Fundamental tradeoffs in estimation of finite-state hidden Markov models

Justin Le and Pushkin Kachroo

April 12, 2018

Abstract

Hidden Markov models (HMMs) constitute a broad and flexible class of statistical models that are widely used in studying processes that evolve over time and are only observable through the collection of noisy data. Two problems are essential to the use of HMMs: state estimation and parameter estimation. In state estimation, an algorithm estimates the sequence of states of the process that most likely generated a certain sequence of observations in the data. In parameter estimation, an algorithm computes the probability distributions that govern the time-evolution of states and the sampling of data. Although algorithms for the two problems are widely researched, relatively little study has been devoted to understanding the tradeoffs between key design variables of these algorithms from a mathematically rigorous viewpoint. In this article, we provide such a study by establishing theorems regarding these tradeoffs. Furthermore, we illustrate the implications of these theorems in practice, highlighting the scope of their applicability and generality. We then suggest directions for future research in this area by bringing attention to the critical assumptions and tools used in the proofs of our theorems.

1 Introduction

In the design of systems for control, signal processing, communications, and many other purposes, a ubiquitous problem arises in which time-evolving processes must be estimated using large amounts of sensor data. In this work, we consider the hidden Markov model (HMM) as a framework for this problem. The HMM consists of two time-evolving states: the hidden state and observable state. The hidden state, which we herein refer to as simply the *state*, represents the underlying real-world process to be estimated. The observable state, which we herein refer to as the *emission*, represents the output of a sensor measuring the underlying process. The state evolves at each moment in time according to a *transition probability*, while the emission is generated by the state at each moment in time according to an *emission probability*.

The probability distributions governing transitions and emissions must be computed by an algorithm, and the process of computing these distributions is referred to as *learning*. The outputs of the learning algorithm are collectively referred to as the *HMM parameters*, and we can thus interchangeably refer to learning as *parameter estimation*. After learning, another algorithm applies the learned parameters to a sequence of emissions in order to estimate the sequence of states that most likely generated this emission-sequence, a process referred to as *decoding*. It is the decoding algorithm that produces an estimate of the real-world process's states, and we can thus interchangeably refer to decoding as *state estimation*. Hence, when modeling with the HMM, practitioners face two algorithmic problems of estimation: parameter estimation and state estimation.

The accuracy of decoding inevitably relies on the effectiveness of the learning algorithm in producing parameters that are representative of the true probability distributions governing the behavior of the real-world process and its sensors. Realistic examples of temporal sequences of states and emissions must be provided to the learning algorithm in order to compute distribution parameters that accurately reflect real-world behaviors. Furthermore, the learning algorithm benefits from having a well-designed objective, one which encourages desirable traits of the parameters to be learned.

The study and application of algorithms for both learning and decoding have seen enormous progress over the past half-century (see, for instance, [4] for an overview), but there is no algorithm which is "best" in any general sense for either task. Rather, a tradeoff is always present between key design variables, and algorithms must be chosen that favor some variables over others. For instance, in the design of algorithms for decoding, one always faces a tradeoff between design variables that include, but are not limited to, (1) a tolerance for error, (2) the number of possible states captured by the model, (3) the amount of data with which to perform decoding, and (4) the level of noise caused by sensors. On the other hand, in the design of algorithms for learning, one faces a tradeoff between variables that include (1) a tolerance for error of parameter estimates with respect to the optimum of some objective, (2) the number of computational steps required to produce an estimate within that tolerance, and (3) the properties of the objective itself. Relatively little study has been devoted to understanding the tradeoffs between such design variables and what they imply in practice. In this work, we attempt to derive and study such tradeoffs.

1.1 Overview

The remainder of this work is organized as follows. In Section 2, we give rigorous mathematical meaning to the notions presented in this chapter. In Sections 3 and 4, we state, discuss, and prove a theorem regarding tradeoffs in decoding. In Sections 5 and 6, we do the same for a theorem regarding tradeoffs in learning. In Section 7, we revisit the issues that arose throughout our analysis to suggest future directions for research.

1.2 Notation

The Euclidean sphere in $\mathbb{R}^n$ is denoted by $\mathbb{S}^{n-1}$. The Frobenius norm and ℓ_2-norm are denoted by $\|\cdot\|_F$ and $\|\cdot\|_2$, respectively. The trace of a matrix is denoted by $\mathrm{Tr}(\cdot)$. The smallest and largest eigenvalues of a matrix M are denoted by $\lambda_{min}(M)$ and $\lambda_{max}(M)$, respectively. We denote the sets $\{r \in \mathbb{R}^m \mid r \geq 0\}$, $\{r \in \mathbb{R} \mid r > 0\}$, and $\{n \in \mathbb{N} \mid n > 0\}$ by the symbols $\mathbb{R}_+$, $\mathbb{R}_{++}$, and $\mathbb{N}_+$, respectively. The abbreviation a.s. denotes *almost surely*. Expectation is denoted by $\mathbb{E}$. When the measure space is understood, probability is denoted generically by $\mathbb{P}$. If a random variable X has Gaussian distribution with mean μ and variance σ^2, we write $X \tilde{\sim} N(\mu, \sigma^2)$. The identity matrix is denoted by I. A standard multivariate Gaussian variable has mean 0 and variance I. The cardinality of a set is denoted by $|\cdot|$.

2 Problem formulation

In this chapter, we formulate the problem of quantifying tradeoffs in the tasks of learning and decoding for the HMM. To do so, we first define an HMM via the notion of a probability kernel. The kernel-centric view provides a framework for constructing an information-theoretic representation of the emissions process, thereby enabling the proof of the main theorem in Section 3, where we study tradeoffs in the decoding problem. Furthermore, probability kernels provide a common language for formalizing the definitions of learning and the error rate of a decoder. To study tradeoffs in the learning problem, we further provide here a construction of the HMM in which the transition and emission probabilities are represented by a parameter vector to be learned via optimization of a regularized log-likelihood objective.

2.1 Hidden Markov models and probability kernels

Definition 2.1 (Finite-state Markov chain)**.** *Let $K(t)$ be a stochastic process with time-parameter $t \in \mathbb{N}$, and suppose that $K(t)$ takes values in a set $\mathcal{K}$ for all t. Then $K(t)$ is called a Markov chain if it satisfies*

$$\begin{aligned}&\mathbb{P}[K(t) = k_t \mid K(t-1) = k_{t-1},\ K(t-2) = k_{t-2},\ \ldots,\ K(0) = k_0] \\ &= \mathbb{P}[K(t) = k_t \mid K(t-1) = k_{t-1}]\end{aligned}$$

a.s. for all t and for all $k_t, k_{t-1}, \ldots, k_0 \in \mathcal{K}$. If $\mathcal{K}$ is finite, then $K(t)$ is referred to as a finite-state Markov chain. $\mathcal{K}$ is referred to as the state-space.

In the above definition, the a.s. requirement ensures that the conditional probability is unique for all t. To see this, note that (for example) the equation

$$\mathbb{P}[K(t) = k_t,\ K(t-1) = k_{t-1}] = \mathbb{P}[K(t) = k_t \mid K(t-1) = k_{t-1}]\mathbb{P}[K(t-1) = k_{t-1}]$$

would not have a unique solution for the conditional probability if it should happen that $\mathbb{P}[K(t-1) = k_{t-1}] = 0$. Such an incident can be prevented by excluding sets of measure zero from $\mathcal{K}$.

For a Markov chain $K(t)$, we define the probability transition matrix (or simply *transition matrix*) by $T_K \equiv T = [T_{ij}]$ for $1 \leq i, j \leq |\mathcal{K}|$, where

$$T_{ij} \equiv \mathbb{P}[K(t) = j | K(t-1) = i].$$

The concept of a transition matrix can be generalized to an uncountable state-space via the notion of a probability kernel, as follows.

Definition 2.2 (Probability kernel)**.** *Let $(\mathcal{A}, \sigma\mathcal{A})$ and $(\mathcal{B}, \sigma\mathcal{B})$ be measurable spaces. The map $\kappa : \mathcal{A} \times \sigma\mathcal{B} \to [0, 1]$ is a probability kernel if it satisfies the following.*

- *For each $a \in \mathcal{A}$, the map $B \mapsto \kappa(a, B)$ is a probability measure on $(\mathcal{B}, \sigma\mathcal{B})$.*
- *For each $B \in \sigma\mathcal{B}$, the map $a \mapsto \kappa(a, B)$ is $\sigma\mathcal{A}$-measurable.*

To gain some intuition on this definition, let us consider the case of $\mathcal{A} = \mathcal{B} = \Xi$ and refer to Ξ as the state-space. The first property in Def. 2.2 suggests a sort of "transition probability" from each state a to a subset of the state-space Ξ. That is, for each a, the kernel assigns a probability to each subset of Ξ. By assigning these probabilities to subsets rather than points, we account for the situation in which Ξ is uncountable.

In fact, we only consider Markov chains with finite state-spaces in this work. However, we will require the notion of a probability kernel, not for the consideration of continuous state-spaces, but rather to consider continuous *emission* spaces in a structure referred to as a hidden Markov model (HMM), as we now define.

Definition 2.3 (Hidden Markov model)**.** *Consider a pair $(K(t), E(t))$, where $K(t)$ is a Markov chain with state-space $\mathcal{K}$, and $E(t)$ is a stochastic process that takes values in a set $\mathcal{E}$ and satisfies the following:*

$$\begin{aligned}
&\mathbb{P}[E(t) \in \mathcal{E}_t,\ E(t-1) \in \mathcal{E}_{t-1},\ \ldots,\ E(0) \in \mathcal{E}_0 \\
&\quad \mid K(t) = k_t,\ K(t-1) = k_{t-1},\ \ldots,\ K(0) = k_0] \\
&= \prod_{\tau=0}^{t} \mathbb{P}[E(\tau) \in \mathcal{E}_\tau \mid K(\tau) = k_\tau],
\end{aligned}$$

a.s. for all t, for all $\mathcal{E}_t, \mathcal{E}_{t-1}, \ldots, \mathcal{E}_0 \subseteq \mathcal{E}$, and for all $k_t, k_{t-1}, \ldots, k_0 \in \mathcal{K}$. Let $(\pi, \mathcal{T}, \mathcal{E})$ be a triplet, where π is a probability mass function, and $\mathcal{T}$ and $\mathcal{E}$ are probability kernels, such that

- *$\pi : \mathcal{K} \to [0, 1]$ defines $\mathbb{P}[K(0) = k_0]$,*
- *$\mathcal{T} : \mathcal{K} \times \sigma\mathcal{K} \to [0, 1]$ defines $\mathbb{P}[K(t) = k_t \mid K(t-1) = k_{t-1}]$ for all t, and*
- *$\mathcal{E} : \mathcal{K} \times \sigma\mathcal{E} \to [0, 1]$ defines $\mathbb{P}[E(t) \in \mathcal{E}_t \mid K(t) = k_t]$ for all t.*

Then $(\pi_L, \mathcal{T}_L, \mathcal{E}_L)$ is referred to as the hidden Markov model of the pair $(K(t), E(t))$.

This definition is adapted from a lemma of [7]. The reader is referred to [7] for a more general measure-theoretic treatment of the HMM. Again, the a.s. requirement ensures uniqueness of conditional probabilities for all t. In the next section, we consider emissions $E(t)$ that take values in an uncountable space and therefore require the notion of a probability kernel in order to be well-defined. We refer to this kernel as the emission kernel of an HMM. For consistency of terminology, we also refer to the transition matrix as a transition "kernel". We assume, throughout this work, that the transition and emission kernels do not depend on t, i.e., that the associated Markov chains are homogeneous [7].

Importantly, observe that any measurable function $f : \mathcal{W} \to \mathcal{X}$ induces a probability kernel

$$\kappa_f(\mathcal{S}|w) = 1(f(w) \in \mathcal{S}), \tag{1}$$

for all $\mathcal{S} \subseteq \mathcal{X}$. We will occasionally use this observation to formalize some important notions in both the decoding and learning problems.

By the a.s. requirement in Def. 2.3, the transition and emission probabilities are unique for all t, and thus, they give rise to the following objects, which we refer to as parameter vectors for both finite and continuous emission-spaces.

Definition 2.4 (HMM parameter vector)**.** *For an HMM with finite state-space $\mathcal{K}$ and finite emission-space $\mathcal{E}$, the transition and emission kernels generate the following variables, referred to as the HMM parameters, for all $k, k' \in \mathcal{K}$ and all $e \in \mathcal{E}$:*

$$\begin{aligned} \pi_k &\equiv \mathbb{P}[K(0) = k], \\ a_{kk'} &\equiv \mathbb{P}[K(t+1) = k' \mid K(t) = k], \\ b_{ke} &\equiv \mathbb{P}[E(t) = e \mid K(t) = k]. \end{aligned}$$

The HMM parameters satisfy the constraints

$$\begin{aligned} \sum_{k \in \mathcal{K}} \pi_k &= 1, \\ \sum_{k' \in \mathcal{K}} a_{kk'} &= 1 \text{ for all } k \in \mathcal{K}, \\ \sum_{e \in \mathcal{E}} b_{ke} &= 1 \text{ for all } k \in \mathcal{K}, \\ \pi_k, a_{kk'}, b_{ke} &\in [0,1] \text{ for all } k, k' \in \mathcal{K},\ e \in \mathcal{E}. \end{aligned} \tag{2}$$

The vector θ with coordinates given by

$$\theta = [\pi_k,\ a_{kk'},\ b_{ke}]^T,\ k, k' \in \mathcal{K},\ e \in \mathcal{E},$$

is referred to as the parameter vector.

Definition 2.5 (HMM parameter vector; white Gaussian emissions)**.** *For an HMM with finite state-space $\mathcal{K}$, emission space $\mathbb{R}^n$, and an emission kernel induced by the multivariate Gaussian distribution $\mathcal{N}(\mu, \sigma^2 I)$, the transition and emission kernels generate the following variables, referred to as the HMM parameters, for all $k, k' \in \mathcal{K}$:*

$$\begin{aligned} \pi_k &\equiv \mathbb{P}[K(0) = k], \\ a_{kk'} &\equiv \mathbb{P}[K(t+1) = k' \mid K(t) = k], \\ \mu &, \text{ as given above}, \\ \sigma^2 &, \text{ as given above}. \end{aligned} \tag{3}$$

The HMM parameters satisfy the constraints

$$\begin{aligned} &\sum_{k\in\mathcal{K}} \pi_k = 1, \\ &\sum_{k'\in\mathcal{K}} a_{kk'} = 1 \text{ for all } k \in \mathcal{K}, \\ &\pi_k, a_{kk'} \in [0,1] \text{ for all } k, k' \in \mathcal{K}, \\ &\sigma^2 > 0. \end{aligned} \tag{4}$$

The vector θ with coordinates given by

$$\theta = [\pi_k,\ a_{kk'},\ \mu,\ \sigma^2]^T,\ k, k' \in \mathcal{K}$$

is referred to as the parameter vector.

Our analysis of the learning problem in Section 5 will be applicable to both of the above definitions of the parameter vector. Note that, due to the summation constraints in these definitions, the space Θ of all possible values of θ is not a vector space. However, we refer to θ as a vector to emphasize that our analysis in Sections 5 and 6 will require norms and metrics defined on a vector space to which Θ is a subset. We refer to this vector space as the enveloping vector space of Θ. Hence, when taking the norm of θ, we clarify here that the norm is applied to the vector θ in the enveloping vector space of Θ rather than the element $\theta \in \Theta$. A similar perspective holds for the enveloping metric space.

2.2 State-space and probability structure

We now particularize the definitions of the previous section to facilitate the information-theoretic approach taken in subsequent chapters.

Let $\mathcal{K} = \{1, \ldots, |\mathcal{K}|\}$. The hidden state $K(t)$ at any t is a random variable on $\mathcal{K}$ with an associated measure space $(\mathcal{K}, 2^{\mathcal{K}}, p_K)$. The observed state $E(t) \equiv Y^m(t)$ at any t is a random vector on $\mathbb{R}^m$, for which a probability density function (p.d.f.) exists.

Now, define a vector K^t with coordinates given by

$$K^t \equiv [K(t),\ K(t-1),\ \ldots,\ K(t-\ell+1)]^T. \tag{5}$$

Note that K^t takes values at random in a set of cardinality $|\mathcal{K}|^\ell$ according to some p.m.f. p_{K^t} induced by p_K. We index this set by a variable $W(t)$. That is, at any t, the index $W(t)$ is a random variable taking values in $\mathcal{W} \equiv \{1, \ldots, |\mathcal{K}|^\ell\}$ according to p_{K^t}. For simplicity, we herein denote the p.m.f. of $W(t)$ as $p_W \equiv p_{K^t}$.

With the above notions, we may now give an assumption that will enable the analysis given in Sections 3 and 4.

Assumption 2.1. *Assume that $Y^m(t)$ is Gaussian with variance $\sigma^2 I = I$ and mean μ that satisfies*

$$\mu = f^m(K(t))$$

for some injection $f^m : \mathcal{K} \to \mathbb{R}^m$.

Define an injection $f : \mathcal{W} \to \mathbb{R}^{\ell m}$ such that, for all t and all $W(t) \in \mathcal{W}$, the coordinates of $f(W(t))$ are the concatenation of the coordinates of vectors in the sequence

$$\{f^m(K(t)),\ f^m(K(t-1)),\ \ldots,\ f^m(K(t-\ell+1))\}. \tag{6}$$

Define a random vector $Y(t)$ on $\mathbb{R}^{\ell m}$ whose coordinates are the concatenation of the coordinates of vectors in the sequence

$$\{Y^m(t),\ Y^m(t-1),\ \ldots,\ Y^m(t-\ell+1)\}.$$

Then, by Assumption 2.1, we may write

$$Y(t) = f(W(t)) + Z,$$

where Z is a standard Gaussian vector on $\mathbb{R}^{\ell m}$. This construction is crucial to the information-theoretic study of tradeoffs in the decoding problem in Section 4.

Because f is injective with finite domain, there exists $P \in \mathbb{R}^{++}$ such that

$$\|f(W(t))\|_2^2 \leq \ell m P.$$

We may interpret P as a (maximal) signal-to-noise ratio by defining an ℓm-length "signal" as $X(t) = f(W(t))$ and observing that, at each time t, the expression

$$\frac{\|X(t)\|_2^2}{\ell m}$$

can be viewed as the average-energy-per-bit of the signal, and the expression

$$\frac{\mathbb{E}[Z^T Z]}{\ell m}$$

can be viewed as the average-energy-per-bit of the noise. Note that the noise energy is equal to 1 here (by Assumption 2.1), but this choice was made without loss of generality to facilitate our analysis in subsequent chapters. (Equivalently, we may fix the signal energy and assume a bound on noise energy instead, but we do not explore this option in our analysis.) As mentioned previously regarding the norm of a parameter vector θ, the quantity $\|f(W(t))\|_2$ makes use of the ℓ_2-norm defined on the enveloping vector space of the codomain of f. Such a norm would be meaningless in the codomain of f because f is injective with a finite domain.

Assumption 2.1 only applies to our study of the decoding problem, in which the emission-space is continuous. Our study of the learning problem will be applicable to both continuous and discrete emission-spaces, as it will only require that the parameter vectors of Defs. 2.4 and 2.5 be well-defined.

2.3 State estimation

In this section, we construct a decoder as a probability kernel that maps an observed state-sequence to a member of $\mathcal{K}$. The decoder further depends on the transition and emission probabilities computed by the learning algorithm, but the effect of this dependence lies beyond the scope of our construction, and we defer its discussion to the end of Section 3. Still, we give here the rigorous notions that will guide that discussion.

Consider an uncountable set $\mathcal{L}$ in which each element is an HMM as defined previously. We denote each HMM in $\mathcal{L}$ as a triplet $L = (\pi_L, \mathcal{T}_L, \mathcal{E}_L)$, where π_L is the probability distribution of hidden states at $t = 0$, $\mathcal{T}_L$ is the transition kernel, and $\mathcal{E}_L$ is the emission kernel.

Let $T \in \mathfrak{T}$ denote a training set, where $\mathfrak{T}$ is the set of all possible training sets (which we intentionally leave ambiguous in definition). Each element of T is an ℓ-length sequence referred to as a *training example* taking the form

$$\{(K(t'), Y^m(t')),\ (K(t'-1), Y^m(t'-1)),\ \ldots,\ (K(t'-\ell+1), Y^m(t'-\ell+1))\}$$

for some $t' \leq t_f$, t_f being a time beyond which no training examples are available.

A learning algorithm (or simply *learner*) is a probability kernel $l : \mathfrak{T} \times \sigma\mathcal{L} \to [0, 1]$. Intuitively, this definition implies that, given a training set $T \in \mathfrak{T}$, the learner assigns a probability to each subset of $\mathcal{L}$. Recall from the previous section that any measurable function induces a probability kernel. Thus, even if a learner is deterministic, if we may represent it as a measurable function, then we may equivalently represent it with the above kernel-based definition. The kernel-based definition then captures the class of learners that are both deterministic and stochastic.

For convenience, we will often abuse notation to denote a learner as the measurable function that induces its kernel (see Eq. 1):

$$l : \mathfrak{T} \to \mathcal{L}.$$

Intuitively, in this notation, l maps a set of training data to a triplet L with some probability.

Given a learned model in $\mathcal{L}$ and a new ℓ-length sequence of observed states of the form

$$\{Y^m(t),\ Y^m(t-1),\ \ldots,\ Y^m(t-\ell+1)\} \tag{7}$$

for some $t > T + \ell - 1$, the decoder of an HMM computes an ℓ-length sequence of hidden states that best "explains" this newly observed sequence (e.g., in the sense of Problem 2 in [20]). The output of the decoder is thus a sequence of the form of Eq. 5.

The decoder can be viewed as a probability kernel $g : \mathbb{R}^{\ell m} \times 2^{\{1,\ldots,|\mathcal{K}|^\ell\}} \to [0, 1]$. Intuitively, for each realization of $Y(t)$, the decoder assigns a probability to each subset of $\{1, \ldots, |\mathcal{K}|^\ell\}$. Thus, it can be interpreted as a probability "transition" from a sequence of observed states to an element of the hidden state-space. With this interpretation, we will herein abuse notation and denote the decoder by the measurable function that induces its kernel (again, see Eq. 1)

$$g : \mathbb{R}^{\ell m} \to \{1, \ldots, |\mathcal{K}|^\ell\}. \tag{8}$$

Naturally, the learner l must influence the probabilities assigned by the decoder, but the scope of this influence remains to be studied, and we return to this issue in subsequent chapters.

Define the average probability of error (in decoding) as

$$\epsilon_{avg} = \frac{1}{|\mathcal{K}|^\ell} \sum_{w \in \mathcal{W}} (1 - P_{Y|X}(g^{-1}(w)|f(w))), \tag{9}$$

where the dependence of W, X, and Y on t is understood implicitly. Note that the averaging is done over the hidden state-space $\mathcal{K}$, and that we are using the notation g in the sense of the map in Eq. 8.

The goal of Sections 3 and 4 is to study an inequality that exhibits the tradeoff between ϵ_{avg}, $|\mathcal{K}|$, n, and P, i.e., the variables of interest in the design of a decoder.

2.4 Parameter estimation

The kernel-centric view of learning in the previous section will be useful in the discussion of how the learner influences the decoder, but it will not be useful in our analysis of tradeoffs in the learning problem. Rather, in the analysis of Sections 5 and 6, we only consider deterministic learners that map a set of training data to an HMM parameter vector. Both Def. 2.4 and Def. 2.5 will be suitable when referring to HMM parameter vectors in this context. Note that the learner of the previous section, which maps training data to a triplet L, is equivalent to the learner we consider here, which maps to a parameter vector θ, because both L and θ encompass information about the transition and emission probabilities. In this sense, L and θ are merely two representations of the same information, one in a measure space and the other in a metric space, respectively.

Within the class of deterministic learners, we further restrict the scope of our analysis to *iterative* learners, which compute the parameter vector θ of either Def. 2.4 or Def. 2.5 by iteratively computing estimates $\theta^{(i)}$ at iteration i according to some optimization objective to be designed by the practitioner.

In the current work, we consider the design of the objective to entail a choice of a regularization matrix R, whose role will be made precise in Section 5. Here, it suffices to mention that R is diagonal, and each of its non-zero elements represents a weight given to a coordinate of θ that encourages the coordinate to tend toward zero. Thus, R enables the practitioner to enforce some prior knowledge regarding θ, motivated by domain knowledge regarding the probability distributions governing the real-world process.

If we denote the optimum of the objective as θ^*, then the performance of the learner can be quantified as the distance of its estimate from θ^* under the ℓ_2 metric. Namely, its error at iteration i is expressed as $\|\theta^{(i)} - \theta^*\|_2$.

The goal of Sections 5 and 6 is to study an inequality that exhibits the tradeoff between $\|\theta^{(i)} - \theta^*\|_2$, i, and R, i.e., the variables of interest in the design of a learner.

2.5 A note on modeling

To make practical sense of the formulation above, we consider here one possible interpretation. For simplicity, at any t, let $f(W(t)) = K^t$ (defined in Eq. 5). Then the emission $Y^m(t)$ at any t may be viewed as m repeated noisy measurements of the underlying state $K(t)$. This situation arises in systems that provide m redundant sensor values of its state at each t, the purpose of redundancy being to counteract the uncertainty caused by noise from the sensor. Concatenation of these m-repeated measurements over an ℓ-length time interval then generates $Y(t)$.

3 Tradeoffs in state-sequence estimation

In this chapter, we state the tradeoffs between the following design parameters:

- the cardinality $|\mathcal{K}|$ of the hidden state-space,
- the tolerance ϵ on the average error probability of a decoder,
- the length ℓ of the emission sequence,
- the dimensionality m of each emission,
- the signal-to-noise ratio P of emissions.

We then discuss its implications in practice and provide plots that aid in interpreting the associated inequalities from a quantitative standpoint. The proof is given in the next chapter.

3.1 The converse theorem

The intuition behind the main result of this chapter is that, given sufficiently large data and an error tolerance ϵ, if a decoder achieves $\epsilon_{avg} \leq \epsilon$, then it must necesssarily satisfy a certain inequality involving the design parameters given in the previous section. In particular, the following theorem holds.

Theorem 3.1. *Let ϵ_{avg} denote the average probability of error of a decoder for an HMM with hidden state-space $\mathcal{K}$, learned with ℓ-length training examples in $\mathbb{R}^m$. For every $\epsilon \in (0, 1]$ and $P \in [0, \infty)$, there exists $N(P, \epsilon) \in \mathbb{N}$ such that, for all $n > N(P, \epsilon)$, the following holds. If a decoder achieves $\epsilon_{avg} \leq \epsilon$ given an ℓ-length sequence of emissions in $\mathbb{R}^m$, then*

$$\log |\mathcal{K}| \leq \frac{n}{2} \log(P + 1) + \frac{1}{2} \log n - nV(P)Q^{-1}(\epsilon) + g_c(P, \epsilon) \tag{10}$$

where

$$n = \ell m,$$
$$V(P) = \frac{P}{2} \frac{P + 2}{(P + 1)^2} \log^2(e),$$

and g_c is a continuous function of both P and ϵ.

We refer to Thm. 3.1 as the converse theorem, by convention of the name used in the information-theoretic work from which we draw the proof in the next chapter. Indeed, we will find in the next chapter that the proof of Thm. 3.1 is identical to that of the converse theorem of finite-blocklength channel coding.

To draw an analogy with the previously established converse theorems of channel coding, Thm. 3.1 can be interpeted as follows. Note that any decoder g must be constructed with respect to design parameters $\mathcal{D} = \{|\mathcal{K}|,\ n,\ P\}$, and we denote such a decoder as $g_{\mathcal{D}}$. Assume that $n > N(P, \epsilon)$. If $\mathcal{D}$ violates the bound 10, then there does not exist any decoder $g_{\mathcal{D}}$ that achieves $\epsilon_{avg} \leq \epsilon$. Hence, Thm. 3.1 can be interpreted as an

impossibility result, analogous to the converse theorems studied in communication theory. We argue that it is a first step toward a result that essentially states, "For sufficiently large n, if $\mathcal{D}$ violates a certain bound, then there does not exist any learner-decoder pair designed for $\mathcal{D}$ that can achieve $\epsilon_{avg} \leq \epsilon$." We return to this topic in the concluding remarks of this chapter.

In the next section, we elaborate on the above interpretation of Thm 3.1 and then plot the bound 10 to highlight some of its characteristics.

3.2 Impossibility and the state-space cardinality bounds

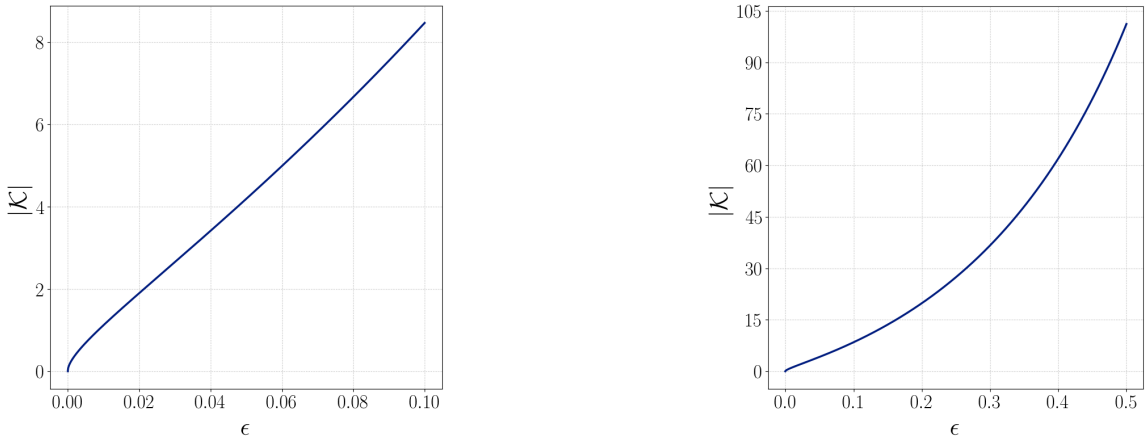
Consider a scenario in which the practitioner is tasked with modeling a real-world process in order to estimate its state-sequence for a given observation sequence. The practioner first specifies a desired set $\mathcal{K}$ of states to be captured in the model and a desired tolerance ϵ of decoding error. An HMM is then learned with ℓ-length training examples that involve elements from $\mathcal{K}$ and corresponding observations that exhibit a known signal-to-noise ratio of P. It is determined empirically that a particular decoder g achieves $\epsilon_{avg} \leq \epsilon$ for a given number m of measurements-per-timestep. Suppose that the resulting $n = \ell m$ is greater than the number $N(P, \epsilon)$ of Thm. 3.1, i.e., the amount of data n (for decoding with g) is sufficiently large so that the bound 10 holds. (Such a number $N(P, \epsilon)$ must exist, according to Thm. 3.1.)

Now, suppose that the practioner aims to modify $\mathcal{K}$ in order to capture more states of the underlying process. Toward this end, the HMM is re-learned with an increased $|\mathcal{K}|$ but with the same choices of ℓ, P, and m previously considered. (Assume that the modified $\mathcal{K}$ does not affect P so that $n > N(P, \epsilon)$ still holds.) The decoder g is likewise redesigned to accommodate the new choice of $\mathcal{K}$. However, if $|\mathcal{K}|$ is increased sufficiently that it now violates the bound 10, then g no longer achieves $\epsilon_{avg} \leq \epsilon$. In fact, violation of the bound 10 ensures that there does not exist any decoder designed for this choice of $\mathcal{K}$, n, and P that can achieve $\epsilon_{avg} \leq \epsilon$. Thus, Thm 3.1 gives a quantitative statement of the intuition that if the hidden state-space is sufficiently "diverse", then there does not exist any decoder that will achieve the desired error tolerance. In this sense, Thm 3.1 can be interpreted as an "impossibility" result, analogous to the converse theorems of channel coding.

In the converse theorems of channel coding, when the transmission rate violates the channel capacity constraint of a given channel, then there does not exist an encoder-decoder pair of that rate that achieves vanishing probability of error (asymptotically in blocklength). Similarly, in the finite-blocklength converse theorem [17] [18], assuming a sufficiently large blocklength n, when the transmission rate violates a bound similar to 10, there does not exist an encoder-decoder pair of that rate that achieves probability of error below a desired ϵ. In the context of channel coding, the transmission rate is defined as the ratio $(\log |\mathcal{K}|)/n$ (using our notation), and the study of information-theoretic limits in this context is centered around this ratio. In the current work, we will thus deviate from this well-studied perspective in order to highlight other aspects of the bound 10 that have not been discussed in previous works.

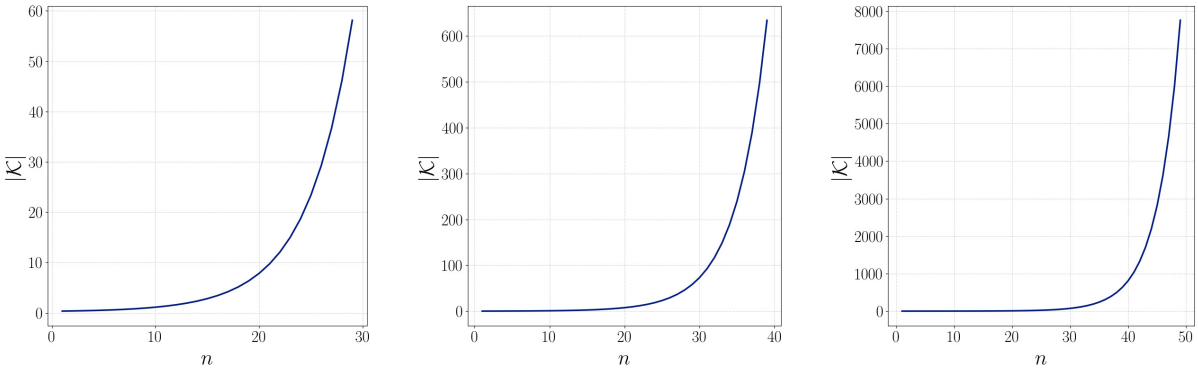
Rather, we consider perspectives relevant to the scenario presented above, in which $|\mathcal{K}|$ is upper bounded by a function of ϵ, n, and P. Fig. 1b shows the bound 10 with respect to ϵ, where $\epsilon \in [0, 0.5]$. A more subtle feature of this bound, however, can be observed when ϵ is restricted to the interval $[0, 0.1]$ as in Fig. 1a, which can be interpreted as the "high-accuracy" regime of decoding that may be relevant to those engineering applications which are especially sensitive to the performance of the state estimation procedure. In this regime, the upper bound on permissible values of $|\mathcal{K}|$ scales almost linearly with ϵ. Here, we find that $P = 1$ is a reasonable constant for illustration because increasing P merely increases the magnitude of the bounds in Fig. 1 without notably affecting its curvature. The same will be true for the remainder of the plots presented here, and a similar reasoning applies to the choices of $n = 10$ and $\epsilon = 0.01$ whenever n and ϵ are fixed.

Fig. 2 shows, from left to right, that the upper bound on $\mathcal{K}$ increases sharply in magnitude with respect to n. Indeed, the figure shows that each increase of 10 in n roughly causes an order-of-magnitude increase in the bound. The ranges of n in the four successive plots of Fig. 2 are chosen to illustrate four relevant regimes of $|\mathcal{K}|$ and n in practice. The appropriate scale of $|\mathcal{K}|$ in practice depends on the underlying process which the HMM is intended to model. Likewise, the appropriate range of $n = \ell m$ depends on how many



(a) $\epsilon \leq 0.1$ (b) $\epsilon \leq 0.5$

Figure 1: Bound 10 plotted as the upper bound on $|\mathcal{K}|$ with respect to ϵ, setting $P = 1$ and $n = 10$.



(a) $n \leq 30$ (b) $n \leq 40$ (c) $n \leq 50$

Figure 2: Detail of Bound 10 with respect to $n = \ell m$, across three different orders-of-magnitude of $\mathcal{K}$ relevant in practice (setting $P = 1$ and $\epsilon = 0.01$).

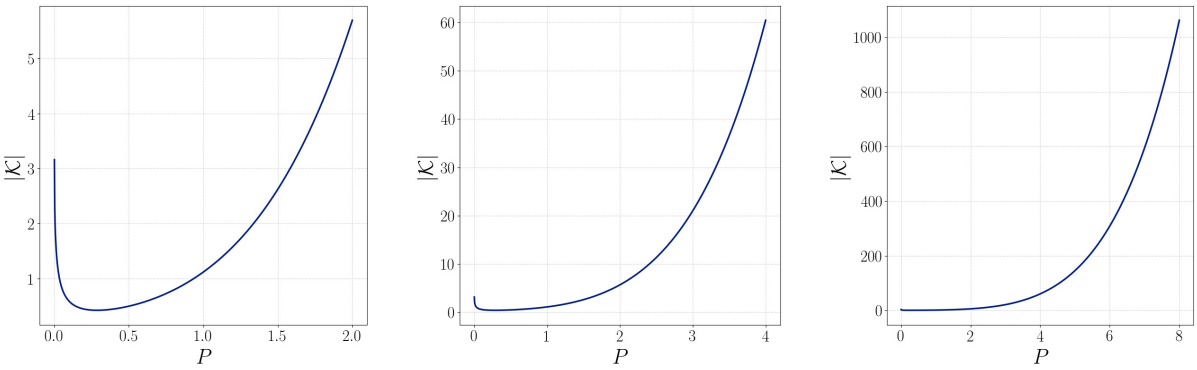
time-steps ℓ are considered in each training example and the number m of measurements to be collected at each time-step (by the interpretation of Section 2).

We take a similar approach when illustrating the bound with respect to P in Fig. 3, but here, the low-SNR regime in Fig. 3a shows that the bound "breaks down" as it approaches $P = 0$. That is, the bound likely does not predict anything useful for low values of P, as it suggests that a more diverse state-space can be supported by the decoder as P decreases from roughly 0.1 to 0. If any useful bounds can be derived in the low-SNR setting, they demand a different approach that lies beyond the scope of the current work.

3.3 Learner-decoder pairs and the role of the transition kernel

In Section 4, we provide the proof of Thm. 3.1. Our approach is to exploit Assumption 2.1 to make the problem amenable to the proof strategy of [17] and [18]. However, there are apparent shortcomings to our approach, which we now discuss. We argue that the main contribution of this chapter is the speculation that follows by observing these shortcomings in Thm. 3.1.

Our adaptation of the proof strategy in [17] and [18] deliberately neglects the Markov property of the process $K(t)$. Indeed, the introduction of $W(t)$ in our formulation allows us to treat the problem of



(a) $P \leq 2$ (b) $P \leq 4$ (c) $P \leq 8$

Figure 3: Detail of Bound 10 with respect to P, across three different orders-of-magnitude of $\mathcal{K}$ relevant in practice (setting $n = 10$ and $\epsilon = 0.01$).

state-sequence estimation as a problem of decoding individual symbols from a finite alphabet $\mathcal{W}$. Such a formulation does not consider the role of the transition kernel in generating the state sequence. It is crucial to note that the design of an HMM decoder almost invariably uses some knowledge of the transition kernel provided by the learner. An important example is the Viterbi algorithm (e.g., see [20]) and its variants. In contrast, the notion of a transition kernel is entirely unnecessary when decoding individual symbols rather than sequences of them.

A more appropriate formulation than that of Section 2 would first entail restriction of the transition behavior to a particular distribution or model (e.g., through the assumption that $K(t)$ is a sum of Rademacher variables). We would then define a modified probability of error ϵ^* that accounts for the decoding of both $K(t)$ and $K(t-1)$. That is, we would consider only the particular case of $\ell = 2$ before attempting to generalize any further. With this new formulation, we speculate that Thm. 3.1 can be extended to encompass the distribution of the transition kernel, just as it currently encompasses the distribution of the emission kernel (through P).

As it stands, the theorem merely shows a tradeoff involving

- the probability of error for decoding $W(t)$ (through the definition of ϵ_{avg}), and
- the variance of emissions (through the definition of P).

By the modified formulation above, we argue that a more appropriate result can be obtained which would exhibit a tradeoff involving

- the probability of error ϵ^* for decoding $K(t)$ and $K(t-1)$ for a given transition kernel (through the definition of ϵ^*),
- the variance of transitions, and
- the variance of emissions.

If we could obtain such a result, we would then pursue its generalization for $\ell > 2$. The ultimate goal would be to arrive at a more complete understanding of state-sequence estimation for HMMs that accounts for the decoder's knowledge about transitions between hidden states.

We further speculate that such a result can be extended to account for the performance of a given learner. To see this, recall the construction of a learner as a probability kernel, as given in Section 2. As it stands, Thm. 3.1 does not consider the fact that g relies on the kernel $l : \mathfrak{T} \times \sigma\mathcal{L} \to [0, 1]$ to compute its estimate of W. Indeed, μ and σ^2 should be estimated by the learner and subsequently employed by g, and in this

sense, it should be possible to relate the decoder's design variables to the design variables of the learner that will be studied in Sections 5 and 6. We speculate that the results arising from such a study could be more fittingly referred to as the fundamental limits of decoding in HMMs, as they would account for learner-decoder pairs rather than decoders alone, just as the fundamental limits of communication systems account for encoder-decoder pairs.

4 Proof of the converse theorem

We now prove the converse theorem of the previous chapter. First, we collect some facts regarding the beta function often considered in the theory of hypothesis testing, as well as the Berry-Esseen theorem that gives the rate at which an i.i.d. sample mean converges to normality. We then prove a general bound on the beta function that we will refer to as the meta-converse bound, by convention of [17] from which we take the strategy of the proof in this chapter. Finally, we apply the Neyman-Pearson lemma and Berry-Esseen theorem in succession to arrive at the result.

4.1 The beta function; the Neyman-Pearson lemma; the Berry-Esseen theorem

In this section, we gather the tools that will be needed in the proof of Theorem 3.1.

Let U be a discrete random variable on $\mathcal{U}$ that can have one of two distributions P and Q. Define a random variable T such that the event $T = 1$ is associated with distribution P and the event $T = 0$ is associated with Q. We call T a hypothesis test between these distributions, and we see that

$$\sum_{u\in\mathcal{U}} P_{T|U}(1|u)Q_U(u)$$

is the test's probability of error under hypothesis Q and

$$\sum_{u\in\mathcal{U}} P_{T|U}(1|u)P_U(u)$$

is its probability of success under hypothesis P. Then we may define the following so-called "beta function" as the infimal error probability under Q given at least α success probability under P:

$$\beta_\alpha(P_U, Q_U) = \inf_{P_{T|U}} \sum_{u\in\mathcal{U}} P_{T|U}(1|u)Q_U(u),$$

where the infimum is taken over all $P_{T|U}$ that satisfy

$$\sum_{u\in\mathcal{U}} P_{T|U}(1|u)P_U(u) \geq \alpha.$$

We will require the following lemma to establish the meta-converse bound of the next subsection.

Lemma 4.1 (Polyanskiy)**.** *[17] If $\beta_\alpha(P_{Y|X=x}, Q_{Y|X=x})$ is independent of $x \in \mathsf{F}$, then for any P_X supported on F,*

$$\beta_\alpha(P_X P_{Y|X}, Q_X Q_{Y|X}) = \beta_\alpha(P_{Y|X=x}, Q_{Y|X=x}). \tag{11}$$

Proof. Please refer to [17]. □

The following is a variant of the Neyman-Pearson lemma as stated in [19] and [18].

Lemma 4.2 (Neyman-Pearson)**.** *Let $\mathcal{P}$ and $\mathcal{Q}$ be probability measures over a space for which a random variable θ is defined. For all $\alpha \in [0, 1]$, there exist real constants $\gamma > 0$ and $\tau \in [0, 1)$ such that*

$$\beta_\alpha(\mathcal{P}, \mathcal{Q}) = \mathcal{Q}[T^*_\alpha = 1] \leq \mathcal{Q}[T = 1], \tag{12}$$

and the optimal test T_α^* is defined by

$$T_\alpha^*(\theta) = \mathbb{1}\left(\frac{d\mathcal{P}}{d\mathcal{Q}} > \gamma\right) + T_\tau \mathbb{1}\left(\frac{d\mathcal{P}}{d\mathcal{Q}} = \gamma\right), \tag{13}$$

where T is any test that satisfies $\mathcal{P}[T=1] \geq \alpha$, $T_\tau \in \{0,1\}$ is 1 with probability τ independent of θ, and the two constants $\gamma > 0$ and $\tau \in [0,1)$ are such that

$$\mathcal{P}[T_\alpha^* = 1] = \alpha. \tag{14}$$

If $\mathcal{P}$ is not absolutely continuous with respect to $\mathcal{Q}$, then extend the quantity $d\mathcal{P}/d\mathcal{Q}$ to equal $+\infty$ over the singular set.

Theorem 4.1 (Berry-Esseen)**.** *[3] For independent random variables $\{X_i\}_{i=1}^n$ with $\mu_i = \mathbb{E}[X_i]$, $\sigma_i^2 = \mathbb{E}[|X_i - \mu_i|^2]$, and $s_i = \mathbb{E}[|X_i - \mu_i|^3]$, it holds true that*

$$\left| \mathbb{P}\left[\frac{\sum_{i=1}^n (X_i - \mu_i)}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \leq c_1 \right] - Q(-c_1) \right| \leq \frac{6\sum_{i=1}^n s_i}{\left(\sum_{i=1}^n \sigma_i\right)^{3/2}}. \tag{15}$$

4.2 Proof of the meta-converse

In this section, we define a hypothesis test on channels and obtain a bound on the associated beta function via the Neyman-Pearson Lemma 4.2. In the next section, this bound will then be expanded via the Berry-Esseen theorem 4.1 to arrive at the main result, Theorem 3.1. Throughout the proof, we omit the dependence of variables on t, as it is understood that the arguments hold for any value of t.

Define the random variables

$$\begin{aligned} X &= bf(W), \\ Y &= X + Z, \\ T &= \mathbb{1}(g(Y) = W), \end{aligned}$$

where $\mathbb{1}(\cdot)$ is the indicator map, and the only notations newly introduced here are those of X and T. We will construct a hypothesis test in which the test variable is T and the "observation" variable is (X, Y), so that the two hypotheses under consideration are the distributions P_{XY} and Q_{XY}, associated with $T = 1$ and $T = 0$, respectively. (This so-called "observation" is unrelated to any notions of observation previously discussed in our framework.) To show that T is indeed such a test, we show that its distribution conditioned on (X, Y) is equivalent under both hypotheses, i.e., that $P_{T|XY} = Q_{T|XY}$.

First, note that Y is independent of W when conditioned on X, and $g(Y)$ is independent of both W and X when conditioned on Y. Then $g(Y)$ is independent of W when conditioned on X and Y, and the four variables form a Markov chain $W \to X \to Y \to g(Y)$. Then we also see that W is independent of Y when conditioned on X. Using these independence properties, under either hypothesis,

$$\begin{aligned} &\mathbb{P}[g(Y) = W | X, Y] \\ &= \sum_{w=1}^{|\mathcal{K}|^\ell} \mathbb{P}[\{g(Y) = w\} \cap \{W = w\} | X, Y] \\ &= \sum_{w=1}^{|\mathcal{K}|^\ell} \mathbb{P}[g(Y) = w | X, Y] \mathbb{P}[W = w | X, Y], \\ &= \sum_{w=1}^{|\mathcal{K}|^\ell} \mathbb{P}[g(Y) = w | X] \mathbb{P}[W = w | Y]. \end{aligned} \tag{16}$$

In the summation of the last line, neither of the factors in each term rely on the choice of joint distribution between X and Y, as a consequence of the above Markov structure. Thus, $\mathbb{P}[g(Y) = W|X, Y] = \mathbb{P}[T = 1|X, Y]$ is invariant under the choice of hypothesis, and T is therefore a valid test with the unique conditional distribution $P_{T|XY} = Q_{T|XY}$.

Let ϵ_1 denote the average probability that $g(Y)$ yields an incorrect estimate of W under P_{XY}. Define ϵ_2 similarly under Q_{XY}. Then, we have that

$$1 - \epsilon_1 = \sum_{x \in B^n(nP)} \sum_{y \in \mathbb{R}^n} P_{T|XY}(1|x, y) P_{XY}(x, y), \tag{17}$$

$$1 - \epsilon_2 = \sum_{x \in B^n(nP)} \sum_{y \in \mathbb{R}^n} P_{T|XY}(1|x, y) Q_{XY}(x, y). \tag{18}$$

We can verify these expressions formally by observing that

$$\begin{aligned}
\epsilon_1 &= \frac{1}{|\mathcal{K}|^\ell} \sum_{w \in \mathcal{W}} (1 - P_{Y|X}(g^{-1}(w)|f(w))) \\
&= \sum_{w \in \mathcal{W}} P_W(w)(1 - P_{Y|X}(g^{-1}(w)|f(w))) \\
&= \sum_{w \in \mathcal{W}} \mathbb{P}(W = w)\mathbb{P}(g(Y) \neq w|W = w) \\
&= \mathbb{P}(T = 0) \\
&= 1 - \mathbb{P}(T = 1) \\
&= 1 - \sum_{x \in B^n(nP)} \sum_{y \in \mathbb{R}^n} P_{T|XY}(1|x, y) P_{XY}(x, y).
\end{aligned}$$

Now, for some $\alpha \in [0, 1]$, define

$$\begin{aligned}
&\beta_\alpha(P_{XY}, Q_{XY}) \\
&= \inf_{P_{T|XY}} \sum_{x \in B^n(nP)} \sum_{y \in \mathbb{R}^n} P_{T|XY}(1|x, y) Q_{XY}(x, y),
\end{aligned} \tag{19}$$

where the infimum is taken over all $P_{T|XY}$ that satisfy

$$\sum_{x \in B^n(nP)} \sum_{y \in \mathbb{R}^n} P_{T|XY}(1|x, y) P_{XY}(x, y) \geq \alpha. \tag{20}$$

To clarify, the infimum is taken over tests that achieve at least α probability of a correct decision under hypothesis P_{XY}. There exists a unique optimal test for which this infimum is achieved, as guaranteed by the Neyman-Pearson lemma (given later), but note that this optimal test is not necessarily the test T constructed above. From this definition, we have

$$\beta_{\epsilon_1}(P_{XY}, Q_{XY}) \leq 1 - \epsilon_2. \tag{21}$$

We wish for P_{XY} to represent the true joint distribution in Eq. 7, while Q_{XY} remains an alternate hypothesis which we are free to construct within reason. Then, to ensure that the previous bound holds for all possible P_{XY}, we loosen the bound over P_{XY}. In contrast, we also tighten the bound over the Q_{XY} that we choose. As a result, we have

$$\inf_{P_X} \sup_{Q_Y} \beta_{\epsilon_1}(P_X P_{Y|X}, P_X \times Q_Y) \leq 1 - \epsilon_2. \tag{22}$$

The optimization above does not include the distribution $P_{Y|X}$, which is already determined by the relation between X and Y in their definitions.

We now restrict x to the sphere $\mathsf{S}^{n-1}(nP)$, although it can be shown that the result can be generalized to the ball $B^n(nP)$ as in [17]. With this restriction, $\beta_\alpha(P_{Y|X=x}, Q_{Y|X=x})$ is independent of x by the radial symmetry of $\mathsf{S}^{n-1}(nP)$, and we may thus apply Lemma 4.1 for $\mathsf{F} = \mathsf{S}^{n-1}(nP)$. Due to the same symmetry, $P_{Y|X=x}$ is also independent of x, so by Lemma 4.1, Eq. 22 becomes

$$\sup_{Q_{Y|X=x}} \beta_\alpha(Q_{Y|X=x}) \leq 1 - \epsilon_2, \tag{23}$$

where we have dropped the first argument from $\beta_\alpha(\cdot,\cdot)$ to emphasize that the bound now only depends on the second argument. Note, however, that the implicit first argument is $P_{Y|X=x}$, the true Gaussian measure of mean x and variance $I_{n\times n}$ induced by Z in Eq. 7.

Here, we will allow $Q_{Y|X=x} = Q_Y$, and we choose Q_Y to be the Gaussian distribution $\mathcal{N}(0, (1+P)I_{n\times n})$ (We refer the reader to chapter 4 of [17] for a thorough justification.)

Since Q_Y has been chosen to be independent of the value of X, the probability $1-\epsilon_2$ becomes the probability that $g(\cdot) = W$ when $g(\cdot)$ chooses with uniform randomness from $\mathcal{K}$. (Here, we denote the output of the decoder under hypothesis Q as $g(\cdot)$, in order to avoid confusion with the output $g(Y)$ under hypothesis P.) Thus, for any p.m.f. p_W, we have

$$\begin{aligned} 1-\epsilon_2 &= \mathbb{P}(g(\cdot) = W) \\ &= \sum_{w\in\mathcal{W}} \mathbb{P}(g(\cdot) = w | W = w) p_W(w) \\ &= \sum_{w\in\mathcal{W}} \frac{1}{|\mathcal{K}|^\ell} 1_W(w) \\ &= \frac{1}{|\mathcal{K}|^\ell}. \end{aligned} \tag{24}$$

Therefore, substituting into Eq. 23 our choice of $Q_{Y|X=x} = Q_Y$, along with Eq. 24, we obtain the meta-converse bound as

$$\sup_{Q_Y} \beta_\alpha(Q_Y) \leq 1 - \epsilon_2. \tag{25}$$

4.3 Computing the Berry-Esseen bound

Temporarily, we now consider $\beta_\alpha(\cdot,\cdot)$ with arbitrary probability measures in order to obtain a general bound via the Neyman-Pearson Lemma that will particularize desirably when we substitute the specific probability distributions in Eq. 25. The particularized bound will be desirable in the sense that it will be amenable to an application of the Berry-Esseen bound, which gives rise to the variables of interest in our main theorem. Finally, we will combine the Berry-Esseen result with the meta-converse of the previous subsection, thus completing the proof of our main theorem.

Define the event

$$\mathsf{E} = \left\{ \frac{d\mathcal{P}}{d\mathcal{Q}} < \gamma \right\}.$$

Then the Radon-Nikodym theorem gives

$$\mathcal{P}[\mathsf{E}] \;=\; \int_{\mathsf{E}} \frac{d\mathcal{P}}{d\mathcal{Q}} d\mathcal{Q} \;<\; \gamma \int_{\mathsf{E}} d\mathcal{Q} \;=\; \gamma \mathcal{Q}[\mathsf{E}]. \tag{26}$$

This "change-of-measure" argument is standard in information theory [16]. Note that the above steps still hold under intersections of E with arbitrary events, a fact which we use in the following argument.

Fix T to be the test defined previously for the hypotheses P_{XY} and Q_{XY}. Whatever value $\mathcal{Q}[T=1]$ that $\beta_\alpha(\mathcal{P},\mathcal{Q})$ achieves under this test cannot be any less than the value it achieves under the unique optimum T^*_α of Lemma 4.2. Therefore, letting $\mathcal{P} = P_{Y|X=x}$ and $\mathcal{Q} = Q_Y$ (as derived using Lemma 4.1), we have

$$\begin{aligned}
\beta_\alpha(\mathcal{P},\mathcal{Q}) &= \mathcal{Q}[T=1] \\
&\geq \mathcal{Q}[T^*_\alpha = 1] \\
&\geq \mathcal{Q}\left[\{T^*_\alpha = 1\} \cap \mathsf{E}\right] \\
&\geq \frac{1}{\gamma}\mathcal{P}[\{T^*_\alpha = 1\} \cap \mathsf{E}] \qquad (27)\\
&\geq \frac{1}{\gamma}\left(\mathcal{P}[T^*_\alpha = 1] - \mathcal{P}[\mathsf{E}^c]\right) \qquad (28)\\
&= \frac{1}{\gamma}\left(\alpha - \mathcal{P}[\mathsf{E}^c]\right). \qquad (29)
\end{aligned}$$

Eq. 27 follows from the Radon-Nikodym theorem as demonstrated above. Eq. 28 follows from the fact that, for any events E_1 and E_2,

$$\begin{aligned}
&\mathbb{P}[\mathsf{E}_1] + \mathbb{P}[\mathsf{E}_2] - \mathbb{P}[\mathsf{E}_1 \cap \mathsf{E}_2] \leq 1, \\
&\mathbb{P}[\mathsf{E}_1 \cap \mathsf{E}_2] \geq \mathbb{P}[\mathsf{E}_1] - \mathbb{P}[\mathsf{E}_2^c]. \qquad (30)
\end{aligned}$$

Eq. 29 follows from Lemma 4.2. The Radon-Nikodym derivative in Eq. 29 can be computed by simple application of the chain rule with respect to Lebesgue measure μ, as follows. Simply observe that $P_{Y|X=x}$ is a Gaussian measure of mean x and variance $I_{n\times n}$ induced by Z in Eq. 7, and observe that Q_Y is a Gaussian measure of mean 0 and variance σ_Y^2, as chosen previously. Then, without loss of generality, let $x = [\sqrt{P}\ \sqrt{P}\ \ldots\ \sqrt{P}]^T$, and we have

$$\begin{aligned}
\frac{dP_{Y|X=x}}{dQ_Y}(y) &= \frac{dP_{Y|X=x}}{d\mu}(y)\frac{d\mu}{dQ_Y}(y) \\
&= \sigma_Y^n \exp\left[-\left(\frac{1}{2}\|y-x\|_2^2 - \frac{1}{2\sigma_Y^2}\|y\|_2^2\right)\right] \\
&= \sigma_Y^n \exp\left[\frac{1}{2}\sum_{i=1}^n\left(\frac{y_i^2}{\sigma_Y^2} - (y_i - \sqrt{P})^2\right)\right]. \\
&= \sigma_Y^n \exp\left[\frac{1}{2\sigma_Y^2}\sum_{i=1}^n\left((\zeta_i - \sqrt{P})^2 - \sigma_Y^2\zeta_i^2\right)\right],
\end{aligned}$$

where we have introduced the zero-mean i.i.d. variable $\zeta_i = y_i - \sqrt{P}$ to facilitate later analysis. In fact, it will soon be convenient to instead use the quantity

$$\begin{aligned}
&-\log\frac{dP_{Y|X=x}}{dQ_Y}(y) \\
&\quad = -n\log\sigma_Y - \frac{\log e}{2\sigma_Y^2}\left[\sum_{i=1}^n\left(P - 2\zeta_i\sqrt{P} - P\zeta_i^2\right)\right], \qquad (31)
\end{aligned}$$

where we have substituted $\sigma_Y^2 = 1+P$ (as chosen previously) inside the summation to simplify terms, but we retain the σ_Y^2 outside the summation to promote clarity in the upcoming calculations.

We now define the quantity γ' in such a way that

$$\gamma = \exp\left(-\gamma' + \frac{n}{2}\log(\sigma_Y^2)\right), \qquad (32)$$

so that the inequality of E^c in Eq. 29 becomes

$$\frac{dP_{Y|X=x}}{dQ_Y} \geq \exp\left(-\gamma' + \frac{n}{2}\log(\sigma_Y^2)\right).$$

Solving for γ' and applying Eq. 31, we obtain

$$\begin{aligned}
\gamma' &\geq \frac{n}{2}\log(\sigma_Y^2) - \log\frac{dP_{Y|X=x}}{dQ_Y}, \\
&= -\frac{\log e}{2\sigma_Y^2}\left[\sum_{i=1}^{n}\left(P - 2\zeta_i\sqrt{P} - P\zeta_i^2\right)\right] \\
&= \frac{n\log e}{2\sigma_Y^2}\left(P\zeta_i^2 + 2\zeta_i\sqrt{P} - P\right), \\
&= \sum_{i=1}^{n} h_i,
\end{aligned}$$

where we define

$$h_i = \frac{\log e}{2\sigma_Y^2}\left(P\zeta_i^2 + 2\zeta_i\sqrt{P} - P\right).$$

Thus, we arrive at

$$\mathcal{P}[\mathsf{E}^c] = \mathcal{P}\left[\sum_{i=1}^{n} h_i \leq \gamma'\right],$$

to which we may then apply the Berry-Esseen theorem, in order to express Eq. 29 in terms of the key variables n and P.

In Thm. 4.1, let $X_i = h_i$, and note that $\mu_i = 0$ for all i by definition of h_i. Also note that σ_i^2 are equivalent for all i, and the same is true for s_i, since the h_i are identically distributed. Now, define

$$\alpha_n = \alpha - \frac{12s_i}{\sqrt{n(\sigma_i^2)^3}} > 0,$$

so that we have

$$n > N(P,\alpha) \equiv \left(\frac{12s_i}{\alpha(\sigma_i^2)^{3/2}}\right)^2.$$

By the definition of h_i, both σ_i^2 and s_i are functions of P, which we may control according to the problem formulation of the previous section. Hence, in the following analysis, we may control α_n and therefore N as desired. Let these quantities be such that

$$\gamma' = -\sqrt{n\sigma_i^2}Q^{-1}(\alpha_n), \tag{33}$$

and in Thm. 4.1, let $c_1 = -Q^{-1}(\alpha_n)$. Then the theorem gives

$$\left|\mathbb{P}\left[\sum_{i=1}^{n} h_i \leq c_1\sqrt{n\sigma_i^2}\right] - Q(-c_1)\right| \leq \frac{6ns_i}{(n\sigma_i)^{3/2}}.$$

Proceeding from this result, we have

$$\begin{aligned}
\mathbb{P}\left[\sum_{i=1}^{n} h_i \leq \gamma'\right] &\leq \alpha_n + \frac{6s_i}{\sqrt{n(\sigma_i^2)^3}} \\
&\leq \alpha - \frac{6s_i}{\sqrt{n(\sigma_i^2)^3}},
\end{aligned}$$

where the last line follows from the definition of α_n. Upon substitution into Eq. 29,

$$\begin{aligned}\beta_\alpha(\mathcal{P},\mathcal{Q}) \;\equiv\; \beta_\alpha(Q_Y) \;&\geq\; \frac{6s_i}{\gamma\sqrt{n(\sigma_i^2)^3}}. \\ &\geq \frac{6s_i}{\sqrt{n(\sigma_i^2)^3}}\exp\left(\gamma' - \frac{n}{2}\log(\sigma_Y^2)\right).\end{aligned}$$

The first equivalence is obvious by the choices of measure that were used in computing the inequality. The last step follows by substituting Eq. 32 for γ. Combining this result with Ineq. 25, taking the logarithm, and rearranging terms,

$$\log\left(\frac{1}{\mathcal{K}^\ell}\right) \geq \log\left(\frac{6s_i}{\sigma_i^3}\right) + \gamma' - \frac{1}{2}\log(n) - n\log(\sigma_Y). \tag{34}$$

The first term in the right-hand side is arranged as such because we soon intend to absorb it into the g_c term in the statement of our theorem.

We now evaluate a lower bound on γ' in terms of $Q^{-1}(\alpha)$, to which we will then substitute $\alpha = 1-\epsilon$ to conclude the proof. Again, recall the dependence of s_i and σ_i^2 on P. Then by our control of P, let $\left[\alpha - \frac{12s_i}{\sqrt{n(\sigma_i^2)^3}},\ \alpha\right] \subset (0,1)$ for all $n > N(P,\alpha)$, and let a be any point in this interval. Taking the Taylor expansion of Eq. 33 over this interval about the point α, we have

$$\begin{aligned}\gamma' &= -\sqrt{n\sigma_i^2}Q^{-1}(\alpha) - (a-\alpha)\sqrt{n\sigma_i^2}\frac{dQ^{-1}}{da}(\alpha) \\ &\geq -\sqrt{n\sigma_i^2}Q^{-1}(\alpha) + \left(\alpha - \frac{12s_i}{\sqrt{n(\sigma_i^2)^3}} - \alpha\right)\sqrt{n\sigma_i^2}\frac{dQ^{-1}}{da}(\alpha) \\ &= -\sqrt{n\sigma_i^2}Q^{-1}(\alpha) + \frac{12s_i}{\sigma_i^2}\frac{dQ^{-1}}{da}(\alpha).\end{aligned} \tag{35}$$

Note that the second term in the expansion has only one negative factor, $(a-\alpha)$, and hence the inequality holds by definition of a. Let us bound the derivative in this result by a function continuous in both P and α in order to absorb terms into the g_c term of our main theorem. Toward this end, define

$$\alpha_1 = \alpha - \frac{12s_i}{\sqrt{N(P,\alpha)(\sigma_i^2)^3}}, \tag{36}$$

so that the interval $[\alpha_1,\ \alpha]$ encloses the interval over which the Taylor expansion was taken (for all $n > N(P,\alpha)$). Since the derivative of Q^{-1} is continuous on this interval, there exists

$$g_1 = \min_{a_1 \in [\alpha_1,\ \alpha]} \frac{dQ^{-1}}{da}(a_1), \tag{37}$$

which is continuous in both P and α. Substituting g_1 into Eq. 35 and combining the result with the γ' of Eq. 34, we have

$$\begin{aligned}&\log\left(\frac{1}{\mathcal{K}^\ell}\right) \\ &\geq -\sqrt{n\sigma_i^2}Q^{-1}(\alpha) - \frac{1}{2}\log(n) - n\log(\sigma_Y) - g_c(P,\alpha),\end{aligned}$$

where

$$g_c(P,\alpha) = -\log\left(\frac{6s_i}{\sigma_i^3}\right) - \frac{12s_i}{\sigma_i^2}g_1$$

is continuous in both P and α. Recall again that the dependence of g_c on P arises from s_i and σ_i, while its dependence on α arises from g_1.

By the definition of α in Eq. 19, we allow

$$Q^{-1}(\alpha) = Q^{-1}(1-\epsilon) = -Q^{-1}(\epsilon), \tag{38}$$

and the previous bound becomes

$$Q^{-1}(\epsilon) \leq \frac{-\log \mathcal{K}^\ell + n\log\sigma_Y + \frac{1}{2}\log n + g_c(P,\epsilon)}{\sqrt{n\sigma_i^2}}. \tag{39}$$

Substituting $\sigma_Y^2 = 1+P$ and computing $\sigma_i^2 = \mathbb{E}[|h_i|^2]$, we obtain the statement of the main theorem.

5 Tradeoffs in parameter estimation

In this chapter, we consider the task introduced in Section 2 of computing the parameter vector θ of an HMM using training data, a process herein referred to as *learning*. We denote the dimension of the parameter vector as p. To particularize the notions of Section 2, we state here the variables whose tradeoffs we intend to study:

- the error $\|\theta^{(i)} - \theta^*\|$ of the learner's parameter estimate at iteration i,
- the number i of iterations,
- properties of a regularization matrix R, namely $\mathrm{Tr}(R)$ and a number δ that restricts its eigenvalues.

In the sections to follow, the specific class of iterative learners to be considered will defined. Furthermore, the matrix R will be defined and the notion of regularization made precise.

5.1 Learning via projected gradient methods

Standard learning techniques for HMMs include the Baum-Welch algorithm and gradient-based methods [20]. Here, we only consider the gradient descent algorithm (or simply *gradient descent*), a method that lends itself well to an analysis of convergence within the setting we consider. Gradient descent computes the HMM parameters by iteratively optimizing an objective function that is typically taken to be the log-likelihood of training data with respect to the HMM parameters. Specifically, defining the objective function to be the negative log-likelihood (NLL)

$$\mathfrak{L}(\theta) = -\log \mathbb{P}(O|\theta),$$

gradient descent attempts to solve the problem

$$\min_{\theta\in\Theta} \mathfrak{L}(\theta) \tag{40}$$

by performing the following update at each iteration i:

$$\begin{aligned} \theta^{(i+1)} &= \mathrm{Proj}_\Theta\left(\theta^{(i)} - \mu\nabla_\theta \mathfrak{L}(\theta)\Big|_{\theta^{(i)}}\right), \\ \mathrm{Proj}_\Theta(\theta) &= \arg\min_{\bar{\theta}\in\Theta} \|\theta - \bar{\theta}\|_{\ell_2}^2, \end{aligned} \tag{41}$$

where μ denotes the learning rate, and the set Θ with the projection operator Proj_Θ accomodates the constraints in Def. 2.4 or Def. 2.5 as appropriate. We refer the reader to [11] for a sampling of alternative methods for enforcing the constraints. In our work, only the projected gradient method of the algorithm 41 is suitable for our analysis.

Note that our analysis in this chapter will not depend on any knowledge regarding the behavior of $\mathfrak{L}(\theta)$ with respect to θ except for its differentiability. This observation suggests that the main theorem of this chapter holds in great generality and is hence relevant to the diverse selection of learning objectives that have been considered in previous literature (see, for instance, [1], [13], [21], [22]). However, we will find that the proof of this theorem will require regularization of the objective in 40 and thus leaves opportunity for further generalization, as discussed in the next section.

5.2 Regularization and the contribution of the log-likelihood gradient

Importantly, we must note that our analysis of the performance of the projected gradient method relies on a regularized form of the objective function considered previously. In particular, we require that the problem 40 must be regularized in the following sense:

$$\min_{\theta\in\Theta} \mathfrak{L}(\theta) + \frac{1}{2}\theta^T R\theta, \tag{42}$$

where $R \in \mathbb{R}^{p\times p}$ is diagonal. Each non-zero component r_{jj} of R (for $j \in \{1, 2, \ldots, p\}$) acts as a tuning parameter to either encourage or discourage the algorithm 41 to produce estimates of parameter θ_j that tend toward zero. Large values of r_{jj} for some j will cause the algorithm to regularize the parameter θ_j heavily relative to another parameter, such as $\theta_{j'}$ for which $r_{j'j'}$ is smaller than r_{jj}.

Although this form of regularization can be useful in many contexts, it is not standard, and therefore, we are unable to argue that the main results of this chapter are fundamental to the problem 40 in any sense. Indeed, beyond the information-theoretic context of Sections 2, 3, and 4, it is not clear what the term "fundamental" encompasses in the context of iterative learning. We argue here that a fundamental tradeoff in iterative learning is one that generalizes to all log-likelihood objectives considered in the learning of HMMs, but the sampling of literature cited in the previous section suggests that the great diversity of learning objectives for HMMs would make this task challenging and far beyond the scope of our current work. Besides variations on the log-likelihood expression, past work has also considered variations on the regularization term (see, for instance, [6] and [14]), but only a regularizer of the form in 42 is suitable for our analysis, and it remains to be seen in future work whether our approach can be adapted for other regularization methods.

We now introduce a modification to the algorithm 41 which will facilitate our analysis. We motivate this decision by mentioning that, as the $\nabla_\theta\mathfrak{L}$ term in 41 fluctuates on each iteration, so too does the bound in our main theorem, as will be seen. To remove this iteration-dependence of the norm, we take only the direction of this term on each step:

$$\begin{aligned} \theta^{(i+1)} &= \theta^{(i)} - \mu\left(R\theta^{(i)} + \xi^{(i)}\right), \\ \xi^{(i)} &= \eta \frac{\nabla_\theta\mathfrak{L}}{\|\nabla_\theta\mathfrak{L}\|_2}\Big|_{\theta^{(i)}}, \end{aligned} \tag{43}$$

and fix the magnitude by $\eta \in [0, \infty)$. In a sense, η serves as a parameter for tuning the "contribution" of this term to the update. By setting η to be the norm in the denominator, we recover the original update step. As we will see, η can also be used to tune the bound in the main theorem. A similar normalization factor to Eq. 43 has been considered in [9].

In the next section, we state the main result of this chapter and discuss its implications.

5.3 The convergence theorem

We now state the main theorem of this chapter.

Theorem 5.1 (Convergence)**.** *Let $\theta^* \in \mathbb{R}^p$ denote the optimal solution to Problem 42. Let $R \in \mathbb{R}^{p\times p}$ have $\lambda_{min}(R) \geq \delta - 1$ and $\lambda_{max}(R) \leq \delta + 1$ for some $\delta \in (4, \infty)$. Using $\mu = 1/\delta$ and the update step in Eq. 43, the algorithm 41 obeys*

$$\|\theta^{(i)} - \theta^*\|_2 \leq \left(\frac{4}{\delta}\right)^i \|\theta^{(0)} - \theta^*\|_2 + \frac{4}{\delta}\left(\sqrt{(\delta+1)\operatorname{Tr}(R)}\|\theta^*\|_2 + \eta\right), \tag{44}$$

for any number of iterations $i \in \mathbb{N}_+$ and initialized at any point $\theta^{(0)} \in \mathbb{R}^p$.

Thm. 5.1 gives an explicit upper bound on the ℓ_2-error of projected gradient descent for the learning problem 42 in terms of the computational cost (i.e., iterations i) and the properties of the regularization matrix R. It also provides the learning rate μ, which depends on the range of eigenvalues of R, for which projected gradient descent converges linearly to the residual (the second term in the right-hand side). Furthermore, it shows the dependence of this residual on the contribution η of the log-likelihood gradient to the update step 43.

5.4 A corollary on the time-complexity of the projected gradient method

Thm. 5.1 shows that, in the limit of iterations i, the estimates converge linearly to a residual of radius $\mu\left(\frac{1}{1-\frac{2}{\delta}}\right)\left(\sqrt{(\delta+1)\operatorname{Tr}(R)}\|\theta^*\|_2 + \eta\right)$. This radius would fluctuate with $\xi^{(i)}$ in the absence of the proposed modification in Eq. 43. The practitioner may be tempted to reduce η in this modified update step in order to reduce the residual in Thm. 5.1, but we caution that sufficiently low values of η can cause gradient descent to diverge from the minimum. Intuitively, setting η to 0, for example, effectively removes the coupling between the objective and the constraint set Θ in Problem 42, and the estimate produced by gradient descent for the resulting unconstrained problem may drastically differ from the optimal solution to the constrained problem. Determining a schedule for η remains a problem for future study.

The following corollary provides the minimum number of iterations required for the estimates produced by the algorithm 41 to converge to the residual within a neighborhood of size ϵ.

Corollary 5.2. *Given an error bound $\epsilon > 0$, gradient descent yields an estimate $\theta^{(i)}$ such that, after a number of iterations $i > \ln(\epsilon)/\ln(2/\delta)$,*

$$\|\theta^{(i)} - \theta^*\|_2 - 4\mu\left(\sqrt{(\delta+1)\operatorname{Tr}(R)}\|\theta^*\|_2 + \eta\right) \leq \epsilon.$$

Proof. First observe that

$$x^{1/\ln(x)} = e. \tag{45}$$

To see this, let y be a real number such that

$$x^{1/\ln(x)} = y. \tag{46}$$

Taking $\ln(\cdot)$, we see that y must be e. Letting $x = 2/\delta$ and raising both sides of Eq. 45 to the power of $\ln(\epsilon)$, we have

$$\left(\frac{4}{\delta}\right)^{\ln(\epsilon)/\ln(x)} = \epsilon.$$

From the statement of the corollary, assume $i > \ln(\epsilon)/\ln(2/\delta)$. Then, noting that $4/\delta < 1$, we have

$$\left(\frac{4}{\delta}\right)^i < \epsilon.$$

Applying this inequality to Thm. 5.1, we have the result. □

6 Proof of the convergence theorem

The proof of the main theorem shows that, the error of parameter estimates produced by projected gradient descent can be expressed as a sum of two terms, the *leading* and the *residual*. The leading term is bounded with the aid of Lemma 6.2, and the residual is bounded with Lemma 6.3, to be given in the next section. Both lemmas follow straightforwardly from a notion that we refer to as (δ, γ)-restrictedness, which we introduce in Def. 6.1. Meanwhile, Lemma 6.1 provides a sufficient condition for Def. 6.1 to be satisfied. Some additional lemmas regarding the properties of Euclidean projections are also needed for handling the optimization constraints of the learning problem. We present all of these tools here before embarking on the main proof.

6.1 (δ, γ)-restrictedness of matrices and useful properties of Euclidean projections

Definition 6.1 ((δ, γ)-restrictedness)**.** *A matrix $M \in \mathbb{R}^{p_1 \times p_2}$ is said to be (δ, γ)-restricted over $V \subseteq \mathbb{R}^{p_2}$ if there exists $\gamma \in \mathbb{R}_+$ such that*

$$\left| \|Mv\|_2^2 - \delta \|v\|_2^2 \right| \leq \gamma,$$

for some $\delta \in \mathbb{R}_{++}$ dependent only on M and for all $v \in V$.

As δ is only dependent on M, the above definition allows the possibility of restricting M w.r.t. the same δ for different choices of V. We employ this observation in a later result.

The following lemma gives a sufficient condition for a matrix to be (δ, γ)-restricted.

Lemma 6.1. *Suppose $p_1 = p_2 \equiv p$ in Def. 6.1. If the eigenvalues of $M^T M$ are bounded above and below by $\delta + 1$ and $\delta - 1$, respectively, then M is (δ, γ)-restricted over $\mathbb{R}^p$.*

Proof. The result follows from the variational characterization of eigenvalues for Hermitian matrices [8]. Since $M^T M$ is Hermitian, letting $\gamma = \|v\|_2^2$ in Def. 6.1 gives

$$v^T M^T M v \leq \lambda_{max}(M^T M) v^T v = (\delta + 1)\gamma,$$

for all $v \in \mathbb{R}^p$, which gives rise to one inequality in Def. 6.1. The other inequality follows from a similar argument using $\lambda_{min}(M^T M)$. □

Although Lemma 6.1 does not require M to be Hermitian, we note that if $M^T M$ is diagonal, then M would indeed be Hermitian (and diagonal), a fact which will later be used.

The next lemma is a convenient restatement of Def. 6.1 which will be crucial to the main proof. Along with Def. 6.1, it plays a role in this article similar to the role played by Gordon's Escape Through the Mesh lemma [5] in a proof that appears in [15], from which we take inspiration for the main strategy in this article.

Lemma 6.2. *Define $\mathcal{V}_- = \{v_- = v_1 - v_2 \mid v_1, v_2 \in \mathbb{S}^{p-1}\}$ and $\mathcal{V}_+ = \{v_+ = v_1 + v_2 \mid v_1, v_2 \in \mathbb{S}^{p-1}\}$. If $M \in \mathbb{R}^{p \times p}$ is (δ, γ_-)-restricted over $\mathcal{V}_-$, then for all v_-,*

$$\frac{1}{\delta}\|Mv_-\|_2^2 \leq \|v_-\|_2^2 + \frac{\gamma_-}{\delta}.$$

Furthermore, if M is (δ, γ_+)-restricted over $\mathcal{V}_+$, then for all v_+,

$$\frac{1}{\delta}\|Mv_+\|_2^2 \geq \max\left\{0,\ \|v_+\|_2^2 - \frac{\gamma_+}{\delta}\right\}.$$

Proof. The first statement follows directly from Def. 6.1. The second also follows directly by noting that the left-hand side can only be positive and is hence the max between 0 and a potentially negative number. □

It was not necessary to introduce the sets $\mathcal{V}_-$ and $\mathcal{V}_+$ in the previous lemma, but it will greatly facilitate later analysis.

The next lemma will allow us to determine the neighborhood in which the estimation error will lie after sufficiently many iterations. As mentioned, we refer to this neighborhood as the residual of the error.

Lemma 6.3. *If $M \in \mathbb{R}^{p\times p}$ is diagonal, positive semidefinite, and (δ,γ)-restricted over $\mathcal{V} \subseteq \mathbb{R}^p$, then*

$$\|M^2 v\|_2 \leq \sqrt{(\delta+1)\operatorname{Tr}(M^2)}\|v\|_2,$$

for all $v \in \mathcal{V}$.

Proof.

$$\begin{aligned}\|M^2 v\|_2 &\leq \sqrt{\|M\|_F^2 \|Mv\|_2^2} \\ &= \sqrt{\operatorname{Tr}(M^2)}\|Mv\|_2.\end{aligned}$$

Applying the definition of (δ,γ)-restrictedness to the last expression gives the result. □

The following lemmas will be useful in working with projection operators.

Lemma 6.4 (Oymak, et al.)**.** *[15] Let $\Theta \subset \mathbb{R}^p$ be a closed and non-empty set that contains 0. Let $\mathcal{C}$ be a closed and non-empty cone such that $\Theta \subset \mathcal{C}$. Then for all $v \in \mathbb{R}^p$,*

$$\|Proj_\Theta(v)\|_{\ell_2} \leq 2\|Proj_{\mathcal{C}}(v)\|_{\ell_2} \tag{47}$$

Proof. Please refer to [15]. □

Lemma 6.5 (Oymak, et al.)**.** *[15] Suppose $\Theta \subset \mathbb{R}^p$ is a closed set. For all $x \in \mathbb{R}^p$, the projection operator of 41 obeys*

$$Proj_\Theta(x+v) = Proj_{\Theta-\{x\}}(v) \tag{48}$$

Proof. Please refer to [15]. □

6.2 Proof of Theorem 5.1

$$\begin{aligned}&\|\theta^{(i+1)} - \theta^*\|_2 \\ &= \|\mathrm{Proj}_\Theta\left(\theta^{(i)} - \mu(R\theta^{(i)} + \xi^{(i)})\right) - \theta^*\|_2 &(49)\\ &= \|\mathrm{Proj}_{\Theta-\{\theta^*\}}\left(\theta^{(i)} - \mu(R\theta^{(i)} + \xi^{(i)}) - \theta^*\right)\|_2 &(50)\\ &\leq 2\|\mathrm{Proj}_{\mathcal{C}}\left(\theta^{(i)} - \mu(R\theta^{(i)} + \xi^{(i)}) - \theta^*\right)\|_2 &(51)\\ &= 2\|\mathrm{Proj}_{\mathcal{C}}\left((I-\mu R)(\theta^{(i)} - \theta^*) - \mu(R\theta^* + \xi^{(i)})\right)\|_2 \\ &= 2\left(\sup_{v_1 \in \mathbb{S}^{p-1}\cap\mathcal{C}} v_1^T(I-\mu R)(\theta^{(i)} - \theta^*) + \mu\left(\|R\theta^*\|_2 + \eta\right)\right), &(52)\end{aligned}$$

where Eq. 49 follows by definition of the update step, Eq. 50 follows from Lemma 6.5, Eq. 51 follows from Lemma 6.4, and Eq. 52 follows from the definition of $\xi^{(i)}$ in 43. Again, the two terms of Eq. 52 are referred to as the leading and the residual.

We now proceed to bound the leading term. First, we observe that it can be expressed as a product involving two unit vectors by normalizing the error:

$$\begin{aligned} &v_1^T(I-\mu R)(\theta^{(i)}-\theta^*) \\ &\quad= \left(v_1^T(I-\mu R)\frac{\theta^{(i)}-\theta^*}{\|\theta^{(i)}-\theta^*\|_2}\right)\|\theta^{(i)}-\theta^*\|_2. \end{aligned}$$

To simplify the presentation, define

$$v_2 = \frac{\theta^{(i)}-\theta^*}{\|\theta^{(i)}-\theta^*\|_2}.$$

Then, note that

$$\begin{aligned} v_1^T(I-\mu R)v_2 &= \frac{1}{4}[(v_1+v_2)^T(I-\mu R)(v_1+v_2) \\ &\quad -(v_1-v_2)^T(I-\mu R)(v_1-v_2)] \\ &= \frac{1}{4}\left[\|v_1+v_2\|_2^2-\mu\|R^{1/2}(v_1+v_2)\|_2^2\right] \\ &\quad -\frac{1}{4}\left[\|v_1-v_2\|_2^2-\mu\|R^{1/2}(v_1-v_2)\|_2^2\right]. \end{aligned} \tag{53}$$

We may now apply Lemma 6.2 using $M = R^{1/2}$, $\gamma_- = \|v_-\|_2^2$, and $\gamma_+ = \|v_+\|_2^2$. These values of γ are valid as we assume the eigenvalues of R to be bounded in the sense of Lemma 6.1. With the chosen value of γ_+, and with $\delta > 2$ as in the statement of the theorem, we observe that

$$\gamma_+ - \frac{\gamma_+}{\delta} > 0. \tag{54}$$

Thus, the second bound in Lemma 6.2 becomes

$$\frac{1}{\delta}\|Mv_+\|_2^2 \geq \|v_+\|_2^2 - \frac{\gamma_+}{\delta}.$$

By setting $\mu = 1/\delta$ in Eq. 53 and applying Lemma 6.2,

$$\begin{aligned} v_1^T(I-\mu R)v_2 &= \frac{1}{4}\left[\gamma_+ - \frac{1}{\delta}\|R^{1/2}v_+\|_2^2\right] \\ &\quad -\frac{1}{4}\left[\gamma_- - \frac{1}{\delta}\|R^{1/2}v_-\|_2^2\right] \\ &\leq \frac{1}{4}\left[\gamma_+ - \frac{1}{\delta}\|R^{1/2}v_+\|_2^2\right] - \frac{1}{4}\left[-\frac{\gamma_-}{\delta}\right] \\ &\leq \frac{1}{4}\left(\frac{\gamma_+ + \gamma_-}{\delta}\right) \\ &\leq \frac{2}{\delta}, \end{aligned} \tag{55}$$

where the last step follows by observing that

$$\sup_{v_-\in\mathcal{V}_-\cap\mathcal{C}} \gamma_- = \sup_{v_+\in\mathcal{V}_+\cap\mathcal{C}} \gamma_+ = 4. \tag{56}$$

Next, by applying Ineq. 55 to Eq. 52 and applying Lemma 6.3 to the residual term, we obtain a bound for the error at iteration $i+1$ given the previous iteration i:

$$\begin{aligned} \|\theta^{(i+1)}-\theta^*\|_2 &\leq 2\left[\left(\frac{2}{\delta}\right)\|\theta^{(i)}-\theta^*\|_2 + \mu\left(\sqrt{(\delta+1)\operatorname{Tr}(R)}\|\theta^*\|_2+\eta\right)\right] \\ &\leq \left(\frac{4}{\delta}\right)\|\theta^{(i)}-\theta^*\|_2 + 2\mu\left(\sqrt{(\delta+1)\operatorname{Tr}(R)}\|\theta^*\|_2+\eta\right) \\ &\equiv \varrho_1\|\theta^{(i)}-\theta^*\|_2 + \mu\varrho_2, \end{aligned}$$

where we have introduced the notation in the last step in order to simplify the following recursive argument.

$$\begin{aligned}\|\theta^{(i+1)} - \theta^*\|_2 &\leq \varrho_1 \left(\varrho_1 \|\theta^{(i-1)} - \theta^*\|_2 + \mu\varrho_2\right) + \mu\varrho_2 \\ &= \varrho_1^2 \|\theta^{(i-1)} - \theta^*\|_2 + \mu\left(\varrho_1 + 1\right)\varrho_2. \end{aligned} \tag{57}$$

After applying the recursion sufficiently many times to produce $\theta^{(0)}$, the first term of Eq. 57 becomes

$$\varrho_1^i \|\theta^{(0)} - \theta^*\|_2,$$

and the second term of Eq. 57 becomes

$$\mu\varrho_2 \sum_{j=1}^{i-1} \varrho_1^j.$$

Note that, for all i, the previous expression can be bounded as

$$\mu\varrho_2 \sum_{j=1}^{i-1} \varrho_1^j \leq 2\mu\varrho_2,$$

because $\varrho_1 < 1$ by the restriction on δ given in the statement of Thm. 5.1. With the appropriate substitutions for ϱ_1 and ϱ_2, this concludes the proof of Thm. 5.1.

7 Conclusion

We have studied tradeoffs between key design variables in the problems of decoding and learning for HMMs.

In particular, in the decoding problem, we saw that, with sufficiently large amounts of data, the achievement of a certain level of error implied that a certain inequality involving the design variables would necessarily be satisfied. We illustrated this bound in multiple regimes and with respect to several combinations of design variables. Furthermore, we showed in the proof of this result that our formulation of the problem played an essential role in enabling the use of contemporary results in information theory. However, we mention here that an important discussion was given in the concluding section of Section 3 about this formulation. In that section, we argued that the formulation neglected crucial aspects of the HMM that should be considered in a comprehensive theory of tradeoffs in decoding, namely the aspects of the transition kernel and the learning algorithm that are expected to influence the decoder in practice. It remains in open problem to account for these aspects in future research on this topic.

In the learning problem, we saw that, with a specific choice of a learning rate that depended on the restrictedness of the regularization matrix, it can be shown that the estimates produced by projected gradient descent would converge to a certain region which also depended on properties of the regularization matrix. Here, we emphasize that the proof of our result depended heavily on the presence of a specific regularization term in the objective optimized by gradient descent. It remains an open problem to study tradeoffs in learning in even greater generality by considering a broader class of objectives. Furthermore, we argue that it would be of interest to consider other learning procedures than gradient descent, such as the Baum-Welch method which sees arguably more frequent use in practice.

The results presented in this work were a first attempt toward truly fundamental tradeoffs. As discussed, the term "fundamental" here suggests the generality of results across all algorithms in decoding (given a class of noise) and across all objectives in learning (given a class of algorithms). Clearly, the meaning was not entirely realized for a number of potential reasons, due (for instance) to the neglect of the role of learning in the decoding problem and the introduction of regularization in the learning problem. Future research will aim to address these shortcomings, using the ideas presented in this work one possible foundation.

References

[1] Brigham Anderson and Andrew Moore. "Active Learning for Hidden Markov Models: Objective Functions and Algorithms". In: *Proceedings of the 22Nd International Conference on Machine Learning*. ICML '05. Bonn, Germany: ACM, 2005, pp. 9–16. ISBN: 1-59593-180-5. DOI: 10.1145/1102351.1102353. URL: http://doi.acm.org/10.1145/1102351.1102353.

[2] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[3] William Feller. *An introduction to probability theory and its applications. Vol. II.* Second edition. New York: John Wiley & Sons Inc., 1971.

[4] Zoubin Ghahramani. "Hidden Markov Models". In: River Edge, NJ, USA: World Scientific Publishing Co., Inc., 2002. Chap. An Introduction to Hidden Markov Models and Bayesian Networks, pp. 9–42. ISBN: 981-02-4564-5. URL: http://dl.acm.org/citation.cfm?id=505741.505743.

[5] Y. Gordon. "On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$". In: *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 1986–87*. Ed. by Joram Lindenstrauss and Vitali D. Milman. Berlin, Heidelberg: Springer Berlin Heidelberg, 1988, pp. 84–106. ISBN: 978-3-540-39235-4. DOI: 10.1007/BFb0081737. URL: http://dx.doi.org/10.1007/BFb0081737.

[6] Robert A. Granat. *A Method of Hidden Markov Model Optimization for Use with Geophysical Data Sets*. 1980.

[7] Ramon van Handel. *Hidden Markov Models: Lecture Notes*. 2008. URL: https://www.princeton.edu/~rvan/orf557/hmm080728.pdf.

[8] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. New York, NY, USA: Cambridge University Press, 1986. ISBN: 0-521-30586-1.

[9] A. N. Iusem. "On the convergence properties of the projected gradient method for convex optimization". en. In: *Computational & Applied Mathematics* 22 (2003), pp. 37–52. ISSN: 1807-0302.

[10] Hamed Karimi, Julie Nutini, and Mark Schmidt. "Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition". In: (2016). URL: http://arxiv.org/abs/1608.04636v2.

[11] Wael Khreich et al. "A Survey of Techniques for Incremental Learning of HMM Parameters". In: *Inf. Sci.* 197 (Aug. 2012), pp. 105–130. ISSN: 0020-0255. DOI: 10.1016/j.ins.2012.02.017. URL: http://dx.doi.org/10.1016/j.ins.2012.02.017.

[12] H. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003. DOI: 10.1007/b97441. URL: http://dx.doi.org/10.1007/b97441.

[13] K. Markov, S. Nakagawa, and S. Nakamura. "Discriminative training of HMM using maximum normalized likelihood algorithm". In: *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*. Vol. 1. 2001, 497–500 vol.1. DOI: 10.1109/ICASSP.2001.940876.

[14] Christoph Neukirchen and Gerhard Rigoll. "Controlling the Complexity of HMM Systems by Regularization". In: NIPS '98. 1998.

[15] Samet Oymak, Benjamin Recht, and Mahdi Soltanolkotabi. "Sharp Time-Data Tradeoffs for Linear Inverse Problems". In: *CoRR* abs/1507.04793 (2015). URL: http://arxiv.org/abs/1507.04793.

[16] Yuriy Polyanskiy and Yihong Wu. *Lecture notes on information theory*. Aug. 2017.

[17] Yury Polyanskiy. "Channel coding: non-asymptotic fundamental limits". PhD thesis. Princeton University, 2010.

[18] Yury Polyanskiy, H. Vincent Poor, and Sergio Verdu. "Channel coding rate in the finite blocklength regime". In: *IEEE Transactions on Information Theory* 56.5 (Apr. 2010). ISSN: 1557-9654. DOI: 10.1109/TIT.2010.2043769.

[19] H. Vincent Poor. *An Introduction to Signal Detection and Estimation (2nd Ed.)* New York, NY, USA: Springer-Verlag New York, Inc., 1994. ISBN: 0-387-94173-8.

[20] Lawrence R. Rabiner. "Readings in Speech Recognition". In: ed. by Alex Waibel and Kai-Fu Lee. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990. Chap. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, pp. 267–296. ISBN: 1-55860-124-4. URL: http://dl.acm.org/citation.cfm?id=108235.108253.

[21] George Saon and Daniel Povey. "Penalty function maximization for large margin HMM training". In: INTERSPEECH '08. 2008.

[22] Fei Sha and Lawrence K. Saul. "Large Margin Hidden Markov Models for Automatic Speech Recognition". In: *Advances in Neural Information Processing Systems 19*. Ed. by B. Schölkopf, J. C. Platt, and T. Hoffman. MIT Press, 2007, pp. 1249–1256. URL: http://papers.nips.cc/paper/3051-large-margin-hidden-markov-models-for-automatic-speech-recognition.pdf.