Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

# Variational Bayes and a problem of reliable communication: II. Infinite systems

**Nigel J Newton**[1,2] **and Sanjoy K Mitter**[2,3]

[1] School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK
[2] Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
E-mail: njn@essex.ac.uk and mitter@mit.edu

Received 15 August 2012
Accepted 19 October 2012
Published 15 November 2012

Online at stacks.iop.org/JSTAT/2012/P11008
doi:10.1088/1742-5468/2012/11/P11008

**Abstract.** We consider a family of estimation problems not admitting conventional analysis because of singularity and measurability issues. We define posterior distributions for the family by a variational technique analogous to that used to define Gibbs measures in statistical mechanics. The family of estimation problems, which arise in the asymptotic analysis of error-control codes, is parametrized by a *code rate*, $R \in (0, \infty)$; this is shown to be analogous to the absolute temperature of statistical mechanics. The family undergoes an (Ehrenfest) first-order phase transition at a critical code rate $C$ (the *channel capacity*), where there is a convex set of posterior distributions. At all other code rates, there is only one posterior distribution; if $R < C$, this is the Dirac measure located at the source sequence, whereas if $R > C$ it has infinite support. In a result reflecting the Dobrushin construction, we show that these *posterior* distributions are *asymptotically consistent* with those of families of finite-sequence error-control codes.

**Keywords:** exact results, source and channel coding, statistical inference

1742-5468/12/P11008+30$33.00

## Contents

## 1. Introduction

This paper uses the techniques of statistical mechanics to define posterior distributions in a family of estimation problems that do not admit conventional analysis because of singularity and measurability issues. The estimation problems arise in the theory of reliable communication, where a random binary sequence (the *source sequence*) has to be communicated over a binary symmetric channel. The latter inverts any bit sent over it with some fixed probability $q \in (0, 1/2)$, and so reliable communication cannot be achieved by the *direct* transmission of the source sequence but requires the use of a code. This maps the source sequence to a binary *code sequence*, which is transmitted over the channel. The code sequence typically includes redundancy, which allows the receiver to correct the more commonly occurring transmission errors introduced by the channel. A (rather crude) example of such a scheme is one in which each source bit is sent $2N + 1$ times, and a majority decision rule is used at the receiver.

In his seminal paper [1], Shannon proposed the use of random block codes, in which $K$-bit source sequences are mapped to $N$-bit code sequences for transmission over the channel. By considering families of such systems, parametrized by $K$, with $N = [R^{-1}K]$ for some fixed *code rate* $R \in (0, \infty)$, he showed that reliable communication can be achieved at all code rates less than a well-defined *channel capacity*, $C$, but not at rates exceeding $C$. In this context, 'reliable communication' means that the probability that the receiver makes source-sequence decoding errors decreases exponentially with $K$. The channel capacity is the value of the *mutual information* between the input and output bits in one use of the channel, maximized over the distribution of the input bit (see, for example, [2] or [3]). In the case of a binary symmetric channel the maximizing distribution is uniform and $C = 1 - b(q)$, where $b{:}[0, 1] \rightarrow [0, 1]$ is the binary entropy function

$$b(\theta) = \begin{cases} -\theta \log \theta - (1-\theta)\log(1-\theta) & \text{if } \theta \in (0,1) \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

**Convention on logarithms.** The base of logarithms throughout this paper is 2; information quantities are, therefore, measured in *bits*. The notation 'exp' is used for the inverse log, i.e. $\exp(x) := 2^x$.

At the receiver, the optimal decoding strategy is to choose the $K$-bit sequence with the *maximum a posteriori probability* of being the source sequence, based on the latter's prior distribution, and observation of the $N$-bit channel output sequence. In this sense, decoding is a problem in Bayesian estimation.

This paper is the second part of a two-part study of this communication problem from the perspective of variational Bayesian methods. The variational interpretation of Bayes' formula is developed in a very general setting in [4], and reviewed in part I [5]. The key result, in the present context, characterizes the posterior distribution as the unique minimizer of a quantity we call the *apparent information* of a probability measure $\tilde{P}$. This comprises two parts: the *relative entropy* of $\tilde{P}$ to the prior, which measures its 'information gain' over the latter, and the average over $\tilde{P}$ of the log-likelihood function associated with the observation.

Part I shows that Shannon's results are connected with *secondary* Bayesian problems, in which the causes of decoding errors (errors in the communication channel and poor outcomes of the random code) are estimated on the basis of observations of the decoding error event or its complement. This gives insight into the dominant cause of errors at different code rates. Part I also finds (large block) scaling limits for the variational information quantities of these secondary estimation problems, as well as those of the *primary* problem of estimating the source sequence. These scaling limits exhibit critical behaviour at particular code rates, including the channel capacity. Shannon's 'reliability function' is recovered as the scaling limit of the apparent information of one of the secondary estimation problems, and the scaling limit of the apparent information in the primary problem illustrates that the channel capacity is associated with an 'information saturation' effect.

In part II, we study a family of infinite-sequence estimation problems (parametrized by the code rate); these reflect the important features of Shannon's result, and provide insight into the fundamental connection between error-control coding and statistical mechanics. They are based on an infinite random code whose outcomes are extremely irregular maps between sequence spaces. This irregularity creates significant problems of analysis, since

J. Stat. Mech. (2012) P11008

the code sequences it produces do not have well-defined distributions. This is an essential feature of any good (infinite) code; it is the infinite-sequence analogue of the property that makes random block codes so effective in error-control coding. The information from each source bit is, in some sense, 'spread' across the entire code sequence. We define posterior distributions for our infinite-sequence estimation problems in a construction that mirrors the variational principle of statistical mechanics. Posterior distributions are analogous to Gibbs measures, in that they minimize *specific* apparent information in the same way that Gibbs measures minimize specific free energy. The prior and posterior distributions are typically mutually singular.

We explore this analogy with statistical mechanics, showing that the code rate, $R$, plays the role of absolute temperature in the estimation problems, and that the latter exhibit (in a very precise sense) an (Ehrenfest) first-order phase transition at the channel capacity. The variational definition gives rise to unique posterior distributions at all code rates except the channel capacity, where any convex combination of two extremes is a posterior distribution. At any code rate below capacity, the posterior distribution is a Dirac measure located at the source sequence. The posterior distributions at code rates above capacity clarify the communication possibilities in that regime.

### 1.1. Lattice systems in statistical mechanics

We review here some of the key ideas from the random field theory of statistical mechanics, since they are the basis of our approach to the infinite-sequence estimation problems discussed above. Readers interested in the full picture are referred to the book by Georgii [6]. We restrict our review to the Ising model, i.e. a random field with index set $\mathbb{Z}^d$, that takes the values ±1 only. The *configuration space* of this random field is the set of all possible outcomes: $\Omega := \{-1,+1\}^{\mathbb{Z}^d}$. Rather than attempting to study the individual *microstates* (outcomes) of such a field, we are often interested in its *macroscopic behaviour* 'at equilibrium'. This is associated with particular probability measures on the configuration space called *Gibbs measures*. The central ansatz of this approach, which may be taken as an axiom in the mathematical theory, is that Gibbs measures are minimizers of *free energy*.

Consider, first, a random field associated with a *finite* lattice, having configuration space $\Omega_A = \{-1,+1\}^{A^d}$, where $A$ is a finite, contiguous set of integers. The *internal energy* of such a system is defined by a function $E : \Omega_A \to \mathbb{R}^+$, which assigns an energy to each microstate. The internal energy of a *macrostate* $\tilde{\Pi}$ (a probability measure on $\Omega_A$) is the average of $E$ over $\tilde{\Pi}$:

$$\langle E, \tilde{\Pi}\rangle := \int_{\Omega_A} E(\omega)\tilde{\Pi}\,(\mathrm{d}\omega),$$

and its *entropy* is the negative of its relative entropy with respect to a product (reference) measure $\lambda^{\otimes A^d}$:

$$\mathcal{S}(\tilde{\Pi}) := -h(\tilde{\Pi}|\lambda^{\otimes A^d}),$$

where $h$ is the relative entropy (Kullback–Leibler divergence) of two measures:

$$h(P|Q) := \begin{cases} \displaystyle\int_{\Omega_A} \log \frac{\mathrm{d}P}{\mathrm{d}Q}(\omega) P\,(\mathrm{d}\omega) & \text{if } P \ll Q \text{ and the integral exists} \\ +\infty & \text{otherwise.} \end{cases} \tag{2}$$

The free energy of the macrostate $\tilde{\Pi}$ is $F(\tilde{\Pi}) := T^{-1}\langle E, \tilde{\Pi}\rangle - S(\tilde{\Pi})$, where $T$ is the absolute temperature. A straightforward calculation shows that the unique minimizer of $F$ at temperature $T$ is the following Gibbs measure:

$$\Pi(B) = \frac{\int_B \exp(-E(\omega)/T)\lambda^{\otimes A^d}\,(\mathrm{d}\omega)}{\int_{\Omega_A} \exp(-E(\omega)/T)\lambda^{\otimes A^d}\,(\mathrm{d}\omega)} \qquad \text{for } B \subseteq \Omega_A. \tag{3}$$

**Remark 1.1.** The connection between Bayesian estimation and statistical mechanics can be seen in its simplest form in the similarity between (3) and Bayes' formula. If $\lambda^{\otimes A^d}$ is a probability measure then it can be regarded as the prior distribution in a Bayesian problem with likelihood function $\exp(-E/T)$ and posterior distribution $\Pi$.

This approach cannot be applied *directly* to infinite lattice systems because of singularity issues. The standard solution to these, pioneered by Dobrushin, Lanford and Ruelle, uses a collection of finite-domain energy functions called a *specification*. For each subset $\Lambda \subseteq \mathbb{Z}^d$, let $\sigma_\Lambda : \Omega \to \{-1,+1\}^\Lambda$ be the co-ordinate map. To each *finite* subset $\Lambda \subset \mathbb{Z}^d$, the specification assigns an energy function $E_\Lambda : \Omega \to \mathbb{R}^+$; this depends on both the *internal variables*, $\sigma_\Lambda(\omega)$, and the *external variables*, $\sigma_{\mathbb{Z}^d\setminus\Lambda}(\omega)$. For any fixed value of the external variables, $E_\Lambda$ expresses the internal energy of the finite subsystem (with configuration space $\{-1,+1\}^\Lambda$) as a function of the internal variables. The external variables are considered to represent part of the *environment* of the finite subsystem. For any fixed environment, a Gibbs measure for the finite subsystem can be defined as in (3). Repeating this procedure for all finite $\Lambda \subset \mathbb{Z}^d$, we obtain a collection of stochastic kernels $\Pi_\Lambda : \Omega \to \mathcal{P}(\mathcal{F})$, where $\mathcal{F}$ is an appropriate $\sigma$-algebra of subsets of $\Omega$ (for example, the product $\sigma$-algebra). Clearly, in order for this to be useful, the stochastic kernels must satisfy some consistency properties. One way of obtaining specifications with such properties is by defining the energy functions $E_\Lambda$ in terms of an *interaction potential* (see chapter 2 of [6]).

**Remark 1.2.** It is convenient in this construction to define the stochastic kernels as maps from $\Omega$ to $\mathcal{P}(\mathcal{F})$, rather than as maps from $\{-1,+1\}^{\mathbb{Z}^d\setminus\Lambda}$ to the set of probability measures on $\{-1,+1\}^\Lambda$; for any $B \subseteq \{-1,+1\}^\Lambda$, $\Pi_\Lambda(\omega)(\sigma_\Lambda^{-1}(B))$ is then the equivalent of $\Pi(B)$ of (3).

A Gibbs measure for the infinite system is a probability measure $\Pi \in \mathcal{P}(\mathcal{F})$ that is *consistent* with the specification in the sense that, for any finite $\Lambda \subset \mathbb{Z}^d$ and any $B \in \mathcal{F}$,

$$\int_\Omega \Pi_\Lambda(\omega)(B)\Pi\,(\mathrm{d}\omega) = \Pi(B). \tag{4}$$

A Gibbs measure is thus a probability measure for which the stochastic kernels $\Pi_\Lambda$ are *regular conditional probabilities*.

This construction is closely related to the Kolmogorov extension theorem for product spaces, the essential difference being that the finite-dimensional *marginal* distributions of Kolmogorov's result are replaced by finite-dimensional *conditional* distributions in the Dobrushin setting. An important consequence of this distinction is the possibility of non-uniqueness of Gibbs measures. In many examples, there are temperatures at which multiple Gibbs measures exist. One of the great successes of the random field theory is its ability to predict and model multiple *phases* of statistical mechanical systems in this way.

A feature of specifications in many applications is *shift invariance*. A shift operator $\theta_K : \Omega \rightarrow \Omega$, with multi-index $K \in \mathbb{Z}^d$, is defined as follows:

$$\theta_K(\omega)_I = \omega_{I+K} \qquad \text{for all } I \in \mathbb{Z}^d.$$

A shift invariant specification is one for which

$$E_{\Lambda+K}(\theta_K(\omega)) = E_\Lambda(\omega) \qquad \text{for all } \omega \in \Omega, \qquad K \in \mathbb{Z}^d \qquad \text{and all finite } \Lambda \subset \mathbb{Z}^d.$$

If a specification is shift invariant then it is natural, when searching for Gibbs measures, to restrict attention to measures sharing this property; although 'symmetry breaking' Gibbs measures are sometimes also of interest.

The variational principle returns to the notion of Gibbs measures as minimizers of free energy, and involves a quantity called the *specific* free energy of a shift invariant measure $\Pi \in \mathcal{P}(\mathcal{F})$. This is defined in terms of large $\Lambda$ limits of the free energies of the finite systems of a shift invariant specification, normalized by the cardinality of the sets $\Lambda$. Under quite relaxed conditions, a shift invariant measure $\Pi$ is a Gibbs measure, as defined by the consistency property (4), if and only if it minimizes the specific free energy (see chapter 15 in [6]).

### 1.2. Overview of sections 2–5

The arrangement of material in the remainder of the paper is as follows. Section 2 introduces the infinite-sequence estimation problems we address, and shows that they do not admit conventional analysis because of the irregularity of the code. It investigates the use of posterior distributions for finite-sequence sub-problems as potential 'specifications' for the infinite-sequence problems (in the sense of section 1.1), and highlights a number of difficulties with this approach. At code rates greater than the channel capacity, the observation-conditional distributions of the finite-sequence sub-problems behave erratically as the sequence length increases; their information gains over the prior manifest themselves in ever changing dependences between the individual source bits. To overcome this problem we introduce, in section 3, a family of one-to-one transformations of the estimand space, thereby introducing an auxiliary estimand sequence $V$. The finite-sequence estimation problems associated with $V$ have the necessary regularity to be used in the construction of posterior distributions by the methods of statistical mechanics.

Posterior distributions for $V$ are defined in section 4 to be minimizers of a specific apparent information. This definition leads to unique posterior distributions at all code rates except the channel capacity. The nature of posterior distributions, and their connections with those of finite block coding systems, are investigated.

Section 5 develops an alternative definition of posterior distributions based on an asymptotic version of the consistency property (4). The 'asymptotics' here involve weak convergence. We show that this definition is fundamentally less satisfactory than the variational definition, since it gives rise to multiple posterior distributions at all code rates less than capacity. This is because it fails to take proper account of information on the 'tail' of $V$.

### 1.3. Related work

The use of *random* block codes, as pioneered by Shannon, is a simple way of defining a family of codes (indexed by source-sequence length) with some common defining feature when studying asymptotics. As a result of the strong law of large numbers, random $(K, N)$ block codes have good performance in the limit of large $K$. However, they require very complex decoders; even the *specification* of the code requires tables that are exponentially large in $K$. Much of the research on coding for noisy channels carried out since 1948 has been on the development of codes that admit simpler decoders, and a major role in this quest has been played by *linear codes*. Until 1993 the practice of error-control coding was dominated by (linear) *convolutional codes* exploiting finite-state machines in the encoders, and dynamic programming in the decoders [7]. However, in the last 18 years, the important classes of *turbo codes* and *low density parity check (LDPC) codes* have dominated the research literature since they admit simple *iterative* decoders, and yet are capable of achieving reliable communication at rates close to capacity [3].

In 1989, Sourlas [8] developed a statistical mechanical interpretation of a type of LDPC error-control code. Since then there has been much research effort to bring the solution techniques for the so-called 'spin glasses' of statistical mechanics to bear on coding theory [9]–[13], [3]. Many of these references concern LDPC codes, which correspond to spin glasses with 'dilute' connectivity—the crucial property leading to low complexity decoding. In this methodology, the spin glass system actually corresponds to the MAP *bit* decoder of the error-control system, i.e. the decoder that maximizes the posterior *marginal* distributions of the individual bits. The MAP sequence decoder is recovered in [8] by the introduction of a new positive 'hyperparameter' $\beta$, which alters the interdependency between the individual bits in the posterior distribution. When $\beta = 1$ no change is made to this dependency, and the spin glass system corresponds to the MAP bit decoder; however, in the limit of large $\beta$, the spin glass system corresponds to the MAP sequence decoder. The interpretation of $\beta$ as an 'inverse temperature' parameter leads to the terminology 'high temperature' decoding for any decoder with a modest value of $\beta$. 'Raising the temperature of the decoder' above that of the MAP sequence decoder (absolute zero) reduces long range dependences in the associated posterior sequence distribution, and so helps to reduce decoder complexity (at the cost of accuracy).

The results of the present paper concern the sequence posterior only, and so are somewhat different from those in [9], [11]–[13], [8]. They do not involve the hyperparameter $\beta$, but show rather that the code rate $R$ plays the role of absolute temperature in the statistical mechanical interpretation of MAP sequence decoding.

### 1.4. Notation

We shall make use of the following notation in the remainder of the paper.

- $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.
- $\mathbb{N}$ is the set of natural numbers.
- For any set $\Lambda$, $|\Lambda|$ is its cardinality.

- $\mathbb{X}$ is the linear space of infinite sequences of bits $(x_k \in \{0,1\};\ k = 1, 2, \ldots)$ over the Galois field $(\{0,1\}, \oplus, \cdot\ )$ and, for each $K \in \mathbb{N}$, $\mathbb{X}_K$ is the linear space of $K$-bit sequences $(x_k \in \{0,1\};\ k = 1, 2, \ldots, K)$.
- $\mathbf{0}$ represents the zero element of $\mathbb{X}$ (or $\mathbb{X}_K$), and $\mathbf{1}$ represents the 'all ones' sequence.
- $\|\cdot\|$ is the 'Hamming norm' on $\mathbb{X}_K$:

$$\|w\| = \|w\|_K := \sum_{k=1}^{K} w_k.$$

- For any $K \in \mathbb{N}$, $T_K : \mathbb{X} \to \mathbb{X}_K$ and $\bar{T}_K : \mathbb{X} \to \mathbb{X}$ are the maps:

$$T_K x := (x_1, x_2, \ldots, x_K) \qquad \text{and} \qquad \bar{T}_K x := (x_{K+1}, x_{K+2}, \ldots)\,.$$

  In the context of a particular value of $K$, the sequences $T_K x$ and $\bar{T}_K x$ will be referred to as the *internal* and *external* sequences, respectively.
- For any $K \in \mathbb{N}$, $\mathcal{X}_K$ is the $\sigma$-algebra of subsets of $\mathbb{X}$ generated by the map $T_K$.
- $\mathcal{X}$ is the product $\sigma$-algebra of subsets of $\mathbb{X}$, i.e. the smallest $\sigma$-algebra containing $\mathcal{X}_K$ for all $K \in \mathbb{N}$.
- $\mathcal{T}$ is the *tail* $\sigma$-algebra on $\mathbb{X}$:

$$\mathcal{T} := \bigcap_{K\in\mathbb{N}} \bar{T}_K^{-1}(\mathcal{X}).$$

- For any $\sigma$-algebra $\mathcal{B}$, $\mathcal{P}(\mathcal{B})$ is the set of probability measures on $\mathcal{B}$.
- For each $\theta \in [0,1]$, $m_\theta$ is the Bernoulli measure on $\{0,1\}$:

$$m_\theta(\{0\}) = 1 - \theta \qquad \text{and} \qquad m_\theta(\{1\}) = \theta,$$

  and $M_\theta$ is the product measure $m_\theta^{\otimes\mathbb{N}}$ on $\mathcal{X}$. Since it occurs so frequently, $M_{1/2}$ will be abbreviated to $M$.
- For each $K \in \mathbb{N}$, $\mu_K : \mathbb{X} \to \mathcal{P}(\mathcal{X})$ is defined by:

$$\mu_K(x)(B) := 2^{-K}|C_{K,x} \cap B|.$$

  $\mu_K(x)$ is the uniform probability measure on $C_{K,x}$, where $C_{K,x}$ is the set of sequences that match $x$ externally:

$$C_{K,x} := \{\tilde{x} \in \mathbb{X} : \ \bar{T}_K\tilde{x} = \bar{T}_K x\}.$$

- For $a \in \mathbb{R}$, $a^+ := \max\{0, a\}$. For $a \in \mathbb{R}^+$, $[a] := \max\{n \in \mathbb{N} \cup \{0\} : n \leq a\}$.

## 2. The infinite-sequence estimation problems

Let $\mathbb{Z} := \{z : \mathbb{X} \to \mathbb{X}\}$ be the set of maps from $\mathbb{X}$ to $\mathbb{X}$ (no matter how irregular), and let $\mathcal{Z}$ be the product $\sigma$-algebra on $\mathbb{Z}$, i.e. that generated by the finite-dimensional sets

$$A_{x,B} := \{z \in \mathbb{Z} : z(x) \in B\} \qquad x \in \mathbb{X},\ B \in \mathcal{X}. \tag{5}$$

Let $Q$ be the product measure $M^{\otimes\mathbb{N}}$, and $e : \mathbb{X}\times\mathbb{Z} \to \mathbb{X}$ the *evaluation map*: $e(x,z) = z(x)$. Let $U : \Omega \to \mathbb{X}$, $\Psi : \Omega \to \mathbb{X}$ and $\Gamma : \Omega \to \mathbb{Z}$ be independent random variables having distributions $M$, $M_q$ and $Q$, respectively. In analogy with Shannon's $(K, N)$ random block coding schemes discussed in section 1, $U$ can be thought of as an infinite *source sequence* to be communicated across a binary symmetric channel having *error sequence* $\Psi$, and $\Gamma$ can be thought of as an infinite *random code*. The channel input and output sequences are then

$$X := e(U,\Gamma) \qquad \text{and} \qquad Y := X \oplus \Psi.$$

Consider the problem of estimating $U$ on the basis of its prior distribution, $M$, observation of the channel output sequence $Y$, and knowledge of $\Gamma$.

**Proposition 2.1.** (i) *The map $e$ is not $M \otimes Q$-measurable.*
(ii) *$M \otimes Q$ can be extended to a measure $\eta \in \mathcal{P}((\mathcal{X} \times \mathcal{Z}) \vee e^{-1}(\mathcal{X}))$ in such a way that, for all $A \in \mathcal{X} \times \mathcal{Z}$ and $B \in \mathcal{X}$,*

$$\eta(A \cap e^{-1}(B)) = \nu(A \times B), \tag{6}$$

*where $\nu \in \mathcal{P}(\mathcal{X} \times \mathcal{Z} \times \mathcal{X})$ is any probability measure having the marginal*

$$\nu(A \times \mathbb{X}) = M \otimes Q(A) \qquad \text{for all } A \in \mathcal{X} \times \mathcal{Z}. \tag{7}$$

**Proof.** See section A.1.

Part (i) of this proposition shows that $X$ is not assigned a distribution through its dependence on $U$ and $\Gamma$. Part (ii) goes further to state that, provided $X^{-1}(\mathcal{X}) \subset \mathcal{F}$, we can choose $\mathbb{P}$ such that the joint distribution of $(U,\Gamma,X)$ is *any* distribution having $(U,\Gamma)$-marginal $M \otimes Q$; in particular we can choose $\mathbb{P}$ so that $X$ is *independent* of $(U,\Gamma)$. For this reason it is impossible to carry out direct Bayesian inference between $Y$ and $U$. This is not a failing of the infinite coding system; it is a natural consequence of the irregularity of the code, which is itself an essential feature of any good (infinite) error-control code.

Shannon's finite random block coding schemes can be recovered from this setup as follows. Let $K, N \in \mathbb{N}$, and consider the $(K,N)$ block coding system, in which the receiver must estimate the finite source sequence $T_K U$ on the basis of the uniform prior $m_{1/2}^{\otimes K}$, the finite channel output sequence $T_N Y$, the code $\Gamma$, *and knowledge of the external source sequence* $\bar{T}_K U$. Since the transmitter and receiver both have access to $\bar{T}_K U$, they are effectively using the *finite* random code $(T_N\Gamma(\omega)(u),\ u \in C_{K,U(\omega)})$ to communicate the finite source sequence $T_K U$.

When considering a family of such problems indexed by $K$ and $N$, it is convenient to define the prior and posterior distributions on the *common* $\sigma$-algebra $\mathcal{X}$ rather than on sets of values of the internal source sequence $T_K U$ (see remark 1.2). The estimation problem associated with the $(K,N)$ block coding system then becomes that of estimating the *infinite* sequence $U$ on the basis of the prior $\mu_K(U)$, the observation $T_N Y$, and the code $\Gamma$. The receiver's knowledge of the external sequence, $\bar{T}_K U$, is then contained in the prior. The posterior distribution is $\Pi_{K,N}(\omega, U(\omega))$, where $\Pi_{K,N} : \Omega \times \mathbb{X} \to \mathcal{P}(\mathcal{X})$ is defined by Bayes' formula:

$$\begin{aligned} \Pi_{K,N}(\omega,u)(B) &:= \frac{\int_B \exp(-H_N(\omega,x))\mu_K(u)\,(\mathrm{d}x)}{\int_{\mathbb{X}} \exp(-H_N(\omega,x))\mu_K(u)\,(\mathrm{d}x)}, \\ H_N(\omega,x) &:= -\rho_N(\omega,x)\log q - (N-\rho_N(\omega,x))\log(1-q), \\ \rho_N(\omega,x) &:= \|T_N Y(\omega) \oplus T_N\Gamma(\omega)(x)\|. \end{aligned} \tag{8}$$

**Remark 2.1.** (i) Although $H_N(\omega,\cdot)$ may not be $\mathcal{X}$-measurable, it *is* $\mu_K(u)$-measurable since the support of $\mu_K(u)$ is the finite set $C_{K,u}$. (The integrals in (8) are finite sums.)

(ii) $\mu_K(u)$ is a regular $(\bar{T}_K U = \bar{T}_K u)$-conditional distribution for $U$. However, we cannot claim that $\Pi_{K,N}(\omega, u)$ is a regular $(\bar{T}_K U = \bar{T}_K u, T_N Y)$-conditional distribution since $H_N$ is not $\mathcal{F} \times \mathcal{X}$-measurable. In what follows, we shall call it the $(K, N, u)$ posterior distribution for $U$.

(iii) $\Pi_{K,N}$ differs from its counterpart in part I in that it depends on the external sequence $\bar{T}_K u$.

For any code rate $R \in (0, \infty)$, let $\mathcal{G}_R$ be the collection of *rate $R$ index pairs*:

$$\mathcal{G}_R := \left\{(K, N) \in \mathbb{N}^2 : \ N = [R^{-1}K]\right\}. \tag{9}$$

Our goal is to construct *rate $R$ posterior distributions* for the infinite sequence $U$ by treating the family of $(K, N, u)$ estimation problems

$$\mathcal{S}_R := \{(\mu_K(u), H_N), (K, N) \in \mathcal{G}_R, u \in \mathbb{X}\}, \tag{10}$$

as the *specification* of a random field in the manner described in section 1.1. However, there are several important differences between $\mathcal{S}_R$ and the specifications of statistical mechanics.

(1) Here, the positions of bits in the sequence $U$ have no bearing on the interactions between them. The topology of the 'lattice of source bits' is determined by $\Gamma$, and this has a distribution that is invariant under arbitrary permutations of source bits.

(2) 'Consistency' methods are less attractive here for a number of reasons. For example, (4) has to be weakened to an *asymptotic* property.

(3) The $(K, N)$ block coding systems are subject to random environments that change as $K$ and $N$ increase. The most significant changes come from the random code $\Gamma$. Sequences $u \neq U(\omega)$ having (by chance outcome of $\Gamma$) unusually high posterior probabilities for one instance of $K$ do not necessarily retain this property as $K$ increases; increasing $K$ by just 1 *doubles* that part of $\Gamma$ entering the environment of the $(K, N)$ system.

## 3. The auxiliary system

The random environment discussed in point 3 above precludes well-defined posterior distributions for the infinite sequence $U$ when $R$ exceeds the channel capacity. We overcome this problem by introducing a family of source-sequence transformations (indexed by $K$ and $N$) that permute the estimand space according to the values of the posterior probabilities of the individual sequences. The resulting transformed posterior distributions have the necessary regularity to enable the study of asymptotics.

For each $K, N \in \mathbb{N}$, let $F_{K,N} : \Omega \times \mathbb{X} \ \rightarrow \ \mathbb{X}$ be a measurable map with the following properties:

(F1) for each $\omega$, $F_{K,N}(\omega, \cdot)$ is a bijection;

(F2) for each $\omega$, $F_{K,N}(\omega, U(\omega)) = \mathbf{1}$;

(F3) for each $(\omega, u)$, $\bar{T}_K F_{K,N}(\omega, u) = \bar{T}_K(u \oplus U(\omega) \oplus \mathbf{1})$;

(F4) for each $(\omega, u, \tilde{u})$ with $u, \tilde{u} \neq U(\omega)$ and $\bar{T}_K u = \bar{T}_K \tilde{u}$

$$H_N(\omega, u) < H_N(\omega, \tilde{u}) \Rightarrow \|T_K F_{K,N}(\omega, u)\| \leq \|T_K F_{K,N}(\omega, \tilde{u})\|.$$

A map $F_{K,N}$ with properties (F1–F4) can be constructed by identifying points in the domain and range with each other as follows: (i) each subset of the domain of $F_{K,N}$ of type $C_{K,u}$ is identified with the subset $C_{K,u\oplus U(\omega)\oplus\mathbf{1}}$ of the range; (ii) the point $U(\omega)$ in the domain is identified with the point $\mathbf{1}$ in the range; (iii) the remaining elements, $\tilde{u} \in C_{K,u}$, of each subset of the domain are arranged in ascending order of $H_N(\omega, \tilde{u})$, and the remaining elements, $v \in C_{K,u\oplus U(\omega)\oplus\mathbf{1}}$, of each subset of the range are arranged in ascending order of $\|T_K v\|$; (iv) domain and range points with the same positions in the orderings of $C_{K,u}$ and $C_{K,u\oplus U(\omega)\oplus\mathbf{1}}$ are identified with each other.

Let $\Pi^F_{K,N} : \Omega \times \mathbb{X} \to \mathcal{P}(\mathcal{X})$ be defined by

$$\Pi^F_{K,N}(\omega, v)(B) = \Pi_{K,N}(\omega, F_{K,N}(\omega, \cdot)^{-1}(v))(F_{K,N}(\omega, \cdot)^{-1}(B)). \tag{11}$$

If $v = F_{K,N}(\omega, u)$ then $\Pi^F_{K,N}(\omega, v)$ is the $(K, N, u)$ posterior distribution $\Pi_{K,N}(\omega, u)$ 'pushed forward' through the map $F_{K,N}(\omega, \cdot)$; it has the following properties:

- $\Pi^F_{K,N}(\omega, v)(C_{K,v}) = 1$;
- $\Pi^F_{K,N}(\omega, \mathbf{1})(\{\mathbf{1}\}) = \Pi_{K,N}(\omega, U(\omega))(\{U(\omega)\})$;
- if $x, \tilde{x} \in C_{K,v} \setminus \{\mathbf{1}\}$ then
$$\|T_K x\| > \|T_K \tilde{x}\| \Rightarrow \Pi^F_{K,N}(\omega, v)(\{x\}) \le \Pi^F_{K,N}(\omega, v)(\{\tilde{x}\}).$$

$\Pi^F_{K,N}(\omega, v)$ is thus a very particular type of probability measure on $C_{K,v}$. With the exception of the sequence $\mathbf{1}$, it is biased towards sequences $x \in C_{K,v}$ for which $\|T_K x\|$ is small. In what follows, we show that $(\Pi^F_{K,N}(\omega, v), (K, N) \in \mathcal{G}_R)$ admits asymptotic analysis in a way that $(\Pi_{K,N}(\omega, u), (K, N) \in \mathcal{G}_R)$ does not. The $(K, N, u)$ posterior distributions $\Pi_{K,N}(\omega, u)$ can be obtained by 'pulling back' $\Pi^F_{K,N}(\omega, v)$ through the maps $F_{K,N}$, and so their properties for large $(K, N) \in \mathcal{G}_R$ can be found from the asymptotics of $\Pi^F_{K,N}$. In the design of the maps $F_{K,N}$, we have been careful to separate the actual source sequence $U(\omega)$ from other sequences with large posterior probabilities, by mapping it to the sequence $\mathbf{1}$. This ensures that the value of the decoding error probability is not lost in the limiting process.

In order to study the asymptotics of the distributions $\Pi^F_{K,N}$, we associate them with the $(K, N, v)$ posterior distributions of an 'auxiliary' error-control system. Let $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ be a probability space on which are defined an *auxiliary source sequence*, $V : \tilde{\Omega} \to \mathbb{X}$, and a family of *auxiliary channel error sequences*, $(\Phi_{K,N} : \tilde{\Omega} \to \mathbb{X},\ K, N \in \mathbb{N})$. These have joint distribution $M \otimes M_q^{\otimes(\mathbb{N}\times\mathbb{N})}$, i.e. they are independent, $V$ having the same distribution as $U$, and the $\Phi_{K,N}$ having the same distribution as $\Psi$. For each $(K, N)$, $V$ is encoded for transmission across a binary symmetric channel with error sequence $\Phi_{K,N}$ by means of the fixed (non $\tilde{\omega}$-random) code

$$\Gamma^\omega_{K,N} := \Gamma(\omega)(F_{K,N}(\omega, \cdot)^{-1}),$$

where $F_{K,N}$ satisfies (F1–F4), and $\omega \in \Omega$ is fixed. The corresponding (auxiliary) channel output sequence is

$$Y^\omega_{K,N} := \Gamma^\omega_{K,N}(V) \oplus \Phi_{K,N}.$$

As with the infinite system of section 2, we cannot use conventional methods to estimate the source sequence $V$ from the output sequences of the various auxiliary channels, because of measurability issues. However, we *can* obtain finite $(K, N)$ block coding systems from the auxiliary setup, just as we did in section 2. In the $(K, N)$ system, the (auxiliary) receiver must estimate the source sequence $V$ on the basis of the prior $\mu_K(V)$, the finite channel output sequence $T_N Y^\omega_{K,N}$ and the code $\Gamma^\omega_{K,N}$. The posterior distribution is $\Pi^\omega_{K,N}(\tilde{\omega}, V(\tilde{\omega}))$, where $\Pi^\omega_{K,N} : \tilde{\Omega} \times \mathbb{X} \to \mathcal{P}(\mathcal{X})$ is defined, as in (8), by Bayes' formula:

$$\begin{aligned} \Pi^\omega_{K,N}(\tilde{\omega}, v)(B) &:= \frac{\int_B \exp(-H^\omega_{K,N}(\tilde{\omega}, x))\mu_K(v)\,(\mathrm{d}x)}{\int_{\mathbb{X}} \exp(-H^\omega_{K,N}(\tilde{\omega}, x))\mu_K(v)\,(\mathrm{d}x)}, \\ H^\omega_{K,N}(\tilde{\omega}, x) &:= -\rho^\omega_{K,N}(\tilde{\omega}, x)\log q - (N - \rho^\omega_{K,N}(\tilde{\omega}, x))\log(1-q), \\ \rho^\omega_{K,N}(\tilde{\omega}, x) &:= \left\| T_N Y^\omega_{K,N}(\tilde{\omega}) \oplus T_N \Gamma^\omega_{K,N}(x) \right\|. \end{aligned} \tag{12}$$

$\Pi^\omega_{K,N}(\tilde{\omega}, v)$ is a '$(K, N, v)$ posterior distribution' for $V$ in the same way that $\Pi_{K,N}(\omega, u)$ is a $(K, N, u)$ posterior distribution for $U$. (See remark 2.1(ii).) As in section 2, we would like to construct *rate $R$ posterior distributions* for $V$ by treating the family

$$\mathcal{S}^\omega_R := \left\{ (\mu_K(v), H^\omega_{K,N}), (K, N) \in \mathcal{G}_R, v \in \mathbb{X} \right\}$$

as the specification of a random field. In general this not possible, for the reasons given at the end of section 2. However, such a construction *is* possible in the special case that the auxiliary channel output sequences take the following specific values:

$$T_N Y^\omega_{K,N}(\tilde{\omega}) = T_N Y(\omega) \qquad \text{for all } K, N \in \mathbb{N}. \tag{13}$$

In this case

$$\begin{aligned} \rho^\omega_{K,N}(\tilde{\omega}, x) &= \rho_N(\omega, F_{K,N}(\omega, \cdot)^{-1}(x)) =: \rho^F_{K,N}(\omega, x), \\ H^\omega_{K,N}(\tilde{\omega}, x) &= H_N(\omega, F_{K,N}(\omega, \cdot)^{-1}(x)) =: H^F_{K,N}(\omega, x), \\ \Pi^\omega_{K,N}(\tilde{\omega}, v) &= \Pi^F_{K,N}(\omega, v), \end{aligned} \tag{14}$$

for all $(K, N)$, where $\Pi^F_{K,N}$ is as defined in (11). The $(K, N, v)$ posterior distributions of the auxiliary system then coincide with the transformed $(K, N, u)$ posterior distributions of the original system, and these admit asymptotic analysis.

From this point on, we shall be interested in the auxiliary system only in the special case that the channel output sequences satisfy (13). In view of (14) we shall use the notation $H^F_{K,N}(\omega, v)$ and $\Pi^F_{K,N}(\omega, v)$ for the $(K, N, v)$ log-likelihood function and posterior distribution of the auxiliary system, rather than the more cumbersome $H^\omega_{K,N}(\tilde{\omega}, v)$ and $\Pi^\omega_{K,N}(\tilde{\omega}, v)$ (with which it would be necessary to remember that $\tilde{\omega}$ satisfies (13)). This choice of notation also makes the connections with the original system of section 2 more transparent.

We finish this section with a proposition that characterizes the asymptotic behaviour of the maps $(F_{K,N}, (K, N) \in \mathcal{G}_R)$. Let $\tau : \mathbb{X} \to [0, 1]$ be defined by

$$\tau(v) := \mathbf{1}_G(v) \lim_K K^{-1} \|T_K v\|, \tag{15}$$

where $G$ is the set on which the limit exists. It is a standard result that $G$ belongs to the tail $\sigma$-algebra $\mathcal{T}$, and that $\tau$ is $\mathcal{T}$-measurable.

**Proposition 3.1.** *For any $R \in (0, \infty)$ and any $v \in \{\mathbf{0}, \mathbf{1}\} \cup \tau^{-1}((0, 1))$,*

$$\mathbb{P}\left( \lim_{(K,N)\in\mathcal{G}_R} N^{-1} \rho^F_{K,N}(\cdot, v) = g(R, \tau(v)) \right) = 1, \tag{16}$$

*where* $g : (0,\infty) \times [0,1] \to [0,1]$ *is defined as follows:*

$$g(r,\lambda) := \begin{cases} b^{-1}((1 - r(1 - b(\lambda)))^{+}) & \text{if } \lambda \in [0,1/2] \\ 1 - g(r, 1-\lambda) & \text{if } \lambda \in [1/2,1) \\ q & \text{if } \lambda = 1. \end{cases} \tag{17}$$

*Here* $b$ *is the binary entropy function of (1), and* $b^{-1} : [0,1] \to [0,1/2]$ *is the inverse of its restriction to* $[0,1/2]$.

**Proof.** See section A.2.

For any $R \in (0,\infty)$, the map $g(R,\cdot) : [0,1) \to [0,1]$ is non-decreasing and has odd symmetry about the point $(1/2,1/2)$. Let $C(\lambda)$ be the capacity of a binary symmetric channel with parameter $\lambda \in [0,1/2]$, i.e. $C(\lambda) = 1 - b(\lambda)$. If $RC(\lambda) \le 1$ then $g(R,\lambda)$ is the value of the parameter of a binary symmetric channel whose capacity is $RC(\lambda)$. If $RC(\lambda) > 1$ then no such channel exists, and $g(R,\lambda) = 0$. If $\lambda \in [1/2,1)$ and $RC(1-\lambda) \le 1$ then $1 - g(R,\lambda)$ is the value of the parameter of a binary symmetric channel whose capacity is $RC(1-\lambda)$. If $RC(1-\lambda) > 1$ then $g(R,\lambda) = 1$. Because of (F2), all the $F_{K,N}$ map the source sequence $U(\omega)$ to the transformed sequence $\mathbf{1}$, and this is why $g(R,\cdot)$ has a discontinuity at $\lambda = 1$.

## 4. The variational method

Our aim in this section is to define posterior distributions for the auxiliary source sequence $V$ (with the special channel output sequence values of (13)) by a method inspired by the variational principle of statistical mechanics. We begin by reviewing the variational version of the $(K,N,v)$ estimation problems of (12), using the notation $H^F_{K,N}$ and $\Pi^F_{K,N}$ as discussed after (14). Let

$$\mathcal{P}_{K,v}(\mathcal{X}) := \{P \in \mathcal{P}(\mathcal{X}) :\ P(C_{K,v}) = 1\}. \tag{18}$$

Let $i^v_{K,N} : \Omega \to (0,\infty)$ and $A^v_{K,N} : \Omega \times \mathcal{P}_{K,v}(\mathcal{X}) \to (0,\infty]$ be defined as follows:

$$\begin{aligned} i^v_{K,N}(\omega) &:= -\log \int_{\mathbb{X}} \exp(-H^F_{K,N}(\omega,x))\mu_K(v)\,(\mathrm{d}x) \\ A^v_{K,N}(\omega,Q) &:= h(Q|\mu_K(v)) + \int_{\mathbb{X}} H^F_{K,N}(\omega,x)Q\,(\mathrm{d}x), \end{aligned} \tag{19}$$

where $h$ is the relative entropy generalizing (2), $i^v_{K,N}(\omega)$ is the *full information* in the observation for the $(K,N,v)$ estimation problem, and $H^F_{K,N}(\omega,x)$ (for $x \in C_{K,v}$) is the *residual information* (i.e. the information remaining in the observation if we already know that $T_K V = T_K x$). $A^v_{K,N}(\omega,Q)$ is the sum of the information gain of the measure $Q$ over the prior, $\mu_K(v)$, and the *average* residual information, and is thus the information *apparently* possessed by an estimator that proposes $Q$ as a posterior distribution for $V$ (see part I or [4] for a fuller discussion of these quantities).

**Proposition 4.1.** *For any* $K,N \in \mathbb{N}$, $\omega \in \Omega$ *and* $v \in \mathbb{X}$,

$$i^v_{K,N}(\omega) = \min_{Q \in \mathcal{P}_{K,v}(\mathcal{X})} A^v_{K,N}(\omega,Q), \tag{20}$$

*and* $\Pi^F_{K,N}(\omega,v)$ *is the unique minimizer.*

**Proof.** See proposition 2.1 in [4].

We define posterior distributions for the auxiliary source sequence $V$, as minimizers of a *specific* apparent information. This is defined in terms of the log-likelihood functions $H^F_{K,N}$ of (14). In the variational principle of statistical mechanics, attention is usually restricted to *shift invariant* measures in order to avoid infinite families of Gibbs measures differing from one another only on finite sets of indices. For the same reason, we shall only look for *exchangeable* posterior distributions for $V$. This restriction is justified by theorem 4.3, below.

**Definition 4.1.** A probability measure $P \in \mathcal{P}(\mathcal{X})$ is said to be exchangeable if $P(T_K^{-1}(\{w\})) = P(T_K^{-1}(\{\pi w\}))$ for all $w \in \mathbb{X}_K$, all bit permutations $\pi : \mathbb{X}_K \to \mathbb{X}_K$, and all $K \in \mathbb{N}$. The subset of $\mathcal{P}(\mathcal{X})$ of exchangeable measures will be denoted $\mathcal{P}_E(\mathcal{X})$.

Clearly, for any $\lambda \in [0,1]$, the product measure $M_\lambda$ is exchangeable and, according to the strong law of large numbers, $M_\lambda(\tau = \lambda) = 1$, where $\tau$ is as defined in (15). In fact, all $P \in \mathcal{P}_E(\mathcal{X})$ are mixtures of such product measures.

**Theorem 4.1.** *(de Finetti) For any $P \in \mathcal{P}_E(\mathcal{X})$:*

*($i$) $P(G) = 1$;*
*($i$) for any $B \in \mathcal{X}$, $P(B|\mathcal{T}) = M_\tau(B)$.*

**Proof.** See, for example, [14] or [15].

It follows from this theorem that any $P \in \mathcal{P}_E(\mathcal{X})$ is fully determined by the distribution, $t_P$, it assigns to the tail variable $\tau$:

$$P(B) = \int_{[0,1]} M_\lambda(B) t_P\,(\mathrm{d}\lambda).$$

The two components of the specific apparent information are the *specific information gain* of $P \in \mathcal{P}_E(\mathcal{X})$ over the prior, $\mathcal{H}(P|M)$, and the *specific residual information*, $\prec H^F_R, P \succ$. The first of these is defined and evaluated as a special case of lemma 4.1.

**Lemma 4.1.** *For any $P \in \mathcal{P}_E(\mathcal{X})$ and $\alpha \in (0,1)$,*

$$\mathcal{H}(P|M_\alpha) := \lim_K K^{-1} h(P_K|M_{\alpha,K}) = \int_{[0,1]} h(m_\lambda|m_\alpha) t_P\,(\mathrm{d}\lambda), \tag{21}$$

*where $P_K$ and $M_{\alpha,K}$ are the restrictions of $P$ and $M_\alpha$, respectively, to $\mathcal{X}_K$, and $h$ is the relative entropy of (2).*

**Proof.** See section A.3.

**Remark 4.1.** Lemma 4.1 shows in particular that, unlike the relative entropies in the defining sequence, the limit $\mathcal{H}(P|M_\alpha)$ is *linear* in $P$. This property reflects that of the specific entropy of shift invariant measures in statistical mechanics [6].

In the definition of the specific internal energy of statistical mechanics, the internal variables, $\sigma_\Lambda(v)$, are integrated out by the appropriate *marginal* of a putative Gibbs measure. The resulting average internal energy is then divided by the cardinality of the set $\Lambda$; sequences of such normalized average energies over increasing sets $\Lambda$ are then shown to have well-defined limits that do not depend on the external variables, [6]. Since our random field contains an inhomogeneity (at $v = \mathbf{1}$) the limit obtained by a procedure of this type *does* depend on the external variables. Because of this we must integrate out the internal variables in $H^F_{K,N}$ by their $\bar{T}_K$*-conditional distribution* under a putative posterior, $P \in \mathcal{P}_E(\mathcal{X})$. After taking limits, we can then integrate out the external variables by their marginal distribution under $P$. Since the internal and external variables are $\mathcal{T}$-conditionally independent under $P$, it is equivalent to integrate out the internal variables by their $\mathcal{T}$-conditional distribution (see proposition 2.4 in [16]). According to theorem 4.1, the latter admits the *regular* form, $m^{\otimes K}_{\tau(v)}$. (NB. We cannot integrate out internal *and* external variables before taking limits because of the measurability problems with $H^F_{K,N}$.)

For any $K, N \in \mathbb{N}$, let $\zeta_{K,N} : \Omega \times \mathbb{X} \rightarrow [0, \infty)$ be defined as follows:

$$\zeta_{K,N}(\omega, v) := \mathbf{1}_G(v) \sum_{x \in C_{K,v}} H^F_{K,N}(\omega, x) m^{\otimes K}_{\tau(v)}(\{T_K x\}). \tag{22}$$

**Lemma 4.2.** *Let $P \in \mathcal{P}_E(\mathcal{X})$ and $R \in (0, \infty)$. For $P$-a.a. $v$,*

$$\mathbb{P}\left(\lim_{(K,N)\in\mathcal{G}_R} K^{-1}\zeta_{K,N}(\cdot, v) = R^{-1}c(R, \tau(v))\right) = 1, \tag{23}$$

*where $c : (0, \infty) \times [0, 1] \rightarrow (0, \infty)$ is defined as follows*

$$c(r, \lambda) := -g(r, \lambda) \log q - (1 - g(r, \lambda)) \log(1 - q). \tag{24}$$

*and $g$ is as in (17).*

**Proof.** See section A.4.

Since $\zeta_{K,N}$ is not $\mathcal{F} \times \mathcal{X}$-measurable, Fubini's theorem does not apply, and we cannot conclude that '$P(K^{-1}\zeta_{K,N}(\omega, \cdot) \rightarrow R^{-1}c(R, \tau)) = 1$ for a.a. $\omega$'. However, $K^{-1}\zeta_{K,N}$ is bounded (by $-R^{-1} \log q$), and so it follows from lemma 4.2 and the bounded convergence theorem that

$$P\left(\lim_{(K,N)\in\mathcal{G}_R} K^{-1}\mathbb{E}\zeta_{K,N} = R^{-1}c(R, \tau)\right) = 1. \tag{25}$$

Here, the $\omega$-dependence of $H_N$ not removed by the transformation $F_{K,N}$ is integrated out before the limit is taken. We base the definition of specific residual information on (25). In doing so, we are replacing $H^F_{K,N}$ by its 'mean-field' approximation, [6]. This is justified by theorem 4.3.

For any $R \in (0, \infty)$, let

$$\mathcal{S}^F_R := \left\{(\mu_K(v), \mathbb{E}H^F_{K,N}), (K, N) \in \mathcal{G}_R, v \in \mathbb{X}\right\}, \tag{26}$$

where $\mathcal{G}_R$ is as in (9). $\mathcal{S}^F_R$ will be called the *auxiliary rate $R$ specification*. In view of (22) and (25), we define the specific residual information of exchangeable measures with respect to $\mathcal{S}^F_R$ as follows:

$$\prec H^F_R, P \succ := R^{-1} \int_{[0,1]} c(R, \lambda) t_P\,(\mathrm{d}\lambda). \tag{27}$$

**Definition 4.2.** (i) For any $R \in (0, \infty)$ the *specific apparent information* of an exchangeable measure $P \in \mathcal{P}_E(\mathcal{X})$ with respect to $\mathcal{S}^F_R$ is

$$\mathcal{A}_R(P) := \mathcal{H}(P|M) + \prec H^F_R, P \succ. \tag{28}$$

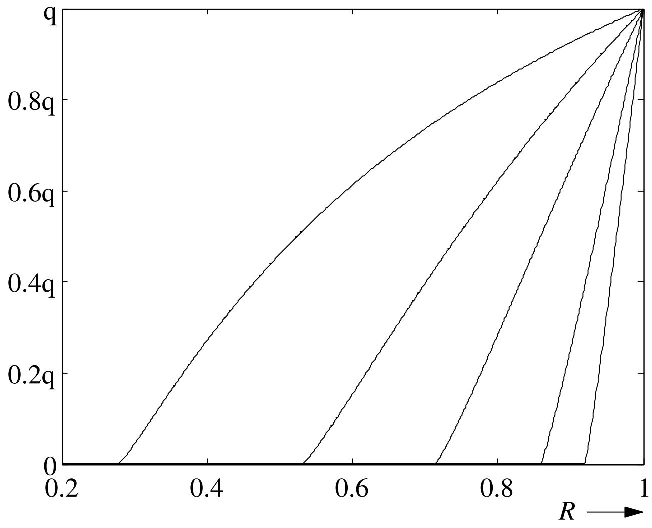

**Figure 1.** The Function $f_q$: $q = 0.2$, 0.1, 0.05, 0.02, 0.01 (left to right).

(ii) For any $R \in (0, \infty)$, an $\mathcal{S}_R^F$-*posterior distribution* for $V$ is any probability measure $P_R^* \in \mathcal{P}_E(\mathcal{X})$, for which

$$\mathcal{A}_R(P_R^*) = \min_{P \in \mathcal{P}_E(\mathcal{X})} \mathcal{A}_R(P). \tag{29}$$

**Theorem 4.2.** (*i*) *The minimum specific apparent information is as follows:*

$$\min_{P \in \mathcal{P}_E(\mathcal{X})} \mathcal{A}_R(P) = \begin{cases} 1 + R^{-1} b(q) & \text{if } R \le C \\ R^{-1} & \text{if } R \ge C. \end{cases} \tag{30}$$

(*i*) *If* $R < C$*, then the unique* $\mathcal{S}_R^F$*-posterior distribution is* $M_1$*.*

(*i*) *If* $R = C$ *and* $P_R^*$ *is an* $\mathcal{S}_R^F$*-posterior distribution, then*

$$P_R^* = \alpha M_0 + (1 - \alpha) M_1 \qquad \text{for some } \alpha \in [0, 1]. \tag{31}$$

(*i*) *If* $R > C$*, then the unique* $\mathcal{S}_R^F$*-posterior distribution is* $M_{f_q(R)}$*, where* $f_q : (0, \infty) \to [0, 1/2)$ *is defined as follows:*

$$f_q(r) := b^{-1}((1 - C/r)^+). \tag{32}$$

**Proof.** See section A.5.

Graphs of the function $f_q$ are shown in figure 1 for a few values of $q$ between 0.01 and 0.2. $f_q$ has a first derivative at all points, and a second derivative at all points except $R = C$. The second derivative grows without limit as $R \downarrow C$.

In the analogy with statistical mechanics, the residual information corresponds to the internal energy of the system, the information gain corresponds to the negative of its entropy, and the apparent information corresponds to its free energy. In this analogy, the function $E_R(\lambda) := R^{-1} c(R, \lambda)$ of (27) evaluates the energy of the 'mesostate' $\tau(v)$. In order to determine the statistical mechanical meaning of the code rate $R$, we need to understand its effect on this function.

Since the channel output bits are $V$-conditionally independent of one another, the log-likelihood function corresponding to a sequence of $N$ channel output bits is the sum of the

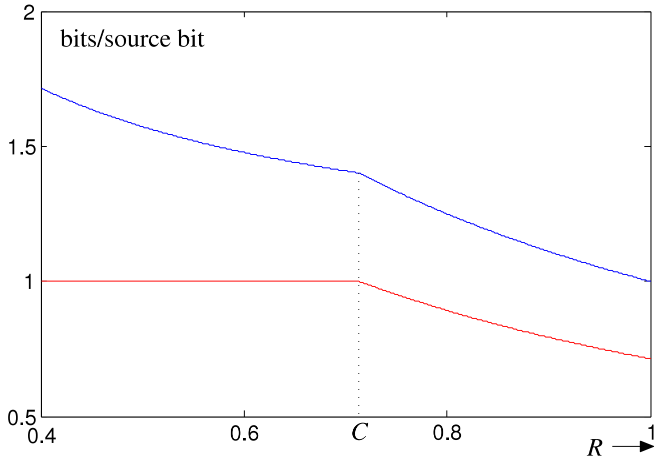

**Figure 2.** Specific information quantities: top curve $\mathcal{A}_R(P_R^*)$; bottom curve $\mathcal{H}(P_R^*|M)$ ($q = 0.05$).

log-likelihood functions corresponding to the individual bits. The latter take values in the set $\{-\log(1-q), -\log q\}$, and so the normalized log-likelihood function $N^{-1}H_{K,N}^F$ takes values in the interval $[-\log(1-q), -\log q]$. However, in lemma 4.2, $H_{K,N}^F$ is normalized by $K$, not $N$, and this accounts for the factor $R^{-1}$ in $E_R$.

Apart from the point $\lambda = 1$, at which it makes a negative jump in value, the function $\lambda \mapsto c(R, \lambda)$ is monotonically increasing. If it were not for this discontinuity, it might appear that the auxiliary system was subject to an $R$-dependent 'external magnetic field' that favoured 'down spins' ($v_k = 0$) over 'up spins' ($v_k = 1$). (Although, even on the sub-domain $[0, 1)$, $c(R, \cdot)$ is not, in general, linear.) However, this effect is a consequence of the very special set of channel output values (13) that we are considering. For a generic set of channel output values, $N^{-1}H_{K,N}^{\omega}(\tilde{\omega}, v)$ behaves in a similar way to $N^{-1}H_N(\omega, u)$ as $(K, N) \in \mathcal{G}_R$ increases: it does not depend in any consistent way on the value of $\tau(v)$, and (unless $v = V$) it does not converge. That part of the $R$-dependency of $E_R$ arising from the function $c(R, \cdot)$ is not, therefore, a generic property of an error-control coding scheme, but the result of a highly contrived set of channel output values. Viewed in a different way, it is the result of 'pushing' the energy functions of the original system of section 2 (which correspond to a generic set of channel output values, and therefore do not exhibit any $\tau(u)$-related bias) through the maps $F_{K,N}$, which introduce an $R$-dependent, $\tau(v)$-related bias.

The only *substantive* way in which $R$ enters the energy function $E_R$, from the point of view of the original system, is thus through the scaling factor $R^{-1}$. In this sense, the code rate is analogous to the *absolute temperature* of statistical mechanics. It controls the number of observation bits per estimand bit, and thus controls the relative weights, in the free energy, given to the internal energy and the entropy.

Graphs of the minimum specific apparent information, $\mathcal{A}_R(P_R^*)$, and the corresponding specific information gain, $\mathcal{H}(P_R^*|M)$, are given in figure 2, for $q = 0.05$. As a function of $R$, $\mathcal{A}_R(P_R^*)$ is continuous and has a derivative at all points except $R = C$, where its left- and right-derivatives differ. In this sense, the auxiliary system undergoes an (Ehrenfest) *first-order phase transition* at $R = C$. In common with problems of statistical mechanics, this manifests itself in the existence of more than one posterior distribution at this value of $R$, [6].

It is perhaps worth re-iterating at this point that posterior distributions for $V$, based on the family of channel output sequences $(T_N Y^{\omega}_{K,N}, (K,N) \in \mathcal{G}_R)$, cannot be constructed by conventional means, since the code sequences $T_N \Gamma^{\omega}_{K,N}(V)$ do not have well-defined distributions. In the absence of conventional tools, we are *defining* posterior distributions for $V$ by (29). Of course, such a definition is only appropriate if it leads to posterior distributions that reflect the important features of Shannon's channel coding theorem. Theorem 4.2 shows that this is the case; in particular, $P^*_R$ is a Dirac measure when $R < C$, but has larger support when $R > C$. The following theorem provides further justification by connecting the posterior distributions $P^*_R$ with their counterparts in the finite $(K, N, v)$ estimation problems.

**Theorem 4.3.** *For any $R \in (0, \infty)$ and $v \in \mathbb{X}$,*

$$\mathbb{P}\left(\lim_{(K,N)\in\mathcal{G}_R} K^{-1} i^v_{K,N} = \lim_{(K,N)\in\mathcal{G}_R} K^{-1} A^v_{K,N}(\cdot, P_{K,v}) = \mathcal{A}_R(P_v)\right) = 1, \tag{33}$$

*where $i^v_{K,N}$ and $A^v_{K,N}$ are as defined in (19), $P_v := \mathbf{1}_W(v) P^*_R + \mathbf{1}_{\mathbb{X}\setminus W}(v) M_{f_q(R)}$, $P_{K,v}$ is the unique element of $\mathcal{P}_{K,v}(\mathcal{X})$ whose restriction to $\mathcal{X}_K$ coincides with that of $P_v$, $\mathcal{P}_{K,v}(\mathcal{X})$ is as defined in (18), and*

$$W := \bigcup_{K\in\mathbb{N}} C_{K,\mathbf{1}}. \tag{34}$$

**Proof.** See section A.6.

If $v \in W$ or $R > C$ then the marginal distribution of the internal sequence under $P_{K,v}$ is equal to its marginal under $P^*_R$, and theorem 4.3 shows that this distribution is 'first-order optimal' for the finite $(K, N, v)$ estimation problems of (19) and (20). If $R < C$ then $\mathcal{A}_R(P^*_R)$ is strictly smaller than $\mathcal{A}_R(M_{f_q(R)})$ and $P^*_R(\{\mathbf{1}\}) = 1$; however, if $v \notin W$, this global optimum is not accessible to the finite $(K, N, v)$ estimation problems. The marginal distribution of the internal sequence under $P_{K,v}$ is then equal to its marginal distribution under $M_{f_q(R)}$, and theorem 4.3 shows that this is first-order optimal for the $(K, N, v)$ problems. The scaling limit of the full information quantities is then equal to the specific apparent information of $M_{f_q(R)}$.

Theorem 4.3 shows that, to first-order accuracy, the minimum apparent information for the $(K, N, v)$ estimation problems is attained by the marginals of $P^*_R$ or $M_{f_q(R)}$. So: (i) the use of a mean-field approximation for $H^F_{K,N}$ in the definition of $\prec H^F_R, P \succ$ introduces only asymptotically insignificant errors; (ii) the use of *finite-exchangeable* measures in the $(K, N, v)$ estimation problems is asymptotically optimal; (iii) the use of finite-exchangeable measures derived from a *common* element of $\mathcal{P}_E(\mathcal{X})$ is asymptotically optimal.

## 5. Asymptotic consistency

In this section, we examine the extent to which the consistency property of Gibbs measures (4) can be carried over to the Bayesian problems of section 3. For any $K \in \mathbb{N}$, $0 \le k \le K$ and $v \in \mathbb{X}$, let

$$B_{K,v}(k) := \{\tilde{v} \in C_{K,v} : \|T_K \tilde{v}\| = k\}. \tag{35}$$

Theorem 4.3 shows that, in restricting attention to finite-exchangeable measures in the $(K, N, v)$ estimation problems, we introduce only asymptotically insignificant information loss. This shows that the information distinguishing individual elements of the sets $B_{K,v}(k)$ is small in comparison with that distinguishing the sets $(B_{K,v}(k), 0 \leq k \leq K)$ themselves. We can eliminate this second-order information by *smoothing* the $(K, N, v)$ posterior distributions in the following way. For any $K, N \in \mathbb{N}$, let $\bar{\Pi}^F_{K,N} : \Omega \times \mathbb{X} \to \mathcal{P}(\mathcal{X})$ be defined by

$$\bar{\Pi}^F_{K,N}(\omega, v)(B) := \sum_{k=0}^{K} \frac{|B \cap B_{K,v}(k)|}{|B_{K,v}(k)|} \Pi^F_{K,N}(\omega, v)(B_{K,v}(k)).$$

It readily follows that

$$\frac{\mathrm{d}\bar{\Pi}^F_{K,N}(\omega, v)}{\mathrm{d}\mu_K(v)} = E_{\mu_K(v)}\left(\frac{\mathrm{d}\Pi^F_{K,N}(\omega, v)}{\mathrm{d}\mu_K(v)} \middle| \mathcal{B}_{K,v}\right),$$

where $\mathcal{B}_{K,v}$ is the $\sigma$-algebra generated by the sets $(B_{K,v}(k),\ 0 \leq k \leq K)$. Of course, we can 'pull back' $\bar{\Pi}^F_{K,N}(\omega, \cdot)$ through the map $F_{K,N}(\omega, \cdot)$ to obtain a smoothed version of the original $(K, N, u)$ posterior distribution $\Pi_{K,N}$:

$$\bar{\Pi}_{K,N}(\omega, u)(B) := \bar{\Pi}^F_{K,N}(\omega, F_{K,N}(\omega, u))(F_{K,N}(\omega, B)).$$

Since the set $B_{K,\mathbf{1}}(K)$ contains only the element $\mathbf{1}$,

$$\Pi_{K,N}(\omega, U(\omega))(\{U(\omega)\}) = \bar{\Pi}_{K,N}(\omega, U(\omega))(\{U(\omega)\});$$

i.e. the smoothing operation does not alter the posterior probability of the *actual* source sequence in the original $(K, N)$ channel coding problem, it merely averages the posterior probabilities of other sequences over sets of such sequences with similar posterior probabilities.

**Theorem 5.1.** *Let $v \in \mathbb{X}$.*

(*i*) *If $R < C$ then*

$$\mathbb{P}\left(\operatorname*{w-lim}_{(K,N)\in\mathcal{G}_R} \bar{\Pi}^F_{K,N}(\cdot, v) = \mathbf{1}_W(v)M_1 + \mathbf{1}_{\mathbb{X}\setminus W}(v)M_0\right) = 1, \tag{36}$$

*where* w-lim *indicates weak convergence with respect to the product topology on $\mathbb{X}$, and $W$ is as defined in (34).*

(*i*) *If $R > C$ then*

$$\mathbb{P}\left(\operatorname*{w-lim}_{(K,N)\in\mathcal{G}_R} \bar{\Pi}^F_{K,N}(\cdot, v) = M_{f_q(R)}\right) = 1. \tag{37}$$

*where $f_q$ is as defined in (32).*

**Proof.** See section A.7.

**Remark 5.1.** If $R < C$ and $v \in W$, the convergence in (36) can be strengthened to convergence in the total variation norm. (This follows directly from (A.27) in the proof, and expresses the reliable communication theorem.)

We cannot expect the $\mathcal{S}_R^F$-posterior distribution(s) to be consistent with the $\bar{\Pi}_{K,N}^F$, in the sense of the Dobrushin construction; however, we can establish a property of *asymptotic* consistency. For any $R \in (0, \infty)$, let

$$\mathcal{K}_R := \left\{ \mathbb{E}\bar{\Pi}_{K,N}^F(\cdot, v),\ (K, N) \in \mathcal{G}_R, v \in \mathbb{X} \right\}. \tag{38}$$

Here, as in (25), the $\omega$-dependence of $\bar{\Pi}_{K,N}^F$ not removed by the transformation $F_{K,N}$ is integrated out (see the comments following (25) on 'mean-field' approximations).

**Definition 5.1.** A probability measure $P \in \mathcal{P}_E(\mathcal{X})$ is said to be *asymptotically consistent* with $\mathcal{K}_R$ if

$$P\left( \underset{(K,N)\in\mathcal{G}_R}{\text{w-lim}} \mathbb{E}\bar{\Pi}_{K,N}^F = M_\tau \right) = 1. \tag{39}$$

**Corollary 5.1.** *(i) If $R < C$ and $P$ is asymptotically consistent with $\mathcal{K}_R$, then*

$$P = \alpha M_0 + (1 - \alpha) M_1 \qquad \text{for some } \alpha \in [0, 1]. \tag{40}$$

*(i) If $R > C$ and $P$ is asymptotically consistent with $\mathcal{K}_R$, then $P = P_R^*$.*

**Proof.** Both parts follow from theorem 5.1 and the bounded convergence theorem.

At any code rate in excess of capacity the unique asymptotically consistent measure is the posterior distribution $P_R^*$ of section 4, and this provides further justification for definition 4.2. However, at code rates less than capacity, any convex combination of $M_0$ and $M_1$ is asymptotically consistent. This is because the information distinguishing between $M_0$ and $M_1$ at such code rates is discarded when $\exp(-H_{K,N}^F)$ is normalized in Bayes' formula. This is the main reason for our preferring the variational definition of a posterior distribution to one based on asymptotic consistency.

## 6. Conclusions

Consider the crude coding scheme discussed in section 1, in which each bit of the source sequence is sent over the channel $2N + 1$ times. Over an infinite sequence of source bits, this scheme suffers the same singularity problem as the random coding scheme of section 2 (but not the measurability problem). However, since the coding scheme does not introduce observation-conditional dependences between the individual source bits, the latter can be independently decoded, and a posterior distribution for the infinite source sequence can be defined as the product of those of the individual bits. This 'Kolmogorov technique', of constructing a posterior distribution on $\mathcal{X}$ from its finite-dimensional marginals, cannot be used with the random binary code since the latter introduces very complex observation-conditional dependences between the source bits. In this paper, we have used a 'Dobrushin technique' to construct posterior distributions from their finite-dimensional *conditional* distributions.

Definition 4.2 and theorem 4.2 extend the Bayesian paradigm to a family of estimation problems not admitting conventional analysis. Moreover, the estimation problems they address are more than simply large-block limits of the conventional $(K, N, v)$ estimation problems of the specification $\mathcal{S}_R^F$. In the latter, it is known *a priori* that $\bar{T}_K V = \bar{T}_k v$, knowledge that becomes, in the limit of large $K$, knowledge of the tail variable $\tau(V)$. However, the estimators of section 4 must estimate $\tau(V)$ along with the internal sequences

$(T_K V,\ K \in \mathbb{N})$. (This is what estimators based on the 'asymptotic consistency' definition of section 5 fail to do.)

The results of section 4 suggest that, for $R \neq C$ and in the limit of large $K$, the combination of rate $R$ encoding, transmission over the binary symmetric channel with parameter $q$, and subsequent decoding, is equivalent to the *direct* transmission of the auxiliary source sequence over a binary symmetric channel with parameter $p = f_q(R)$ (graphs of which are shown in figure 1). The information gain arising from the observation manifests itself as information gain on the *individual bits* of $V$, which remain independent under $P^*_R$. This is in contrast with the original system where all the information gain manifests itself as a dependency between the bits of $U$ (see proposition 3.1 in part I). At code rates less than capacity $p = 0$ and the equivalent direct channel is error free. At code rates greater than capacity $p$ increases rapidly; it coincides with $q$ when $R = 1$, and approaches $1/2$ as $R \rightarrow \infty$.

## Acknowledgments

The authors would like to thank an anonymous referee, whose comments have led to a substantial improvement in the presentation of the material of section 3.

This work was supported by NSF Grant CCF-0836720, 'Collaborative Research: CDI-Type II: Discovery of Succinct Dynamical Relationships in Large Scale Biological Data Sets', NSF Grant ECCS-0801549, 'Control over Networks', and Siemens Corporate Research Grant, 'Advanced Control Methods for Complex Networked Systems'.

## Appendix. Proofs

### A.1. Proof of proposition 2.1

For any $K \in \mathbb{N}$, let $(w^i;\ 1 \leq i \leq 2^K)$ be the $2^K$ distinct elements of $\mathbb{X}_K$ and, for any $1 \leq i \leq 2^K$, let $D_i := (T_K e)^{-1}(\{w^i\})$. For any $E \subseteq \mathbb{X}$, let $\mathcal{X}^E$ be the $\sigma$-algebra of subsets of $\mathbb{Z}$ generated by the sets $(A_{x,B},\ x \in E, B \in \mathcal{X})$, where $A_{x,B}$ is as defined in (5). Let $C \in \mathcal{X} \times \mathcal{Z}$; then it is a standard result (see, for example, lemma 3.5.2 in [17]) that $C \in \mathcal{X} \times \mathcal{X}^E$ for some countable $E$. If $C \subset D_i$ then, according to Fubini's theorem,

$$\begin{aligned}
M \otimes Q(C) &= \int_{\mathbb{X}} \int_{\mathbb{Z}} \mathbf{1}_C(x,z) Q\,(\mathrm{d}z) M\,(\mathrm{d}x) \\
&= \int_{\mathbb{X}\setminus E} \int_{\mathbb{Z}} \mathbf{1}_C(x,z) Q\,(\mathrm{d}z) M\,(\mathrm{d}x) \\
&= \int_{\mathbb{X}\setminus E} \int_{\mathbb{Z}} \mathbf{1}_C(x,z) \mathbf{1}_{D_i}(x,z) Q\,(\mathrm{d}z) M\,(\mathrm{d}x) \\
&= \int_{\mathbb{X}\setminus E} \int_{\mathbb{Z}} \mathbf{1}_C(x,z) Q\,(\mathrm{d}z) \int_{\mathbb{Z}} \mathbf{1}_{D_i}(x,z) Q\,(\mathrm{d}z) M\,(\mathrm{d}x) \\
&= 2^{-K} M \otimes Q(C),
\end{aligned}$$

where the second step uses the fact that $M(E) = 0$, and the fourth step uses the $Q$-independence of $e(x,\cdot)$ and $e(\tilde{x},\cdot)$ for $x \neq \tilde{x}$. Thus $M \otimes Q(C) = 0$, and this shows that the outer measure $(M \otimes Q)^*((\mathbb{X} \times \mathbb{Z}) \setminus D_i)$ is 1. A similar argument with $C \in \mathcal{X} \times \mathcal{Z}$ such that $C \subset (\mathbb{X} \times \mathbb{Z}) \setminus D_i$ shows that $(M \otimes Q)^*(D_i) = 1$, and this proves part (i).

For $K \in \mathbb{N}$, let

$$\mathcal{B}_K := (\mathcal{X} \times \mathcal{Z}) \vee \{D_i, 1 \le i \le 2^K\}.$$

It follows from a recursion on example 1.2.7 in [17] that, for any set $B \in \mathcal{B}_K$, there exist sets $(C_i \in \mathcal{X} \times \mathcal{Z},\ 1 \le i \le 2^K)$ such that

$$B = \bigcup_i (C_i \cap D_i). \tag{A.1}$$

Suppose that $(\tilde{C}_i \in \mathcal{X} \times \mathcal{Z},\ 1 \le i \le 2^K)$ is another sequence with the same property; then, since the union in (A.1) is disjoint, $C_i \cap D_i = \tilde{C}_i \cap D_i$ for all $i$. Thus $C_i \triangle \tilde{C}_i \subset (\mathbb{X} \times \mathbb{Z}) \setminus D_i$, and so

$$M \otimes Q(C_i \triangle \tilde{C}_i) = 0 \qquad \text{for all } i. \tag{A.2}$$

Let $\nu$ be as in the statement of the proposition, and let $\tau_1 : \mathbb{X} \times \mathbb{Z} \times \mathbb{X} \ \to \ \mathbb{X} \times \mathbb{Z}$ and $\tau_2 : \mathbb{X} \times \mathbb{Z} \times \mathbb{X} \ \to \ \mathbb{X}$ be the following co-ordinate maps:

$$\tau_1(u, z, x) = (u, z) \qquad \text{and} \qquad \tau_2(u, z, x) = x.$$

Since $(\mathbb{X}, \mathcal{X})$ is complete and separable there exists a regular $\tau_1$-conditional distribution for $\tau_2$, $\nu_{\tau_2|\tau_1} : \mathbb{X} \times \mathbb{Z} \times \mathcal{X} \ \to \ [0, 1]$. It follows from (A.2) that the following definition for $\phi_K \in \mathcal{P}(\mathcal{B}_K)$ is unambiguous: for $B \in \mathcal{B}_K$ and $C_i$ as in (A.1)

$$\phi_K(B) := \sum_i \int_{C_i} \nu_{\tau_2|\tau_1}((u, z), e(D_i)) M \otimes Q\, (\mathrm{d}(u, z)).$$

Now $\mathbb{X} \times \mathbb{Z} = \cup_i D_i$, and so

$$\begin{aligned} \phi_K(\mathbb{X} \times \mathbb{Z}) &= \sum_i \int_{\mathbb{X}} \nu_{\tau_2|\tau_1}((u, z), e(D_i)) M \otimes Q\, (\mathrm{d}(u, z)) \\ &= \int_{\mathbb{X}} \sum_i \nu_{\tau_2|\tau_1}((u, z), e(D_i)) M \otimes Q\, (\mathrm{d}(u, z)) \\ &= 1. \end{aligned} \tag{A.3}$$

Suppose that $(B_j \in \mathcal{B}_K,\ j \in \mathbb{N})$ is a sequence of disjoint sets. It follows from (A.1) that there exist sets $(C_{j,i} \in \mathcal{X} \times \mathcal{Z},\ 1 \le i \le 2^K, j \in \mathbb{N})$ such that, for each $j$,

$$B_j = \cup_i (C_{j,i} \cap D_i),$$

and, for $\tilde{j} \ne j$ and all $i$, $C_{\tilde{j},i} \cap C_{j,i} \subset (\mathbb{X} \times \mathbb{Z}) \setminus D_i$, so that $M \otimes Q(C_{\tilde{j},i} \cap C_{j,i}) = 0$. Now

$$\begin{aligned} \phi_K(\cup_j B_j) = \phi_K(\cup_j \cup_i (C_{j,i} \cap D_i)) &= \phi_K(\cup_i (\cup_j C_{j,i}) \cap D_i) \\ &= \sum_i \int_{\cup_j C_{j,i}} \nu_{\tau_2|\tau_1}((u, z), e(D_i)) M \otimes Q\, (\mathrm{d}(u, z)) \\ &= \sum_i \sum_j \int_{C_{j,i}} \nu_{\tau_2|\tau_1}((u, z), e(D_i)) M \otimes Q\, (\mathrm{d}(u, z)) \\ &= \sum_j \phi_K(B_j). \end{aligned} \tag{A.4}$$

It follows from (A.3) and (A.4) that $\phi_K$ is a probability measure on $\mathcal{B}_K$.

Let $\mathcal{B}_\infty := \cup_K \mathcal{B}_K$; then any $B \in \mathcal{B}_\infty$ also belongs to $\mathcal{B}_K$ for some $K \in \mathbb{N}$. From this, and the fact that $\mathcal{X} \times \mathcal{Z} \subset \mathcal{B}_1 \subset \mathcal{B}_2 \subset \cdots$, it easily follows that $\mathcal{B}_\infty$ is an algebra and that $\phi_\infty : \mathcal{B}_\infty \to [0,1]$, defined by $\phi_\infty(B) = \phi_K(B)$ for $K$ sufficiently large, is a probability measure. It follows from Lebesgue's extension theorem (see, for example, Theorem 1.5.6 in [17]) that $\phi_\infty$ extends uniquely to a probability measure on $(\mathcal{X} \times \mathcal{Z}) \vee e^{-1}(\mathcal{X})$, and this is the required extension. □

### A.2. Proof of proposition 3.1

We begin with a lemma. Fix $R \in (0, \infty)$. For each $(K,N) \in \mathcal{G}_R$, $\omega \in \Omega$, $v \in \mathbb{X}$, $\lambda \in [0,1]$ and $\theta \in [-1,1]$, let

$$\begin{array}{rcl} D_K(\lambda, v) & := & \{x \in C_{K,v} \setminus \{\mathbf{1}\} : \|T_K x\| \le K\lambda\}, \\ A_K(\omega, \theta, v) & := & \{x \in C_{K,v} \setminus \{\mathbf{1}\} : \rho^F_{K,N}(\omega, x) \le N\theta\}, \\ \theta^-_K(\omega, \lambda, v) & := & \sup\{\theta \in [-1,1] : A_K(\omega,\theta,v) \subset D_K(\lambda,v)\} \in [0,1] \\ \theta^+_K(\omega, \lambda, v) & := & \inf\{\theta \in [-1,1] : A_K(\omega,\theta,v) \supseteq D_K(\lambda,v)\} \in [0,1]. \end{array} \tag{A.5}$$

**Lemma A.1.** *For any $v \in \mathbb{X}$ and any $\lambda \in [0,1)$,*

$$\mathbb{P}\left( \lim_{(K,N)\in\mathcal{G}_R} \theta^-_K(\cdot, \lambda, v) = \lim_{(K,N)\in\mathcal{G}_R} \theta^+_K(\cdot, \lambda, v) = g(R, \lambda) \right) = 1, \tag{A.6}$$

*where $g$ is as defined in (17).*

**Proof.** If there is an $x$ in $A_K(\omega,\theta,v)\backslash D_K(\lambda,v)$ then $\rho^F_{K,N}(\omega,x) \le N\theta$ and $\|T_K x\| > \|T_K \tilde{x}\|$ for all $\tilde{x} \in D_K(\lambda, v)$; it then follows from (F4) that

$$\rho^F_{K,N}(\omega, \tilde{x}) \le \rho^F_{K,N}(\omega, x) \le N\theta \qquad \text{for all } \tilde{x} \in D_K(\lambda, v),$$

so that $A_K(\omega,\theta,v) \supset D_K(\lambda,v)$. If, on the other hand, there is an $x$ in $D_K(\lambda,v)\backslash A_K(\omega,\theta,v)$ then $\|T_K x\| \le K\lambda$ and $\rho^F_{K,N}(\omega,x) > \rho^F_{K,N}(\omega,\tilde{x})$ for all $\tilde{x} \in A_K(\omega,\theta,v)$; it then follows from (F4) that

$$\|T_K\tilde{x}\| \le \|T_K x\| \le K\lambda \qquad \text{for all } \tilde{x} \in A_K(\omega,\theta,v),$$

so that $A_K(\omega,\theta,v) \subset D_K(\lambda,v)$. Thus, either $A_K(\omega,\theta,v)$ and $D_K(\lambda,v)$ are equal, or one is a strict subset of the other. So

$$\begin{array}{l} \theta^-_K(\omega,\lambda,v) = \sup\{\theta : |A_K(\omega,\theta,v)| < |D_K(\lambda,v)|\} \\ \theta^+_K(\omega,\lambda,v) = \inf\{\theta : |A_K(\omega,\theta,v)| \ge |D_K(\lambda,v)|\}. \end{array} \tag{A.7}$$

For any $x \ne \mathbf{1}$, $\rho^F_{K,N}(\cdot, x)$ has the binomial distribution with parameters $(N, 1/2)$ and so, for any $\theta \in [0,1]$,

$$\begin{aligned} \mathbb{E}|A_K(\cdot,\theta,v)| &= \sum_{x \in C_{K,v}\setminus\{\mathbf{1}\}} \sum_{n=0}^{[N\theta]} \mathbb{P}(\rho^F_{K,N}(\cdot,x) = n) \\ &= \left(2^K - \mathbf{1}_{\{\mathbf{1}\}}(\bar{T}_K v)\right) \sum_{n=0}^{[N\theta]} \binom{N}{n} 2^{-N}, \end{aligned}$$

$$
\begin{aligned}
\mathbb{E}|A_K(\cdot,\theta,v)|^2 &= \sum_{x,\tilde{x}\in C_{K,v}\setminus\mathbf{1}} \sum_{n,\tilde{n}=0}^{[N\theta]} \mathbb{P}(\rho^F_{K,N}(\cdot,x)=n, \rho^F_{K,N}(\cdot,\tilde{x})=\tilde{n}) \\
&\le (\mathbb{E}|A_K(\cdot,\theta,v)|)^2 + \mathbb{E}|A_K(\cdot,\theta,v)|,
\end{aligned}
$$

and

$$
\operatorname{var}\left(\frac{|A_K(\cdot,\theta,v)|}{\mathbb{E}|A_K(\cdot,\theta,v)|}\right) \le \frac{1}{\mathbb{E}|A_K(\cdot,\theta,v)|} \le \frac{2^N}{(2^K-1)} \binom{N}{[N\theta]}^{-1}. \tag{A.8}
$$

Suppose that $\lambda,\theta\in[0,1/2]$; then

$$
\binom{K}{[K\lambda]} \le |D_K(\lambda,v)| \le (K\lambda+1)\binom{K}{[K\lambda]},
$$

$$
(2^K-1)\binom{N}{[N\theta]}2^{-N} \le \mathbb{E}|A_K(\cdot,\theta,v)| \le (N\theta+1)\binom{N}{[N\theta]}2^{K-N},
$$

and so, according to lemma A.1 in part I,

$$
\lim N^{-1}\log\left(\frac{|D_K(\lambda,v)|}{\mathbb{E}|A_K(\cdot,\theta,v)|}\right) = Rb(\lambda)-R-b(\theta)+1. \tag{A.9}
$$

If $\theta\in(g(R,0),1/2]$ then it follows from lemma A.1 in part I that the term on the right-hand side of (A.8) decreases exponentially rapidly in $N$, with rate $b(\theta)+R-1>0$. The infinite sequence of variances in (A.8) thus has finite sum and so, according to the moment form of the first Borel–Cantelli lemma,

$$
\mathbb{P}\left(\frac{|A_K(\cdot,\theta,v)|}{\mathbb{E}|A_K(\cdot,\theta,v)|} \quad\rightarrow\quad 1\right) = 1.
$$

Together with (A.9) and the definition of $g$, this shows that, for $\mathbb{P}$-a.a. $\omega$,

$$
\lim N^{-1}\log\left(\frac{|D_K(\lambda,v)|}{|A_K(\omega,\theta,v)|}\right)\begin{cases} = b\circ g(R,\lambda)-b(\theta) & \text{if } \lambda\ge\lambda_0 \\ <0 & \text{otherwise,}\end{cases} \tag{A.10}
$$

where $\lambda_0 := b^{-1}((1-R^{-1})^+)$.

If $\theta\in[-1,g(R,0))$ then lemma 2.1(iii) in part I shows that

$$
\lim N^{-1}\log\mathbb{P}\left(\min_{x\in C_{K,v}\setminus\mathbf{1}} \rho^F_{K,N}(\cdot,x)\le N\theta\right) < 0.
$$

($g(R,0)$ is equal to $\theta_{\mathrm{GV}}(R)$ of part I.) So $\sum_K \mathbb{P}(A_K(\cdot,\theta,v)\ne\emptyset)<\infty$, and the first Borel–Cantelli lemma shows that

$$
\mathbb{P}\left(\bigcup_{L=1}^{\infty}\bigcap_{K=L}^{\infty} A_K(\cdot,\theta,v)=\emptyset\right) = 1. \tag{A.11}
$$

We now claim that

$$
\begin{aligned}
&\mathbb{P}\left(\liminf\theta^-_K(\cdot,\lambda,v)=g(R,\lambda)\right)=1 \qquad \text{for all } \lambda\in[0,1/2] \\
&\mathbb{P}\left(\limsup\theta^+_K(\cdot,\lambda,v)=g(R,\lambda)\right)=1 \qquad \text{for all } \lambda\in[0,1/2).
\end{aligned} \tag{A.12}
$$

If $\lambda > \lambda_0$ then $g(R,\lambda) > g(R,0)$, and (A.12) follows from (A.7) and the equation in (A.10). If $\lambda \leq \lambda_0$ then $g(R,\lambda) = g(R,0)$, and (A.12) follows from (A.7), the inequality in (A.10), and (A.11).

Suppose, now, that $\lambda \in [1/2, 1)$. Arguments similar to those used above, but with $C_{K,v} \setminus D_K(\lambda, v)$ substituted for $D_K(\lambda, v)$ and $C_{K,v} \setminus A_K(\omega,\theta,v)$ substituted for $A_K(\omega,\theta,v)$, show that, for any $\theta \in [1/2, 1 - g(R,0))$ and $\mathbb{P}$-a.a. $\omega$,

$$\lim N^{-1} \log\left(\frac{|C_{K,v} \setminus D_K(\lambda,v)|}{|C_{K,v} \setminus A_K(\omega,\theta,v)|}\right) \begin{cases} = b \circ g(R,\lambda) - b(\theta) & \text{if } \lambda \leq 1 - \lambda_0 \\ <0 & \text{otherwise,} \end{cases} \tag{A.13}$$

which shows that the equations in (A.12) are also true for $\lambda \in (1/2, 1)$ and $\lambda \in [1/2, 1)$, respectively. This completes the proof. □

**Proof of Proposition 3.1 .** Since $\theta_K^-(\omega, 0, \mathbf{0}) \leq N^{-1}\rho_{K,N}^F(\omega, \mathbf{0}) \leq \theta_K^+(\omega, 0, \mathbf{0})$, (16) in the special case that $v = \mathbf{0}$ follows directly from lemma A.1. Since $\rho_{K,N}^F(\cdot, \mathbf{1})$ $(=\|T_N\Psi\|)$ is the sum on $N$ independent Bernoulli random variables with common mean $q$, (16) in the special case that $v = \mathbf{1}$ follows from the strong law of large numbers. Suppose, then, that $v \in \tau^{-1}((0,1))$, and let $\lambda := \tau(v)$. For any $n > \max\{\lambda^{-1}, (1-\lambda)^{-1}\}$, there exists an $L_{v,n} < \infty$ such that $v \in D_K(\lambda + n^{-1}, v) \setminus D_K(\lambda - n^{-1}, v)$ for all $K \geq L_{v,n}$, and so, for all such $K$ and all $\omega$,

$$\theta_K^-(\omega, \lambda - n^{-1}, v) \leq N^{-1}\rho_{K,N}^F(\omega, v) \leq \theta_K^+(\omega, \lambda + n^{-1}, v).$$

It thus follows from lemma A.1 that $\mathbb{P}(B_{\lambda,n,v}) = 1$, where

$$\begin{aligned} B_{\lambda,n,v} := \{\omega :\ & g(R, \lambda - n^{-1}) \leq \liminf N^{-1}\rho_{K,N}^F(\omega, v) \leq \limsup N^{-1}\rho_{K,N}^F(\omega, v) \\ & \leq\ g(R, \lambda + n^{-1})\}. \end{aligned}$$

Since $g(R, \cdot)$ is continuous on $(0,1)$

$$B_{\lambda,n,v} \downarrow B_{\lambda,v} := \{\omega :\ N^{-1}\rho_{K,N}^F(\omega, v)\ \rightarrow\ g(R,\lambda)\},$$

and since $\mathbb{P}$ is $\sigma$-additive $\mathbb{P}(B_{\lambda,v}) = 1$, which completes the proof of (16) for the cases $\lambda \in \tau^{-1}((0,1))$. □

### A.3. Proof of lemma 4.1

For any $K \in \mathbb{N}$ and $\lambda \in [0,1]$, let $Q_{K,\lambda}$ be the binomial distribution with parameters $(K, \lambda)$, and let $\hat{Q}_{K,\lambda}$ be the distribution of the $[0,1]$-valued random variable $K^{-1}Z$, where $Z$ has distribution $Q_{K,\lambda}$. Let $Q_K := \int_{[0,1]} Q_{K,\lambda} t_P\,(\mathrm{d}\lambda)$; then

$$\frac{\mathrm{d}P_K}{\mathrm{d}M_{\alpha,K}}(x) = \frac{\mathrm{d}Q_K}{\mathrm{d}Q_{K,\alpha}}(\|T_K x\|),$$

and so

$$\begin{aligned} h(P_K|M_{\alpha,K}) = h(Q_K|Q_{K,\alpha}) &= h(Q_K|\nu_K) - \log(K+1) \\ &- \sum_{k=0}^{K}\left(\log\binom{K}{k} - K b_\alpha(k/K)\right) Q_K(\{k\}), \end{aligned}$$

where $\nu_K$ is the uniform probability measure on $\{0, 1, \ldots, K\}$, and $b_\alpha : [0, 1] \to (0, \infty)$ is defined by

$$b_\alpha(s) := -s \log \alpha - (1 - s) \log(1 - \alpha). \tag{A.14}$$

So

$$\begin{aligned} \liminf K^{-1} h(P_K | M_{\alpha,K}) &\geq -\lim \sum_{k=0}^{K} \left( K^{-1} \log \binom{K}{k} - b_\alpha(k/K) \right) Q_K(\{k\}) \\ &= -\lim \sum_{k=0}^{K} (b(k/K) - b_\alpha(k/K)) Q_K(\{k\}) \\ &= \lim \int_{[0,1]} \int_{[0,1]} h(m_s | m_\alpha) \hat{Q}_{K,\lambda}\,(\mathrm{d}s) t_P\,(\mathrm{d}\lambda) \\ &= \int_{[0,1]} h(m_\lambda | m_\alpha) t_P\,(\mathrm{d}\lambda), \end{aligned} \tag{A.15}$$

where we have used the non-negativity of $h(Q_K | \nu_K)$ in the first step, and lemma A.1 of part I in the second step. The final step follows from the continuity and boundedness of the map $[0, 1] \ni s \mapsto h(m_s | m_\alpha) \in \mathbb{R}^+$, the fact that $\hat{Q}_{K,\lambda}$ converges weakly to the Dirac measure at $\lambda$, and the bounded convergence theorem.

Since $h(\cdot | M_{\alpha,K})$ is convex, Jensen's inequality shows that

$$\begin{aligned} K^{-1} h(P_K | M_{\alpha,K}) &\leq K^{-1} \int_{[0,1]} h(M_{\lambda,K} | M_{\alpha,K}) t_P\,(\mathrm{d}\lambda) \\ &= \int_{[0,1]} h(m_\lambda | m_\alpha) t_P\,(\mathrm{d}\lambda). \end{aligned}$$

Together with (A.15), this completes the proof. □

### A.4. Proof of lemma 4.2

Since $P(\{\mathbf{0}, \mathbf{1}\} \cup \tau^{-1}((0, 1))) = 1$, it suffices to consider only $v$ in this set. Let $b_q$ be as in (A.14). Now $N^{-1} \zeta_{K,N}(\omega, v) = b_q(N^{-1} \rho^F_{K,N}(\omega, v))$ if $v = \mathbf{0}$ or $\mathbf{1}$, and so (23) follows directly from proposition 3.1 in these cases. Suppose, then, that $v \in \tau^{-1}((0, 1))$. Now

$$\begin{aligned} \zeta_{K,N}(\omega, v) &= \sum_{k=0}^{K} \sum_{x \in B_{K,v}(k)} H^F_{K,N}(\omega, x) m^{\otimes K}_{\tau(v)}(\{T_K x\}) \\ &= \sum_{k=0}^{K} \hat{H}^F_{K,N}(\omega, k/K, v) Q_{K,\tau(v)}(\{k\}) \\ &= \int_{[0,1]} \hat{H}^F_{K,N}(\omega, s, v) \hat{Q}_{K,\tau(v)}\,(\mathrm{d}s), \end{aligned}$$

where $B_{K,v}$ is as defined in (35), $Q_{K,\tau(v)}$ and $\hat{Q}_{K,\tau(v)}$ are as defined in the proof of lemma 4.1, and

$$\hat{H}^F_{K,N}(\omega, s, v) := |B_{K,v}([Ks])|^{-1} \sum_{x \in B_{K,v}([Ks])} H^F_{K,N}(\omega, x).$$

Now $0 \le N^{-1}\hat{H}^F_{K,N}(\omega,s,v) \le -\log q$ and $\hat{Q}_{K,\tau(v)}$ converges weakly to the Dirac measure at $\tau(v)$, and so, for any $\epsilon < \min\{\tau(v), 1-\tau(v)\}/2$ and any $\omega$,

$$\lim N^{-1}\left(\zeta_{K,N}(\omega,v) - \frac{1}{\hat{Q}_{K,\tau(v)}(I_{v,\epsilon})}\int_{I_{v,\epsilon}} \hat{H}^F_{K,N}(\omega,s,v)\hat{Q}_{K,\tau(v)}\,(\mathrm{d}s)\right) = 0, \tag{A.16}$$

where $I_{v,\epsilon} := (\tau(v)-\epsilon, \tau(v)+\epsilon)$.

For any $s \in I_{v,\epsilon}$ and any $K > \epsilon^{-1}$, $B_{K,v}([Ks]) = D_K(s,v) \setminus D_K(s-K^{-1},v)$, where $D_K$ is as defined in (A.5), and so

$$b_q \circ \theta^-_K(\omega,\tau(v)-2\epsilon,v) \le N^{-1}\hat{H}^F_{K,N}(\omega,s,v) \le b_q \circ \theta^+_K(\omega,\tau(v)+\epsilon,v),$$

where $\theta^\pm_K$ are as in (A.5). Together with lemma A.1, the continuity of $c(R,\cdot)$ on $(0,1)$ and (A.16), this proves (23) for the cases $v \in \tau^{-1}((0,1))$. □

### A.5. Proof of theorem 4.2

It follows from lemmas 4.1 and 4.2 that $\mathcal{A}_R(P) = \int_{[0,1]} \mathcal{A}_R(M_\lambda)t_P\,(\mathrm{d}\lambda)$, where

$$\begin{aligned}\mathcal{A}_R(M_\lambda) &= 1 - b(\lambda) + R^{-1}c(R,\lambda)\\ &= 1 - b(\lambda) + R^{-1}b\circ g(R,\lambda) + R^{-1}h(m_{g(R,\lambda)}|m_q).\end{aligned} \tag{A.17}$$

Let $\lambda_0 := (1-R^{-1})^+$. If $\lambda \in [\lambda_0, 1-\lambda_0)$ then

$$\mathcal{A}_R(M_\lambda) = R^{-1}(1 + h(m_{g(R,\lambda)}|m_q));$$

whereas, if $R > 1$ and $\lambda \in [0,\lambda_0)$ then

$$\mathcal{A}_R(M_\lambda) = 1 - b(\lambda) - R^{-1}\log(1-q);$$

if $R > 1$ and $\lambda \in (1-\lambda_0, 1)$ then

$$\mathcal{A}_R(M_\lambda) = 1 - b(\lambda) - R^{-1}\log q;$$

finally, if $\lambda = 1$ then

$$\mathcal{A}_R(M_1) = 1 + R^{-1}b(q). \tag{A.18}$$

A comparison of these expressions reveals that the minimum value of $\mathcal{A}_R(M_\lambda)$ is given by the right-hand side of (30), and is achieved only by the values $\lambda = 1$ for $R \le C$, and $\lambda = f_q(R)$ for $R \ge C$. ($g(R, f_q(R)) = q$ if $R \ge C$.) This completes the proof of all parts of the theorem. □

### A.6. Proof of theorem 4.3

For any $(K,N) \in \mathcal{G}_R$, $\lambda \in [0,1]$ and $\omega \in \Omega$, let

$$S_K(\omega,\lambda,v) := \int_{B_{K,v}([K\lambda])} \exp(-H^F_{K,N}(\omega,x))\mu_K(v)\,(\mathrm{d}x), \tag{A.19}$$

where $B_{K,v}$ is as defined in (35). If $\lambda = 0$ then $B_{K,v}([K\lambda]) = D_K(0,v)$, where $D_K$ is as defined in (A.5); whereas if $\lambda \in (0,1)$ and $K > \lambda^{-1}$ then $B_{K,v}([K\lambda]) = D_K(\lambda,v) \setminus D_K(\lambda - K^{-1},v)$. In either case, for any $\epsilon > 0$ and $K > \epsilon^{-1}$,

$$\alpha^-_K(\omega,(\lambda-\epsilon)^+,v) \le -K^{-1}\log S_K(\omega,\lambda,v) \le \alpha^+_K(\omega,\lambda,v),$$

where

$$\alpha_K^{\pm}(\omega,\lambda,v) := 1 - K^{-1}\log\binom{K}{[K\lambda]} + NK^{-1}b_q \circ \theta_K^{\pm}(\omega,\lambda,v),$$

and $\theta_K^{\pm}$ are as defined in (A.5). It thus follows from lemma A.1 here, lemma A.1 of part I, the continuity of $c(R,\cdot)$ on $[0,1)$, and (A.17) that

$$\mathbb{P}(-K^{-1}\log S_K(\cdot,\lambda,v) \;\to\; 1 - b(\lambda) + R^{-1}c(R,\lambda) = \mathcal{A}_R(M_\lambda)) = 1. \quad \text{(A.20)}$$

Suppose now that $\lambda = 1$. If $v \notin W$ and $K > 1$, then $B_{K,v}([K\lambda]) = D_K(1,v) \setminus D_K(1 - K^{-1},v)$, and arguments similar to those used to prove (A.20) show that

$$\mathbb{P}\left(\liminf -K^{-1}\log S_K(\cdot,1,v) \geq \lim_{\lambda\uparrow 1}\mathcal{A}_R(M_\lambda)\right) = 1. \quad \text{(A.21)}$$

If $v \in W$ and $K$ is sufficiently large, then $B_{K,v}(K)$ comprises the single element $\mathbf{1}$ and $S_K(\omega,1,v) = \exp(-H^F_{K,N}(\omega,\mathbf{1}) - K)$, and so according to proposition 3.1 with $v = \mathbf{1}$, and (A.18)

$$\mathbb{P}\left(-K^{-1}\log S_K(\cdot,1,v) \;\to\; 1 + R^{-1}c(R,1) = \mathcal{A}_R(M_1)\right) = 1. \quad \text{(A.22)}$$

For any $\epsilon \in (0,1/2)$, let

$$J_{R,v,\epsilon,K} := \begin{cases} \{0 \leq k \leq K :\, k/K \in I_{R,\epsilon}\} & \text{if } R > C \quad \text{or} \quad v \notin W \\ \{0 \leq k \leq K-1\} & \text{if} R < C \quad \text{and} \quad v \in W \\ \{0 \leq k \leq K-1 :\, k/K \in I_{R,\epsilon}\} & \text{if } R = C \quad \text{and} \quad v \in W, \end{cases} \quad \text{(A.23)}$$

where $I_{R,\epsilon} := \{s \in [0,1] :\, |s - f_q(R)| > \epsilon\}$, and let

$$\Lambda_{R,v} := \begin{cases} \{f_q(R)\} & \text{if } R > C \quad \text{or} \quad v \notin W \\ \{1\} & \text{if } R < C \quad \text{and} \quad v \in W \\ \{0,1\} & \text{if } R = C \quad \text{and} \quad v \in W. \end{cases} \quad \text{(A.24)}$$

It follows from the arguments following (A.17) that

$$\inf_K \min_{k\in J_{R,v,\epsilon,K}} \mathcal{A}_R(M_{k/K}) > \max_{\lambda\in\Lambda_{R,v}} \mathcal{A}_R(M_\lambda),$$

which, together with (A.20)–(A.22) shows that, for any $v$ and any $\lambda \in \Lambda_{R,v}$,

$$\mathbb{P}\left(\liminf K^{-1}\log\left(\frac{S_K(\cdot,\lambda,v)}{\sum_{k\in J_{R,v,\epsilon,K}} S_K(\cdot,k/K,v)}\right) > 0\right) = 1. \quad \text{(A.25)}$$

Now

$$\exp(-i^v_{K,N}(\omega)) = \sum_{k=0}^{K} S_K(\omega,k/K,v),$$

and so it follows from (A.20)–(A.25), and the continuity of the map $[0,1) \ni \lambda \mapsto \mathcal{A}_R(M_\lambda)$, that

$$\mathbb{P}\left(\lim K^{-1} i^v_{K,N} = \mathcal{A}_R(P_v)\right) = 1.$$

It thus remains to prove that

$$\mathbb{P}\left(\lim K^{-1}\xi^{v}_{K,N}(\cdot, P_{K,v}) = \prec H^{H}_{R}, P_v \succ\right) = 1, \tag{A.26}$$

where

$$\xi^{v}_{K,N}(\omega, P_{K,v}) := \int_{\mathbb{X}} H^{F}_{K,N}(\omega, x) P_{K,v}\,(\mathrm{d}x).$$

If $R \le C$ then

$$\xi^{v}_{K,N}(\omega, P_{K,v}) = H^{F}_{K,N}(\omega, v_{K,\mathbf{0}}) t_{P_v}(\{0\}) + H^{F}_{K,N}(\omega, v_{K,\mathbf{1}}) t_{P_v}(\{1\}),$$

where $v_{K,\mathbf{0}}, v_{K,\mathbf{1}} \in C_{K,v}$, $T_K v_{K,\mathbf{0}} = \mathbf{0}$, and $T_K v_{K,\mathbf{1}} = \mathbf{1}$. Now

$$b_q \circ \theta^{-}_{K}(\omega, 0, v) \le N^{-1} H^{F}_{K,N}(\omega, v_{K,\mathbf{0}}) \le b_q \circ \theta^{+}_{K}(\omega, 0, v),$$

and, if $v \in W$, $H^{F}_{K,N}(\omega, v_{K,\mathbf{1}}) = H^{F}_{K,N}(\omega, \mathbf{1})$ for all sufficiently large $K$, and so (A.26) for the case $R \le C$ follows from lemma A.1 and proposition 3.1.

If $R > C$ then

$$\xi^{v}_{K,N}(\omega, P_{K,v}) = \int_{[0,1]} \hat{H}^{F}_{K,N}(\omega, s, v) \hat{Q}_{K,\lambda}\,(\mathrm{d}s),$$

where $\hat{Q}_{K,\lambda}$ and $\hat{H}^{F}_{K,N}$ are as defined in the proofs of lemmas 4.1 and 4.2, and $\lambda = f_q(R) \in (0,1)$. Equation (A.26) now follows from arguments similar to those used, for the case $v \in \tau^{-1}((0,1))$, in the proof of lemma 4.2. □

### A.7. Proof of theorem 5.1

It follows from the definition of $\Pi^{F}_{K,N}$ that, for any $(K,N) \in \mathcal{G}_R$, $\omega \in \Omega$ and $\lambda, s \in [0,1]$,

$$\frac{\Pi^{F}_{K,N}(\omega, v)(B_{K,v}([K\lambda]))}{\Pi^{F}_{K,N}(\omega, v)(B_{K,v}([Ks]))} = \frac{S_K(\omega, \lambda, v)}{S_K(\omega, s, v)},$$

where $B_{K,v}$ and $S_K$ are as defined in (35) and (A.19). So (A.25) shows that, for any $\epsilon \in (0, 1/2)$,

$$\mathbb{P}\left(\Pi^{F}_{K,N}(\cdot, v)\left(\cup_{k \in J_{R,v,\epsilon,K}} B_{K,v}(k)\right) \;\to\; 0\right) = 1, \tag{A.27}$$

where $J_{R,v,\epsilon,K}$ is as defined in (A.23).

Let $L \in \mathbb{N}$ and $E_L := T_L^{-1}(\{\mathbf{1}\})$, then since $\bar{\Pi}^{F}_{K,N}(\omega, v) \circ T_K^{-1}$ is finite-exchangeable, for any $K \ge L$,

$$\begin{aligned}\bar{\Pi}^{F}_{K,N}(\omega, v)(E_L) &= \sum_{k=0}^{K} \bar{\Pi}^{F}_{K,N}(\omega, v)\left(E_L | B_{K,v}(k)\right) \bar{\Pi}^{F}_{K,N}(\omega, v)(B_{K,v}(k)) \\ &= \sum_{k=L}^{K} \frac{k}{K} \frac{k-1}{K-1} \cdots \frac{k-L+1}{K-L+1} \Pi^{F}_{K,N}(\omega, v)(B_{K,v}(k)).\end{aligned}$$

Now, for any $1 \le i \le k - L + 1$,

$$\left| \frac{k-i}{K-i} - \frac{k}{K} \right| = \frac{(K-k)i}{K(K-i)} \le \frac{L}{K-L},$$

and so

$$\lim\left(\bar{\Pi}^F_{K,N}(\omega,v)(E_L)-\sum_{k=0}^{K}(k/K)^L\Pi^F_{K,N}(\omega,v)(B_{K,v}(k))\right)=0.$$

Together with (A.27), this shows that, for $\mathbb{P}$-a.a. $\omega$,

$$\lim\bar{\Pi}^F_{K,N}(\omega,v)(E_L)=\begin{cases}f_q(R)^L & \text{if } R>C \quad \text{or} \quad v\notin W\\ 1 & \text{if } R<C \quad \text{and} \quad v\in W.\end{cases} \tag{A.28}$$

Since the collection of finite-dimensional sets in $\mathcal{X}$ is a countable $\pi$-system that generates the product topology, it is convergence determining (see, for example, theorem 2.2 in [18]). This establishes (36) and (37), since all such sets can be expressed as finite unions, intersections and complements of the sets $(E_L, L\in\mathbb{N})$. □

## References

[1] Shannon C E, *A mathematical theory of communication*, 1948 *Bell Syst. Tech. J.* **27** 379
Shannon C E, 1948 *Bell Syst. Tech. J.* **27** 623

[2] Cover T M and Thomas J A, 2006 *Elements of Information Theory* (New York: Wiley)

[3] Richardson T and Urbanke R, 2008 *Modern Coding Theory* (Cambridge: Cambridge University Press)

[4] Mitter S K and Newton N J, *A variational approach to nonlinear estimation*, 2003 *SIAM J. Control Optim.* **42** 1813

[5] Newton N J and Mitter S K, *Variational Bayes in a problem of reliable communication I: finite systems*, 2010 *Commun. Inform. Syst.* **10** 155

[6] Georgii H-O, 1988 *Gibbs Measures and Phase Transitions, Studies in Mathematics 9* (Berlin: de Gruyter)

[7] Verdú S, *Fifty years of Shannon theory*, 1998 *IEEE Trans. Inform. Theor.* **44** 2057

[8] Sourlas N, *Spin glass models as error correcting codes*, 1989 *Nature* **339** 693

[9] Fedrigo M, *A large deviations approach to Shannon random coding*, 2005 *Project Report for Diploma di Perfezionamento in Matematica per le Tecnologie Industriale* (Pisa: Scuola Normale)

[10] Franz S, Leone M, Montanari A and Ricci-Tersenghi F, *Dynamic phase transition for decoding algorithms*, 2002 *Phys. Rev.* E **66** 046120

[11] Kabashima Y and Saad D, *Statistical mechanics of low density parity check codes*, 2004 *J. Phys. A: Math. Gen.* **37** R1

[12] Merhav N, *Relations between random coding exponents and statistical physics of random codes*, 2009 *IEEE Trans. Inform. Theor.* **55** 83

[13] Montanari A and Sourlas N, *The statistical mechanics of turbo codes*, 2000 *Eur. Phys. J.* B **18** 107

[14] Feller W, 1968 *An Introduction to Probability Theory and its Applications* (New York: Wiley)

[15] Kingman J F C, *The 1977 Wald memorial lectures: uses of exchangeability*, 1978 *Ann. Probab.* **6** 183

[16] van Putten C and van Schuppen J H, *Invariance properties of the conditional independence relation*, 1985 *Ann. Probab.* **13** 934

[17] Bogachev V I, 2007 *Measure Theory* vol 1 (Berlin: Springer)

[18] Billingsley P, 1999 *Convergence of Probability Measures* (New York: Wiley)