COMMUNICATIONS IN INFORMATION AND SYSTEMS
Vol. 10, No. 3, pp. 155-182, 2010

002

VARIATIONAL BAYES AND A PROBLEM OF RELIABLE COMMUNICATION I: FINITE SYSTEMS*

NIGEL J. NEWTON† AND SANJOY K. MITTER‡

Abstract. This paper is the first in a two-part study of a variational Bayesian method and its application to a problem of reliable communication. The variational method expresses a Bayesian posterior distribution as the unique minimizer of a quantity dubbed *apparent information*. This has the same nature as the *free energy* of statistical mechanics. The minimum apparent information coincides with the *full information* of the observation.

Reliable communication over an error prone channel can be achieved by the use of random block coding, as originally proposed by Shannon. The *primary* Bayesian problem in this context, is that of estimating the transmitted block from observations of the output of the channel. Scaling limits for the various information quantities are derived for this problem; these show that the primary problem undergoes a second-order phase transition, in a very precise sense, at the channel capacity; the code rate is shown to play the role of absolute temperature.

Shannon's *reliability function* is recovered as the scaling limit of the full information of a *secondary* Bayesian problem, in which the channel noise and random code are estimated from the observation of a block decoding error. This secondary problem undergoes a third-order phase transition at a second critical code rate.

1. Introduction. This paper is the first part in a two-part study investigating a connection between Bayesian estimation and statistical mechanics in the context of a problem in telecommunications. This connection can be seen in the formulae for the posterior distribution of a Bayesian problem and the Gibbs measure for a finite statistical mechanical system. These have identical structure if the range space of the estimand, its prior distribution and the (negative) log-likelihood function of the Bayesian problem are identified with the phase space, the reference measure and the (temperature-scaled) energy function of the statistical mechanical system. Gibbs distributions can be defined as minimizers of a so called *free energy*. In the Bayesian setting, the posterior distribution is the unique minimizer of a quantity we call the *apparent information*.

We consider the classical problem of reliably communicating a sequence of bits over an error prone channel. This not only provides example applications of the

* 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".

† School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK, and Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: njn@essex.ac.uk

‡ Department of Electrical Engineering and Computer Science, and Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: mitter@mit.edu

variational method, but also avoids too much abstraction. Part I considers the communication of blocks of data of *finite* size K, and shows that Shannon's classical results on reliable communication [20] are connected with secondary Bayesian problems, in which the sources of errors are estimated on the basis of observation of the error event. This gives insight into the dominant cause of errors in different regimes. Part I also finds (large K) scaling laws for the variational quantities in these secondary problems, as well as for those of the *primary* Bayesian problem, in which the transmitted data is estimated at the receiver. Both primary and secondary problems are shown to exhibit critical behaviour at key code rates.

Part II [15] develops the notion of posterior distributions for a family of communication problems (parametrized by the code rate) in which the source is an *infinite* sequence of bits. In this case, posterior distributions cannot be defined in the conventional Bayesian way (likelihood $\times$ prior, suitably normalized) because of measurability issues. Instead, we define them to be minimizers of a *specific* apparent information, in a construction that mirrors the Dobrushin construction of statistical mechanics [3], [19], [9]. The family of communication problems is shown to undergo a second order phase transition at the channel capacity, in the sense that there is more than one posterior distribution at this code rate. Posterior distributions for the infinite problems can be used to investigate the properties of large, finite problems; their properties both below and above capacity are examined in detail in Part II.

Sections 1.1–1.4 of this introduction review the ingredients that will be used in this study, and section 1.5 places the study in the context of other results in this field. Section 2 applies the variational Bayesian method to obtain scaling limits for various information quantities related to reliable communication such as the *reliability function*, and Section 3 evaluates scaling limits for the primary Bayesian problem. Sections 1.1 and 1.2 review the variational formulae in an abstract setting. Readers not interested in such generalities could skip these sections on a first reading, and refer back to them when they are later used in more specific estimation problems.

Convention on Logarithms. The base of logarithms throughout this paper is 2, unless explicitly stated otherwise; all information quantities are, therefore, measured in *bits*. The notation "exp" is frequently used for the inverse log; i.e. $\exp(x) := 2^x$. Furthermore, $\log 0 := -\infty$ and $0 \log 0 := 0$.

1.1. Regular Bayesian Estimation. The essence of Bayesian estimation is the construction of *posterior* distributions for estimands from *prior* and *observation* information. In all but the simplest, discrete problems such posterior distributions are defined only up to sets of observation values of probability zero. This is highly unsatisfactory from the point of view of applications, where one often wishes to compute a posterior distribution corresponding to a single outcome of the observation. The standard (usually implicit) solution to this problem is to construct *regular* ver-

sions of posterior distributions having some continuity property with respect to the observation.

Suppose that the estimand, U, and observation, Y, are random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, that take values in metric spaces $\mathbb{U}$ and $\mathbb{Y}$, respectively. Let $\mathcal{U}$ and $\mathcal{Y}$ be the Borel σ-algebras of subsets of these spaces, let P_U, P_Y and $P_{U,Y}$ be the marginal and joint distributions of U and Y, and suppose that

$$P_{U,Y} \ll P_U \otimes P_Y. \tag{1}$$

Let ϕ_Y be a σ-finite (reference) measure on $\mathcal{Y}$ such that $P_{U,Y} \ll P_U \otimes \phi_Y$. ($P_Y$ is an example of such a measure; however, where one exists, it is common to choose for ϕ_Y a *uniform* measure such as Lebesgue measure on the real line.) Let $\tilde{Q}$ be a version of the density (Radon-Nikodym derivative) $dP_{U,Y}/d(P_U \otimes \phi_Y)$, and let

$$\bar{\mathbb{Y}} := \left\{ y \in \mathbb{Y} : \ 0 < \int_U \tilde{Q}(u,y) P_U(du) < \infty \right\}; \tag{2}$$

then it can easily be shown that $\bar{\mathbb{Y}} \in \mathcal{Y}$ and $P_Y(\bar{\mathbb{Y}}) = 1$. Let

$$Q(u,y) := \tilde{Q}(u,y)\mathbf{1}_{\bar{\mathbb{Y}}}(y) + \mathbf{1}_{\mathbb{Y}\setminus\bar{\mathbb{Y}}}(y), \tag{3}$$

and, for any $B \in \mathcal{U}$, let

$$P_{U|Y}(B,y) := \frac{\int_B Q(u,y) P_U(du)}{\int_{\mathbb{U}} Q(u,y) P_U(du)}; \tag{4}$$

then $P_{U|Y}$ is a regular conditional distribution for U:

(R1) $P_{U|Y}(\,\cdot\,, y)$ is a probability measure on $\mathcal{U}$ for each y;

(R2) $P_{U|Y}(B, \,\cdot\,)$ is $\mathcal{Y}$-measurable for each B;

(R3) for any $B \in \mathcal{U}$ and any $C \in \mathcal{Y}$, $\int_C P_{U|Y}(B,y) P_Y(dy) = P_{U,Y}(B \times C)$.

($Q(\,\cdot\,, y)$ can be interpreted as the likelihood function for the observed value $Y = y$.)

REMARK 1.1. *By exploiting (1), this construction of a regular conditional distribution avoids the usual Polish space assumptions on* $\mathbb{U}$. *(See, for example, [4].) In fact it does not even rely on the metric nature of* $(\mathbb{U},\mathcal{U})$ *and* $(\mathbb{Y},\mathcal{Y})$, *and can be applied to general measurable spaces. However, even when* $\mathbb{U}$ *and* $\mathbb{Y}$ are *metric spaces, it is not always clear that a continuous version of* $y \mapsto P_{U|Y}(\,\cdot\,, y)$ *(with respect, for example, to the weak topology in the range space) can be constructed without further structure. In many special cases, such as those investigated here, continuity is obtained by construction. See [1] and [2] for a problem of nonlinear filtering, and [23] for a more abstract result.*

1.2. Variational Bayes. In order to make the information quantities we shall use finite, we strengthen (1), requiring the mutual information between U and Y to be finite:

$$I(U\,;\,Y) := \int_{\mathbb{U}\times\mathbb{Y}} \log\left(\frac{dP_{U,Y}}{d(P_U \otimes P_Y)}(u,y) \right) P_{U,Y}(d(u,y)) < \infty. \tag{5}$$

Let

$$\tilde{\mathbb{Y}} := \bar{\mathbb{Y}} \cap \left\{ y \in \mathbb{Y} : \int_{\mathbb{U}} (x \log x)(\tilde{Q}(u,y)) P_U(du) < \infty \right\},$$

where $\tilde{Q}$ and $\bar{\mathbb{Y}}$ are as in section 1.1; then it easily follows from (5) that $\tilde{\mathbb{Y}} \in \mathcal{Y}$ and $P_Y(\tilde{\mathbb{Y}}) = 1$. Replacing $\bar{\mathbb{Y}}$ by $\tilde{\mathbb{Y}}$ in (3), we obtain a (new) likelihood function Q, for which

$$\int_{\mathbb{U}} (x \log x)(Q(u,y)) P_U(du) < \infty \quad \text{for all } y \in \mathbb{Y}. \tag{6}$$

This gives rise, as in (4), to a (new) regular conditional distribution $P_{U|Y}$, which will be used in all that follows. We denote by $H : \mathbb{U} \times \mathbb{Y} \to (-\infty, +\infty]$ the corresponding (negative) log-likelihood function:

$$H(u,y) := -\log Q(u,y). \tag{7}$$

Let $\mathcal{P}(\mathcal{U})$ be the set of probability measures on $\mathcal{U}$, and $\mathcal{M}(\mathbb{U})$ the set of measurable $(-\infty, +\infty]$-valued functions on $(\mathbb{U}, \mathcal{U})$. For $\tilde{P}_U \in \mathcal{P}(\mathcal{U})$ and $\tilde{H} \in \mathcal{M}(\mathbb{U})$, let

$$\begin{aligned}
h(\tilde{P}_U \,|\, P_U) &:= \int_{\mathbb{U}} (x \log x) \left(\frac{d\tilde{P}_U}{dP_U}(u) \right) P_U(du) \quad \text{if } \tilde{P}_U \ll P_U \\
&\quad +\infty \qquad\qquad\qquad\qquad\qquad\quad \text{otherwise,} \\
\langle \tilde{H}\,,\, \tilde{P}_U \rangle &:= \int_{\mathbb{U}} \tilde{H}(u) \tilde{P}_U(du) \quad \text{if the integral exists} \\
&\quad +\infty \qquad\qquad\qquad\quad \text{otherwise,} \\
i(\tilde{H}) &:= -\log \int_{\mathbb{U}} \exp(-\tilde{H}(u)) P_U(du) \quad \text{if the integral is nonzero} \\
&\quad -\infty \qquad\qquad\qquad\qquad\quad \text{otherwise.}
\end{aligned} \tag{8}$$

h is the relative entropy (Kullback Leibler divergence) of two measures. In the context of probability measures, it can be thought of as the *information gain* of the first measure over the second. (The information gain $h(P_{U|Y}(\,\cdot\,, y) \,|\, P_U)$ is finite by virtue of (6).) If we interpret $\exp(-\tilde{H})$ as a likelihood function for U, associated with some (unspecified) observation, then $\tilde{H}(u)$ is the *residual (side) information* in that observation if we already know that $U = u$, and $i(\tilde{H})$ is the *full information* in that observation, i.e. the information in the observation if the only other knowledge we have of U is its prior P_U. (See [14] for a fuller discussion of these information quantities.)

The following proposition characterizes $h(P_{U|Y}(\,\cdot\,, y) \,|\, P_U)$ in terms of $i(H(\,\cdot\,, y))$ and vice versa; a simple proof appears (as Proposition 2.1) in [14].

PROPOSITION 1.1. *Suppose that (5) is satisfied, and that H and $P_{U|Y}$ are as defined in (7) and (4) (with Q as defined in this section). Then, for any $y \in \mathbb{Y}$:*

(i)

$$i(H(\,\cdot\,, y)) = \min_{\tilde{P}_U \in \mathcal{P}(\mathcal{U})} \left\{ h(\tilde{P}_U \,|\, P_U) + \langle H(\,\cdot\,, y)\,,\, \tilde{P}_U \rangle \right\}; \tag{9}$$

(ii)

$$h(P_{U|Y}(\,\cdot\,,y)\,|\,P_U) = \max_{\tilde{H}\in\mathcal{M}(\mathbb{U})} \left\{ i(\tilde{H}) - \langle \tilde{H}\,,\, P_{U|Y}(\,\cdot\,,y)\rangle \right\}; \tag{10}$$

(iii) $P_{U|Y}(\,\cdot\,,y)$ *is the unique minimizer in (9);*

(iv) if $\hat{H}$ *is a maximizer in (10), then there exists a real constant* c *such that* $\mathbb{P}(\hat{H}(U) = H(U,y) + c) = 1$.

The term to be minimized in (9) is called in [14] the *apparent information* of the distribution $\tilde{P}_U$. It is the sum of the information gain of $\tilde{P}_U$ over the prior and the average residual information in the true likelihood function, and is thus the information *apparently* possessed by an estimator that proposes $\tilde{P}_U$ as a posterior distribution for U. This is greater than or equal to the full information in the actual observation, with equality if and only if $\tilde{P}_U = P_{U|Y}(\,\cdot\,,y)$. The two components of the apparent information show the tension between accommodating the prior and posterior information.

The term to be maximized in (10) is called in [14] the *compatible information* of the function $\tilde{H}$. It is the difference between the full information and the average (over the true posterior) residual information in the likelihood function $\exp(-\tilde{H})$; it is thus the information in $\exp(-\tilde{H})$ compatible with the true posterior. This is less than or equal to the information gain of the true posterior, with equality if and only if $\exp(-\tilde{H})$ is equivalent to the true likelihood function in the sense of part (iv).

REMARK 1.2. *The compatible information, like the relative entropy, is an* absolute *information quantity, in the sense that it does not depend on the choice of any reference measure. The full, residual and apparent information, on the other hand, are* differential *information quantities that depend on* ϕ_Y. *It is because of this that there are multiple maximizers in (10), but only one minimizer in (9).*

REMARK 1.3. *The negative of the relative entropy,* $-h(\tilde{P}_U\,|\,P_U)$, *can be thought of as the differential entropy of the probability measure* $\tilde{P}_U$ *with respect to the reference measure* P_U. *With this interpretation, the principle of apparent information minimization is one of "controlled" entropy maximization, in which the log-likelihood function* $H(\,\cdot\,,y)$ *is the control. If the probability of the observed value* y *is non-zero,* $P_Y(\{y\}) > 0$, *(as is almost surely the case, for example, if* $\mathbb{Y}$ *is discrete) then the principle of apparent information minimization coincides with that of entropy maximization over* joint *probability measures on* $\mathcal{U}\times\mathcal{Y}$, *subject to the* constraint $Y = y$. *In fact, if* $\tilde{P}_{U,Y}$ *is a probability measure satisfying this constraint then, for any* $B\in\mathcal{U}$ *and* $C\in\mathcal{Y}$,

$$\tilde{P}_{U,Y}(B\times C) = \tilde{P}_U(B)\mathbf{1}_C(y).$$

The differential entropy of $\tilde{P}_{U,Y}$ *with respect to* $P_{U,Y}$ *is* $-h(\tilde{P}_{U,Y}\,|\,P_{U,Y})$. *In order for this to be finite, we must choose* $\tilde{P}_{U,Y}\ll P_{U,Y}$, *and this ensures that* $P_{Y|U}(\{y\},u) > 0$

for $\tilde{P}_U$-almost all u. So

$$\frac{d\tilde{P}_{U,Y}}{dP_{U,Y}}(u,\tilde{y}) = \frac{d\tilde{P}_U}{dP_U}(u)\frac{\mathbf{1}_{\{y\}}(\tilde{y})}{P_{Y|U}(\{y\},u)} \quad \text{for } \tilde{P}_{U,Y}\text{−a.a. } (u,\tilde{y}),$$

and

$$\begin{aligned} -h(\tilde{P}_{U,Y} \,|\, P_{U,Y}) &= -h(\tilde{P}_U \,|\, P_U) - \int \log\left(\frac{\mathbf{1}_{\{y\}}(\tilde{y})}{P_{Y|U}(\{y\},u)}\right)\tilde{P}_{U,Y}(d(u,\tilde{y})) \\ &= -\left(h(\tilde{P}_U \,|\, P_U) + \langle H(\,\cdot\,,y)\,,\,\tilde{P}_U\rangle\right), \end{aligned}$$

which is the negative of the apparent information. However, this interpretation is not *valid if $P_Y(\{y\}) = 0$ (as is the case, for example, if $\mathbb{Y}$ is the real line and P_Y has a density). In this case the differential entropy of $\tilde{P}_{U,Y}$ is $-\infty$ for all probability measures satisfying the constraint $Y = y$. In this sense, the principle of apparent information minimization is a generalization of that of constrained entropy maximization as championed by Jaynes [10].*

We would argue that the variational characterization of the posterior distribution in (9) is *more fundamental* than that in (4). Not only does it define the correct posterior distribution, but it also says something about the consequences of getting this distribution wrong. The *information excess* of a putative posterior (its apparent information minus that of the true posterior) is a non-application-specific measure of the error in an incorrect posterior, which may occur, for example, in parametric approximations of posterior distributions. Furthermore, as discussed in [14], the *information deficit* in the inverse problem of Proposition 1.1 parts (ii) and (iv) (the true compatible information minus that of an incorrect log-likelihood function) is a fundamental measure of the error associated with the use of an incorrect likelihood function in the Bayes formula. (It can also be adapted to show the effect of an incorrect prior [14].) Such errors may have their origin in modelling or measurement errors. This view is supported by the connection between the apparent information and the free energy of statistical mechanics. The concept of free energy minimization is more fundamental to statistical mechanics than the particular formula for a Gibbs measure to which it gives rise.

1.3. Gibbs Measures in Statistical Mechanics. Statistical mechanics is a discipline that studies large systems exhibiting random behaviour. One of its earliest appearances was in the theory of heat, [12]. Consider, for example, a sealed cylinder containing n molecules of a gas. This system is "large" in the sense that it has $6n$ degrees of freedom corresponding to the three position co-ordinates and three components of momentum of each of the n molecules, and n is large (eg. 10^{20}). The dynamical equations of the system can be found from its *energy function* $E : \mathbb{R}^{3n} \times \mathbb{R}^{3n} \to \mathbb{R}^+$, which defines the energy of the system in the state (q,p), where q is the $3n$-vector of position co-ordinates, and p is the $3n$-vector of momentum components

of the molecules. The *Hamiltonian* dynamical equations are as follows:

$$\left.\begin{aligned} \dot{q}_i(t) &= \frac{\partial E}{\partial p_i}(q(t), p(t)) \\ \dot{p}_i(t) &= -\frac{\partial E}{\partial q_i}(q(t), p(t)). \end{aligned}\right\} \quad \text{for } i = 1, 2, \ldots, 3n \tag{11}$$

E is typically the sum of two terms: a *kinetic* term, which is quadratic in p, and a *potential* term, which depends only on q, and models the interactive forces between pairs of molecules, and between individual molecules and the walls of the cylinder.

It would seem that one could fully determine the state of the gas at time $t > 0$ from a knowledge of its state at time 0 by solving (11). In practice, however, n is too large to make this feasible. Even if this were not so, there are still fundamental questions regarding the accuracy of (11) on very small time and space scales, and its sensitivity to small perturbations in $(q(0), p(0))$. The classical approach to such problems is to switch to a *macroscopic* description of the system, involving the quantities: *absolute temperature*, T, *volume*, V, and *pressure*, P. The *ideal gas law* states that

$$PV = \alpha T, \tag{12}$$

where α is a constant depending on the number and mass of the molecules, [17]. The origin of (12) is physical experimentation, and it is the job of statistical mechanics to bridge the gap between it and (11). The *stochastic dynamics* approach considers the "phase space" process $((q(t), p(t)); t \in \mathbb{R})$ to be random, and to have an evolving distribution, $(\Pi_t, t \in \mathbb{R})$. The central ansatz of this approach is that, when the gas is in *thermodynamic equilibrium*, Π minimizes the *free energy* $F : \mathcal{P}(\mathcal{B}^{6n}) \to [0, +\infty]$, defined as follows:

$$F(\tilde{\Pi}) := \frac{1}{T}\langle E\,,\,\tilde{\Pi}\rangle - S(\tilde{\Pi}), \tag{13}$$

where $\langle E\,,\,\tilde{\Pi}\rangle$ is the *internal energy* (defined in the same way as $\langle \tilde{H}\,,\,\tilde{P}\rangle$ in (8)), $S : \mathcal{P}(\mathcal{B}^{6n}) \to [-\infty, \infty)$ is the *entropy*,

$$S(\tilde{\Pi}) := -h(\tilde{\Pi}\,|\,\lambda^{\otimes 6n}), \tag{14}$$

and $\lambda^{\otimes m}$ is Lebesgue measure in $\mathbb{R}^m$. The minimum free energy, $F(\Pi)$, or (equivalently) the *partition function*, $\exp(-F(\Pi))$, can be used to derive the macroscopic laws of the system, such as (12). (See, for example, [8] or [17].)

A straightforward calculation shows that, in this finite setting, the minimizer of free energy is the following *Gibbs distribution*:

$$\Pi(B) = \frac{\int_B \exp(-E(q,p)/T)\lambda^{\otimes 6n}(d(q,p))}{\int_{\mathbb{R}^{6n}} \exp(-E(q,p)/T)\lambda^{\otimes 6n}(d(q,p))} \quad \text{for } B \in \mathcal{B}^{6n}. \tag{15}$$

The connection between Bayesian estimation and statistical mechanics can be seen in its simplest form in the similarity between (4, 9) and (15, 13). These formulae

are identical if $\lambda^{\otimes 6n}$ is identified with the prior distribution P_U, and E/T is identified with the log-likelihood function H. Although the Bayes and Gibbs formulae, (4) and (15), completely characterize the posterior and Gibbs distributions, $P_{U|Y}$ and Π, respectively, their variational counterparts say much more about the nature of $P_{U|Y}$ and Π. However, the true power of the variational method is that it permits extension to problems not admitting the simple forms (4) and (15). This aspect will be investigated in Part II.

1.4. Error Control Coding. The problem we consider is that of reliably communicating information across an error prone channel. This domain has been extensively researched since the seminal paper by Shannon, [20]. We have chosen a particularly simple problem from this domain to illustrate our ideas. Nevertheless, these are applicable to much wider class of communication problems. We begin by introducing some notation.

DEFINITION 1.1.

(i) $(\Omega, \mathcal{F}, \mathbb{P})$ *is a complete probability space on which all random quantities are defined.*

(ii) $\mathbb{N}$ *is the set of natural numbers.*

(iii) For each $n \in \mathbb{N}$, $J_n := \{0, 1, \ldots, n\}$.

(iv) $\mathbb{X}$ *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*-sequences of bits* $(x_k \in \{0,1\};\, k = 1, 2, \ldots, K)$.

(v) $\|\cdot\|$ *is the Hamming norm on* $\mathbb{X}_K$*:*

$$\|v\| = \|v\|_K := \sum_{k=1}^{K} v_k. \tag{16}$$

(vi) For each $K \in \mathbb{N}$, $T_K : \mathbb{X} \to \mathbb{X}_K$ *is the* truncation map*:*

$$T_K x := (x_1, x_2, \ldots, x_K)\,. \tag{17}$$

(vii) The symbol $\leftrightarrow$ *is the* join *operator: for* $x \in \mathbb{X}_K$ *and* $y \in \mathbb{X}$

$$x \leftrightarrow y := (x_1, x_2, \ldots, x_K, y_1, y_2, \ldots). \tag{18}$$

(viii) $\mathcal{X}$ *is the product* σ*-algebra on* $\mathbb{X}$*, i.e. the* σ*-algebra of subsets of* $\mathbb{X}$ *generated by the finite-dimensional sets* $\{T_K^{-1}(\{v\});\, v \in \mathbb{X}_K, K \in \mathbb{N}\}$. *For each* $K \in \mathbb{N}$, $\mathcal{X}_K$ *is the* σ*-algebra of* all *subsets of* $\mathbb{X}_K$.

(ix) For any σ*-algebra* $\mathcal{B}$, $\mathcal{P}(\mathcal{B})$ *is the set of probability measures on* $\mathcal{B}$.

(x) For each $\theta \in [0,1]$, Q_θ *is the Bernoulli measure on* $\mathcal{X}_1$*:*

$$Q_\theta(\{0\}) = 1 - \theta \quad \text{and} \quad Q_\theta(\{1\}) = \theta. \tag{19}$$

The noisy channel coding problem considered here is that of *reliably* transmitting a *source* sequence $U \in \mathbb{X}$ across a *Binary Symmetric Channel* (BSC). The latter is characterized by the random *error sequence* $\Psi : \Omega \to \mathbb{X}$, which is an iid sequence of bits having the Bernoulli distribution Q_q for some $q \in (0, 1/2)$:

$$\mathbb{P}(\Psi_n = x) = Q_q(\{x\}) \quad \text{for } x \in \{0,1\} \text{ and } n \in \mathbb{N}. \tag{20}$$

The channel inverts the n'th bit transmitted over it if and only if $\Psi_n = 1$.

Reliable communication can be achieved by the use of *block coding* with a *random code ensemble* [7], [20]. In the present context, the latter is a random map $\Gamma : \Omega \times \mathbb{X} \to \mathbb{X}$, for which the family of random variables $(\Gamma(\,\cdot\,, u)_n,\ u \in \mathbb{X}, n \in \mathbb{N})$ is independent of Ψ, and iid with common distribution $Q_{1/2}$. In a (K, N) block coding scheme, the first K bits of the source sequence U form a *source word* $T_K U$ that is encoded into an N-bit *codeword* $X_{K,N}$ as follows:

$$X_{K,N}(\omega) := T_N \Gamma(\omega, T_K U \leftrightarrow \mathbf{0}), \tag{21}$$

where $\mathbf{0}$ is the zero element of $\mathbb{X}$. It is this codeword that is transmitted bit-by-bit across the channel, resulting in the channel output word

$$Y_{K,N}(\omega) := X_{K,N}(\omega) \oplus T_N \Psi(\omega). \tag{22}$$

REMARK 1.4. *The use of random block codes, as pioneered by Shannon [20], is a simple way of defining a* sequence *of codes (indexed by K) with some common defining feature, when finding scaling limits. The randomness of a code is an additional potential source of communication errors, since two different words $v, \tilde{v} \in \mathbb{X}_K$ may give rise, by chance, to very close or even identical codewords $T_N\Gamma(\omega, v \leftrightarrow \mathbf{0})$ and $T_N\Gamma(\omega, \tilde{v} \leftrightarrow \mathbf{0})$. However, as a result of the strong law of large numbers, random block codes have good error performance in the limit of large K. This fact is verified in Theorem 2.1 below.*

The receiver knows $Y_{K,N}$ and Γ, and must compute an estimate of the source word $T_K U$. Of course, the Bayesian approach makes the assumption that $T_K U$ has some *prior* distribution $P_{U,K} \in \mathcal{P}(\mathcal{X}_K)$, and computes the appropriate posterior distribution, $\Pi_{K,N} : \Omega \to \mathcal{P}(\mathcal{X}_K)$, as in (4):

$$\Pi_{K,N}(\omega)(B) := \frac{\sum_{v \in B} q^{\rho_{K,N}(\omega,v)} (1-q)^{N - \rho_{K,N}(\omega,v)} P_{U,K}(\{v\})}{\sum_{v \in \mathbb{X}_K} q^{\rho_{K,N}(\omega,v)} (1-q)^{N - \rho_{K,N}(\omega,v)} P_{U,K}(\{v\})} \quad \text{for } B \in \mathcal{X}_K. \tag{23}$$

Here $\rho_{K,N} : \Omega \times \mathbb{X}_K \to J_N$ is the Hamming distance between the channel output word and the codeword corresponding to v:

$$\rho_{K,N}(\omega, v) := \left\| Y_{K,N}(\omega) \oplus T_N \Gamma(\omega, v \leftrightarrow \mathbf{0}) \right\|. \tag{24}$$

In what follows, $P_{U,K}$ will always be the *uniform* prior $Q_{1/2}^{\otimes K}$. The maximum a-posteriori probability (MAP) estimator of $T_K U$ then takes the form

$$\hat{X}_K(\omega) := \arg\max_{v \in \mathbb{X}_K} \Pi_{K,N}(\omega)(\{v\}) = \arg\min_{v \in \mathbb{X}_K} \rho_{K,N}(\omega, v).$$

REMARK 1.5. *In section 3 we shall regard $Y_{K,N}$ as being the observation in this primary Bayesian problem, and Γ as being a random parametrization. As in (4), $\Pi_{K,N}$ depends on the outcome of the observation, $Y_{K,N}(\omega)$; however, it also depends on the outcome of the code, $\Gamma(\omega, \cdot)$. Both of these dependencies appear in (23) through ω.*

We consider *sequences* of such block coding problems indexed by K, with $N = [R^{-1}K]$ for some *code rate* $R \in (0, \infty)$. (We assume, throughout, that $K > R$ so that $N \geq 1$.) The classical result of Shannon [20] states that reliable communication can be achieved at all code rates less than a well defined *channel capacity* C, but not at rates exceeding C. By "reliable communication" we mean that the error probability of the MAP estimator can be made arbitrarily small by the use of sufficiently large values of K (for a fixed value of R). It is shown in [20] that the channel capacity C is the value of the mutual information between the input and output bits of the error prone channel, maximized over the distribution of the input bits. In the case of the binary symmetric channel, the maximizing distribution is $Q_{1/2}$, and $C = 1 - b(q)$, where $b : [0, 1] \to [0, 1]$ is the *binary entropy function*

$$b(\theta) = -\theta \log \theta - (1 - \theta) \log(1 - \theta). \tag{25}$$

(See, for example, [7].)

Even though reliable communication, in this sense, is not possible at code rates exceeding C, useful communication may still be possible. For example, it may be acceptable, in a particular application, for decoding errors to occur, provided that the *size* of the errors is not too great. (The latter may be defined by some distortion metric on the source space, $\mathbb{X}_K$.) The decoding strategy would then be to pick the word from the source space that minimized the posterior mean distortion. The minimum average distortion would typically be smaller than that arising from pure guesswork, even if the code rate were greater than C, and this is the reason for including such code rates in this study. The use of a code rate in excess of 1 would be appropriate if we were prepared to accept a larger average distortion than that arising from the raw use of the channel. (Such code rates would have obvious advantages in the effective transmission rate they achieved over the channel.) The posterior distributions for the infinite systems considered in Part II are useful in the study of such issues.

1.5. Discussion. Although random block codes have good error performance in the limit of large K, they require very computationally demanding decoders. In fact, even the *specification* of the code requires tables of a size that is 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 far simpler decoders, and a major role in this quest has been played by *linear codes*, for which the non-random equivalent of $T_N\Gamma(\omega, T_K U \leftrightarrow \mathbf{0})$ of section 1.4 is a linear map from $\mathbb{X}_K$ to $\mathbb{X}_N$. 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, [22]. Because of the resulting linear correlation structure in the codewords, such codes cannot be used at rates very close to capacity. However, in the last 15 years, the important classes of *turbo codes* and *low density parity check (LDPC) codes* have dominated the research literature because they admit simple *iterative* decoders, and yet are capable of achieving reliable communication at rates close to capacity [18].

In 1989, Sourlas [21] 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, [5], [6], [11], [13], [16], [18]. 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 problem, i.e. the decoder that maximizes the posterior *marginal* distributions of the individual bits. The MAP word decoder is recovered in [21] by the introduction of a new positive "hyper parameter", β, which has the effect of altering the dependency between the individual bits in the posterior distribution. When $\beta = 1$ no change is made to this dependency, and so the spin glass system corresponds to the MAP bit decoder; however, in the limit of large β, the spin glass system corresponds to the MAP word decoder. The interpretation of β as an "inverse temperature" parameter leads to the terminology "high temperature" decoding for any decoder with a small value of β. "Raising the temperature of the decoder" above that of the MAP word decoder (absolute zero) reduces long range dependencies in the associated posterior word distribution, and so helps to reduce decoder complexity (at the cost of accuracy). In the context of the channel coding problem of section 1.4, the spin glass system corresponds to the MAP bit decoder based on an erroneous model that assumes a value

$$\tilde{q} := \frac{q^\beta}{q^\beta + (1-q)^\beta},$$

for the channel error probability.

The results of the present paper and Part II [15] concern the true word posterior only, and so are somewhat different from those in [5], [11], [13], [16] and [21]. They do not involve the hyper parameter β, but show rather that the code rate R plays the role of temperature in the statistical mechanical interpretation of MAP word decoding.

2. Reliable Communication. The *region of reliable communication* for the channel coding problem of section 1.4 is the range of code rates, R, for which the error probability of the MAP word decoder vanishes with increasing K. Within this region, decoding errors are rare events and can be characterized by their large deviation

rate function, which is known as the *reliability function* in the coding literature. The reliability function of the binary symmetric channel with random coding is well known (see, for example, [7]); however, we show here that it arises naturally in the variational Bayes formulae for secondary estimation problems in which the observation is the decoding error event, or its complement. For example, because the decoding error event is rare in the region of reliable communication, its observation bears significant information on the channel error sequence, Ψ, and/or the random code, Γ. The reliability function specifies the *full information* in this observation as a function of the code rate R. By computing the *information gains* that this observation provides on Ψ and Γ, we can identify the dominant source of errors.

We consider a sequence of (K, N) block codes, indexed by K $(> R)$, with $N = [R^{-1}K]$ for some fixed code rate $R \in (0, \infty)$. For each K, we define the MAP *word decoding error event* as follows:

$$E_K := \{\omega \in \Omega : \ \mu_K(\omega) \leq \nu_K(\omega)\}, \tag{26}$$

where

$$\begin{aligned} \mu_K(\omega) &:= \min_{v \in A_K} \rho_{K,N}(\omega, v), \\ A_K &:= \mathbb{X}_K \setminus \{T_K U\}, \\ \nu_K(\omega) &:= \rho_{K,N}(\omega, T_K U). \end{aligned} \tag{27}$$

REMARK 2.1. *E_K is the event that the channel output word is at least as close (in Hamming distance) to the codeword corresponding to some $v \in A_K$ as it is to the transmitted codeword. If more than one $v \in \mathbb{X}_K$ minimizes this distance then there is more than one MAP estimate of $T_K U$, and it is not clear what the receiver's strategy should be. Our definition of E_K is "pessimistic" in the sense that it contains the error events for* all *MAP strategies that produce a single estimate. To use a strict inequality in (26) would be "optimistic", since E_K would then be the error event for a receiver that always chose $T_K U$ when it was one of a set of MAP estimates. However, with one exception (mentioned at the very end of this section), the results that follow are not sensitive to such distinctions, and remain true if the inequality in (26) is made strict.*

Straightforward arguments show that the family $\{(Y_{K,N} \oplus T_N \Gamma(\,\cdot\,, v \leftrightarrow \mathbf{0}))_n,\ v \in A_K,\ 1 \leq n \leq N\}$ is independent of Ψ, and iid with common distribution $Q_{1/2}$. In particular, ν_K and μ_K are independent.

DEFINITION 2.1.

(i) Let $f : D_f \to \mathbb{R}$ and $g : D_g \to \mathbb{R}$ for some intervals of the real line, D_f and D_g. We say that (f, g) is a Fenchel-Legendre pair *if f is convex, g is concave,*

$$g(t) = \inf_{\theta \in D_f} \{f(\theta) + \theta t\} \quad \text{for all } t \in D_g \tag{28}$$

and

$$f(\theta) = \sup_{t \in D_g} \{g(t) - \theta t\} \quad \text{for all } \theta \in D_f. \tag{29}$$

(ii) For any code rate $R \in (0, \infty)$,

$$\theta_{\mathrm{GV}}(R) := b^{-1}((1-R)^+), \tag{30}$$

where b^{-1} *is the inverse of the binary entropy function of (25), when the latter is restricted to the sub-domain* $[0, 1/2]$, *and, for any* $\alpha \in \mathbb{R}$, $\alpha^+ := \max\{0, \alpha\}$.

(iii) The functions $f_\nu : [0,1] \to [0,\infty)$, $f_\mu : [0,1] \times (0,\infty) \to (-\infty, 1)$, $g_\nu : \mathbb{R} \to (-\infty, 1)$ *and* $g_\mu : \mathbb{R} \times (0,\infty) \to (-\infty, 1)$ *are as follows:*

$$\begin{aligned} f_\nu(\theta) &:= h(Q_\theta \,|\, Q_q) = \theta \log\left(\frac{\theta}{q}\right) + (1-\theta)\log\left(\frac{1-\theta}{1-q}\right) \\ f_\mu(\theta, R) &:= h(Q_\theta \,|\, Q_{1/2}) - R \;=\; 1 - b(\theta) - R \\ g_\nu(t) &:= -\log(1 - q + q2^{-t}) \\ g_\mu(t, R) &:= 1 - \log(1 + 2^{-t}) - R. \end{aligned} \tag{31}$$

Remark 2.2. *If* $R \le 1$ *then* $\theta_{\mathrm{GV}}(R)$ *is the unique* $\theta \in [0, 1/2]$ *for which* $h(Q_\theta \,|\, Q_{1/2}) = R$. *When* K *is large,* $N\theta_{\mathrm{GV}}(R)$ *approximates the Hamming distance between codewords of a deterministic code for which the Gilbert-Varshamov bound is* 2^K. *(See, for example, [7].) The channel capacity,* C, *is the unique value of* R *for which* $\theta_{\mathrm{GV}}(R) = q$.

The following lemma summarizes some large deviation estimates for the sequences $(\nu_K;\, K \in \mathbb{N})$ and $(\mu_K;\, K \in \mathbb{N})$ of (27). It is a standard result; however, an outline proof is included in the Appendix for the sake of completeness. In this Lemma, and all subsequent results, $N = [R^{-1}K]$.

Lemma 2.1.

(i) The pair (f_ν, g_ν) *is a Fenchel-Legendre pair.*

(ii) For any $R \in (0,\infty)$, *the pair* $(f_\mu(\,\cdot\,, R), g_\mu(\,\cdot\,, R))$ *is a Fenchel-Legendre pair.*

(iii) For any $R \in (0,\infty)$ *and any* $\theta \in [0,1]$,

$$\begin{aligned} \lim_K N^{-1} \log \mathbb{P}(\nu_K \le N\theta) &= -f_\nu(\theta)\mathbf{1}_{[0,q]}(\theta), \\ \lim_K N^{-1} \log \mathbb{P}(\nu_K \ge N\theta) &= -f_\nu(\theta)\mathbf{1}_{[q,1]}(\theta), \\ \lim_K N^{-1} \log \mathbb{P}(\mu_K \le N\theta) &= -f_\mu(\theta, R)\mathbf{1}_{[0,\theta_{\mathrm{GV}}(R))}(\theta). \end{aligned} \tag{32}$$

(iv) For any $R \in (0,\infty)$ *and any* $\theta \in [0, 1/2)$,

$$\lim_K N^{-1} \log(-\log \mathbb{P}(\mu_K > N\theta)) = -f_\mu(\theta, R). \tag{33}$$

(v) If $R = 1$ *then* $\mathbb{P}(\mu_K > 0) \to e^{-1}$.

Consider the Bayesian problems of estimating ν_K and μ_K (jointly or separately) from the observation that the decoding error event E_K has occurred. The prior and posterior distributions in these problems will be denoted $P^K_{\nu,\mu}$, P^K_ν, P^K_μ, $P^K_{\nu,\mu|E}$, $P^K_{\nu|E}$ and $P^K_{\mu|E}$. The log-likelihood functions are as follows:

$$\begin{aligned} H^K_{\nu,\mu|E}(n,m) &:= -\log \mathbb{P}(E_K \,|\, \nu_K = n, \mu_K = m) \\ &= 0 \qquad \text{if } n \geq m \\ &\quad +\infty \quad \text{otherwise,} \\ H^K_{\nu|E}(n) &:= -\log P^K_\mu(J_n), \\ H^K_{\mu|E}(m) &:= -\log P^K_\nu(J_N \setminus J_{m-1}), \end{aligned} \tag{34}$$

where we have used the fact that $P^K_{\nu,\mu} = P^K_\nu \otimes P^K_\mu$ as discussed after Remark 2.1, and J_n is as in Definition 1.1. The *full information* of E_K (in the sense of section 1.2) in all three problems has the common value

$$i(H^K_{\nu,\mu|E}) = i(H^K_{\nu|E}) = i(H^K_{\mu|E}) = -\log \mathbb{P}(E_K). \tag{35}$$

Consider also the Bayesian problems of estimating ν_K and μ_K from the observation that the error event E_K has *not* occurred. The posterior distributions in these problems will be denoted $P^K_{\nu,\mu|\bar{E}}$, $P^K_{\nu|\bar{E}}$ and $P^K_{\mu|\bar{E}}$. (We define $\bar{E}_K := \Omega \setminus E_K$.) The log-likelihood functions are as follows:

$$\begin{aligned} H^K_{\nu,\mu|\bar{E}}(n,m) &:= -\log \mathbb{P}(\bar{E}_K \,|\, \nu_K = n, \mu_K = m) \\ &= 0 \qquad \text{if } n < m \\ &\quad +\infty \quad \text{otherwise,} \\ H^K_{\nu|\bar{E}}(n) &:= -\log P^K_\mu(J_N \setminus J_n), \\ H^K_{\mu|\bar{E}}(m) &:= -\log P^K_\nu(J_{m-1}). \end{aligned} \tag{36}$$

Once again, the full information of $\bar{E}_K$ in all three problems has a common value:

$$i(H^K_{\nu,\mu|\bar{E}}) = i(H^K_{\nu|\bar{E}}) = i(H^K_{\mu|\bar{E}}) = -\log \mathbb{P}(\bar{E}_K). \tag{37}$$

The variational Bayes formulae of section 1.2 express the *information gains* on (ν,μ), ν and μ arising from the observation of E_K or $\bar{E}_K$ in terms of the full information quantities of (35) and (37) and the mean values of the log-likelihood functions of (34) and (36).

Of course, in the context of a *sequence* of coding problems, indexed by K, all these information quantities depend on K. The following theorem shows that this dependency is asymptotically a *linear* increase in all cases, and evaluates scaling limits. Its proof, which makes repeated use of Proposition 1.1, is given in the appendix.

THEOREM 2.1.

(i) For any $R \in (0,\infty)$, the full information of the observation sequence $(E_K, K \in \mathbb{N})$ admits the following scaling limit:

$$\begin{aligned}\lim_K N^{-1} i(H^K_{\nu,\mu|E}) &= \mathcal{I}_E(R)\\ &:= \big(f_\nu(\theta^*(R)) + f_\mu(\theta^*(R),R)\big)\mathbf{1}_{(0,C]}(R),\end{aligned} \tag{38}$$

where

$$\theta^*(R) := \min\left\{\frac{\sqrt{q}}{\sqrt{(1-q)}+\sqrt{q}},\, \theta_{\mathrm{GV}}(R)\right\}. \tag{39}$$

(ii) For any $R \in (0,1]$, the full information of the observation sequence $(\bar{E}_K, K \in \mathbb{N})$ admits the following scaling limit:

$$\begin{aligned}\lim_K N^{-1} i(H^K_{\nu,\mu|\bar{E}}) &= \mathcal{I}_{\bar{E}}(R)\\ &:= f_\nu(\theta_{\mathrm{GV}}(R))\mathbf{1}_{[C,1]}(R),\end{aligned} \tag{40}$$

(iii) For any $R \in (0,\infty)$, the information gains on (ν_K, μ_K), ν_K and μ_K, arising from the observation of E_K admit the following scaling limits:

$$\begin{aligned}&\lim_K N^{-1} h(P^K_{\nu,\mu|E} \,|\, P^K_{\nu,\mu}) = \mathcal{I}_E(R),\\ &\lim_K N^{-1} h(P^K_{\nu|E} \,|\, P^K_{\nu}) = \mathcal{H}_{\nu\,|\,E}(R) := f_\nu(\theta^*(R))\mathbf{1}_{(0,C]}(R),\\ &\lim_K N^{-1} h(P^K_{\mu|E} \,|\, P^K_{\mu}) = \mathcal{H}_{\mu\,|\,E}(R) := f_\mu(\theta^*(R),R)\mathbf{1}_{(0,C]}(R),\end{aligned} \tag{41}$$

where $\theta^*(R)$ is as defined in (39).

(iv) For any $R \in (0,1]$, the information gains on (ν_K, μ_K), ν_K and μ_K, arising from the observation of $\bar{E}_K$ admit the following scaling limits:

$$\begin{aligned}\lim_K N^{-1} h(P^K_{\nu,\mu|\bar{E}} \,|\, P^K_{\nu,\mu}) &= \lim_K N^{-1} h(P^K_{\nu|\bar{E}} \,|\, P^K_{\nu})\\ &= \mathcal{I}_{\bar{E}}(R)\\ \lim_K N^{-1} h(P^K_{\mu|\bar{E}} \,|\, P^K_{\mu}) &= 0.\end{aligned} \tag{42}$$

REMARK 2.3. *$\mathcal{I}_E$ and $\mathcal{I}_{\bar{E}}$ are the* large deviation rate functions *for the events E_K and $\bar{E}_K$, respectively. For large K,*

$$\mathbb{P}(E_K) \approx \exp(-\mathcal{I}_E(R)N) \quad \text{and} \quad \mathbb{P}\left(\bar{E}_K\right) \approx \exp(-\mathcal{I}_{\bar{E}}(R)N). \tag{43}$$

The scaling limits are plotted against R in Figures 1 and 2 for the channel error probability $q = 0.05$. As well as showing the significance of the channel capacity C $(= 0.7136)$, Figure 1 also shows a secondary critical code rate

$$R^* := 1 - b\left(\frac{\sqrt{q}}{\sqrt{(1-q)}+\sqrt{q}}\right) = 0.3057. \tag{44}$$

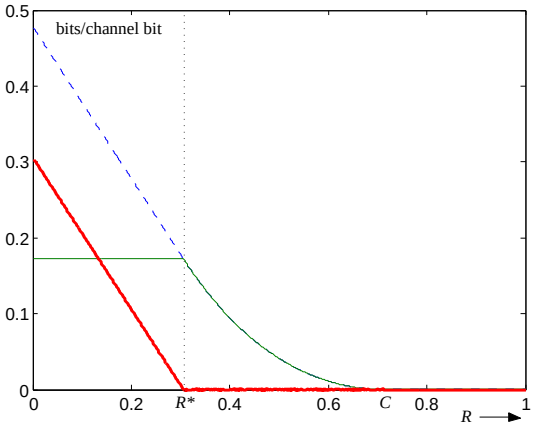

FIG. 1. *Scaling Limits for the Observation Sequence* (E_K)*: dashed line* $\mathcal{I}_E$*, thin line* $\mathcal{H}_{\nu\,|\,E}$*, thick line* $\mathcal{H}_{\mu\,|\,E}$*.* ($q = 0.05$)

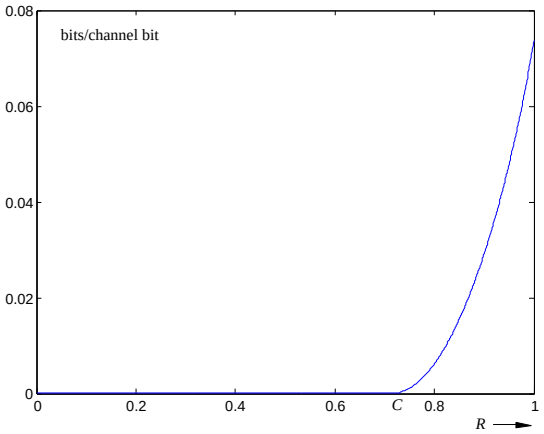

FIG. 2. *Scaling Limit for the Observation Sequence* ($\bar{E}_K$)*:* $\mathcal{I}_{\bar{E}}$*.* ($q = 0.05$)

This is the value at which the minimum in (39) switches from the first to the second term. This switch causes discontinuities in the first derivatives of $\mathcal{H}_{\nu\,|\,E}$ and $\mathcal{H}_{\mu\,|\,E}$, and in the second derivative of $\mathcal{I}_E$ at $R = R^*$. Since the latter corresponds to the scaling limit of the free energy of a statistical mechanical system (in the analogy of section 1.3), these secondary Bayesian estimation problems can be said to exhibit a third-order phase transition at this code rate.

For code rates less than C, E_K carries asymptotically significant information on the sequence (ν_K, μ_K), but $\bar{E}_K$ does not; this is because $\bar{E}_K$ is not a rare event in this range of code rates, and so its observation does not add significantly to the prior information in $P^K_{\nu,\mu}$. For code rates greater than C this distinction is reversed—the error event E_K then has a high probability of occurrence and its complement $\bar{E}_K$

is rare, and so bears significant information on (ν_K, μ_K). The second critical code rate, R^*, is that above which neither E_K nor $\bar{E}_K$ carries asymptotically significant information on μ_K. This shows that, for code rates greater than R^*, the randomness of the code does not make a significant contribution to whether or not decoding errors occur. For rates in this range, large deviations of ν_K are the dominant source of rare decoding errors when $R < C$, and rare decoding successes when $R > C$. For rates less than R^*, large deviations of *both* ν_K and μ_K occur in typical error events.

For code rates greater than 1 the scaling limits associated with the observation E all take the value 0; however, the scaling limits associated with $\bar{E}$ become dependent on the arbitrary choice that E contains any ω for which $\nu_K(\omega) = \mu_K(\omega)$. (See Remark 2.1.)

3. The Primary Bayesian Problem. The primary Bayesian problem is that of estimating the source word $T_K U$ from the channel output $Y_{K,N}(\omega)$ of (22). Assuming a uniform prior ($P_{U,K} = Q_{1/2}^{\otimes K}$), the log-likelihood function, $H_{K,N} : \Omega \times \mathbb{X} \to [0, \infty)$, the full information $i_{K,N} : \Omega \to [0, \infty)$, and the information gain $h_{K,N} : \Omega \to [0, \infty)$ are as follows:

$$H_{K,N}(\omega, v) := -\rho_{K,N}(\omega, v) \log q - (N - \rho_{K,N}(\omega, v)) \log(1 - q) \tag{45}$$

$$i_{K,N}(\omega) := -\log \sum_{v \in \mathbb{X}_K} q^{\rho_{K,N}(\omega,v)} (1-q)^{N - \rho_{k,N}(\omega,v)} 2^{-K} \tag{46}$$

$$h_{K,N}(\omega) := h(\Pi_{K,N}(\omega) \,|\, Q_{1/2}^{\otimes K}), \tag{47}$$

where $\rho_{K,N}$ is as defined in (24) and $\Pi_{K,N}$ is as defined in (23).

The following theorem shows that $i_{K,N}$ and $h_{K,N}$ scale linearly with (large) K, and finds scaling limits. It is proved in the appendix.

THEOREM 3.1.

(i) For almost all ω, the full information in the observation sequence $(Y_{K,N}, K \in \mathbb{N})$ admits the following scaling limit:

$$\lim_K K^{-1} i_{K,N}(\omega) = \mathcal{I}(R) := \begin{cases} 1 + R^{-1} b(q) & \text{if } R \le C \\ R^{-1} & \text{if } R \ge C. \end{cases} \tag{48}$$

(ii) For almost all ω, the information gain for the sequence of primary problems admits the following scaling limit:

$$\lim_K K^{-1} h_{K,N}(\omega) = \mathcal{H}(R) := \begin{cases} 1 & \text{if } R \le C \\ R^{-1} C & \text{if } R \ge C. \end{cases} \tag{49}$$

The limits $\mathcal{I}(R)$ and $\mathcal{H}(R)$ are plotted against R in Figure 3 for the fixed channel error probability $q = 0.05$. In the region of reliable communication ($R < C$) $\mathcal{H}(R) = 1$,

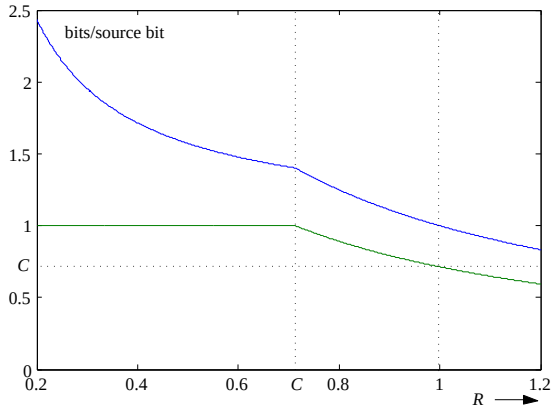

FIG. 3. *Scaling Limit Profiles for the Primary Problem: top line $\mathcal{I}(R)$, bottom line $\mathcal{H}(R)$. ($q = 0.05$)*

indicating that the scaling limit of the information gain at the receiver matches the entropy rate of the source. For code rates above capacity this is no longer true.

This behaviour has the following statistical mechanical interpretation. For each $\omega \in \Omega$ and $K \in \mathbb{N}$, let $\Sigma_{K,\omega}$ be an abstract statistical mechanical system with phase space $\mathbb{X}_K$, energy function $E_{K,\omega} : \mathbb{X}_K \to \mathbb{R}^+$, and entropy function $S_K : \mathcal{P}(\mathcal{X}_K) \to [-\infty, 0]$ defined as follows:

$$\begin{aligned} E_{K,\omega}(v) &:= RH_{K,N}(\omega, v) \\ S_K(\tilde{\Pi}) &:= -h(\tilde{\Pi} \,|\, Q_{1/2}^{\otimes K}). \end{aligned} \tag{50}$$

The full information of the primary Bayesian problem, $i_{K,N}(\omega)$, is equal to the minimum free energy of $\Sigma_{K,\omega}$; in fact it follows from (9) that

$$\begin{aligned} i_{K,N}(\omega) &= h(\Pi_{K,N} \,|\, Q_{1/2}^{\otimes K}) + \langle H_{K,N}(\omega, \cdot\,) \,,\, \Pi_{K,N}(\omega) \rangle \\ &= \min_{\tilde{\Pi} \in \mathcal{P}(\mathcal{X}_K)} \{ R^{-1} \langle E_{K,\omega} \,,\, \tilde{\Pi} \rangle - S_K(\tilde{\Pi}) \}. \end{aligned}$$

In the limit of large K, this becomes

$$\mathcal{I}(R) = R^{-1} \prec E \,,\, \Pi \succ -\mathcal{S}(\Pi), \tag{51}$$

where $\prec E \,,\, \Pi \succ$ and $\mathcal{S}(\Pi)$ are the scaling limits for the sequences of internal energy and entropy, respectively, of the statistical mechanical systems ($\Sigma_{K,\omega}$, $K \in \mathbb{N}$):

$$\begin{aligned} \lim_K K^{-1} \langle E_{K,\omega} \,,\, \Pi_{K,N}(\omega) \rangle &= \prec E \,,\, \Pi \succ = \; b(q) \quad \text{for a.a. } \omega \\ \lim_K K^{-1} S_K(\Pi_{K,N}(\omega)) &= \mathcal{S}(\Pi) = -\mathcal{H}(R) \qquad \text{for a.a. } \omega. \end{aligned}$$

Crucially, $\prec E \,,\, \Pi \succ$ does not depend on R, which shows that R plays the role of absolute temperature in $\Sigma_{K,\omega}$ for large K. (Cf. (13).)

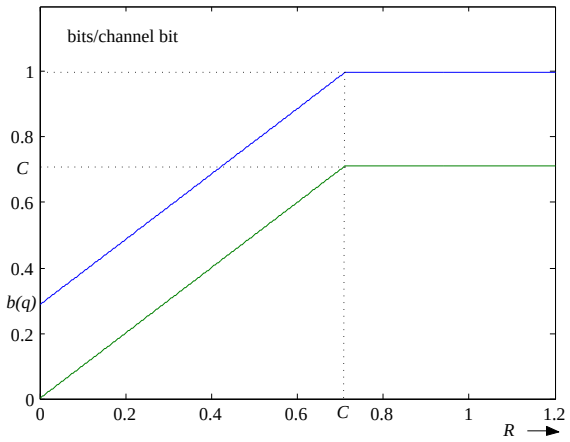

FIG. 4. *Scaling Limit Profiles normalized by N: top line $R\mathcal{I}(R)$, bottom line $R\mathcal{H}(R)$. ($q = 0.05$)*

When K is large, the normalized minimum free energy of $\Sigma_{K,\omega}$ is approximately $\mathcal{I}(R)$, which has a discontinuous derivative at $R = C$. In this sense, $\Sigma_{K,\omega}$ exhibits a *second order phase transition* at the temperature $R = C$. This separates the two distinct phases of *reliable communication* at low temperatures, and *failure of communication* at high temperatures. The phase transition is a large K limiting property of the sequence of systems $(\Sigma_{K,\omega},\ K \in \mathbb{N})$. In Part II [15] it is shown also to be a property of an infinite system that represents an extension of the Bayesian paradigm. The distinct phases below and above the critical temperature are examined there in detail.

The scaling limits $\mathcal{I}(R)$ and $\mathcal{H}(R)$ are measured in *bits per source bit*. If $i_{K,N}$ and $h_{K,N}$ were normalized by N instead of K then the resulting scaling limits would be measured in *bits per channel bit*. Graphs of $R\mathcal{I}(R)$ and $R\mathcal{H}(R)$ against R are shown in Figure 4. This shows how failure of communication at code rates exceeding C is a result of *saturation* in the entropy rate of the observation sequence $(Y_{K,N},\ K \in \mathbb{N})$; the latter has a maximum value of 1 bit per channel bit, which is not enough to carry the rates of entropy of both the source and the channel error sequences when $R > C$. (The rate of entropy of the channel error sequence is $b(q)$ bits per channel bit at all values of R.)

Theorem 3.1(ii) shows that the observation continues to supply information on U, even when the code rate exceeds the channel capacity. The posterior distribution, $\Pi_{K,N}$, is thus far from being uniform, and it is natural to ask whether it contains more information on individual bits than the (uniform) prior. This is answered in the negative in the following proposition, which is proved in the appendix. The information gain at rates above capacity thus concerns only *dependencies* between the bits of U. This is unfortunate, since in many applications the individual bits have

value in their own right. We return to this issue in Part II.

PROPOSITION 3.1. *For any $R > C$, any $k \in \mathbb{N}$ and almost all ω,*

$$\lim_K \Pi_{K,N}(\omega)(\{v \in \mathbb{X}_K : v_k = U_k\}) = 1/2. \tag{52}$$

Appendix A. Proofs.

We first prove a simple lemma on combinatorics and information. We require a slightly stronger version of this (uniformity over J_N) than that normally appearing in the literature.

LEMMA A.1. *For J_N as in Definition 1.1,*

$$\lim_N \max_{n \in J_N} \left| N^{-1} \log \binom{N}{n} - b(n/N) \right| \to 0, \tag{53}$$

where b is the binary entropy function of (25).

Proof. The term in the max is clearly zero if $n = 0$ or N. Suppose, then, that $n \in J_N \setminus \{0, N\}$. Stirling's formula shows that

$$\log_e(n!) = (n + 1/2)\log_e(n) - n + 1/2 \log_e(2\pi) + \alpha_n,$$

where $0 < \alpha_n < 1/n$, and so

$$\log_e \binom{N}{n} = -n\log_e(n/N) - (N-n)\log_e((N-n)/N) + 1/2 \log_e(N/n(N-n)) + (\alpha_N - \alpha_n - \alpha_{N-n} - 1/2 \log_e(2\pi)).$$

Changing the base of logarithms, it follows that

$$\max_{n \in J_N} \left| N^{-1} \log \binom{N}{n} - b(n/N) \right| \le \frac{1}{2N \log_e 2} \max_{n \in J_N \setminus \{0,N\}} \left| \log_e \left(\frac{N}{n(N-n)} \right) + 2(\alpha_N - \alpha_n - \alpha_{N-n} - \log_e(2\pi)) \right|,$$

and (53) follows. □

A.1. Proof of Lemma 2.1. Parts (i) and (ii) are easily verified, and the first two equations in (32) follow from an application of Cramér's theorem to the iid Bernoulli sequence $(\Psi_n,\ n = 1, 2, \ldots, N)$, the sum of which has moment generating function

$$\Phi_N(t) := \mathbb{E} \exp\left(t \sum_{n=1}^N \Psi_n \right) = (1 - q + q2^t)^N = \exp(-N g_\nu(-t)).$$

(See, for example, [4].) Similarly, for $v \in A_K$ and $\theta \in [0, 1/2]$,

$$\lim_K N^{-1} \log \mathbb{P}(\rho_{K,N}(\,\cdot\,, v) \le N\theta) = -h(Q_\theta \,|\, Q_{1/2}).$$

If $\theta \in [0, 1/2)$ then, for any $v \in A_K$,

$$\begin{aligned} 2^{-K} \log \mathbb{P}(\mu_K > N\theta) &= (1 - 2^{-K}) \log \mathbb{P}(\rho_{K,N}(\,\cdot\,, v) > N\theta) \\ &= (1 - 2^{-K}) \log(1 - \mathbb{P}(\rho_{K,N}(\,\cdot\,, v) \le N\theta)) \\ &= -\log e\, \mathbb{P}\left(\rho_{K,N}(\,\cdot\,, v) \le N\theta\right)(1 + \epsilon_K), \end{aligned} \tag{54}$$

where $\limsup_K N^{-1} \log |\epsilon_K| < 0$ and the last step is based on the second-order Taylor expansion of $\log_e(1 - x)$ about the point $x = 0$. So

$$-R + \lim_K N^{-1} \log(-\log \mathbb{P}(\mu_K > N\theta)) = -h(Q_\theta \,|\, Q_{1/2}), \tag{55}$$

and this proves (33). It also proves the third equation in (32) for the case $\theta > \theta_{\mathrm{GV}}$. If $\theta < \theta_{\mathrm{GV}}$ then it follows from (55) that

$$-R + \lim_K N^{-1} \log \mathbb{P}(\mu_K \le N\theta) = -h(Q_\theta \,|\, Q_{1/2}),$$

and this proves the third equation in (32) for the case $\theta < \theta_{\mathrm{GV}}$. The case $\theta = \theta_{\mathrm{GV}}$ follows from this since $\mathbb{P}(\mu_K \le N\theta_{\mathrm{GV}}) \ge \mathbb{P}(\mu_K \le N(\theta_{\mathrm{GV}} - \epsilon))$ for any $\epsilon > 0$.

It follows from (54) that

$$\log \mathbb{P}(\mu_K > 0) = \log(1/e) 2^{K-N} (1 + \epsilon_K),$$

and, in the special case that $K = N$, this proves part (v). □

A.2. Proof of Theorem 2.1. According to Proposition 1.1(iii), the minimizer of apparent information in the Bayesian problem for (ν_K, μ_K) given E_K is the posterior distribution, $P^K_{\nu,\mu|E}$, and so

$$< H^K_{\nu,\mu|E}\,,\, P^K_{\nu,\mu|E} >= 0 \quad \text{and} \quad h(P^K_{\nu,\mu|E} \,|\, P^K_{\nu,\mu}) = i(H^K_{\nu,\mu|E}) \tag{56}$$

It follows from the independence of ν_K and μ_K, the chain rule of relative entropy (see, for example, [4]) and Jensen's inequality that

$$h(P^K_{\nu|E} \,|\, P^K_\nu) + h(P^K_{\mu|E} \,|\, P^K_\mu) \le h(P^K_{\nu,\mu|E} \,|\, P^K_{\nu,\mu}).$$

Together with the non-negativity of relative entropy and (56), this shows that

$$0 \le h(P^K_{\nu|E} \,|\, P^K_\nu) + h(P^K_{\mu|E} \,|\, P^K_\mu) \le h(P^K_{\nu,\mu|E} \,|\, P^K_{\nu,\mu}) = i(H^K_{\nu,\mu|E}). \tag{57}$$

Similarly

$$0 \le h(P^K_{\nu|\bar{E}} \,|\, P^K_\nu) \le h(P^K_{\nu|\bar{E}} \,|\, P^K_\nu) + h(P^K_{\mu|\bar{E}} \,|\, P^K_\mu) \le h(P^K_{\nu,\mu|\bar{E}} \,|\, P^K_{\nu,\mu}) = i(H^K_{\nu,\mu|\bar{E}}), \tag{58}$$

The remainder of the proof finds upper bounds on $i(H^K_{\nu,\mu|E})$ and $i(H^K_{\nu,\mu|\bar{E}})$, and tighter lower bounds on $h(P^K_{\nu|E} \,|\, P^K_\nu)$, $h(P^K_{\mu|E} \,|\, P^K_\mu)$ and $h(P^K_{\nu|\bar{E}} \,|\, P^K_\nu)$.

Upper bounds. For each $\theta \in [0,1]$, let $H_{K,\theta}, \bar{H}_{K,\theta} : J_N^2 \to [0,\infty]$ be as follows

$$\begin{aligned} H_{K,\theta}(n,m) &:= 0 \qquad \text{if } n \geq N\theta \text{ and } m \leq N\theta \\ &\quad\ +\infty \qquad \text{otherwise} \\ \bar{H}_{K,\theta}(n,m) &:= 0 \qquad \text{if } n \leq N\theta \text{ and } m > N\theta \\ &\quad\ +\infty \qquad \text{otherwise.} \end{aligned}$$

Since $H^K_{\nu,\mu|E}(n,m) \leq H_{K,\theta}(n,m)$ for all $n,m \in J_N$,

$$i(H^K_{\nu,\mu|E}) \leq i(H_{K,\theta}) = -\log \mathbb{P}(\nu_K \geq N\theta) - \log \mathbb{P}(\mu_K \leq N\theta). \tag{59}$$

Together with the second and third equations in (32), this provides the upper bound

$$\limsup_K N^{-1} i(H^K_{\nu,\mu|E}) \leq \min_{\theta\in[0,1]} \{f_\nu(\theta)\mathbf{1}_{[q,1]}(\theta) + f_\mu(\theta,R)\mathbf{1}_{[0,\theta_{\mathrm{GV}})}(\theta)\}. \tag{60}$$

That $\theta^*(R)$ minimizes the right-hand side here for all $R \in (0,\infty)$ is easily verified. Part (iii) and (38), for the case $R \geq C$ (where the right-hand side of (60) is zero), follow directly from (57) and (60).

Similarly, for any $R \in (0,1]$,

$$\limsup_K N^{-1} i(H^K_{\nu,\mu|\bar{E}}) \leq f_\nu(\theta_{\mathrm{GV}})\mathbf{1}_{(C,1]}(R), \tag{61}$$

where we have used Lemma 2.1(v) in the case $R = 1$. Part (iv) and (40), for the case $R \leq C$, follow directly from (58) and (61).

Lower bounds. For each $t \in \mathbb{R}$, let $H_{K,t} : J_N \to \mathbb{R}$ be as follows:

$$H_{K,t}(n) := tn.$$

Consider the Bayesian problem of estimating ν_K given the observation that E_K has occurred. The full information of $H_{K,t}$ in this context is

$$i_\nu(H_{K,t}) = -\log \mathbb{E}\exp(-t\nu_K) = -\log(1-q+q2^{-t})^N = Ng_\nu(t). \tag{62}$$

For any $\theta \in [0,1]$ and any $\epsilon > 0$, let

$$J_N(\theta,\epsilon) := \{n \in J_N : |n/N - \theta| < \epsilon\} \quad \text{and} \quad \bar{J}_N(\theta,\epsilon) := J_N \setminus J_N(\theta,\epsilon). \tag{63}$$

Now $P^K_{\nu|E}(\{n\}) = \mathbb{P}(E_K)^{-1}\,\mathbb{P}\,(\nu_K = n)\,\mathbb{P}\,(\mu_K \leq n)$ and

$$N^{-1}\min\{\mathbb{P}(\nu_K \leq n),\, \mathbb{P}(\nu_K \geq n)\} \leq \mathbb{P}(\nu_K = n) \leq \min\{\mathbb{P}(\nu_K \leq n),\, \mathbb{P}(\nu_K \geq n)\},$$

and so it follows from Lemma 2.1(iii) that, for any $\theta \in [0,1]$,

$$\lim_K N^{-1}\log\left(P^K_{\nu|E}(\{[N\theta]\})\,\mathbb{P}\,(E_K)\right) = -f_\nu(\theta) - f_\mu(\theta,R)\mathbf{1}_{[0,\theta_{\mathrm{GV}})}(\theta).$$

Suppose that $R \in (0, C]$, then $\theta^*(R)$ is the unique maximizer of the right-hand side here and, since the latter is strictly convex,

$$\lim_K N^{-1} \log \left(\frac{P^K_{\nu|E}(\bar{J}_N(\theta^*, \epsilon))}{P^K_{\nu|E}(J_N(\theta^*, \epsilon))} \right) < 0 \quad \text{for any } \epsilon > 0. \tag{64}$$

Let $\hat{P}_{K,\epsilon}$ be the following approximation to $P^K_{\nu|E}$:

$$\begin{aligned} \hat{P}_{K,\epsilon}(\{n\}) := P^K_{\nu|E}(\{n\})/P^K_{\nu|E}(J_N(\theta^*, \epsilon)) \quad & \text{if } n \in J_N(\theta^*, \epsilon) \\ 0 \quad & \text{otherwise;} \end{aligned}$$

then

$$t(\theta^* - \epsilon) \le \liminf_K N^{-1} \langle H_{K,t}\,,\, \hat{P}_{K,\epsilon} \rangle \le \limsup_K N^{-1} \langle H_{K,t}\,,\, \hat{P}_{K,\epsilon} \rangle \le t(\theta^* + \epsilon),$$

and, because of (64),

$$\lim_K N^{-1} \left(\langle H_{K,t}\,,\, P^K_{\nu|E} \rangle - \langle H_{K,t}\,,\, \hat{P}_{K,\epsilon} \rangle \right) = 0.$$

Since these expressions are true for all $\epsilon > 0$,

$$\lim_K N^{-1} \langle H_{K,t}\,,\, P^K_{\nu|E} \rangle = t\theta^* \quad \text{for all } R \in (0, C]. \tag{65}$$

A similar argument shows that

$$\lim_K N^{-1} \langle H_{K,t}\,,\, P^K_{\mu|E} \rangle = t\theta^* \quad \text{for all } R \in (0, C]. \tag{66}$$

It follows from (10), (62) and (65) that

$$\begin{aligned} \liminf_K N^{-1} h(P^K_{\nu|E} \,|\, P^K_\nu) &\ge \sup_t \lim_K N^{-1} \left(i_\nu(H_{K,t}) - \langle H_{K,t}\,,\, P^K_{\nu|E} \rangle \right) \\ &= \sup_t (g_\nu(t) - t\theta^*) \\ &= f_\nu(\theta^*) \quad \text{for all } R \in (0, C]. \end{aligned} \tag{67}$$

Consider next the Bayesian problem of estimating μ_K given the observation that E_K has occurred. The full information of $H_{K,t}$ in this context is

$$i_\mu(H_{K,t}) = -\log \mathbb{E} \exp(-t\mu_K).$$

Now

$$\begin{aligned} \mathbb{E} \exp(-t\mu_K) &\le \sum_{n \in J_N} 2^{-tn}\, \mathbb{P}\,(\mu_K \le n) \\ &= \sum_{n \in J_N} 2^{-tn}\, \mathbb{P}\,(\cup_{v \in A_K} \{\omega \in \Omega :\, \rho_{K,N}(\omega, v) \le n\}) \\ &< \sum_{n \in J_N} 2^{K-tn}\, \mathbb{P}\,(\rho_{K,N}(\omega, v) \le n) \quad \text{for any } v \in A_K \\ &= 2^{K-N}(1 + 2^{-t})^N, \end{aligned}$$

and so

$$\liminf_K N^{-1} i_\mu(H_{K,t}) \geq 1 - R - \log(1 + 2^{-t}) = g_\mu(t, R). \tag{68}$$

It follows from (10), (66) and (68) that

$$\begin{aligned} \liminf_K N^{-1} h(P^K_{\mu|E} \,|\, P^K_\mu) &\geq \sup_t \liminf_K N^{-1} \left(i_\mu(H_{K,t}) - \langle H_{K,t} \,,\, P^K_{\mu|E} \rangle \right) \\ &\geq \sup_t (g_\mu(t, R) - t\theta^*) \\ &= f_\mu(\theta^*, R) \quad \text{for all } R \in (0, C]. \end{aligned} \tag{69}$$

Combining the lower bounds (67) and (69) with the upper bound (60) and (57) proves part (iii) and (38) for the case $R < C$.

Similar arguments to those used to prove (67) provide the following lower bound

$$\liminf_K N^{-1} h(P^K_{\nu|\bar{E}} \,|\, P^K_\nu) \geq f_\nu(\theta_{\mathrm{GV}}, q) \mathbf{1}_{[C,1]}(R), \tag{70}$$

and this, together with (61) and (58) proves part (iv) and (40) for the case $R \in (C, 1]$.

□

A.3. Proof of Theorem 3.1. We identify dominant terms in the sum in (46) when it is expressed in the following form:

$$i_{K,N}(\omega) = -\log \left(\alpha_K(\omega) + \sum_{n \in J_N} \beta_K(\omega, n) \right).$$

Here,

$$\begin{aligned} \alpha_K(\omega) &:= q^{\rho_{K,N}(\omega, T_K U)} (1 - q)^{N - \rho_{K,N}(\omega, T_K U)} \, 2^{-K}, \\ \beta_K(\omega, n) &:= M_K(\omega, n) q^n (1 - q)^{N-n} \, 2^{-K}, \\ M_K(\omega, n) &:= \mathrm{card}\{v \in A_K : \rho_{K,N}(\omega, v) = n\}, \end{aligned} \tag{71}$$

and A_K is as defined in (27).

The limiting behaviour of α_K is given by the strong law of large numbers:

$$\lim_K N^{-1} \log \alpha_K = -b(q) - R \quad \text{a.s.} \tag{72}$$

In order to find the limiting behaviour of $\sum \beta_K$, we first find an upper bound. To make this bound uniform in n we set up a sequence that "scans" the values $n = 0, 1, \ldots, N$ between each incrementation of K; for any $l \in \mathbb{N}$, let

$$\begin{aligned} K_l &:= \max \left\{ K \in \mathbb{N} : \sum_{k=1+[R]}^{K-1} ([R^{-1}k] + 1) < l \right\}, \\ N_l &:= [R^{-1} K_l], \\ n_l &:= l - 1 - \sum_{k=1+[R]}^{K_l - 1} ([R^{-1}k] + 1), \\ \gamma_l(\omega) &:= \beta_{K_l}(\omega, n_l), \end{aligned}$$

where a sum with a void index set is, by definition, zero. Now

$$\begin{aligned}\mathbb{E}M_K(\,\cdot\,,n) &= \sum_{v\in A_K}\mathbb{P}(\rho_{K,N}(\,\cdot\,,v)=n)\\ &= (2^K-1)\left(\begin{array}{c}N\\ n\end{array}\right)2^{-N},\end{aligned}\tag{73}$$

and so, according to Lemma A.1,

$$\lim_l\left(N_l^{-1}\log\mathbb{E}\gamma_l+1+h(Q_{\theta_l}\,|\,Q_q)\right)=0,$$

where $\theta_l := n_l/N_l$. So, for any $\epsilon>0$,

$$\sum_{l=1}^{\infty}\mathbb{E}\gamma_l\exp\left(N_l(1+h(Q_{\theta_l}\,|\,Q_q)-\epsilon)\right)<\infty.$$

It thus follows from the moment form of the first Borel-Cantelli lemma that

$$\gamma_l\exp\left(N_l(1+h(Q_{\theta_l}\,|\,Q_q)-\epsilon)\right)\to 0\quad\text{a.s. ,}$$

and so

$$\limsup_l\left(N_l^{-1}\log\gamma_l+h(Q_{\theta_l}\,|\,Q_q)\right)<\epsilon-1\quad\text{a.s.}$$

Since this is true for all $\epsilon>0$

$$\limsup_l\left(N_l^{-1}\log\gamma_l+h(Q_{\theta_l}\,|\,Q_q)\right)\le -1\quad\text{a.s. ,}$$

and so

$$\limsup_K\max_{n\in J_N}\left\{N^{-1}\log\beta_K(\,\cdot\,,n)+h(Q_{n/N}\,|\,Q_q)\right\}\le -1\quad\text{a.s. ,}\tag{74}$$

and, since the relative entropy term here is non-negative,

$$\limsup_K N^{-1}\log\max_{n\in J_N}\beta_K(\,\cdot\,,n)\le -1\quad\text{a.s.}\tag{75}$$

It follows from the continuity and strict convexity of f_ν that, for any $\epsilon>0$, there exists a $\delta=\delta_\epsilon>0$ such that

$$\inf_K\inf_{n\in\bar J_N(q,\epsilon)}h(Q_{n/N}\,|\,Q_q)>h(Q_{q+\delta}\,|\,Q_q),$$

where $\bar J_N$ is as defined in (63). It thus follows from (74) that

$$\limsup_K\max_{n\in\bar J_N(q,\epsilon)}N^{-1}\log\beta_K(\,\cdot\,,n)<-1-h(Q_{q+\delta}\,|\,Q_q)\quad\text{a.s.}\tag{76}$$

Next we refine (75) when n is restricted to the set $\Theta_{N,\epsilon} := J_N(1/2, 1/2 - \theta_{\rm GV} - \epsilon)$, where $0 < \epsilon < 1/2 - \theta_{\rm GV}$. If $v, \tilde{v} \in A_K$ and $v \neq \tilde{v}$ then $\rho_{K,N}(\,\cdot\,, v)$ and $\rho_{K,N}(\,\cdot\,, \tilde{v})$ are iid, and so

$$\begin{aligned}\mathbb{E}M_K(\,\cdot\,, n)^2 &= \sum_{v,\tilde{v}\in A_K} \mathbb{P}\big(\rho_{K,N}(\,\cdot\,, v) = \rho_{K,N}(\,\cdot\,, \tilde{v}) = n\big) \\ &\leq \mathbb{E}M_K(\,\cdot\,, n) + (\mathbb{E}M_K(\,\cdot\,, n))^2,\end{aligned}$$

and

$$\mathrm{var}\{\beta_K(\,\cdot\,, n)/\,\mathbb{E}\,\beta_K(\,\cdot\,, n)\} \leq \left((2^K - 1)\begin{pmatrix} N \\ n \end{pmatrix}\right)^{-1} 2^N.$$

Lemma A.1 thus shows that

$$\limsup_l N_l^{-1} \log\Big(\mathbf{1}_{\Theta_{N_l,\epsilon}}(n_l)\mathrm{var}\{\gamma_l/\,\mathbb{E}\,\gamma_l\}\Big) \leq 1 - R - b(\theta_{\rm GV} + \epsilon) < 0,$$

and so

$$\sum_{l=1}^{\infty} \mathbf{1}_{\Theta_{N_l,\epsilon}}(n_l)\mathrm{var}\{\gamma_l/\,\mathbb{E}\,\gamma_l\} < \infty,$$

and, according to the moment form of the first Borel-Cantelli lemma,

$$\max_{n\in\Theta_{N,\epsilon}} \left(\beta_K(\,\cdot\,, n)/\,\mathbb{E}\,\beta_K(\,\cdot\,, n) - 1\right)^2 \to 0 \quad \text{a.s.} \tag{77}$$

This, together with Lemma A.1 and (73), shows that, for any $\theta \in (\theta_{\rm GV}, 1 - \theta_{\rm GV})$,

$$\lim_K N^{-1} \log \beta_K(\,\cdot\,, [N\theta]) = -1 - h(Q_\theta \,|\, Q_q) \quad \text{a.s.} \tag{78}$$

Now $\alpha_K + \Sigma_n \beta_K(\,\cdot\,, n)$ admits the following bounds:

$$\begin{aligned}\alpha_K \mathbf{1}_{(0,C]}(R) + \beta_K(\,\cdot\,, [Nq])\mathbf{1}_{(C,\infty)}(R) &\leq \alpha_K + \sum_{n\in J_N} \beta_K(\,\cdot\,, n) \\ &\leq (N+2)\max\left\{\alpha_K,\, \max_{n\in J_N} \beta_K(\,\cdot\,, n)\right\},\end{aligned}$$

and so, from (72), (75), (76), (78), and the continuity of b,

$$\lim_K N^{-1} i_{K,N} = (R + b(q))\mathbf{1}_{(0,C]}(R) + \mathbf{1}_{(C,\infty)}(R) \quad \text{a.s.}\,, \tag{79}$$

which proves (48).

For any $\epsilon > 0$, let

$$D_{K,\epsilon}(\omega) := \left\{v \in \mathbb{X}_K :\, q - \epsilon < N^{-1}\rho_{K,N}(\omega, v) < q + \epsilon\right\};$$

then

$$\Pi_{K,N}(\,\cdot\,)(D_{K,\epsilon}) = \frac{\alpha_K \mathbf{1}_{(q-\epsilon,q+\epsilon)}(N^{-1}\nu_K) + \sum_{n\in J_N(q,\epsilon)} \beta_K(\,\cdot\,, n)}{\alpha_K + \sum_{n\in J_N} \beta_K(\,\cdot\,, n)},$$

and it follows from the strong law of large numbers, (72), (75), (76) and (78) that

$$\lim_K \Pi_{K,N}(\,\cdot\,)(D_{K,\epsilon}) = 1 \quad \text{a.s.}$$

Since this is true for all $\epsilon > 0$,

$$\lim_K N^{-1}\langle H_{K,N}\,,\,\Pi_{K,N}\rangle = b(q) \quad \text{a.s.},$$

and (49) follows from parts (ii) and (iv) of Proposition 1.1. □

A.4. Proof of Proposition 3.1. For any $K \geq k$, let $B_{K,k} := \{v \in \mathbb{X}_K : v_k = U_k\}$; then

$$\Pi_{K,N}(\omega)(B_{K,k}) = \frac{\alpha_K(\omega) + \sum_{n\in J_N}\tilde{\beta}_K(\omega,n)}{\alpha_K(\omega) + \sum_{n\in J_N}\beta_K(\omega,n)},$$

where α_K and β_K are as defined in (71), and

$$\tilde{\beta}_K(\omega,n) = \mathrm{card}\{v \in A_K \cap B_{K,k} : \rho_{K,N}(\omega,v) = n\}q^n(1-q)^{N-n}2^{-K}.$$

Now

$$\begin{aligned}\mathbb{E}\tilde{\beta}_K(\,\cdot\,,n) &= \sum_{v\in A_K\cap B_{K,k}} \mathbb{P}(\rho_{K,N}(\,\cdot\,,v) = n)q^n(1-q)^{N-n}2^{-K}\\ &= (2^K-1)^{-1}(2^{K-1}-1)\,\mathbb{E}\,\beta_K(\,\cdot\,,n),\end{aligned}$$

and arguments identical to those used above show that (77) and (78) remain valid if $\tilde{\beta}$ is substituted for β. It follows from these limits and (72) that, for any $0 < \epsilon < q - \theta_{\mathrm{GV}}$,

$$\begin{aligned}\lim_K \Pi_{K,N}(\omega)(B_{K,k}) &= \lim_K \frac{\sum_{n\in J_N(q,\epsilon)}\tilde{\beta}_K(\omega,n)}{\sum_{n\in J_N(q,\epsilon)}\beta_K(\omega,n)}\\ &= \lim_K \frac{\sum_{n\in J_N(q,\epsilon)}\mathbb{E}\tilde{\beta}_K(\,\cdot\,,n)}{\sum_{n\in J_N(q,\epsilon)}\mathbb{E}\beta_K(\,\cdot\,,n)}\\ &= 1/2,\end{aligned}$$

and this completes the proof. □

REFERENCES

[1] J. M. C. Clark, *The design of robust approximations to the stochastic differential equations of nonlinear filtering*, in: Communication Systems and Random Process Theory, J. K. Skwirzynski (ed.), NATO Advanced Study Institute Series, Sijthoff and Noordhoff, Alphen aan den Rijn, 1978, pp. 721–734.

[2] J. M. C. Clark and D. Crisan, *On a robust version of the integral representation formula of nonlinear filtering*, Probability Theory and Related Fields, 133 (2005), pp. 43–56.

[3] R. L. Dobrushin, *The description of a random field by means of conditional probabilities, and conditions of its regularity*, Theory Probab. Appl., 13 (1968), pp. 197–224.

[4] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, Wiley, New York, 1997.

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

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

[7] R. G. Gallager, *Information Theory and Reliable Communication*, Wiley, 1968.

[8] G. Gallavotti, *Statistical Mechanics—A Short Treatise*, Springer-Verlag, 1999.

[9] H-O. Georgii, *Gibbs Measures and Phase Transitions*, de Gruyter, 1988.

[10] E. T. Jaynes, *Probability Theory: the Logic of Science*, Cambridge University Press, 2003.

[11] Y. Kabashima and D. Saad, *Statistical mechanics of low density parity check codes*, J. Phys. A, 37 (2004), pp. R1–R43.

[12] J. C. Maxwell, *Theory of Heat*, Longmans, London, 1871.

[13] N. Merhav, *Relations between random coding exponents and statistical physics of random codes*, IEEE Trans. Information Th. 55 (2009), pp. 83–92.

[14] S. K. Mitter and N. J. Newton, *A Variational approach to Nonlinear Estimation*, SIAM J. Control Optim., 42 (2003), pp. 1813–1833.

[15] N. J. Newton and S. K. Mitter, *Variational Bayes and a problem of reliable communication II: infinite systems*, submitted for publication.

[16] A. Montanari and N. Sourlas, *The statistical mechanics of turbo codes*, European Phys. J. B, 18 (2000), pp. 107–119.

[17] O. Penrose, *Foundations of Statistical Mechanics*, Pergamon, Oxford, 1970.

[18] T. Richardson and R. Urbanke, *Modern Coding Theory*, Cambridge University Press, 2008.

[19] D. Ruelle, *A variational formulation of equilibrium statistical mechanics and the Gibbs phase rule*, Commun. Math. Physics, 5 (1967), pp. 324–329.

[20] C. E. Shannon, *A mathematical theory of communication*, Bell System Technical Journal, 27 (1948), pp. 379–423 and 623–656.

[21] N. Sourlas, *Spin glass models as error correcting codes*, Nature, 339 (1989), pp. 693–695.

[22] S. Verdú, *Fifty years of Shannon theory*, IEEE Trans. Information Theory, 44 (1998), pp. 2057–2078.

[23] S. Zabell, *Continuous versions of regular conditional distributions*, Annals of Probab., 7 (1979), pp. 159–165.