

Statistical learning theory

Carlos Cotrini
Department of Computer Science
ETH Zürich
Switzerland

March 1, 2021

Introduction

This is a script for the course statistical learning theory at ETH. In this introduction, we explain what statistical learning is about and what you will learn in this course.

What is learning?

Let us consider a fictitious scenario. You are given a handful of patients in a hospital. Some of which have been diagnosed with pneumonia. Each patient is represented as a sequence of values, where each value describes a relevant property of the patient, like the body temperature or the number of days for which the patient has been coughing.

With enough training and discipline, you can start to recognize patterns among these patients. For example, patients with a high temperature and who have been coughing for long are likely to have pneumonia. We can call this pattern a *model* of reality. The process by which a human constructs a model can be called (human) *learning*.

It has been demonstrated that machines can perform a process very similar to human learning. One can design an algorithm that takes as input patient representations and produces from them a (mathematical) *model* that can effectively diagnose pneumonia on patients that were not part of the initial set of patients. We can call this process *machine learning*. The act of computing the model is called *training*.

What is statistical learning?

Statistical learning is a subfield of machine learning where the algorithm trains models by searching for a model that minimizes a given cost function. An example of statistical learning is linear regression. You compute a line that

models the behavior of an observed set of points by searching for the line that minimizes the sum of the squared distances from each point to the line.

A popular method of statistical learning is *empirical risk minimization*, which we illustrate with the example of pneumonia. In this method, we first define a *cost function* R that measures (i) the cost of misdiagnosing pneumonia and (ii) the cost of failing to diagnose pneumonia. Observe that the two costs may not necessarily be the same. A simple probabilistic argument shows that to minimize costs in the long run, the computer program should train a model that minimizes *the expected cost*. This is defined as the expected value $\mathbb{E}_X [R(c, X)]$ of $R(c, X)$, where $R(c, X)$ indicates the cost that model c incurs when analyzing a random patient X . The expected cost is often impossible to evaluate, because we cannot access the distribution of X . So we instead minimize an empirical estimate, called the *empirical cost*, using a sample X_1, \dots, X_N from X 's distribution. In our context, this sample is the set of patients that we use to build our model.

Empirical risk minimization advocates then to design algorithms that output models c^* that minimize

$$\frac{1}{N} \sum_{i \leq N} R(c, X_i). \quad (1)$$

Statistical learning theory is the study of algorithms that train models via empirical risk minimization.

What will you learn in this course?

This is a course that studies and proposes methods for statistical learning. Some of the ideas proposed here are intended to establish an *information theory of algorithms*, where algorithms can be seen as communication channels. An algorithm takes as input a random variable, representing a training set, and outputs a random variable, representing a model.

This representation implies that when given as input a training set, the algorithm produces a *posterior distribution* over a set of possible models. A fundamental question is then *how do we validate that the algorithm is producing the output that we desire?*

This course proposes answers to this question. Along the way, you will learn the following:

- How to design probability distributions over hypothesis classes.

- Statistical learning requires an approach of score maximization rather than an approach of cost minimization.
- The concept of entropy plays a key role when defining a notion of algorithm validation.

Outline

The course starts in the middle of February with an introduction to statistical learning. Afterwards, in March, we present *maximum entropy and posterior agreement*, a method for designing and validating algorithms for statistical learning. We then study some instantiations of maximum entropy in the case of clustering. In April, we study other approaches for clustering and show their connection to maximum entropy. In May, we study some popular techniques for approximating intractable distributions that arise when following the maximum-entropy approach. Finally, we propose information-theoretic foundations for posterior agreement.

This script is planned to be completed throughout the next years. For this year, the script focuses on maximum entropy and posterior agreement.

Acknowledgements and contributions

This script is a remake of a previous script that was maintained by former members of the institute of machine learning at ETH: Alexey Gronskiy, Sergio Solórzano, Francesco Locatello, and David Tedaldi. This new version contains the following contributions:

- An earlier introduction to posterior agreement.
- A new toy problem, the random array problem, which illustrates the effectiveness of maximum entropy and posterior agreement over empirical risk minimization.
- A new relation between the holdout and posterior agreement, discovered by Buhmann.
- Calculations and experiments by Kenneth Rose [11, 12] on deterministic annealing.
- An analysis by Rose on the temperatures at which phase transitions occur in deterministic annealing.

- A presentation of the Laplace method and its application for least-angle clustering, proposed by Buhmann.
- A new chapter discussing how to explore Gibbs distributions via sampling.
- An alternative information-theoretic motivation for posterior agreement, including a detailed recap on Shannon's channel coding theorem.

Remarks

This is the first draft of this new version of the script and it is an independent and voluntary attempt to provide more supporting material to you. Therefore, you are kindly asked to be patient to typos and terse explanations. This does not mean, however, that you have to deal with them. You are encouraged to report typos in the git repo and post in Piazza whenever you encounter some part is unclear to you. I'd be very happy to discuss them with you.

Chapter 1

Maximum entropy and posterior agreement

1.1 Introduction

We present here an alternative statistical learning method called ME+PA (maximum entropy + posterior agreement). This method is robust and can be applied to a wide variety of learning problems. In addition, it has been experimentally demonstrated that it generalizes better than empirical risk minimization techniques in different domains, including brain image analysis [6, 13], access control [5, 3], minimum spanning trees [8], and predicting rankings in chess tournaments [2]. In particular, the notion of posterior agreement captures a notion similar to that of mutual information, when seeing learning algorithms as communication channels that use models to communicate information about an underlying phenomenon.

1.2 Overview of ME+PA

Before making a formal presentation, we give an intuitive overview. ME+PA consists of the following steps:

Step 1: Hypothesis class Define a hypothesis class \mathcal{C} . This indicates the class of candidate models you want to train. Examples of hypothesis classes are the class of linear regression models, the class of classification trees, and the class of neural networks with a specific architecture.

Step 2: Cost function The next step is to define a cost function R . This function takes as input a model $c \in \mathcal{C}$ and an observation X and produces a cost value $R(c, X)$ measuring how well the model fits the given observation. Since cost functions are usually defined by humans, we require cost functions to be non-negative.

Step 3: Maximum entropy For an arbitrary observation X , we now define a posterior probability distribution $p(\cdot | X)$ over \mathcal{C} , conditioned on X . For $c \in \mathcal{C}$, $p(c | X)$ should measure how much we believe that c is the right model for X . This belief is quantified by $R(c, X)$. The lower this cost is, the more we believe that c is the right model for X . We argue later in Section 1.4 that a natural posterior distribution that fulfils this requirement is the *Gibbs distribution induced by R* :

$$p(c | X) \propto \exp\left(-\frac{1}{T}R(c, X)\right). \quad (1.1)$$

where $T > 0$ is a hyper-parameter called the *temperature*.

Step 4: Model selection The last step is to select an adequate temperature T . This can be done using standard techniques like cross validation.

An alternative model selection technique that we study in this course is *posterior agreement*. For this, we use two independent observations X' and X'' and define the *posterior agreement kernel* between X' and X'' :

$$\kappa(X', X'') := \sum_{c \in \mathcal{C}} p(c | X)p(c | X''). \quad (1.2)$$

This kernel can be used to select a value for T from a set of candidate temperatures \mathcal{T} .

$$T^*, R^* = \arg \max_{T \in \mathcal{T}, R \in \mathcal{R}} \kappa(X', X'') \quad (1.3)$$

Finally, the two Gibbs distributions $p(\cdot | X')$ and $p(\cdot | X'')$ can be aggregated together to yield a final distribution over models

$$p^*(c | X', X'') \propto p(c | X')p(c | X''). \quad (1.4)$$

Observe that this distribution gives high probability only to those models c for which both $p(c | X')$ and $p(c | X'')$ are high. Therefore, $p^*(c | X', X'')$ is more robust to noise interference coming from either X' or X'' .

1.3 Formalization

We assume that we are interested in computing a model for a *phenomenon* of interest. The phenomenon is represented with a probability distribution p over an *instance space* \mathcal{X} . We assume that p is unknown to us.

We perceive the phenomenon through *observations*, which we represent with elements in \mathcal{X} drawn from p . In particular, we are interested in complex phenomena where making an observation is expensive, and in the worst case, we only have two independent observations X' and X'' . Depending on the context, we sometimes view X' and X'' instead as random variables with distribution p .

We also assume given a finite¹ hypothesis class \mathcal{C} comprising all candidate models.

Definition 1. A cost function is a function $R : \mathcal{C} \times \mathcal{R} \rightarrow [0, \infty)$.

Definition 2. For a cost function R and an observation X , a Gibbs distribution (induced by R and X) is defined by

$$p(c | X) \propto \exp\left(-\frac{1}{T}R(c, X)\right), \quad (1.5)$$

where $T > 0$ is a hyper-parameter called the temperature.

1.4 Maximum entropy

In this section, we explain the rationale behind the Gibbs distribution. Remember that for an observation X , we want to have a distribution $p(\cdot | X)$ such that, for $c \in \mathcal{C}$, $p(c | X)$ measures how much we believe that c is the right model for X . The base for that measure is the cost $R(c, X)$. The lower this cost is, the more we believe that c is the right model for X . We can therefore elicit the following requirement:

Requirement 1. For any $X \in \mathcal{X}$ and any $c_1, c_2 \in \mathcal{C}$, $R(c_1, X) \leq R(c_2, X) \Leftrightarrow P(c_1 | X) \leq P(c_2 | X)$.

There are too many distributions that fulfil this requirement. To narrow the set of candidate distributions, we follow the *maximum entropy principle* [9, 10].

¹ME+PA can also be applied to continuous hypothesis classes, but the foundations of this method have only been established for finite hypothesis classes.

The maximum entropy principle states that when choosing between two distributions that equally fulfil some criteria, choose the one with the highest entropy. This is equivalent to choosing the distribution that is “closest” to the uniform distribution with respect to the Kullback-Leibler divergence. By aiming for the “most uniform” distribution, we aim for a distribution that does not show arbitrary preferences for one model over the other. The distribution should only show preferences based on the cost function.

Requirement 2. $P(\cdot | X)$ shall have maximum entropy.

Finally, we add some regularity requirements.

Requirement 3. $P(\cdot | X)$ is a regular distribution. That is, $P(c | X) \geq 0$, for any $c \in \mathcal{C}$, $\sum_{c \in \mathcal{C}} P(c | X) = 1$, and $\mathbb{E}_{C \sim p(\cdot | X)}[R(C, X)] < \infty$.

These requirements define a constrained optimization problem

$$\max_p H[p] \tag{1.6}$$

$$s.t. \quad p(c_1) \leq p(c_2), \text{ for any } c_1, c_2 \in \mathcal{C} \text{ with } R(c_1, X) \leq R(c_2, X), \tag{1.7}$$

$$p(c) \geq 0, \text{ for } c \in \mathcal{C}, \tag{1.8}$$

$$\sum_{c \in \mathcal{C}} p(c) = 1, \tag{1.9}$$

$$\mathbb{E}_{C \sim p}[R(C, X)] = \mu. \tag{1.10}$$

Here, μ is a hyper-parameter ensuring that the expectation is finite. We will later see that its exact value is not important and can be chosen using the posterior agreement principle.

We solve this problem by using Lagrange multipliers, forgetting about the inequality constraints, and then showing that the solution happens to fulfil those inequality constraints. The Lagrangian in this case is

$$\mathcal{L}(p, \lambda_1, \lambda_2) = H[p] + \lambda_1 \left(1 - \sum_{c \in \mathcal{C}} p(c) \right) + \lambda_2 (\mathbb{E}_{C \sim p}[R(C, X)] - \mu). \tag{1.11}$$

Therefore, for $c \in \mathcal{C}$,

$$\frac{\partial \mathcal{L}}{\partial p(c)} = -1 - \log p(c) + \lambda_1 + \lambda_2 R(c, X). \tag{1.12}$$

Setting this expression equal to zero and solving for $p(c)$, while using the constraint that $\sum_c p(c) = 1$, yields that

$$p(c) \propto \exp(\lambda_2 R(c, X)). \tag{1.13}$$

Observe that $p(\cdot)$ already fulfils the constraint (1.8). If we enforce the constraint (1.7), then λ_2 must be zero or negative. Furthermore, if $\lambda_2 = 0$, which is the case when $\mu = 1/|\mathcal{C}|\sum_c R(c, X)$, then $p(\cdot)$ is the uniform distribution over \mathcal{C} , which is not interesting for us as $p(\cdot)$ does not discriminate the different models in \mathcal{C} in this case. We can therefore agree that $\lambda_2 < 0$. To make this distribution look more like the Gibbs distribution used in statistical physics to model particle systems, we define $T := 1/\lambda_2$. Hence, the solution of the constrained optimization problem is

$$p(c) \propto \exp\left(-\frac{1}{T}R(c, X)\right), \text{ for some } T > 0. \quad (1.14)$$

We refrain from computing the exact value for T as it is analytically very difficult and unnecessary. The value of T is defined by μ , which is a hyper-parameter. Hence, we just leave T as another hyper-parameter. T can also be seen as a concentration hyper-parameter, defining how concentrated $p(\cdot)$ is around the global minima of $R(c, X)$. To see this, we have the following exercise.

Exercise 1. *Prove that*

$$\lim_{T \rightarrow \infty} p(c) = \frac{1}{|\mathcal{C}|}. \quad (1.15)$$

$$\lim_{T \rightarrow 0} p(c) = \begin{cases} a & \text{if } c \in \arg \min_{c \in \mathcal{C}} R(c, X) \\ 0 & \text{otherwise.} \end{cases} \quad (1.16)$$

Here, a is a constant value.

Observe that p is unique. Any other p' that solves the constrained problem must also have the form given by Equation (1.14). Hence, the difference between p' and p must be in their corresponding values for T , but observe that a change in T induces a change in $\mathbb{E}_{\mathcal{C}}[R(C, X)]$. As a result, it is impossible that $p \neq p'$ if both happen to solve the constrained optimization problem and have $\mathbb{E}_{C \sim p}[R(C, X)] = \mathbb{E}_{C \sim p'}[R(C, X)] = \mu$.

The results in this section can be naturally adapted in some cases when \mathcal{C} is not discrete, like \mathbb{R}^m . In this case, one of the main challenges is to ensure that the Gibbs distribution has a bounded normalization constant. This is usually satisfied if the cost function goes to ∞ fast enough as $\|c\| \rightarrow \infty$.

Exercise 2. *Let $\mathcal{C} = \mathbb{R}^m$ and \mathcal{X} denote all subsets of pairs $(x, y) \in \mathbb{R}^m \times \mathbb{R}$. Consider the sum-of-squares cost function used for linear regression, which is defined for $c \in \mathcal{C}$ and $X \in \mathcal{X}$ as follows:*

$$R(c, X) = \sum_{(x, y) \in \mathcal{X}} (y - c^\top x)^2 + \lambda \|c\|^2. \quad (1.17)$$

Demonstrate that the solution to the constrained optimization problem given by Equations 1.6—1.10 is given by

$$p(\cdot | X) \propto \exp\left(-\frac{1}{T}R(c, X)\right), \quad \text{with } T > 0.$$

Observe that without the regularization term $\lambda \|c\|^2$, it would not be possible to satisfy all the constraints.

1.5 Posterior agreement

Once the Gibbs distribution $p(\cdot | X)$ for a cost function R and an observation X has been defined, the last step is to select a temperature $T > 0$. One way to do this is via standard model selection techniques like cross-validation. However, we investigate in this course a new technique inspired by information theory, called *posterior agreement*.

Definition 3. For a cost function R and a temperature $T > 0$. The posterior agreement kernel (induced by R and T) is a function $\kappa : \mathcal{X} \times \mathcal{X} \rightarrow [0, 1]$ such that, for any two observations $X', X'' \in \mathcal{X}$,

$$\kappa(X', X'') = \sum_{c \in C} p(c | X') p(c | X''). \quad (1.18)$$

Here, $p(\cdot | X')$ and $p(\cdot | X'')$ are two Gibbs distributions induced by R and with temperature T .

Definition 4 (The combined Gibbs distribution). Assume given two Gibbs distributions induced by a same cost function, but two different observations X' and X'' , and having the same temperature. The combined Gibbs distribution is defined as follows

$$p(c | X', X'') \propto p(c | X') p(c | X''). \quad (1.19)$$

Definition 5 (The posterior agreement principle). Assume given two independent observations X' and X'' of a phenomenon. Whenever making a choice ω from a set Ω of options concerning a particular learning algorithm, one must choose the option that maximizes the posterior agreement kernel:

$$\omega^* = \arg \max_{\omega \in \Omega} \kappa(X', X''). \quad (1.20)$$

1.6 Relation between the holdout and ME+PA

Many training algorithms require hyper-parameters to define the hypothesis class \mathcal{C} . A common technique to select those hyper-parameters is *the hold-out*.

In the hold-out, we draw two samples X' and X'' , usually called the *training sample* and the *testing sample* respectively. Models are trained in X' and then evaluated in X'' . A natural evaluation metric is *the out-of-sample error*:

$$\mathbb{E}_{C|X'} [R(C, X'')] = \mathbb{E}_{C|X'} [-T \log p(C | X'')] \quad (1.21)$$

$$\propto - \int p(c | X') \log p(c | X'') dc. \quad (1.22)$$

Here, $p(C | X')$ and $p(C | X'')$ are Gibbs distributions induced by R and some temperature T .

The hold-out advocates to select the hyper-parameter that minimizes this error.

In contrast, posterior agreement advocates to select the hyper-parameter that maximizes the posterior agreement kernel induced by R and some temperature T :

$$\mathbb{E}_{C|X'} [p(C | X'')]. \quad (1.23)$$

We argue here that the posterior agreement kernel is less sensitive to noise than the out-of-sample error. To see this suppose that $X' = X + \Delta'$ and $X'' = X + \Delta''$, where $X \in \mathbb{R}^m$ and that Δ', Δ'' are distributed according to a Gaussian. Furthermore, we assume that $\|\Delta'\| < 1$ and $\|\Delta''\| < 1$.

Exercise 3. Show that

$$\mathbb{E}_{C|X'} [R(C, X'')] = \mathbb{E}_{C|X} [R(C, X)] + O(\|\Delta\|) \text{ and} \quad (1.24)$$

$$\mathbb{E}_{C|X'} [p(C | X'')] = \mathbb{E}_{C|X} [R(C, X)] + O(\|\Delta\|^2). \quad (1.25)$$

Use the following second-order Taylor approximation:

$$p(C | X') = p(C | X) + (\nabla_X p(C | X))^\top \Delta' + O(\|\Delta'\|^2). \quad (1.26)$$

1.7 The random array

We present an artificial toy example, called *the random array* that illustrates the power of ME+PA. In particular, it illustrates the ways in which empirical risk minimization overconfidently generalizes when there is insufficient data and noise interference.

1.7.1 Setup

Consider a random array $\mathfrak{X} = (\mathfrak{X}_0, \dots, \mathfrak{X}_{n-1})$ with range \mathbb{R}^n . Suppose that $\mathfrak{X}_i \sim \mathcal{N}(i, \sigma_i)$. For convenience, we let σ_i be such that $\mathbf{P}(\mathfrak{X}_i < 0) = 0.05$. The random array \mathfrak{X} is our phenomenon.

Assume that we only have two observations X' and X'' , consisting of two arrays drawn at random from \mathfrak{X} 's distribution. We want to estimate, from just X' and X'' ,

$$i^* := \arg \min_{i \leq n} \mathbb{E}_{\mathfrak{X}}[\mathfrak{X}_i]. \quad (1.27)$$

Take a moment to convince yourself that $i^* = 0$. To give some intuition on the problem, consider Figure 1.1, which shows a realization of X' and X'' .

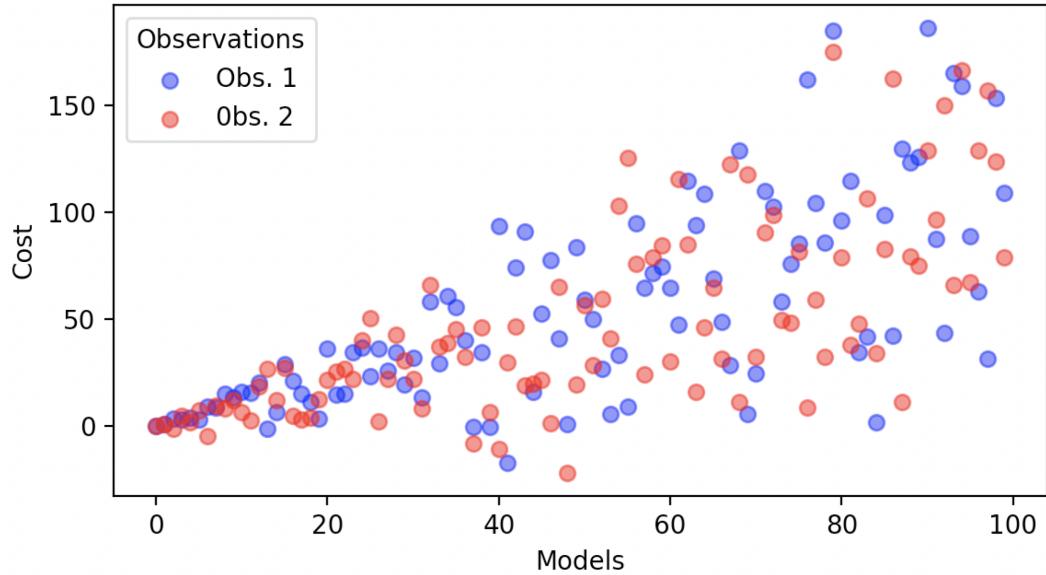


Figure 1.1

1.7.2 What happens if we use statistical learning to solve this problem?

Statistical learning proposes to train models via *empirical risk minimization (ERM)*. In ERM, we define an instance space \mathcal{X} , a hypothesis class \mathcal{C} , and a cost function R .

ERM trains a model in $c \in \mathcal{C}$ that minimizes the *expected cost* $\mathbb{E}_X[R(c, X)]$, where X is a random observation of the phenomenon. However, this requires

the distribution p behind the phenomenon, which we assume to be unavailable. In consequence, statistical learning advocates to approximate this expected cost with the *empirical cost* $\frac{1}{n} \sum_{i \leq n} R(c, X_i)$, where X_1, \dots, X_n is a sample of p .

We define the *empirical risk minimizer* as a model \hat{c} that minimizes the empirical cost. That is,

$$\hat{c} = \arg \min_{c \in \mathcal{C}} \sum_{i \leq n} R(c, X_i). \quad (1.28)$$

We assume for simplicity that \hat{c} always exist, although this is not true in general.

If we follow standard ERM, then we would estimate i^* with

$$\hat{i} := \arg \min_{i \leq n} X'_i + X''_i.$$

That is, the popular wisdom in statistical learning advocates to estimate i^* with the index where the minimum of $X' + X''$ is. However, this approach fails in probability as n increases. This is because, with high probability, there is some $i > 0$ for which $X'_i + X''_i < X'_0 + X''_0$.

1.7.3 What happens if we do ME+PA?

ME+PA correctly estimates i^* much more often than ERM. To demonstrate this, we set $n = 1000$ and conducted an experiment where we drew at random 1000 pairs (X', X'') of independent observations from \mathfrak{X} 's distribution. For each pair, we executed ERM and ME+PA to estimate i^* . We counted how often each value $i \leq n$ was chosen as the estimate by each method. Figures 1.2 and 1.3 demonstrate that in 20% of the cases, ME+PA selects i^* as the estimate, whereas ERM never picks it.

1.7.4 Application of ME+PA

We now illustrate the steps to apply ME+PA to this problem.

First, the hypothesis class is $\mathcal{C} = \{0, \dots, n - 1\}$ and the set \mathcal{X} is \mathbb{R}^n .

Let $X \in \mathcal{X}$ be an observation. The cost function is $R(c, X) = X_c$. The Gibbs distribution is given by $p(c | X) \propto \exp(-1/TX_c)$. For the two given observations X' and X'' , a combined Gibbs distribution is $p(c | X', X'') \propto \exp(-1/T(X'_c + X''_c))$.

We now need to define a value for T . We guess a set of candidate values and, for each of them, we compute the posterior agreement kernel $\kappa(X', X'')$.

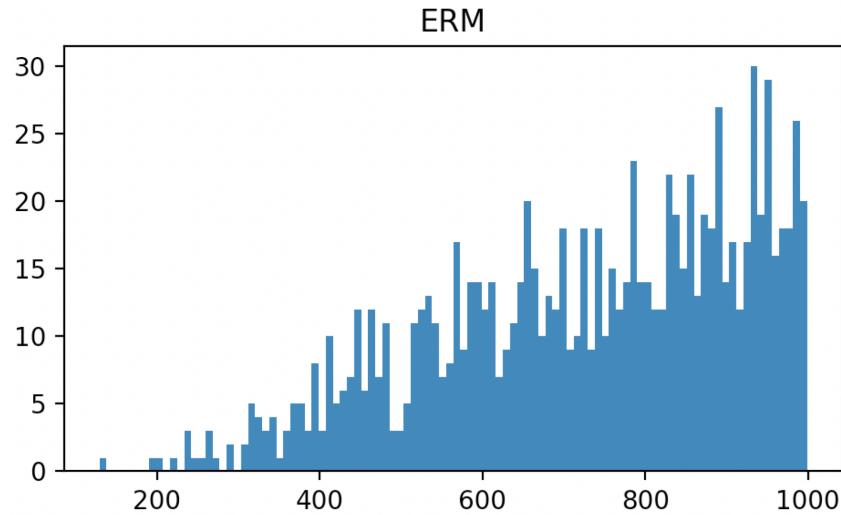


Figure 1.2: Histogram counting how often each estimate in $\{0, \dots, 999\}$ was picked by ERM as the estimate for i^* across 1000 trials. Observe that ERM never picked the correct estimate.

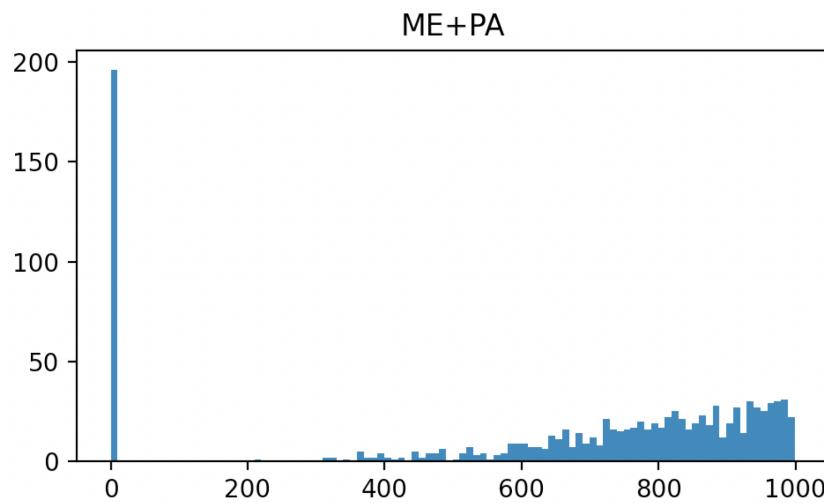


Figure 1.3: Histogram counting how often each estimate in $\{0, \dots, 999\}$ was picked by ME+PA as the estimate for i^* across 1000 trials. In contrast to ERM, ME+PA picks the correct estimate around 20% of the time.

We then pick the value that yielded the maximum posterior agreement kernel. Finally, we just choose the \hat{c} with largest $p(\hat{c} | X', X'')$.

1.8 Conclusion

A popular statistical learning approach for training models is empirical risk minimization (ERM). In this approach, you define a cost function and aim for a model that minimizes the expected cost. In most practical cases, minimizing this expected cost is not possible as we do not have access to the distribution governing that expectation. Instead, we approximate it with an empirical estimate, and then compute the model that minimizes this empirical estimate. This chapter questions this approach in two ways.

ERM trains suboptimal models: First, in many practical cases, the cost function is non-convex and has several local minima, as in the case of the K -means cost function or cost functions used for training neural networks. As a result, empirical risk minimization produces only one model that is a local minimum of the cost function. We argue that such models may be undesirable, as we observed in the K -means algorithm.

Lack of information: ERM produces only single models. These single models may not give a proper overview of the nature of the phenomenon and may be misleading, as we observed in the random array problem.

To address these limitations, we propose a new approach: maximum entropy and posterior agreement. In maximum entropy, rather than producing a single model, we produce a maximum-entropy distribution over models. This distribution is produced in a way that models with low cost get more mass than models with high cost. The distribution offers then a better overview of the landscape of models that fit the phenomenon. These distributions require a value for a hyper-parameter, called the temperature. This parameter can be chosen using standard model selection techniques. In our case, we propose an alternative, called posterior agreement.

Chapter 2

Simulated annealing and methods for exploring Gibbs distributions

Let \mathcal{C} and \mathcal{X} be a hypothesis class and an instance space, respectively. Let also R be a cost function. We are interested in the distribution $p(\cdot | X)$, for a given $X \in \mathcal{X}$. In practice, however, the complexity of R makes often $p(\cdot | X)$ *analytically intractable*: there is no known analytical technique to use $p(\cdot | X)$ to make predictions. In other cases, the size of \mathcal{C} is so large that computing the normalization constant of this distribution *computationally intractable*: there is no known efficient algorithm for computing the value of $p(c | X)$, for every $c \in \mathcal{C}$.

This chapter explores methods for computationally handling and eventually extracting useful models from $p(\cdot | X)$, all of them based in sampling. These methods were originally proposed for optimization, but we use them here to produce samples of $p(\cdot | X)$ that are concentrated around a local minimum of $R(\cdot, X)$. Moreover, in the case of a particular technique called *simulated annealing*, it has been formally demonstrated that when its hyper-parameters are adjusted properly, the samples concentrate around a global minimum of $R(\cdot, X)$.

2.1 Chapter overview

We have seen how the Gibbs distribution concentrates on the global minima of $R(\cdot, X)$ as the temperature converges to zero. This is advantageous over ERM methods, as it gives us more candidate models aside from the empirical risk

minimizer that not only have a lower cost but can also generalize better. Unfortunately, in practice, this Gibbs distribution is intractable, mainly because of its normalization constant.

One way to go around this intractability is via sampling. There are methods for sampling intractable distributions that are effective and popular. We show how to use sampling to explore the Gibbs distribution and choose models from it.

A popular method to sample intractable distributions is by MCMC sampling, which we study in Section 2.3. However, MCMC is impractical for sampling Gibbs distributions. We will see in Section 2.4 that the lower the temperature of the Gibbs distribution, the more time MCMC requires to produce a useful sample, as consecutive samples tend to accumulate in the neighborhood of a local minima of the cost function. We will also see that when the temperature is high MCMC works well. Unfortunately, in this case the Gibbs distribution does not substantially discriminate between good and bad models.

We study simulated annealing (SA) in Section 2.5 and see how it overcomes this limitation of MCMC *by decreasing the temperature during the MCMC sampling*. SA starts by setting the temperature to a very large value, so that the Gibbs distribution is almost uniform and MCMC efficiently produces samples. After some samples, SA alternates between decreasing the current temperature and drawing more samples. We will see, albeit informally, that the alternation between sampling and temperature decreasing avoids that SA gets stuck in the neighborhood of a bad local minimum. It can be formally proven that when the temperature decreasing schedule is done properly, SA converges to the cost function's global minimum.

Simulated annealing offers two advantages over the standard empirical risk minimization (ERM). First, it can converge to a global minimum, or at least to a local minimum of good quality. Second, the sampling nature of SA gives not only one but several models to choose from. In contrast, ERM tends to produce only one model that is just local minima of the cost function. As seen in the random array problem from Section 1.4, this model may not generalize well.

In the case of clustering, the stochastic nature of SA makes it superior to ERM methods like K-means, as we will study in Chapter 3. Moreover, we will see that in several instances, SA can be improved to a more exact and simpler method, called *deterministic annealing* (DA).

2.2 Applications of sampling

Sampling gives us a way to explore and to choose a model from $p(\cdot | X)$, when \mathcal{C} is a vector space, or at least a space that admits a notion of an average: draw a sample from $p(\cdot | X)$ and then propose the average of that sample as a model.

Instead of sampling to compute a single model, one can also use sampling to estimate the probability of an event of interest. Consider the case of clustering. Assume that you have a sample X of points and that you have calculated the Gibbs distribution $p(\cdot | X)$.

Assume now that you want to estimate the probability $\mathbf{P}(x \in C_k | X)$ that a given point x is in a particular cluster C_k . We can estimate this probability as follows:

$$\mathbf{P}(x \in C_k | X) = \sum_{c \in \mathcal{C}} \mathbf{P}(x \in C_k | c)p(c | X). \quad (2.1)$$

However, as we will see in Chapter 3, this distribution is computationally intractable. We can easily approximate this probability if we have a sample c_1, c_2, \dots, c_m from the Gibbs distribution as follows:

$$\mathbf{P}(x \in C_k | X) \approx \sum_{t \leq \ell} \frac{1}{m} \mathbf{1}\{c_t(x) = C_k\}, \quad (2.2)$$

where

$$\mathbf{1}\{c_t(x) = C_k\} \begin{cases} 1 & \text{if } c_t \text{ assigns } x \text{ to cluster } C_k \\ 0 & \text{otherwise.} \end{cases} \quad (2.3)$$

2.3 MCMC sampling

MCMC sampling is one of the most popular methods for sampling from intractable distributions, especially when the source of intractability is the normalization constant. We give here a basic recap of MCMC. For a more thorough overview, please refer to Bishop’s “Pattern recognition and machine learning” or Murphy’s “Machine learning: a probabilistic perspective”.

2.3.1 Markov chain Monte Carlo

Given a distribution p , MCMC sampling is usually done by defining a parameterized family of conditional *transition distributions* $\pi(\cdot | c)$, for an arbitrary $c \in \mathcal{C}$. This distribution gives rise to a graph G whose vertices are \mathcal{C} and where

there is an edge between any two $c, c' \in \mathcal{C}$ iff $\pi(c' | c) > 0$. Your task is then to find a family $\{\pi(\cdot | c)\}_{c \in \mathcal{C}}$ fulfilling the requirements below.

Theorem 6. *Let p be a distribution over \mathcal{C} and let $\{\pi(\cdot | c)\}_{c \in \mathcal{C}}$ be a parameterized family of conditional distributions on \mathcal{C} meeting the following requirements.*

- G_π is connected.
- There is an edge from a vertex to itself in G_π .
- $\pi(c' | c)p(c) = \pi(c | c')p(c')$, for any two $c, c' \in \mathcal{C}$.

Let C_0, C_1, \dots be a sequence of random variables such that, for $1 \leq t$, C_t 's distribution is given by $\pi(\cdot | C_{t-1})$. Then, for $c \in \mathcal{C}$,

$$\lim_{t \rightarrow \infty} \mathbf{P}(C_t = c) = p(c). \quad (2.4)$$

The proof is found in standard probability textbooks and is studied in the course *probabilistic artificial intelligence*.

MCMC sampling exploits this theorem to efficiently produce samples from p as follows. First, pick an arbitrary $C_0 \in \mathcal{C}$. Then for $t = 1, 2, \dots$, sample C_t from $\pi(\cdot | C_{t-1})$. The samples eventually look like samples from p if the conditions below are met.

2.3.2 The Metropolis-Hastings trick

MCMC gives a simpler recipe to sample an intractable distribution p : just come up with a transition distribution fulfilling the three requirements from Theorem 6. The hardest requirement is the last one, called detailed balance, as it demands a good understanding of p in order to fulfil that equation. Fortunately, there is a convenient trick proposed by Metropolis and Hastings that easily yields such a transition distribution.

The trick works for distributions where the intractability comes from the normalization constant, which is often the case in Bayesian methods and in our Gibbs distributions. So let us assume that $p(c) \propto f(c)$, where $f(c)$ is tractable.

Definition 7. *Metropolis-Hastings trick*

1. Define a parameterized family of conditional proposal distributions $\{q(\cdot | c) \mid c \in \mathcal{C}\}$ such that G_q is connected and every vertex in G_q has an edge to itself.
2. Define, for $c, c' \in \mathcal{C}$, the accepting probability

$$A(c', c) := \min \left\{ 1, \frac{q(c | c') p(c')}{q(c' | c) p(c)} \right\} = \min \left\{ 1, \frac{q(c | c') f(c')}{q(c' | c) f(c)} \right\}. \quad (2.5)$$

The transition distribution is then

$$\pi(c' | c) := \begin{cases} q(c' | c) A(c', c) & \text{if } c \neq c' \text{ and} \\ 1 - \sum_{c' \neq c} q(c' | c) A(c', c) & \text{otherwise.} \end{cases} \quad (2.6)$$

Exercise 4. Show that $\pi(\cdot | c)$ fulfils the requirements from Theorem 1.

Definition 8. The algorithm that performs MCMC sampling using the Metropolis-Hastings trick is called Metropolis-Hastings algorithm. In the particular case when $q(c|c') = q(c'|c)$, for $c, c' \in \mathcal{C}$, the algorithm is called Metropolis algorithm.

Exercise 5. In our context, we usually choose proposal distributions $q(\cdot | \cdot)$ that are symmetric. That is, $q(c' | c) = q(c | c')$ for $c, c' \in \mathcal{C}$. Show that, in this case, the Metropolis transition distribution $\pi(\cdot | c)$ for a Gibbs distribution $p(c | X) \propto \exp(-\frac{1}{T}R(c, X))$ simplifies to the following:

- If $c \neq c'$ and $R(c', X) > R(c, X)$, then $\pi(c' | c) = q(c' | c)\rho(c, c')$, where $\rho(c, c') = \exp\left(\frac{1}{T}(R(c, X) - R(c', X))\right)$.
- Otherwise, if $c \neq c'$ and $R(c', X) \leq R(c, X)$, then $\pi(c' | c) = q(c' | c)$.
- Otherwise,

$$\pi(c' | c) = 1 - \sum_{c':R(c',X) \leq R(c,X)} q(c' | c) - \sum_{c':R(c',X) > R(c,X)} q(c' | c)\rho(c, c'). \quad (2.7)$$

For $c \in \mathcal{C}$, we can then draw a sample C from $\pi(\cdot | c)$ in two easy steps:

1. Draw a sample \tilde{C} from $q(\cdot | c)$.
2. If $R(\tilde{C}, X) > R(c, X)$, then draw a sample b from a Bernoulli distribution with mean $\exp\left(\frac{1}{T}(R(c) - R(\tilde{C}))\right)$. If $b = 1$, then $C := \tilde{C}$. Otherwise, $C := c$.

2.3.3 MCMC sampling of Gibbs distributions

We now summarize all our insights so far. Assume given a Gibbs distribution $p(c \mid X) \propto \exp\left(-\frac{1}{T}R(c, X)\right)$. The normalization constant $\sum_c \exp\left(-\frac{1}{T}R(c, X)\right)$, also called *the partition function*, is intractable in most of the interesting cases. We are interested in models that have a high probability mass according to this distribution, as we believe that some of them generalize very well. To retrieve these models without dealing with the intractability of $p(c \mid X)$, we do the following.

1. Define a symmetric proposal distribution $q(\cdot \mid c)$ such that G_1 is connected and every vertex in G_q has an edge to itself.
2. Let C_0 be an arbitrary element in C .
3. For $t = 1, 2, \dots$, draw a sample C_t from $\pi(\cdot \mid C_{t-1})$ as explained above.

By Theorem 6, the samples produced by this algorithm eventually become samples of $p(\cdot \mid X)$.

It is important to emphasize that MCMC and the Metropolis-Hastings trick are not restricted only to Gibbs distributions. These sampling techniques have existed for decades and are still very popular today for dealing with intractable distributions.

2.4 MCMC sampling as a randomized traversal algorithm

Before we continue, we give a toy example that gives us some intuitions of how MCMC sampling works for Gibbs distributions. We argue that MCMC sampling of $p(\cdot \mid X)$ can be understood with a parable of a treasure hunter that explores the hypothesis class \mathcal{C} in a random way, searching for local minima of $R(\cdot, X)$. The sequence of locations visited by the hunter correspond to the sequence of samples drawn via MCMC. When the temperature is low, this hunter lingers in local minima of $R(\cdot, X)$ that are close to the hunter's starting position. When the temperature is high, the hunter roams over \mathcal{C} , showing little preference over the different local minima of $R(\cdot, X)$.

In the next section, we explain why this is problematic in practice: for the Gibbs distribution to be useful for us, it must have a sufficiently low temperature so that its probability mass is concentrated on valuable local minima of $R(\cdot, X)$. However, such a low temperature prevents MCMC to



Figure 2.1

leave bad local minima and makes the success of MCMC very sensitive to the choice of the initial sample, making MCMC impractical for our analyzing Gibbs distributions. Afterwards, we show how simulated annealing solves this issue.

2.4.1 Illustrative example

Let $C = \{0, 1, \dots, 24\}$ and consider the cost function $R(\cdot, X)$ in Figure 2.1, where X is some arbitrary observation. The horizontal axis describes the models in \mathcal{C} . The vertical axis denotes $R(c, X)$, for some $c \in \mathcal{C}$. Notice that the minimizer of $R(\cdot, X)$ is $c^* = 19$.

Figures 2.2a–2.2d depict four Gibbs distributions over \mathcal{C} with different temperatures. Observe that the higher the temperature, the more uniform the Gibbs distribution is. In contrast, the lower the temperature, the more concentrated on c^* the distributions is.

2.4.2 MCMC as a treasure hunter

Let us now do MCMC sampling over one of these Gibbs distributions. We first define a family of proposal $c \in \mathcal{C}$ via the following randomized algorithm. Assume that you have sampled c . To draw the next sample, flip a coin. If heads comes out, then the next sample is c again. Otherwise, flip a coin again. If heads comes out, the next sample is $\min(24, c + 1)$; otherwise, it is



Figure 2.2: Four different Gibbs distributions induced by the cost function from Figure 2.1. Observe that the lower the temperature, the more concentrated the distribution is on the empirical risk minimizer c^* . Furthermore, the number of samples needed to reach c^* from $c_0 = 0$ increases dramatically as the temperature decreases.

$\max(0, c - 1)$. More formally,

$$q(c' | c) = \begin{cases} 1/4 & \text{if } c' = c + 1 \text{ and } c < 24, \\ 1/4 & \text{if } c' = c + 1 \text{ and } c > 0, \\ 1/2 & \text{if } c' = c, \\ 3/4 & \text{if } c' = c = 0, \\ 3/4 & \text{if } c' = c = 24. \end{cases} \quad (2.8)$$

One can then imagine the MCMC sampling of a Gibbs distribution as a treasure hunter that randomly traverses the landscape of the cost function in search for a local minimum of $R(\cdot, X)$. The hunter starts at some location C_0 . Then it uses the proposal distribution $q(\cdot | C_0)$ to choose the location C_1 where it will go next. If $R(C_1, X) \leq R(C_0, X)$, the hunter eagerly moves to that location. Otherwise, the hunter ponders whether to risk forgoing the current location, with the hopes that moving to this higher cost location leads to a location better than the current one. The pondering is simulated by drawing a sample B from a Bernoulli distribution with mean $\exp\left(\frac{1}{T}(R(C_0, X) - R(C_1, X))\right)$. Observe that the higher $R(C_1, X)$ is with respect to $R(C_0, X)$, the less motivated is the hunter to move to that location. This motivation is also affected by the temperature. The lower the temperature, the less motivated is the hunter to move to locations higher than the current one.

2.4.3 The tradeoff induced by the temperature

Equipped with the intuition of the MCMC sampling algorithm as a treasure hunter, we report in Figures 2.2a–2.2d the number of samples that we need to draw via MCMC sampling until c^* is sampled for the first time, when the first sample is $C_0 = 0$. In other words, we report the number of locations that were visited by the hunter before reaching c^* , assuming that the hunter starts from $C_0 = 0$. Observe that the lower the temperature, the more samples are needed to draw before sampling c^* . This is simply because in order to reach c^* from C_0 , the hunter must go through three local minima before reaching c^* . At each local minimum, the hunter must ponder and choose to forgo the current local minimum. The choice to forgo becomes more unlikely the more we lower the temperature. This is why the lower the temperature, the more iterations are needed by the hunter to reach the c^* for the first time.

The tradeoff induced by the temperature is problematic in practice. We want a high temperature, so that the hunter is motivated to explore the landscape of the cost function. However, a high temperature reduces the ability of the Gibbs distribution to discriminate between good and bad models because

the distribution looks more like a uniform distribution. So we also want a low temperature, but this negatively affects the hunter’s motivation, making the hunter resistant to leave bad local minima.

2.5 Simulated annealing

Simulated annealing overcomes this tradeoff with a natural solution: start at a high temperature and then alternate between MCMC sampling and decreasing the temperature. When the temperature is high, the hunter roams through the model space. By gradually decreasing the temperature, we hope that the hunter starts to linger more and more on valuable local minima. When the temperature becomes significantly low, we hope that the hunter stays in that local minima. Algorithm 1 describes simulated annealing.

Simulated annealing offers two advantages. First, the gradual temperature decrease allows the hunter to escape bad local minima, while searching for the cost function’s global minima. So in practice, one can expect that the cost of simulated annealing’s solutions are lower than those from classical ERM schemes like gradient descent, as they do not have procedures to escape bad local minima. Second, the sampling nature of simulated annealing gives us not one but many solutions that are in the “neighborhood” of the empirical risk minimizer.

2.6 Conclusion

Computing Gibbs distributions induced by practical cost functions is intractable. The intractability comes from the normalization constant, as it often requires to sum over an exponentially large hypothesis class or to integrate a transcendental function.

This chapter presents how to use sampling to handle this intractability. A sample of models drawn from the Gibbs distribution can be used to estimate one single model or to estimate the probability of an event of interest.

We then studied a popular sampling method, MCMC, and derived from it the Metropolis(-Hastings) algorithm, a convenient and straightforward procedure for sampling distributions with intractable normalization constants.

We then observed, that these algorithms often get stuck in local minima when sampling from Gibbs distributions. This can be problematic when the cost function underlying the Gibbs distribution is non-convex and has many local minima. To overcome this issue, simulated annealing was proposed.

Algorithm 1 Simulated annealing

```

1: Let
2:    $\epsilon > 0$  be a temperature threshold.
3:    $R(\cdot, X)$  be a cost function,
4:   reduce be a function for decreasing the temperature.
5:    $N \in \mathbb{N}$  be the number of samples drawn before decreasing the temper-
   ature.
6:    $q(\cdot | c)$ , for  $c \in \mathcal{C}$ , a family of proposal distributions.
7:
8: function SIMANN( $\epsilon, R(\cdot, X), \text{reduce}, \{q(\cdot | c) | c \in \mathcal{C}\}$ )
9:    $T \leftarrow \infty$                                  $\triangleright \infty$  is a sufficiently large value.
10:   $c_0^T \leftarrow \$$                              $\triangleright$  Fix the first element in the sample.
11:  while  $T > \epsilon$  do
12:    for  $t = 0 \dots N$  do
13:       $\tilde{c} \leftarrow \text{sample } q(\cdot | c_t^T)$ 
14:      if  $R(\tilde{c}, X) \leq R(c_t^T, X)$  then
15:         $c_{t+1}^T \leftarrow \tilde{c}$ 
16:      else
17:         $b \leftarrow \text{sample Bernoulli } (\exp(\frac{1}{T}(R(c_t^T, X) - R(\tilde{c}, X))))$ 
18:        if  $b = 1$  then
19:           $c_{t+1}^T \leftarrow \tilde{c}$ 
20:        else
21:           $c_{t+1}^T \leftarrow c_t^T$ 
22:        end if
23:      end if
24:    end for
25:     $c_0 \leftarrow c_N^T$                              $\triangleright$  Fix the first element for the next sample
26:     $T \leftarrow \text{reduce}(T)$ 
27:     $c_0^T \leftarrow c_0$ 
28:  end while
29:  return  $(c_t^T)_{T,t}$ 
30: end function

```

Simulated annealing is obtained from the Metropolis algorithm by letting the temperature decrease during the sampling procedure. Simulated annealing has been successful for optimizing non-convex cost functions. By initially sampling at high temperatures and then decreasing the temperature, simulated annealing manages to escape bad local minima and “get stuck” in better quality minima of the cost function.

Chapter 3

Deterministic annealing

3.1 The problem of centroid-based clustering

In this chapter, we consider the problem of centroid-based clustering. In this problem, an observation is a sample $X = \{x_1, \dots, x_N\} \subseteq \mathbb{R}^d$ of points and our objective is to compute a set $\{\theta_1, \dots, \theta_K\} \subseteq \mathbb{R}^d$ of K centroids and a *cluster assignment* function $c : \{1, \dots, N\} \rightarrow \{1, \dots, K\}$, where $K \in \mathbb{N}$ is a hyperparameter. We let the hypothesis class \mathcal{C} be all possible cluster assignment functions with fixed K .

As cost function, we choose the K -means cost function:

$$R(c, \theta, X) := \sum_{i \leq n} \|x_i - \theta_{c(i)}\|^2. \quad (3.1)$$

Note that, in contrast to K -means, we neither require θ_k to be the mean of all points in X assigned to cluster k nor that $c(x)$ is the index of the centroid closest to x .

3.2 Chapter outline

In this chapter, we study how to apply simulated annealing to the problem of clustering, and derive a simplified version called deterministic annealing, proposed by Kenneth Rose [12]. We compare deterministic annealing against a popular ERM approach, K -means, and empirically show that deterministic annealing often provides solutions that generalize better than K -means. Furthermore, combining deterministic annealing with posterior agreement gives a method to estimate the number of clusters while executing deterministic annealing.

Deterministic annealing

Deterministic annealing originates from applying simulated annealing to clustering. We can summarize simulated annealing in this context in the following three steps.

1. Pick a first model c and an initial (high) temperature.
2. Use MCMC to draw a sample from $p(\cdot \mid X)$, the Gibbs distribution induced by $R(\cdot, \cdot, X)$, starting from c as the first model in the sample.
3. Decrease the temperature, let c be the last model in the MCMC sample, and go to Step 2.

We see in Section 3.3 that the Gibbs distribution induced by the cost function $R(\cdot, \cdot, X)$ is actually tractable. This means that we do not need to do MCMC sampling. We can instead directly compute the Gibbs distribution. The three steps become the following procedure, called *deterministic annealing*:

1. Pick an initial set of centroids and an initial (high) temperature.
2. Compute the Gibbs distribution $p(\cdot \mid X)$, for θ arbitrary.
3. Decrease the temperature and go to Step 2.

In Section 3.4, we show that the centroids can then be computed as those that maximize the Gibbs distribution's entropy. However, this maximization is intractable, so we use the EM algorithm to efficiently compute an approximation. We combine all these insights to formulate the algorithm for deterministic annealing in Section 3.5.

In Section 3.6, we provide some experimental results comparing deterministic annealing against K -means.

In Section 3.7, we describe an interesting behavior of deterministic annealing. When the temperature is high, there is only one cluster containing all points and whose centroid is just the sample mean. As the temperature decreases, this cluster decomposes into several clusters that continue to decompose as the temperature keeps decreasing. One could argue that at a sufficiently low temperature, each point becomes its own cluster.

Finally, in Section 3.8, we exploit this behavior and posterior agreement to determine the number of clusters while executing deterministic annealing. This yields an extra advantage of deterministic annealing over other standard clustering methods like K -means which demand to set the number of clusters in advance.

3.3 Tractability of the Gibbs distribution

We assume that our hypothesis class \mathcal{C} is the set of all cluster assignments $c : \{1, \dots, N\} \rightarrow \{1, \dots, K\}$. We define then the family of Gibbs distributions induced by the K -means cost function as all distributions over \mathcal{C} of the form

$$p(\cdot | \theta, X) \propto \exp\left(-\frac{1}{T} R(c, \theta, X)\right). \quad (3.2)$$

Here, $T > 0$ is a hyper-parameter denoting the temperature and $\theta \in \mathbb{R}^{K \times d}$ are parameters denoting the cluster centroids.

The centroids are chosen to be hyper-parameters just for convenience. One could treat them as part of the hypothesis class, but this substantially complicates the analysis.

Theorem 9. *A Gibbs distribution induced by the K -means cost function factorizes as follows:*

$$p(c | \theta, X) = \prod_{i \leq N} p(i \mapsto c(i) | \theta, X), \quad (3.3)$$

where

$$p(i \mapsto k | \theta, X) \propto \exp\left(-\frac{1}{T} \|x_i - \theta_k\|^2\right). \quad (3.4)$$

Proof. The Gibbs distribution is

$$p(c | \theta, X) = \frac{\exp\left(-\frac{1}{T} \sum_{i \leq N} \|x_i - \theta_{c(i)}\|^2\right)}{\sum_{c \in \mathcal{C}} \exp\left(-\frac{1}{T} \sum_{i \leq N} \|x_i - \theta_{c(i)}\|^2\right)}. \quad (3.5)$$

The numerator can be rewritten as follows:

$$\exp\left(-\frac{1}{T} \sum_{i \leq N} \|x_i - \theta_{c(i)}\|^2\right) = \prod_{i \leq N} \exp\left(-\frac{1}{T} \|x_i - \theta_{c(i)}\|^2\right). \quad (3.6)$$

We now apply the combinatorial trick that we studied in the lecture to rewrite the denominator. Unfortunately, I do not see a better way to explain this trick without the diagrams from the lecture.

$$\sum_{c \in \mathcal{C}} \exp\left(-\frac{1}{T} \sum_{i \leq N} \|x_i - \theta_{c(i)}\|^2\right) = \dots = \prod_{i \leq N} \sum_{k \leq K} \exp\left(-\frac{1}{T} \|x_i - \theta_k\|^2\right). \quad (3.7)$$

Putting these results together yields that

$$p(c \mid \theta, X) = \frac{\prod_{i \leq N} \exp\left(-\frac{1}{T} \|x_i - \theta_{c(i)}\|^2\right)}{\prod_{i \leq N} \sum_{k \leq K} \exp\left(-\frac{1}{T} \|x_i - \theta_k\|^2\right)} \quad (3.8)$$

$$= \prod_{i \leq N} \frac{\exp\left(-\frac{1}{T} \|x_i - \theta_{c(i)}\|^2\right)}{\sum_{k \leq K} \exp\left(-\frac{1}{T} \|x_i - \theta_k\|^2\right)} \quad (3.9)$$

$$= \prod_{i \leq N} p(i \mapsto c(i) \mid \theta, X). \quad (3.10)$$

□

Theorem 9 shows that we can compute $p(c \mid \theta, X)$ in just $O(NK)$ -time. You just have to compute $\exp\left(-\frac{1}{T} \|x_i - \theta_k\|^2\right)$, which is $O(1)$ -time, for $i \leq N$ and $k \leq K$. Hence, computing $p(i, c(i) \mid \theta, X)$ takes $O(K)$ -time, for $i \leq N$, yielding a total computing time of $O(NK)$.

3.4 Computing the centroids

The calculations from the previous section do not tell us the values of the centroids. Remember that we follow the maximum-entropy principle, so we compute the centroids that maximize the entropy of $p(\cdot \mid \theta, X)$. In this section, we use C to denote a random cluster assignment whose distribution is $p(\cdot \mid \theta, X)$.

We assume here that $\mathbb{E}_C [R(C, X)] = \mu$, for some fixed $\mu \in \mathbb{R}^+$ and that the temperature hyper-parameter of $p(\cdot \mid \theta, X)$ depends exclusively on μ .

Lemma 1.

$$\frac{\partial H[p(\cdot \mid \theta, X)]}{\partial \theta} = \mathbb{E}_{C \sim p(\cdot \mid \theta, X)} \left[\frac{\partial}{\partial \theta} R(C, \theta, X) \right], \text{ for } k \leq K. \quad (3.11)$$

Proof. We leave the proof details as an exercise and provide just a vague proof. For convenience, all expectations are with respect to a random cluster assignment $C \sim p(\cdot \mid \theta, X)$.

First, show that

$$H[p(\cdot \mid \theta, X)] = \frac{1}{T} \mathbb{E}[R(C, \theta, X)] + \mathbb{E} \left[\log \sum_{c \in \mathcal{C}} \exp\left(-\frac{1}{T} R(c, \theta, X)\right) \right]. \quad (3.12)$$

Recall that $\mathbb{E}[R(C, \theta, X)]$ is fixed to be constant, by construction. Hence, the first term on the right-hand side of the equation above is constant with respect to θ . Moreover, the second term does not depend on the random variable C . We get then that

$$H[p(\cdot | \theta, X)] = \text{const} + \log \sum_{c \in \mathcal{C}} \exp\left(-\frac{1}{T} R(c, \theta, X)\right). \quad (3.13)$$

Take the derivative with respect to θ and demonstrate the desired result. \square

For an event A , we now denote by $\mathbf{1}_A$ the indicator function. That is,

$$\mathbf{1}_A(\omega) \begin{cases} 1 & \text{if } \omega \in A \text{ and} \\ 0 & \text{otherwise.} \end{cases} \quad (3.14)$$

Lemma 2. *Let C denote a random cluster assignment and $C(i)$, for $i \leq N$, the cluster that C assigns to x_i . If $R(\cdot, \cdot, X)$ is the K -means cost function, then*

$$\frac{\partial}{\partial \theta_k} R(C, \theta, X) = 2\theta_k \sum_{i \leq N} \mathbf{1}\{C(i) = k\} - 2 \sum_{i \leq N} x_i \mathbf{1}\{C(i) = k\}. \quad (3.15)$$

Proof. The proof is a straightforward use of vector calculus and is left as an exercise. \square

Theorem 10. *The set θ^* of centroids that maximize the entropy of $p(\cdot | \theta, X)$ must satisfy the following conditions.*

$$\theta_k^* = \frac{\sum_{i \leq N} p(i \mapsto k | \theta^*, X) x_i}{\sum_{i \leq N} p(i \mapsto k | \theta^*, X)}, \text{ for } k \leq K, \quad (3.16)$$

Proof. In order to compute θ^* , our plan is to solve the equations

$$\frac{\partial H[p(\cdot | \theta, X)]}{\partial \theta_k} = 0, \text{ for } k \leq K. \quad (3.17)$$

So we start by applying Lemma 1 on the left-hand side and then plugging Equation 3.15. This results in the following equations:

$$\mathbb{E}_{C \sim p(\cdot | \theta, X)} \left[\theta_k \sum_{i \leq N} \mathbf{1}\{C(i) = k\} - \sum_{i \leq N} x_i \mathbf{1}\{C(i) = k\} \right] = 0, \text{ for } k \leq K. \quad (3.18)$$

Apply now linearity of the expectation and the fact that $\mathbb{E}[\mathbf{1}_A] = \mathbf{P}(A)$ to get that

$$\theta_k \sum_{i \leq N} p(C(i) = k \mid \theta, X) - \sum_{i \leq N} x_i \mathbf{P}(C(i) = k \mid \theta, X) = 0, \text{ for } k \leq K. \quad (3.19)$$

The desired condition follows from this equation and the fact that

$$\mathbf{P}(C(i) = k \mid \theta, X) = p(i \mapsto k \mid \theta, X). \quad (3.20)$$

□

Unfortunately, Equation 3.16 does not give us a closed formula to compute θ^* . We can still attempt to estimate θ^* iteratively, with the following procedure.

1. Set $t \leftarrow 0$ and $\theta(t)$ to an arbitrary value.
2. Set $t \leftarrow t + 1$ and

$$\theta^{(t+1)} \leftarrow \frac{\sum_{i \leq N} p(i \mapsto k \mid \theta^{(t)}, X) x_i}{p(i \mapsto k \mid \theta^{(t)}, X)}. \quad (3.21)$$

This procedure converges to a local maximum of $H[p(\cdot \mid \theta, X)]$.

Exercise 6. Demonstrate that the procedure above also results from applying the EM-algorithm to the following maximization problem:

$$\max_{\theta} \log \sum_{c \in \mathcal{C}} p(X, c \mid \theta), \quad (3.22)$$

where $p(X, c \mid \theta) := p(c \mid \theta, X)p(X)$ and $p(X)$ is the pdf of the phenomenon where X comes from. Use this to demonstrate why the procedure above converges.

3.5 Deterministic annealing

We now collect all insights from this chapter and use them to propose an improved version of simulated annealing, called *deterministic annealing*. Algorithm 2 provides the details. The idea of this procedure is the following.

1. We set a high value for the temperature and start with arbitrary centroids.

2. We alternate between (i) computing the maximum-entropy distribution for a random cluster assignment while fixing the centroids and (ii) computing the centroids that maximize the entropy of that distribution. This alternation is repeated until convergence of the centroids.
3. We reduce the temperature and repeat Step 2. To avoid having repeated centroids, we add a small amount of noise.

We remark some differences between deterministic annealing and simulated annealing:

- There is no MCMC sampling. This is because the Gibbs distribution induced by the K -means cost is tractable. So we can just compute it directly.
- The centroids are treated as parameters of the Gibbs distribution and not as part of the hypothesis class. This is done mainly for convenience. A Gibbs distribution that also treats the centroids as random variables makes the whole procedure intractable.

3.6 Experimental results

Figure 3.1, from Rose [11, 12], compares deterministic annealing with K -means on a particular dataset X which is a sample from a mixture of six Gaussians with different means and different covariance matrices. K -means depends on the initialization, so it was run 25 times. The result shown in the figure was obtained only once, and in more than 80% of the runs, K -means was stuck in a local minimum. In contrast, deterministic annealing is independent of the initialization and finds the global minimum more often than K -means. For both K -means and deterministic annealing, 6 centroids were used.

3.7 Phase transitions

We now study the behavior of the centroids as the temperature decreases. We will see that, at the beginning when the temperature is high, all centroids are equal to the sample mean. As a result one can assume that there is only one cluster. As the temperature decreases, this cluster eventually decomposes into a few clusters, which continue decomposing into more clusters as the temperature keeps decreasing.

Algorithm 2 Deterministic annealing

```

1: Let
2:    $\epsilon > 0$  be a temperature threshold,
3:   reduce( $\cdot$ ) be a function for decreasing the temperature.
4:   close( $\cdot, \cdot$ ) be a function that evaluates if two matrices are close.
5: function DETANN( $\epsilon$ , reduce, close)
6:    $T \leftarrow \infty$                                  $\triangleright \infty$  is a sufficiently large value.
7:    $\theta \leftarrow \$$                                  $\triangleright$  Define arbitrary initial centroids.
8:   while  $T > \epsilon$  do
9:     repeat
10:     $\theta_0 \leftarrow \theta$ 
11:    Compute  $p(i \mapsto k \mid \theta_0, X)$ , for  $i \leq N$  and  $k \leq K$ 
12:     $\theta \leftarrow \frac{\sum_{i \leq N} p(i \mapsto k \mid \theta_0, X) x_i}{\sum_{i \leq N} p(i \mapsto k \mid \theta_0, X)}$ 
13:    until close( $\theta, \theta_0$ )
14:    Add a small amount of noise to each centroid in  $\theta$ .
15:     $T \leftarrow \text{reduce}(T)$ 
16:   end while
17:   return  $\theta, \{p(i \mapsto k \mid \theta, X) \mid i \leq N, k \leq K\}$ 
18: end function

```

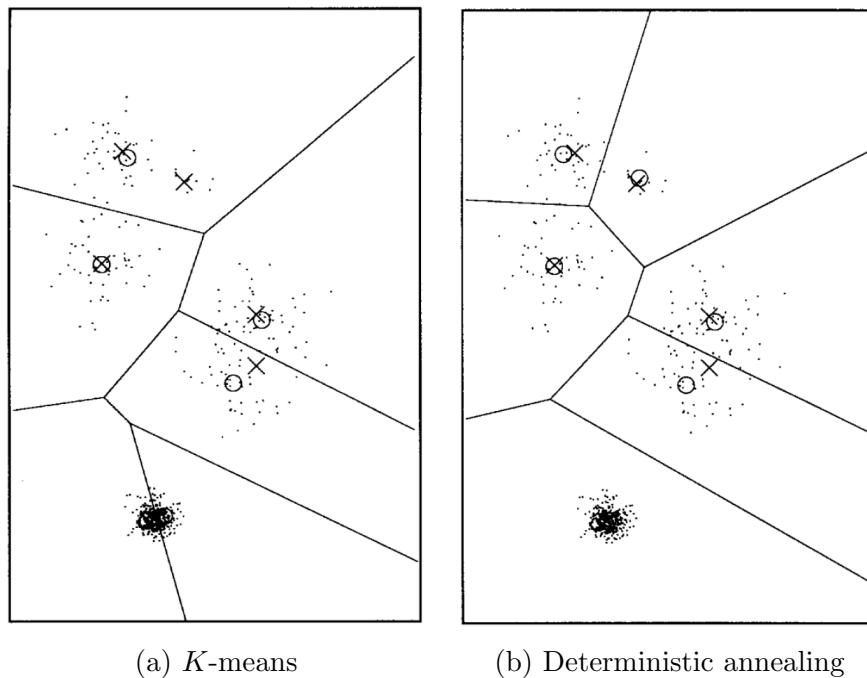


Figure 3.1: Comparison of (a) K -means and (b) deterministic annealing. The \times marks are the means of the Gaussian components. The \bigcirc marks are the centroids computed by the algorithm.

Figure 3.2, from Rose [12], illustrates this behavior. The four subfigures show the centroids (\times marks) at four different points of the execution of deterministic annealing with 9 centroids. Here, X is a sample from a mixture of 9 Gaussians. The subfigures are sorted by decreasing temperature. The curves in the graphic represent points that have all the same probability of belonging to one particular cluster. Observe how the apparent number of centroids change as the temperature decreases. In Figure 3.2a, there seems to be only three centroids, but what actually happens is that some of the 9 centroids are very close to each other. Figure 3.2b shows the centroids after the temperature has decreased. There seems to be now 5 centroids. In Figure 3.2c we have a lower temperature and now 7 apparent centroids. Finally, in Figure 3.2d, all 9 centroids are at different locations and close to the means of the Gaussian mixture.

We now investigate this behavior analytically. We do this through a sequence of lemmas whose proofs are left as exercises.

Theorem 11.

$$\lim_{T \rightarrow \infty} p(i \mapsto k \mid \theta, X) = \frac{1}{K}. \quad (3.23)$$

$$\lim_{T \rightarrow 0} p(i \mapsto k \mid \theta, X) = \begin{cases} 1 & \text{if } \theta_k \text{ is the centroid closest to } x_k \text{ and} \\ 0 & \text{otherwise.} \end{cases} \quad (3.24)$$

Furthermore, when $T \rightarrow \infty$, the set θ^∞ of centroids that maximize the entropy of $H[p(\cdot \mid \theta, X)]$ contains only the sample mean of all points in X .

Proof. This is a consequence of Exercise 1 and Theorem 10. \square

Theorem 12. At $T = 2\lambda$, there is a transition from one to several clusters, where λ is the maximum eigenvalue of the sample covariance matrix of X . This is the matrix $\frac{1}{n}XX^\top$, where X is seen as a matrix whose columns are all points in the observation.

The rest of this section is devoted to prove lemmas that will help prove this theorem.

Lemma 3. For $R(\cdot, \theta, X)$, the K -means cost function, maximizing $H[p(\cdot \mid \theta, X)]$ with respect to θ is equivalent to minimizing

$$F(\theta, X) := -T \sum_{i \leq N} \log \left(\sum_{k \leq K} \exp \left(-\frac{1}{T} \|x_i - \theta_{c(i)}\|^2 \right) \right). \quad (3.25)$$

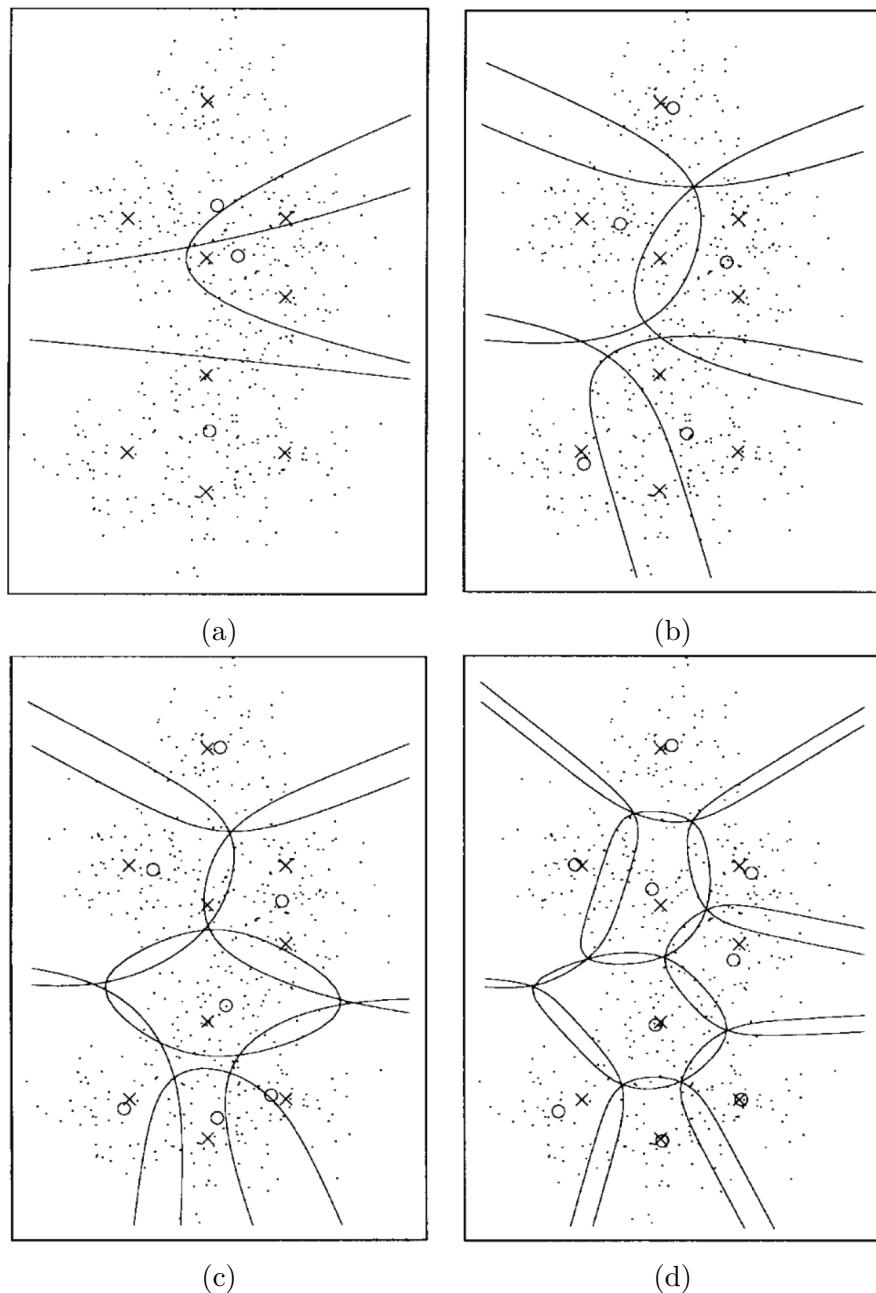


Figure 3.2

Proof. The proof is similar to that of Lemma 1 and involves Theorem 9. \square

$F(\theta, X)$ is called *the Gibbs free energy (induced by $R(\cdot, \theta, X)$)*.

Without loss of generality, we assume that $\theta^\infty = 0$. Observe that when T is sufficiently high, the Hessian H of F evaluated at θ^∞ is positive definite. So new clusters arise when H is no longer positive definite. Our plan is to demonstrate that H stops being positive definite when $T = \frac{1}{2\lambda}$.

The Hessian H can be seen as a block matrix where each block is the Hessian matrix:

$$\frac{\partial F}{\partial \theta_\ell \partial \theta_k} \Big|_{\theta^\infty} = \begin{pmatrix} \frac{\partial F}{\partial \theta_{\ell 1} \partial \theta_{k 1}} \Big|_{\theta^\infty} & \cdots & \frac{\partial F}{\partial \theta_{\ell 1} \partial \theta_{k d}} \Big|_{\theta^\infty} \\ \vdots & \ddots & \vdots \\ \frac{\partial F}{\partial \theta_{\ell d} \partial \theta_{k 1}} \Big|_{\theta^\infty} & \cdots & \frac{\partial F}{\partial \theta_{\ell d} \partial \theta_{k d}} \Big|_{\theta^\infty} \end{pmatrix}. \quad (3.26)$$

From now on, we restrict the domain of $\frac{\partial F}{\partial \theta_\ell \partial \theta_k}$ to a ball \mathcal{B} centered at the origin with a sufficiently small radius so that $\theta_k^\top \theta_\ell$ is insignificant, for every $k, \ell \leq K$. In particular, $\theta_{km} \theta_{\ell m'}$ is insignificant, for every $k, \ell \leq K$ and every $m, m' \leq d$.

Lemma 4. *Inside \mathcal{B} ,*

$$\frac{\partial F}{\partial \theta_k} \approx -\frac{2}{T} \sum_{i \leq N} \frac{A_{ik}}{B_i}, \quad (3.27)$$

where

$$A_{ik} = (x_i - \theta_k) \left(1 + \frac{2}{T} x_i^\top \theta_i \right) \quad \text{and} \quad B_i = \sum_{k' \leq K} \left(1 + \frac{2}{T} x_i^\top \theta_{k'} \right). \quad (3.28)$$

Proof. Start from

$$\frac{\partial F}{\partial \theta_k} = -\frac{2}{T} \sum_{i \leq N} p(i, k | \theta, X) (x_i - \theta_k). \quad (3.29)$$

Observe that, inside \mathcal{B} ,

$$\|x_i - \theta_k\|^2 = \|x_i\|^2 - 2x_i^\top \theta_k + \|\theta_k\|^2 \approx \|x_i\|^2 - 2x_i^\top \theta_k. \quad (3.30)$$

Hence,

$$\exp \left(-\frac{1}{T} \|x_i - \theta_k\|^2 \right) \approx \exp \left(-\frac{1}{T} \|x_i\|^2 \right) \exp \left(\frac{2}{T} x_i^\top \theta_k \right) \quad (3.31)$$

$$\approx \exp \left(-\frac{1}{T} \|x_i\|^2 \right) \left(1 + \frac{2}{T} x_i^\top \theta_k \right). \quad (3.32)$$

$$(3.33)$$

Therefore,

$$p(i, k \mid \theta, X) = \frac{\exp\left(-\frac{1}{T}\|x_i - \theta_k\|^2\right)}{\sum_{k' \leq K} \exp\left(-\frac{1}{T}\|x_i - \theta_{k'}\|^2\right)} \quad (3.34)$$

$$\approx \frac{\exp\left(-\frac{1}{T}\|x_i\|^2\right) \left(1 + \frac{2}{T}x_i^\top \theta_k\right)}{\sum_{k' \leq K} \exp\left(-\frac{1}{T}\|x_i\|^2\right) \left(1 + \frac{2}{T}x_i^\top \theta_{k'}\right)} \quad (3.35)$$

$$= \frac{\left(1 + \frac{2}{T}x_i^\top \theta_k\right)}{\sum_{k' \leq K} \left(1 + \frac{2}{T}x_i^\top \theta_{k'}\right)}. \quad (3.36)$$

The result follows from plugging (3.36) in Equation (3.29). \square

Lemma 5.

$$\frac{\partial F}{\partial \theta_\ell \partial \theta_k} \Big|_{\theta^\infty} = \underbrace{\frac{2N}{K} \mathbb{I}\{\ell = k\} \left(I - \frac{2}{TN} XX^\top\right)}_P + \underbrace{\frac{2}{K^2} XX^\top}_Q. \quad (3.37)$$

Here, $\mathbb{I}\{\ell = k\}$ is the identity matrix when $\ell = k$ and the zero matrix otherwise.

Proof. For $m, m' \leq d$, apply the quotient rule to show that

$$\frac{\partial F}{\partial \theta_{\ell m'} \partial \theta_{km}} \Big|_{\theta^\infty} = -2 \sum_{i \leq N} \left(\frac{1}{B_i} \left| \frac{\partial A_{ikm}}{\partial \theta_{\ell m'}} \right|_{\theta^\infty} \right) + 2 \sum_{i \leq N} \frac{A_{ikm} \Big|_{\theta^\infty}}{\left(B_i \Big|_{\theta^\infty}\right)^2} \left(\frac{\partial B_i}{\partial \theta_{\ell m'}} \Big|_{\theta^\infty} \right). \quad (3.38)$$

Observe now that

$$A_{ik} \Big|_{\theta^\infty} = x_i, \quad B_i \Big|_{\theta^\infty} = K, \quad \text{and} \quad \frac{\partial B_i}{\partial \theta_\ell} \Big|_{\theta^\infty} = x_i. \quad (3.39)$$

Therefore,

$$\frac{\partial F}{\partial \theta_\ell \partial \theta_k} \Big|_{\theta^\infty} = -\frac{2}{K} \left(\frac{\partial}{\partial \theta_\ell} \sum_{i \leq N} A_{ik} \right) \Big|_{\theta^\infty} + \frac{2}{K^2} \sum_{i \leq N} x_i x_i^\top. \quad (3.40)$$

Show now that

$$\sum_{i \leq N} A_{ik} = \left(\frac{2}{T} XX^\top - NI \right) \theta_k. \quad (3.41)$$

This implies that

$$\left(\frac{\partial}{\partial \theta_\ell} \sum_{i \leq N} A_{ik} \right) \Big|_{\theta^\infty} = NI \mathbb{I}\{\ell = k\} \left(\frac{2}{TN} XX^\top - I \right). \quad (3.42)$$

Plugging this into Equation 3.40 yields the desired result. \square

Observe that, with high probability, Q is positive semi-definite.

Lemma 6. *Let Q be a positive semi-definite matrix. Let H be a block matrix where the block (i, j) is given by*

$$H[i, j] = \mathbb{I}\{\ell = k\} P + Q. \quad (3.43)$$

Then H is positive-definite iff P is.

Proof. See Rose's doctoral thesis [11]. \square

We now prove Theorem 12. Recall that the temperature at which θ^∞ stops being the unique minimum is the temperature at which the Hessian matrix H evaluated at θ^∞ stops begin positive definite. By Lemma 6, this occurs when

$$P = I - \frac{2}{TN} XX^\top \quad (3.44)$$

is no longer positive definite, which occurs when

$$0 = \left| I - \frac{2}{TN} XX^\top \right| = \left| \frac{T}{2} I - \frac{1}{N} XX^\top \right|. \quad (3.45)$$

That is, when $T/2$ is an eigenvalue of $\frac{1}{N} XX^\top$. Since R starts at a very large value and is gradually decreased, we conclude that the temperature at which more clusters appear is at $T = 2\lambda$ where λ is the largest eigenvalue of $\frac{1}{N} XX^\top$. This concludes the proof.

3.8 Deterministic annealing and posterior agreement

There is still an open question regarding deterministic annealing. What is the number of clusters that we should use? More generally, at which temperature we should stop the execution of deterministic annealing? The posterior agreement principle recommends to have two observations X' and X'' and compute the temperature that maximizes the posterior agreement kernel

$$\kappa(X', X'') = \sum_c p(c | X') p(c | X''). \quad (3.46)$$

Observe that for the two probabilities to make sense, we must have $X' = \{x'_1, \dots, x'_M\}$ and $X'' = \{x''_1, \dots, x''_N\}$ of equal size. We assume from now on

that this is the case. Furthermore, we assume that x'_i and x''_i , for $i \leq N$, are two observations of one same object. The more general scenario is studied by Buhmann [1].

An analytical method to maximize $\kappa(X', X'')$ with respect to the temperature is unknown. Moreover, the posterior agreement kernel seems to be computationally intractable, as it is a sum over K^N objects. Fortunately, the factorization of the Gibbs distribution makes it computationally tractable.

Exercise 7. Show that

$$\sum_c p(c | X') p(c | X'') = \prod_{i \leq N} \sum_{k \leq K} p(i \mapsto k | X') p(i \mapsto k | X''). \quad (3.47)$$

Show also that computing this takes $O(NK)$ -time.

Our plan is then to select a set of candidate temperatures and pick the one with the highest $\kappa(X', X'')$ as the stopping temperature. We can implement this plan during the execution of deterministic annealing. Observe that the while loop of Algorithm 2 gives as a sequence of decreasing temperatures. Therefore, we can compute $\kappa(X', X'')$ at each iteration with the current temperature. We can keep track of these kernel values. The value of $\kappa(X', X'')$ has been observed empirically to be low when the temperature is high, then it increases as the temperature is lowered, and eventually it starts decreasing. Hence, we can keep track of how the kernel values change as deterministic annealing decreases the temperature. After observing, for a certain amount of iterations, that $\kappa(X', X'')$ is no longer increasing and starts decreasing, we can stop the annealing and choose as stopping temperature the one that yielded the largest $\kappa(X', X'')$.

Chapter 4

Laplace's method

In this chapter, we study an analytical tool to handle intractable Gibbs distributions: *Laplace's method*. We use this method to analytically approximate the normalization constant of certain Gibbs distributions, which is often the main source of intractability. The essence of the method is to rewrite the normalization constant as an integral of the form

$$\int \exp\left(-\frac{1}{T}f(y)\right) dy, \quad (4.1)$$

where f is a convex function (or at least something similar to it). Afterwards, we exploit the fact that for sufficiently low T , this integral is approximated by $\exp(-\frac{1}{T}f(y_0))$, where y_0 is the minimum of f .

We illustrate this method in two scenarios. The first one is just a toy example. The second one shows how to cluster documents represented as vectors using the bag of words model. In this case, we wish then to cluster documents not by their Euclidean distance, but by how much the words occurring in the respective documents overlap. For two documents x and y represented as vectors, this overlap can be quantified by their *cosine similarity*: $x^\top y / \|x\| \|y\|$.

4.1 Laplace's method

Theorem 13. Let $f : \mathbb{R}^D \rightarrow \mathbb{R}$ be a twice continuously differentiable function. If f has a unique minimum y_0 and $H := \frac{\partial^2 f}{\partial y^2}|_{y_0}$ (the Hessian of f evaluated at y_0) is positive definite then

$$\lim_{M \rightarrow \infty} \frac{\int \exp(-Mf(y)) dy}{\exp(-Mf(y_0)) \left| \frac{MH}{2\pi} \right|^{-1/2}} = 1. \quad (4.2)$$

Intuitively, Laplace's method suggests to approximate $\int \exp(-Mf(y)) dy$ with $\exp(-Mf(y_0))$ if M is sufficiently large.

Figure 4.1 illustrates the behavior of $\exp(-Mf(y))$, with $f(y) = -1 + y^2 + \sin(5y)$. Observe that although $\int \exp(-Mf(y)) dy$ is analytically tractable, as $M \rightarrow \infty$, the area under the curve is mostly around the minimum of f .

The proof of Theorem 13 relies on showing that

$$\lim_{M \rightarrow \infty} \frac{\int \exp(-Mf(y)) dy}{\int \exp(-MT_f^2(y)) dy} = 1, \quad (4.3)$$

where $T_f^2(y)$ is the 2nd order Taylor approximation of f at y_0 evaluated at y :

$$T_f^2(y) = f(y_0) + (y - y_0)^\top \frac{\partial f}{\partial y}|_{y_0} + \frac{1}{2}(y - y_0)^\top H(y - y_0) \quad (4.4)$$

$$= f(y_0) + \frac{1}{2}(y - y_0)^\top H(y - y_0). \quad (4.5)$$

4.2 Illustrative example

Let $\mathcal{C} = \{-1, +1\}^N$, for some $N \in \mathbb{N}$. We now illustrate how to exploit Laplace's method to approximate the intractable Gibbs distribution

$$p(c) := \frac{1}{Z} \exp\left(-\frac{1}{T}R(c)\right), \quad (4.6)$$

where

$$Z := \sum_{c \in \mathcal{C}} \exp\left(-\frac{1}{T}R(c)\right) \quad \text{and} \quad R(c) := -c^\top c. \quad (4.7)$$

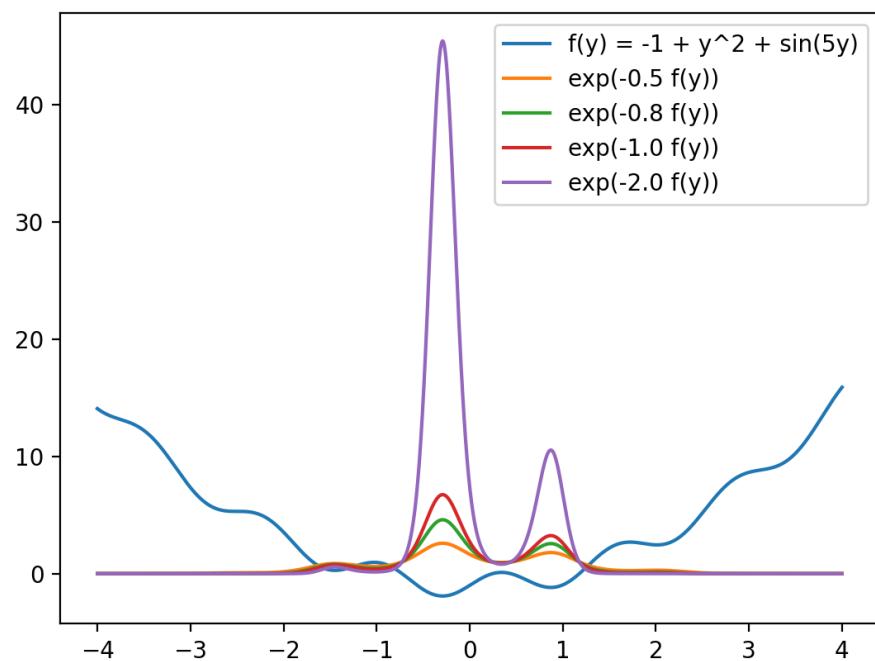


Figure 4.1

The normalization constant is a sum over 2^N elements and becomes intractable for large N . The following four exercises give a recipe for approximating this constant using Laplace's method.

Exercise 8. [Complete the square] Show that

$$\exp\left(\frac{1}{T}c^\top c\right) = (\pi T)^{-d/2} \int \exp\left(-\frac{1}{T}y^\top y\right) \prod_{i \leq N} \exp\left(\frac{2}{T}y_i c_i\right) dy. \quad (4.8)$$

Hint: Start here

$$(\pi T)^{d/2} = \int \exp\left(-\frac{1}{T}\|y - c\|^2\right) dy. \quad (4.9)$$

Exercise 9. [Rewrite normalization constant] Complete the following derivation

$$Z = \sum_c \exp\left(\frac{1}{T}c^\top c\right) \quad (4.10)$$

$$= (\pi T)^{-d/2} \int \exp\left(-\frac{1}{T}y^\top y\right) \sum_c \prod_{i \leq N} \exp\left(\frac{2}{T}y_i c_i\right) dy \quad (4.11)$$

$$= \dots \quad (4.12)$$

$$= (\pi T)^{-d/2} \int \exp\left(-\frac{1}{T}\left(y^\top y - T \sum_{i \leq N} \log\left(2 \cosh\left(\frac{2}{T}y_i\right)\right)\right)\right) dy. \quad (4.13)$$

Hint: Apply the combinatorial trick from Theorem 9. Recall also that

$$2 \cosh(x) = \exp(x) + \exp(-x). \quad (4.14)$$

Exercise 10. [Apply Laplace's method] Show that, when $T > 2$, the conditions to apply Laplace's method hold. As a result, the integral in Equation 4.13 could be approximated to 1. Is this a good estimation?

It is still possible to apply Laplace's method when $T \leq 2$, and approximate the integral above to 1. However, doing this formally requires a more thorough understanding of the proof of Laplace's method. An appendix describing this approximation is left for a future version of this script.

4.3 Using Laplace's method to approximate intractable normalization constants

The section before illustrates the following analytical strategy to estimate normalization constants:

1. **Rewrite cost function** Rewrite your cost function as the squared norm of some vector, so that

$$\exp\left(-\frac{1}{T}R(c, X)\right) = \text{const} \exp(g(c)^\top g(c)). \quad (4.15)$$

2. **Complete the square** Recall that

$$\pi^{d/2} = \int \exp(-\|y - g(c)\|^2) dy. \quad (4.16)$$

Use this to derive that

$$\exp(g(c)^\top g(c)) = \pi^{-d/2} \int \exp(-y^\top y + 2y^\top g(c)) dy. \quad (4.17)$$

3. **Rewrite normalization constant** Show that the normalization constant can be expressed as follows

$$Z = \sum_c \exp\left(-\frac{1}{T}R(c, X)\right) \quad (4.18)$$

$$= \dots \quad (4.19)$$

$$= \text{const} \int \exp(-Mf(y)) dy. \quad (4.20)$$

4. **Apply Laplace's method** Verify that the conditions for Laplace's method hold and apply it to estimate Z with $\text{const} \exp(-Mf(y_0))$, where y_0 is the minimum of f .

4.4 Least-angle clustering

We now apply Laplace's method for a more interesting problem. Assume given a set $X = \{x_1, x_2, \dots, x_N\} \subseteq \mathbb{R}^d$ of unit vectors. We wish to cluster them in K clusters so that any pair of vectors are assigned to the same cluster if they point in similar directions.

We model a cluster assignment as a matrix $M \in \{0, 1\}^{N \times K}$, where

$$M_{ik} \begin{cases} 1 & \text{if } x_i \text{ is assigned to cluster } k \\ 0 & \text{otherwise.} \end{cases} \quad (4.21)$$

Therefore, it is required that $\sum_{k \leq K} M_{ik} = 1$, for any $i \leq N$. Observe that for $i, j \leq N$ and $k \leq K$, $M_{ik}M_{jk} = 1$ iff x_i and x_j are assigned to the same cluster.

Our goal is to use maximum entropy to train a cluster assignment, using the following cost function

$$R(M, X) = \frac{1}{N} \sum_{i < j \leq N} \sum_{k \leq K} M_{ik}M_{jk}x_i^\top x_j. \quad (4.22)$$

Observe that the normalization constant of the Gibbs distribution induced by this cost function is a sum over K^N elements, making it intractable. The next four sections illustrate how to manipulate the normalization constant so we can apply Laplace's method.

4.5 Rewrite cost function

Exercise 11. *Show that*

$$R(M, X) = -\frac{1}{2N} \sum_{k \leq K} \left\| \sum_{i \leq N} M_{ik}x_i \right\|^2 + \frac{1}{2}. \quad (4.23)$$

Proof.

$$R(M, X) = -2 \left(\frac{1}{2N} \sum_{i < j \leq N} \sum_{k \leq K} M_{ik} M_{jk} x_i^\top x_j \right) \quad (4.24)$$

$$= -\frac{1}{2N} \sum_{i < j \leq N} \sum_{k \leq K} M_{ik} M_{jk} x_i^\top x_j - \frac{1}{2N} \sum_{i < j \leq N} \sum_{k \leq K} M_{ik} M_{jk} x_i^\top x_j \quad (4.25)$$

$$= -\frac{1}{2N} \sum_{i < j \leq N} \sum_{k \leq K} M_{ik} M_{jk} x_i^\top x_j - \frac{1}{2N} \sum_{j < i \leq N} \sum_{k \leq K} M_{ik} M_{jk} x_i^\top x_j \quad (4.26)$$

$$= -\frac{1}{2N} \sum_{i < j \leq N} \sum_{k \leq K} M_{ik} M_{jk} x_i^\top x_j - \frac{1}{2N} \sum_{j < i \leq N} \sum_{k \leq K} M_{ik} M_{jk} x_i^\top x_j + \quad (4.27)$$

$$- \frac{1}{2N} \sum_i \sum_{k \leq K} M_{ik} M_{ik} x_i^\top x_i + \frac{1}{2N} \sum_i \sum_{k \leq K} M_{ik} M_{ik} x_i^\top x_i \quad (4.28)$$

$$= -\frac{1}{2N} \sum_{k \leq K} \sum_{i, j \leq N} M_{ik} M_{jk} x_i^\top x_j + \frac{1}{2N} \sum_i \sum_{k \leq K} M_{ik} M_{ik} x_i^\top x_i \quad (4.29)$$

$$= -\frac{1}{2N} \sum_{k \leq K} \sum_{i, j \leq N} M_{ik} M_{jk} x_i^\top x_j + \frac{1}{2N} N. \quad (4.30)$$

$$= -\frac{1}{2N} \sum_{k \leq K} \sum_{i, j \leq N} M_{ik} M_{jk} x_i^\top x_j + \frac{1}{2}. \quad (4.31)$$

$$= -\frac{1}{2N} \sum_{k \leq K} \left\| \sum_{i \leq N} M_{ik} x_i \right\|^2 + \frac{1}{2}. \quad (4.32)$$

□

This result only implies that

$$\exp \left(-\frac{1}{T} R(M, X) \right) = e^{-\frac{1}{2T}} \prod_{k \leq K} \exp \left(-\frac{1}{2N} \left\| \sum_{i \leq N} M_{ik} x_i \right\|^2 \right). \quad (4.33)$$

This is not an expression of the form $\text{const} \exp(g(c)^\top g(c))$, as recommended in Section 4.3. However, we show next how to adapt the cost function so that we can still follow all the other steps from Section 4.3.

4.6 Complete the square

Recall the Gibbs distribution

$$p(c | X) = \frac{1}{Z_X} \exp\left(-\frac{1}{T} R(M, X)\right),$$

where Z_X is the partition function.

Exercise 12. *Prove that*

$$Z_X = \sum_{M \in \{0,1\}^{N \times K}} e^{-\frac{1}{2T}} \prod_{k \leq K} \exp\left(\frac{1}{2NT} \left\| \sum_{i \leq N} M_{ik} x_i \right\|^2\right). \quad (4.34)$$

Proof.

$$\begin{aligned} Z_X &= \sum_M \exp\left(-\frac{1}{T} R(M, X)\right) \\ &= \sum_M \exp\left(\frac{1}{2NT} \sum_{k \leq K} \left\| \sum_{i \leq N} M_{ik} x_i \right\|^2 - \frac{1}{2T}\right) \\ &= e^{-\frac{1}{2T}} \sum_M \prod_{k \leq K} \exp\left(\frac{1}{2NT} \left\| \sum_{i \leq N} M_{ik} x_i \right\|^2\right). \end{aligned}$$

□

Observe that the partition function requires to compute a sum over $2^{N \times K}$ elements and it is not clear how to do that sum analytically or in a computationally efficient way. Hence, computing the Gibbs distribution takes exponential time. This makes ME + PA very inefficient as these methods rely on computing the Gibbs distribution.

We now use the technique of completing the square.

Exercise 13. *For $k \leq K$, define a dummy variable y_k . Show that*

$$\exp\left(\frac{1}{2NT} \left\| \sum_{i \leq N} M_{ik} x_i \right\|^2\right) \quad (4.35)$$

$$= \left(\frac{2\pi T}{N}\right)^{\frac{D}{2}} \int \exp\left(-\frac{N}{2T} \|y_k\|^2 + \frac{1}{T} \sum_{i \leq N} M_{ik} x_i^\top y_k\right) dy_k \quad (4.36)$$

Hint: Start from here

$$\int \exp \left(-\frac{N}{2T} \left\| y_k - \frac{1}{N} \sum_{i \leq N} M_{ik} x_i \right\|^2 \right) dy_k.$$

Observe that this is the normalization constant for a multivariate Gaussian distribution.

Proof. Using the fact that this is the normalization constant of a Gaussian, we have that

$$\left(\frac{2\pi T}{N} \right)^{-\frac{D}{2}} \quad (4.37)$$

$$= \int \exp \left(-\frac{N}{2T} \left\| y_k - \frac{1}{N} \sum_{i \leq N} M_{ik} x_i \right\|^2 \right) dy_k \quad (4.38)$$

$$= \int \exp \left(-\frac{N}{2T} \left[\|y_k\|^2 - \frac{2}{N} y_k^\top \sum_{i \leq N} M_{ik} x_i + \frac{1}{N^2} \left\| \sum_{i \leq N} M_{ik} x_i \right\|^2 \right] \right) dy_k \quad (4.39)$$

$$= \exp \left(-\frac{N}{2T} \frac{1}{N^2} \left\| \sum_{i \leq N} M_{ik} x_i \right\|^2 \right) \int \exp \left(-\frac{N}{2T} \left[\|y_k\|^2 - \frac{2}{N} y_k^\top \sum_{i \leq N} M_{ik} x_i \right] \right) dy_k \quad (4.40)$$

$$= \exp \left(-\frac{1}{2NT} \left\| \sum_{i \leq N} M_{ik} x_i \right\|^2 \right) \int \exp \left(-\frac{N}{2T} \|y_k\|^2 + \frac{1}{T} y_k^\top \sum_{i \leq N} M_{ik} x_i \right) dy_k. \quad (4.41)$$

We can rearrange the resulting equation to obtain Equation 4.36. \square

4.7 Rewrite the normalization constant

Exercise 14. *Using Equations 4.34 and 4.36, show that*

$$Z_X = \alpha \int \exp \left(-\frac{N}{2T} \sum_{k \leq K} \|y_k\|^2 + \sum_{i \leq N} \log \sum_{k \leq K} \exp \left(\frac{1}{T} x_i^\top y_k \right) \right) d\mathbf{y}, \quad (4.42)$$

where

$$\alpha := e^{-\frac{1}{2T}} \left(\frac{2\pi T}{N} \right)^{\frac{KD}{2}}, \text{ and} \quad (4.43)$$

$$\int g(y_1, \dots, y_K) d\mathbf{y} := \int \int \dots \int g(y_1, \dots, y_K) dy_1 dy_2 \dots dy_K. \quad (4.44)$$

Hint: Prove and use the following auxiliary results.

$$\prod_{k \leq K} \int f_k(y_k) dy_k = \int \prod f_k(y_k) d\mathbf{y}.$$

$$\sum_{M \in \{0,1\}^{N \times K}} \prod_{i \leq N} \exp \left(\frac{1}{T} \sum_{k \leq K} M_{ik} x_i^\top y_k \right) = \prod_{i \leq N} \sum_{k \leq K} \exp \left(\frac{1}{T} x_i^\top y_k \right).$$

Proof. Using Equation 4.36, Equation 4.34 becomes

$$Z_X = e^{-\frac{1}{2T}} \sum_{M \in \{0,1\}^{N \times K}} \prod_{k \leq K} \exp \left(\frac{1}{2NT} \left\| \sum_{i \leq N} M_{ik} x_i \right\|^2 \right) \quad (4.45)$$

$$= e^{-\frac{1}{2T}} \sum_{M \in \{0,1\}^{N \times K}} \prod_{k \leq K} \left(\frac{2\pi T}{N} \right)^{\frac{D}{2}} \int \exp \left(-\frac{N}{2T} \|y_k\|^2 + \frac{1}{T} \sum_{i \leq N} M_{ik} x_i^\top y_k \right) dy_k \quad (4.46)$$

$$= e^{-\frac{1}{2T}} \left(\frac{2\pi T}{N} \right)^{\frac{KD}{2}} \sum_{M \in \{0,1\}^{N \times K}} \prod_{k \leq K} \int \exp \left(-\frac{N}{2T} \|y_k\|^2 + \frac{1}{T} \sum_{i \leq N} M_{ik} x_i^\top y_k \right) dy_k. \quad (4.47)$$

Observe now the following property.

$$\begin{aligned} \prod_{k \leq K} \int f_k(y_k) dy_k &= \left(\int f_1(y_1) dy_1 \right) \left(\int f_2(y_2) dy_2 \right) \dots \left(\int f_K(y_K) dy_K \right) \\ &= \int f_1(y_1) f_2(y_2) \dots f_K(y_K) d\mathbf{y} \\ &= \int \prod f_k(y_k) d\mathbf{y}. \end{aligned}$$

Applying this property above gives us the following.

$$Z_X \tag{4.48}$$

$$= \alpha \sum_M \int \prod_{k \leq K} \exp \left(-\frac{N}{2T} \|y_k\|^2 + \frac{1}{T} \sum_{i \leq N} M_{ik} x_i^\top y_k \right) d\mathbf{y} \tag{4.49}$$

$$= \alpha \sum_M \int \exp \left(-\frac{N}{2T} \sum_{k \leq K} \|y_k\|^2 + \frac{1}{T} \sum_{k \leq K} \sum_{i \leq N} M_{ik} x_i^\top y_k \right) d\mathbf{y} \tag{4.50}$$

$$= \alpha \sum_M \int \exp \left(-\frac{N}{2T} \sum_{k \leq K} \|y_k\|^2 \right) \exp \left(\frac{1}{T} \sum_{i \leq N} \sum_{k \leq K} M_{ik} x_i^\top y_k \right) d\mathbf{y} \tag{4.51}$$

$$= \alpha \sum_M \int \exp \left(-\frac{N}{2T} \sum_{k \leq K} \|y_k\|^2 \right) \prod_{i \leq N} \exp \left(\frac{1}{T} \sum_{k \leq K} M_{ik} x_i^\top y_k \right) d\mathbf{y} \tag{4.52}$$

$$= \alpha \int \sum_M \left[\exp \left(-\frac{N}{2T} \sum_{k \leq K} \|y_k\|^2 \right) \prod_{i \leq N} \exp \left(\frac{1}{T} \sum_{k \leq K} M_{ik} x_i^\top y_k \right) \right] d\mathbf{y} \tag{4.53}$$

$$= \alpha \int \exp \left(-\frac{N}{2T} \sum_{k \leq K} \|y_k\|^2 \right) \sum_M \prod_{i \leq N} \exp \left(\frac{1}{T} \sum_{k \leq K} M_{ik} x_i^\top y_k \right) d\mathbf{y}. \tag{4.54}$$

Given $M \in \{0, 1\}^{N \times K}$, we now define an auxiliary function $c_M : \{1, \dots, N\} \rightarrow \{1, \dots, K\}$. For $i \leq N$, $c_M(i) = k$ iff $M_{ik} = 1$. Observe that c_M is well defined as $M \in \{0, 1\}^{N \times K}$ and $\sum_k M_{ik} = 1$. As a result, $\sum_k M_{ik} x_i^\top y_k = x_i^\top y_{c(i)}$. Observe now that

$$\sum_{M \in \{0, 1\}^{N \times K}} \prod_{i \leq N} \exp \left(\frac{1}{T} \sum_{k \leq K} M_{ik} x_i^\top y_k \right) \tag{4.55}$$

$$= \sum_{c: \{1, \dots, N\} \rightarrow \{1, \dots, K\}} \prod_{i \leq N} \exp \left(\frac{1}{T} x_i^\top y_{c(i)} \right) \tag{4.56}$$

$$= \prod_{i \leq N} \sum_{k \leq K} \exp \left(\frac{1}{T} x_i^\top y_k \right). \tag{4.57}$$

If we apply this to the equation above, we get

$$Z_X \quad (4.58)$$

$$= \alpha \int \exp \left(-\frac{N}{2T} \sum_{k \leq K} \|y_k\|^2 \right) \prod_{i \leq N} \sum_{k \leq K} \exp \left(\frac{1}{T} x_i^\top y_k \right) d\mathbf{y} \quad (4.59)$$

$$= \alpha \int \exp \left(-\frac{N}{2T} \sum_{k \leq K} \|y_k\|^2 \right) \exp \left(\log \left(\prod_{i \leq N} \sum_{k \leq K} \exp \left(\frac{1}{T} x_i^\top y_k \right) \right) \right) d\mathbf{y} \quad (4.60)$$

$$= \alpha \int \exp \left(-\frac{N}{2T} \sum_{k \leq K} \|y_k\|^2 + \sum_{i \leq N} \log \sum_{k \leq K} \exp \left(\frac{1}{T} x_i^\top y_k \right) \right) d\mathbf{y} \quad (4.61)$$

□

4.8 Apply Laplace's method

Let \mathbf{y} be the vector of real variables resulting from the concatenation of y_1, y_2, \dots, y_K . Equation 4.61 can then be rewritten as

$$Z_X = e^{-\frac{1}{2T}} \left(\frac{2\pi T}{N} \right)^{\frac{KD}{2}} \int \exp \left(-\frac{N}{T} f_X(\mathbf{y}) \right) d\mathbf{y}, \quad (4.62)$$

where

$$f_X(\mathbf{y}) = \frac{1}{2} \sum_{k \leq K} \|y_k\|^2 - \frac{T}{N} \sum_{i \leq N} \log \sum_{k \leq K} \exp \left(\frac{1}{T} x_i^\top y_k \right). \quad (4.63)$$

Exercise 15. Show how to compute \mathbf{y}^* .

Proof. We obtain \mathbf{y}^* by solving

$$\frac{\partial f_X}{\partial \mathbf{y}} = 0 \quad \Leftrightarrow \quad \frac{\partial f_X}{\partial y_k} = 0, \text{ for all } k \leq K. \quad (4.64)$$

$$\begin{aligned} \frac{\partial f_X}{\partial y_k} &= \frac{1}{2} \frac{\partial f_X}{\partial y_k} \sum_{k' \leq K} \|y_{k'}\|^2 - \frac{T}{N} \sum_{i \leq N} \frac{\partial f_X}{\partial y_k} \log \sum_{k' \leq K} \exp \left(\frac{1}{T} x_i^\top y_{k'} \right) \\ &= y_k - \frac{T}{N} \sum_{i \leq N} \frac{\frac{1}{T} x_i^\top \exp \left(\frac{1}{T} x_i^\top y_k \right)}{\sum_{k' \leq K} \exp \left(\frac{1}{T} x_i^\top y_{k'} \right)} \end{aligned}$$

Hence, for $k \leq K$, $\frac{\partial f_X}{\partial y_k} = 0$ implies that

$$y_k = \sum_{i \leq N} \frac{x_i \exp\left(\frac{1}{T} x_i^\top y_k\right)}{N \sum_{k' \leq K} \exp\left(\frac{1}{T} x_i^\top y_{k'}\right)}.$$

This transcendental equation, can be solved iteratively, using an EM-like procedure. \square

Exercise 16. Show how to compute $\left.\frac{\partial^2 f_X}{\partial y^2}\right|_{y^*}$.

Proof. Recall that for a function $g : \mathbb{R}^d \rightarrow \mathbb{R}$, the Hessian is defined as a matrix whose (i, j) entry is $\frac{\partial^2 g}{\partial y_i \partial y_j}$. Hence, $\frac{\partial^2 f_X}{\partial y^2}$ is the following block matrix:

$$\frac{\partial^2 f_X}{\partial y^2} = \begin{pmatrix} \frac{\partial^2 f_X}{\partial y_1 \partial y_1} & \cdots & & \frac{\partial^2 f_X}{\partial y_1 \partial y_K} \\ \vdots & \ddots & \ddots & \vdots \\ & & \frac{\partial^2 f_X}{\partial y_k \partial y_\ell} & \\ & \ddots & \ddots & \\ \frac{\partial^2 f_X}{\partial y_K \partial y_1} & \cdots & & \frac{\partial^2 f_X}{\partial y_K \partial y_K} \end{pmatrix}.$$

Please be careful to observe that $\frac{\partial^2 f_X}{\partial y_k \partial y_\ell}$ is a *matrix*! In particular, if $k = \ell$, then one can show that

$$\frac{\partial^2 f_X}{\partial y_k^2} = I - \frac{1}{TN} \sum_{i \leq N} x_i x_i^\top P_{ik} (1 - P_{ik}),$$

where

$$P_{ik} = \frac{\exp\left(\frac{1}{T} x_i^\top y_k\right)}{\sum_k \exp\left(\frac{1}{T} x_i^\top y_k\right)}.$$

When $k \neq \ell$, one can show that

$$\frac{\partial^2 f_X}{\partial y_k \partial y_\ell} = \frac{1}{TN} \sum_{i \leq N} x_i x_i^\top P_{ik} P_{i\ell}.$$

\square

It can be shown that $\left.\frac{\partial^2 f_X}{\partial y^2}\right|_{y^*}$ is positive definite, when N is sufficiently large and T is not too low¹. One can then apply Laplace's method to do the following approximation:

$$Z_X = e^{-\frac{1}{2T}} \left(\frac{2\pi T}{N} \right)^{\frac{K_D}{2}} \exp\left(-\frac{N}{T} f_X(y^*)\right). \quad (4.65)$$

¹A more precise statement will be done for next year. I still have not found a simple way to demonstrate this.

Chapter 5

Histogram clustering

This chapter illustrates how deterministic annealing can be extended to cluster objects that do not necessarily belong to an Euclidean space. In particular, we present here *histogram clustering*, a method inspired in deterministic annealing, to cluster objects that can be represented as histograms.

Figure 5.1 shows a grayscale image that has been partitioned into patches by a cyan grid. We assume all patches to be squared and equally sized. We want to cluster the patches by how similar they look to each other. One could argue that there are five clusters, as there are five predominant textures in the images. Our goal is to apply deterministic annealing to cluster patches of an image so that patches of similar textures are clustered together.

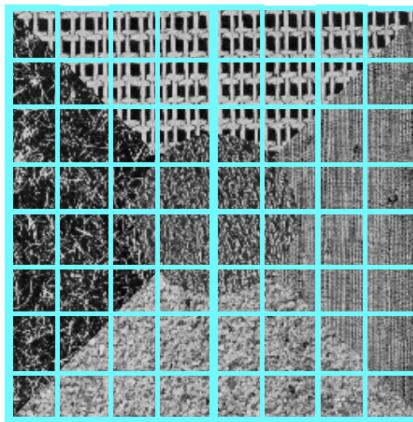


Figure 5.1

The first steps are then to define the hypothesis class and the cost function. To do this, we start by representing patches as *histograms*.

Definition 14. A (grayscale) histogram is a function $f : \{0, \dots, 255\} \rightarrow [0, 1]$ whose domain is the set of all possible grayscale values and such that $\sum_{b \in \{0, \dots, 255\}} f(b) = 1$. The histogram representing a grayscale patch x is the histogram $\hat{p}(\cdot | x)$ where $\hat{p}(b | x)$ counts the fraction of pixels in the patch with value b .

We argue that two patches x and x' should be clustered together if their representing histograms $\hat{p}(\cdot | x)$ and $\hat{p}(\cdot | x')$ are similar.

Histogram clustering is not restricted to clustering patches of an image. The ideas in this chapter can be extended to other domains like clustering documents. One can represent a document x as a histogram $f : \mathcal{W} \rightarrow [0, 1]$, where \mathcal{W} denotes a set of words appearing in a text corpus. The document x can be represented by the following probability mass function (pmf):

$$f_x(w) = \frac{\text{Number of times word } w \text{ occurs in } x}{\text{Total occurrences of all words occurring in } x}. \quad (5.1)$$

The ideas in this chapter extend easily to any setting where the objects to be clustered can be represented as such functions. This gives rise to the *histogram clustering problem*.

Definition 15. Let \mathfrak{X} be a set of objects that are perceived only through a set \mathcal{Y} of features. Assume given a function $n : \mathfrak{X} \times \mathcal{Y} \rightarrow \mathbb{N}$ that counts $(x, y) \in \mathfrak{X} \times \mathcal{Y}$ the number $n(x, y)$ of times that feature y has been observed for object x . The histogram $\hat{p}(\cdot | x) : \mathcal{Y} \rightarrow [0, 1]$ of an object $x \in \mathfrak{X}$ is the probability mass function given by

$$\hat{p}(y | x) = \frac{n(x, y)}{\sum_{y \in \mathcal{Y}} n(x, y)}, \text{ for } y \in \mathcal{Y}. \quad (5.2)$$

The problem of histogram clustering is to cluster the objects according to histograms.

Observe that for the problem of clustering image patches, the objects are patches and the features are gray values. Also, for the problem of clustering documents, the objects are documents and the features are words.

5.1 The hypothesis class and the cost function

For the sake of intuition, we present how to do histogram clustering in the context of clustering patches of an image, but this presentation easily extends to the general case.

Definition 16. Let \mathcal{X} be the set of all images patches occurring in a given image and fix $K \in \mathbb{N}$. The hypothesis class \mathcal{C} induced by \mathcal{X} and K is the set of all functions $c : \mathcal{X} \rightarrow \{1, \dots, K\}$ mapping patches to cluster indices.

We define a cost function that is analogous to the K -means cost function from Chapter 3. This cost function is parameterized by a set $Q = \{q_1, \dots, q_K\}$ of histograms that will act as “centroids”. The cost function takes then as input a function $c \in \mathcal{C}$, the set Q , and an image X partitioned into patches $\{x_1, \dots, x_N\}$. This function measures, for $i \leq N$, how much $\hat{p}(\cdot | x_i)$ diverges from $q_{c(i)}(\cdot)$. We choose the popular Kullback-Leibler divergence to measure this:

$$D_{KL}[\hat{p}(\cdot | x_i) \| q_k(\cdot)] = \sum_{y \in \{0, \dots, 255\}} \hat{p}(y | x_i) \log \left(\frac{\hat{p}(y | x_i)}{q_k(y)} \right). \quad (5.3)$$

Based on the K -means cost function from Chapter 3, we now propose the following cost function.

Definition 17.

$$R(c, Q, X) = \sum_{x \in \mathcal{X}} D_{KL}[\hat{p}(\cdot | x_i) \| q_k(\cdot)]. \quad (5.4)$$

5.2 The Gibbs distribution

5.2.1 Efficient computation of the Gibbs distribution

Lemma 7. Any Gibbs distribution $p(\cdot | X)$ induced by R and X factorizes as follows

$$p(c | Q, X) = \prod_{x \in X} p(x \mapsto c(x) | Q, X), \quad (5.5)$$

where

$$p(x \mapsto c(x) | Q, X) \propto \exp \left(-\frac{1}{T} D_{KL}[\hat{p}(\cdot | x) \| q_{c(x)}(\cdot)] \right). \quad (5.6)$$

Proof. The proof is very similar to the one for Theorem 9 and is left as an exercise to the reader. \square

This factorization leads to an efficient way to compute a Gibbs distribution for histogram clustering in $O(NK)$ -time.

5.2.2 Computing the centroids

The centroids that maximize the entropy of a Gibbs distribution can be computed analogously to the way done in Chapter 3. Once again, we assume that the temperature hyper-parameter depends exclusively on the Gibbs distribution and not on the centroids.

Exercise 17. *Show that the set Q^* of centroids that maximize the entropy of $p(\cdot | Q, X)$ must satisfy the conditions:*

$$q_k^*(\cdot) = \frac{\sum_{x \in X} p(x \mapsto k | Q^*, X) \hat{p}(\cdot | x)}{\sum_{x \in X} p(x \mapsto k | Q^*, X)}, \quad \text{for } k \leq K, \quad (5.7)$$

Observe that this does not give a closed form solution to analytically compute Q^* . So we recur to an EM-like procedure to iteratively estimate Q^* .

Exercise 18. *[Open problem!] Does this EM-like procedure converge to a local maximum?*

5.3 Histogram clustering via deterministic annealing

Instantiating deterministic annealing for our cost function and Gibbs distribution yields Algorithm 3

Figure 5.2 illustrates the behavior of the centroids Q when executing Algorithm 3 on the image in Figure 5.1. the patches in this execution have smaller size than those shown in Figure 5.1. Observe how, as the temperature decreases, the number of apparent centroids increase, following a behavior similar to the one observed in deterministic annealing.

5.4 Histogram clustering as a probabilistic generative approach

We can also use histogram clustering to produce a probabilistic model for our observation. Such a model offers some benefits. We can, for example, synthesize new observations by drawing samples from this model or define an outlier detection system.

We define a probabilistic model for a random image. In this model, the image is seen as a sample $(x_1, y_1), \dots, (x_\ell, y_\ell)$ drawn from a distribution, where

Algorithm 3 Histogram clustering via deterministic annealing

```

1: Let
2:    $\epsilon > 0$  be a temperature threshold,
3:    $\text{reduce}(\cdot)$  be a function for decreasing the temperature.
4:    $\text{close}(\cdot, \cdot)$  be a function that evaluates if two matrices are close.
5: function DETANN( $\epsilon$ ,  $\text{reduce}$ ,  $\text{close}$ )
6:    $T \leftarrow \infty$                                  $\triangleright \infty$  is a sufficiently large value.
7:    $Q \leftarrow \$$                                  $\triangleright$  Define arbitrary initial centroids.
8:   while  $T > \epsilon$  do
9:     repeat
10:     $Q_0 \leftarrow Q$ 
11:    Compute  $p(x \mapsto k | Q_0, X)$ , for  $x \in X$  and  $k \leq K$ 
12:     $Q \leftarrow \frac{\sum_{x \in X} p(x \mapsto k | Q_0, X) \hat{p}(\cdot | x)}{\sum_{x \in X} p(x \mapsto k | Q_0, X)}$ , for  $k \leq K$ .
13:   until  $\text{close}(Q, Q_0)$ 
14:   Add a small amount of noise to each centroid in  $Q$ .
15:    $T \leftarrow \text{reduce}(T)$ 
16: end while
17: return  $Q, \{p(x \mapsto k | Q, X) | x \in X, k \leq K\}$ 
18: end function

```

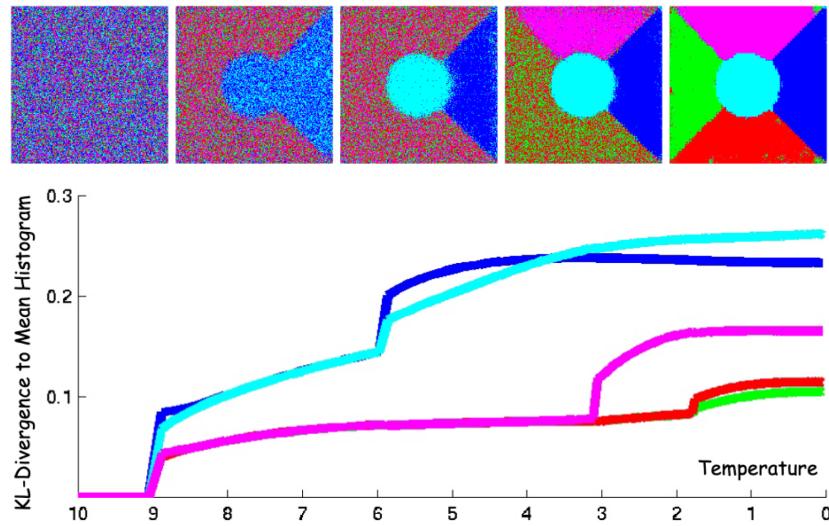


Figure 5.2: Histogram clustering for Figure 5.1. The x axis shows the temperature. The y axis measures the KL-divergence between each centroid histogram and the sample mean histogram. The graphics are from Joachim Buhmann.

x_r and y_r are a patch and a grayscale value, respectively. Let \mathcal{X} be the set of patches, $Q = \{q_1, \dots, q_K\}$ be a set of centroid histograms, \mathcal{Y} be a set of grayscale values, and $c : \mathcal{X} \rightarrow \{1, \dots, K\}$ be a function. Consider now the following procedure.

1. Draw a patch x from \mathcal{X} .
2. Draw a grayscale value y from \mathcal{Y} with probability $q_{c(x)}(y)$.

This procedure induces the following pmf over $\mathcal{X} \times \mathcal{Y}$:

$$p(x, y) = p(x)q_{c(x)}(y), \text{ for } (x, y) \in \mathcal{X} \times \mathcal{Y}. \quad (5.8)$$

As \mathcal{X} and \mathcal{Y} are finite, one can represent $p(\cdot)$ and $q_k(\cdot)$, for $k \leq K$, as pmfs of multinomial distributions. As a result, the pmf $p(\cdot, \cdot)$ is parameterized by $p(\cdot)$, c , and $q_k(\cdot)$, for $k \leq K$. We denote by θ all these parameters. We can estimate these parameters via maximum-likelihood estimation (MLE).

Exercise 19. Start our MLE by showing that the log-likelihood \mathcal{L}_θ of a sample $\{|(x_1, y_1), \dots, (x_\ell, y_\ell)|\}$ ¹ is:

$$\ell \sum_{x \in \mathcal{X}} \hat{p}(x) \log p(x) + \ell \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \hat{p}(x, y) \log q_{c(x)}(y), \quad (5.9)$$

where

$$\hat{p}(x, y) := \frac{n(x, y)}{\ell} = \frac{1}{\ell} \sum_{r \leq \ell} \mathbf{1}\{(x_r, y_r) = (x, y)\}, \text{ and} \quad (5.10)$$

$$\hat{p}(x) := \sum_{y \in \mathcal{Y}} \hat{p}(x, y). \quad (5.11)$$

Hint: Show that the likelihood of the sample is

$$\prod_{x \in \mathcal{X}} \prod_{y \in \mathcal{Y}} (p(x)q_{c(x)}(y))^{\ell \hat{p}(x, y)}. \quad (5.12)$$

Exercise 20. Show that

$$\max_{\theta} \mathcal{L}_\theta = \max_{p(\cdot)} \ell \sum_{x \in \mathcal{X}} \hat{p}(x) \log p(x) - \quad (5.13)$$

$$\min_{\substack{q_1(\cdot), \dots, q_K(\cdot) \\ c}} \sum_{r \leq \ell} D_{KL} [\hat{p}(\cdot | x_r) \mid q_{c(x_r)}(\cdot)]. \quad (5.14)$$

¹We use these brackets to denote a *multiset*. This is necessary as \mathcal{X} and \mathcal{Y} are finite and we may have repeated pairs in the sample.

It's easy to show that the maximization problem solved by $(\hat{p}(x))_{x \in \mathcal{X}}$. The minimization problem is equivalent to minimizing $R(c, Q, X)$ with respect to c and $Q = \{q_k(\cdot) \mid k \leq K\}$, which is precisely what we do in this chapter. Hence, histogram clustering yields not only a clustering of histograms, but also a probabilistic model.

Chapter 6

Information bottleneck

To be done in 2022.

Chapter 7

Constant-shift embedding

To be done in 2022.

Chapter 8

Mean-field approximation

To be done in 2022.

Chapter 9

Pairwise clustering

To be done in 2022.

Chapter 10

An information-theoretic foundation for posterior agreement

10.1 Introduction

This chapter proposes an information-theoretic foundation for posterior agreement, originally proposed by Joachim Buhmann [1] and studied by Alex Gronskiy as part of his doctoral thesis [7]. Posterior agreement can be understood as a method for *validating algorithms* that solve stochastic optimization problems of the form

$$\min_c \mathbb{E}_X [R(c, X)].$$

Here, $R(c, X)$ is the result of evaluating a cost function on a candidate solution c on an instance of the problem described by the random variable X . The expectation is computed with X 's distribution.

In posterior agreement, we assume that an algorithm is a function that computes, from a given observation X' , a *posterior distribution* $p(\cdot | X')$ over a finite space \mathcal{C} of feasible solutions. Every algorithm can be converted into such an algorithm, even if it is deterministic. In this case, the posterior distribution is just a distribution that assigns probability mass one to the algorithm's output.

Posterior agreement assesses the performance of an algorithm by computing the *expected log posterior agreement*:

$$\mathbb{E}_{X', X''} \log (|\mathcal{C}| \kappa(X', X'')), \quad (10.1)$$

where $\kappa(X', X'')$ is the *posterior agreement kernel*:

$$\kappa(X', X'') := \sum_{c \in \mathcal{C}} p(\cdot | X') p(\cdot | X'').$$

The expected log posterior agreement requires the joint probability distribution of X' and X'' , which is often unknown. One only has access to a handful of observations, at least two: X' and X'' . In this case, one can use the *empirical log posterior agreement*:

$$\log(|\mathcal{C}| \kappa(X', X'')). \quad (10.2)$$

Moreover, the empirical log posterior agreement is intended to be a metric to compare different algorithms. Therefore, it is often sufficient to measure $\kappa(X', X'')$.

We show how posterior agreement can compare, for example, Prim's, Kruskal's, and the reverse-delete algorithm for computing spanning trees for graphs whose edge weights are defined by random variables, as in the example above. Posterior agreement can also compare among different cost functions and hyperparameter values, when training machine learning models.

Therefore, posterior agreement advocates that, in the context of stochastic optimization, *algorithms should aim for maximizing the posterior agreement kernel, given two instances of the problem*, rather than minimizing the cost function for either instance or an aggregate of these instances.

ToDo: Organization of these notes.

10.2 Motivation

Figure 10.1 gives two graphs modeling a city with four main locations: North, East, South, and West. There is a road connecting any two different locations and, depending on the day, traversing that road by car takes some amount of time. The number labeling each edge indicates that time. Observe that the time varies with each day.

Consider the minimum spanning tree of each of these two graphs. For Monday's graph, the tree is the one with North at the root and all other locations as children of North. For Tuesday's graph, the tree is the one with South at the root and all other locations as children of South. We call these two trees the North and South trees, respectively.

Can we estimate from this information what Wednesday's graph's minimum spanning tree will look like?

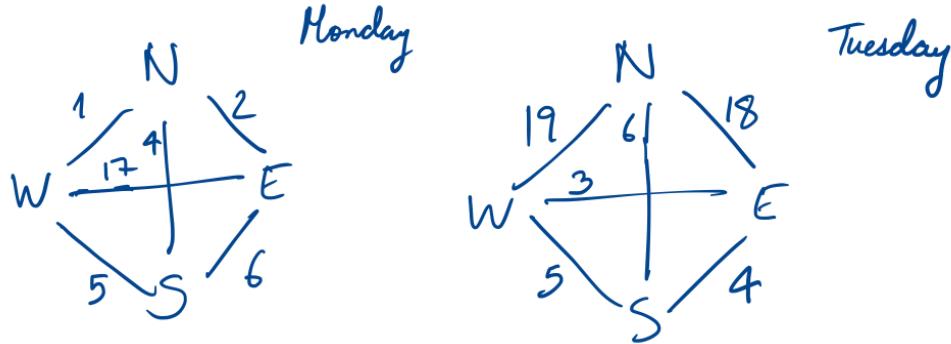


Figure 10.1: A graph with its observed edge weights for two different days.

The city graph can be understood as a random variable X . More precisely, it is a collection of 6 random variables, each of them defining an edge's weight on a particular day. Hence, the weight $R(c, X)$ of a tree c in X , which is the sum of the edge weights in c , is also a random variable. We can then formulate the problem of the minimum spanning tree for our case as the following stochastic optimization problem:

$$\arg \min_{c \in \mathcal{C}} \mathbb{E} [R(c, X)], \quad (10.3)$$

where \mathcal{C} denotes all spanning trees of the city graph.

In practice, the main challenge in this type of optimization problems is that we do not know X 's probability distribution. One natural way to go around this issue and approximately solve Problem 10.3 is to substitute $\mathbb{E} [R(c, X)]$ with an *empirical estimate*:

$$\mathbb{E} [R(c, X)] \approx \sum_{i \leq N} R(c, X_i),$$

where $\{X_1, \dots, X_n\}$ is a sample of observed values of X . When n is sufficiently large and each X_i is independent from the others, then by the law of large numbers, this becomes a good approximation of $\mathbb{E} [R(c, X)]$. We call the *empirical risk minimizer* the solution of the problem:

$$\arg \min_{c \in \mathcal{C}} \sum_{i \leq n} [R(c, X_i)], \quad (10.4)$$

We present later scenarios where the empirical risk minimizer is not the best solution we can obtain, especially when we have only very few observations. Carlos: Demonstrate this. One example is the problem of computing the

minimum of an array $X = (X_1, \dots, X_n)$ of random variables. Given just two observations X^1 and X^2 , approximating $\min_i \mathbb{E} X_i$ with $\min_i X_i^1 + X_i^2$ is not the best we can do.

10.3 Overview

Posterior agreement originates from formalizing an algorithm \mathcal{A} as a communication channel by which a sender and a receiver communicate outputs from \mathcal{A} . The robustness of an algorithm can be measured by the capacity of that communication channel, where the capacity defines the maximum number of distinguishable messages that can be communicated through the channel.

Shannon's channel coding theorem

Consider a channel by which a sender can transmit bits to a receiver. Assume that the channel is noisy in the sense that a bit can be flipped during transmission with probability $\epsilon < 0.5$. The sender and the receiver agree on transmitting only bitstrings of length n . We call these bitstrings *codewords*. Observe that if $\epsilon = 0$, then the sender can reliably communicate 2^n different messages to the receiver, by agreeing in advance with the receiver on a way to encode each message as one codeword of length n . This correspondence is called a *code*.

In practice, $\epsilon > 0$. Therefore, the sender and the receiver need to agree on a code that is robust to the channel's noise. We now show two examples of codes:

- They could agree on just 2 messages, encoded as $00\dots 0$ and $11\dots 1$, respectively. If the codeword length is sufficiently large, then with high probability, less than half of the bits will be flipped during transmission. As a result, the receiver can almost surely identify the message from the received codeword, by just counting the frequency of each bit in the codeword.
- They agree on sending 2^n messages, each encoded with a unique codeword of length n . With high probability, some of the bits will be flipped during transmission and the receiver will fail to identify the message that the sender tried to communicate.

Observe that these two codes represent two extremes, as $n \rightarrow \infty$. On one hand, the first code communicates only 2 messages, but with high probability

of success. On the other hand, the second code communicates 2^n messages, but with low probability of success. One can also imagine other codes that strike a balance between the number of messages to be communicated and the success probability.

Shannon's coding theorem answers the following question. *What is the code that maximizes the number of messages that the sender can communicate to the receiver, while attaining a probability of success close to 1, as $n \rightarrow \infty$?* When $\epsilon > 0$, this number is clearly below 2^n , but since $\epsilon < 0.5$, this number must be positive. Shannon shows that this number is 2^{nc} , where c is the channel's capacity, a quantity that is defined by the channel. Therefore, channels with higher capacity allow the communication of more messages at the same codeword length.

Posterior agreement

Posterior agreement originates from modeling an algorithm \mathcal{A} as a communication channel $C_{\mathcal{A}}$, where a sender communicates outputs from \mathcal{A} to a receiver. We argue that the capacity of $C_{\mathcal{A}}$ is defined by \mathcal{A} 's robustness to noise in the input. That is, if \mathcal{A} is robust to noise, then it is possible to communicate many more messages through $C_{\mathcal{A}}$ than when \mathcal{A} is sensitive to noise.

We give an overview of this argument next. A rigorous argument is given in Section 10.5.

We assume given an *instance space* \mathcal{X} , comprising all possible observations, and a *solution space* \mathcal{C} , comprising all possible solutions. A phenomenon is then a probability distribution over \mathcal{X} . We assume that algorithms intending to solve Problem 10.3 receive in the input an observation X' and output a distribution $p(\cdot | X')$ over \mathcal{C} .

Example 1. *In the minimum spanning tree problem, a codeword is a spanning tree, a phenomenon is a distribution governing a graph's edge weights, and an observation is a graph whose edge weights are drawn from a fixed distribution.*

Example 2. *In the centroid-based clustering problem, a codeword is a cluster assignment function and a set of centroids, an observation is a set of points to be clustered, and a phenomenon is a distribution where the points are drawn from.*

For our analysis, we assume that each instance space contains all observations of a given “size” $n \in \mathbb{N}$. This size n is a notion that measures the observations' and phenomena's complexity. For example, in the minimum spanning tree problem, an instance space contains only all weighted graphs with a fixed number n of vertices.

For an algorithm \mathcal{A} , we define a communication channel $C_{\mathcal{A}}$ that works as follows. To use the channel, a sender picks an instance X' , drawn from a phenomenon p_X , computes and inputs $p(\cdot | X')$ to the channel. The channel replaces $p(\cdot | X')$ with $p(\cdot | X'')$, where X'' is a fresh new instance drawn from p_X . The channel outputs $p(\cdot | X'')$ to the receiver.

We now emphasize the key insight of this modeling. If \mathcal{A} is robust to the fluctuations in X' , then there should not be much difference between $p(\cdot | X')$ and $p(\cdot | X'')$. In contrast, if \mathcal{A} is very sensitive to the fluctuations in X' , then $p(\cdot | X')$ and $p(\cdot | X'')$ may be very different. Hence, \mathcal{A} 's robustness to noise defines how many different “messages” can we send through this channel. We conclude then that \mathcal{A} 's robustness is measured by the capacity of this channel $C_{\mathcal{A}}$. This capacity, as we show in Section 10.5, can be estimated by the expected log posterior agreement. For this reason, we argue that algorithms intended to solve Problem 10.3 shall be measured by their expected log posterior agreement.

Observe the following analogies to Shannon's coding theory. Channels have a capacity that define the maximum number of distinguishable messages that can be communicated. Channels with higher capacity are preferable, as they allow more different messages to be communicated. Analogously, we argue that algorithms can be modeled as channels and, therefore, have a capacity, which we later show to be the expected log posterior agreement. Hence, we argue that algorithms with higher expected log posterior agreement are preferable, as they allow more different messages to be communicated.

Protocol overview

The protocol describes a code by which a sender can communicate messages to a receiver. Let \mathcal{A} be an algorithm intending to solve Problem 10.3. In our case, a message is a phenomenon and a codeword is the output $p(\cdot | X')$ of \mathcal{A} when given an observation X' from a phenomenon as input. Analogous to Shannon's coding theorem, the sender and the receiver aim to maximize the number of different messages that the sender can communicate, while ensuring that the receiver's probability of success goes to 1 as $n \rightarrow \infty$.

Figure 10.2 gives an overview of the protocol. The sender must describe a phenomenon q to a receiver through a noisy channel. The sender makes an observation X' from q , uses \mathcal{A} to compute $p(\cdot | X')$, and sends through the channel to the receiver. The channel is noisy and we represent its noise by replacing $p(\cdot | X')$ with $p(\cdot | X'')$, where X'' is another observation from q . The receiver succeeds if he is able, using $p(\cdot | X'')$, to distinguish X' from observations from other different phenomena.

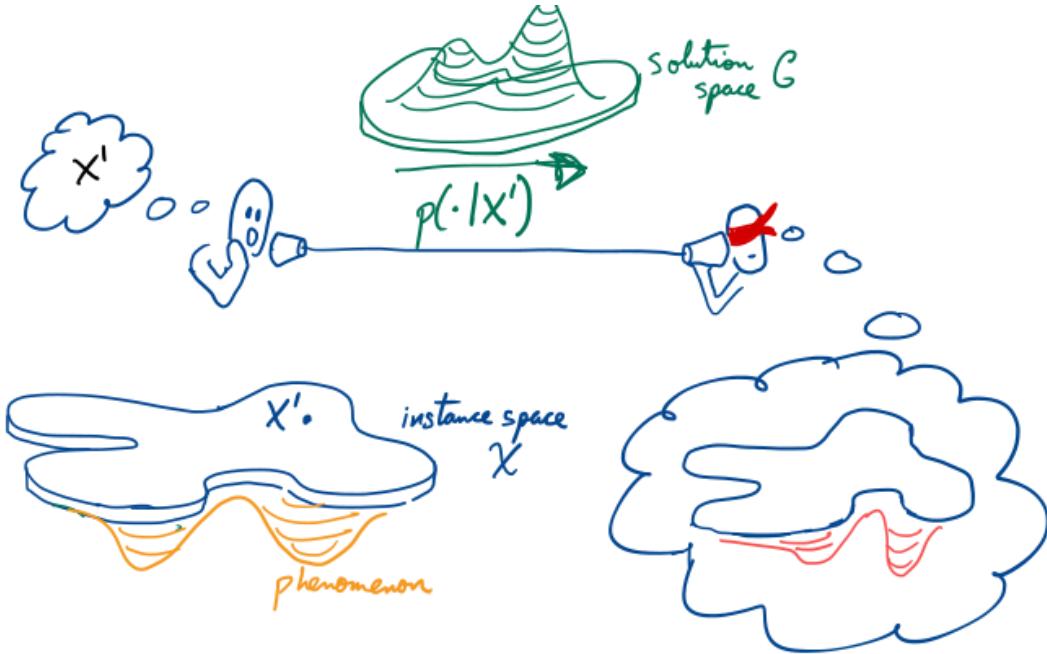


Figure 10.2

We remark that this protocol is just a thought experiment. The protocol is computationally impossible for many interesting optimization problems. This is because the protocol requires that we know the underlying distribution behind the observations of one phenomenon. This is not possible in most of the cases. Nonetheless, the protocol provides a formal justification and motivation for posterior agreement.

Protocol example

We now give a more precise overview by showing how posterior agreement works with a simple example. Let \mathcal{A}_1 and \mathcal{A}_2 be two algorithms that estimate the mean of a univariate distribution, given only a sample from that distribution. For example, \mathcal{A}_1 fits a Gaussian to the sample via maximum-likelihood estimation whereas \mathcal{A}_2 does the same, but only using only the sample maximum and minimum. Suppose that both algorithms output Gaussian distributions that indicate where they believe that the mean is.

The protocol works as follows. Fix $n \in \mathbb{N}$, which denote the size of all the observations in the instance space. In our case, n denotes the sample size. The sender and the receiver are given the algorithm under evaluation \mathcal{A}_i and then they choose the size $m_n \in \mathbb{N}$ of the set of messages that they the sender

will attempt to communicate. Recall that they want to choose m_n as large as possible. However, a large m_n increases the probability P_n that the receiver fails to recognize the message from the codeword transmitted by the sender. We later see that the best choice for m_n is defined by \mathcal{A}_i .

The protocol proceeds then as follows:

1. (Figure 10.3) The sender and the receiver agree on a set of $m := m_n$ phenomena, which we represent with the probability distributions q_1, q_2, \dots, q_m .
2. (Figure 10.3) The sender and the receiver together make one observation X'_i of each phenomenon q_i , with $i \leq m$. In this case, an observation is a sample of points from q_i . Afterwards, they use \mathcal{A}_i to compute a distribution $p(\cdot | X')$ for each observation X' .
3. (Figure 10.4) An observation X' is chosen out of these m observations uniformly at random. X' is given to the sender, but kept secret from the receiver.
4. (Figure 10.5) The sender sends $p(\cdot | X')$ to the receiver through a *noisy communication channel*. This channel replaces $p(\cdot | X')$ with $p(\cdot | X'')$, where X'' is a fresh new observation from the same phenomenon where X' comes from.
5. (Figure 10.6) The receiver gets $p(\cdot | X'')$ and must now guess which observation in $\{X'_1, \dots, X'_m\}$ the sender chose in Step 3. For this, the receiver uses the natural approach of guessing the observation \hat{X} for which $p(\cdot | \hat{X})$ overlaps the most with $p(\cdot | X'')$. In other words, the receiver guesses the observation \hat{X} that fulfills:

$$\kappa(\hat{X}, X') \geq \kappa(Y, X'), \text{ for all } Y.$$

6. If $\hat{X} = X'$, the receiver has succeeded.

Receiver's probability of failure

We show in Section 10.5.4 that the receiver's failure probability is bounded above by

$$\exp(-\mathbb{E}_{X', X''} [\log(|\mathcal{C}| \kappa(X', X''))] + \epsilon \log |\mathcal{C}| + \log m),$$

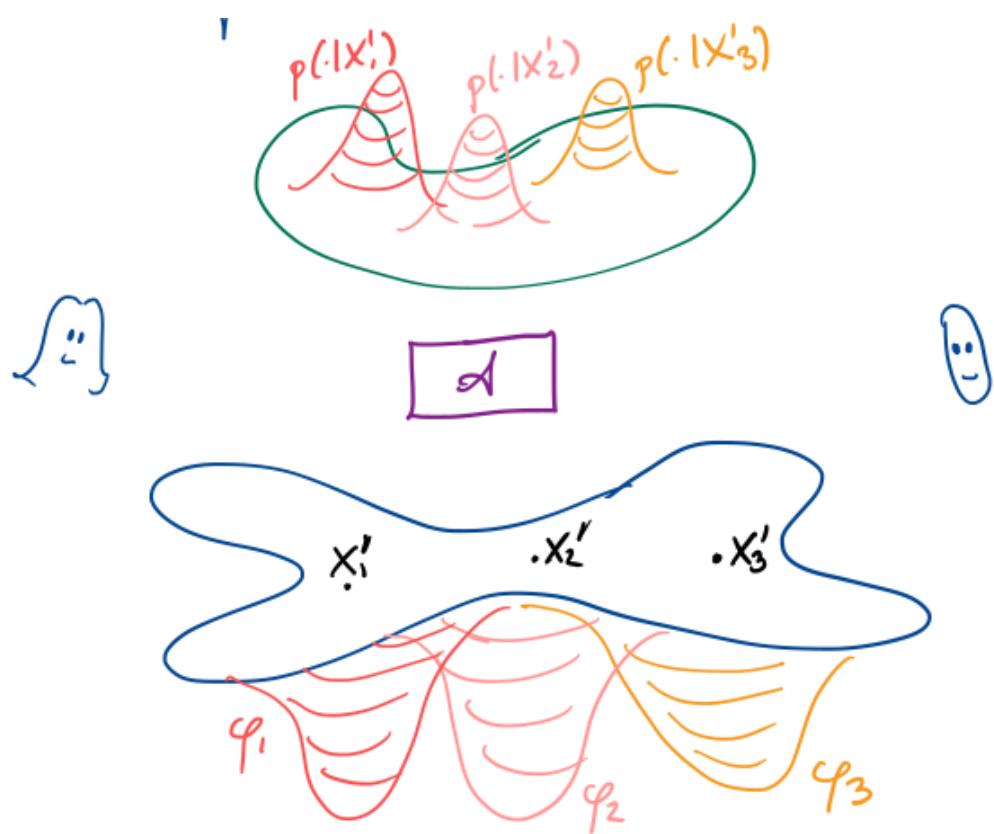


Figure 10.3

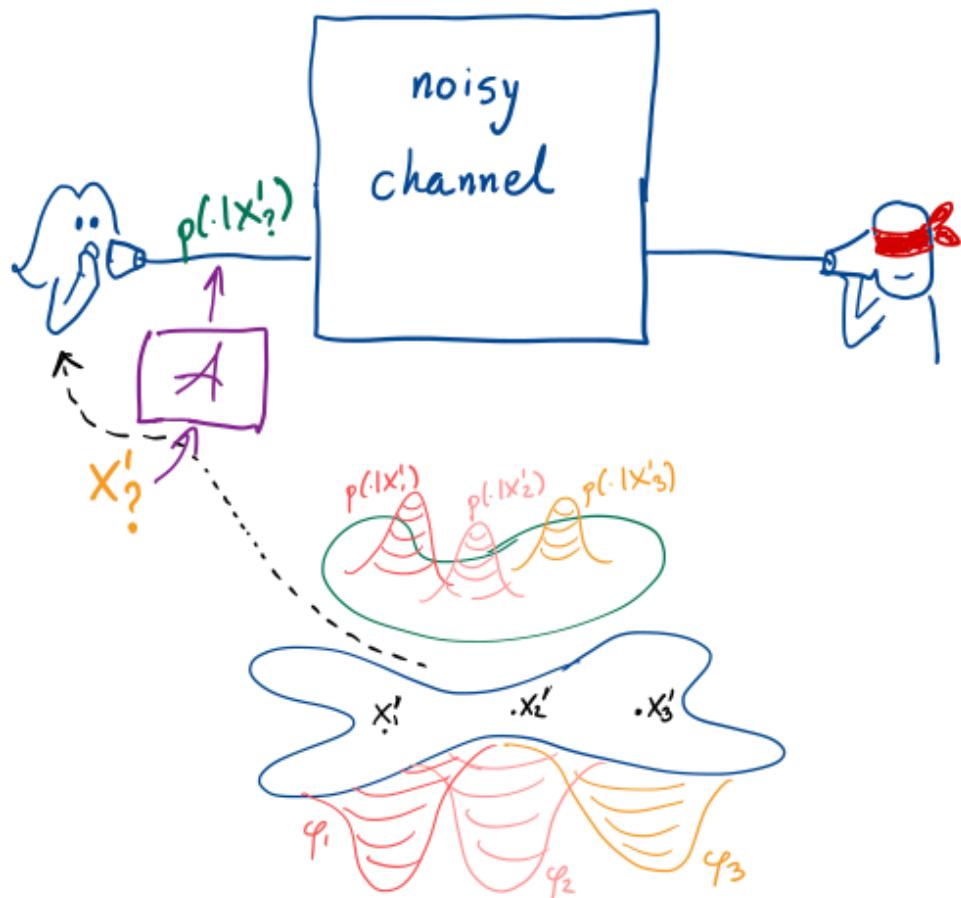


Figure 10.4

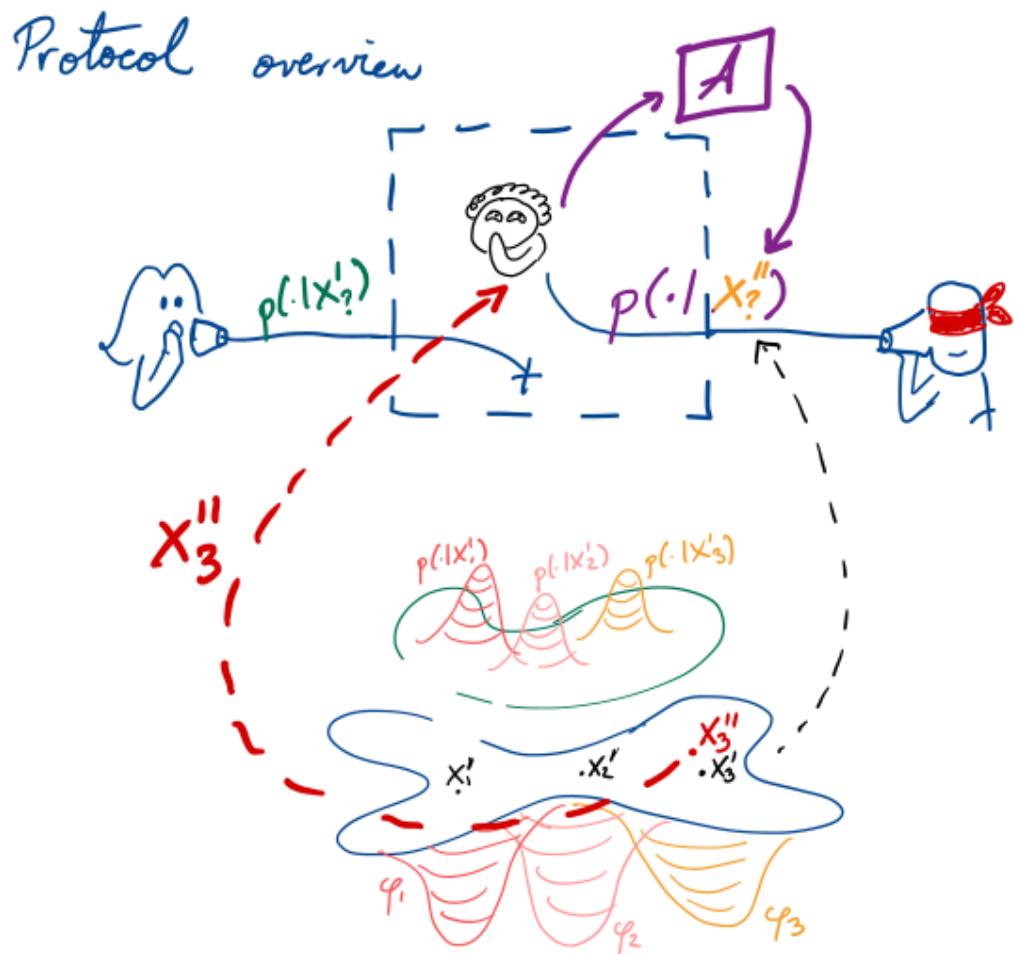


Figure 10.5



Figure 10.6

where $\epsilon > 0$ is arbitrary, \mathcal{C} is the solution space, and m is the number of observations defined in Step 1.

Assume now that $\log |\mathcal{C}| = \Omega(n)$. This is a reasonable assumption, as for many interesting problems, \mathcal{C} grows exponentially on n . Observe that the algorithm can only influence $\kappa(X', X'')$. If ϵ is sufficiently small and the algorithm ensures that

$$\mathbb{E}_{X', X''} \log(|\mathcal{C}| \kappa(X', X'')) - \log m = \Omega(n),$$

then the receiver's failure probability becomes 0 as $n \rightarrow \infty$. The algorithm can ensure this by *maximizing the expected log posterior agreement*. The larger this quantity is, the higher m can be and the more messages the sender can communicate to the receiver.

10.4 Shannon's channel coding theorem

To properly formulate and analyze the communication protocol above, we build upon Shannon's channel coding theorem. This theorem measures how much information we can optimally send through a communication channel. This section is mainly based on Chapter 7 from Thomas and Cover [4].

10.4.1 Channels

We understand a channel as a medium by which one sender can send symbols from a fixed set \mathfrak{A} to a receiver. We allow channels to be noisy, meaning that the symbol a can be altered to another symbol b with probability $p(b | a)$ during the transmission through the channel. We do not allow the receiver to give feedback to the sender.

Definition 18. A (noisy) channel is a pair $(\mathfrak{A}, \{p(\cdot | a)\}_{a \in \mathfrak{A}})$, where \mathfrak{A} is a set and $p(\cdot | a)$, for $a \in \mathfrak{A}$, is a conditional distribution on \mathfrak{A} .

For convenience, we sometimes write just \mathcal{P} to denote the family of distributions $\{p(\cdot | a)\}_{a \in \mathfrak{A}}$. Observe that we assume that transmissions are independent from each other. What the receiver gets does not influence what the user sends or the channel's conditional probabilities in the future.

Example 3. The binary channel is the channel with $\mathfrak{A} = \{0, 1\}$ and $p(b | a) = \mathbb{I}\{a = b\}$, for $a, b \in \mathfrak{A}$, where \mathbb{I} is the indicator function. This is a channel where there is no noise interference.

Example 4. The noisy binary channel is the binary channel, but with

$$p(b | a) \begin{cases} 1 - \epsilon & \text{if } a = b \text{ and} \\ \epsilon & \text{if } a \neq b. \end{cases}$$

Unless stated otherwise, we assume that $0 < \epsilon \leq 0.5$.

Example 5. The typewriter channel is the channel with $\mathfrak{A} = \{a, b, \dots, z\}$ and $p(b | a) = \mathbb{I}\{a = b\}$, for $a, b \in \mathfrak{A}$, where \mathbb{I} is the indicator function.

Example 6. The noisy typewriter channel is the channel with $\mathfrak{A} = \{a, b, \dots, z\}$, but with

$$p(b | a) \begin{cases} 1 - \epsilon & \text{if } a = b, \\ \epsilon & \text{if } a = b + 1 \bmod 26. \end{cases}$$

10.4.2 Channel capacity

Which of the channels above sends *the most information per transmission*? Intuitively, a letter has more information than a bit and the presence of noise affects the amount of information we send in one transmission. Indeed, sending one letter through the typewriter channel provides more information than sending one letter through a noisy typewriter channel. Also, a letter has more information than a bit. Hence, we say that the typewriter channel has *the most capacity*: one transmission through this channel carries in average more information than one transmission through any of the other three channels. Similarly, we say that the noisy binary channel has *the least capacity*.

We now formally define channel capacity.

Definition 19. For a channel $(\mathfrak{A}, \{p(\cdot | a)\}_{a \in \mathfrak{A}})$, its capacity is

$$\max_{p(\cdot)} I(S; \hat{S}),$$

where S and \hat{S} are random variables denoting a symbol input to the channel and the output symbol, respectively, when S is distributed according to $p(\cdot)$. We refer to $I(S; \hat{S})$ as the channel's input-output mutual information and the joint distribution of S and \hat{S} as the input-output distribution.

Example 7. For the binary channel, we can reliably send one bit per transmission. This corresponds to the channel's capacity,

$$\max_{p(\cdot)} I(S; \hat{S}) = \max_{p(\cdot)} \{H(S) - H(\hat{S} | S)\} = \max_{p(\cdot)} H(S) = \max_{p(\cdot)} H(S) = 1.$$

The second equality follows from the fact that $\hat{S} = S$, so $H(\hat{S} | S) = 0$. The last equality follows from the fact that the distribution that maximizes the entropy of a Bernoulli random variable is the uniform distribution, which yields an entropy of 1 bit.

Example 8. A similar line of reasoning shows that the typewriter channel's capacity is $\max_{p(\cdot)} H(S) = \log 26 \approx 4.7$. This corresponds to the intuition that we can reliably send one letter per transmission, which contains around 4.7 bits of information.

Example 9. Consider now the noisy typewriter with $\epsilon = 0.5$. How many bits can we reliably send per transmission? Observe that one way to reliably send information is by agreeing to only send letters at even positions in the alphabetic order. In that way, if you receive, for example, **a** or **b**, you know for sure that the sender input **a** to the channel. However, by sending only the “even” letters, you need to double the efforts with respect to the typewriter channel without noise. As a consequence, the noisy typewriter has less capacity. One can actually show that $\max_{p(\cdot)} I(S; \hat{S}) = -1 + \log 26$, where a maximizing $p(\cdot)$ is the uniform distribution over the “even” letters.

10.4.3 Codes

Intuition on codes and rates

Carlos: Why aren't we taking into account loss bits or duplicated bits? This can happen during transmission.

Consider the noisy binary channel. By sending several bits in a specific pattern to the channel, one can come up with sophisticated ways to transmit complex information through the channel, like images or spreadsheets. We illustrate this by showing how to use the binary channel to send letters in $\{a, b, \dots, z\}$. The sender and the receiver must first agree on a *code* for those letters. One such code, which we call *naïve code*, encodes the letter **a** as the codeword 00000, **b** as 00001, and so on. That is, the codeword for the i -th alphabet letter is number i in base 2, written as a bit string of length 5. In a similar fashion, we can conceive codes for communicating more complex data like images and spreadsheets.

Unfortunately, the code mentioned in the previous paragraph is sensitive to the channel's noise. If we send the codeword for **a**, the receiver may get the codeword for **b** with probability $\epsilon(1 - \epsilon)^4$, yielding a *communication error*. Information theory has come up with smarter codes that reduce the probability of such a communication error, but at the cost of longer codewords. For

example, we can use a code, which we call *the 5-redundant code*. This code encodes **a** as 00000 and **b** as 11111. The receiver would then take the received codeword \hat{w} and search for the codeword w that closest codeword with respect to the Hamming distance. The receiver would then assume that the sender sent the letter associated to w . For example, if the receiver gets 11010, then she assumes that the sender sent 11111, which is the codeword for **b**. This code is more robust to noise than the naïve alphabet code. In comparison with the naïve code, more bits need to be flipped by noise in order to get a communication error. This is less likely than having one bit flipped in the naïve code's codewords.

Unfortunately, the robustness comes at the price of less messages. If we use codewords of length n only. Using the naïve code, the sender can communicate 2^n different messages. However, using the n -redundant code, the sender can communicate only 2. Assuming that the noise does not cause a communication error, we manage to transmit at a *rate* of one bit per transmission in the naïve code and one bit per n transmissions in the n -redundant code. In the limit, as $n \rightarrow \infty$, the naïve code attains a rate of 1 bit per transmission, but a probability of a communication error equal to 1. On the other hand, the n -redundant code attains a rate of 0 bits per transmission, but a probability of a communication error equal to 0. This comparison is summarized in Figure 10.7

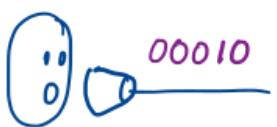
Intuition on Shannon's channel coding theorem

We started by using the binary noisy channel to send bits and we are now devising codes to send a finite set of letters through the binary noisy channel. In a similar manner, we can devise codes to send words through the binary noisy channel, by creating a code that maps words to bit strings of fixed length.[Carlos: Why are we talking only of finite sentences \$S\$?](#) In practice, I want to send sentences of variable lengths. Also, why don't we allow encodings of different lengths?

We can continue in this way to create codes to send images of a fixed size, videos of a fixed length, and so on.[Carlos: Why the obsession with fixed size?](#) All this is done by using longer codewords.

Shannon's coding theorem concerns with the problem of finding sustainable strategies for building codes. By sustainable, we mean that the strategy should yield codes that fulfill two conditions.

- First, the codes attain and maintain a positive rate r of bits of information per transmission as $n \rightarrow \infty$.
- Second, the probability of a communication error goes to zero as $n \rightarrow \infty$.

	Naïve	n -redundant
codeword length		
# messages	2^n	2^n
prob. fail $n \rightarrow \infty$	1	0
bits/trans $n \rightarrow \infty$	1	0

How can we get positive rate and prob.fail=0 as $n \rightarrow \infty$?

Figure 10.7

Shannon's channel coding theorem states that there is a sustainable strategy for building codes, as long as the targeted rate r is below the channel's capacity. For a fixed codeword length n , the maximum number of messages that can be communicated with this strategy is $\lfloor 2^{nr} \rfloor$.

Formalization

For the definitions below, let $M, n \in \mathbb{N}$ and let $(\mathfrak{A}, \mathcal{P})$ be a channel.

Definition 20. An (M, n) -code is a pair (Enc, Dec) of functions with $Enc : \{1, 2, \dots, M\} \rightarrow \mathfrak{A}^n$ and $Dec : \mathfrak{A}^n \rightarrow \{1, 2, \dots, M\}$.

Definition 21. The rate of a (M, n) -code is

$$\frac{\log M}{n}.$$

Observe that if a (M, n) -code has rate r , then $M = 2^{nr}$.

Definition 22. For an (M, n) -code (Enc, Dec) , its probability of a communication error is

$$\frac{1}{M} \sum_{i \leq M} \mathbf{P} \left(Dec(\hat{W}) \neq i \mid W = Enc(i) \right),$$

where $\mathbf{P} \left(Dec(\hat{W}) \neq i \mid W = Enc(i) \right)$ is the probability that the receiver decodes something different to i , given that we sent $Enc(i)$ through the channel.

Intuitively, the probability of a communication error is the probability that the receiver decodes a wrong message when we send him the codeword of a message chosen uniformly at random.

Example 10. The n -redundant code discussed above is an example of a $(2^n, n)$ -code, whose rate is $\log 2^n/n = 1$.

Definition 23. A rate r is attainable if there is a sequence of $(\lfloor 2^{nr} \rfloor, n)$ -codes, indexed by n , such that the probability of a communication error goes to zero as $n \rightarrow \infty$.

10.4.4 Shannon's coding theorem

Typicality

We now recall some important notions from Thomas and Cover [4]

Theorem 24. (*The asymptotic equipartition property*) Let S, S_1, S_2, \dots, S_n be identically and independently distributed random variables with distribution $p(\cdot)$ over a space \mathcal{S} , then

$$-\frac{1}{n}p(S_1, \dots, S_n) \rightarrow H(S), \quad \text{in probability as } n \rightarrow \infty.$$

This theorem follows from the weak law of large numbers. In our context, S, S_1, S_2, \dots, S_n denote symbols from a channel's alphabet and $p(\cdot)$ is $\arg \max_p I(S; \hat{S})$. That is, $p(\cdot)$ is the distribution that achieves the channel's capacity.

Definition 25. For $n \in \mathbb{N}$ and $\epsilon > 0$, the typical set $A_\epsilon^{(n)}$ with respect to $p(\cdot)$ is the set of sequences $(s_1, \dots, s_n) \in \mathcal{S}^n$ such that

$$H(S) - \epsilon \leq -\frac{1}{n} \log p(s_1, s_2, \dots, s_n) \leq H(S) + \epsilon.$$

A sequence in $A_\epsilon^{(n)}$ is called a typical sequence.

In our context, codewords will consist of typical sequences. The next theorem justifies the following intuitions:

1. All typical sequences have approximately the same probability $\approx 2^{-nH(S)}$.
2. If you draw a sequence in \mathcal{S}^n , using $p(\cdot)$, then the resulting sequence is typical with high probability.
3. $|A_\epsilon^{(n)}| \approx 2^{nH(S)}$.

Theorem 26.

1. If $(s_1, s_2, \dots, s_n) \in A_\epsilon^{(n)}$, then $2^{-n(H(S)+\epsilon)} \leq p(s_1, \dots, s_n) \leq 2^{-n(H(S)-\epsilon)}$.
2. $\mathbf{P}(A_\epsilon^{(n)}) > 1 - \epsilon$, for sufficiently large n .
3. $(1 - \epsilon)2^{n(H(S)-\epsilon)} \leq |A_\epsilon^{(n)}| \leq 2^{n(H(S)+\epsilon)}$.

The reader can take it as an exercise to proof these claims. The proofs are in Thomas and Cover [4].

Definition 27. Let $n \in \mathbb{N}$ and let $p_{S\hat{S}}(\cdot, \cdot)$ be the joint distribution of two random variables S and \hat{S} , whose ranges are \mathcal{S} and $\hat{\mathcal{S}}$, respectively. The set $A_\epsilon^{(n)}$ of jointly typical sequences with respect to $p_{S\hat{S}}$ is the set of pairs $(\mathbf{s}^n, \hat{\mathbf{s}}^n)$ of sequences that fulfill the following:

1. $\left| -\frac{1}{n} \log p_{S^n}(\mathbf{s}^n) - H(S) \right| < \epsilon.$
2. $\left| -\frac{1}{n} \log p_{\hat{S}^n}(\hat{\mathbf{s}}^n) - H(\hat{S}) \right| < \epsilon.$
3. $\left| -\frac{1}{n} \log p_{S^n \hat{S}^n}(\mathbf{s}^n, \hat{\mathbf{s}}^n) - H(S, \hat{S}) \right| < \epsilon.$

A pair in $A_\epsilon^{(n)}$ is called a jointly typical pair of sequences.

We clarify that, for $\mathbf{s}^n = (s_1, \dots, s_n)$ and $\hat{\mathbf{s}}^n = (\hat{s}_1, \dots, \hat{s}_n)$,

$$\begin{aligned} p_{S^n}(\mathbf{s}^n) &= \prod_{i \leq n} p_S(s_i), \\ p_{\hat{S}^n}(\hat{\mathbf{s}}^n) &= \prod_{i \leq n} p_{\hat{S}}(\hat{s}_i) = \prod_{i \leq n} \sum_s p_S(s) p_{\hat{S}|S}(\hat{s}_i | s), \text{ and} \\ p_{S^n \hat{S}^n}(\mathbf{s}^n, \hat{\mathbf{s}}^n) &= \prod_{i \leq n} p_{S \hat{S}}(s_i, \hat{s}_i) = \prod_{i \leq n} p_S(s_i) p_{\hat{S}|S}(\hat{s}_i | s_i). \end{aligned}$$

In the context of communication via a channel, $p_{S^n \hat{S}^n}$ represents the joint distribution of \mathbf{S}^n and $\hat{\mathbf{S}}^n$, where

- \mathbf{S}^n denotes a random codeword, where each symbol was chosen at random according to the distribution $p_S = \arg \max_p I(S; \hat{S})$ that achieves channel capacity.
- $\hat{\mathbf{S}}^n$ denotes a codeword that the channel would output, after we send \mathbf{S}^n as input.

We call $p_{S^n \hat{S}^n}$ the *codeword input-output distribution*.

In the context of communication via a channel, the following theorem justifies the following intuitions:

1. Suppose that we build a codeword \mathbf{s}^n at random by choosing each of its symbols at random according to p_S , the distribution that attains channel capacity. Then we send \mathbf{s}^n through the channel and let $\hat{\mathbf{s}}^n$ be the output codeword. Then $(\mathbf{s}^n, \hat{\mathbf{s}}^n)$ is jointly typical with high probability.
2. Suppose now that we build another codeword \mathbf{q}^n at random using the same procedure. Then it is very *unlikely* that $(\mathbf{q}^n, \mathbf{y}^n)$ is jointly typical.

Theorem 28.

1. $\mathbf{P} \left((\mathbf{S}^n, \hat{\mathbf{S}}^n) \in A_\epsilon^{(n)} \right) \rightarrow 1, \text{ as } n \rightarrow \infty.$
2. If $\mathbf{Q}^n \sim p_{S^n}(\cdot)$ and $\hat{\mathbf{S}}^n \sim p_{\hat{S}^n}(\cdot)$ (i.e., they are drawn independently at random from the marginal distributions of $p_{S^n \hat{S}^n}$), then

$$(1 - \epsilon)2^{-n(I(S; \hat{S}) + 3\epsilon)} \leq \mathbf{P} \left((\mathbf{Q}^n, \hat{\mathbf{S}}^n) \in A_\epsilon^{(n)} \right) \leq 2^{-n(I(S; \hat{S}) - 3\epsilon)}$$

These two intuitions justify the effectiveness of a very simple code, called *Shannon's random code*.

Shannon's random code

Theorem 29. *A rate is attainable iff it is below the channel's capacity.*

We only focus here on proving the following direction: if a rate is below the channel's capacity, then it is attainable, as it illustrates how to propose a $(\lfloor 2^{nr} \rfloor, n)$ -code for communicating $\lfloor 2^{nr} \rfloor$ messages.

To prove this, we present, for $n > 1$, a $(\lfloor 2^{nr} \rfloor, n)$ -code whose probability of a communication error is at most $p_n = 2^{-n(cap - 3\epsilon - r)}$, where ϵ is chosen to be sufficiently small. Hence, if $r < cap$, we get that the probability of a communication error goes to zero as $n \rightarrow \infty$.

The code's encoder function *Enc* is defined as follows. For a message $m \leq \lfloor 2^{nr} \rfloor$, we define *Enc*(m) as a string in \mathbf{s}^n where each symbol was drawn from a distribution $p^* = \arg \max_{p(\cdot)} I(S; \hat{S})$. That is, a distribution that maximizes the channel's input-output mutual information and attains the channel's capacity.

The code's decoder function *Dec* is defined as follows. Given the string \mathbf{s}^n output by the channel, *Dec* goes through each message m and tests if $(\text{Enc}(m), \mathbf{s}^n)$ is jointly typical with respect to the codeword input-output distribution $p_{S^n \hat{S}^n}$. *Dec* outputs the first message for which this test succeeds. If no message succeeds on the test, then *Dec* outputs an arbitrary message.

Carlos: Remind the reader that the sender and the receiver agree in advance on *Enc* and *Dec*. Therefore, the receiver also knows both *Enc* and the set $\{1, 2, \dots, \lfloor 2^{nr} \rfloor\}$ of messages.

Theorem 30. *The probability $\mathbf{P}(\mathcal{E})$ of a communication error for Shannon's random code goes to 0 as $n \rightarrow \infty$.*

Proof. Let \mathcal{K} be a random variable representing a possible code. Then

$$\mathbf{P}(\mathcal{E}) = \sum_{\mathcal{K}} \mathbf{P}(\mathcal{K}) P_e(\mathcal{K}),$$

where $P_e(\mathcal{K})$ is the probability of a communication error for code \mathcal{K} . Observe now that

$$\begin{aligned}\mathbf{P}(\mathcal{E}) &= \sum_{\mathcal{K}} \mathbf{P}(\mathcal{K}) P_e(\mathcal{K}) \\ &= \sum_{\mathcal{K}} \mathbf{P}(\mathcal{K}) \frac{1}{[2^{nr}]} \sum_{w \leq [2^{nr}]} \mathbf{P}\left(Dec(\hat{\mathbf{S}}^n) \neq w \mid \mathbf{S}^n = Enc(w)\right) \\ &= \frac{1}{[2^{nr}]} \sum_{\mathcal{K}} \sum_{w \leq [2^{nr}]} \mathbf{P}(\mathcal{K}) \mathbf{P}\left(Dec(\hat{\mathbf{S}}^n) \neq w \mid \mathbf{S}^n = Enc(w)\right).\end{aligned}$$

Observe now that all codewords were chosen independently at random, so

$$\mathbf{P}\left(Dec(\hat{\mathbf{S}}^n) \neq w \mid \mathbf{S}^n = Enc(w)\right) = \mathbf{P}\left(Dec(\hat{\mathbf{S}}^n) \neq 1 \mid \mathbf{S}^n = Enc(1)\right),$$

for $w > 1$. Hence,

$$\begin{aligned}\mathbf{P}(\mathcal{E}) &= \frac{1}{[2^{nr}]} \sum_{\mathcal{K}} \sum_{w \leq [2^{nr}]} \mathbf{P}(\mathcal{K}) \mathbf{P}\left(Dec(\hat{\mathbf{S}}^n) \neq w \mid \mathbf{S}^n = Enc(w)\right) \\ &= \sum_{\mathcal{K}} \mathbf{P}(\mathcal{K}) \mathbf{P}\left(Dec(\hat{\mathbf{S}}^n) \neq 1 \mid \mathbf{S}^n = Enc(1)\right) \\ &= \mathbf{P}(\mathcal{E} \mid w = 1).\end{aligned}$$

This means that the probability of a communication error is equal to the probability of a communication error, assuming that the sender sent the codeword for message 1.

Note that, in Shannon's random code, the event of a communication error implies at least one of the following events: the received codeword $\hat{\mathbf{S}}^n(1)$ is not jointly typical with the codeword $\mathbf{S}^n(1)$ for message 1 or the received codeword $\hat{\mathbf{S}}^n(1)$ is jointly typical with the codeword $\mathbf{S}^n(w)$ for a message $w > 1$. More precisely,

$$\mathbf{P}(\mathcal{E} \mid M = 1) = \mathbf{P}\left(\begin{array}{l} (\mathbf{S}^n(1), \hat{\mathbf{S}}^n(1)) \notin A_\epsilon^{(n)} \text{ or} \\ (\mathbf{S}^n(2), \hat{\mathbf{S}}^n(1)) \in A_\epsilon^{(n)} \text{ or} \\ (\mathbf{S}^n(3), \hat{\mathbf{S}}^n(1)) \in A_\epsilon^{(n)} \text{ or} \\ \vdots \\ (\mathbf{S}^n([2^{nr}]), \hat{\mathbf{S}}^n(1)) \in A_\epsilon^{(n)}. \end{array}\right).$$

By the union bound,

$$\mathbf{P}(\mathcal{E} \mid M = 1) \leq \mathbf{P}\left(\left(\mathbf{S}^n(1), \hat{\mathbf{S}}^n(1)\right) \notin A_\epsilon^{(n)}\right) + \sum_{w>1} \mathbf{P}\left(\left(\mathbf{S}^n(w), \hat{\mathbf{S}}^n(1)\right) \in A_\epsilon^{(n)}\right).$$

We now apply Theorem 28, which implies the following:

- $\mathbf{P}\left(\left(\mathbf{S}^n(1), \hat{\mathbf{S}}^n(1)\right) \notin A_\epsilon^{(n)}\right) \rightarrow 1$, as $n \rightarrow \infty$. In other words, $\mathbf{P}\left(\left(\mathbf{S}^n(1), \hat{\mathbf{S}}^n(1)\right) \notin A_\epsilon^{(n)}\right) \rightarrow 0$, as $n \rightarrow \infty$.
- $\mathbf{P}\left(\left(\mathbf{S}^n(w), \hat{\mathbf{S}}^n(1)\right) \in A_\epsilon^{(n)}\right) \leq 2^{-n(I(S;\hat{S})-3\epsilon)}$. This is because, for $w > 1$, $\mathbf{S}^n(1)$ and $\mathbf{S}^n(w)$ were independently drawn from p_{S^n} and, therefore, $\hat{\mathbf{S}}^n(1)$ and $\mathbf{S}^n(w)$ were independently drawn from $p_{\hat{S}^n}$ and p_{S^n} , respectively.

Using these observations we get that

$$\begin{aligned} \mathbf{P}(\mathcal{E} \mid M = 1) &\leq \mathbf{P}\left(\left(\mathbf{S}^n(1), \hat{\mathbf{S}}^n(1)\right) \notin A_\epsilon^{(n)}\right) + \sum_{w>1} \mathbf{P}\left(\left(\mathbf{S}^n(w), \hat{\mathbf{S}}^n(1)\right) \in A_\epsilon^{(n)}\right) \\ &\leq \mathbf{P}\left(\left(\mathbf{S}^n(1), \hat{\mathbf{S}}^n(1)\right) \notin A_\epsilon^{(n)}\right) + \sum_{w>1} 2^{-n(I(S;\hat{S})-3\epsilon)} \\ &= \mathbf{P}\left(\left(\mathbf{S}^n(1), \hat{\mathbf{S}}^n(1)\right) \notin A_\epsilon^{(n)}\right) + (\lfloor 2^{nr} \rfloor - 1) 2^{-n(I(S;\hat{S})-3\epsilon)} \\ &\leq \mathbf{P}\left(\left(\mathbf{S}^n(1), \hat{\mathbf{S}}^n(1)\right) \notin A_\epsilon^{(n)}\right) + 2^{nr} 2^{-n(I(S;\hat{S})-3\epsilon)} \\ &= \mathbf{P}\left(\left(\mathbf{S}^n(1), \hat{\mathbf{S}}^n(1)\right) \notin A_\epsilon^{(n)}\right) + 2^{-n(I(S;\hat{S})-r-3\epsilon)}. \end{aligned}$$

Observe that we chose p_S as $\arg \max_p I(S; \hat{S})$, so $I(S; \hat{S}) = cap$, the channel's capacity. So, if the rate r is below the channel's capacity and ϵ is sufficiently small, then $\mathbf{P}(\mathcal{E} \mid M = 1) \rightarrow 0$ as $n \rightarrow \infty$. \square

10.5 Communication protocol for algorithm validation

We now formalize the communication protocol where posterior agreement originates. We first explain an *ideal variant* where we know X 's probability distribution p_X . Furthermore, we assume that p_X is from a parameterized family \mathfrak{P} of probability distributions. Afterwards, we explain the *empirical variant* where we do not know anything about p_X and, even worse, we only have *two observations* X' and X'' drawn from p_X . Recall that, in practice, we only have access to observations and have no information of the nature of the distribution where these observations came from.

10.5.1 Assumptions

For our statements to hold, we make the following assumptions.

Exponential solution space: We assume that \mathcal{C} is discrete and that $\log |\mathcal{C}| = \Theta(n)$, where n measures the “size” of an observation from a phenomenon. For example, n measures the number of edges in a graph or the number of data-points in a clustering instance. Intuitively, we assume that the solution space’s size grows exponentially in n .

Probabilistic outputs: We assume that any algorithm for stochastic optimization, when given an input X' from an instance space \mathcal{X} , outputs a distribution $p(\cdot | X')$ over a discrete solution space \mathcal{C} . Every algorithm can be thought to be as such, even when it only outputs a fixed value $c_{X'} \in \mathcal{C}$, when given X' as input. In this case, $p(c | X') = 1$ if $c = c_{X'}$ and 0 otherwise.

10.5.2 Ideal variant

Let \mathcal{A} be an algorithm intended for solving Problem 10.3. Posterior agreement originates from a communication protocol between a sender and a receiver. The sender must communicate, using \mathcal{A} ’s outputs, the nature of a phenomenon to the receiver. The receiver must be able, using the information sent from the sender, to identify the phenomenon. The communication from the sender to the receiver is done through a noisy channel, whose noise is defined by \mathcal{A} and the phenomenon.

Messages Fix $n \in \mathbb{N}$. Present the instance space \mathcal{X} to the sender and the receiver. Then agree draw at random a set $\mathcal{F} \subseteq \mathfrak{P}$ of m phenomena and, for each $p \in \mathcal{F}$, draw an observation X' of size n . Let $\mathcal{M} = \{X'_1, \dots, X'_m\}$. Agreeing on \mathcal{F} is often not possible in practice, so we explain in the empirical variant how to deal with this. This \mathcal{M} constitutes the set of possible messages that the sender may try to communicate to the receiver.

For the moment, we leave the value of m undefined. We leave for later to figure out what is the maximum value of m that we can use. The choice of m must still ensure that the probability of a communication error goes to 0, as $n \rightarrow \infty$.

Code Present algorithm \mathcal{A} to the sender and the receiver. For each $X' \in \mathcal{M}$, use \mathcal{A} to compute $p(\cdot | X')$. Define the code $(Enc_{\mathcal{A}}, Dec_{\mathcal{A}})$ as follows:

- $Enc_{\mathcal{A}}$ encodes $X' \in \mathcal{M}$ as the codeword $p(\cdot | X')$.
- $Dec_{\mathcal{A}}$ decodes a probability distribution $p(\cdot | Y)$ over \mathcal{C} as any $\hat{X} \in \mathcal{M}$ such that $k(Y, \hat{X}) \geq k(Y, X)$, for all $X \in \mathcal{M}$. Here,

$$k(Y, X') := \sum_c p(c | Y) p(c | X').$$

Channel The channel is the pair $(\{p(\cdot | X')\}_{X' \in \mathcal{M}}, \mathcal{P})$ where \mathcal{P} is defined by the following probabilistic procedure. Assume that the sender inputs $p(\cdot | X')$ to the channel. The channel then replaces X' with a fresh new observation X'' from the same phenomenon where X' comes from. Then it uses the algorithm \mathcal{A} to compute $p(\cdot | X'')$ and outputs that to the receiver.

Communication A message X' is selected uniformly at random from \mathcal{M} and without the receiver's knowledge. The sender then sends the codeword $p(\cdot | X')$ through the channel. The receiver gets $p(\cdot | X'')$ and uses the decoding function to guess which message the sender sent. The receiver succeeds by correctly guessing X' .

Observe how the algorithm and the phenomenon define the channel's noise. Hence, the algorithm can be evaluated by *the capacity* of the resulting channel. This capacity is measured by its maximum attainable rate, which is obtained by figuring out how to maximize the number m of messages that can be used in the protocol, while making the probability of a communication error go to zero as n , the size of the observations, go to infinity. We carry this analysis in the empirical variant.

10.5.3 Empirical variant

The capacity of the channel built above cannot be computed, as we often only have access to a set of observations and not to the phenomena behind them. This means that we cannot compute the set \mathcal{M} of messages, as described in the ideal variant. For this reason, we use an empirical variant, where we assume that we are given at least two observations X' and X'' of a same phenomenon.

Messages (Figure 10.8) We create a set of messages by producing *transformed copies* of X' . We take a set \mathbb{T} of transformations, where each transformation transforms instances into instances. Afterwards, we draw m transformations $\tau_1, \tau_2, \dots, \tau_m$ from \mathbb{T} uniformly at random. In this way, we can create

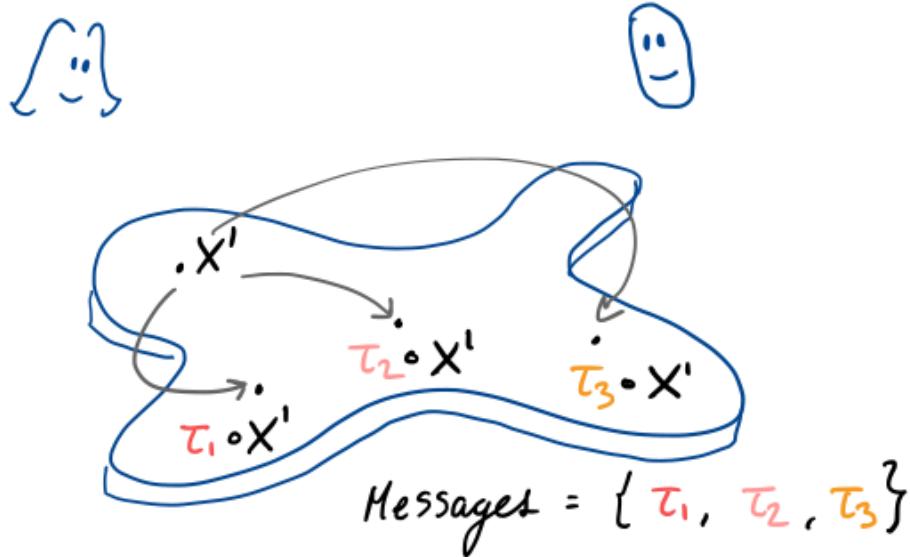


Figure 10.8

a set $\mathcal{M} = \{\tau_1 \circ X', \tau_2 \circ X', \dots, \tau_m \circ X'\}$ of messages that are analogous to the set of messages that we created in the ideal variant. We impose the following requirement on \mathbb{T} so that \mathcal{M} looks like the set \mathcal{M} that we would obtain in the ideal variant.

- $\sum_{\tau} p(c \mid \tau \circ X') \in \left[\frac{|\mathbb{T}|}{|\mathcal{C}|}(1 - \rho), \frac{|\mathbb{T}|}{|\mathcal{C}|}(1 + \rho) \right]$, for some small $\rho > 0$. The reason for this assumption is that we do not want the probability mass of all these codewords to be concentrated in a narrow subset of \mathcal{C} , as this reduces the number of different messages that the sender can communicate to the receiver. This can be ensured that, for each $c \in \mathcal{C}$, the total probability mass c gets is $\sum_{\tau} p(c \mid \tau \circ X') \approx \frac{|\mathbb{T}|}{|\mathcal{C}|}$.
- The transformations do not alter an instance's randomness.

Observe that such a set of transformations does not necessarily always exist. For example, if the algorithm under evaluation always produces the same constant distribution $p(\cdot \mid X')$, independent of X' , then no set of transformations will be able to achieve the requirement above. In this case, however, there is no need to build a channel to evaluate such an algorithm, as it is clear that the algorithm is not producing any value from the data.

Code (Figure 10.9) The encoding and decoding functions are the analogous of the ideal variant. The codeword of a message $\tau \in \mathcal{M}$ is $p(\cdot \mid \tau \circ X')$. When

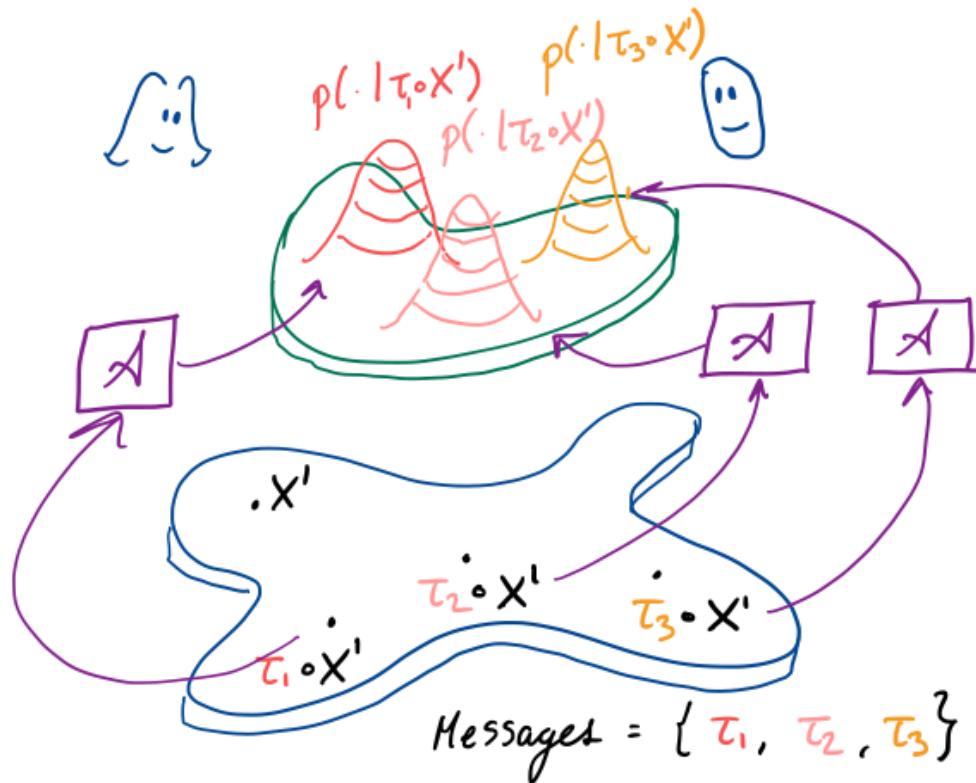


Figure 10.9

given a codeword $p(\cdot | Y)$, the decoding function outputs the message τ for which $k(Y, \tau \circ X') \geq k(Y, \sigma \circ X')$, for every $\sigma \in \mathcal{M}$.

Channel (Figure 10.10) When the sender inputs $p(\cdot | \tau \circ X')$ to the channel, the channel outputs $p(\cdot | \tau \circ X'')$, as in the ideal variant.

Protocol The sender and the receiver agree on a set \mathbb{T} of transformations and use the algorithm under evaluation \mathcal{A} to compute the code's encoding and decoding functions. A set $\mathcal{M} = \{\tau_1, \dots, \tau_m\}$ of m messages is drawn uniformly at random from \mathbb{T} . A message $\tau \in \mathcal{M}$ is selected uniformly at random and without the receiver's knowledge. The sender then sends the codeword $p(\cdot | \tau \circ X')$ through the channel. The receiver gets $p(\cdot | \tau \circ X'')$ and uses the decoding function to guess which message the sender sent (Figure 10.6). The receiver succeeds by correctly guessing τ .

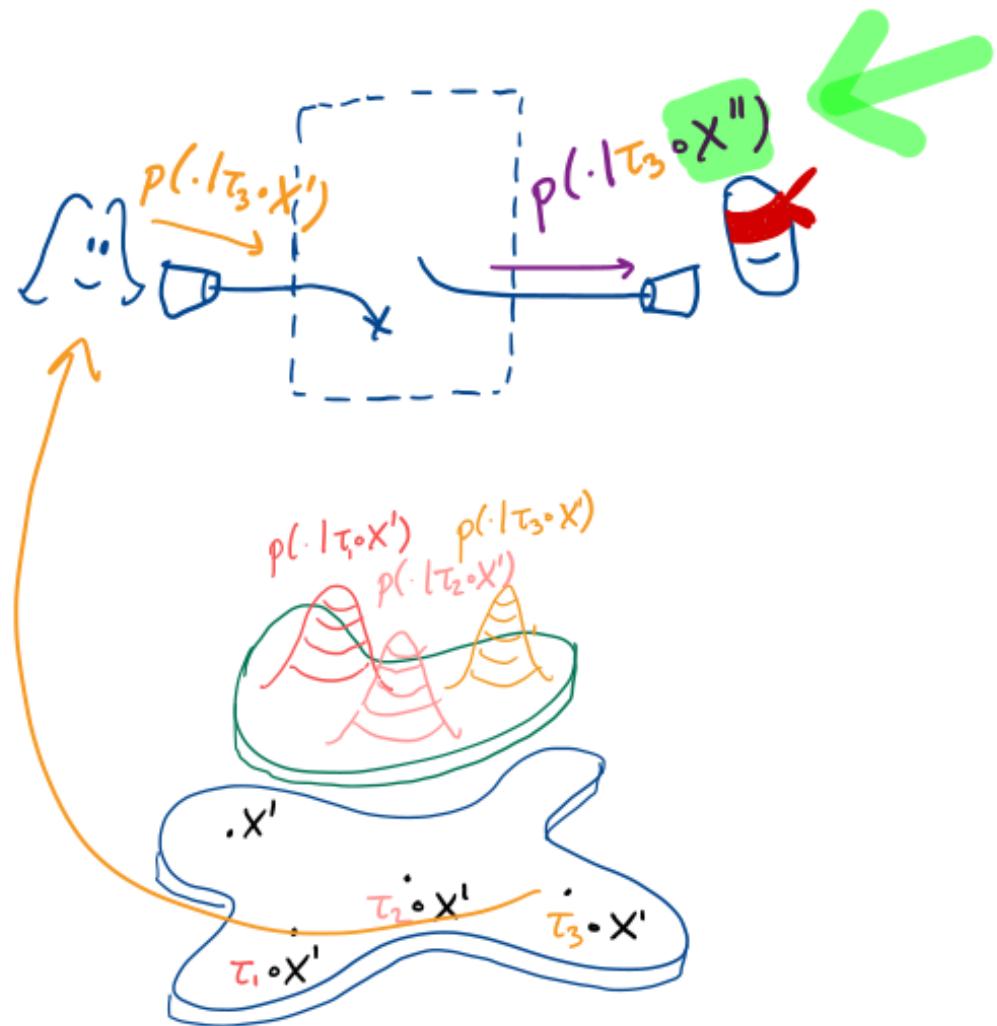


Figure 10.10

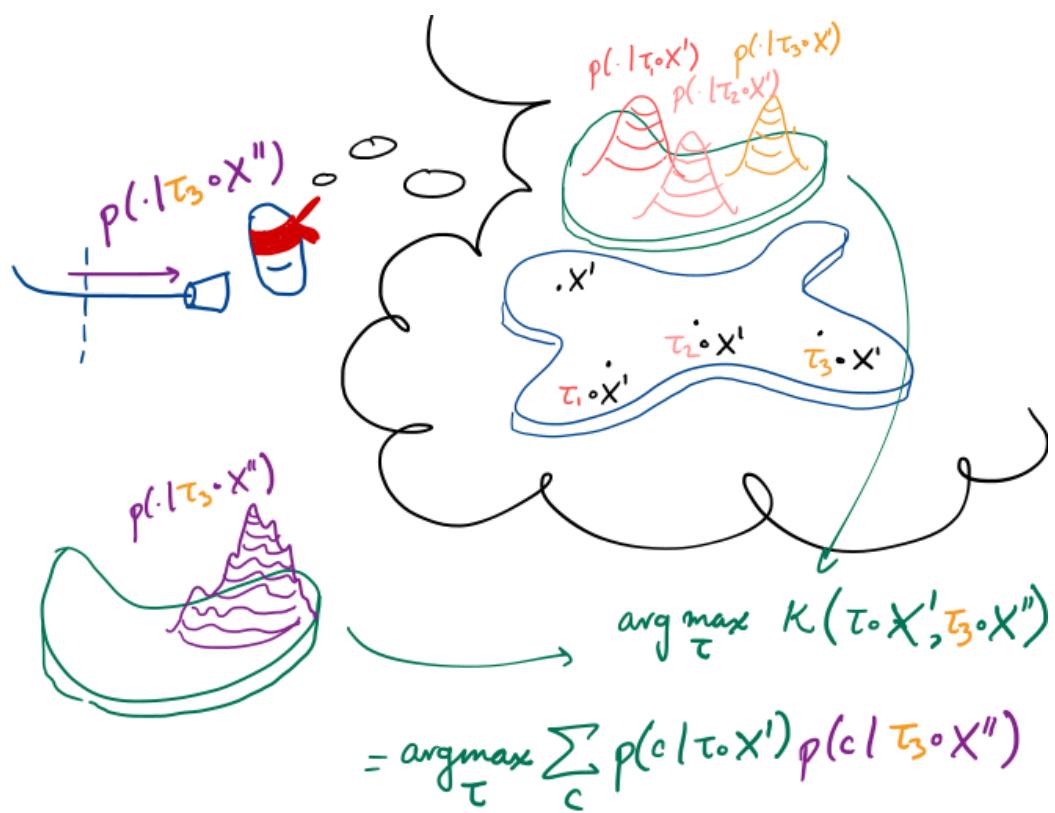


Figure 10.11

10.5.4 Probability of a communication error

Theorem 31. *The probability of a communication error is bounded above by*

$$P_{(n)} = \exp(-I + \log m + \epsilon \log |\mathcal{C}|), \quad (10.5)$$

where

$$I := \mathbb{E}_{X', X''} [\log(|\mathcal{C}| k(X', X''))] \quad (10.6)$$

One can show that $P_{(n)} \rightarrow 0$ as $n \rightarrow \infty$ by choosing ϵ sufficiently small and ensuring that $I - \log m = \Omega(n)$.

Proof. Let τ_s be the message sent by the sender and let $\hat{\tau}$ be the message guessed by the receiver. Just as in the proof of Theorem 30, one can show that $\mathbf{P}(\hat{\tau} \neq \tau_s) = \mathbf{P}(\hat{\tau} \neq id)$, where id is some arbitrary transformation. Without loss of generality, we assume that id is the identity transformation.

Using the definition of communication error, we get that

$$\mathbf{P}(\hat{\tau} \neq id) = \mathbf{P}\left(\max_{\tau \neq id} \kappa(\tau \circ X', X'') \geq \kappa(X', X'') \mid id\right).$$

Applying the union bound, we get

$$\mathbf{P}(\hat{\tau} \neq id) \leq \sum_{\tau \neq id} \mathbf{P}(\kappa(\tau \circ X', X'') \geq \kappa(X', X'') \mid id).$$

We now rewrite probabilities as expectations

$$\begin{aligned} \mathbf{P}(\hat{\tau} \neq id) &\leq \sum_{\tau \neq id} \mathbf{P}(\kappa(\tau \circ X', X'') \geq \kappa(X', X'') \mid id) \\ &= \sum_{\tau \neq id} \mathbb{E}_{X', X''} [\mathbb{E}_\tau [\mathbb{I}\{\kappa(\tau \circ X', X'') \geq \kappa(X', X'')\} \mid id, X', X'']] \\ &= \sum_{\tau \neq id} \mathbb{E}_{X', X''} [\mathbf{P}(\kappa(\tau \circ X', X'') \geq \kappa(X', X'') \mid id, X', X'')]. \end{aligned}$$

Here, \mathbb{I} is the indicator function.

We now apply Markov's inequality.

$$\begin{aligned} \mathbf{P}(\hat{\tau} \neq id) &\leq \sum_{\tau \neq id} \mathbb{E}_{X', X''} [\mathbf{P}(\mathbb{I}\{\kappa(\tau \circ X', X'') \geq \kappa(X', X'')\} \mid id, X', X'')] \\ &\leq \sum_{\tau \neq id} \mathbb{E}_{X', X''} \left[\frac{\mathbb{E}_\tau [\kappa(\tau \circ X', X'') \mid id, X', X'']}{\kappa(X', X'')} \right]. \end{aligned}$$

We now compute an upper bound for the numerator.

$$\begin{aligned}
\mathbb{E}_\tau [\kappa(\tau \circ X', X'') \mid X', X''] &= \mathbb{E}_\tau \left[\sum_c p(c \mid \tau \circ X') p(c \mid X'' \mid X', X'') \right] \\
&= \sum_c p(c \mid X'') \mathbb{E}_\tau [p(c \mid \tau \circ X')] \\
&= \sum_c p(c \mid X'') \sum_\tau \frac{1}{|\mathbb{T}|} p(c \mid \tau \circ X') \\
&\leq \frac{1}{|\mathbb{T}|} \frac{|\mathbb{T}|}{|\mathcal{C}|} (1 + \rho) \sum_c p(c \mid X'') = (1 + \rho) \frac{1}{|\mathcal{C}|}.
\end{aligned}$$

For the last inequality, we used the assumption that $\sum_\tau p(c \mid \tau \circ X') \in \left[\frac{|\mathbb{T}|}{|\mathcal{C}|} (1 - \rho), \frac{|\mathbb{T}|}{|\mathcal{C}|} (1 + \rho) \right]$.

We have then that

$$\begin{aligned}
\mathbf{P}(\hat{\tau} \neq id) &\leq \sum_{\tau \neq id} \mathbb{E}_{X', X''} \left[\frac{1 + \rho}{|\mathcal{C}| \kappa(X', X'')} \right] \\
&= (1 + \rho)(m - 1) \mathbb{E}_{X', X''} \exp(-\log(|\mathcal{C}| \kappa(X', X''))) \\
&\leq m(1 + \rho) \mathbb{E}_{X', X''} \exp(-\log(|\mathcal{C}| \kappa(X', X''))) \\
&= (1 + \rho) \mathbb{E}_{X', X''} \exp(-\log(|\mathcal{C}| \kappa(X', X''))) + \log m \\
&\approx \mathbb{E}_{X', X''} \exp(-\hat{I} + \log m).
\end{aligned}$$

Here, for simplicity, we assumed $1 + \rho \approx 1$.

To provide an upper bound to $\mathbf{P}(\hat{\tau} \neq id)$, we must bound the behavior of the random variable $\hat{I} = \log(|\mathcal{C}| \kappa(X', X''))$. To do this, we assume that \hat{I} satisfies *an asymptotic equipartition property* in the sense that, as $n \rightarrow \infty$, $\hat{I} \rightarrow I$, where $I = \mathbb{E}_{X', X''} \log(|\mathcal{C}| \kappa(X', X''))$, the expected log posterior agreement. Under this assumption, for every $\epsilon, \delta > 0$, there is $n_0 \in \mathbb{N}$ such that for any $n > n_0$,

$$\mathbf{P} \left(|\hat{I} - I| \leq \epsilon \log |\mathcal{C}| \right) > 1 - \delta.$$

ToDo: Explain why this assumption is reasonable. In particular, explain its connection to typicality.

With this assumption, we can derive the following:

$$\begin{aligned}
\mathbf{P}(\hat{\tau} \neq id) &\leq \mathbb{E}_{X',X''} \exp(-\hat{I} + \log m) \\
&= \mathbb{E}_{X',X''} \exp(-\hat{I} + (I - I) + \log m) \\
&= \mathbb{E}_{X',X''} \exp(-I + (I - \hat{I}) + \log m) \\
&\leq \mathbb{E}_{X',X''} \exp(-I + |I - \hat{I}| + \log m) \\
&\leq \mathbb{E}_{X',X''} \exp(-I + \epsilon \log |\mathcal{C}| + \log m) \\
&= \exp(-I + \epsilon \log |\mathcal{C}| + \log m).
\end{aligned}$$

This concludes the proof. \square

Recall that the goal is to be able to maximize the number of distinguishable messages that can be sent through the channel. Hence, we must aim to make both m and I as large as possible. The algorithm can only influence I and, therefore, good algorithms *shall maximize the expected log posterior agreement*.

Computing I requires the underlying distribution of X' and X'' , which we assume to be unknown. In this case, we can approximate I with *the empirical log posterior agreement*

$$\frac{1}{L} \sum_{\ell \leq L} [\log(|\mathcal{C}| k(X'_\ell, X''_\ell))], \quad (10.7)$$

where $\{X'_1, X''_1, \dots, X'_L, X''_L\}$ is a set of observations.

Finally, we remark some analogies with Shannon's channel coding theorem. The quantity $\frac{1}{n} \mathbb{E}_{X',X''} \log(|\mathcal{C}| \kappa(X', X''))$ plays the role the input-output mutual information. The value $\log m/n$ plays the role of the code rate.

Bibliography

- [1] J. M. Buhmann. Information theoretic model validation for clustering. In *2010 IEEE international symposium on information theory*, pages 1398–1402. IEEE, 2010.
- [2] L. Busse, M. H. Chehreghani, and J. M. Buhmann. Approximate sorting. In *German Conference on Pattern Recognition*, pages 142–152. Springer, 2013.
- [3] C. Cotrini, L. Corinzia, T. Weghorn, and D. Basin. The next 700 policy miners: A universal method for building policy miners. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 95–112, 2019.
- [4] T. M. Cover. *Elements of information theory*. John Wiley & Sons, 1999.
- [5] M. Frank, A. P. Streich, D. Basin, and J. M. Buhmann. A probabilistic approach to hybrid role mining. In *Proceedings of the 16th ACM conference on Computer and communications security*, pages 101–111, 2009.
- [6] N. S. Gorbach, M. Tittgemeyer, and J. M. Buhmann. Pipeline validation for connectivity-based cortex parcellation. *NeuroImage*, 181:219–234, 2018.
- [7] A. Gronskiy. *Statistical Mechanics and Information Theory in Approximate Robust Inference*. PhD thesis, ETH Zurich, 2018.
- [8] A. Gronskiy and J. M. Buhmann. How informative are minimum spanning tree algorithms? In *2014 IEEE International Symposium on Information Theory*, pages 2277–2281. IEEE, 2014.
- [9] E. T. Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

- [10] E. T. Jaynes. Information theory and statistical mechanics. II. *Physical review*, 108(2):171, 1957.
- [11] K. Rose. *Deterministic annealing, clustering, and optimization*. PhD thesis, California Institute of Technology, 1991.
- [12] K. Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, 86(11):2210–2239, 1998.
- [13] V. Wegmayr and J. M. Buhmann. Entrack: Probabilistic spherical regression with entropy regularization for fiber tractography. *International Journal of Computer Vision*, pages 1–25, 2020.

Appendix A

Information theory

To be done in 2022.

Appendix B

The EM-algorithm

B.1 Introduction

We motivate and present the EM (expectation-maximization) algorithm, an algorithm used for approximately computing parameter values for probability distributions in maximum-likelihood estimation. The reader is expected to have knowledge of undergraduate probability theory and to be familiar with maximum-likelihood estimation.

The presentation is divided in three parts. Section B.2 presents an optimization problem regarding a vegan flea. We present an algorithm that finds an approximate solution and provides intuitions to understand why the EM-algorithm works. Section B.3 presents the problem of building a simple movie recommendation system. We show that movie ratings can be understood as samples from a probabilistic model that is defined by a set of multivariate Bernoulli distributions. Estimating the parameters of these distributions via maximum-likelihood turns out to be very hard using analytical methods. In Section B.4, we show that this maximum-likelihood estimation problem is an instance of the vegan-flea optimization problem and derive the EM-algorithm from the approximation algorithm presented in Section B.2.

B.2 The vegan-flea optimization problem

B.2.1 The dog

We introduce a two-dimensional dog, depicted in Figure B.1a. Although, in practice, dogs are three-dimensional entities, a two-dimensional dog makes easier the presentation of some ideas later. Observe that this two-dimensional

dog has only two legs, since two legs suffice to keep balance, and one eye, since there is no need for perspectives in a two-dimensional space.

Figure B.1b shows some of the dog's cardiovascular system. Observe that a blood vessel of a two-dimensional being does not have the shape of a cylinder. They still naturally expand when a surge of blood flows from a heart's pump. For the rest of these notes we focus only on a small area of this figure; namely, the tiny green square shown in the figure.

Figure B.1c shows the area marked by the green square in detail. We have placed some Cartesian axes there for reference. There is a flea at coordinates (0.25, 6). The dog's skin is the brown curve, and the upper border of a blood vessel is the red curve. Observe that, in a two-dimensional space, skins are lines instead of surfaces.

B.2.2 The skin and the blood vessel's upper border

We now mathematically model the skin and the blood vessel's upper border. Let $\text{skin} : [0, 1] \times [0, \infty) \rightarrow \mathbb{R}$, and $\text{vessel} : [0, 1] \times [0, \infty) \rightarrow \mathbb{R}$ be two functions. For $t \in [0, \infty)$, let $\text{skin}(\cdot, t) : [0, 1] \rightarrow \mathbb{R}$ be the function that maps x to $\text{skin}(x, t)$. We define the function $\text{vessel}(\cdot, t)$ analogously. Intuitively, for $t \in [0, \infty)$, the functions $\text{skin}(\cdot, t)$ and $\text{vessel}(\cdot, t)$ describe the skin and the blood vessel's upper border at time t , as depicted in Figure B.2. Hence, $\text{skin}(x, t) \geq \text{vessel}(x, t)$, for any $(x, t) \in [0, 1] \times [0, \infty)$. Observe that $\text{skin}(\cdot, t)$ and $\text{vessel}(\cdot, t)$ vary with t . This is to model the fact that blood flows through the vessel and, consequently, makes the skin surface and the blood vessel's upper border vary with time. The animated .gif file attached with these notes¹ illustrate the setting. We strongly encourage the reader to look the .gif file before proceeding.

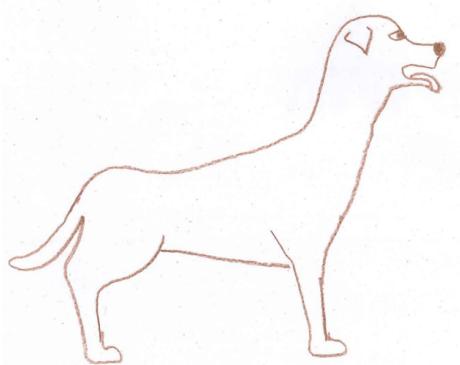
Assumption 1. *We assume that for any $x \in [0, 1]$ and any two time points $t_1, t_2 \in [0, \infty)$, $\text{skin}(x, t_1) - \text{vessel}(x, t_1) = \text{skin}(x, t_2) - \text{vessel}(x, t_2)$.*

This assumption states that the skin surface changes by the same amount that the blood vessel's upper border changes. This allows us to define a function d such that $d(x) = \text{skin}(x, t) - \text{vessel}(x, t)$, for any $x \in [0, 1]$ and $t \in [0, \infty)$.

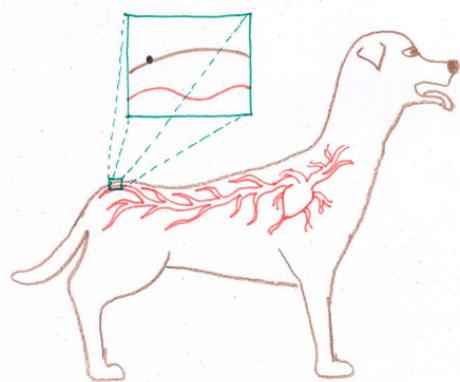
Blood flows periodically through the vessel and therefore makes the vessel shape change. Moreover, we assume the following:

Assumption 2. *For any $x \in [0, 1]$ and any $t \in [0, \infty)$, there is $t' \geq t$ such that $\text{vessel}(x, t') = \max_{x'} \text{vessel}(x', t')$.*

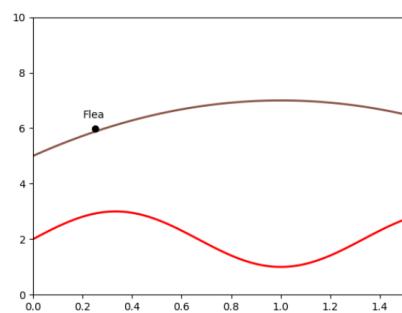
¹https://people.inf.ethz.ch/ccarlos/assets/em/vegan_flea.gif



(a) A two-dimensional dog.



(b) A part of the dog's (two-dimensional) cardiovascular system.



(c) The flea, the skin, and the upper border of a blood vessel.

Figure B.1

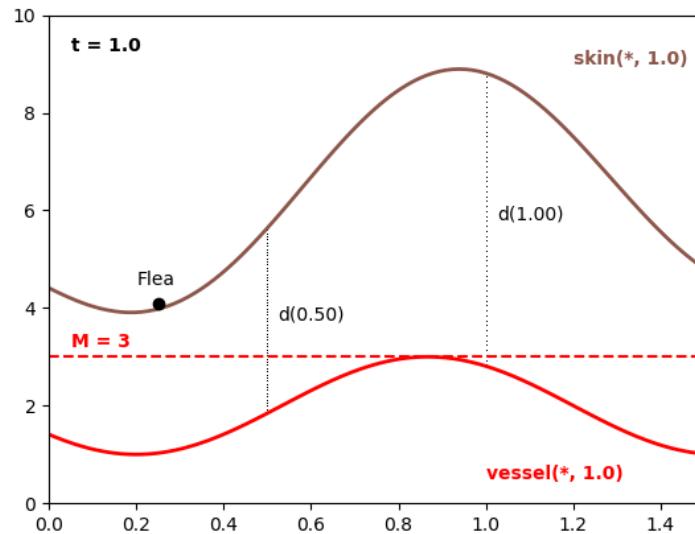
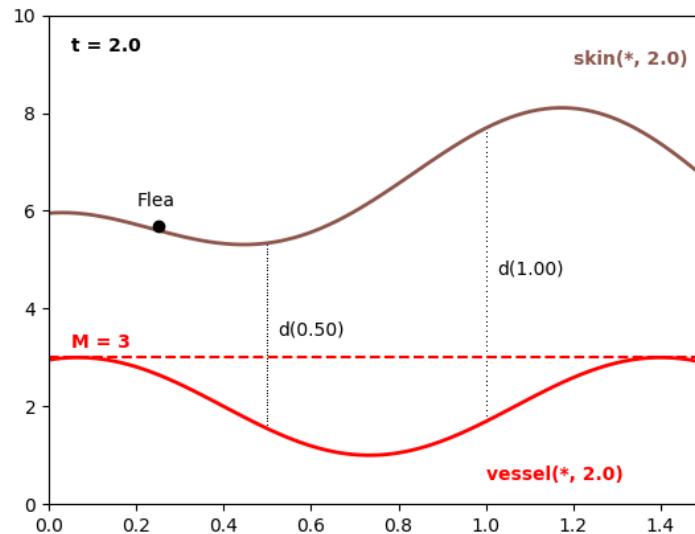
(a) Skin and blood vessel at $t = 1$ (b) Skin and blood vessel at $t = 2$

Figure B.2: An illustration of the functions $\text{skin}(\cdot, t)$ and $\text{vessel}(\cdot, t)$, for $t \in \{1, 2\}$.

If you observe the .gif animation, you can see that we defined a constant M . You can also see that, for any x and any t , $vessel(x, t) \leq M$ and that there is $t' \geq t$ such that $vessel(x, t') = M$. We could then make Assumption 2 stronger by stating that, for any $x \in [0, 1]$ and any $t \in [0, \infty)$, $vessel(\cdot, t)$ is bounded by M and that there is $t' \geq t$ such that $vessel(x, t) = M$. However, Assumption 2 is enough for our purposes. Moreover, it is weaker and, hence, more general.

B.2.3 The vegan flea

Imagine now that there is a flea resting on the skin surface at $(x_0, skin(x_0, 0))$, for some $x_0 \in [0, 1]$. The flea has decided to become vegan and wishes to be as far away from the blood vessel as possible, to avoid the temptation of the blood. More precisely, the flea's goal is the following.

Objective 1. Compute a value x^* that maximizes d .

A look at the .gif file shows that $x^* = 1.0$. This is easy for us as we, three-dimensional creatures, have an omniscient view of the flea's universe. The flea, however, cannot see that as she knows nothing about $vessel$. In spite of this, we illustrate how the flea can partially achieve its objective.

We make two assumptions about the flea's computation abilities.

Assumption 3. For any $t \in [0, \infty)$, the flea can efficiently compute

$$x^* = \arg \max_{x'} skin(x, t).$$

This assumption bases on the idea that the flea can see the dog's skin and can therefore maximize $skin(\cdot, t)$. The next assumption states that the flea, located at $(x_0, (skin(x_0, 0)))$, can identify the moment t' when $vessel(x_0, t') = M$, the maximum of $vessel(\cdot, t)$.

Assumption 4. For any $x \in [0, 1]$ and any $t \in [0, \infty)$, the flea can efficiently compute some $\hat{t} \geq t$ such that $vessel(x, \hat{t}) = \max_{x'} vessel(x', \hat{t})$.

We give some justification for this assumption. Blood flows through the vessel in a periodic way and $skin(\cdot, t)$ changes in the same way as $vessel(\cdot, t)$ does. Hence, the flea can learn the blood pulse and then wait for a time \hat{t} where $vessel(x_0, \hat{t}) = M$.

B.2.4 An approximate maximization algorithm

We describe a strategy by which the flea can compute a value x^* such that $d(x^*) \geq d(x_0)$, where $(x_0, \text{skin}(x_0, 0))$ is the flea's current position.

[**E-step**] The flea waits for a time \hat{t} at which $\text{vessel}(x_0, \hat{t}) = \max_x \text{vessel}(x, \hat{t})$ (Figure B.3a). This is possible by Assumption 4.

[**M-step**] At time \hat{t} , the flea computes a value x^* such that

$$x^* = \arg \max_x \text{skin}(x, \hat{t}).$$

(Figure B.3b). This is possible by Assumption 3.

[**Move**] The flea moves to $(x^*, \text{skin}(x^*, \hat{t}))$ (Figure B.3c).

Figure B.4 illustrates why $d(x^*) \geq d(x_0)$. Observe that $\text{skin}(x^*, \hat{t}) \geq \text{skin}(x_0, \hat{t})$, since x^* is a maximum of $\text{skin}(\cdot, \hat{t})$. Observe also that $\text{vessel}(x^*, \hat{t}) \leq \text{vessel}(x_0, \hat{t})$, since $\text{vessel}(x_0, \hat{t})$ is a maximum of $\text{vessel}(\cdot, \hat{t})$. Hence, $d(x^*) = \text{skin}(x^*, \hat{t}) - \text{vessel}(x^*, \hat{t}) \geq \text{skin}(x_0, \hat{t}) - \text{vessel}(x_0, \hat{t}) = d(x_0)$.

Notice that the flea can set $x_0 = x^*$ and repeat this procedure to find another value x^{**} such that $d(x^{**}) \geq d(x^*)$. The flea can repeat this process as long as the computed values increase d .

We summarize these insights into Algorithm 4, which computes a sequence of values x_0, x_1, \dots such that $d(x_0) \leq d(x_1) \leq \dots$ Observe that this algorithm converges.

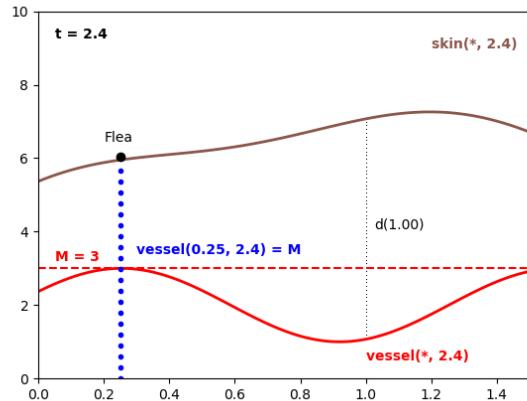
One can also relax the assumptions above so that the algorithm works even when vessel and skin 's domains are a product $\mathcal{X} \times \mathcal{T}$ of any two sets \mathcal{X} and \mathcal{T} .

With an argumentation similar to the one above, one can prove that $d(x^*)$ never decreases between two iterations of Algorithm 4's loop.

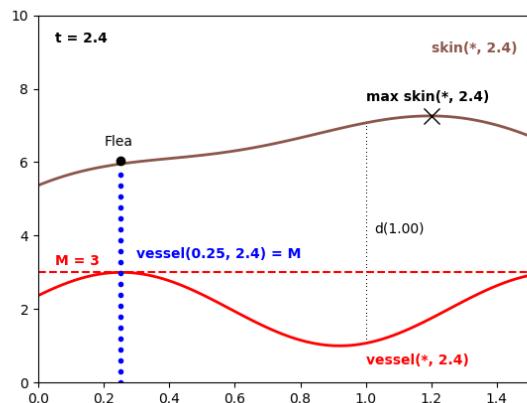
B.3 Building a movie-recommendation system with a mixture of multivariate Bernoulli distributions

Table B.1 shows the ratings that 10 (fictitious) individuals gave to 6 popular movies. To keep it simple, we assume only binary ratings (good or bad). After a close look to the ratings, the reader can see that Alice and Bob have the exact same taste for movies: *sci-fi* movies. The next four people have

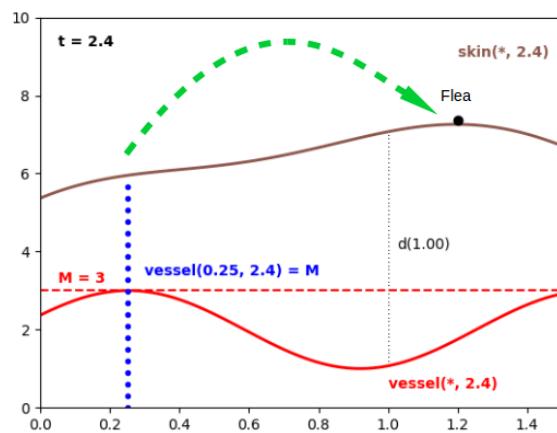
B.3. BUILDING A MOVIE-RECOMMENDATION SYSTEM WITH A MIXTURE OF MULTIVARIABLES



(a) Step 1.



(b) Step 2.



(c) Step 3.

Figure B.3

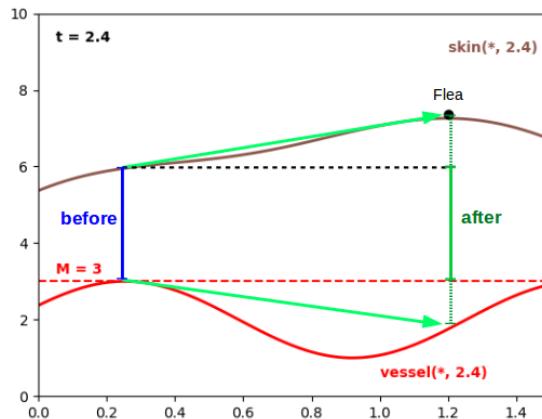


Figure B.4

	Star Wars	Star Trek	Titanic	Pretty Woman	007	Mission Impossible
Alice	✓	✓	✗	✗	✗	✗
Bob	✓	✓	✗	✗	✗	✗
Carlos	✗	✗	✓	✓	✗	✗
David	✗	✗	✓	✓	✗	✗
Ellen	✗	✗	✓	✗	✗	✓
Fabian	✗	✗	✓	✓	✗	✗
Gabriel	✓	✓	✗	✗	✓	✓
Hector	✓	✗	✗	✗	✓	✓
Ian	✓	✓	✗	✗	✓	✓
Zelya	✓	✗	✓	✗	✗	✓
John	?	?	?	✓	?	?

Table B.1: Ratings from 10 individuals for 6 movies. According to the table, everyone who likes Pretty Woman also liked Titanic. Therefore, it is likely that John would also like Titanic.

Algorithm 4

Require: \mathcal{X} and \mathcal{T} two sets and real functions d , $vessel$, and $skin$ satisfying the following.

A1 $d(x) = skin(x, t) - vessel(x, t)$, for any $x \in \mathcal{X}$ and $t \in \mathcal{T}$.

A2 For any $x \in \mathcal{X}$, one can efficiently compute $\hat{t} \in \mathcal{T}$ such that $vessel(x, \hat{t}) = \max_{x'} vessel(x', \hat{t})$.

A3 For any $t \in \mathcal{T}$, one can efficiently compute $\arg \max_x skin(x, t)$.

```

1: function DISTANCEMAX( $vessel : \mathcal{X} \times \mathcal{T} \rightarrow \mathbb{R}$ ,  $skin : \mathcal{X} \times \mathcal{T} \rightarrow \mathbb{R}$ )
2:   Choose any  $x_0 \in \mathcal{X}$ .
3:   for  $i = 0, 1, \dots$  do
4:     [E-step] Compute  $\hat{t} \in \mathcal{T}$  s.t.  $vessel(x_0, \hat{t}) = \max_x vessel(x, \hat{t})$ .
5:     [M-step] Compute  $x^* = \arg \max_x skin(x, \hat{t})$ .
6:     Print  $x^*$  and  $d(x^*)$ .
7:      $x_0 \leftarrow x^*$ .
8:   end for
9: end function
```

a strong interest for *romantic* movies. The next three people like sci-fi and action movies. Zelya's tastes seem to be different from everyone else.

Consider now John, he likes Pretty Woman, but does not like Star trek. He has not seen any of the other movies. What movie could we recommend to him? Since he likes Pretty Woman and everyone who liked Pretty Woman also liked Titanic, we can recommend him to watch Titanic.

From all 2^6 combinations of ratings, the table contains only a few of them. Moreover, a large majority of the people in the table seem to belong to one of very few taste categories: sci-fi, romantic, or sci-fi+action. Real life is not so different: a large majority of people can be partitioned into very few categories and people within a same category have very similar preferences. To recommend a movie to someone, we estimate the category where this person belongs and then search for a movie that people in this category liked.

B.3.1 Movie ratings as samples from probability distributions

We can view Table B.1 as the result of a sampling process. Initially, the table was empty and then one person at random appeared (Alice, in our case) and filled the first row of the table. Then Bob appeared and so on. To sample

the film ratings of one person, we first *sample the category* where this person belongs and then, for each movie, we *sample the rating* this person gave to that movie, conditioned on the person belonging to the sampled category. This sampling process is then defined by the following probability distributions:

- A distribution over categories.
- For each category and each movie, a distribution defining the probability that a person in the category likes the movie.

From these two distributions, we build a new probability distribution with which we can answer the following question: *if a person watched and liked movies m_1, m_2, \dots, m_k , how likely is that she will like a movie m' that she has not seen?* This probability distribution constitutes then our recommendation system. To decide which movie to recommend, we take the ratings the person has given to previously watched movies. Then, for each movie in the database she has not seen, we compute the probability that she likes that movie. Finally, we recommend the movie that she will most likely like.

We first formalize the two distributions mentioned above. Suppose we have K categories and D movies. We identify categories with the numbers $1, 2, \dots, K$ and movies with the numbers $1, 2, \dots, D$. We can model the distribution of K categories using a discrete distribution with a set $\{\nu_1, \nu_2, \dots, \nu_K\}$ of K parameters that add up to 1. For the k -th category and the movie j , we define a value μ_{kj} indicating what the probability is that a person in the k -th category likes movie j . We leave the values ν_k and μ_{kj} , for $k \leq K$ and $j \leq D$, undefined for the moment.

Having defined these distributions, we can now assign a probability to the ratings a person gave to all movies in the database. We model these ratings with a vector $x \in \{0, 1\}^D$, where x_j , for $j \leq D$, indicates whether the movie was rated good ($x_j = 1$) or bad ($x_j = 0$). We leave as an exercise to show that the probability $p(x)$ of a vector $x \in \{0, 1\}^D$ is as follows:

$$p(x) = \sum_{k \leq K} \left(\nu_k \prod_{j \leq D} \mu_{kj}^{x_j} (1 - \mu_{kj})^{1-x_j} \right). \quad (\text{B.1})$$

We can now define probabilities for movie-rating databases. We model a movie-rating database with a matrix $X \in \{0, 1\}^{N \times D}$. Each row X_i , for $i \leq N$, represents a person and each entry $X_{i,j}$, for $j \leq D$, represents the rating person i gave to movie j . Assuming that the ratings of two different people are

B.3. BUILDING A MOVIE-RECOMMENDATION SYSTEM WITH A MIXTURE OF MULTIVARIATE NORMAL DISTRIBUTIONS

independent, we can show that the probability $p(X)$ is given by the following.

$$p(X) = \prod_{i \leq N} p(X_i) = \prod_{i \leq N} \sum_{k \leq K} \left(\nu_k \prod_{j \leq D} \mu_{kj}^{X_{i,j}} (1 - \mu_{kj})^{1-X_{i,j}} \right). \quad (\text{B.2})$$

B.3.2 Maximum-likelihood estimation

We now choose values for ν_k and μ_{kj} , for $k \leq K$ and $j \leq D$. Here, we use *maximum-likelihood estimation*, which argues that the best values for our parameters are those that maximize $p(X)$, for X the movie database we have. For computational reasons, one searches instead for the parameters that maximize $\log p(X)$. Using basic logarithm properties, we can show that

$$\log p(X) = \sum_{i \leq N} \log \left(\sum_{k \leq K} \left(\nu_k \prod_{j \leq D} \mu_{kj}^{x_j} (1 - \mu_{kj})^{1-x_j} \right) \right). \quad (\text{B.3})$$

The value $\log p(X)$ is called X 's *log likelihood*.

Finding the parameter values that maximize this log likelihood is difficult, even with approximation methods. Nonetheless, it is possible to find a set of parameter values that locally maximize the log likelihood, using Algorithm 4. To do this, we introduce a new set $\mathbf{Z} = \{\mathbf{Z}(i) \mid i \leq N\}$ of random variables. For $i \leq N$, $\mathbf{Z}(i)$ indicates to which category person i belongs. Assume for a moment that, in addition of X , we also know, for $i \leq N$, the category $Z(i)$ where person i belongs. One can show that (X, Z) 's log likelihood is given by the following.

$$\log p(X, Z) = \sum_{i \leq N} \begin{pmatrix} \log \nu_{Z(i)} + \\ x_{i,Z(i)} \log \mu_{Z(i),j} + \\ (1 - x_{i,Z(i)}) \log (1 - \mu_{Z(i),j}) \end{pmatrix}. \quad (\text{B.4})$$

The log likelihood of (X, Z) is much easier to maximize with respect to the parameters than X 's log likelihood; it can be maximized using standard calculus. However, observe that the movie database does not tell us to which category each person belongs, so we do not know Z and there is no clear way how to obtain it.

It is common that log likelihood maximization problems become easier when we introduce additional random variables to the probabilistic model. It is also common that the values that such random variables take are not available. For these situations, the EM-algorithm was proposed.

B.4 Derivation of the EM-algorithm

We generalize the problem we addressed in the previous section. Let \mathbf{X} and \mathbf{Z} be random variables and let X be an observed value for \mathbf{X} . Assume that the joint pdf $p(\cdot, \cdot | \theta)$ for (\mathbf{X}, \mathbf{Z}) is parameterized by θ , which can take values in Θ . Our goal is to compute

$$\arg \max_{\theta \in \Theta} \log p(X | \theta). \quad (\text{B.5})$$

Assume now that it is preferable to work with $\log p(X, Z)$ than with $\log p(X)$, for any Z in \mathbf{Z} 's range. If we knew the value Z that \mathbf{Z} took when we obtained $\mathbf{X} = X$, then we could state that

$$\log p(X | \theta) = \log p(X, Z | \theta) - \log p(Z | X, \theta). \quad (\text{B.6})$$

This identity follows from the definition of conditional pdfs. We can make Z irrelevant by computing expectations on both sides with respect to some pdf \tilde{p} for \mathbf{Z} . We leave for later the problem of defining \tilde{p} .

$$\begin{aligned} \int \tilde{p}(Z) \log p(X | \theta) dZ &= \int \tilde{p}(Z) \log p(X, Z | \theta) dZ - \int \tilde{p}(Z) \log p(Z | X, \theta) dZ \\ &= \mathbb{E}_{\tilde{p}(\mathbf{Z})} [\log p(X, \mathbf{Z} | \theta)] - \mathbb{E}_{\tilde{p}(\mathbf{Z})} [\log p(\mathbf{Z} | X, \theta)]. \end{aligned}$$

Observe that $\log p(X | \theta)$ does not depend on Z or \tilde{p} . Therefore, the left-hand side equals $\log p(X | \theta)$. As a result,

$$\log p(X | \theta) = \mathbb{E}_{\tilde{p}(\mathbf{Z})} [\log p(X, \mathbf{Z} | \theta)] - \mathbb{E}_{\tilde{p}(\mathbf{Z})} [\log p(\mathbf{Z} | X, \theta)]. \quad (\text{B.7})$$

We now show that the maximization problem in Equation B.5 is an instance of the vegan-flea problem. Let $\mathcal{X} = \Theta$ and \mathcal{T} be the set of all pdfs for \mathbf{Z} . For $\theta \in \Theta$ and $\tilde{p} \in \mathcal{T}$, let

$$\begin{aligned} d(\theta) &= \log p(X | \theta) \\ \text{skin}(\theta, \tilde{p}) &= \mathbb{E}_{\tilde{p}(\mathbf{Z})} [\log p(X, \mathbf{Z} | \theta)] \\ \text{vessel}(\theta, \tilde{p}) &= \mathbb{E}_{\tilde{p}(\mathbf{Z})} [\log p(\mathbf{Z} | X, \theta)]. \end{aligned}$$

We now derive sufficient conditions for the assumptions [A1], [A2], and [A3] to hold.

A1 This assumption follows from Equation B.7, so no condition is necessary.

A2 In our case, this assumption means the following: for any $\theta \in \Theta$, one can efficiently compute $\tilde{p} \in \mathcal{T}$ such that for any $\theta' \in \Theta$,

$$\mathbb{E}_{\tilde{p}(\mathbf{Z})} [\log p(\mathbf{Z} | X, \theta)] \geq \mathbb{E}_{\tilde{p}(\mathbf{Z})} [\log p(\mathbf{Z} | X, \theta')] . \quad (\text{B.8})$$

We can fulfill this inequality by setting $\tilde{p}(\mathbf{Z}) = p(\mathbf{Z} | X, \theta)$. This follows from Gibbs's inequality, which states that for any two pdfs p and q for a random variable \mathbf{Z} ,

$$\mathbb{E}_p(\mathbf{Z}) [\log p(\mathbf{Z})] \geq \mathbb{E}_p(\mathbf{Z}) [\log q(\mathbf{Z})] . \quad (\text{B.9})$$

Hence, for **[A2]** to hold, we require the pdf $p(\mathbf{Z} | X, \theta)$ to be efficiently computable.

A3 This assumption requires that, for any $\tilde{p} \in \mathcal{T}$, we can efficiently compute

$$\arg \max_{\theta \in \Theta} \mathbb{E}_{\tilde{p}(\mathbf{Z})} [\log p(X, \mathbf{Z} | \theta)] .$$

In summary, to apply Algorithm 4 to compute $\arg \max_{\theta} \log p(X | \theta)$, we require the following.

AE1 One can efficiently compute the pdf $p(\mathbf{Z} | X, \theta)$.

AE2 One can efficiently compute $\arg \max_{\theta \in \Theta} \mathbb{E}_{\tilde{p}(\mathbf{Z})} [\log p(X, \mathbf{Z} | \theta)]$, for any pdf \tilde{p} for \mathbf{Z} .

Instantiating Algorithm 4 to our particular problem, we obtain *the EM algorithm*.

Algorithm 5

Require:

- Θ a set of parameters.
- A joint pdf $p(\mathbf{X}, \mathbf{Z} | \theta)$ over two random variables \mathbf{X} and \mathbf{Z} , governed by a parameter θ that ranges over Θ .
- A value X in \mathbf{X} 's range.

AE1 One can efficiently compute the pdf $p(\mathbf{Z} | X, \theta)$.**AE2** One can efficiently compute $\arg \max_{\theta \in \Theta} \mathbb{E}_{\tilde{p}(\mathbf{Z})} [\log p(X, \mathbf{Z} | \theta)]$, for any pdf \tilde{p} for \mathbf{Z} .

```

1: function EM( $X, p(\mathbf{X}, \mathbf{Z} | \theta), \Theta$ )
2:   Choose any  $\theta_0 \in \Theta$ .
3:   for  $i = 0, 1, \dots$  do
4:     [E-step] Compute  $p(\mathbf{Z} | X, \theta_i)$ .
5:     [M-step] Compute  $\theta_{i+1} = \arg \max_{\theta} \mathbb{E}_{p(\mathbf{Z}|X,\theta_i)} [\log p(X, \mathbf{Z} | \theta)]$ .
6:     Print  $\theta_{i+1}$  and  $\log p(X | \theta_{i+1})$ .
7:   end for
8: end function

```
