Fundamental tradeoffs in inference of finite-state hidden Markov processes

Justin Le, *Student Member, IEEE,* and Pushkin Kachroo, *Senior Member, IEEE*

***Abstract*—We study the tradeoff between variables considered in the design of state-sequence estimators of hidden Markov models. The variables to be considered are: the average rate of error in inferring a sequence of hidden states given a sequence of observed states, the length of these sequences, the dimensionality of the observed state-space, and the cardinality of the hidden state-space. The result can be viewed as a probabilistic corollary to the fundamental limit of channel coding in the finite blocklength regime, given an appropriate interpretation of our proposed model. The implications and requirements of this interpretation remain open to study.**

I. 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 real-time sensor data. In this article, we consider the popular hidden Markov model (HMM) as a framework for this problem. As we allow the observed state-space to be real-valued, we use the term *emission kernel* rather than emission matrix when referring to the probability of an observed state conditioned on a hidden state. Although we take the hidden state-space to be discrete, we use the term *transition kernel* for consistency.

To construct the HMM, examples of temporal sequences of these two states must be provided to an algorithm (i.e., the learning algorithm) in order to compute the parameters of the kernels. After learning, a decoder (or state estimator) applies the learned model to a new sequence of observed states in order to estimate the sequence of hidden states that most likely generated these observations. The study and application of these algorithms have seen enormous progress over the past half-century (see, for instance, [3] for an overview), and there is no solution to the learning problem which is "best" in any general sense. For countable hidden state-spaces, the Viterbi algorithm yields an optimum to the decoding problem, but whether this solution is equal to the true state sequence depends on the rare situation in which the learned model parameters accurately capture the probabilistic relations of the real-world processes. Similar remarks can be made for any other state-estimation procedure. However, relatively little study has been devoted to understanding how the design parameters of the HMM trade off with each other quantitatively and what this tradeoff implies for practicing engineers.

In this article, we demonstrate one such tradeoff. In particular, we derive an inequality that exhibits the relationship between a number of essential design parameters shared by all HMMs: the length of state-sequences, the dimensionalities of the spaces to which these states belong, and the expected rate of error of the inference algorithm.

Our main result gives a precise quantitative statement of a conjecture that arises from basic intuition:

The average error rate of a state estimator is universally lower-bounded by a function which increases with the hidden state-space dimensionality and which decreases with the observed state-space dimensionality and the duration of observation sequences.

Roughly speaking, the intuition is that high-dimensional states are difficult to estimate, but the difficulty should be relieved by the availability of more data (whether in terms of the number of sensors or the resolution of each sensor). Naturally, a sufficiently large amount of data may instead cause the opposite effect, but this regime of data intake is beyond the scope of this work.

The above intuition is further bolstered by results in non-asymptotic information theory. For example, for channel coding in the finite-blocklength regime, the fundamental tradeoff is known between the blocklength, the decoder error rate, and the size of the source alphabet [6]. This observation motivates the proof of our main result, in which we show one possible interpretation of these variables in the context in decoding the hidden state sequence of an HMM. The requirements and implications of this interpretation are not straightforward, however, and we leave most of this discussion to future work.

II. Problem formulation

A. State-space and probability structure

Consider an uncountable set $\mathcal{L}$ of discrete-time hidden Markov models, each of which we denote as a pair $L = (\mathcal{T}_L, \mathcal{E}_L)$ consisting of a transition kernel and emission kernel, respectively, each of which is defined for all time t. For a given $L \in \mathcal{L}$, the hidden state $K_L(t)$ at any $t \in \mathbb{N}/\{0\}$ is a random variable $K : \mathcal{K} \to \{1, \ldots, |\mathcal{K}|\}$ over the space $(\mathcal{K}, 2^{\mathcal{K}}, p_K)$, where $\mathcal{K}$ is a fixed finite set unknown to us. Furthermore, for each L, the observable state $Y_L^m(t)$ at any t is a random vector taking values in $\mathbb{R}^m$, for which a probability density function (p.d.f.) exists. Since $Y_L^m(t)$ is dependent only on $K_L(t)$ (by the structure of the HMM), and their marginal probabilities are specified as above, it suffices to specify the conditional probabilities between $K_L(t)$ and $Y_L^m(t)$ (i.e., the emission kernel) in order to uniquely specify the joint distribution between them. Similarly, it suffices to give a conditional probabilities between hidden states over time (i.e., the transition kernel) in order to specify their joint probability structure at any point in time.

J. Le and P. Kachroo are with the Department of Electrical and Computer Engineering, University of Nevada, Las Vegas. E-mail: justin.le@unlv.edu.

Updated Nov. 4, 2017.

B. Learning and decoding

The learning procedure computes a model in $\mathcal{L}$ given a set of ℓ-length sequences, where each sequence has the form $\{(Y^m(t), K(t)) : t \leq T\}$. This set of sequences is referred to as the training set. A single sequence from this set is referred to as a training example. T is a time beyond which no training examples are available.

Given the learned model and a new ℓ-length sequence of observed states $\{Y^m(t) : t > T\}$, the decoder computes an ℓ-length sequence of hidden states that best "explains" this newly observed sequence (e.g., in the sense of Problem 2 in [8]). In this paper, roughly speaking, we ask: on average, how well does any inference procedure perform under any learning procedure? More precisely, how does inference error vary with the hyperparameters ℓ, m, and $|\mathcal{K}|$, under any procedures of learning and inference? We attempt to study this question in a manner that is entirely agnostic to the specific procedures that can be used for these tasks.

C. Assumptions

Note that a sequence $K^t = \{K(t),\ K(t-1),\ \ldots,\ K(t-\ell+1)\}$ belongs to a set of cardinality $|\mathcal{K}|^\ell$. We index this set by

$$W(t) \equiv W(t, K^t; \ell).$$

At any t, $W(t)$ is a random variable taking values in $\{1, \ldots, |\mathcal{K}|^\ell\}$ according to some p.m.f. p_W.

Fix a marginal distribution for the observed states. Then, note that fixing $L \in \mathcal{L}$ would fix p_K and thus fix p_W. Then we see that

$$S_W \equiv \max_{w \in \{1,\ldots,|\mathcal{K}|^\ell\}} p_W(w) \tag{1}$$

is a random variable over the space $(\mathcal{L}, \sigma\mathcal{L}, \mu_L)$, where we assume that μ_L is such that the p.d.f. p_S exists for S_W. Here, we view the values of p_W as the values of a real-valued random vector over $\mathcal{L}$ having unit ℓ_1-norm. We assume that the fixed marginal distribution of observed states is such that

$$Y(t) \equiv b(t) f(W(t)) + Z(t), \tag{2}$$

where $Y(t) \in \mathbb{R}^{\ell m}$ is the concatenation of the sequence $\{Y^m(t),\ Y^m(t-1),\ \ldots,\ Y^m(t-\ell+1)\}$, $Z(t)$ at any t is a standard Gaussian vector, independent over time, taking values in $\mathbb{R}^{\ell m}$, $f : \mathcal{K} \to \mathbb{R}^{\ell m}$ is injective, and $b(t) \in \mathbb{R}^{++}$ ensures that, for all t,

$$\|b(t) f(W(t))\|_2^2 \leq \ell m P, \tag{3}$$

for some fixed $P \in \mathbb{R}^{++}$ which is known.

The inference algorithm, which we denote by $g : \mathbb{R}^{\ell m} \to \{1, \ldots, |\mathcal{K}|^\ell\}$, is viewed as a transition kernel (unrelated to any HMM kernels) that maps a realization of $Y(t)$ to an element of $\{1, \ldots, |\mathcal{K}|^\ell\}$ with some probability determined by the learning and inference procedures.

Define the average probability of error (in inference) at any t as

$$\epsilon = 1 - \frac{1}{|\mathcal{K}|^\ell} \sum_{w=1}^{|\mathcal{K}|^\ell} \mathbb{P}[\{W(t) = w\} \cap \{g(Y(t)) = w\}]. \tag{4}$$

We will require the following assumption on S_W.

Assumption II.1. *For a given $\mathcal{K}$ with at least two elements, there exists $\ell^* \in \mathbb{N}/\{0\}$ and an indexing procedure to generate W such that*

$$\mathbb{E}[S_W] \leq \frac{1 + c_1}{|\mathcal{K}|^\ell}, \tag{5}$$

where the expectation is taken over $\mathcal{L}$, and the real constant $c_1 \in (0, 1)$ can be tuned using the training set.

D. Technical overview

We study the tradeoffs between the following:

- the generalization error ϵ of a hidden Markov model in inferring a sequence of hidden states given a sequence of observed states
- the length ℓ of the above sequences
- the dimensionality m of the observed state-space
- the cardinality $|\mathcal{K}|$ of the hidden state-space

Toward this end, we consider an HMM with a finite hidden state-space and continuous observed states taking values in $\mathbb{R}^m$. Ultimately, the tradeoff is a probabilistic corollary to the non-asmymptotic fundamental limit of channel coding derived in [6], in which the probabilistic nature arises from the uncertainty associated with drawing HMMs at random from a collection $\mathcal{L}$.

We now state the main result and sketch the proof.

Theorem II.1. *Instantiate the prevailing notions, and choose ℓ and $\mathcal{K}$. Then, there exists $N(P, \epsilon) \in \mathbb{N}$ such that, for all $n > N(P, \epsilon)$,*

$$\epsilon \geq Q\left[\frac{\log \mathcal{E} + nC(P) + \frac{1}{2}\log n + g_c(P, \epsilon)}{\sqrt{nV(P)}}\right] \tag{6}$$

holds with probability at least

$$1 - \exp\left(-\frac{2c_0^2}{(1 - 1/|\mathcal{K}|^\ell)^2}\right). \tag{7}$$

where

$$n = \ell m, \tag{8}$$

$$\mathcal{E} = \mathbb{E}[S_W] + c_0, \tag{9}$$

$$C(P) = \frac{1}{2}\log(1 + P), \tag{10}$$

$$V(P) = \frac{P}{2}\frac{P+2}{(P+1)^2}\log^2(e), \tag{11}$$

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

First, we represent the emissions of an HMM as the Gaussian channel considered in the channel coding problem [6]. Then, we obtain the meta-converse through a Markov chain representation

as in [5] but with a modification involving a Hoeffding inequality to account for the non-uniformity in source distribution (i.e., the hidden state index distribution of the HMM). The remainder of the derivation follows straightforwardly from the strategy in [5]. The meta-converse can be further bounded by the Neyman-Pearson lemma to pass to a likelihood ratio (or Radon-Nikodym derivative) that evaluates to an expression to which the Berry-Esseen theorem is applicable.

E. Notation

Logarithms are taken to be base 2. $B^n(r)$ denotes the ball in $\mathbb{R}^n$ with radius r centered at the origin. The σ-algebra generated by a set A is denoted by σA.

III. Proof of Theorem II.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. (We discuss the temporal aspects of our work in a later section.) Aside from Lemma III.2, the proof is adapted from the strategies found in [6] and [5], in which a similar approach was developed for obtaining fundamental limits of channel coding in the finite blocklength regime.

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 "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,

$$\begin{aligned} &\mathbb{P}[g(Y) = W|X, Y] \\ &= \sum_{w=1}^{|\mathcal{K}|^\ell} \mathbb{P}[\{g(Y) = w\} \cap \{W = w\}|X, Y] \\ &= \sum_{w=1}^{|\mathcal{K}|^\ell} \mathbb{P}[g(Y) = w|X, Y]\mathbb{P}[W = w|X, Y], \\ &= \sum_{w=1}^{|\mathcal{K}|^\ell} \mathbb{P}[g(Y) = w|X]\mathbb{P}[W = w|Y]. \end{aligned} \tag{12}$$

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 probability that $g(Y)$ yields an incorrect estimate of W under P_{XY}. Define ϵ_2 similarly under Q_{XY}. Specifically,

$$1 - \epsilon_1 = \sum_{x \in B^n(nP)} \sum_{y \in \mathbb{R}^n} P_{T|XY}(1|x, y)P_{XY}(x, y), \tag{13}$$

$$1 - \epsilon_2 = \sum_{x \in B^n(nP)} \sum_{y \in \mathbb{R}^n} P_{T|XY}(1|x, y)Q_{XY}(x, y). \tag{14}$$

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

$$\begin{aligned} &\beta_\alpha(P_{XY}, Q_{XY}) \\ &= \inf_{P_{T|XY}} \sum_{x \in B^n(nP)} \sum_{y \in \mathbb{R}^n} P_{T|XY}(1|x, y)Q_{XY}(x, y), \end{aligned} \tag{15}$$

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

$$\sum_{x \in B^n(nP)} \sum_{y \in \mathbb{R}^n} P_{T|XY}(1|x, y)P_{XY}(x, y) \geq \alpha. \tag{16}$$

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

$$\beta_\alpha(P_{XY}, Q_{XY}) \leq 1 - \epsilon_2. \tag{17}$$

We wish for P_{XY} to represent the true joint distribution in Eq. 2, 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

$$\inf_{P_X} \sup_{Q_{XY}} \beta_\alpha(P_X P_{Y|X}, Q_{XY}) \leq 1 - \epsilon_2. \tag{18}$$

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 $\mathsf{S}^{n-1}(nP)$ and later show how our final result can be generalized to the ball $B^n(nP)$. With this restriction, we have that $\beta_\alpha(P_{Y|X=x}, Q_{Y|X=x})$ is independent of x by the radial symmetry of $\mathsf{S}^{n-1}(nP)$, and we may thus apply the following lemma due to [5] for $\mathsf{F} = \mathsf{S}^{n-1}(nP)$.

Lemma III.1 (Polyanskiy)**.** *[5] 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,*

$$\beta_\alpha(P_X P_{Y|X}, Q_X Q_{Y|X}) = \beta_\alpha(P_{Y|X=x}, Q_{Y|X=x}). \tag{19}$$

Due to the same symmetry, $P_{Y|X=x}$ is also independent of x, so by Lemma III.1, Eq. 18 becomes

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

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. 2.

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 [5] for a justification, since we make no changes to the argument.)

We will need the following lemma to bound ϵ_2.

Lemma III.2. *Define S_W as in Assumption II.1, and let $c_0 \in (0, 1-1/|\mathcal{K}|^\ell)$ for a given $\mathcal{K}$. Then,*

$$\mathbb{P}[S_W - \mathbb{E}[S_W] \geq c_1] \leq \exp\left(-\frac{2c_0^2}{(1-1/|\mathcal{K}|^\ell)^2}\right). \quad (20)$$

Proof: Suppose that $S_W < 1/|\mathcal{K}|^\ell$. Then a contradiction arises, for p_W would sum to a number less than 1, and it would no longer qualify as a probability mass function. Thus, S_W takes values a.s. in the interval $[1/|\mathcal{K}|^\ell, 1]$. The statement then follows trivially from the standard Hoeffding bound for a single random variable (e.g., see [1]). ∎

Naturally, Lemma III.2 can be strengthened by Bennett's inequality or by some further gymnastics under additional assumptions, but a simple Hoeffding-type result will suffice in demonstrating the key technique we propose in this work.

Since Q_Y has been chosen to be independent of X, the probability $1-\epsilon_2$ becomes the probability that $g(Y) = W$ when $g(Y)$ chooses with uniform randomness from $\mathcal{K}$. At most, this probability is S_W (by definition of S_W). Thus, Lemma III.2 gives

$$\sup_{Q_Y} \beta_\alpha(Q_Y) \leq \mathbb{E}(S_W) + c_0, \quad (21)$$

with probability at least

$$1 - \exp\left(-\frac{2c_0^2}{(1-1/|\mathcal{K}|^\ell)^2}\right). \quad (22)$$

A. Computing the Berry-Esseen bound

Temporarily, we now consider $\beta_\alpha(\cdot,\cdot)$ with arbitrary probability measures in order to obtain a general bound. The bound will then particularize when we substitute the specific probability distributions in Eq. 21. We may then apply the Berry-Esseen theorem to this particular bound. Finally, in the sequel, we will combine the Berry-Esseen result with the bound on $\beta_\alpha(\cdot,\cdot)$ of the previous section, thus completing the proof of our main theorem.

The following is a variant of the Neyman-Pearson lemma as stated in [7] and [6], which we will need.

Lemma III.3 (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*

$$\beta_\alpha(\mathcal{P},\mathcal{Q}) = \mathcal{Q}[T^*_\alpha = 1] \leq \mathcal{Q}[T = 1], \quad (23)$$

and the optimal test T^*_α is defined by

$$T^*_\alpha(\theta) = \mathbb{1}\left(\frac{d\mathcal{P}}{d\mathcal{Q}} > \gamma\right) + T_\tau \mathbb{1}\left(\frac{d\mathcal{P}}{d\mathcal{Q}} = \gamma\right), \quad (24)$$

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

$$\mathcal{P}[T^*_\alpha = 1] = \alpha. \quad (25)$$

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.

Define the event

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

Then the Radon-Nikodym theorem gives

$$\mathcal{P}[\mathsf{E}] = \int_{\mathsf{E}} \frac{d\mathcal{P}}{d\mathcal{Q}} d\mathcal{Q} < \gamma \int_{\mathsf{E}} d\mathcal{Q} = \gamma \mathcal{Q}[\mathsf{E}]. \quad (26)$$

This "change-of-measure" argument is standard in information theory [4]. 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 III.3. Therefore, letting $\mathcal{P} = P_{Y|X=x}$ and $\mathcal{Q} = Q_Y$ (as derived using Lemma III.1), we have

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

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,

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

Eq. 29 follows from Lemma III.3. 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. 2, 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 $x = [\sqrt{P}\ \sqrt{P}\ \dots\ \sqrt{P}]^T$, 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

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

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

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

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.

Theorem III.1 (Berry-Esseen)**.** *[2] 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*

$$\left|\mathbb{P}\left[\frac{\sum_{i=1}^n(X_i-\mu_i)}{\sqrt{\sum_{i=1}^n\sigma_i^2}} \leq c_1\right] - Q(-c_1)\right| \leq \frac{6\sum_{i=1}^n s_i}{\left(\sum_{i=1}^n\sigma_i\right)^{3/2}}. \quad (33)$$

In Thm. III.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

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

so that we have

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

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

$$\gamma' = -\sqrt{n\sigma_i^2}Q^{-1}(\alpha_n), \quad (34)$$

and in Thm. III.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. 21, taking the logarithm, and rearranging terms,

$$\log(\mathbb{E}[S_W]+c_0) \geq \log\left(\frac{6s_i}{\sigma_i^3}\right) + \gamma' - \frac{1}{2}\log(n) - n\log(\sigma_Y). \quad (35)$$

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 $\left[\alpha - \frac{12 s_i}{\sqrt{n(\sigma_i^2)^3}},\ \alpha\right] \subset (0,1)$ for all $n > N(P,\alpha)$, and let a be any point in this interval. Taking the Taylor expansion of Eq. 34 over this interval about the point α, we have

$$\begin{aligned}
\gamma' &= -\sqrt{n\sigma_i^2}Q^{-1}(\alpha) - (a-\alpha)\sqrt{n\sigma_i^2}\frac{dQ^{-1}}{da}(\alpha) \\
&\geq -\sqrt{n\sigma_i^2}Q^{-1}(\alpha) + \left(\alpha - \frac{12 s_i}{\sqrt{n(\sigma_i^2)^3}} - \alpha\right)\sqrt{n\sigma_i^2}\frac{dQ^{-1}}{da}(\alpha) \\
&= -\sqrt{n\sigma_i^2}Q^{-1}(\alpha) + \frac{12 s_i}{\sigma_i^2}\frac{dQ^{-1}}{da}(\alpha). \qquad (36)
\end{aligned}$$

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

$$\alpha_1 = \alpha - \frac{12 s_i}{\sqrt{N(P,\alpha)(\sigma_i^2)^3}}, \qquad (37)$$

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

$$g_1 = \min_{a_1 \in [\alpha_1,\ \alpha]} \frac{dQ^{-1}}{da}(a_1), \qquad (38)$$

which is continuous in both P and α. Substituting g_1 into Eq. 36 and combining the result with the γ' of Eq. 35, we have

$$\begin{aligned}
&\log(\mathbb{E}[S_W] + c_0) \\
&\geq -\sqrt{n s_i}Q^{-1}(\alpha) - \frac{1}{2}\log(n) - n\log(\sigma_Y) - g_c(P,\alpha),
\end{aligned}$$

where

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

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. 15, we allow

$$Q^{-1}(\alpha) = Q^{-1}(1-\epsilon) = -Q^{-1}(\epsilon), \qquad (39)$$

and the previous bound becomes

$$Q^{-1}(\epsilon) \leq \frac{\log \mathcal{E} + n\log\sigma_Y + \frac{1}{2}\log n + g_c(P,\epsilon)}{\sqrt{n\sigma_i^2}}. \qquad (40)$$

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

References

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

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

[3] 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.

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

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

[6] 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.

[7] 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.

[8] 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.