p(x) = P(X \in A)\,, \tag{11}$$

where A is in the joint sigma-algebra, X is the joint outcome and P is defined as the probability measure on the joint sigma-algebra. Clearly, the following holds for the density function

$$p(x_1, \ldots, x_n) \geq 0\,, \quad \int_{\mathbb{R}^{mn}} dx_1 \ldots dx_n\; p(x_1, \ldots, x_n) = 1\,. \tag{12}$$

Note that there is not much difference between what we considered for the single random variable. All the differences are hidden in the way we construct the Kolmogorov triplet.

Interesting differences appear when we start "disassembling" the joint distribution back into individual random variables and study relations between these variables. Consider the following function

$$F(x) \coloneqq \sum_{y \in \mathbb{Z}} P(X = x, Y = y)\,, \tag{13}$$

where $P(X = x, Y = y)$ is the probability function of random variables X and Y and we sum over all possible values of Y. Note that we are summing over disjoint outcomes [Kirill:indeed, we can't observe Y to take two different values and our outcome space is the product of outcome spaces of X, Y]. We can rewrite it as follows

$$F(x) \coloneqq \sum_{y \in \mathbb{Z}} P(X = x, Y = y) = P(\cup_{y \in \mathbb{Z}}(X = x, Y = y)) = P((X = x, Y \in \Omega_y)) = P(X = x)\,, \tag{14}$$

where in the last transition we drop the dependency on Y since it always holds. This motivates the following definition.

Definition 7 (Marginal distribution)**.** *For the discrete random variables X, Y, the marginal distributions of X and Y are defined as follows*

$$P(x) := \sum_y P(x,y)\,,\ \ P(y) := \sum_x P(x,y)\,. \tag{15}$$

For the continuous random variables X, Y, the probability density function of the marginal distribution of X is defined as follows

$$p(x) := \int_{\Omega_y} dy\; p(x,y)\,,\ \ p(y) := \int_{\Omega_x} dx\; p(x,y)\,. \tag{16}$$

You can think about marginalization as of projection of the joint distribution to one of the "planes" corresponding to some random variable. [Kirill:Nobody stops you from defining a new plane in the joint space and projecting to it] In the space of observations it corresponds to "forgetting" or "erasing" the information about Y, e.g. if you particles hitting a detector, you can register only one of the coordinates, which corresponds to the marginalization of the joint distribution of coordinates.

In practice, one of the most important concepts is the conditioning. Intuitively, it corresponds to querying your database with something like "select all the musicians that are 27 years old". More formally, you can consider all the events $\{(X = i, Y = y)\}_{i\in\mathbb{Z}}$, i.e. we fix the value of Y at y and make a collection of all such events by varying X. Note that clearly $P(X = i, Y = y) \geq 0$ for all i, hence it is almost a valid probability function, the only problem is that it doesn't normalize to 1, i.e.

$$\sum_{i\in\mathbb{Z}} P(X = i, Y = y) < 1\,, \tag{17}$$

because y is fixed and we can't sum over all events in the joint outcome space Ω. This function, though, makes total sense, and we can think about events of the type "$X < 0$ and $Y = 6$". The solution to this problem is to define a new random variable that corresponds to the "slice" of the entire outcome space where $Y = y$. This leads to the following *definition*!

Definition 8 (Conditional distribution)**.** *For the discrete random variables X, Y, the conditional distribution of X for $Y = y$ is defined as*

$$P(X = x \,|\, Y = y) := \frac{P(X = x, Y = y)}{P(Y = y)}\,. \tag{18}$$

For the continuous random variables X, Y, the probability density function of the conditional distribution of X for $Y = y$ is defined as

$$p(x \,|\, y) := \frac{p(x,y)}{p(y)}\,. \tag{19}$$

The vertical line in the notation separates the random variables from the conditions. In other words, $p(x \,|\, y)$ defines a random variable on X but not on y, this is easy to see by evaluating the normalization constants of the conditional distribution.

Corollary 1 (Normalization of conditional distribution)**.** *Conditional distribution normalizes to 1, i.e. defines a valid random variable.*

Proof. By the straightforward accounting exercises, we have

$$\sum_x P(x \,|\, y) = \sum_x \frac{P(x,y)}{P(y)} = \frac{1}{P(y)} \sum_x P(x,y) = \frac{1}{P(y)} P(y) = 1\,, \tag{20}$$

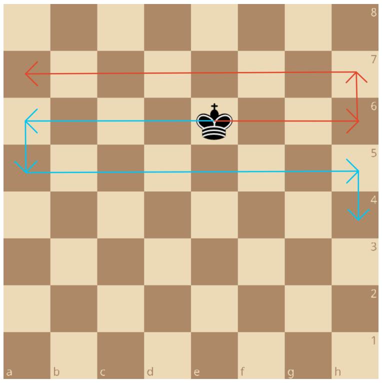

Figure 2: Example of the process that generates independent distributions (horizontal and vertical coordinates of the king) but is completely deterministic and predictable

and

$$\int dx\, p(x \mid y) = \frac{1}{p(y)} \int dx\, p(x, y) = \frac{1}{p(y)} p(y) = 1\,. \tag{21}$$

□

[Kirill:note that a lot of results translate seamlessly from the discrete case to the continuous (which i have been abusing above by reasoning in the discrete case and then applying for the continuous). This is not always the case, but we can't keep writing everything for both cases all the time. There is no default choice, or a "better" choice, the answer might be disappointing for you, but you have to know both and to know all the bridges between both. Thus, by making progress in one of the worlds you can jump to another and get some insights there.]

Note that for different y, according to Definition 8, we have different distributions $p(x \mid y)$. However, what if it is not the case? In other words, consider the special case when

$$p(x \mid y) = p(x)\,, \ \forall y. \tag{22}$$

Then it makes sense to call random variables X and Y independent, because the distribution of X is not anyhow affected by our choice of y. Using Definition 8, we have

$$\frac{p(x, y)}{p(y)} = p(x)\,, \implies p(x, y) = p(x)p(y)\,. \tag{23}$$

Definition 9 (Independent random variables)**.** *Two random variables X, Y are called independent if their joint pdf is a product of the marginal pdfs, i.e.*

$$p(x, y) = p(x)p(y)\,. \tag{24}$$

[Kirill:Please note that this is not some philosophical definition of independence and you can make some conclusions based on this. This is the mathematical definition that means exactly what is written. See example below]

Example 2. *Consider a king on a chess board going forward and backward as shown in Figure 2. After a big number of steps, the distribution of X and Y coordinates of its position is going to be uniform. However, it doesn't mean that we can't predict where the king will move next.*

In theory, everything looks symmetric and there is no conceptual difference between $p(x \mid y)$ and $p(y \mid x)$. However, in practice, we usually know only one of these conditional probabilities and we would like to find the other.

Let's say we know the distribution of X through density $p(x)$ and we know how Y depends on X, i.e. we know $p(y \,|\, x)$. Oftentimes we are interested in "inverting the function" $p(y \,|\, x)$, i.e. in finding $p(x \,|\, y)$ [Kirill:sometimes finding the posterior distribution is called the inverse problem]. This can be done as follows

$$p(x \,|\, y) = \frac{p(x,y)}{p(y)} = \frac{p(x,y)}{\int_{\Omega_x} dx \; p(x,y)} = \frac{p(y \,|\, x)p(x)}{\int_{\Omega_x} dx \; p(y \,|\, x)p(x)} \,, \tag{25}$$

where we used the definitions of marginal and conditional distributions. People refer to this result as Bayes' theorem, as follows.

Theorem 1 (Bayes' theorem)**.** *Conditional density $p(x \,|\, y)$ can be written using the conditional density $p(y \,|\, x)$ and the marginal density $p(x)$ as follows*

$$p(x \,|\, y) = \frac{p(y \,|\, x)p(x)}{\int_{\Omega_x} dx \; p(y \,|\, x)p(x)} \,. \tag{26}$$

This course is devoted to different practical solutions of probabilistic inverse problems based on this result.

Exercise 3 (Medical Tests)**.** *Consider the test t for some disease d, which conditional distribution $p(t \,|\, d)$ is defined in the table below. Assuming that we know the marginal distribution of d ($p(d = 1) = 0.5$) for the disease, find $p(d = 1 \,|\, t = 1)$*

	$p(t = 1 \,\vert\, d)$	$p(t = 1 \,\vert\, d)$
d=1	*0.99*	*0.01*
d=0	*0.01*	*0.99*

How the result changes if the disease is very rare ($p(d = 1) = 10^{-4}$)?

Example 3. *Consider the discrete distribution $P(Y = 0) = P(Y = 1) = 1/2$ and the continuous distribution $X \,|\, Y$ which has the density $\mathcal{N}(x \,|\, y, \sigma^2)$. Using the Bayes' theorem, we have*

$$\begin{aligned}
P(Y = 1 \,|\, x) &= \frac{p(x \,|\, y = 1)P(Y = 1)}{p(x \,|\, y = 0)P(Y = 0) + p(x \,|\, y = 1)P(Y = 1)} && (27)\\
&= \frac{\mathcal{N}(x \,|\, 1, \sigma^2)}{\mathcal{N}(x \,|\, 0, \sigma^2) + \mathcal{N}(x \,|\, 1, \sigma^2)} = \frac{\exp\!\left(-\frac{1}{2\sigma^2}(x-1)^2\right)}{\exp\!\left(-\frac{1}{2\sigma^2}(x-1)^2\right) + \exp\!\left(-\frac{1}{2\sigma^2}x^2\right)} && (28)\\
&= \frac{1}{1 + \exp\!\left(-\frac{1}{2\sigma^2}[x^2 - x^2 + 2x - 1]\right)} = \frac{1}{1 + \exp\!\left(-\frac{1}{2\sigma^2}[2x - 1]\right)} && (29)
\end{aligned}$$

The resulting probability depends on x and is described by the sigmoid function (see Figure 3)

Definition 10 (Bernoulli random variable)**.** *The random variable X defined on two states $\Omega = \{0, 1\}$ [Kirill:the states can be anything, ofc] defines the Bernoulli random variable, i.e.*

$$X = \begin{cases} 1\,, & \text{with prob. } \theta\,, \\ 0\,, & \text{with prob. } (1-\theta)\,, \end{cases} \tag{30}$$

where $\theta \in [0, 1]$ is a parameter. For $\theta = 1/2$, some people call this random variable "fair coin".

Oftentimes, we have a number of *independent* Bernoulli random variables and we want to reason about how many of them equal 1 and how many of them equal 0. First, we will rewrite the probability function of a single Bernoulli variable as follows

$$P(x \,|\, \theta) = \theta^x (1-\theta)^{1-x} \,, \tag{31}$$

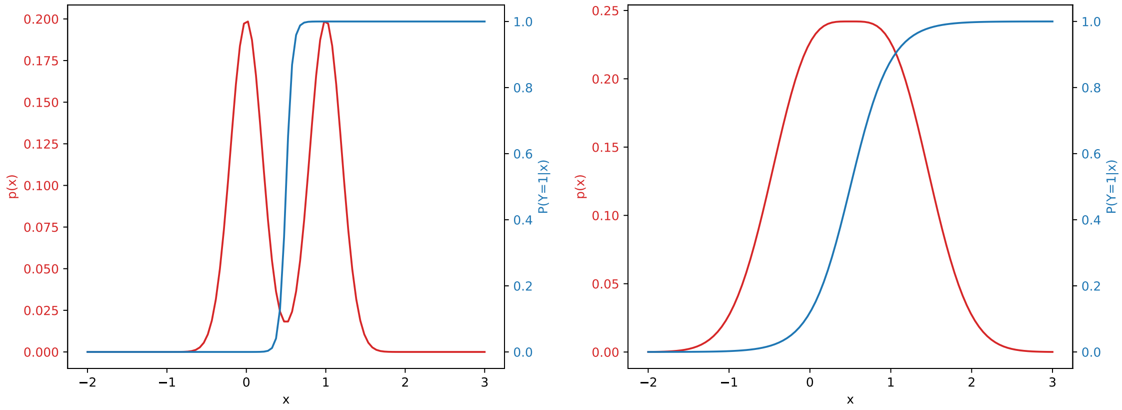

Figure 3: Density $p(x)$ and the probability $P(Y = 1 \,|\, x)$ for $\sigma = 0.2$ (left) and $\sigma = 0.5$ (right).

which, as you can see, corresponds to Equation (30). However, now we can multiply this probability functions according to Definition 9, i.e.

$$P(x_1, \ldots, x_n \,|\, \theta) = \theta^{\sum_i x_i} (1-\theta)^{\sum_i (1-x_i)}\,, \tag{32}$$

which serves as a motivation to another random variable in the following definition.

Definition 11 (Binomial random variable)**.** *Consider n independent Bernoulli random variables with the parameter θ. The Binomial random variable X is defined as the* ***number*** *of Bernoulli variables that are equal 1, i.e.*

$$P(X = k \,|\, \theta, n) = C_n^k \theta^k (1-\theta)^{n-k}\,. \tag{33}$$

Exercise 4. *Prove the formula from Definition 11.*

Definition 12 (Poisson random variable)**.** *Poisson random variable X with the rate λ is defined as*

$$P(X = k \,|\, \lambda) = \exp(-\lambda)\frac{\lambda^k}{k!}. \tag{34}$$

[Kirill:usually the Poisson random variable defines some sequence of events happening randomly in time (e.g., earthquakes) that's why it makes sense to call its parameter the "rate".]

Exercise 5. *Prove additivity of the Poisson random variables, i.e. for $X_1 \sim Poiss(\lambda_1)$ and $X_2 \sim Poiss(\lambda_2)$, prove that $X_1 + X_2 \sim Poiss(\lambda_1 + \lambda_2)$.*

1.3 Probabilistic Graphical model

We can always reason about all the present random variables by writing the joint distribution of all of them. However, oftentimes, there is a lot of structure in the dependencies (according to Definition 9) between the variables. It is very convenient to depict the relations between the variables graphically, that's why the community has agreed on the following definition.

Definition 13 (Probabilistic Graphical model)**.** *Acyclic directed graph is defines a probabilistic graphical model in the following way*

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \,|\, \texttt{parents of } x_i)\,. \tag{35}$$

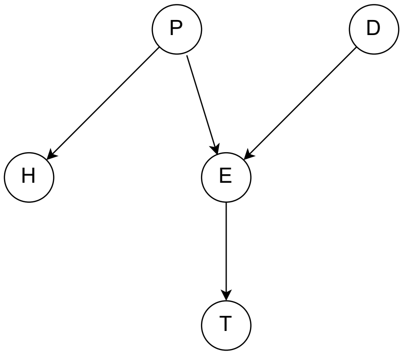

Figure 4: Probabilistic graphical model for Exercise 7.

For instance, the graph in Figure 4 defines the following joint distribution

$$p(H, P, E, D, T) = p(T \mid E)p(E \mid P, D)p(H \mid P)p(P)p(D). \tag{36}$$

Exercise 6. *Draw the probabilistic graphical model for the Markov chain probabilistic model*

$$p(x_1, \dots, x_n) = p(x_0) \prod_{i=0}^{n-1} p(x_{i+1} \mid x_i)\,, \tag{37}$$

for the auto-regressive probabilistic model

$$p(x_1, \dots, x_n) = p(x_0) \prod_{i=0}^{n-1} p(x_{i+1} \mid x_i, x_{i-1} \dots, x_0)\,. \tag{38}$$

Exercise 7. *The model consists of the following random variables: the student went to a party $P \in \{0, 1\}$, the student has a depression $D \in \{0, 1\}$, the student has a headache $H \in \{0, 1\}$, results of the exam are good $E \in \{0, 1\}$, the teacher is happy $T \in \{0, 1\}$. The relations between these variables are given in Figure 4 and the conditional probabilities are defined as follows*

P	D	$P(E = 0 \mid P, D)$
1	1	0.999
0	1	0.9
1	0	0.9
0	0	0.01

P	$P(H = 1 \mid P)$
1	0.999
0	0.9

E	$P(T = 1 \mid E)$
1	0.5
0	0.95

$$P(P = 1) = 0.2$$
$$P(D = 1) = 0.4.$$

Find the following probabilities $P(P = 1 \mid H = 1)$, $P(P = 1 \mid T = 0)$, $P(P = 1 \mid T = 0, H = 1)$.

1.4 Maximum Likelihood

In practice, we don't usually know the probabilistic model of some phenomenon, and we would like to design such a model. The design process of a probabilistic model consists of two main stages:

1. Choosing the parametric family of the probabilistic model, i.e. defining how exactly the density value (or probability) $p(x \mid \theta)$ depends on the parameters θ.

2. Estimating the parameters θ from the given dataset of observations $\{x_1, \dots, x_N\}$.

Let's assume that we know exactly the parametric family $p(x \,|\, \theta)$, i.e. we know for sure that the data we are modeling was generated according to this distribution. Then one can use the maximum likelihood principle to estimate the parameters θ from the data.

Definition 14 (Maximum likelihood)**.** *For the given parametric family $p(x \,|\, \theta)$, and the empirical data $\{x_1, \ldots, x_N\}$, Maximum Likelihood Estimator (MLE) of parameters $\theta \in \Theta$ is defined as*

$$\theta_{MLE} = \operatorname*{argmax}_{\theta \in \Theta} p(\{x_1, \ldots, x_N\} \,|\, \theta)\,, \tag{39}$$

where the density (or probability) $p(\{x_1, \ldots, x_N\} \,|\, \theta)$ is called likelihood.

If the data was acquired through repeating the same experiment over and over again with the same conditions, it is reasonable to assume [Kirill:this is a very big and crucial assumption for statistics and machine learning] that the observations $\{x_1, \ldots, x_N\}$ are iid, i.e.

$$p(\{x_1, \ldots, x_N\} \,|\, \theta) = \prod_{i=1}^{N} p(x_i \,|\, \theta)\,. \tag{40}$$

Under this assumption it's convenient to introduce the following definition.

Definition 15 (Log-likelihood)**.** *For the given parametric family $p(x \,|\, \theta)$, and the iid samples $\{x_1, \ldots, x_N\}$, log-likelihood is defined as*

$$\ell(\theta) = \log p(\{x_1, \ldots, x_N\} \,|\, \theta) = \log \prod_{i=1}^{N} p(x_i \,|\, \theta) = \sum_{i=1}^{N} \log p(x_i \,|\, \theta)\,. \tag{41}$$

The maximum likelihood estimator yields closed-form solution for many parametric distribution just by writing down the necessary conditions for the maximum.

Example 4 (MLE for the Binomial distribution)**.** *Consider $\mathcal{D} = \{x_1, \ldots, x_N\}$ — iid samples from the Binomial distribution Binom$(x \,|\, n, \theta)$. The log-likelihood for θ is*

$$\ell(\theta) = \sum_{i=1}^{N} \log\big(C_n^{x_i} \theta^{x_i} (1-\theta)^{n-x_i}\big) = \sum_{i=1}^{N} (\log C_n^{x_i} + x_i \log\theta + (n - x_i)\log(1-\theta))\,. \tag{42}$$

From the maximum likelihood principle, we have

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\ell(\theta) = \sum_{i=1}^{N}\left(x_i \frac{1}{\theta} + (n - x_i)\frac{1}{1-\theta}\right), & (43)\\ (1-\theta)\sum_i x_i &= \theta \sum_i (n - x_i) & (44)\\ \theta_{MLE} &= \frac{\sum_i x_i}{\sum_i n} = \frac{1}{N}\sum_i \frac{x_i}{n}\,. & (45) \end{aligned}$$

It is the average success rate over all the experiments.

The following theorem is one of the main motivations for using MLE.

Theorem 2 (Convergence of MLE)**.** *Consider the set of iid samples $\mathcal{D} = \{x_1, \ldots, x_N\}$ from the categorical distribution $P(x \,|\, \theta^*)$*

$$P(x = k \,|\, \theta^*) = \theta_k^*\,,\ \theta_k^* \geq 0\ \forall k\,. \tag{46}$$

The maximum likelihood estimator of θ (under the ground true probabilistic model) converges (in the number of samples) to the ground true parameters, i.e.

$$\lim_{N\to\infty} \theta_{MLE} = \theta^* \,. \tag{47}$$

Proof. The log-likelihood of the dataset is

$$\ell(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log P(x_i \,|\, \theta) = \frac{1}{N}\sum_{k=1}^{M} N_k \log P(x_i = k \,|\, \theta) \tag{48}$$

$$= \frac{1}{N}\sum_{k=1}^{M} N_k \log \theta_k = \sum_{k=1}^{M} \frac{N_k}{N} \log \theta_k \,, \tag{49}$$

where N_k is the number of samples that equal k, i.e. $N_k = |\{x \in \mathcal{D} \,|\, x = k\}|$. From the maximum likelihood principle, we have

$$\theta_{\text{MLE}} = \operatorname{argmax} \ell(\theta) \,, \;\; \text{s.t.} \;\; \sum_{k=1}^{M} \theta_k = 1 \,,\; \theta_k \geq 0 \;\forall k \,. \tag{50}$$

The corresponding Largangian is

$$\mathcal{L}(\theta) = \ell(\theta) + \lambda \sum_{k=1}^{M} (\theta_k - 1) \,, \tag{51}$$

and the necessary condition for the optimum is

$$\forall k \,,\; 0 = \frac{\partial \ell(\theta)}{\partial \theta_k} + \lambda = \frac{1}{\theta_k}\frac{N_k}{N} + \lambda \implies \theta_k = \frac{N_k}{N} \,. \tag{52}$$

Thus, we have

$$\theta_{\text{MLE}} = \begin{bmatrix} \dots \\ \frac{N_k}{N} \\ \dots \end{bmatrix} . \tag{53}$$

For the limit of the infinite number of samples, we have

$$\lim_{N\to\infty} (\theta_{\text{MLE}})_k = \lim_{N\to\infty} \frac{N_k}{N} = P(X = k \,|\, \theta^*) = \theta^*_k \,. \tag{54}$$

□

Thus, we have demonstrated that if we know the probabilistic model $p(x \,|\, \theta)$ exactly, then for any given precision of the parameter estimation θ we can collect the dataset $\mathcal{D}$ large enough to satisfy this precision.

Exercise 8. *Consider $x_1, \dots, x_N$ — iid samples from the Normal distribution $\mathcal{N}(x \,|\, \mu, \sigma)$. Find the maximum likelihood estimation of μ, σ.*

Exercise 9. *Consider $x_1, \dots, x_N$ — iid samples from the Poisson distribution Poiss(λ). Find the maximum likelihood estimation of λ.*

Proposition 1 (MLE for regression). *Consider the dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. Assuming the following probabilistic model for the dataset*

$$y = f(x; \theta) + \varepsilon \,, \varepsilon \sim \mathcal{N}(\varepsilon \,|\, 0, \sigma^2) \,, \tag{55}$$

where f is some known function. find the maximum likelihood estimation of θ and σ

Proof. First, let's rewrite the probabilistic model in a more familiar form

$$p(y \mid x, \theta, \sigma) = \mathcal{N}(y \mid f(x;\theta), \sigma^2)\,. \tag{56}$$

Note that here y is conditioned on the features x, parameters of the mean predictor θ and the standard deviation σ. The log-likelihood of this probabilistic model is

$$\ell(\theta, \sigma) = \frac{1}{N}\sum_{i=1}^{N} \log \mathcal{N}(y_i \mid f(x_i;\theta), \sigma^2) = \frac{1}{N}\sum_{i=1}^{N}\left[-\frac{1}{2\sigma^2}(y_i - f(x_i;\theta))^2 - \frac{1}{2}\log(2\pi\sigma^2)\right]. \tag{57}$$

Hence, we have

$$\begin{aligned} \theta_{\text{MLE}} = &\ \operatorname*{argmax}_{\theta} \ell(\theta,\sigma) = \operatorname*{argmax}_{\theta} \ell(\theta,\sigma)\frac{1}{N}\sum_{i=1}^{N}\left[-\frac{1}{2\sigma^2}(y_i - f(x_i;\theta))^2 - \frac{1}{2}\log(2\pi\sigma^2)\right] & (58)\\ = &\ \operatorname*{argmin}_{\theta} \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{2\sigma^2}(y_i - f(x_i;\theta))^2\right] = \operatorname*{argmin}_{\theta} \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i;\theta))^2\,. & (59) \end{aligned}$$

Thus, MLE is equivalent to Mean-Squared Error (MSE) loss function. For σ_{MLE}, we have

$$\begin{aligned} 0 = \frac{\partial \ell(\sigma)}{\partial \sigma} = &\ \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{\sigma^3}(y_i - f(x_i;\theta))^2 - \frac{1}{\sigma}\right] = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sigma^3}(y_i - f(x_i;\theta))^2 - \frac{1}{\sigma}\,, & (60)\\ \sigma^2_{\text{MLE}} = &\ \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i;\theta_{\text{MLE}}))^2\,. & (61) \end{aligned}$$

Thus, the MLE for σ equals the average error that persists for θ_{MLE}. □

Exercise 10. *MLE for Laplace residuals*

Exercise 11. *MLE for logistic regression*

2 Basics of Bayesian statistics

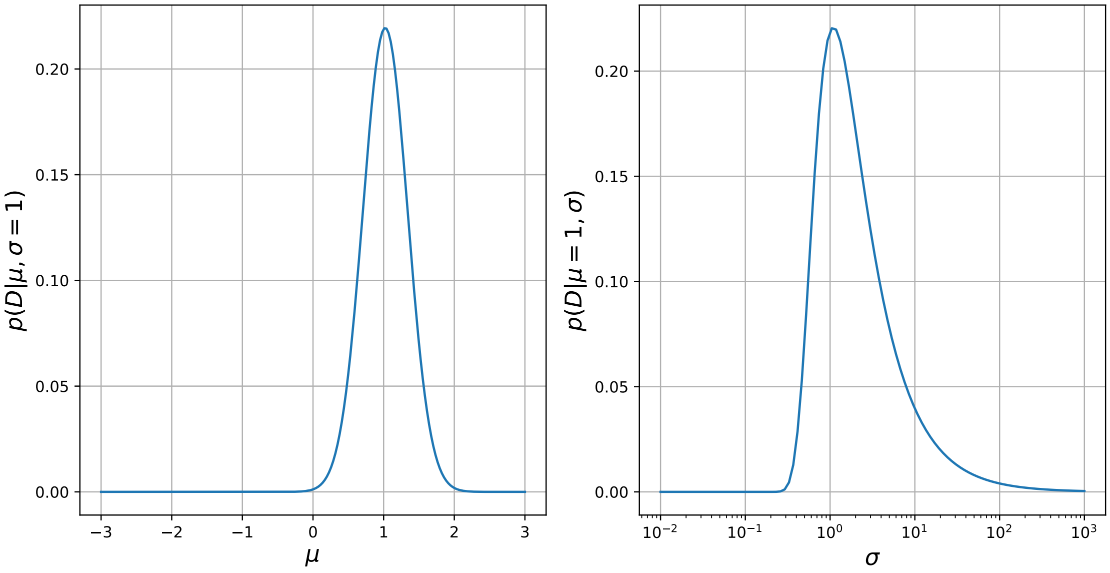

Figure 5: Likelihood of the dataset $\mathcal{D}$ of 10 samples from $\mathcal{N}(x\,|\,\mu = 1, \sigma^2 = 0.5^2)$.

Maximum likelihood gives us a point-estimate of the parameters of our probabilistic model, and if the model is correct [Kirill:no] then we know that with the number of data points our estimate converges to the ground true value. However, what if we plot the likelihood for other values of parameters? In Figure 5, we see that the likelihood function depending on the parameter also looks like some density, i.e. it is a positive function that vanishes at the infinity so it might be normalizable.

Indeed, consider the likelihood of some dataset $\mathcal{D}$, that corresponds to the normal model $p(x\,|\,\theta) = \mathcal{N}(x\,|\,\mu, \sigma^2)$. For the likelihood w.r.t μ (and some fixed σ) it's very straightforward

$$p(\mathcal{D}\,|\,\mu, \sigma = 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^2\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\sum_i x_i^2 - 2\mu\sum_i x_i + N\mu^2\right)\right) \tag{62}$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\sqrt{N}^2}{2}\left(\mu^2 - 2\mu\hat{\mu} + \hat{\mu}^2 - \hat{\mu}^2 + \frac{1}{N}\sum_i x_i^2\right)\right) \tag{63}$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2(1/\sqrt{N})^2}(\mu - \hat{\mu})^2 - \hat{\mu}^2 + \frac{1}{N}\sum_i x_i^2\right) \propto \mathcal{N}\left(\mu\,|\,\hat{\mu}, \frac{1}{N}\right), \tag{64}$$

where $\hat{\mu} = \frac{1}{N}\sum_i x_i$. Thus, we see that the likelihood is proportional to the normal distribution w.r.t. μ. However, it is the normal distribution w.r.t μ because it doesn't normalize to 1. [Kirill:in some cases the function can be not normalizable in principle]

For the likelihood w.r.t. σ (and fixed μ), we have

$$p(\mathcal{D}\,|\,\mu = 0, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N} x_i^2\right) \propto \left(\frac{1}{\sigma^2}\right)^{N/2} \exp\left(-\left(\frac{1}{2}\sum_{i=1}^{N} x_i^2\right)\frac{1}{\sigma^2}\right) \tag{65}$$

$$= \gamma^{N/2} \exp\left(-\left(\frac{1}{2}\sum_{i=1}^{N} x_i^2\right)\gamma\right) \propto \mathcal{G}\left(\gamma\,|\,N/2 + 1, \frac{1}{2}\sum_{i=1}^{N} x_i^2\right), \tag{66}$$

where we introduced the parameter $\gamma = 1/\sigma^2$. The likelihood function is proportional to the density of the gamma distribution (see below), but it's not the gamma distribution because it doesn't normalize to 1.

Definition 16 (Gamma distribution)**.** *Consider the density proportional to*

$$\mathcal{G}(x \,|\, a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)\,, \tag{67}$$

where $\Gamma(\cdot)$ is the gamma function.

To sum up, we see that the likelihood already defines something like a distribution on the set of parameters. Thus, we should consider treating parameters as random variables!

2.1 Bayesian reasoning

When treating parameters as random variables, let's start with the likelihood model, i.e. let's assume that the following holds

$$p(\mathcal{D} \,|\, \theta) = \prod_i p(x_i \,|\, \theta)\,, \tag{68}$$

and the model $p(x \,|\, \theta)$ is given. Remember that the entire information about x and θ is in their joint distribution, whose density we can write down as follows (using Definition 8)

$$\underbrace{p(x, \theta)}_{\text{joint}} = \underbrace{p(x \,|\, \theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}\,. \tag{69}$$

Thus, if we know or introduce the prior distribution $p(\theta)$, we know everything about the random variables x and θ.

Namely, using the definition of the conditional distribution (or Theorem 1), we can write down the following

$$p(\theta \,|\, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D} \,|\, \theta)p(\theta)}{\int d\theta\; p(\mathcal{D} \,|\, \theta)p(\theta)}\,. \tag{70}$$

Definition 17 (Posterior Distribution)**.** *Distribution corresponding to the density $p(\theta \,|\, \mathcal{D})$ from Equation (70) is called the posterior distribution.*

Let's try choosing prior $p(\theta)$ such that we know the posterior distribution $p(\theta \,|\, \mathcal{D})$ immediately.

Example 5. *For the likelihood, consider the normal distribution with the parameter μ and fixed σ*

$$p(x \,|\, \theta) = \mathcal{N}(x \,|\, \mu, \sigma^2 = 1)\,. \tag{71}$$

From Equation (64), we already have

$$p(\mathcal{D} \,|\, \mu, \sigma = 1) \propto \mathcal{N}\left(\mu \,|\, \hat{\mu}, \frac{1}{N}\right) \propto \exp\left(-\frac{1}{2(1/\sqrt{N})^2}(\mu - \hat{\mu})^2\right), \tag{72}$$

where $\hat{\mu} = \frac{1}{N}\sum_i x_i$. The posterior distribution is defined as

$$p(\mu \,|\, \mathcal{D}) \propto \exp\left(-\frac{1}{2(1/\sqrt{N})^2}(\mu - \hat{\mu})^2\right) p(\mu)\,. \tag{73}$$

Let's choose the prior distribution $p(\mu)$ so that the functional form of the posterior distribution is the same as of the prior distribution. That is, consider $p(\mu) = \mathcal{N}(\mu \,|\, m, s^2)$

$$\begin{aligned}
p(\mu \,|\, \mathcal{D}) \propto\;& \exp\left[-\frac{1}{2(1/\sqrt{N})^2}(\mu - \hat{\mu})^2 - \frac{1}{2s^2}(\mu - m)^2\right] \propto \exp\left[-\frac{1}{2}\left(\left(N + \frac{1}{s^2}\right)\mu^2 - 2\mu\left(N\hat{\mu} + \frac{m}{s^2}\right)\right)\right] && (74)\\
=\;& \exp\left[-\frac{1}{2}\left(\frac{Ns^2+1}{s^2}\right)\left(\mu^2 - 2\mu\left(\frac{Ns^2\hat{\mu} + m}{Ns^2+1}\right)\right)\right] && (75)\\
\propto\;& \exp\left[-\frac{1}{2}\left(\frac{Ns^2+1}{s^2}\right)\left(\mu - \left(\frac{Ns^2\hat{\mu} + m}{Ns^2+1}\right)\right)^2\right] = \mathcal{N}\left(\mu \,|\, \frac{Ns^2\hat{\mu} + m}{Ns^2+1}, \frac{s^2}{Ns^2+1}\right) && (76)
\end{aligned}$$

Note that

$$\lim_{N\to\infty} \frac{Ns^2\hat{\mu} + m}{Ns^2 + 1} = \lim_{N\to\infty} \hat{\mu} = \lim_{N\to\infty} \mu_{\text{MLE}}\,, \tag{77}$$

and for large N, we have

$$\frac{s^2}{Ns^2 + 1} \approx \frac{1}{N}\,, \tag{78}$$

as we had before for the likelihood w.r.t. σ.

By finding the prior such that the posterior is easy to find we have motivated the following definition.

Definition 18 (Conjugate distribution)**.** *For the likelihood $p(x \,|\, \theta)$, if the parametric family of the posterior $p(\theta \,|\, x)$ is the same as the parametric family of the prior $p(\theta)$ these distributions are called conjugate. Oftentimes, people just say conjugate prior $p(\theta)$ to the likelihood $p(x \,|\, \theta)$.*

Example 6. *For the likelihood, consider the normal distribution with the parameter σ and fixed μ*

$$p(x \,|\, \theta) = \mathcal{N}(x \,|\, \mu = 0, \sigma^2)\,. \tag{79}$$

From before, we have

$$p(\mathcal{D} \,|\, \mu = 0, \sigma) \propto \mathcal{G}\left(\gamma \,|\, N/2 + 1, \frac{1}{2}\sum_{i=1}^{N} x_i^2\right)\,, \tag{80}$$

The posterior distribution is defined as

$$p(\gamma \,|\, \mathcal{D}) \propto \gamma^{N/2} \exp\left(-\left(\frac{1}{2}\sum_{i=1}^{N} x_i^2\right)\gamma\right) p(\gamma)\,. \tag{81}$$

Let's choose the prior distribution $p(\gamma)$ so that the functional form of the posterior distribution is the same as of the prior distribution. That is, consider $p(\gamma) = \mathcal{G}(\gamma \,|\, a, b)$

$$p(\gamma \,|\, \mathcal{D}) \propto \gamma^{N/2} \exp\left(-\left(\frac{1}{2}\sum_{i=1}^{N} x_i^2\right)\gamma\right) \gamma^{a-1}\exp(-b\gamma) \propto \mathcal{G}\left(\gamma \,|\, N/2 + a, b + \frac{1}{2}\sum_{i=1}^{N} x_i^2\right)\,. \tag{82}$$

Exercise 12. *Consider $x_1, \ldots, x_N$ — iid samples from the exponential density function*

$$p(x \,|\, \lambda) = \lambda \exp(-\lambda x)\,, \quad x \geq 0\,, \lambda > 0. \tag{83}$$

Find the conjugate prior $p(\lambda)$ and the corresponding posterior $p(\lambda \,|\, \mathcal{D})$.

Exercise 13. *Consider $x_1, \ldots, x_N$ — iid samples from the Poisson distribution Poiss(λ)*

$$P(X = k \,|\, \lambda) = \exp(-\lambda)\frac{\lambda^k}{k!}. \tag{84}$$

Find the conjugate prior $p(\lambda)$ and the corresponding posterior $p(\lambda \,|\, \mathcal{D})$.

Exercise 14. *Consider $x_1, \ldots, x_N$ — iid samples from the Bernoulli distribution Bernoulli(θ)*

$$p(x \,|\, \theta) = \theta^x (1 - \theta)^{(1-x)}. \tag{85}$$

Find the conjugate prior $p(\theta)$ and the corresponding posterior $p(\theta \,|\, \mathcal{D})$. Does it matter if instead we consider one sample from the Binomial with probability of success θ and number of trials N?

[Kirill:I'll just put this here]

Definition 19 (Beta distribution)**.** *Consider the density proportional to*

$$\mathcal{B}(x\,|\,a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1} = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}\,, \tag{86}$$

where $\Gamma(\cdot)$ and $B(\cdot)$ are the gamma and beta functions correspondingly.

When it's hard to find the entire posterior distribution, one can use its mode, which is easier to find.

Definition 20 (Maximum A posteriori (MAP))**.** *The mode of the posterior distribution is called the Maximum A posteriori (MAP) estimate and is defined as follows*

$$\theta_{MAP} = \operatorname*{argmax}_{\theta\in\Theta} p(\theta\,|\,\mathcal{D}) = \operatorname*{argmax}_{\theta\in\Theta} \frac{p(\mathcal{D}\,|\,\theta)p(\theta)}{p(\mathcal{D})} = \operatorname*{argmax}_{\theta\in\Theta} p(\mathcal{D}\,|\,\theta)p(\theta)\,. \tag{87}$$

Finally, to make new prediction, one can take into account all the possible parameters from the posterior distribution. Namely, let's say we want to get the distribution of observations x after observing the dataset $\mathcal{D}$, then we can write [Kirill:absolutely mindless mathematical tautology]

$$p(x\,|\,\mathcal{D}) = \int d\theta\; p(x,\theta\,|\,\mathcal{D}) = \int d\theta\; p(x\,|\,\theta,\mathcal{D})p(\theta\,|\,\mathcal{D})\,, \tag{88}$$

after we assume that all the information about the dataset is in our parameters θ. In other words, having the parameters θ is all we need for making the probabilistic model, i.e. $p(x\,|\,\theta,\mathcal{D}) = p(x\,|\,\theta)$. Under this strong assumption, we proceed as

$$p(x\,|\,\mathcal{D}) = \int d\theta\; p(x\,|\,\theta)p(\theta\,|\,\mathcal{D}) = \mathbb{E}_{p(\theta\,|\,\mathcal{D})}p(x\,|\,\theta)\,, \tag{89}$$

i.e. the probabilistic model of x, after observing the dataset $\mathcal{D}$ is defined as an average of probabilities/densities over all possible values of parameters sampled from the posterior distribution.

Definition 21 (Predictive Distribution)**.** *Given the likelihood $p(x\,|\,\theta)$ and the posterior $p(\theta\,|\,\mathcal{D})$, one defines the predictive distribution as follows*

$$p(x\,|\,\mathcal{D}) = \int d\theta\; p(x\,|\,\theta)p(\theta\,|\,\mathcal{D}) = \mathbb{E}_{p(\theta\,|\,\mathcal{D})}p(x\,|\,\theta)\,. \tag{90}$$

2.2 Exponential family

There is a large and important family of distributions that allows for the Bayesian inference.

Definition 22 (Exponential family)**.** *The exponential family of random variables is all random variables which density has the following functional form*

$$p(x\,|\,\theta) = \frac{f(x)}{g(\theta)}\exp\big(\theta^T u(x)\big)\,. \tag{91}$$

First, let's play around with the definition. For instance, let's see what we can get from the normalization of the density.

$$\int dx\, p(x\,|\,\theta) = \int dx\, \frac{f(x)}{g(\theta)} \exp\bigl(\theta^T u(x)\bigr) = 1\,, \tag{92}$$

$$g(\theta) = \int dx\, f(x) \exp\bigl(\theta^T u(x)\bigr)\,. \tag{93}$$

It is reasonable to call $g(\theta)$ a normalization constant of this distribution. Note that the derivative of the normalization constant has interesting properties

$$\frac{\partial}{\partial\theta_i} g(\theta) = \int dx\, f(x) \frac{\partial}{\partial\theta_i} \exp\bigl(\theta^T u(x)\bigr) = \int dx\, f(x) \exp\bigl(\theta^T u(x)\bigr) u_i(x) = g(\theta) \int dx\, p(x\,|\,\theta) u_i(x)\,, \tag{94}$$

$$\frac{\partial}{\partial\theta_i} \log g(\theta) = g(\theta) \int dx\, p(x\,|\,\theta) u_i(x) = \mathbb{E}_{p(x\,|\,\theta)} u_i(x)\,. \tag{95}$$

Thus, we have

Proposition 2. *For the density $p(x\,|\,\theta)$ from the exponential family, we have*

$$\nabla_\theta \log g(\theta) = \mathbb{E}_{p(x\,|\,\theta)} u(x)\,. \tag{96}$$

Exercise 15. *For the exponential family $p(x\,|\,\theta) = \frac{f(x)}{g(\theta)} \exp\bigl(\theta^T u(x)\bigr)$, find the expression for*

$$\frac{\partial^2}{\partial\theta_i\theta_j} \log g(\theta) = ? \tag{97}$$

Let's find the maximum likelihood estimator for this family

$$\log p(\mathcal{D}\,|\,\theta) = \sum_{i=1}^N \bigl[\log f(x_i) - \log g(\theta) + \theta^T u(x_i)\bigr]\,, \tag{98}$$

$$\nabla_\theta \log p(\mathcal{D}\,|\,\theta) = \sum_{i=1}^N [-\nabla_\theta \log g(\theta) + u(x_i)] = 0\,, \tag{99}$$

$$\nabla_\theta \log g(\theta) = \frac{1}{N} \sum_{i=1}^N u(x_i)\,. \tag{100}$$

Using Proposition 2, we have that, for θ_{MLE}, the expectation of statistics $u(x)$ equals empirical expectation of statistics $u(x)$ on the observed data, i.e.

$$\mathbb{E}_{p(x\,|\,\theta_{\text{MLE}})} u(x) = \frac{1}{N} \sum_{i=1}^N u(x_i)\,. \tag{101}$$

Note we need only values of $u(x_i)$ on our data to completely determine the maximum likelihood estimation of our parameters θ. This motivates the following definition

Definition 23 (Sufficient statistics)**.** *When the density $p(x\,|\,\theta)$ can be factorized as*

$$p(x\,|\,\theta) = p_1(x) p_2(u(x), \theta)\,, \tag{102}$$

the function $u(x)$ is called sufficient statistics.

Example 7. *Consider the normal density*

$$\mathcal{N}(x\,|\,\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Bigl(-\frac{1}{2\sigma^2}(x-\mu)^2\Bigr) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Bigl(-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\Bigr)\,. \tag{103}$$

From the definition of exponential family, we can write

$$u(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}, \quad \theta = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}, \quad g(\theta)^{-1} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) = \sqrt{\frac{-\theta_2}{\pi}} \exp\left(\frac{\theta_1^2}{4\theta_2}\right). \tag{104}$$

Thus, we have a different parameterization of the normal distribution. Let's find the expectation from this parametric form.

$$\frac{\partial}{\partial\theta_1} \log g(\theta) = \; -\frac{\theta_1}{2\theta_2} = \mu\,, \tag{105}$$

$$\frac{\partial}{\partial\theta_2} \log g(\theta) = \; -\frac{1}{2\theta_2} + \frac{\theta_1^2}{4\theta_2^2} = \sigma^2 + \mu^2\,. \tag{106}$$

Hence, we see that we can find expectations of sufficient statistics without integration [*Kirill:integration is principally much harder than differentiation*]

When we write down some density in the form of Definition 22, the parameters θ are called *natural parameters*.

Let's find the conjugate distributions to the exponential family

$$p(\theta \,|\, \mathcal{D}) \propto \prod_i p(x_i \,|\, \theta) p(\theta) = \left[\prod_i \frac{f(x_i)}{g(\theta)}\right] \exp\left(\sum_i \theta^T u(x_i)\right) p(\theta) \tag{107}$$

$$= \left[\prod_i f(x_i)\right] \exp\left(\sum_i \theta^T u(x_i)\right) \frac{p(\theta)}{g(\theta)^N}\,. \tag{108}$$

Based on this form, we can choose the prior

$$p(\theta \,|\, \eta, \nu) = \frac{1}{h(\eta,\nu)} \frac{1}{g(\theta)^\nu} \exp\left(\theta^T \eta\right). \tag{109}$$

Then

$$p(\theta \,|\, \mathcal{D}) \propto \frac{1}{g(\theta)^N} \exp\left(\sum_i \theta^T u(x_i)\right) \frac{1}{h(\eta,\nu)} \frac{1}{g(\theta)^\nu} \exp\left(\theta^T \eta\right) \propto \frac{1}{g(\theta)^{\nu+N}} \exp\left(\theta^T \left(\eta + \sum_i u(x_i)\right)\right) \tag{110}$$

$$p(\theta \,|\, \mathcal{D}) \propto p(\theta \,|\, \eta + \sum_i u(x_i), \nu + N)\,. \tag{111}$$

Exercise 16. *Represent the gamma distribution*

$$\mathcal{G}(x \,|\, a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)\,, \tag{112}$$

as a density from the exponential family, find the expectations of the sufficient statistics.

Exercise 17. *Represent the beta distribution*

$$\mathcal{B}(x \,|\, a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}\,, \tag{113}$$

as a density from the exponential family, find the expectations of the sufficient statistics.

2.3 Maximum Entropy Principle (Jaynes, 1957)

Consider a distribution over a finite set of outcomes $\{1, \ldots, M\}$ with probabilities $\{p_1, \ldots, p_M\}$. Clearly, when $p_1 = 1$ and $p_i = 0\,, i = 2, \ldots, M$ the distribution becomes degenerate and we can say for sure that all the outcomes are going to be 1. This motivates the following question: can we introduce some function $\mathcal{H}(\{p_1, \ldots, p_M\})$ that measures the 'uncertainty' of the given distribution. To answer these question let's start with the following three axioms:

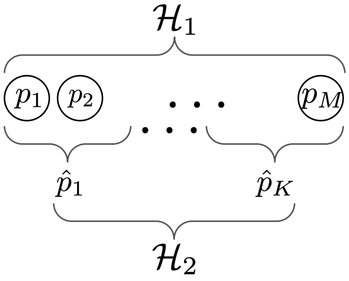

Figure 6: Invariance of the entropy w.r.t. grouping events into disjoint sets.

1. The function $\mathcal{H}(\{p_1, \dots, p_M\})$ has to be continuous w.r.t. p_i. [Kirill:indeed, we don't expect any discontinuous transitions]

2. For $p_1 = \dots = p_M = 1/M$, the function has to be increasing w.r.t. M. [Kirill:that means that more outcomes introduce more uncertainty]

3. The function $\mathcal{H}(\{p_1, \dots, p_M\})$ has to be invariant over the different outcomes in the sigma algebra [Kirill:there are multiple ways to compute the uncertainty of the same distribution by grouping the events in different groups. uncertainty has to be invariant w.r.t. all these ways.]. Indeed, consider K disjoint sets of outcomes (see Figure 6) where the total probability of events in the set k is $\hat{p}_k$. The outcomes $\{x_{k1}, \dots, x_{kM_k}\}$ within the set k define new distribution with probabilities $\{p_{k1}/\hat{p}_k, \dots, p_{kM_k}/\hat{p}_k\}$. The distribution over the sets is $\{\hat{p}_1, \dots, \hat{p}_K\}$. Thus, the uncertainty over the original events $\{1, \dots, M\}$ is can be evaluated as the uncertainty over $\{\hat{p}_1, \dots, \hat{p}_K\}$ and then the expectation over uncertainties within every subgroup, i.e.

$$\mathcal{H}(\{\hat{p}_1, \dots, \hat{p}_K\}) + \sum_{k=1}^{K} \hat{p}_k \mathcal{H}(\{p_{k1}/\hat{p}_k, \dots, p_{kM_k}/\hat{p}_k\}) = \mathcal{H}(\{p_1, \dots, p_M\})\,. \tag{114}$$

Let's consider a distribution $\{p_1, \dots, p_M\}$, where the probabilities are given as rational numbers

$$p_i = \frac{m_i}{\sum_{j=1}^{M} m_j}\,. \tag{115}$$

Let's 'ungroup' every event x_i into m_i events with uniform probabilities and use the invariance axiom

$$\mathcal{H}(\{p_1, \dots, p_M\}) + \sum_{i=1}^{M} p_i \mathcal{H}(\{1/m_i, \dots, 1/m_i\}) = \mathcal{H}(\{1/\sum_{i=1}^{M} m_i, \dots, 1/\sum_{i=1}^{M} m_i\})\,. \tag{116}$$

It's useful to introduce the following function

$$A(M) \coloneqq \mathcal{H}(\{1/M, \dots, 1/M\})\,, \tag{117}$$

and note that, for $m_i = m\,,\ \forall i$, Equation (116) becomes

$$A(M) + \sum_{i=1}^{M} \frac{m}{mM} A(m) = A(mM) \tag{118}$$

$$A(M) + A(m) = A(mM)\,. \tag{119}$$

Clearly, for $m = 1$,

$$A(M) + A(1) = A(M) \implies A(1) = 0\,. \tag{120}$$

For $m = M$,

$$A(m) + A(m) = A(m^2) \implies nA(m) = A(m^n)\,. \tag{121}$$

Thus, we see that $A(m) = \log m$ and using Equation (116), we have

$$\mathcal{H}(\{p_1, \dots, p_M\}) = -\sum_{i=1}^M p_i \log n_i + \log \sum_{j=1}^M \log n_j \; = -\sum_{i=1}^M p_i \log p_i\,. \tag{122}$$

This motivates the following definition.

Definition 24 (Entropy)**.** *For a distribution over discrete outcomes $\{1, \dots, M\}$ with probabilities $\{p_1, \dots, p_M\}$, entropy is defined as*

$$\mathcal{H}(\{p_1, \dots, p_M\}) = -\sum_{i=1}^M p_i \log p_i\,. \tag{123}$$

Once we introduce the measure of uncertainty, or the measure of information, we can start asking questions about the distributions with the most uncertainty or the least information. For example, consider a dataset $\mathcal{D}$ of observations $\{x_i\}_{i=1}^N$, and the statistics $u_k(x)$ evaluated on it, i.e.

$$\mathbb{E}_{x\sim\mathcal{D}} u_k(x) = \frac{1}{N}\sum_{i=1}^N u_k(x_i) = \mu_k\,, \quad k = 1, \dots, K\,. \tag{124}$$

Then we can look for *the distribution that explains the data and doesn't explain anything else*, as defined in the following definition.

Definition 25 (Maximum entropy principle)**.** *For the dataset $\mathcal{D}$ with statistics μ_k, the maximum entropy principle defines the distribution as a solution to the following optimization problem*

$$\max_p \mathcal{H}(\{p_1, \dots, p_M\})\,, \quad \textit{s.t.} \quad \sum_{i=1}^M p_i u_k(i) = \mu_k\,, k = 1, \dots, K\,. \tag{125}$$

Let's find the solution of this optimization problem. First, the Lagrangian corresponding to the constrained optimization problem is

$$\begin{aligned} \mathcal{L}(\lambda, p) =&\; \mathcal{H}(\{p_1, \dots, p_M\}) + \sum_{k=1}^K \lambda_k \left(\sum_{i=1}^M p_i u_k(i) - \mu_k\right) & (126)\\ =&\; -\sum_{i=1}^M p_i \log p_i + \sum_{k=1}^K \lambda_k \left(\sum_{i=1}^M p_i u_k(i) - \mu_k\right)\,. & (127)\end{aligned}$$

We have to solve

$$\min_\lambda \max_p \mathcal{L}(\lambda, p)\,. \tag{128}$$

Taking the derivative w.r.t. p, we have

$$\frac{\partial \mathcal{L}}{\partial p_j} = -\log p_j - 1 + \sum_k \lambda_k u_k(j) = 0 \implies p_j \propto \exp\left(\sum_k \lambda_k u_k(j)\right)\,, \tag{129}$$

$$p_j = \frac{1}{Z_\lambda}\exp\left(\sum_k \lambda_k u_k(j)\right)\,, \quad Z_\lambda = \sum_j^M \exp\left(\sum_k \lambda_k u_k(j)\right)\,. \tag{130}$$

Thus, we have

$$\min_\lambda \max_p \mathcal{L}(\lambda, p) = \min_\lambda \log Z_\lambda - \sum_{i=1}^M p_i \sum_k \lambda_k u_k(i) + \sum_{k=1}^K \lambda_k \left(\sum_{i=1}^M p_i u_k(i) - \mu_k\right) \tag{131}$$

$$= \min_\lambda -\mathbb{E}_{x\sim\mathcal{D}}\left[\sum_{k=1}^K \lambda_k u_k(x) - \log Z_\lambda\right] = \max_\lambda \mathbb{E}_{x\sim\mathcal{D}} \log p_x\,, \tag{132}$$

which, as we see, is equivalent to maximum likelihood in the exponential family.

The same reasoning applies for the continuous random variables as follows.

Definition 26 (Differential Entropy)**.** *For a continuous random variable with the density $q(x)$, differential entropy is defined as*

$$\mathcal{H}(q) = -\int dx\; q(x)\log q(x) = -\mathbb{E}_{q(x)} \log q(x)\,. \tag{133}$$

Let's see what the maximum entropy principle tells us for the differential entropy.

$$\max_q \mathcal{H}(q(x))\,, \quad \text{s.t.} \quad \int dx\; q(x) u_n(x) = \mu_n\,, n = 1, \ldots, N\,. \tag{134}$$

The Lagrangian corresponding to the problem is

$$\mathcal{L}(q, \lambda) = \mathcal{H}(q) + \sum_{n=1}^N \lambda_n \left(\int dx\; q(x) u_n(x) - \mu_n\right) \tag{135}$$

$$= -\int dx\; q(x)\log q(x) + \sum_{n=1}^N \lambda_n \left(\int dx\; q(x) u_n(x) - \mu_n\right)\,, \tag{136}$$

and the dual optimization problem is

$$\min_\lambda \max_q \mathcal{L}(q, \lambda)\,. \tag{137}$$

Let's write down the extremum condition for the inner optimization problem $\min_q \mathcal{L}(q, \lambda)$

$$\frac{\delta\mathcal{L}}{\delta q} = -\log q(x) + \sum_n \lambda_n u_n(x) = 0 \implies q(x) \propto \exp\left(\sum_n \lambda_n u_n(x)\right)\,. \tag{138}$$

From the normalization condition, we have

$$q(x\,|\,\lambda) = \frac{1}{Z_\lambda}\exp\left(\sum_n \lambda_n u_n(x)\right)\,, \quad Z_\lambda = \int dx\; \exp\left(\sum_n \lambda_n u_n(x)\right)\,. \tag{139}$$

$$\mathcal{L}(q, \lambda) = \log Z_\lambda - \int dx\; q(x\,|\,\lambda)\left(\sum_n \lambda_n u_n(x)\right) + \sum_{n=1}^N \lambda_n \left(\int dx\; q(x\,|\,\lambda) u_n(x) - \mu_n\right) \tag{140}$$

$$= \log Z_\lambda - \sum_{n=1}^N \lambda_n \mu_n = \log Z_\lambda - \sum_{n=1}^N \lambda_n \mathbb{E}_{x\sim\mathcal{D}} \mu_n(x) = -\mathbb{E}_{x\sim\mathcal{D}} \log q(x\,|\,\lambda)\,. \tag{141}$$

Thus, we have the following optimization problem

$$\min_\lambda \max_q \mathcal{L}(q, \lambda) = \min_\lambda -\log q(\mathcal{D}\,|\,\lambda) = \max_\lambda \log q(\mathcal{D}\,|\,\lambda)\,, \tag{142}$$

and we can make the following statement.

Theorem 3. *Maximum entropy principle corresponds to the maximum likelihood in the exponential family.*

Exercise 18. *Find the density $q(x)$ of the random variable defined on $x \geq 0$ that has the biggest entropy and a given expectation μ .*

Exercise 19. *Find the density $q(x)$ of the random variable defined on $x \in \mathbb{R}$ that has the biggest entropy, given expectation μ, and given variance σ^2 .*

2.4 Bayesian Model Selection

Design of the probabilistic model is the first step of Bayesian reasoning and oftentimes the choice of the model is not obvious. We considered the priors that give us answers that are easy to compute, but what if a different prior fits better our purposes or which parameters of the prior should we choose? Let's look how the Bayesian inference looks like if we introduce model variable m which we use to control the choice of the prior distribution (could the functional form or its parameters or both), i.e.

$$p(\theta \,|\, \mathcal{D}, m) = \frac{p(\mathcal{D} \,|\, \theta) p(\theta \,|\, m)}{p(\mathcal{D} \,|\, m)} \,. \tag{143}$$

Note that the denominator, which we usually refer to as a normalization constant of the posterior, i.e.

$$p(\mathcal{D} \,|\, m) = \int d\theta \; p(\mathcal{D} \,|\, \theta) p(\theta \,|\, m) \,, \tag{144}$$

is actually a likelihood of m (indeed, it is the conditional distribution of the observed data $\mathcal{D}$ for model m). Then we can choose the model according to the maximum likelihood principle (see Definition 14), i.e.

$$m_{\text{MLE}} = \operatorname*{argmax}_m \underbrace{p(\mathcal{D} \,|\, m)}_{\text{evidence}} \,, \tag{145}$$

where the quantity $p(\mathcal{D} \,|\, m)$ is called *evidence*. After choosing the model, we can define our posterior distribution as

$$p(\theta \,|\, \mathcal{D}, m_{\text{MLE}}) \,. \tag{146}$$

Alternatively, one can do the Bayesian inference on the model variable [Kirill:indeed, Bayesian inference is a way of reasoning, which can be applied to any variables. we can keep on doing this hierarchical inference any number of times]. That is, we assume the uniform distribution over models $p(m) = \text{Uniform}[M]$ (let's say we have M different models) [Kirill:of course, we can introduce more complicated priors]. Then the posterior is the marginalization over m, i.e.

$$p(\theta \,|\, \mathcal{D}) = \sum_m p(\theta, m \,|\, \mathcal{D}) = \sum_m p(\theta \,|\, \mathcal{D}, m) p(m \,|\, \mathcal{D}) \,, \quad \text{where} \quad p(m \,|\, \mathcal{D}) \propto p(\mathcal{D} \,|\, m) p(m) \propto p(\mathcal{D} \,|\, m) \,. \tag{147}$$

Note that the posterior becomes the mixture of posteriors under different models m weighted proportionally to their evidence $p(\mathcal{D} \,|\, m)$.

Example 8 (Model selection)**.** *Consider three random variables X, Y, Z which joint observations are given in the following table.*

	$x = 0$		$x = 1$	
	$z = 0$	$z = 1$	$z = 0$	$z = 1$
$y = 0$	132	19	52	11
$y = 1$	9	0	97	6

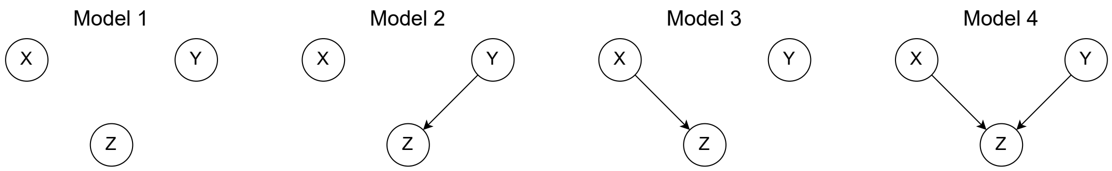

Figure 7: Models for Example 8.

2.5 Bayesian Linear Regression

Let's consider the classical ML example — linear regression. The likelihood for linear regression is given as follows

$$p(y \,|\, x, \theta) = \mathcal{N}(y \,|\, w^T x, \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2}(y - w^T x)^2\right). \tag{148}$$

It's useful to think in terms of the log-likelihood, though, i.e.

$$\log p(y \,|\, x, \theta) = \log \mathcal{N}(y \,|\, w^T x, \beta^{-1}) = -\frac{\beta}{2}(y - w^T x)^2 + \frac{1}{2}\log\frac{\beta}{2\pi}. \tag{149}$$

First, let's find the maximum likelihood estimator $\theta_{\text{ML}} = \text{argmax}_\theta \log p(\mathcal{D} \,|\, x, \theta)$ for the dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$. That is, the derivative of the likelihood is

$$\nabla_w \sum_i \log \mathcal{N}(y_i \,|\, w^T x_i, \beta^{-1}) = -\nabla_w \sum_i \frac{\beta}{2}(y_i - w^T x_i)^2 = \beta \sum_i (y_i - w^T x_i) x_i = 0\,, \tag{150}$$

where $x_i \in \mathbb{R}^d$ is a feature-vector of the i-th object. It's convenient for us to introduce matrix X such that X_{ij} is the j-th coordinate of vector x_i (of i-th object) and vector y that consists of all the labels y_i. Thus, we have

$$\sum_i y_i X_{ik} - \sum_i \sum_j w_j X_{ij} X_{ik} = 0\,,\ \forall\, k\,, \tag{151}$$

$$X^T y - X^T X w = 0 \implies w_{\text{MLE}} = (X^T X)^{-1} X^T y\,. \tag{152}$$

Note that we got the classic formula for the linear regression. Analogously, we have

$$\frac{\partial}{\partial \beta}\left[\sum_i \log \mathcal{N}(y_i \,|\, w^T x_i, \beta^{-1})\right] = -\sum_{i=1}^N \frac{1}{2}(y_i - w^T x_i)^2 + \frac{N}{2\beta} = 0 \tag{153}$$

$$\frac{1}{\beta_{\text{MLE}}} = \frac{1}{N}\sum_{i=1}^N (y_i - w^T x_i)^2 = \frac{1}{N}\|y - Xw\|^2\,. \tag{154}$$

Now, let's introduce the prior distribution for the weights w. That is, we assume that the weights are sampled (apriori) from the Normal distribution with the diagonal covariance matrix as follows

$$w \sim \mathcal{N}(w \,|\, 0, A^{-1})\,, \quad \text{where} \quad A = \begin{bmatrix} \alpha_1 & & 0 \\ & \ddots & \\ 0 & & \alpha_d \end{bmatrix}. \tag{155}$$

The posterior distribution for this model is defined as

$$p(\theta \,|\, \mathcal{D}) \propto \prod_{i=1}^N p(y_i \,|\, x_i, \theta) p(\theta) = \prod_{i=1}^N \mathcal{N}(y_i \,|\, w^T x_i, \beta^{-1}) \mathcal{N}(w \,|\, 0, A^{-1})\,. \tag{156}$$

$$F(w) \coloneqq \prod_{i=1}^N \mathcal{N}(y_i \,|\, w^T x_i, \beta^{-1}) \mathcal{N}(w \,|\, 0, A^{-1})\,. \tag{157}$$

In this context, it is much easier to work with the logarithm of the density rather than with the density itself. Thus, we have

$$\begin{aligned}\log F(w) = & \sum_{i=1}^{N} \log \mathcal{N}(y_i \,|\, w^T x_i, \beta^{-1}) + \log \mathcal{N}(w \,|\, 0, A^{-1}) && (158)\\ = & -\sum_{i=1}^{N} \frac{\beta}{2}(y_i - w^T x_i)^2 + \frac{N}{2}\log\frac{\beta}{2\pi} - \frac{1}{2}w^T A w + \frac{1}{2}\log\det(A) - \frac{d}{2}\log 2\pi && (159)\\ = & -\frac{\beta}{2}\|y - Xw\|^2 + \frac{N}{2}\log\beta - \frac{1}{2}w^T A w + \frac{1}{2}\log\det(A) - \frac{d+N}{2}\log 2\pi\,. && (160)\end{aligned}$$

Note that this a quadratic function, which means that the Taylor expansion of this function is going to contain only the first 3 terms, i.e.

$$\begin{aligned}\log F(w) = & \log F(w_{\text{MAP}}) + \langle \nabla \log F(w_{\text{MAP}}), w - w_{\text{MAP}}\rangle + && (161)\\ & + \frac{1}{2}(w - w_{\text{MAP}})^T \nabla^2_{ww} \log F(w_{\text{MAP}})(w - w_{\text{MAP}}) + \underbrace{\frac{\partial^3}{\partial w^3} + \ldots}_{=0}\,. && (162)\end{aligned}$$

Note that we make the expansion at the point w_{MAP} which is the Maximum A posteriori estimate, i.e.

$$w_{\text{MAP}} = \operatorname*{argmax}_{w} \log F(w)\,. \tag{163}$$

Hence, $\nabla \log F(w_{\text{MAP}}) = 0$ is the necessary condition for being an optimum. Thus, we have

$$\log F(w) = \log F(w_{\text{MAP}}) + \frac{1}{2}(w - w_{\text{MAP}})^T \nabla^2_{ww} \log F(w_{\text{MAP}})(w - w_{\text{MAP}})\,. \tag{164}$$

Let's find w_{MAP} first. Clearly, we have

$$\nabla_w \log F(w) = \beta X^T(y - Xw) - Aw = 0 \tag{165}$$

$$\beta X^T y - \beta X^T X w - Aw = 0 \implies w_{\text{MAP}} = \beta(\beta X^T X + A)^{-1} X^T y = (X^T X + \beta^{-1} A)^{-1} X^T y\,, \tag{166}$$

And the second derivative is

$$\nabla^2_{ww} \log F(w) = -\beta X^T X - A\,. \tag{167}$$

Remember that

$$p(w \,|\, \mathcal{D}) \propto F(w) \implies \log p(w \,|\, \mathcal{D}) = \log F(w) + \text{const}\,, \tag{168}$$

where the constant can be defined from the normalization of the density (indeed, $p(w \,|\, \mathcal{D})$ has to be a valid density). However, taking this integral is absolutely unnecessary since we already see that

$$p(w \,|\, \mathcal{D}) \propto \exp\left(-\frac{1}{2}(w - w_{\text{MAP}})^T (\beta X^T X + A)(w - w_{\text{MAP}})\right)\,. \tag{169}$$

Thus, we have the following posterior distribution

$$p(w \,|\, \mathcal{D}) = \mathcal{N}\left(w \,|\, \underbrace{(X^T X + \beta^{-1} A)^{-1} X^T y}_{w_{\text{MAP}}}, (\beta X^T X + A)^{-1}\right)\,. \tag{170}$$

Let's find the evidence

$$p(\mathcal{D} \,|\, \alpha, \beta) = \int dw\; p(\mathcal{D} \,|\, w, \beta) p(w \,|\, \alpha)\,. \tag{171}$$

Clearly, using the previously introduced notation, we have

$$
\begin{aligned}
p(\mathcal{D}\,|\,\alpha,\beta) = & \int dw\; F(w) = F(w_{\text{MAP}}) \int dw\;\; \exp\left[-\frac{1}{2}(w-w_{\text{MAP}})^T(\beta X^T X + A)(w-w_{\text{MAP}})\right] && (172)\\
= &\; F(w_{\text{MAP}}) \frac{(2\pi)^{d/2}}{\sqrt{\det(\beta X^T X + A)}} && (173)
\end{aligned}
$$

Thus, we have to find $F(w_{\text{MAP}})$. Let's do it

$$
\begin{aligned}
\log F(w_{\text{MAP}}) + \text{const} &= -\frac{\beta}{2}\|y - Xw_{\text{MAP}}\|^2 - \frac{1}{2} w_{\text{MAP}}^T A w_{\text{MAP}} && (174)\\
&= -\frac{\beta}{2}\big(y^T y - 2y^T X w_{\text{MAP}} + w_{\text{MAP}}^T X^T X w_{\text{MAP}}\big) - \frac{1}{2} w_{\text{MAP}}^T A w_{\text{MAP}} && (175)\\
&= -\frac{\beta}{2}\big(y^T y - 2y^T X w_{\text{MAP}}\big) - \frac{1}{2} w_{\text{MAP}}^T \big(\beta X^T X + A\big) w_{\text{MAP}} && (176)\\
&= -\frac{\beta}{2}\Big(y^T y - 2\beta y^T X \big(\beta X^T X + A\big)^{-1} X^T y\Big) - \frac{\beta^2}{2} y^T X \big(\beta X^T X + A\big)^{-1} X^T y && (177)\\
&= -\frac{\beta}{2} y^T \Big(I - \beta X \big(\beta X^T X + A\big)^{-1} X^T\Big) y && (178)\\
&= -\frac{\beta}{2} y^T \Big(I - X \big(X^T X + \beta^{-1} A\big)^{-1} X^T\Big) y && (179)
\end{aligned}
$$

For the last expression, we need the Woodbury matrix identity

$$
(A + UCV)^{-1} = A^{-1} - A^{-1} U \big(C^{-1} + V A^{-1} U\big)^{-1} V A^{-1}\,. \tag{180}
$$

Thus, we have

$$
\log F(w_{\text{MAP}}) = -\frac{\beta}{2} y^T \big(I + \beta^{-1} X A^{-1} X^T\big)^{-1} y + \frac{N}{2}\log\beta + \frac{1}{2}\log\det(A) - \frac{d+N}{2}\log 2\pi\,. \tag{181}
$$

Finally, the evidence is

$$
p(\mathcal{D}\,|\,\alpha,\beta) = \frac{(2\pi)^{d/2}\beta^{N/2}\sqrt{\det(A)}}{(2\pi)^{(d+N)/2}\sqrt{\det(\beta X^T X + A)}} \exp\Big(-\frac{\beta}{2} y^T \big(I + \beta^{-1} X A^{-1} X^T\big)^{-1} y\Big)\,, \tag{182}
$$

for which, we can use

$$
\det(A + UV) = \det(I + V A^{-1} U)\det(A) \tag{183}
$$

and get

$$
\beta^d \det\big(X^T X + \beta^{-1} A\big) = \beta^d \det(\beta^{-1} A)\det(I + \beta^{-1} X A X^T) = \det(A)\det(I + \beta^{-1} X A X^T)\,. \tag{184}
$$

Thus, we have

$$
p(\mathcal{D}\,|\,\alpha,\beta) = \frac{\beta^{N/2}}{(2\pi)^{N/2}\sqrt{\det(I + \beta^{-1} X A X^T)}} \exp\Big(-\frac{\beta}{2} y^T \big(I + \beta^{-1} X A^{-1} X^T\big)^{-1} y\Big)\,, \tag{185}
$$

which we can then optimize as follows

$$
\alpha^*, \beta^* = \underset{\alpha,\beta}{\operatorname{argmax}}\; p(\mathcal{D}\,|\,\alpha,\beta)\,. \tag{186}
$$

To find the optimum of this function, it is easier to consider the variational bound of this function

Definition 27 (Variational Bound)**.** *We call the function $g(x,\xi)$ a variational bound of function $f(x)$ if*

1. $\forall\, x, \xi \;\; f(x) \geq g(x,\xi)\,,$
2. $\forall\, x\; \exists\, \xi_x \;: f(x) = g(x,\xi_x)\,.$

$$
\begin{aligned}
\log p(\mathcal{D} \mid \alpha, \beta) = & -\frac{\beta}{2}\|y - Xw_{\text{MAP}}\|^2 - \frac{1}{2}w_{\text{MAP}}^T A w_{\text{MAP}} + & (187)\\
& + \frac{N}{2}\log\beta + \frac{1}{2}\log\det(A) - \frac{N}{2}\log 2\pi - \frac{1}{2}\log\det(\beta X^T X + A) & (188)\\
\geq & -\frac{\beta}{2}\|y - Xw\|^2 - \frac{1}{2}w^T A w + & (189)\\
& + \frac{N}{2}\log\beta + \frac{1}{2}\log\det(A) - \frac{N}{2}\log 2\pi - \frac{1}{2}\log\det(\beta X^T X + A) \coloneqq \hat{F}(w, \alpha, \beta) & (190)
\end{aligned}
$$

$$
\frac{\partial}{\partial w}\hat{F}(w, \alpha, \beta) = -\beta X^T(y - Xw) = 0 \implies w = w_{\text{MAP}}\,. \tag{191}
$$

$$
\begin{aligned}
\frac{\partial}{\partial \alpha_j}\hat{F}(w, \alpha, \beta) = & -\frac{1}{2}w_j^2 + \frac{1}{2}\frac{1}{\alpha_j} - \frac{1}{2}\left\langle(\beta X^T X + A)^{-1}, \frac{\partial}{\partial \alpha_j}(\beta X^T X + A)\right\rangle & (192)\\
= & -\frac{1}{2}w_j^2 + \frac{1}{2}\frac{1}{\alpha_j} - \frac{1}{2}(\beta X^T X + A)^{-1}_{jj} = 0 & (193)\\
\alpha_j = & \frac{1}{w_j^2 + (\beta X^T X + A)^{-1}_{jj}} & (194)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial \beta}\hat{F}(w, \alpha, \beta) = & -\frac{1}{2}\|y - Xw\|^2 + \frac{N}{2\beta} - \frac{1}{2}\left\langle(\beta X^T X + A)^{-1}, \frac{\partial}{\partial \beta}(\beta X^T X + A)\right\rangle & (195)\\
= & -\frac{1}{2}\|y - Xw\|^2 + \frac{N}{2\beta} - \frac{1}{2\beta}\langle(\beta X^T X + A)^{-1}, \beta X^T X \pm A\rangle & (196)\\
= & -\frac{1}{2}\|y - Xw\|^2 + \frac{N}{2\beta} - \frac{1}{2\beta}(d - \sum_j \alpha_j(\beta X^T X + A)^{-1}_{jj}) = 0 & (197)\\
\beta = & \frac{N - d + \sum_j \alpha_j(\beta X^T X + A)^{-1}_{jj}}{\|y - Xw\|^2} & (198)
\end{aligned}
$$

Exercise 20. *Find the predictive distribution, i.e.*

$$
p(y \mid x, \alpha, \beta) = \int dw\, \mathcal{N}(y \mid w^T x, \beta^{-1})p(w \mid \mathcal{D}) = ? \tag{199}
$$

2.6 Bayesian Logistic Regression, Laplace Approximation

Consider the Logistic Regression for classification problem $y \in \{-1, 1\}$

$$
p(y \mid x, w) = \frac{1}{1 + \exp(-yw^T x)} \tag{200}
$$

With the normal prior

$$
p(w) = \mathcal{N}(w \mid 0, A^{-1})\,, \quad \text{where} \quad A = \begin{bmatrix} \alpha_1 & & 0 \\ & \ddots & \\ 0 & & \alpha_d \end{bmatrix}. \tag{201}
$$

The posterior distribution is

$$
p(w \mid \mathcal{D}) \propto \prod_i p(y_i \mid x_i, w)\mathcal{N}(w \mid 0, A^{-1}) \tag{202}
$$

$$\log p(w\,|\,\mathcal{D}) + \log Z = \sum_i \log p(y_i\,|\,x_i, w) - \frac{1}{2}w^T A w + \frac{1}{2}\log\det(A) - \frac{d}{2}\log(2\pi) \tag{203}$$

$$= -\sum_i \log(1+\exp(-y_i w^T x_i)) - \frac{1}{2}w^T A w + \frac{1}{2}\log\det(A) - \frac{d}{2}\log(2\pi) := \log F(w)\,. \tag{204}$$

Let's do the same expansion as we did for the linear regression

$$\log F(w) = \log F(w_{\text{MAP}}) + \langle \nabla \log F(w_{\text{MAP}}), w - w_{\text{MAP}}\rangle + \tag{205}$$

$$+ \frac{1}{2}(w - w_{\text{MAP}})^T \frac{\partial^2}{\partial w^2}\log F(w_{\text{MAP}})(w - w_{\text{MAP}}) + \underbrace{\frac{\partial^3}{\partial w^3} + \dots}_{\neq 0}\,. \tag{206}$$

The last terms are not zero anymore, but this is an approximation. Even for this, it's not trivial to find w_{MAP}

$$\nabla_w \log F(w) = -\sum_i \frac{\exp(-y_i w^T x_i)}{1+\exp(-y_i w^T x_i)}(-y_i x_i) - Aw = \sum_i \frac{1}{1+\exp(y_i w^T x_i)} y_i x_i - Aw \tag{207}$$

$$\frac{\partial^2}{\partial w^2}\log F(w) = -\sum_i \frac{\exp(y_i w^T x_i)}{(1+\exp(y_i w^T x_i))^2} x_i x_i^T - A = -X^T R(w) X - A\,, \tag{208}$$

where $R(w)$ is the diagonal matrix with the elements

$$R(w)_{ii} = \frac{\exp(y_i w^T x_i)}{(1+\exp(y_i w^T x_i))^2}\,. \tag{209}$$

Thus, we approximate the posterior distribution as follows

$$p(w\,|\,D) \propto F(w) \propto \mathcal{N}(w\,|\,w_{\text{MAP}}, (X^T R(w_{\text{MAP}})X + A)^{-1})\,. \tag{210}$$

Of course, this approximation is not specific to the logistic regression but a general technique that can be summarized as follows.

Definition 28 (Laplace approximation)**.** *For a given density $p(x)$ the Laplace approximation is*

$$p(x) \approx \mathcal{N}(x\,|\,\mu, \Sigma)\,,\ \ \mu = \operatorname*{argmax}_x p(x)\,,\ \ \Sigma^{-1} = -\nabla^2_{xx}\log p(x)\,. \tag{211}$$

$$p(\mathcal{D}\,|\,A) = \int dw\ \prod_i p(y_i\,|\,x_i, w)\mathcal{N}(w\,|\,0, A^{-1}) \tag{212}$$

3 Bayesian Variational Inference

3.1 Kullback-Leibler divergence

The philosophy of the Bayesian variational inference is "if we cannot find the distribution analytically because it is too complicated, let us approximate it with a simpler distribution by minimizing some notion of discrepancy between distributions."

There are numerous ways for measuring discrepancy or divergence between distributions and all of them have their pros and cons. One of the most important ways comes from the information theory and is defined as follows.

Definition 29 (Kullback-Leibler divergence)**.** *For two discrete distributions p and q defined on the same set of outcomes $\{1, \ldots, M\}$, the Kullback-Leibler (KL) divergence is defined as*

$$D_{\mathrm{KL}}(p, q) = \sum_{i=1}^{M} p_i \log \frac{p_i}{q_i}\,. \tag{213}$$

Analogously, for two densities $p(x)$ and $q(x)$, and $x \in \Omega$, we have

$$D_{\mathrm{KL}}(p(x), q(x)) = \int_{\Omega} dx\; p(x) \log \frac{p(x)}{q(x)}\,. \tag{214}$$

[Kirill:it is tightly related to the concept of entropy and can be interpreted as the number of excess bits required for encoding the distribution $p(x)$ with the distribution $q(x)$.]

One can easily see that the KL-divergence is always positive

$$\begin{aligned} D_{\mathrm{KL}}(p(x), q(x)) &= \int_{\Omega} dx\; p(x) \log \frac{p(x)}{q(x)} && (215)\\ &= -\int_{\Omega} dx\; p(x) \log \frac{q(x)}{p(x)} && (216)\\ &\geq -\log\left(\int_{\Omega} dx\; p(x) \frac{q(x)}{p(x)}\right) = -\log 1 = 0\,, && (217) \end{aligned}$$

where we used the Jensen's inequality for probabilities. Namely, for any convex function $f(x)$ and any function $\varphi(x)$, we have

$$\int dx\; p(x) f(\varphi(x)) \geq f\left(\int dx\; p(x)\varphi(x)\right). \tag{218}$$

Furthermore, if $q(x) = p(x)$ it equals zero, indeed

$$D_{\mathrm{KL}}(p(x), p(x)) = \int_{\Omega} dx\; p(x) \log \frac{p(x)}{p(x)} = \int_{\Omega} dx\; p(x) \log 1 = \log 1 = 0\,. \tag{219}$$

Finally, one can prove that if $D_{\mathrm{KL}}(p(x), q(x)) = 0$ then $p(x) = q(x)$.

Example 9. *Assume we observe some data $\mathcal{D} = \{x_1, \ldots, x_N\}$ that was generated from the distribution $x \sim p(x)$. Let's approximate $p(x)$ by solving*

$$\theta^* = \operatorname*{argmin}_{\theta} D_{\mathrm{KL}}(p(x), q(x \,|\, \theta))\,. \tag{220}$$

First, let's rewrite the KL-divergence as follows

$$D_{\mathrm{KL}}(p(x), q(x \,|\, \theta)) = \int dx\; p(x) \log \frac{p(x)}{q(x \,|\, \theta)} = \int dx\; p(x) \log p(x) - \int dx\; p(x) \log q(x \,|\, \theta)\,. \tag{221}$$

Note that the first term does not depend on θ. Thus, we have

$$\operatorname*{argmin}_{\theta} D_{\mathrm{KL}}(p(x), q(x \,|\, \theta)) = \operatorname*{argmin}_{\theta} - \int dx\; p(x) \log q(x \,|\, \theta) = \operatorname*{argmax}_{\theta} \mathbb{E}_{x \sim p(x)} \log q(x \,|\, \theta)\,. \tag{222}$$

In other words, we see that minimizing $D_{\mathrm{KL}}(p(x), q(x \,|\, \theta))$ is equivalent to the maximum likelihood principle from Definition 14.

Exercise 21. *Find the KL-divergence between two multivariate Gaussians with the same covariance matrix but different expectations, i.e.*

$$D_{\mathrm{KL}}(\mathcal{N}(\mu_1, \Sigma), \mathcal{N}(\mu_2, \Sigma)) = ? \tag{223}$$

Exercise 22 (Mode covering/seeking behaviour)**.** *Find the mean of the normal distribution that minimizes the KL-divergence between a Gaussian and some density $p(x)$, i.e.*

$$\operatorname*{argmin}_{\mu} D_{\mathrm{KL}}(p(x), \mathcal{N}(x \,|\, \mu, \Sigma)) = ? \tag{224}$$

How the problem changes if we flip the arguments of the KL-divergence, i.e. we consider

$$\operatorname*{argmin}_{\mu} D_{\mathrm{KL}}(\mathcal{N}(x \,|\, \mu, \Sigma), p(x)) = ? \tag{225}$$

Exercise 23. *Prove*

$$D_{\mathrm{KL}}(q, p) \leq \chi^2(q, p)\,, \quad \textit{where} \quad \chi^2(q, p) = \mathbb{E}_{p(x)} \left(\frac{q(x)}{p(x)} - 1 \right)^2 . \tag{226}$$

Exercise 24. *Prove that for any p_1, q_1, p_2, q_2 and $\alpha \in [0, 1]$, we have*

$$D_{\mathrm{KL}}(\alpha q_1 + (1-\alpha) q_2, \alpha p_1 + (1-\alpha) p_2) \leq \alpha D_{\mathrm{KL}}(q_1, p_1) + (1-\alpha) D_{\mathrm{KL}}(q_2, p_2)\,. \tag{227}$$

3.2 Evidence Lower Bound (ELBO)

Let's take any approximation $q(\theta)$ and evaluate the KL-divergence between this approximation and the posterior distribution. First, we just write the definition of the KL-divergence and the definition of the posterior distribution, i.e.

$$\begin{aligned} D_{\mathrm{KL}}(q(\theta), p(\theta \,|\, \mathcal{D})) &= \int d\theta\; q(\theta) \log \frac{q(\theta)}{p(\theta \,|\, \mathcal{D})} = \int d\theta\; q(\theta) \log \frac{q(\theta) p(\mathcal{D})}{p(\mathcal{D} \,|\, \theta) p(\theta)} &(228)\\ &= -\int d\theta\; q(\theta) \log p(\mathcal{D} \,|\, \theta) + \int d\theta\; q(\theta) \log \frac{q(\theta)}{p(\theta)} + \log p(\mathcal{D})\,. &(229) \end{aligned}$$

This is what we want to minimize. However, usually people take few more steps, i.e.

$$\begin{aligned} D_{\mathrm{KL}}(q(\theta), p(\theta \,|\, \mathcal{D})) &= -\mathbb{E}_{q(\theta)} \log p(\mathcal{D} \,|\, \theta) + D_{\mathrm{KL}}(q(\theta), p(\theta)) + \log p(\mathcal{D}) &(230)\\ \log p(\mathcal{D}) &= \mathbb{E}_{q(\theta)} \log p(\mathcal{D} \,|\, \theta) - D_{\mathrm{KL}}(q(\theta), p(\theta)) + D_{\mathrm{KL}}(q(\theta), p(\theta \,|\, \mathcal{D})) &(231)\\ \log \underbrace{p(\mathcal{D})}_{\text{evidence}} &\geq \underbrace{\mathbb{E}_{q(\theta)} \log p(\mathcal{D} \,|\, \theta) - D_{\mathrm{KL}}(q(\theta), p(\theta))}_{\text{Evidence Lower Bound}}\,, &(232) \end{aligned}$$

where the last transition follows from the positivity of the KL-divergence.

It's a good moment to recall Definition 27 and see the following fact.

Theorem 4. *Evidence Lower Bound is a variational bound.*

Proof. Indeed, we already showed that

$$\forall \mathcal{D}\,, q(\theta) \quad \log p(\mathcal{D}) \geq \; \mathbb{E}_{q(\theta)} \log p(\mathcal{D} \,|\, \theta) - D_{\mathrm{KL}}(q(\theta), p(\theta))\,. \tag{233}$$

Thus, we just have to show the second requirement for a variational bound. Namely, we have to show

$$\forall \mathcal{D}\,, \exists\, q(\theta) \; : \log p(\mathcal{D}) = \; \mathbb{E}_{q(\theta)} \log p(\mathcal{D} \,|\, \theta) - D_{\mathrm{KL}}(q(\theta), p(\theta))\,. \tag{234}$$

From Equation (231), we can easily see that the equality is achieved at $q(\theta) = p(\theta \,|\, \mathcal{D})$, i.e. by perfectly fitting the posterior. Then, we have

$$\mathbb{E}_{p(\theta \,|\, \mathcal{D})} \log p(\mathcal{D} \,|\, \theta) - D_{\mathrm{KL}}(p(\theta \,|\, \mathcal{D}), p(\theta)) \; = \int d\theta \; p(\theta \,|\, \mathcal{D}) \log \frac{p(\theta)p(\mathcal{D} \,|\, \theta)}{p(\theta \,|\, \mathcal{D})} \tag{235}$$

$$= \int d\theta \; p(\theta \,|\, \mathcal{D}) \log p(\mathcal{D}) = \log p(\mathcal{D})\,. \tag{236}$$

□

Let's say we have parameterized the variational posterior $q(\theta \,|\, \varphi)$. Then, we can find the optimal parameters φ^* by solving the following optimization problem

$$\varphi^* = \operatorname*{argmax}_{\varphi} \mathbb{E}_{q(\theta \,|\, \varphi)} \log p(\mathcal{D} \,|\, \theta) - D_{\mathrm{KL}}(q(\theta \,|\, \varphi), p(\theta))\,, \tag{237}$$

i.e. by maximizing the Evidence Lower Bound. Usually we choose $q(\theta \,|\, \varphi)$ and $p(\theta)$ such that the KL-divergence can be found in the analytic form. However, the first term $\mathbb{E}_{q(\theta \,|\, \varphi)} \log p(\mathcal{D} \,|\, \theta)$ always requires a numerical estimation of the gradient (in order to use the gradient-based optimization algorithms). Thus, we have a question how de we estimate the gradient $\nabla_\varphi \mathbb{E}_{q(\theta \,|\, \varphi)} f(\theta)$.

Taking the gradient w.r.t. the parameters of the distribution is a big problem in probabilistic modeling

$$\nabla_\varphi \mathbb{E}_{q(x \,|\, \varphi)} f(x) =? \tag{238}$$

We will consider two approaches to this. The first one is straightforward

$$\nabla_\varphi \mathbb{E}_{q(x \,|\, \varphi)} f(x) = \; \int dx \; \nabla_\varphi q(x \,|\, \varphi) f(x) = \int dx \; q(x \,|\, \varphi) f(x) \nabla_\varphi \log q(x \,|\, \varphi) \tag{239}$$

$$= \; \mathbb{E}_{q(x \,|\, \varphi)} f(x) \nabla_\varphi \log q(x \,|\, \varphi)\,. \tag{240}$$

Thus, we get the **score function gradient estimator**. Note that

$$\mathbb{E}_{q(x \,|\, \varphi)} \nabla_\varphi \log q(x \,|\, \varphi) = \int dx \; \nabla_\varphi q(x \,|\, \varphi) = \nabla_\varphi \int dx \; q(x \,|\, \varphi) = \nabla_\varphi 1 = 0\,. \tag{241}$$

Thus, we can introduce so-called **baseline** as follows

$$\nabla_\varphi \mathbb{E}_{q(x \,|\, \varphi)} f(x) = \mathbb{E}_{q(x \,|\, \varphi)} f(x) \nabla_\varphi \log q(x \,|\, \varphi) = \mathbb{E}_{q(x \,|\, \varphi)} (f(x) - b) \nabla_\varphi \log q(x \,|\, \varphi)\,, \tag{242}$$

where $b \in \mathbb{R}$ is a constant w.r.t. x. In practice, the average value of f minimizes the variance of the estimator, i.e.

$$\nabla_\varphi \mathbb{E}_{q(x \,|\, \varphi)} f(x) \simeq \frac{1}{N} \sum_{i=1}^N \left(f(x_i) - \frac{1}{N-1} \sum_{j=1, j\neq i}^N f(x_j) \right) \nabla_\varphi \log q(x_i \,|\, \varphi)\,. \tag{243}$$

Exercise 25. *Prove that the estimator in Equation (243) is unbiased.*

The second approach assumes that we can reparameterize $q(x \,|\, \varphi)$ as follows

$$g(\varepsilon; \varphi) \sim q(x \,|\, \varphi)\,, \text{ where } \varepsilon \sim q(\varepsilon)\,, \tag{244}$$

i.e. there exist some, usually very simple, density $q(\varepsilon)$ that does not depend on parameters φ and the corresponding diffeormorphism (differentiable one-to-one function) $g(\varepsilon;\varphi)$ that maps samples from $q(\varepsilon)$ to samples from $q(x\,|\,\varphi)$. Then, one can write

$$\mathbb{E}_{q(x\,|\,\varphi)} f(x) = \mathbb{E}_{q(\varepsilon)} f(g(\varepsilon;\varphi)) \tag{245}$$

$$\nabla_\varphi \mathbb{E}_{q(x\,|\,\varphi)} f(x) = \mathbb{E}_{q(\varepsilon)} \nabla_\varphi f(g(\varepsilon;\varphi)) = \mathbb{E}_{q(\varepsilon)} \frac{\partial g(\varepsilon;\varphi)}{\partial \varphi} \nabla_x f(x)\bigg|_{x=g(\varepsilon;\varphi)} . \tag{246}$$

This is called the **reparameterization trick** or the **pathwise gradient estimator**.

In general, the reparameterization trick, when possible, is always better than the score gradient estimator. Intuitively, this can be explained by the fact that score estimator does not use any information about the gradients of the function $f(x)$ (which carry a lot of information if the function is smooth).

3.3 Latent Variables, EM-algorithm

Extending the state-space with additional variables is one of the fundamental ideas in many disciplines [Kirill:the best example illustrating this is presented in "Ordinary Differential Equations" by V.Arnold in the problem about two circular wagons]. In our case, when parameterizing a distribution on the state-space $\mathcal{X}$, we can always write

$$\int dz\; p(x,z\,|\,\theta) = p(x\,|\,\theta)\,, \tag{247}$$

i.e. we introduce another state-space $\mathcal{Z}$ and introduce the joint distribution on $\mathcal{X}\times\mathcal{Z}$, then we say that our parameterization $p(x\,|\,\theta)$ is simply the marginal of the joint distribution. [Kirill:note that there is infinitely many joint distributions on $\mathcal{X}\times\mathcal{Z}$ that have the marginal $p(x\,|\,\theta)$]

Definition 30 (Latent Variables)**.** *In probabilistic inference, the extended space $\mathcal{Z}$ is called the **latent space** and the variables in this space are called **latent variables**.*

In particular, one can define the joint distribution as follows

$$p(x,z\,|\,\theta) = p(x\,|\,z;\theta)p(z\,|\,\theta)\,, \tag{248}$$

i.e. by defining the distribution over $z\in\mathcal{Z}$ and the conditional distribution on $x\in\mathcal{X}$. Then, applying the maximum likelihood principle, we get

$$\text{Log-Likelihood}(\theta) = \mathbb{E}_{x\sim p_{\text{data}}(x)} \log p(x\,|\,\theta) = \mathbb{E}_{x\sim p_{\text{data}}(x)} \log\big[\mathbb{E}_{z\sim p(z\,|\,\theta)} p(x\,|\,z;\theta)\big]\,. \tag{249}$$

This, however, is very hard to optimize because we usually can't estimate numerically the logarithm of the expectation. Let's apply the variational principle to this. Recall that in Example 9 we derive the variational formulation of maximum likelihood as the miminization of the KL-divergence. Let's apply the same technique, but for the joint distribution $p(x,z\,|\,\theta)$. The obvious problem here is that we don't have the data distribution for the latent variables z, so let's use the posterior $p(z\,|\,x;\theta)$ of our model $p(x\,|\,z;\theta)p(z\,|\,\theta)$. That is

$$D_{\mathrm{KL}}(p_{\text{data}}(x)p(z\,|\,x;\theta), p(x,z\,|\,\theta)) = \int dxdz\; p_{\text{data}}(x)p(z\,|\,x;\theta)\log\frac{p_{\text{data}}(x)p(z\,|\,x;\theta)}{p(x,z\,|\,\theta)} \tag{250}$$

$$= \int dxdz\; p_{\text{data}}(x)p(z\,|\,x;\theta)\log\frac{p_{\text{data}}(x)}{p(x\,|\,\theta)} = D_{\mathrm{KL}}(p_{\text{data}}(x), p(x\,|\,\theta))\,. \tag{251}$$

Thus, we have that the minimization of the KL on the extended space is equivalent to maximum likelihood if we "impute" the latent variables of the target data with the posterior distribution. The last step in this sequence is that we would like to get rid of the differentiation through the "imputed" latent variables by introducing the variational upper bound. Namely, let's say we take some other posterior for generating the latents of the data, i.e. instead of $D_{\mathrm{KL}}(p_{\text{data}}(x)p(z\,|\,x;\theta), p(x,z\,|\,\theta))$, we consider

$$D_{\mathrm{KL}}(p_{\text{data}}(x)\underbrace{p(z\,|\,x;\eta)}_{\text{different parameters!}}, p(x,z\,|\,\theta)) = \int dxdz\; p_{\text{data}}(x)p(z\,|\,x;\eta)\log\frac{p_{\text{data}}(x)p(z\,|\,x;\eta)}{p(x,z\,|\,\theta)} \tag{252}$$

$$= \int dxdz\; p_{\text{data}}(x)p(z\,|\,x;\eta)\log\frac{p_{\text{data}}(x)}{p(x\,|\,\theta)} + \int dxdz\; p_{\text{data}}(x)p(z\,|\,x;\eta)\log\frac{p(z\,|\,x;\eta)}{p(z\,|\,x;\theta)} \tag{253}$$

$$= D_{\mathrm{KL}}(p_{\text{data}}(x), p(x\,|\,\theta)) + \mathbb{E}_{p_{\text{data}}(x)} D_{\mathrm{KL}}(p(z\,|\,x;\eta), p(z\,|\,x;\theta)) \tag{254}$$

$$= D_{\mathrm{KL}}(p_{\text{data}}(x)p(z\,|\,x;\theta), p(x,z\,|\,\theta)) + \mathbb{E}_{p_{\text{data}}(x)}\underbrace{D_{\mathrm{KL}}(p(z\,|\,x;\eta), p(z\,|\,x;\theta))}_{\geq 0}, \tag{255}$$

where in the last transition we use Equation (251). Thus, we have derived the following variational upper bound

$$\begin{aligned} \forall\,\theta,\eta,\;\; & D_{\mathrm{KL}}(p_{\text{data}}(x)p(z\,|\,x;\theta), p(x,z\,|\,\theta)) \leq D_{\mathrm{KL}}(p_{\text{data}}(x)p(z\,|\,x;\eta), p(x,z\,|\,\theta))\,, \\ \forall\,\theta,\;\exists\eta=\theta,\;\; & D_{\mathrm{KL}}(p_{\text{data}}(x)p(z\,|\,x;\theta), p(x,z\,|\,\theta)) = D_{\mathrm{KL}}(p_{\text{data}}(x)p(z\,|\,x;\eta), p(x,z\,|\,\theta))\,. \end{aligned} \tag{256}$$

Thus, we have proved the following result.

Proposition 3 (Variational Formulation of the EM-algorithm)**.** *The following optimization problems are equivalent*

$$\max_\theta \mathbb{E}_{x\sim p_{data}(x)}\log\left[\mathbb{E}_{z\sim p(z\,|\,\theta)}p(x\,|\,z;\theta)\right] \iff \min_\theta\min_\eta D_{\mathrm{KL}}(p_{data}(x)p(z\,|\,x;\eta), p(x,z\,|\,\theta))\,. \tag{257}$$

Definition 31 (EM-algorithm)**.** *The iterative optimization of the following variational upper bound w.r.t. η and θ is called the Expectation Maximization (EM) algorithm*

$$D_{\mathrm{KL}}(p_{data}(x)p(z\,|\,x;\theta), p(x,z\,|\,\theta)) \leq D_{\mathrm{KL}}(p_{data}(x)p(z\,|\,x;\eta), p(x,z\,|\,\theta))\,. \tag{258}$$

Expectation step (E-step). is the optimization w.r.t. the parameters η. This optimization problem is trivial due to the variational bound above. Indeed, simply setting $\eta = \theta$ minimizes the bound. That is why, oftentimes, people say that at the E-step we find the posterior distribution

$$p(z\,|\,x_i;\theta)\,,\;\; x_i \sim p_{\text{data}}(x)\,, \tag{259}$$

which then we can use to estimate the objective by integrating w.r.t. z or sampling the latent variables

$$z_i \sim p(z\,|\,x_i;\theta)\,,\;\; x_i \sim p_{\text{data}}(x)\,. \tag{260}$$

Maximization step (M-step). is the *minimization* of the derived variational upper bound on the KL-divergence w.r.t. θ, i.e.

$$\theta^* = \operatorname*{argmin}_\theta D_{\mathrm{KL}}(p_{\text{data}}(x)p(z\,|\,x;\eta), p(x,z\,|\,\theta))\,. \tag{261}$$

Well, why do people call it the maximization step then? It's very simple — there is another interpretation of this optimization problem. Recall Equation (251), then we can write

$$D_{\mathrm{KL}}(p_{\text{data}}(x), p(x\,|\,\theta)) \leq D_{\mathrm{KL}}(p_{\text{data}}(x)p(z\,|\,x;\eta), p(x,z\,|\,\theta)) \tag{262}$$

$$\mathbb{E}_{p_{\text{data}}(x)}[\log p_{\text{data}}(x) - \log p(x\,|\,\theta)] \leq \mathbb{E}_{p_{\text{data}}(x)p(z\,|\,x;\eta)}\big[\log p_{\text{data}}(x) + \log p(z\,|\,x;\eta) \tag{263}$$

$$- \log p(x,z\,|\,\theta)\big] \tag{264}$$

$$\mathbb{E}_{p_{\text{data}}(x)p(z\,|\,x;\eta)}[-\log p(z\,|\,x;\eta) + \log p(x,z\,|\,\theta)] \leq \mathbb{E}_{p_{\text{data}}(x)}\log p(x\,|\,\theta)\,. \tag{265}$$

Let's rearrange the terms on the left by decomposing the joint density $p(x,z\,|\,\theta) = p(x\,|\,z;\theta)p(z\,|\,\theta)$

$$\underbrace{\mathbb{E}_{p_{\text{data}}(x)p(z\,|\,x;\eta)}\log p(x\,|\,z;\theta) - \mathbb{E}_{p_{\text{data}}(x)}D_{\mathrm{KL}}(p(z\,|\,x;\eta), p(z\,|\,\theta))}_{\text{lower bound}} \leq \underbrace{\mathbb{E}_{p_{\text{data}}(x)}\log p(x\,|\,\theta)}_{\text{marginal likelihood}}\,. \tag{266}$$

Namely, for the M-step, eliminating the terms that are independent of θ, we have the following result.

Proposition 4 (M-step). *The following optimization problems are equivalent*

$$\operatorname*{argmin}_{\theta} D_{\mathrm{KL}}(p_{data}(x)p(z \,|\, x;\eta), p(x, z \,|\, \theta)) = \operatorname*{argmax}_{\theta} \mathbb{E}_{p_{data}(x)p(z \,|\, x;\eta)} \log p(x \,|\, z;\theta)p(z \,|\, \theta)\,. \tag{267}$$

3.4 Variational Auto-Encoder (VAE) (Kingma u. a., 2013)

Finding the posterior distribution $p(z \,|\, x;\theta)$ might be infeasible in practice. Indeed, if the likelihood model is a neural network, e.g.

$$p(x \,|\, z;\theta) = \mathcal{N}(x \,|\, \mu(z;\theta), \sigma^2(z;\theta))\,, \tag{268}$$

where $\mu(z;\theta)$ and $\sigma^2(z;\theta))$ are outputs of the network, then finding $p(z \,|\, x;\theta)$ in closed form is impossible. Even sampling $z \sim p(z \,|\, x;\theta) \propto p(x \,|\, z;\theta)p(z \,|\, \theta)$ might be very complicated task.

The idea is to approximate $p(z \,|\, x;\theta)$ with another parametric model $q(z \,|\, x;\eta)$. We already have all the necessary tools for this! Indeed, when proving the variational bound Equation (256) we did not use anywhere that the posterior with different parameters η has the same functional family as the true posterior. Thus, we can write the following

$$D_{\mathrm{KL}}(p_{\text{data}}(x) \underbrace{p(z \,|\, x;\theta)}_{\text{true posterior}}, p(x \,|\, z;\theta)p(z \,|\, \theta)) \leq D_{\mathrm{KL}}(p_{\text{data}}(x) \underbrace{q(z \,|\, x;\eta)}_{\text{approximation}}, p(x \,|\, z;\theta)p(z \,|\, \theta))\,, \tag{269}$$

i.e. instead of using the posterior from the same functional family but with different parameters (from the previous iteration) as in the EM-algorithm, we approximate the posterior with a completely independent model $q(z \,|\, x;\eta)$. This is a variational bound in the following sense

$$\begin{gathered}\forall\,\theta\,, q(z \,|\, x;\eta)\,,\quad D_{\mathrm{KL}}(p_{\text{data}}(x)p(z \,|\, x;\theta), p(x \,|\, z;\theta)p(z \,|\, \theta)) \leq D_{\mathrm{KL}}(p_{\text{data}}(x)q(z \,|\, x;\eta), p(x \,|\, z;\theta)p(z \,|\, \theta))\,,\\ \forall\,\theta\,,\ \exists q(z \,|\, x;\eta) = p(z \,|\, x;\theta)\,,\quad D_{\mathrm{KL}}(p_{\text{data}}(x)p(z \,|\, x;\theta), p(x, z \,|\, \theta)) = D_{\mathrm{KL}}(p_{\text{data}}(x)p(z \,|\, x;\eta), p(x, z \,|\, \theta))\,.\end{gathered} \tag{270}$$

Clearly, the same result holds for the lower variational bound on the marginal log-likelihood, i.e.

$$\mathbb{E}_{p_{\text{data}}(x)q(z \,|\, x;\eta)} \log p(x \,|\, z;\theta) - \mathbb{E}_{p_{\text{data}}(x)} D_{\mathrm{KL}}(p(z \,|\, x;\eta), p(z \,|\, \theta)) \leq\ \mathbb{E}_{p_{\text{data}}(x)} \log p(x \,|\, \theta)\,. \tag{271}$$

Exercise 26 (VAE with Gaussians). *Consider the following parametric model*

$$\underbrace{p(x \,|\, z;\theta) = \mathcal{N}(x \,|\, \mu(z;\theta), \sigma^2(z;\theta))}_{decoder},\ \underbrace{p(z \,|\, \theta) = \mathcal{N}(0, 1)}_{prior},\ \underbrace{q(z \,|\, x;\eta) = \mathcal{N}(z \,|\, \mu(x;\eta), \sigma^2(x;\eta))}_{encoder}\,. \tag{272}$$

For this model, write down the training objective for both the encoder and the decoder, i.e. the KL upper bound or the log-likelihood lower bound. Which integrals you can take and which of them you have to estimate via Monte Carlo? Which gradient estimator you should use to estimate the gradients w.r.t. θ and η?

3.5 Importance Weighted Auto-Encoder (IWAE) (Burda u. a., 2015)

The log-likelihood lower bound is usually derived in the following way. First, one uses the definition of the marginal density and introduce a fictional density q, which can be anything, i.e.

$$p(x \,|\, \theta) = \int dz\ p(x, z \,|\, \theta) = \int dz\ q(z \,|\, x;\eta)\frac{p(x, z \,|\, \theta)}{q(z \,|\, x;\eta)}\,,\quad \forall\ q(z \,|\, x;\eta)\,. \tag{273}$$

Then one can use the Jensen's inequality for the logarithm and get

$$\log p(x \,|\, \theta) \geq \int dz\ q(z \,|\, x;\eta) \log \frac{p(x, z \,|\, \theta)}{q(z \,|\, x;\eta)} = \mathbb{E}_{q(z \,|\, x;\eta)} \log \frac{p(x, z \,|\, \theta)}{q(z \,|\, x;\eta)}\,,\quad \forall\ q(z \,|\, x;\eta)\,. \tag{274}$$

When we do this there are

- ✓ good news: we get a lower bound on the log-likelihood amenable for optimization;
- ✗ bad news: the bound is tight only for the true posterior, i.e. $q(z\,|\,x;\eta) = p(z\,|\,x;\theta)$.

Note that the estimate in Equation (273) holds for any q and there is no bounds/errors introduced there yet. [Kirill:of course the problem of optimizing the marginal density directly is that the density is log-concave. for instance, consider the gaussian, the gradient of the density vanishes very fast once you go far from the mode. however, the gradient of the log-density is simply x and does not vanish anywhere (formally, except the mode).] This spurred the following idea, what if we could draw more samples in Equation (273) first and only *then* take the logarithm and use Jensen's inequality? Indeed, we can write

$$\begin{aligned}
p(x\,|\,\theta) &= \frac{1}{K}\sum_{i=1}^{K}\int dz_i\; p(x,z_i\,|\,\theta) = \frac{1}{K}\sum_{i=1}^{K}\int dz_i\; q(z_i\,|\,x;\eta)\frac{p(x,z_i\,|\,\theta)}{q(z_i\,|\,x;\eta)} && (275)\\
&= \frac{1}{K}\sum_{i=1}^{K}\int \prod_{j=1}^{K} dz_j\; q(z_j\,|\,x;\eta)\frac{p(x,z_i\,|\,\theta)}{q(z_i\,|\,x;\eta)} = \int \prod_{j=1}^{K} dz_j\; q(z_j\,|\,x;\eta)\frac{1}{K}\sum_{i=1}^{K}\frac{p(x,z_i\,|\,\theta)}{q(z_i\,|\,x;\eta)}\,, && (276)
\end{aligned}$$

i.e. we just "copied" the expression K times and labeled the variables differently, which, of course, does not change the value of the expression. On the second row of derivations we use the fact that the integral expression depends only on z_i, so all the other densities over z_j integrate to 1. [Kirill:note that we could choose completely different densities $q^i(z_i\,|\,x;\eta_i)$ but we keep the same density for simplicity] The final step is to take the logarithm and to apply Jensen's inequality, i.e.

$$\log p(x\,|\,\theta) \geq \int \prod_{j=1}^{K} dz_j\; q(z_j\,|\,x;\eta)\log\frac{1}{K}\sum_{i=1}^{K}\frac{p(x,z_i\,|\,\theta)}{q(z_i\,|\,x;\eta)} = \mathbb{E}_{z_1,\ldots,z_K\sim q}\log\frac{1}{K}\sum_{i=1}^{K}\frac{p(x,z_i\,|\,\theta)}{q(z_i\,|\,x;\eta)}\,. \qquad (277)$$

Thus, we have derived the IWAE lower bound on the marginal density. This bound can be used to train or to evaluate auto-encoding models.

Proposition 5 (IWAE estimator)**.** *The IWAE estimator of the marginal density is*

$$\log p(x\,|\,\theta) \geq \mathbb{E}_{z_1,\ldots,z_K\sim q}\log\frac{1}{K}\sum_{i=1}^{K}\frac{p(x,z_i\,|\,\theta)}{q(z_i\,|\,x;\eta)}\,,\ \textit{where } z_i\sim q(z_i\,|\,x;\eta)\,. \qquad (278)$$

Exercise 27. *Prove that the bound becomes tighter with the number of samples K.*

3.6 Denoising Diffusion Probabilistic Models (DDPM) (Sohl-Dickstein u. a., 2015)

First, I would like to give a small motivation for introducing diffusion models. VAE had a tremendous success in the ML/DL community and quickly started spreading to other fields. It was very simple to understand, implement, and train while allowing for a qualitatively new level of generative modeling. Quickly, it was overshadowed by Generative Adversarial Networks (GANs), but always remained an actively studied/utilized model. Despite this, VAE had limitations, which were formalized as posterior collapse to the prior and ELBO not giving tight enough bounds etc.

[Kirill:Informally, there was a common intuition that you can't get a good density model by slapping a gaussian on top of network's output, and this fact was much more influential than all the mathematical arguments. That's why people started building hierarchical VAEs, i.e. putting a VAE inside a VAE.]

Following DDPM (Sohl-Dickstein u. a., 2015; Ho u. a., 2020) we define the density model as the marginal following backward (decoding) process

$$\begin{gathered}
p(x_0\,|\,\theta) = \int dx_1,\ldots,dx_T\; p(x_T)\prod_{t=1}^{T}p(x_{t-1}\,|\,x_t;\theta)\,,\\
\text{where } p(x_T) = \mathcal{N}(0,1)\,,\quad p(x_{t-1}\,|\,x_t;\theta) = \mathcal{N}(x_{t-1}\,|\,\mu(x_t;\theta),g_t^2)\,.
\end{gathered} \qquad (279)$$

After defining a density model we, as always, can write down the lower bound on the log-density as follows

$$\log p(x_0 \,|\, \theta) \;= \log \int dx_1 \dots dx_T \; p(x_T) \prod_{t=1}^{T} p(x_{t-1} \,|\, x_t; \theta) \tag{280}$$

$$= \log \int dx_1 \dots dx_T \; \prod_{t=1}^{T} q(x_t \,|\, x_{t-1}) \frac{p(x_T) \prod_{t=1}^{T} p(x_{t-1} \,|\, x_t; \theta)}{\prod_{t=1}^{T} q(x_t \,|\, x_{t-1})} \tag{281}$$

$$\geq \mathbb{E}_{\prod_{t=1}^{T} q(x_t \,|\, x_{t-1})} \log \frac{p(x_T) \prod_{t=1}^{T} p(x_{t-1} \,|\, x_t; \theta)}{\prod_{t=1}^{T} q(x_t \,|\, x_{t-1})} \tag{282}$$

$$= \mathbb{E}_{\prod_{t=1}^{T} q(x_t \,|\, x_{t-1})} \left[\log p(x_T) + \sum_{t=1}^{T} \log \frac{p(x_{t-1} \,|\, x_t; \theta)}{q(x_t \,|\, x_{t-1})} \right]. \tag{283}$$

Note that this holds for any forward (encoding) process $\prod_{t=1}^{T} q(x_t \,|\, x_{t-1})$. To progress further, we need to introduce the marginals of this process as follows

$$q(x_\tau \,|\, x_0) \coloneqq \int dx_1 \dots dx_{\tau-1} \prod_{t=1}^{\tau} q(x_t \,|\, x_{t-1}), \;\; \text{and} \;\; q(x_0 \,|\, x_0) \coloneqq 1\,. \tag{284}$$

Using this definition and the lower bound from Equation (283), we have

$$\log p(x_0 \,|\, \theta) \geq \; \mathbb{E}_{\prod_{t=1}^{T} q(x_t \,|\, x_{t-1})} \left[\log p(x_T) + \sum_{t=1}^{T} \log \frac{p(x_{t-1} \,|\, x_t; \theta)}{q(x_t \,|\, x_{t-1})} \right] \tag{285}$$

$$= \; \mathbb{E}_{q(x_T \,|\, x_0)} \log p(x_T) + \sum_{t=1}^{T} \mathbb{E}_{q(x_{t-1} \,|\, x_0) q(x_t \,|\, x_{t-1})} \log \frac{p(x_{t-1} \,|\, x_t; \theta)}{q(x_t \,|\, x_{t-1})}\,. \tag{286}$$

The final step in the derivation is to *define* $q(x_{t-1} \,|\, x_t, x_0)$. Indeed, for the joint distribution $q(x_t \,|\, x_{t-1}) q(x_{t-1} \,|\, x_0)$, we can define (using Definition 8) the following conditional distribution

$$q(x_{t-1} \,|\, x_t, x_0) \coloneqq \frac{q(x_t \,|\, x_{t-1}) q(x_{t-1} \,|\, x_0)}{\int dx_{t-1} \; q(x_t \,|\, x_{t-1}) q(x_{t-1} \,|\, x_0)} = \frac{q(x_t \,|\, x_{t-1}) q(x_{t-1} \,|\, x_0)}{q(x_t \,|\, x_0)}\,. \tag{287}$$

Continuing with this definition, we get

$$\log p(x_0 \,|\, \theta) \geq \; \mathbb{E}_{q(x_T \,|\, x_0)} \log p(x_T) - \sum_{t=1}^{T} \mathbb{E}_{q(x_t \,|\, x_0)} D_{\mathrm{KL}}(q(x_{t-1} \,|\, x_t, x_0), p(x_{t-1} \,|\, x_t; \theta)) + \tag{288}$$

$$+ \sum_{t=1}^{T} \mathbb{E}_{q(x_{t-1} \,|\, x_0) q(x_t \,|\, x_{t-1})} \log \frac{q(x_{t-1} \,|\, x_0)}{q(x_t \,|\, x_0)}\,. \tag{289}$$

For the last term, we have

$$\sum_{t=1}^{T} \mathbb{E}_{q(x_{t-1} \,|\, x_0) q(x_t \,|\, x_{t-1})} \log \frac{q(x_{t-1} \,|\, x_0)}{q(x_t \,|\, x_0)} = \; \sum_{t=0}^{T-1} \mathbb{E}_{q(x_t \,|\, x_0)} \log q(x_t \,|\, x_0) - \sum_{t=1}^{T} \mathbb{E}_{q(x_t \,|\, x_0)} \log q(x_t \,|\, x_0) \tag{290}$$

$$= \; \mathbb{E}_{q(x_0 \,|\, x_0)} \log q(x_0 \,|\, x_0) - \mathbb{E}_{q(x_T \,|\, x_0)} \log q(x_T \,|\, x_0) \tag{291}$$

$$= \; - \mathbb{E}_{q(x_T \,|\, x_0)} \log q(x_T \,|\, x_0)\,. \tag{292}$$

Finally, we have the following result

Theorem 5 (Backward Process Lower Bound)**.** *For the backward process from Equation (279) and any forward process $q(x_t \,|\, x_{t-1})$, we have the following lower bound on the log-density*

$$\log p(x_0 \,|\, \theta) \geq \; - KL(q(x_T \,|\, x_0), p(x_T)) - \sum_{t=1}^{T} \mathbb{E}_{q(x_t \,|\, x_0)} D_{\mathrm{KL}}(q(x_{t-1} \,|\, x_t, x_0), p(x_{t-1} \,|\, x_t; \theta))\,. \tag{293}$$

Proposition 6 (Variational formulation)**.** *For the backward process from Equation (279) and any forward process $q(x_t\,|\,x_{t-1})$, we can derive the following upper bound on the KL-divergence*

$$D_{\mathrm{KL}}(p_{data}(x_0), p(x_0\,|\,\theta)) \le\ KL(q(x_T), p(x_T)) + \sum_{t=1}^{T} \mathbb{E}_{q(x_t)} D_{\mathrm{KL}}(q(x_{t-1}\,|\,x_t), p(x_{t-1}\,|\,x_t;\theta))\,, \tag{294}$$

where $p_{data}(x_0)$ is the density of the data distribution and the marginals $q(x_\tau)$ are defined as

$$q(x_\tau) \coloneqq \int dx_0\; q(x_\tau\,|\,x_0) p_{data}(x_0)\,. \tag{295}$$

Exercise 28. *Derive the variational formulation.*

Theoretically, using this lower bound on the log-likelihood or its variational formulation in terms of the KL, one could parameterize both q and p and solve the optimization problem w.r.t. the models' parameters. In practice, however, a naive parameterization of this does not work

✗ indeed, to get the marginal $q(x_\tau\,|\,x_0)$ one has to apply the parameteric model multiple times what creates a lot of numerical instabilities;

✗ furthermore, there exist infinite number of marginal sequences $q(x_\tau)$ between $p_{\text{data}}(x_0)$ and $p(x_T)$, hence, multiple forward and backward processes.

Sohl-Dickstein u. a. (2015) solved these problems by fixing the forward (encoding) process $q(x_t\,|\,x_{t-1})$ to the *diffusion process*, which bears the following benefits

✓ there is unique sequence of marginals $q(x_\tau)$, hence the marginals of the backward process are also fixed [Kirill:still exists infinitely many conditionals].

✓ sampling from $q(x_\tau\,|\,x_0)$ can be performed directly, i.e. without simulating multiple transition kernels $q(x_t\,|\,x_{t-1})$;

✓ from physics we know that the functional form of the reverse (backward) process is the same as the forward diffusion process (**?**). [Kirill:thus, for the first time in history, it was mathematically OK to slap a gaussian on networks' outputs.]

In particular, the diffusion process (its time discretization) is defined by the following forward transition kernel

$$q(x_t\,|\,x_{t-1}) = \mathcal{N}(x_t\,|\,\alpha_t x_{t-1}, \sigma_t^2 \mathbb{1})\,, \tag{296}$$

that is, to generate next x_t we have to scale x_{t-1} by some scalar α_t and add some noise

$$x_t = \alpha_t x_{t-1} + \sigma_t \varepsilon\,,\;\; \varepsilon \sim \mathcal{N}(\varepsilon\,|\,0,1)\,. \tag{297}$$

For this process, we would like to find the following marginal

$$q(x_t\,|\,x_0) = ? \tag{298}$$

However, first, let's answer a simpler question $q(x_t\,|\,x_{t-2}) = ?$ That is,

$$\begin{aligned} x_t &= \alpha_t x_{t-1} + \sigma_t \varepsilon = \alpha_t(\alpha_{t-1} x_{t-2} + \sigma_{t-1}\varepsilon') + \sigma_t\varepsilon\,,\;\; \varepsilon,\varepsilon' \sim \mathcal{N}(0,\mathbb{1}) \qquad (299)\\ &= \alpha_t\alpha_{t-1} x_{t-2} + \alpha_t\sigma_{t-1}\varepsilon' + \sigma_t\varepsilon\,. \qquad (300) \end{aligned}$$

Thus, we have

$$q(x_t\,|\,x_{t-2}) = \mathcal{N}(x_t\,|\,\alpha_t\alpha_{t-1}x_{t-2}, ((\alpha_t\sigma_{t-1})^2 + \sigma_t^2)\mathbb{1})\,. \tag{301}$$

Applying this proposition iteratively, we get the following formula.

Proposition 7. *For the transition probability $q(x_t \,|\, x_{t-1}) = \mathcal{N}(x_t \,|\, \alpha_t x_{t-1}, \sigma_t^2 \mathbb{1})$, we have the following formula for the marginals*

$$q(x_\tau \,|\, x_0) = \mathcal{N}\left(x_t \middle| \prod_{t=1}^{\tau} \alpha_t x_0, \sum_{t=1}^{\tau} \sigma_t^2 \prod_{i=t+1}^{\tau} \alpha_i^2 \mathbb{1}\right). \tag{302}$$

To evaluate the objective we need the following conditional density.

Exercise 29. *For the transition probability $q(x_t \,|\, x_{t-1}) = \mathcal{N}(x_t \,|\, \sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbb{1})$, derive $q(x_{t-1} \,|\, x_t, x_0)$ using the definition*

$$q(x_{t-1} \,|\, x_t, x_0) \coloneqq \frac{q(x_t \,|\, x_{t-1}) q(x_{t-1} \,|\, x_0)}{\int dx_{t-1}\; q(x_t \,|\, x_{t-1}) q(x_{t-1} \,|\, x_0)} = \frac{q(x_t \,|\, x_{t-1}) q(x_{t-1} \,|\, x_0)}{q(x_t \,|\, x_0)}. \tag{303}$$

The final step is to derive the KL-divergence

Exercise 30. *For the transition probability $q(x_t \,|\, x_{t-1}) = \mathcal{N}(x_t \,|\, \sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbb{1})$ and parameterized model of the backward process $p(x_{t-1} \,|\, x_t) = \mathcal{N}(x_{t-1} \,|\, \mu_t(x_t; \theta), \beta_t \mathbb{1})$, find*

$$D_{\mathrm{KL}}(q(x_{t-1} \,|\, x_t), p(x_{t-1} \,|\, x_t; \theta)) =? \tag{304}$$

4 Monte Carlo methods

Monte Carlo methods are usually used to numerically estimate the integrals of the form

$$\int dx\; p(x)f(x) = \mathbb{E}_{p(x)}f(x)\,. \tag{305}$$

The key insight here is that instead of approximating the expression on the left using a regular grid or quadratures, one can draw samples distributed as $p(x)$ and approximate

$$\mathbb{E}_{p(x)}f(x) \simeq \frac{1}{N}\sum_{i=1}^{N} f(x_i)\,, x_i \sim p(x)\,. \tag{306}$$

To understand the behaviour of this estimate one can use the Central Limit Theorem.

Theorem 6 (Central Limit Theorem)**.** *Consider a random variable X with the density $p(x)$, mean μ, and variance σ^2. The following convergence result takes place*

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)\,,\;\; \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i\,, x_i \sim p(x)\,, \tag{307}$$

where $\xrightarrow{d}$ denotes the convergence in distribution.

Using this estimator one can, for instance, estimate the predictive distribution of a Bayesian model as follows

$$\int d\theta\; p(x\,|\,\theta)p(\theta\,|\,\mathcal{D}) = \mathbb{E}_{p(\theta\,|\,\mathcal{D})}p(x\,|\,\theta) \simeq \frac{1}{n}\sum_{i=1}^{n} p(x\,|\,\theta_i)\,,\; \theta_i \sim p(\theta_i\,|\,\mathcal{D})\,. \tag{308}$$

Hence, instead of approximating the posterior distribution as we did in Variational Inference (Section 3), one can reduce the problem to sampling from $p(\theta\,|\,\mathcal{D})$.

The central question of this section is "How to sample from a given density?"

Let's start with the simplest case of a 1D variable with a known Cumulative Density Function (CDF). That is, for the random variable X its CDF is defined as

$$F_X(t) = \mathbb{P}(X \leq t)\,. \tag{309}$$

Proposition 8 (Inverse CDF Sampling)**.** *Define the random variable Y as follows*

$$Y = F_X^{-1}(U)\,,\; U \sim [0,1]\,, \tag{310}$$

where the random variable U is uniformly distributed in $[0,1]$. Y and X have the same distribution.

Proof. To prove the equivalence of distributions let's evaluate the CDF of Y. We have

$$F_Y(t) = \mathbb{P}(Y \leq t) = \mathbb{P}(F_X^{-1}(U) \leq t) = \mathbb{P}(U \leq F_X(t)) = F_X(t)\,. \tag{311}$$

□

4.1 Importance Sampling

Before diving into the discussion of numerous ways to sample from a given density, let's see what we could possibly do if we cannot sample from the given density $p(x)$ but instead we can sample from some other density $q(x)$. Can we estimate the integral in that case? Well, let's write down the following identity

$$\mu := \int dx\; p(x)f(x) = \int dx\; q(x)\frac{p(x)}{q(x)}f(x) = \mathbb{E}_{q(x)}\underbrace{\frac{p(x)}{q(x)}}_{w(x)}f(x)\,, \tag{312}$$

where $w(x)$ is the density ratio that is called the importance weight. This identity suggests the following Monte Carlo estimator

$$\int dx\; p(x)f(x) = \mathbb{E}_{q(x)} \underbrace{\frac{p(x)}{q(x)}}_{w(x)} f(x) \approx \frac{1}{n}\sum_{i=1}^{n} \frac{p(x_i)}{q(x_i)} f(x_i)\,, \quad x_i \sim q(x)\,. \tag{313}$$

Note that the estimator is itself a random variable, so one can wonder what is the mean of this random variable or the expected error.

$$\mathbb{E}_{x_1,\ldots,x_n\sim q(x)} \frac{1}{n}\sum_{i=1}^{n} \frac{p(x_i)}{q(x_i)} f(x_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{x_i\sim q(x)} \frac{p(x_i)}{q(x_i)} f(x_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{x_i\sim q(x)} \frac{p(x_i)}{q(x_i)} f(x_i) = \mu\,. \tag{314}$$

That is, we have shown that the expected value of the Importance Sampling estimator matches the ground true value μ of the integral.

Proposition 9 (Unbiasedness of Importance Sampling)**.** *The Importance Sampling estimator is unbiased, i.e.*

$$\mathbb{E}_{x_1,\ldots,x_n\sim q(x)} \frac{1}{n}\sum_{i=1}^{n} \frac{p(x_i)}{q(x_i)} f(x_i) = \int dx\; p(x)f(x)\,. \tag{315}$$

The unbiasedness of an estimator is an important property, especially, for the optimization objectives. In particular, the unbiased estimator allows for the unbiased estimate of the gradients which ubiquitously used for the convergence results of the Stochastic Gradient Descent (SGD).

Another crucial property of a Monte Carlo estimate is the consistency of the estimator.

Proposition 10 (Consistency of Importance Sampling)**.** *The Importance Sampling estimator is consistent, i.e. for $x_i \sim q(x_i)$, we have*

$$\overset{p}{\lim_{n\to\infty}} \frac{1}{n}\sum_{i=1}^{n} \frac{p(x_i)}{q(x_i)} f(x_i) = \int dx\; p(x)f(x)\,, \tag{316}$$

where $\lim^p$ denotes convergence in probability.

The variance of the IS estimator is

$$\mathbf{D}\left[\frac{1}{n}\sum_{i=1}^{n} \frac{p(x_i)}{q(x_i)} f(x_i)\right] = \frac{1}{n^2}\sum_{i=1}^{n} \mathbf{D}\left[\frac{p(x)}{q(x)} f(x)\right] = \frac{1}{n}\mathbb{E}_{q(x)}\left[\frac{p(x)}{q(x)} f(x) - \mu\right]^2 = \frac{1}{n}\mathbb{E}_{q(x)}\left[\frac{p(x)}{q(x)} f(x)\right]^2 - \frac{\mu^2}{n}\,. \tag{317}$$

Let's consider the functional derivative of this variance w.r.t. $q(x)$

$$\frac{\delta}{\delta q}\mathbf{D}\left[\frac{p(x)}{q(x)} f(x)\right] = \frac{\delta}{\delta q}\int dx\; \frac{1}{q(x)}[p(x)f(x)]^2 = \frac{-1}{q(x)^2}[p(x)f(x)]^2\,. \tag{318}$$

Since $q(x)$ is a density, the variation of the variance is zero when the derivative is constant w.r.t. x (up to a measure zero). Hence, the minimal variance is achieved

$$\operatorname*{argmin}_q \mathbf{D}\left[\frac{p(x)}{q(x)} f(x)\right] \propto p(x)|f(x)|\,. \tag{319}$$

Proposition 11 (Optimal Proposal for Importance Sampling)**.** *The proposal density $q(x)$ that minimizes the variance of the Monte Carlo estimate in Equation* (313) *is*

$$q(x) = \frac{p(x)|f(x)|}{\int dx\; p(x)|f(x)|}\,. \tag{320}$$

In practice, the main challenge is to sample from an *unnormalized density*, i.e., for the target density $p(x) = \hat{p}(x)/Z_p$, we know only $\hat{p}(x)$ and don't know the normalization constant Z_p. The same applies for the proposal, i.e. we know only $q(x) = \hat{p}(x)/Z_q$. In this case, we can estimate the normalization constant if we consider $f(x) \equiv 1$. Indeed,

$$1 = \int dx\; p(x) = \mathbb{E}_{q(x)} \frac{Z_q}{Z_p} \frac{\hat{p}(x)}{\hat{q}(x)} \tag{321}$$

$$\frac{Z_p}{Z_q} = \mathbb{E}_{q(x)} \frac{\hat{p}(x)}{\hat{q}(x)} \approx \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{p}(x_i)}{\hat{q}(x_i)}\,, \quad x_i \sim q(x)\,, \tag{322}$$

Using this expression for the estimation of Z_q/Z_p, we get the following expression

$$\int dx\; p(x)f(x) \approx \frac{1}{n}\sum_{i=1}^{n} \frac{p(x_i)}{q(x_i)} f(x_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_q}{Z_p}\frac{\hat{p}(x_i)}{\hat{q}(x_i)} f(x_i) \approx \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{p}(x_i)/\hat{q}(x_i)}{\sum_j \hat{p}(x_j)/\hat{q}(x_j)} f(x_i)\,, \quad x_i \sim q(x)\,. \tag{323}$$

This motivates the following estimator.

Definition 32 (Self-Normalized Importance Sampling)**.** *For the unnormalized target density $p(x) \propto \hat{p}(x)$ and unnormalized proposal distribution $q(x) \propto \hat{q}(x)$, Self-Normalized Importance Sampling (SNIS) is the following estimator*

$$\int dx\; p(x)f(x) \approx \sum_{i=1}^{n} w_i f(x_i)\,, w_i = \frac{\hat{p}(x_i)/\hat{q}(x_i)}{\sum_j \hat{p}(x_j)/\hat{q}(x_j)} \quad x_i \sim q(x)\,. \tag{324}$$

Exercise 31. *Prove that SNIS is a biased but consistent estimator.*

4.2 Discrete-Space Markov Chains

As we discussed before, the main question of Monte Carlo methods is the design of algorithms sampling from the given target distribution. One of the main tools in this design is Markov Chains. However, to learn how to use this tool, we have to start with studying Markov Chains. This section is aimed at developing the intuition for Markov Chains.

Definition 33 (Markov Process)**.** *Consider a sequence of random variables $X_0, \ldots, X_t, \ldots$. This sequence is called a Markov process if for any t the joint density of the random variables up to T factorizes as follows*

$$p(x_0, \ldots, x_T) = p(x_0) \prod_{t=1}^{T} p(x_t | x_{t-1}) \tag{325}$$

[Kirill:interestingly, Kolmogorov was calling these processes stochastically determined and Follmer called them memory-less processes, it took some time for the development of common convention Markov Processes]

For the discrete state-space, we assume that every x_t takes values in some finite amount of states that we can number as $1, \ldots, N$. Then, the distribution of x_0 can simply described by an N-dimensional vector q, i.e.

$$p(x_0 = i) = q_i\,, \quad q_i \geq 0\,, \;\forall\, i \quad \sum_{i=1}^{N} q_i = 1\,. \tag{326}$$

Furthermore, the conditional distribution $p(x_t|x_{t-1})$, which is usually called the transition probability can be described as the following matrix

$$p(x_t = i \,|\, x_{t-1} = j) = P_{ij}\,, \quad P_{ij} \geq 0\,, \;\forall\, i,j \quad \sum_{i=1}^{N} p(x_t = i \,|\, x_{t-1} = j) = \sum_{i=1}^{N} P_{ij} = 1\,. \tag{327}$$

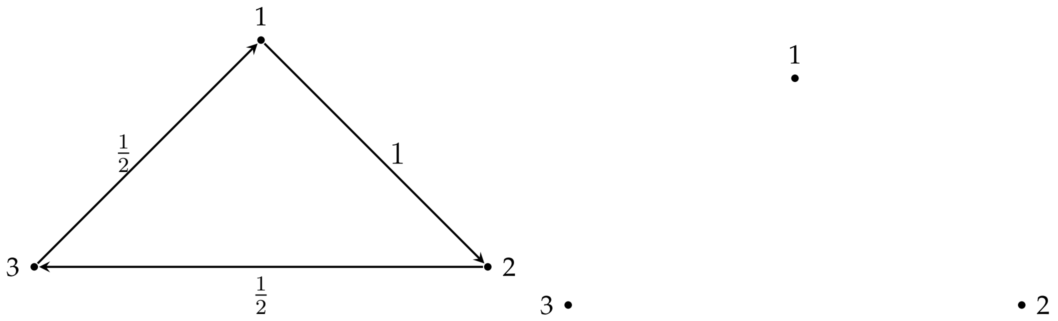

Figure 8: Diagrams of different Markov Chains: left is the diagram for the Markov chain in Example 10, right is the diagram corresponding to the identity matrix.

In particular, if the Markov Process is time-homogeneous (the tranistion kernel does not change in time), i.e. $p(x_t = i \,|\, x_{t-1} = j) = p(x_s = i \,|\, x_{s-1} = j)\,,\ \forall\, t, s$, then we can use the same matrix P for all the transition kernels.

Example 10. *Let's consider a Markov Chain on three states depicted in Figure 8, the transition matrix for this chain is the following*

$$P = \begin{bmatrix} p(x_t = 1 \,|\, x_{t-1} = 1) & p(x_t = 1 \,|\, x_{t-1} = 2) & p(x_t = 1 \,|\, x_{t-1} = 3) \\ p(x_t = 2 \,|\, x_{t-1} = 1) & p(x_t = 2 \,|\, x_{t-1} = 2) & p(x_t = 2 \,|\, x_{t-1} = 3) \\ p(x_t = 3 \,|\, x_{t-1} = 1) & p(x_t = 3 \,|\, x_{t-1} = 2) & p(x_t = 3 \,|\, x_{t-1} = 3) \end{bmatrix} = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{bmatrix}. \tag{328}$$

Natural question for Markov Chains is the marginal distributions after n steps. Let's start from a single step starting from some marginal distribution $p(x_0) = q$.

$$p(x_1 = i) = \sum_j p(x_1 = i \,|\, x_0 = j) p(x_0 = j) = \sum_j P_{ij} q_j = (Pq)_i\,, \tag{329}$$

i.e. the marginal distribution at the next step $p(x_1)$ can be obtained simply by multiplying the transition matrix by the marginal distribution of the initial state $p(x_0)$ from the right. Thus, we know how to start calculating the marginals, but what if we consider the process in the middle of the chain? Here, we can wonder what's the transition probability between t and $t-2$, i.e. skipping one step ahead. For this transition probability, we have

$$p(x_t = i \,|\, x_{t-2} = k) = \sum_j p(x_t = i \,|\, x_{t-1} = j) p(x_{t-1} = j \,|\, x_{t-2} = k) = \sum_j P_{ij} P_{jk} = (PP)_{ik}\,, \tag{330}$$

i.e. the transition kernel [Kirill:same as transition probability matrix. it is oftentimes called kernel, especially, in the continuous state-space] from $t-2$ to t is defined as matrix multiplication of the transition matrix P. [Kirill:how did we come up with this formula for the transition kernel at first place? it's a good excercise to practice.]

Exercise 32. *Prove the opposite result that for any row-stochastic matrix P*

$$P_{ij} \geq 0\,,\ \forall\, i, j\ \ \sum_{i=1}^{N} P_{ij} = 1\,, \tag{331}$$

we have a valid transition kernel that transforms any distribution into a distribution.

Exercise 33. *Using the previous two facts prove the following statements by induction*

$$p(x_n = i \,|\, x_0 = k) = (P^n)_{ik}\,,\ \ p(x_n = i) = (P^n q)_i\,. \tag{332}$$

Using these facts we would like to analyse the asymptotic behaviour of our chain after many transitions. In order to do this, we have to evaluate P^n, which can be done if we represent the matrix P in its eigenbasis, i.e., for simplicity, let's assume that eigenvectors of P form an orthonormal basis, then we can represent P as follows

$$P = U\Lambda U^T\,, \text{ where } \Lambda = \begin{bmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_N \end{bmatrix}\,, \quad U = \begin{bmatrix} e_1 & \dots & e_N \end{bmatrix}\,, \tag{333}$$

Λ is the diagonal matrix with eigenvalues on the diagonal and U is the matrix, which columns are eigenvectors of P. Clearly, for this decomposition, we have

$$PP = U\Lambda U^T U\Lambda U^T = U\Lambda^2 U^T\,, \quad P^n = U\Lambda^n U^T\,, \tag{334}$$

i.e. raising P to the power of n just raises the diagonal matrix Λ to the power of n.

Let's get back to the transition matrix from Example 10, and look at its eigenvalues.

$$\det\left(\begin{bmatrix} -\lambda & 1/2 & 1/2 \\ 1 & -\lambda & 1/2 \\ 0 & 1/2 & -\lambda \end{bmatrix}\right) = -\lambda\left(\lambda^2 - \frac{1}{4}\right) - 1\left(-\frac{\lambda}{2} - \frac{1}{4}\right) \tag{335}$$

$$= \left(\lambda + \frac{1}{2}\right)\left[-\lambda\left(\lambda - \frac{1}{2}\right) + \frac{1}{2}\right] = \left(\lambda + \frac{1}{2}\right)^2 (\lambda - 1) \tag{336}$$

Thus, we have $\lambda_1 = 1$ and, alright, here we have the eigenvalue $\lambda_{2,3} = -1/2$ of multiplicity 2, hence we can't diagonalize Λ, but irregardless, let's say that we start from q that can be represented in our eigenbasis, i.e.

$$q = \alpha_1 e_1 + \alpha_2 e_2 + \alpha_3 e_3\,, \tag{337}$$

where e_i are the eigenvectors and α_i are the coordinates of q in this basis.

$$P^n q = \alpha_1 P^n e_1 + \alpha_2 P^n e_2 + \alpha_3 P^n e_3 = \alpha_1 \lambda_1^n e_1 + \alpha_2 \lambda_2^n e_2 + \alpha_3 \lambda_3^n e_3 \tag{338}$$

$$= \alpha_1 e_1 + \alpha_2 (-1/2)^n e_2 + \alpha_3 (-1/2)^n e_3 \tag{339}$$

$$P^n q \xrightarrow[n\to\infty]{} \alpha_1 e_1\,. \tag{340}$$

Thus, we see that the marginal distribution $p(x_n)$ of this Markov Chain always converges to the distribution proportional to the eigenvector e_1 corresponding to $\lambda_1 = 1$. In particular,

$$\begin{bmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{bmatrix} e_1 = e_1 \implies \begin{bmatrix} b + c = 2a \\ 2a + c = 2b \\ b = 2c \end{bmatrix} \implies \begin{bmatrix} a = 3/2 \\ b = 2 \\ c = 1 \end{bmatrix} \implies e_1 = \begin{bmatrix} 3/9 \\ 4/9 \\ 2/9 \end{bmatrix}\,. \tag{341}$$

To summarize this example, we observe that for any starting distribution q our markov chain converges to the distribution e_1, which satisfies

$$Pe_1 = e_1\,. \tag{342}$$

Of course, convergence to e_1 is not a coincidence and, more generally, we are interested in convergence to the stationary points of our dynamics.

Definition 34 (Stationarity Distribution)**.** *For the transition matrix P, the stationary distribution π is*

$$P\pi = \pi\,. \tag{343}$$

Clearly, the convergence of the Markov Chain depends on its spectrum, hence, we would like to state something general about the spectrum of Markov Chains. We start with the following result.

Proposition 12 (Eigenvalues of the Transition Matrix)**.** *For any transition matrix of a Markov Chain, there exist eigenvalue $\lambda_1 = 1$ and all the eigenvalues are bounded by 1, i.e.*

$$\lambda_1 = 1 \geq |\lambda_i| \,, \; \forall \, i = 2, \ldots, N \,. \tag{344}$$

Proof. We are going to analyse the eigenvalues of P^T because they are the same as of P. Indeed, the characteristic polynomials of these matrices are the same

$$\det(P - \lambda \mathbf{1}) = \det(P^T - \lambda \mathbf{1}). \tag{345}$$

Clearly, for P^T, we have

$$(P^T \mathbf{1})_i = \sum_j (P^T)_{ij} \mathbf{1}_j = \sum_j P_{ji} = 1 \implies P^T \mathbf{1} = \mathbf{1} \,. \tag{346}$$

Thus, there exist eigenvalue $\lambda = 1$. This is also the largest eigenvalue among all of them. Indeed, using the Perron-Frobenius theorem (or Gershgorin theorem) we have the following inequality for the norm of the largest eigenvalue

$$\min_i \sum_j P^T_{ij} \leq |\lambda_1| \leq \max_i \sum_j P^T_{ij} \,, \tag{347}$$

$$1 \leq |\lambda_1| \leq 1 \implies |\lambda_1| = 1 \,, \tag{348}$$

where we denote λ_1 as the largest eigenvalue. That is the absolute value of the largest eigenvalue is always 1. Since there exist $\lambda = 1$, we can always choose $\lambda_1 = 1$ as the largest eigenvalues. [Kirill:in general, of course, there might be other eigenvalues with the absolute value 1] □

Thus, we see that we can always pull of the same trick as in our example for the first eigenvalue. Indeed, let's write the starting distribution in the eigenbasis

$$q = \sum_{i=1}^N \alpha_i e_i \,, \quad P^n q = \sum_{i=1}^N \alpha_i \lambda_i^n e_i = \alpha_1 e_1 + \sum_{i=2}^N \alpha_i \lambda_i^n e_i \,. \tag{349}$$

However, we can't guarantee that all the other λ_i except λ_1 will vanish. Indeed, an obvious example is the identity matrix, for which we have all the eigenvalues equal zero and all the vectors being stationary points.

The issue with the identity matrix becomes obvious if we think about markov chains in terms of the diagram from Figure 8. Clearly, the identity matrix does not make any transitions, i.e. if we start from some state we are going to stay in this state [Kirill:also, we can clearly see that the transition graph is disconnected]. To eliminate these uninteresting Markov Chains from consideration, the community came up with the following definition.

Definition 35 (Irreducibility)**.** *If for any i, j there exists $n \in \mathbb{N}$ such that $p(x_n = i \,|\, x_0 = j) > 0$, the Markov Chain is called irreducible.*

This definition is broad enough to cover a lot of interesting Markov Chains and at the same time strong enough to guarantee the existence of the stationary distribution.

Theorem 7 (Unique Stationary)**.** *For irreducible Markov Chain P there exist unique stationary distribution π such that $P\pi = \pi$.*

Theorem 8 (Ergodic Theorem)**.**

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n P^i q = \pi \tag{350}$$

Note that the irreducibility does not imply the convergence of the markov chain to the stationary distribution. Indeed, consider the looped chain that transitions everything to the state with the next index, i.e. $p(x_t = i + 1 \,|\, x_{t-1} = i) = 1$ and $p(x_t = 1 \,|\, x_{t-1} = N) = 1$. One has to require the aperiodicity, but this is beyond our current scope.

4.3 Metropolis-Hastings Test

In the previous section we analysed the asymptotic behavior of given Markov Chain. However, the central question of MCMC field is

> *How do we design a Markov Chain with a given stationary distribution π?*

With more details, one has to guarantee the following three properties:

1. Stationarity [Kirill:the kernel preserves the target distribution]
2. Irreducibility [Kirill:the kernel can reach any point in the space where the target density is positive]
3. Aperiodicity [Kirill:the kernel does not contain periodic loops]

However, the last two properties are usually guaranteed by selecting the kernel has positive density over the entire state-space. Here, we focus on the stationarity condition, that can be formalized as follows.

Definition 36 (Stationarity Condition)**.** *For the Markov chain $k(x' \,|\, x)$, the stationary density π satisfies*

$$\int dx\; k(y\,|\,x)\pi(x) = \pi(y)\,. \tag{351}$$

This is not evident how to design kernel $k(y\,|\,x)$ that satisfies the stationarity condition! What we have is the information about x π and some standard distributions (e.g. uniform, normal, Poisson etc.). The main idea here is to take any $k(y\,|\,x)$ and augment it with an accept/reject step. Namely, let's call the sample $y \sim k(y\,|\,x)$ as the proposal point. Based on the current state x and the proposed state y, we want to make a decision whether we move to y or stay in x. Mathematically, we have $(x,y) \sim k(y\,|\,x)\pi(x)$ and for the next state x' we choose x with probability $g(x,y)$ or y with probability $(1-g(x,y))$, let's denote the density of x' as $q(x')$, then we have

$$\begin{aligned}
q(x') =&\; \int dxdy\; (g(x,y)\delta(x'-y) + (1-g(x,y))\delta(x'-x))k(y\,|\,x)\pi(x) && (352)\\
=&\; \int dx\; g(x,x')k(x'\,|\,x)\pi(x) + \int dy\; (1-g(x',y))k(y\,|\,x')\pi(x') && (353)\\
=&\; \pi(x') + \int dx\; g(x,x')k(x'\,|\,x)\pi(x) - \int dy\; g(x',y)k(y\,|\,x')\pi(x')\,. && (354)
\end{aligned}$$

Clearly, nobody stops us from renaming variables inside the integral. Also, for the stationarity, we require $q(x') = \pi(x')$, then we have

$$\begin{aligned}
q(x') =&\; \pi(x') + \int dy\; (g(y,x')k(x'\,|\,y)\pi(y) - g(x',y)k(y\,|\,x')\pi(x')) && (355)\\
\pi(x') =&\; \pi(x') + \int dy\; (g(y,x')k(x'\,|\,y)\pi(y) - g(x',y)k(y\,|\,x')\pi(x')) && (356)\\
0 =&\; \int dy\; (g(y,x')k(x'\,|\,y)\pi(y) - g(x',y)k(y\,|\,x')\pi(x'))\,, && (357)
\end{aligned}$$

In particular, this can be satisfied by the following choice

$$\begin{aligned}
g(y,x')k(x'\,|\,y)\pi(y) =&\; g(x',y)k(y\,|\,x')\pi(x')\,,\ \forall\, y,x'\,, && (358)\\
g(x',y) =&\; g(y,x')\frac{k(x'\,|\,y)\pi(y)}{k(y\,|\,x')\pi(x')}\,,\ \forall\, y,x'\,. && (359)
\end{aligned}$$

The final step in the reasoning is to recall that $g(x',y)$ is a probability (the acceptance probability); hence, $0 \le g(x',y) \le 1$, and we can take

$$g(x',y) = \min\left\{1, \frac{k(x'\,|\,y)\pi(y)}{k(y\,|\,x')\pi(x')}\right\},\ \forall\, y,x'\,, \tag{360}$$

Algorithm 1 Metropolis-Hastings test

```
Require: proposal kernel $k(y \,|\, x)$, starting point $x_0$, target density $\pi(x)$
  for iterations $i \in [0, n)$ do
    sample proposal $y \sim k(y \,|\, x_i)$
    evaluate the probability $P = \min\left\{1, \frac{\pi(y)k(x_i \,|\, y)}{\pi(x_i)k(y \,|\, x_i)}\right\}$
    $x_{i+1} \leftarrow \begin{cases} y, & \text{with probability } \; P \\ x_i, & \text{with probability } \; (1-P) \end{cases}$
  end for
  return samples $\{x_i\}_{i=1}^n$
```

which guarantees the stationarity condition.

Exercise 34. *Verify that Equation (360) guarantees the stationarity by substituting it into the transition kernel Equation (352).*

Finally, we demonstrate the pseudo-code for the Metropolis-Hastings accept/reject test in Algorithm 1.

4.4 Metropolis-Hastings-Green Test

In the previous section, we consider generating the proposal point stochastically and then derive the test for its acceptance. Here, we start from a different idea — generating the proposal point via a diffeomorphism (differentiable bijection with differentiable inverse) $f(x)$. That is, the transition kernel is

$$k(y \,|\, x) = g(x)\delta(y - f(x)) + (1 - g(x))\delta(y - x)\,, \tag{361}$$

where $g(x)$ is the acceptance probability. Then the stationarity condition (Definition 36) is

$$\int dx \; k(y \,|\, x)\pi(x) = \; \pi(y) \tag{362}$$

$$\int dx \; [g(x)\delta(y - f(x)) + (1 - g(x))\delta(y - x)]\pi(x) = \; \pi(y) \tag{363}$$

$$\int dz \; g(f^{-1}(z))\pi(f^{-1}(z))\delta(y - z)\left|\frac{\partial f^{-1}(z)}{\partial z}\right| + (1 - g(y))\pi(y) = \; \pi(y) \tag{364}$$

$$g(f^{-1}(y))\pi(f^{-1}(y))\left|\frac{\partial f^{-1}(y)}{\partial y}\right| = \; g(y)\pi(y)\,, \tag{365}$$

where we have used the change of variables $z = f(x)$. Analogously to the reasoning in the previous section, we have to choose function $g(y)$ that satisfies the equation above and is a probability $(0 \leq g(y) \leq 1)$. A reasonable guess is to choose

$$g(y) = \min\left\{1, \frac{\pi(f^{-1}(y))}{\pi(y)}\left|\frac{\partial f^{-1}(y)}{\partial y}\right|\right\}. \tag{366}$$

The stationarity condition in Equation (365) then becomes

$$\begin{aligned}
\min\left\{\pi(f^{-1}(y))\left|\frac{\partial f^{-1}(y)}{\partial y}\right|, \pi(f^{-2}(y))\left|\frac{\partial f^{-1}(x)}{\partial x}\right|_{x=f^{-1}(y)}\left|\frac{\partial f^{-1}(y)}{\partial y}\right|\right\} &= \min\left\{\pi(y), \pi(f^{-1}(y))\left|\frac{\partial f^{-1}(y)}{\partial y}\right|\right\} \\
\min\left\{\pi(f^{-1}(y))\left|\frac{\partial f^{-1}(y)}{\partial y}\right|, \pi(f^{-2}(y))\left|\frac{\partial f^{-2}(y)}{\partial y}\right|\right\} &= \min\left\{\pi(y), \pi(f^{-1}(y))\left|\frac{\partial f^{-1}(y)}{\partial y}\right|\right\},
\end{aligned} \tag{367}$$

where we use the notation for iterative application of the function, i.e. $f^2(x) = f(f(x))$, $f^{-2}(x) = f^{-1}(f^{-1}(x))$, and $f^0(x) = x$. The last condition holds in two cases: (i) when f preserves the target density $\pi(f^{-1}(y))\left|\frac{\partial f^{-1}(y)}{\partial y}\right| =$

Algorithm 2 Metropolis-Hastings-Green test

```
Require: proposal kernel $k(y\,|\,x)$, starting point $x_0$, target density $\pi(x)$, involution $f(x,y) = f^{-1}(x,y)$
  for iterations $i \in [0,n)$ do
    sample $y \sim k(y\,|\,x_i)$
    evaluate the probability $P = \min\left\{1, \frac{\bar{\pi}(f(x_i,y))}{\bar{\pi}(x_i,y)}\left|\frac{\partial f(x_i,y)}{\partial(x_i,y)}\right|\right\}$
    $(x_{i+1},y) \leftarrow \begin{cases} f(x_i,y), & \text{with probability } \; P \\ (x_i,y), & \text{with probability } \; (1-P) \end{cases}$
  end for
  return samples $\{x_i\}_{i=1}^n$
```

$\pi(y)$, which is very hard to satisfy because for every π we have to design some specific f (ii) when $f^{-2}(x) = f^0(x) = x$, i.e. the function is an involution. Indeed, then Equation (367) becomes

$$\min\left\{\pi(f^{-1}(y))\left|\frac{\partial f^{-1}(y)}{\partial y}\right|, \pi(y)\right\} \;=\; \min\left\{\pi(y), \pi(f^{-1}(y))\left|\frac{\partial f^{-1}(y)}{\partial y}\right|\right\}, \tag{368}$$

which is an identity. Finally, we get the following kernel

$$k(y\,|\,x) = \min\left\{1, \frac{\pi(f(x))}{\pi(y)}\left|\frac{\partial f(x)}{\partial x}\right|\right\}\delta(y - f(x)) + \left(1 - \min\left\{1, \frac{\pi(f(x))}{\pi(y)}\left|\frac{\partial f(x)}{\partial x}\right|\right\}\right)\delta(y - x)\,, \tag{369}$$

where we use $f^{-1}(x) = f(x)$ because f must be an involution.

The great news are that it is relatively easy to design function which is an involution. Indeed, for any invertible $f(x)$, we can define an involution on the extended space x, d

$$\bar{f}(x,d) = \begin{cases} (f(x), -d)\,, & \text{if } d = 1 \\ (f^{-1}(x), -d)\,, & \text{if } d = 1\,, \end{cases} \tag{370}$$

where the binary variable $d \in \{-1, 1\}$ defines the direction in which we apply the function.

Exercise 35. *Prove that $\bar{f}$ from Equation (370) is an involution.*

However, the bad news are that involution always iterates between two points, i.e.

$$x \to f(x) \to f^2(x) = x \to f^3(x) = f(x) \to f^4(x) = x \to \dots\,. \tag{371}$$

In order to cover the entire state-space, one can introduce auxiliary random variables that extend the state-space [Kirill:the proposal distribution in the previous section is exactly the same thing, the latent variables have the same flavour in Section 3.3]. That is, instead of sampling from $\pi(x)$ let's sample from the extended target distribution

$$\bar{\pi}(x,y) = \pi(x)k(y\,|\,x)\,, \tag{372}$$

where $k(y\,|\,x)$ can be any distribution which we can efficiently sample from and evaluate the density. Thus, by resampling y at every iteration, we can avoid jumping between two points and get the Metropolis-Hastings-Green test (see Algorithm 2). Choosing different $k(y\,|\,x)$ and different involutions f one can describe many existing MCMC algorithms (**?**).

4.5 Langevin Dynamics

Consider the following Stochastic Differential Equation (SDE)

$$dx_t = v_t(x_t)dt + \sigma_t dW_t\,,\; x_{t=0} = x_0\,, \tag{373}$$

where $v_t(x_t)$ is a vector field, dW_t is the standard Wiener process, and σ_t is the noise scale. To get some intuition, it is useful to consider the discretization of this equation in time dt. The following discretization is called Euler integration-scheme

$$x_{t+dt} = x_t + v_t(x_t)dt + \sigma_t\sqrt{dt}\varepsilon\,,\ \varepsilon \sim \mathcal{N}(\varepsilon \,|\, 0, 1)\,. \tag{374}$$

As you can see, unlike the intergration of the Ordinary Differential Equations (ODEs), the integration of SDEs involves sampling a random variable $\varepsilon \sim \mathcal{N}(\varepsilon \,|\, 0, 1)$ at every step. Therefore, for a system that starts with some state x_0, we can reason about the distribution of the coordinates x_t, i.e. x_t is a random variable that is defined as the integration of Equation (373) starting from x_0.

Let's denote the density of x_t as $p_t(x)$. Then we will introduce the following fact without proof.

Theorem 9 (The Fokker-Planck Equation)**.** *The density of the random variable x_t defined by Equation (373) changes according to the following Partial Differential Equation (PDE)*

$$\frac{\partial p_t(x)}{\partial t} = -\langle\nabla, p_t(x)v_t(x)\rangle + \frac{\sigma_t^2}{2}\Delta p_t(x)\,. \tag{375}$$

This equation allows us to analyse the dynamics of distributions for given $v_t(x)$ and σ_t .

Let's ask the following question: can we find the vector field $v_t(x)$ and noise schedule σ_t such that they preserve the target density $\pi(x)$? In other words, if the density equals $p_t(x) = \pi(x)$ we want to find such $v_t(x), \sigma_t$ that the time-derivative is zero, i.e.

$$\frac{\partial p_t(x)}{\partial t} = -\langle\nabla, p_t(x)v_t(x)\rangle + \frac{\sigma_t^2}{2}\Delta p_t(x) \tag{376}$$

$$0 = -\langle\nabla, \pi(x)v_t(x)\rangle + \frac{\sigma_t^2}{2}\Delta \pi(x) \tag{377}$$

$$\langle\nabla, \pi(x)v_t(x)\rangle = \frac{\sigma_t^2}{2}\langle\nabla, \nabla\pi(x)\rangle \tag{378}$$

$$\langle\nabla, \pi(x)v_t(x)\rangle = \left\langle\nabla, \pi(x)\frac{\sigma_t^2}{2}\nabla\log\pi(x)\right\rangle. \tag{379}$$

From the last equation, we see that, in particular, $v_t(x) = \frac{\sigma_t^2}{2}\nabla\log\pi(x)$ is a solution of the equation. [Kirill:there are many other solutions of this equation, but we don't discuss them here] This vector field defines the following family of SDEs

Definition 37 (The Langevin dynamics)**.** *For the given target density $\pi(x)$, the Langevin dynamics refers to the following family of SDEs*

$$dx_t = \frac{\sigma_t^2}{2}\nabla\log\pi(x)dt + \sigma_t dW_t\,,\ \forall\ \sigma_t\,. \tag{380}$$

We have already demonstrated that the Langevin dynamics preserves $\pi(x)$, now we want to analyse its convergence to the target $\pi(x)$ from any starting density $p_{t=0}(x)$.

Theorem 10 (Convergence of the Langevin dynamics)**.** *The KL-divergence between the current density of the Langevin dynamics $p_t(x)$ and the target density $\pi(x)$ changes as follows*

$$\frac{\partial}{\partial t}D_{\mathrm{KL}}(p_t(x), \pi(x)) = -\frac{\sigma_t^2}{2}\int dx\; p_t(x)\left\|\nabla\log\frac{\pi(x)}{p_t(x)}\right\|^2 \leq 0\,. \tag{381}$$

Proof. First, let's rewrite the Fokker-Planck equation

$$\frac{\partial p_t(x)}{\partial t} = -\langle\nabla, p_t(x)v_t(x)\rangle + \frac{\sigma_t^2}{2}\Delta p_t(x) \tag{382}$$

$$\frac{\partial p_t(x)}{\partial t} = -\left\langle\nabla, p_t(x)\left(v_t(x) - \frac{\sigma_t^2}{2}\nabla\log p_t(x)\right)\right\rangle, \tag{383}$$

then let's put in the corresponding values of $v_t(x), \sigma_t$

$$\frac{\partial p_t(x)}{\partial t} = -\left\langle \nabla, p_t(x)\left(\frac{\sigma_t^2}{2}\nabla \log \frac{\pi(x)}{p_t(x)}\right)\right\rangle . \tag{384}$$

Then, we take the time-derivative of the KL-divergence

$$\begin{aligned}
\frac{\partial}{\partial t} D_{\text{KL}}(p_t(x), \pi(x)) = \;& \frac{\partial}{\partial t}\int dx\; p_t(x) \log \frac{p_t(x)}{\pi(x)} = \int dx\; \frac{\partial p_t(x)}{\partial t} \log \frac{p_t(x)}{\pi(x)} + \int dx\; p_t(x) \frac{\pi(x)}{p_t(x)}\frac{1}{\pi(x)}\frac{\partial p_t(x)}{\partial t} \\
= \;& \int dx\; \frac{\partial p_t(x)}{\partial t} \log \frac{p_t(x)}{\pi(x)} = -\int dx\; \left\langle \nabla, p_t(x)\left(\frac{\sigma_t^2}{2}\nabla \log \frac{\pi(x)}{p_t(x)}\right)\right\rangle \log \frac{p_t(x)}{\pi(x)} \qquad (385) \\
= \;& \int dx\; p_t(x) \left\langle \frac{\sigma_t^2}{2}\nabla \log \frac{\pi(x)}{p_t(x)}, \nabla \log \frac{p_t(x)}{\pi(x)}\right\rangle \qquad (386) \\
= \;& -\frac{\sigma_t^2}{2}\int dx\; p_t(x) \left\|\nabla \log \frac{\pi(x)}{p_t(x)}\right\|^2 \leq 0 , \qquad (387)
\end{aligned}$$

where to get Equation (386) we used integration by parts. □

4.6 Hamiltonian/Hybrid Monte Carlo (HMC) (?)

Inspired by the Metropolis-Hastings-Green test (see Algorithm 2) we can design the test using f that preserves the density and the volume. Such dynamics are described by the Hamiltonian mechanics, which we quickly recall here. The equations of motion in the Hamiltonian mechanics are as follows

$$\frac{dx}{dt} = \nabla_v H(x, v) , \tag{388}$$

$$\frac{dv}{dt} = -\nabla_x H(x, v) . \tag{389}$$

In particular, the Hamiltonian mechanics described the following famous example.

Example 11. *Newton's mechanics is described by the following Hamiltonian*

$$H(x, v) = \frac{m}{2}\|v\|^2 + U(x) , \tag{390}$$

where m is the mass of the system, $U(x)$ is the potential energy. Indeed, the equations then become

$$\frac{dx}{dt} = mv , \tag{391}$$

$$\frac{dv}{dt} = -\nabla_x U(x) . \tag{392}$$

The candidate for the proposal function f in the Metropolis-Hastings-Green test is the flow defined by the Hamiltonian dynamics. Flow is a transformation that takes the initial point of the trajectory and propagates it until time t as follows

$$\varphi_t(x_0, v_0) = (x_t, v_t) , \quad \frac{d}{dt}\varphi_t(x_0, v_0) = \left(\frac{dx_t}{dt}, \frac{dv_t}{dt}\right) = (\nabla_v H(x, v), -\nabla_x H(x, v)) . \tag{393}$$

The flow φ_t of the Hamiltonian dynamics has two important properties: energy conservation (constant density) and volume conservation (the determinant of Jacobian equals 1). Formally, these properties are stated in the following propositions.

Proposition 13 (Energy Conservation)**.** *For the integral curve x_t, v_t, the Hamiltonian is constant along the curve, i.e.*

$$\forall\; t, s , \quad H(x_t, v_t) = H(x_s, v_s) . \tag{394}$$

Proof.

$$\frac{d}{dt}H(x_t, v_t) = \left\langle \nabla_x H(x_t, v_t), \frac{dx}{dt} \right\rangle + \left\langle \nabla_v H(x_t, v_t), \frac{dv}{dt} \right\rangle \tag{395}$$

$$= \langle \nabla_x H(x_t, v_t), \nabla_v H(x, v) \rangle + \langle \nabla_v H(x_t, v_t), -\nabla_x H(x, v) \rangle = 0\,. \tag{396}$$

□

Proposition 14 (Volume Preserving)**.** *The flow conserves the volume,*

$$\left| \frac{\partial \varphi_t(x, v)}{\partial x, v} \right| = 1 + \int_0^t d\tau \; \mathrm{div}(\nabla_v H(x_\tau, v_\tau), -\nabla_x H(x_\tau, v_\tau)) = 1\,, \tag{397}$$

Proof.

$$\mathrm{div}(\nabla_v H(x_\tau, v_\tau), -\nabla_x H(x_\tau, v_\tau)) = \langle \nabla_x, \nabla_v H(x, v) \rangle + \langle \nabla_v, -\nabla_x H(x, v) \rangle = 0\,. \tag{398}$$

□

Now, once we see these nice properties of Hamiltonian mechanics, we can propose the following scheme. For the target density $\pi(x)$, we define the extended target density $\bar{\pi}(x, v)$ as follows

$$\bar{\pi}(x, v) \propto \exp\left(-\frac{1}{2}\|v\|^2 + \log \pi(x) \right) = \exp(-H(x, v))\,, \quad \text{where} \;\; H(x, v) \coloneqq \frac{1}{2}\|v\|^2 - \log \pi(x)\,. \tag{399}$$

Note that here we introduce Hamiltonian $H(x, v)$ with the potential $-\log \pi(x)$. Clearly, if we can sample from $\bar{\pi}(x, v)$, we can sample from $\pi(x)$, because

$$\int dv \; \bar{\pi}(x, v) = \pi(x)\,. \tag{400}$$

Now, once we found the Hamiltonian $H(x, v)$, we can sample random $v \sim \mathcal{N}(v \,|\, 0, 1)$ simulate the Hamiltonian dynamics with $H(x, v)$ and use it as a proposal for sampling. Then we can decide on accept/reject using the Metropolis-Hasting-Green test Algorithm 2 and repeat. The final step for figuring out the algorithm is the evaluation of the Jacobian of the integration scheme for the Hamiltonian dynamics.

Definition 38 (Leap-Frog Integrator)**.** *For the Hamiltonian $H(x, v) = \frac{1}{2}\|v\|^2 + U(x)$, the following integration scheme is called Leap-Frog* (*velocity-Verlet integrator*)

$$v_{t+dt/2} = v_t - \frac{dt}{2}\nabla U(x_t)\,, \tag{401}$$

$$x_{t+dt} = x_t + dt v_{t+dt/2}\,, \tag{402}$$

$$v_{t+dt} = v_{t+dt/2} - \frac{dt}{2}\nabla U(x_{t+dt})\,. \tag{403}$$

The most important property of this integrator is that it preserves the volume, as formalized in the following proposition.

Proposition 15. *Leap-Frog integrator conserves the volume, i.e. the absolute value of the Jacobian determinant is* 1

$$\left| \frac{\partial(x_{t+dt}, v_{t+dt})}{\partial(x_t, v_t)} \right| = 1\,. \tag{404}$$

Proof. The determinant of the Jacobian is defined as

$$\left| \frac{\partial(x_{t+dt}, v_{t+dt})}{\partial(x_t, v_t)} \right| = \begin{vmatrix} \frac{\partial x_{t+dt}}{\partial x_t} & \frac{\partial x_{t+dt}}{\partial v_t} \\ \frac{\partial v_{t+dt}}{\partial x_t} & \frac{\partial v_{t+dt}}{\partial v_t} \end{vmatrix} = \begin{vmatrix} \frac{\partial x_{t+dt}}{\partial x_t} & \frac{\partial x_{t+dt}}{\partial v_t} \\ -\frac{dt}{2}\nabla^2 U(x_t) - \frac{dt}{2}\frac{\partial \nabla U(x_{t+dt})}{\partial x_{t+dt}}\frac{\partial x_{t+dt}}{\partial x_t} & \mathbb{1} - \frac{dt}{2}\frac{\partial \nabla U(x_{t+dt})}{\partial x_{t+dt}}\frac{\partial x_{t+dt}}{\partial v_t} \end{vmatrix} \tag{405}$$

$$= \begin{vmatrix} -\frac{dt}{2}\nabla^2 U(x_t) - \frac{dt}{2}\frac{\partial \nabla U(x_{t+dt})}{\partial x_{t+dt}}\frac{\partial x_{t+dt}}{\partial x_t} & \mathbb{1} - \frac{dt}{2}\frac{\partial \nabla U(x_{t+dt})}{\partial x_{t+dt}}\frac{\partial x_{t+dt}}{\partial v_t} \\ \frac{\partial x_{t+dt}}{\partial x_t} & \frac{\partial x_{t+dt}}{\partial v_t} = dt \cdot \mathbb{1} \end{vmatrix}\,. \tag{406}$$

Algorithm 3 Hamiltonian Monte Carlo

```
Require: starting point $x_0$, target density $\pi(x)$, step-size $dt$
  for iterations $j \in [0, n)$ do
    sample $v \sim \mathcal{N}(v \,|\, 0, 1)$
    for $i \in [0, n)$ do
      $(x_{i+1}, v_{i+1}) \leftarrow \texttt{Leap-Frog}_{dt}(x_i, v_i)$
    end for
    $P = \min\Big\{1, \frac{\bar{\pi}(x_n, v_n)}{\bar{\pi}(x_0, y_0)}\Big\} = \min\Big\{1, \exp\Big(-\frac{1}{2}\|v_n\|^2 - \frac{1}{2}\|v_0\|^2 + \log\pi(x_n) - \log\pi(x_0)\Big)\Big\}$
    $x_{j+1} \leftarrow \begin{cases} x_n, & \text{with probability } \; P \\ x_0, & \text{with probability } \; (1-P) \end{cases}$
  end for
  return samples $\{x_j\}_{j=1}^n$
```

In the last expression we have to have $(-1)^d$ where d is the dimensionality, but we are interested only in the absolute value of the determinant. Now, we want to use the fact (see proof in Proposition 18)

$$\det\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \det(D)\det(A - BD^{-1}C)\,. \tag{407}$$

Then, we have

$$\begin{aligned}
\left|\frac{\partial(x_{t+dt}, v_{t+dt})}{\partial(x_t, v_t)}\right| &= \det(dt\cdot\mathbb{1})\det\bigg(-\frac{dt}{2}\nabla^2 U(x_t) - \frac{dt}{2}\frac{\partial\nabla U(x_{t+dt})}{\partial x_{t+dt}}\frac{\partial x_{t+dt}}{\partial x_t} - & (408)\\
&\quad - \left(\mathbb{1} - \frac{dt}{2}\frac{\partial\nabla U(x_{t+dt})}{\partial x_{t+dt}}\frac{\partial x_{t+dt}}{\partial v_t}\right)\left(\frac{\partial x_{t+dt}}{\partial v_t}\right)^{-1}\frac{\partial x_{t+dt}}{\partial x_t}\bigg) & (409)\\
&= \det(dt\cdot\mathbb{1})\det\bigg(-\frac{dt}{2}\nabla^2 U(x_t) - \left(\frac{\partial x_{t+dt}}{\partial v_t}\right)^{-1}\frac{\partial x_{t+dt}}{\partial x_t}\bigg) & (410)\\
&= \det(dt\cdot\mathbb{1})\det\bigg(-\frac{dt}{2}\nabla^2 U(x_t) - \frac{1}{dt}\left(\mathbb{1} - \frac{dt^2}{2}\nabla^2 U(x_t)\right)\bigg) = 1\,. & (411)
\end{aligned}$$

□

To formalize the algorithm let's define the following function

$$\begin{aligned}
v_{t+dt/2} &= v_t - \frac{dt}{2}\nabla U(x_t)\,, & (412)\\
x_{t+dt} &= x_t + dt v_{t+dt/2}\,, & (413)\\
v_{t+dt} &= v_{t+dt/2} - \frac{dt}{2}\nabla U(x_{t+dt})\,, & (414)\\
\texttt{Leap-Frog}_{dt}(x_t, v_t) &:= x_{t+dt}, v_{t+dt}\,. & (415)
\end{aligned}$$

Finally, the algorithm is presented in Algorithm 3.

References

[Burda u. a. 2015] Burda, Yuri ; Grosse, Roger ; Salakhutdinov, Ruslan: Importance weighted autoencoders. In: *arXiv preprint arXiv:1509.00519* (2015)

[Ho u. a. 2020] Ho, Jonathan ; Jain, Ajay ; Abbeel, Pieter: Denoising diffusion probabilistic models. In: *Advances in neural information processing systems* 33 (2020), S. 6840–6851

[Jaynes 1957] Jaynes, Edwin T.: Information theory and statistical mechanics. In: *Physical review* 106 (1957), Nr. 4, S. 620

[Kingma u. a. 2013] Kingma, Diederik P. ; Welling, Max u. a.: *Auto-encoding variational bayes*. 2013

[Sohl-Dickstein u. a. 2015] Sohl-Dickstein, Jascha ; Weiss, Eric ; Maheswaranathan, Niru ; Ganguli, Surya: Deep unsupervised learning using nonequilibrium thermodynamics. In: *International conference on machine learning* pmlr (Veranst.), 2015, S. 2256–2265

A Recall of some facts

A.1 Basic Concepts from Statistics

Theorem 11 (Central Limit Theorem)**.** *Consider a random variable X with the density $p(x)$, mean μ, and variance σ^2. The following convergence result takes place*

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)\,, \quad \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i\,, x_i \sim p(x)\,, \tag{416}$$

where $\xrightarrow{d}$ denotes the convergence in distribution.

Proof. Recall the basic properties of characteristic functions □

A.2 Differentiation

Definition 39 (Differentiable function)**.** *The function $f : \mathcal{X} \to \mathbb{R}$ is differentiable at point $x \in \mathcal{X}$ if there exist linear operator $L_x : \mathcal{X} \to \mathbb{R}$ such that*

$$\forall\, h \in \mathcal{X} \;: f(x+h) = f(x) + L_x[h] + o(\|h\|)\,. \tag{417}$$

The operator $L_x[h]$ is called differential and is usually denoted as $df(x)[h]$, i.e. we have

$$f(x+h) = f(x) + df(x)[h] + o(\|h\|)\,. \tag{418}$$

Linear operators in Hilbert spaces can be represented as a scalar product. Namely, if $df(x)[h]$ is linear w.r.t. h we can consider applying this operator to the basis vectors e_i, then we have

$$df(x)[h] = df(x)\left[\sum_i h_i e_i\right] = \sum_i h_i \underbrace{df(x)[e_i]}_{\partial f(x)/\partial x_i} = \langle h, \nabla f(x)\rangle\,, \tag{419}$$

where we denote $df(x)[e_i]$ as $\partial f(x)/\partial x_i$ and call it partial derivative. The vector of partial derivatives is called gradient and is denoted as

$$\nabla f(x) = \begin{bmatrix} \partial f(x)/\partial x_1 \\ \vdots \\ \partial f(x)/\partial x_d \end{bmatrix}. \tag{420}$$

A.3 Matrix Calculus

Proposition 16 (Woodbury-Morrison formula)**.** *The following identity holds*

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U\big(C^{-1} + VA^{-1}U\big)^{-1}VA^{-1}\,. \tag{421}$$

Proof.

$$\begin{aligned}
(A+UCV)^{-1} &= A^{-1} - A^{-1}U\big(C^{-1} + VA^{-1}U\big)^{-1}VA^{-1} && (422)\\
I &= (A+UCV)A^{-1} - (A+UCV)A^{-1}U\big(C^{-1} + VA^{-1}U\big)^{-1}VA^{-1} && (423)\\
I &= \big(I + UCVA^{-1}\big) - U\big(I + CVA^{-1}U\big)\big(C^{-1} + VA^{-1}U\big)^{-1}VA^{-1} && (424)\\
I &= \big(I + UCVA^{-1}\big) - UC\big(C^{-1} + VA^{-1}U\big)\big(C^{-1} + VA^{-1}U\big)^{-1}VA^{-1} && (425)\\
I &= \big(I + UCVA^{-1}\big) - UCVA^{-1} = I && (426)
\end{aligned}$$

□

Proposition 17 (Determinant lemma)**.**

$$\det(A+UV) = \det(I + VA^{-1}U)\det(A) \tag{427}$$

Proposition 18 (Block Determinant)**.**

$$\det\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \det(D)\det(A - BD^{-1}C) \tag{428}$$

Proof. Let's note that

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} 1 & 0 \\ -D^{-1}C & D^{-1} \end{bmatrix} = \begin{bmatrix} A - BD^{-1}C & BD^{-1} \\ 0 & 1 \end{bmatrix}. \tag{429}$$

Then, we have

$$\det\begin{bmatrix} A & B \\ C & D \end{bmatrix}\det\begin{bmatrix} 1 & 0 \\ -D^{-1}C & D^{-1} \end{bmatrix} = \det\begin{bmatrix} A - BD^{-1}C & BD^{-1} \\ 0 & 1 \end{bmatrix} \tag{430}$$

$$\det\begin{bmatrix} A & B \\ C & D \end{bmatrix}\det(D)^{-1} = \det(A - BD^{-1}C) \tag{431}$$

$$\det\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \det(D)\det(A - BD^{-1}C)\,. \tag{432}$$

□

B Exam

Question 1. 1. *Consider $x_1, \ldots, x_N$ — iid samples from the Poisson distribution Poiss*(λ) $= P(X = k \,|\, \lambda) = \exp(-\lambda)\frac{\lambda^k}{k!}$. *Find the maximum likelihood estimation of* λ.

2. *The Metropolis-Hastings algorithm. For example: write down the transition kernel, prove stationarity of the kernel, what are the practical guidelines for the choosing the proposal?*

Question 2. 1. *Consider $x_1, \ldots, x_N$ — iid samples from the exponential density function*

$$p(x \,|\, \lambda) = \lambda \exp(-\lambda x)\,, \quad x \geq 0\,, \lambda > 0.$$

Find the conjugate prior $p(\lambda)$ and the corresponding posterior $p(\lambda \,|\, \mathcal{D})$.

2. *Denoising Diffusion Probabilistic Models. For example: prove the likelihood lower bound.*

Question 3. 1. *Consider $x_1, \ldots, x_N$ — iid samples from the Poisson distribution Poiss*(λ)

$$P(X = k \,|\, \lambda) = \exp(-\lambda)\frac{\lambda^k}{k!}.$$

Find the conjugate prior $p(\lambda)$ and the corresponding posterior $p(\lambda \,|\, \mathcal{D})$.

2. *Hamiltonian Monte Carlo. For example: what is Hamiltonian mechanics?, prove that Leap-Frog preserves the volume.*

Question 4. 1. *For the exponential family $p(x \,|\, \theta) = \frac{f(x)}{g(\theta)} \exp\big(\theta^T u(x)\big)$, find the expression for*

$$\frac{\partial^2}{\partial \theta_i \theta_j} \log g(\theta) = ?$$

2. *Conjugate prior. For example: give the definition and prove that , how one could do continual learning with conjugate distributions?*

Question 5. 1. *Represent the gamma distribution*

$$\mathcal{G}(x \,|\, a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)\,,$$

as a density from the exponential family, find the expectations of the sufficient statistics.

2. *Importance Sampling. For example: what is importance sampling, what are the properties of the estimator, is it possible to do importance sampling if the target density is unnormalized?*

Question 6. 1. *For the transition kernel*

$$\bar{k}(x' \,|\, x) = \int dy\ (g(x,y)\delta(x' - y) + (1 - g(x,y))\delta(x' - x))k(y \,|\, x)\pi(x)\,,$$

verify that

$$g(x,y) = \min\left\{1, \frac{k(x \,|\, y)\pi(y)}{k(y \,|\, x)\pi(x)}\right\}$$

guarantees the stationarity of $\bar{k}$ w.r.t. $\pi(x)$.

2. *Variational Auto Encoder. For example: define the probabilistic model of VAE, prove the lower bound on the likelihood.*

Question 7. 1. *Find the density $q(x)$ of the random variable defined on $x \in \mathbb{R}$ that has the biggest entropy, given expectation μ, and given variance σ^2.*

2. *Discrete Markov Chains. For example: how to find the stationary distribution?, is the stationary distribution unique?, how to find the marginal of the n-th step?*

Question 8. 1. *Prove*

$$D_{\mathrm{KL}}(q,p) \leq \chi^2(q,p)\,, \quad \textit{where} \quad \chi^2(q,p) = \mathbb{E}_{p(x)}\left(\frac{q(x)}{p(x)} - 1\right)^2 .$$

2. *Exponential family. For example: define the exponential family, defin the sufficient statistics, demonstrate the property of the normalization constant derivative, what's the conjugate prior?*

Question 9. 1. *Prove that the following optimization problems are equivalent*

$$\max_\theta \mathbb{E}_{x\sim p_{data}(x)} \log\big[\mathbb{E}_{z\sim p(z\,|\,\theta)} p(x\,|\,z;\theta)\big] \iff \min_\theta \min_\eta D_{\mathrm{KL}}(p_{data}(x)p(z\,|\,x;\eta), p(x,z\,|\,\theta))\,.$$

2. *Model Selection. For example: how one can choose prior?, what is evidence?, how one can estimate evidence?*

Question 10. 1. *Given the likelihood model $p(\mathcal{D}\,|\,\theta)$ and the prior $p(\theta)$, for the variational posterior $q(\theta)$, derive the evidence lower bound*

$$\log p(\mathcal{D}) \geq ?$$

2. *Baeysian ML models. For example: Bayesian linear regression, Bayesian Logistic Regression, Laplace approximation.*