Some large linear systems in communications: estimation, information and statistical physics

Dongning Guo
Department of Electrical Engineering and Computer Science, Northwestern University
2145 Sheridan Road, Evanston, IL 60208, USA.
E-mail: dGuo@northwestern.edu

Abstract. Linear systems are frequently used to model physical channels in communication problems. In such systems, the goal is to recover the information bearing input signal to the linear transformation based on noisy observations of the output. This paper investigates the performance of linear systems randomly selected from some ensembles in the large-dimension limit. Previous analysis of such systems using statistical physics techniques is strengthened, where it is shown that the posterior of each individual input symbol is asymptotically identical to the posterior of a Gaussian channel with the same input. Interestingly, the replica symmetry solution takes an identical form as the rigorous result obtained for an ensemble of sparse linear systems. Some challenges to the information theory and statistical physics communities are highlighted.

1. Introduction
Linear systems are ubiquitous in communications for two main reasons: First, the input and output of a physical channel with electromagnetic waveform signaling exhibit a linear relationship in most cases of interest; secondly, the signal processing units are often linear by design for feasibility and tractability, even though linear processing is not always optimal.

A linear system in noise can be described as the following. Let $X = [X_1, \ldots, X_K]^T$ denote the input vector. Let $A = \text{diag}(A_1, \ldots, A_K)$ contain the amplitudes and the $L \times K$ matrix $S$ describe the linear transformation. The output of the transformation, $SAX$, is observed through arbitrary parallel independent noisy channels of the same type:

$$p_{Y|X,S} = \prod_{l=1}^{L} f(y_l|s Ax_l)$$

where $f(y|w) = p_{Y|W}(y|w)$ denotes the conditional probability density function of each scalar channel, and $(\cdot)_l$ denotes the $l$-th element of a vector. A familiar special case of the system with additive Gaussian noise is described by

$$Y = SAX + N$$

This paper adopts the following notational convention unless noted otherwise: Deterministic and random variables are denoted by lowercase and uppercase letters respectively; scalars, vectors and matrices are distinguished using normal, bold and underlined bold fonts respectively.
where \( \mathbf{N} \) is a Gaussian vector.

Communication systems described by the above linear models typically fall into one of the following three categories as well as their compositions:

(i) **Vector channels** in one symbol interval. Examples include code-division multiple access (CDMA) and multiantenna systems, which is often referred to as multiple-input multiple-output (MIMO) systems. In particular, (1)–(2) describe a CDMA systems where \( \mathbf{A}_k \) and \( \{\mathbf{S}_{lk}\} \) represent the amplitude and spreading sequence of user \( k \) respectively. The same model applies to MIMO systems, where \( \mathbf{A} = \mathbf{I} \) and \( \mathbf{S}_l \) denotes the gain between the \( k \)-th transmit antenna and the \( l \)-th receive antenna.

(ii) **Dispersive channels** in a block of \( L \) symbol intervals. Examples include inter-symbol interference (ISI) channels and multipath channels. In particular, the linear model (2) succinctly characterizes the following channel:

\[
Y_l = \sum_{i=0}^{m-1} h_i X_{l-i} + N_l
\]

where \([h_0, \ldots, h_{m-1}]\) represents the finite impulse response. Correspondingly, \( \mathbf{A} = \mathbf{I} \) in (2) and \( \mathbf{S}_l = h_{l-k} \) if \( 0 \leq l - k < m \) and \( \mathbf{S}_l = 0 \) otherwise.

(iii) **Multiterminal networks** in one symbol interval, where \((X_i, Y_i)\) represents the input–output pair of the \( i \)-th terminal. While the multi-hop aspect of networks is not described explicitly in (1), the model does encompass the well-known multiaccess, broadcast, relay, and interference channels.

The focus of this paper is on the estimation and information theoretic aspects of the vector channels in category (i). As we shall see, such channels have been completely and analytically characterized in the large-system limit in terms of simple scalar Gaussian channels. In other words, a single-letter characterization is known for category (i). On the contrary, a satisfactory characterization of the channels in categories (ii) and (iii) turns out to be very challenging and remains widely open.

The goal of communication through noisy channels is in general to recover the information bits embedded in the transmitted signal based on some noisy observation. Consider the models (1) and (2). For simplicity, we assume the symbols \( \{X_k\} \) to be independent and identically distributed (i.i.d.) and to take values in the alphabet \( \chi \subset \mathbb{R} \), which may be discrete or continuous. Let \( P_X \) denote the cumulative distribution function (cdf) of \( X_k \), which is of zero mean and finite variance.\(^2\)

Given the linear channel \( \mathbf{S}, \mathbf{A} \), the input statistics \( P_X \), and the channel characteristics \( f(\cdot|\cdot) \), the task of a detector is to produce an estimate of each \( X_k \) as a function of the observation. Oftentimes, the detector is followed by a decoder for error-control codes, so that a soft estimate of the symbols is most desirable. A generic optimal detector produces the posterior distribution of \( X_k \) given the output \( Y \) and the channel matrix \( \mathbf{S} \mathbf{A} \), i.e., 
\( \{P_{X_k|Y, \mathbf{S} \mathbf{A}}(x|Y, \mathbf{S} \mathbf{A})\} \), which is a sufficient statistic for \( X_k \) and upon which any deterministic decision can be derived. For example, the minimum mean-square error (MMSE) estimator computes the conditional mean \( \mathbf{E}\{X_k|Y, \mathbf{S} \mathbf{A}\} \) and the maximum a posteriori (MAP) detector outputs \( \hat{X}_k = \arg\max_{x \in \chi} P\{X_k = x|Y, \mathbf{S} \mathbf{A}\} \) in case of discrete alphabet \( \chi \).

If the alphabet \( \chi \) is discrete (resp. continuous), the soft detection output can be represented as a probability mass (resp. density) function. In practice, the alphabet is often very simple and the soft estimate admits many equivalent forms. For example, if \( \chi = \{+1, -1\} \), the optimal soft output is simply the a posteriori probability of \( X_k = +1 \) (or \(-1\)) and an equivalent form

\(^2\) The analysis can in principle be generalized to heterogeneous distributions of the independent symbols \( X_k \) or even some joint distribution \( P_X \).
of the soft output is the log-likelihood ratio (LLR) of the \textit{a posteriori} probability of the two hypotheses.

For the convenience of establishing the scalar channel characterization of the vector channels, we introduce a new concept called the \textit{equivalent statistic} in section 2. The equivalence between different estimation problems is then defined.

In section 3, we provide the key results obtained using the \textit{replica method}, which was originally developed in statistical physics. Previous characterization of the (random) vector channels in the large-system limit is strengthened, where it is shown that the posterior of individual input symbols is asymptotically identical to the posterior of a Gaussian channel with the same input. The replica analysis is not known to be rigorous and a complete justification of the result remains open.

In contrast, section 4 describes rigorous results obtained in [1] for similar vector channels where the linear transformation is chosen randomly from an ensemble of asymptotically sparse systems. Surprisingly, the result takes the same form as those obtained in section 3.

In section 5, we point out some obstacles which so far have prevented a precise characterization of the vector channels in general. We also discuss the challenges in the dealings with the dispersive channels and the multiterminal networks. Section 6 concludes the paper.

2. Equivalent statistics and equivalent channels

All that is needed for estimating each symbol $X_k$ using the observation $Y$ is evidently included in the posterior $P_{X_k|Y,SA}(\cdot|Y,SA)$, which is a distribution on $\chi$ and generally takes a very complicated structure. The goal here is to provide a much simpler statistic of the same quality for $X_k$. For that purpose, we introduce several equivalence relations in the following. Throughout this paper, let $P_{X|Y}(\cdot|Y)$ denote a random posterior cdf on $\chi$ dependent on $Y$ for technical convenience.

\textbf{Definition 1} Consider any $(X, Y, Z)$ with a joint distribution. We call $Z$ an equivalent statistic of $Y$ for $X$ if $P_{X|Z}(\cdot|Z) = P_{X|Y}(\cdot|Y)$, i.e., $Y$ and $Z$ give identical posterior of $X$.

If, as a function of $Z$, $Y = f(Z)$ is an equivalent statistic of $Z$ for $X$, then $Y$ is a \textit{sufficient statistic} of $Z$ for $X$, i.e., $X$ is conditionally independent of $Z$ given $Y$. For example, let $W, X, Y$ be Bernoulli random variables where $W$ is independent of $X$. Let $Z = Y + 2W$. Then $Y$ and $Z$ are equivalent for $X$ and in fact $Y$ is a sufficient statistic of $Z$.

In the situation where the statistic $Z$ is not equivalent to $Y$ in the sense of Definition 1, but $Z$ and $Y$ are of the same quality statistically, we define the following weaker equivalence.

\textbf{Definition 2} We call $Z$ a statistically equivalent statistic of $Y$ for $X$, or we say that $Z$ and $Y$ are statistically equivalent for $X$, if $P_{X|Z}(\cdot|Z)$ and $P_{X|Y}(\cdot|Y)$ are identically distributed conditioned on $X$.

For example, let $X, U, V$ be i.i.d. Bernoulli random variables and let $Y = X + U$ and $Z = X + V$. Then $Y$ and $Z$ are statistically equivalent which should be evident by symmetry. But $Y$ and $Z$ are in general not equivalent, which can be shown by considering a particular realization of $(X, U, V)$ as $(1, 1, 0)$, in which case $X = 1$ can be inferred from $Y = 2$ but not from $Z = 1$.

Consider the channel $X \mapsto Y$ characterized by the conditional distribution $P_{Y|X}$. To an estimator which only has access to $Y$, a sufficient statistic is $P_{X|Y}(\cdot|Y)$, which is the posterior distribution as a function of $Y$. Suppose for every realization $X = x \in \chi$, the posterior $P_{X|Y}(\cdot|Y)$ and $P_{X|Z}(\cdot|Z)$ are identically distributed (the distribution of the two distributions are identical), where $Z$ is another observable, then the two posteriors are not distinguishable to any third party with access to $(Y, Z)$ but no access to $X$. 


Definition 3. We say $Y_K$ is asymptotically equivalent to $Z$ for $X$ if, conditioned on $X$,

$$P_{X|Y_K}(\cdot|Y_K) \rightarrow P_{X|Z}(\cdot|Z) \text{ in probability}$$

as $K \rightarrow \infty$. Alternatively, if the convergence in (4) holds in distribution instead of in probability, $Y_K$ is said to be asymptotically statistically equivalent to $Z$. The subscript $K$ is often dropped when the index is clear from the context.

The notion of (statistical) equivalence is used to study the performance of the vector channels in sections 3 and 4. In particular, we find a scalar asymptotically (statistically) equivalent statistic of $(Y, S A)$ for each individual symbol $X_k$. In fact, the statistic is as simple as $X_k$ itself corrupted by Gaussian noise.

3. Replica analysis of large random linear systems

The performance of vector channels with optimal estimation is a hard problem. Any estimation error measure is determined by the posterior $P_{X|Y}$. Although $P_X$ and $P_{Y|X, SA}$ show very simple structure, the posterior $P_{X_k|Y}$ for each $k$ is highly complicated in general because $P_Y$ is a mixture of Gaussian distributions. With an additional constraint of linear detection, however, the estimate of the input becomes simple and the performance of the Gaussian vector channel (2) is tractable (see e.g., [2]).

The first successful treatment of a non-trivial vector channel with non-linear detection at arbitrary signal-to-noise ratio (SNR) is due to Tanaka [3], who applied the replica method to obtain the optimal bit-error-rate of CDMA in the large-system limit. The result was generalized in [4] where it is shown that the vector channel with generic posterior mean estimation can be decoupled into a bank of scalar Gaussian channels, one for each input symbol.

This section first describes the ensemble of vector channels studied in [3–5]. We then strengthen the previous results, where it is shown that the posterior of the vector channel can in fact be decoupled into the posterior of scalar channels. The detailed analysis follows [4] and is omitted.

3.1. Large random linear systems

In problems such as MIMO and CDMA, the linear transformation is often random by nature or by design. A typical case of interest is where $\sqrt{L} \cdot S$ is randomly selected from an ensemble in which the entries $S_{lk}$ are i.i.d. with zero mean and unit variance. The amplitudes $\{A_k\}$ are assumed to be i.i.d. with distribution $P_A$. Note that the amplitudes $A_k$ may be different for different symbols; thus fading is inherently considered in the model.

Of interest is the average performance of the vector channel over the ensemble. A common approach for the analysis is to investigate the average performance in the large-system limit where $K, L \rightarrow \infty$ with their ratio converging to a positive number $K/L \rightarrow \beta$, known as the system load. The input distribution $P_X$, the amplitude distribution $P_A$, and the system load $\beta$ remain fixed. The distribution of the independent entries $S_{lk}$ turns out to be irrelevant as long as their fourth order moment is finite.

3.2. The decoupling principle

Let us introduce the canonical scalar Gaussian channel:

$$Z = \sqrt{g}X + N$$

where $X \sim P_X$ and $g$ denotes the gain of the channel in SNR. Throughout this paper, $N \sim \mathcal{N}(0, 1)$ denotes a standard Gaussian random variable, and $P_{X|Z,g}(\cdot|z;g)$ denotes the specific
cdf of the posterior distribution of the input $X$ given $Z = z$ according to the Gaussian model (5), which is parameterized by $g$. Let

$$\text{var} \{U | V\} = E \left\{ (U - E \{U | V\})^2 \right\}$$

(6)

stand for the average conditional variance (or equivalently the MMSE) of estimating a random variable $U$ given observation $V$, which may contain multiple elements.

**Proposition 1** Assume replica symmetry holds. For every $k = 1, 2, \ldots$, the statistic $(Y, \mathbf{SA})$ is asymptotically statistically equivalent to $Z_k = \sqrt{\eta} A_k X_k + N$ for $X_k$ in the large-system limit, where the efficiency $\eta$ satisfies the following fixed-point equation

$$\eta^{-1} = 1 + \beta \text{var} \{AX|\sqrt{\eta}AX + N, A\}.$$  

(7)

In other words, For every $x \in \chi$ and $k = 1, 2, \ldots$, conditioned on $X_k = x$ and $A_k = a$, the following (random) posterior cdf converges

$$P_{X_k|Y, \mathbf{SA}}(\cdot | Y, \mathbf{SA}) \rightarrow P_{X|Z,g}(\cdot | \sqrt{\eta}ax + N; a^2\eta)$$

(8)

in distribution as $K, L \to \infty$ with $K/L \rightarrow \beta$. In case (7) has more than one solution, $\eta$ is chosen to minimize

$$I(X; \sqrt{\eta}AX + N | A) + \frac{1}{2\beta} (\eta - 1 - \log \eta).$$

(9)

The conditional mutual information in (9) is obtained as an average over the distribution $P_A$:

$$I(X; \sqrt{\eta}AX + N | A) = \int_{0}^{\infty} I(X; \sqrt{\eta}aX + N) dP_A(a).$$

(10)

The essence of Proposition 1 is the following characterization of the linear system: From the viewpoint of an individual symbol, the posterior of the symbol given the output vector is increasingly similar to that under a simple scalar setting described by (5) as the system becomes large. Indeed, in the large-system limit, the vector channel can be decoupled into a bank of scalar Gaussian channels with the same parameter $\eta$. Note that the statistic $\sqrt{\eta} A_k X_k + N$ may not be computed from $(Y, \mathbf{SA})$; rather, the statistics are asymptotically of the same quality.

The parameter $\eta$ is interpreted as the loss of efficiency due to multiaccess interference because it would be equal to 1 in absence of interfering symbols. It is important to note that the efficiency $\eta$ does not depend on any specific amplitudes in the large system; rather, it depends only on the distribution $P_A$.

Proposition 1 implies that any estimate for $X_k$ computed from $(Y, \mathbf{SA})$ has an equivalent counterpart as a function of $Z_k$. Consider the optimal estimate in mean-square sense, which is the conditional mean, denoted by $\langle X_k \rangle = E \{X_k | Y, \mathbf{SA}\}$. Proposition 1 strengthens the following result which was obtained in [4].

**Corollary 1** The conditional mean estimate $\langle X_k \rangle$, which is a function of $(Y, \mathbf{SA})$, is asymptotically identically distributed as $E \{X_k | \sqrt{\eta} A_k X_k + N, A_k\}$ conditioned on $(X_k, A_k)$, where $\eta$ satisfies (7). Indeed, the distribution of the output $(X_k)$ conditioned on $X_k = x \in \chi$ being transmitted with amplitude $A_k = a$ converges to the distribution of $(X) = E \{X | \sqrt{\eta} X + N\}$ conditioned on $X = x$ being transmitted with $g = a^2\eta$,

$$P_{(X_k)|X_k,A_k}(\cdot|x,a) \rightarrow P_{(X)|X,g}(\cdot|x,a^2\eta)$$

(11)

for all $a$ and all $x \in \chi$ except for the points of discontinuity of $P_X$.

3 Note that the observation $(Y, \mathbf{SA})$ is implicitly indexed by the system size $(K, L)$, which goes to infinity in the large-system limit.

4 Throughout this paper, the unit of information is nats and all logarithms are natural.
Note that the fixed-point equation (7) obtained using the replica method has at least one solution for $\eta$. The equation may have multiple solutions, depending on $P_X$, $P_A$ and $\beta$. Among those solutions, it is believed that the (thermodynamically) dominant solution gives the smallest value of (9), which is essentially $I(X; Y|SA)$ as we shall see in Proposition 2. This is the solution that carries relevant operational meaning in the communication problem. Note also that for sufficiently small $\beta$, however, the slope of the right-hand side as a function of $\eta$ is negligible, so that the solution to (7) is unique, so that there is no phase transition.

Following Proposition 1, the mutual information between each symbol and the observation is easily characterized as the following.

**Corollary 2** For every $k$ and $a$, in the large-system limit,

$$I(X_k; Y|SA, A_k = a) \rightarrow I(X; \sqrt{\eta} aX + N).$$  

(12)

The input–output mutual information of the vector channel is given in the following result.

**Proposition 2** Under replica symmetry, the input–output mutual information per symbol of the vector channel converges in the large-system limit:

$$\frac{1}{K} I(X; Y|SA) \rightarrow I(X; \sqrt{\eta} AX + N|A) + \frac{\eta - 1 - \log \eta}{2\beta}.$$  

(13)

In comparison to the average of (12), the extra term is proportional to a Kullback-Leibler divergence between two Gaussian distributions [4], which is always positive.

Proposition 2 is an outcome of the chain rule of mutual information applied to the input–output mutual information of the vector channel

$$I(X; Y|SA) = \sum_{k=1}^{K} I(X_k; Y|SA, X_{k+1}, \ldots, X_K).$$  

(14)

Each mutual information in the sum is a single-symbol mutual information over the vector channel conditioned on the symbols of higher indexes. Using a successive-decoding argument, the posterior for symbol $k$ is identical to the posterior in a system with $k$ instead of $K$ symbols. The equivalent scalar channel for each symbol is Gaussian by Proposition 1. The efficiency of symbol $k$ is a function of the effective load $\beta_k = k/L$, and the mutual information converges to $I\left(X; \sqrt{\eta(\beta_k)} AX + N|A\right)$. In view of (14),

$$\frac{1}{K} I(X; Y|SA) = \frac{1}{K} \sum_{k=1}^{K} I\left(X_k; \sqrt{\eta(\beta_k)} AX + N|A\right)$$  

(15)

$$\rightarrow \frac{1}{\beta} \int_{0}^{\beta} I\left(X; \sqrt{\eta(\xi)} AX + N|A\right) d\xi$$  

(16)

as $K \rightarrow \infty$ where the last equation is by definition of the Riemann integral. Proposition 2 can be shown using an interesting relationship between the mutual information and the conditional variance in Gaussian channels [6, Theorem 1].

We remark that the replica symmetry solution can be understood as follows: If one insists on a simple scalar-channel characterization (i.e., single-letter characterization) of the large linear system, even if it is not precise, then the replica symmetry solution is the natural answer.

Finally, the replica symmetry solution obtained in section 3 is not known to be rigorous and in fact the symmetry can be broken under certain circumstances [5]. Numerical studies show that the replica symmetry solution is a good approximation for systems of moderate size with a wide range of parameters.
4. Rigorous analysis of a sparse linear system

This section considers an ensemble of sparse linear systems, where the density of nonzero entries in the transformation \( S \) vanishes as the system dimension becomes large. A set of rigorous results is obtained, where the key formulas turn out to be essentially identical to those given in section 3. The key to the rigorous analysis is that, under the sparsity constraint, the marginal posterior of each input symbol can be computed using belief propagation (BP), which leads to a converging iterative formula.

The sparsity constraint is in part inspired by capacity-approaching low-density parity-check (LDPC) codes and the success of iterative decoding techniques. In [7], Yoshida and Tanaka proposed a family of sparse CDMA and analyzed the performance using the replica method. Montanari and Tse proposed in [8] a different ensemble of sparse CDMA, and justified Tanaka’s formula [3] for the first time in special case without resorting to replicas. Some recent study of sparse CDMA are found in [9] due to Raymond and Saad and [1] due to Guo and Wang.

4.1. Sparse linear systems

Similar to described in section 3.1, we consider a sequence of ensembles indexed by the user number \( K \) (implicit in notation). For any fixed \((K, L)\), let \( S \) be randomly constructed as follows (see also [1]). First, a binary incidence matrix \( H_{L \times K} = (H_{lk}) \) is randomly picked from the so-called doubly Poisson ensemble to be described shortly. For all \((l, k)\) with \( H_{lk} = 0\), we set \( S_{lk} = 0\). For all \((l, k)\) with \( H_{lk} = 1\), \( S_{lk} \) are drawn i.i.d. with some distribution of zero mean and unit variance. The normalization factor for each column \( S_k \) is \( \sqrt{\Lambda_k} \), where \( \Lambda_k = \sum_{l=1}^{L} H_{lk} \) is the symbol (node) degree of \( X_k \), which is also the number of chips over which \( X_k \) is spread.

We define \( \bar{\Lambda} = \frac{1}{K} \sum_{k=1}^{K} \Lambda_k \) as the average symbol degree. Similarly, the chip (node) degree and the average chip degree are defined by \( \Gamma_l = \sum_{k=1}^{K} H_{lk} \) and \( \bar{\Gamma} = \frac{1}{L} \sum_{l=1}^{L} \Gamma_l \) respectively. Finally, the doubly Poisson ensemble for \( H \) is such that each entry \( H_{lk} \) is i.i.d. Bernoulli with \( P \{ H_{lk} = 1 \} = \bar{\Gamma} / K \). The results in this section can be extended to accommodate more general ensembles [10], as demonstrated in [11].

The model (2) remains valid except that the \( k \)-th column of \( S \) is now \( \frac{1}{\sqrt{\Lambda_k}} [S_{k1}, \ldots, S_{kL}]^\top \). Note that \( \{A_k\} \) are i.i.d. with distribution \( P_A \). A factor graph illustrated in figure 1 can be constructed to fully describe the probability law of the system, so that the estimation problem is equivalent to statistical inference on the graph.

\[ X_1 \quad X_2 \quad X_3 \quad \ldots \quad X_K \]
\[ Y_1 \quad Y_2 \quad Y_3 \quad Y_4 \quad \ldots \quad Y_{L-1} \quad Y_L \]

\( s_{11} \quad s_{L, K} \)

Figure 1. The Forney-style factor graph for the sparse linear system. Squares and circles correspond to the observations \( Y_l \) and symbols \( X_k \) respectively.

This section considers the large-sparse-system limit, in which \( K, L, \bar{\Gamma} \to \infty \) with \( K/L \to \beta < \infty \) and \( \bar{\Gamma} = o(K^{1/(4t)}) \), the last condition of which ensures that the bipartite graph is free of cycles of length smaller than \( t \) in probability. More detailed description of the ensemble of interest can be found in [11].
4.2. The asymptotic performance of BP

Belief propagation [12] is an iterative message-passing algorithm for computing the marginal posterior distributions on Bayesian inference networks. If applied to an inference problem with a factor graph free of cycles, BP produces the exact marginal posterior distribution and is thus optimal in every information-theoretic sense. BP is also frequently applied to inference networks with cycles and is known to produce good approximation of the posterior marginals in practice, even though its suboptimal performance is notoriously difficult to quantify.

In this work, the BP algorithm for general input (with distribution $P_{X}$) can be derived from the factor graph representation (refer to [1] for details). Consider the ensemble described in section 4.1 and the problem of inferring about any individual symbol $X_{k}$. For any number of iterations $t$, there is no cycle in the subgraph induced by all nodes within $2t - 1$ distance from $X_{k}$ with probability 1 in the large-system limit. After $t$ iterations, the output of BP is a posterior distribution for $X_{k}$ computed based on all observations at chip nodes within distance $2t - 1$ to $X_{k}$ on the factor graph, denoted by $Y_{k}^{(t)}$. At each intermediate node, the LLR from the children are aggregated and passed to the parent.

Since the node degree increases without bound in the large-system limit, the aggregate LLR converges to a Gaussian random variable due to central limit theorem. This allows the performance to be analytically determined using density evolution for any given ensemble. It can be shown that the quality of detection can be simply described by the variance of the likelihood ratio. Interestingly, in the viewpoint of each symbol, the vector channel combined with the BP detector is asymptotically equivalent to a scalar Gaussian channel, where the collective impact of interfering symbols is a degradation in the SNR of the desired symbol. The degradation, i.e., the efficiency, can be obtained from an iterative formula.

With slight abuse of notation, let $P_{X_{k}}(\cdot | Y_{k}^{(t)} , S)$ denote the cdf of the approximate posterior of $X_{k}$ given $Y_{k}^{(t)}$ and the channel state. For simplicity, we omit the adjective “approximate” as long as it is clear from the context that the output of the BP-based detection is referred to. A key result in [1] states that the posterior computed for each symbol using BP after $t$ iterations essentially converges to the posterior of a scalar Gaussian channel.

Theorem 1 ([1]) Fix the number of iterations $t$. For every $k$,

$$P_{X_{k}|Y_{k}^{(t)} , S A}(\cdot | Y_{k}^{(t)} , S A) \rightarrow P_{X|Z;g}(\cdot | h(Y_{k}^{(t)} , S A) ; \eta A_{k}^{2})$$

in probability in the large-sparse-system limit for some $\eta^{(t)} \in [0, 1]$ and some function $h(\cdot)$ such that $h(Y_{k}^{(t)} , S A) \sim N(\sqrt{\eta^{(t)} a x} , 1)$ conditioned on $X_{k} = x$ and $A_{k} = a$. Precisely, $\eta^{(0)} = 0$ and $\eta^{(t)} , t = 1, 2, \ldots$ are determined by the following recursion:

$$\frac{1}{\eta^{(t)}} = 1 + \beta \text{var} \left\{ AX \left| \sqrt{\eta^{(t-1)}} AX + N, A \right. \right\}. \tag{18}$$

Theorem 1 states that the problem of estimating each individual symbol $X_{k}$ using $t$ iterations of BP is asymptotically equivalent to that of estimating the same symbol through a scalar Gaussian channel with SNR equal to $\eta^{(t)} A_{k}^{2}$, e.g., with a degradation $\eta^{(t)}$ in the input SNR. The parameter $\eta^{(t)}$ is termed the efficiency of the BP detector after $t$ iterations. This type of information equivalence is referred to as the decoupling principle (see, e.g., [4,5]). The function $h(\cdot)$ finds essentially a Gaussian sufficient statistic for the inference problem through BP.

4.3. Optimal detection and its relation to BP

Theorem 2 ([1]) Suppose (7) has a unique fixed point $\eta$. Then for every $k$,

$$P_{X_{k}|Y , S A}(\cdot | Y , S A) \rightarrow P_{X|Z;g}(\cdot | h(Y , S A) ; A_{k}^{2} \eta) \tag{19}$$
in probability in the large-sparse-system limit, where \( h(\cdot) \) is such that \( h(Y) \sim \mathcal{N}(\sqrt{\eta} ax, 1) \) conditioned on \( X_k = x \) and \( A_k = a \).

Theorem 2 states that if (7) has a unique solution, then the problem of estimating each \( X_k \) given the entire observation \( Y \) is also asymptotically equivalent to estimating the same symbol through a scalar Gaussian channel. The efficiency is determined by fixed-point equation (7), which corresponds to iterative formula (18), and was originally obtained for the non-sparse ensemble in section 3 using the replica method. It can be shown that \( \lim_{t \to \infty} \eta^{(t)} = \eta \) in case (7) has a unique solution.

Theorem 2 can be established using the following sandwiching argument. Recall that BP is an optimal detection rule based on the limited observations \( Y^{(t)} \) on the subtree of depth \( 2t - 1 \) rooted at the each desired symbol \( X_k \). Define \( X^{(t)} \) as the symbols on the same subtree. Let us define a genie-aided BP (gBP) as the optimal detection based on \( Y^{(t)} \) where all entries of \( X \) not in \( X^{(t)} \) are revealed to the BP estimator by a genie. This effectively guarantees independence of the leaves on the subtree of depth \( 2t - 1 \). One can show that the a posteriori detector for symbol \( X_k \) based on \( Y \) is a physically degraded detection rule with respect to gBP while the classical BP represents further degradation. Since the performances of BP and gBP sandwich the performance of optimal detection, the latter will also face a scalar Gaussian channel in the large-system limit as described in Theorem 2. Moreover, by the similar argument as in the last paragraph, with probability one, the actual posterior distribution is identical to the same posterior distributions computed by BP and gBP in the large-sparse-system limit.

**Corollary 3 ([1])** If (7) has a unique fixed point, then for every \( k \) as \( t \to \infty \),

\[
\lim_{\bar{\Gamma} \to \infty} \left| \frac{P_{X_k|Y^{(t)}}}{S^A(x|Y^{(t)}, S^A)} - \frac{S^A(x|Y, S^A)}{P_{X_k|Y, S^A}} \right| \to 0, \quad \text{in probability.} \tag{20}
\]

Corollary 3 implies that essentially the same posterior about each symbol is obtained either using \( Y^{(t)} \) or using \( Y \) in large sparse linear systems as the number of iterations \( t \) becomes large.

Theorem 2 characterizes the optimal performance one can hope to achieve when the solution of (7) is unique. If the system load \( \beta \) is relatively large, the solution to (7) may not be unique. Let \( \eta_0 \) and \( \eta_1 \) denote the smallest and the largest fixed point of (7) respectively. The performance of BP is characterized by \( \eta_0 \) because iterative formula (18) leads to \( \eta_0 = 0 \). It is not known whether the decoupling principle still holds true. It can be shown that the quality of optimal detection is inferior than the posterior of the scalar Gaussian channel \( X \mapsto \sqrt{\eta_1} AX + N \).

5. Discussion

The replica result in section 3 suggests that the entire observation \( (Y, S^A) \) is asymptotically statistically equivalent to some conditional Gaussian variable given the input (in the posterior sense). The finding is fully consistent with the rigorous results in section 4 made available by introducing some sparsity constraint. In fact, for the sparse systems, BP computes a conditionally Gaussian asymptotically equivalent statistic, which is in fact asymptotically a sufficient statistic. If the system is not sparse, BP does not lead to an equivalent statistic and it is not known how to compensate for the loopy structure in the factor graph.

The major puzzle remains to be solved is either a rigorous justification of the replica method, with or without replica symmetry, or an alternative rigorous analysis of the large linear systems if the replica method is flawed. Moreover, the gap between the performance of a relatively small system and the large-system performance remains to be quantified.

Another important question is how the performance of sparse linear systems compares with that of dense systems. According to discussion in this section, the difference appears to be small,
which adds to the practical appeal of using sparse systems because they admit low-complexity near-optimal estimation. In fact, the iterative formula obtained in section 4 leads to exactly the fixed-point equation for the efficiency obtained using the non-rigorous replica method in section 3. The insignificance of the difference is also buttressed by numerical results (see, e.g., [4, 9]). A precise quantification of the difference between the i.i.d. ensemble and the sparse ensemble is not available, however, due to lack of a rigorous and manageable expression of the performance of optimal detection of the former ensemble.

The dispersive channels discussed in section 1 carries a structure which is not as easy to deal with as the i.i.d. ensemble. In fact, the linear transformation is essentially a circular matrix with a banded diagonal. The capacity of the dispersive channel with Gaussian noise is known if the input is Gaussian, but remains open for practical input constellations.

The multiterminal network model appears to be an even harder problem. The theory for general networks, important for applications such as mobile ad hoc networks, is far from mature. In fact, the capacity of all but the multiaccess channels is in general open.

6. Conclusion

This paper presents some strengthened results on large random linear systems obtained using the replica method. It is shown that the linear system can be essentially decoupled in the large-system limit. That is, in terms of the posterior for each individual input symbol, the vector channel is equivalent to a scalar Gaussian channel. The results are then compared with some recent rigorous results obtained for large sparse linear systems, which admits optimal estimation using belief propagation, and hence a simple scalar-channel characterization.

The observations made in this work is analogous to a comparison that can be made between random linear block codes and LDPC codes, which are both asymptotically optimal, while the latter admits low-complexity decoding using BP.

Acknowledgments

This research was supported in part by the NSF CAREER Award CCF-0644344 and DARPA IT-MANET program under Grant W911NF-07-1-0028.

References

[1] Guo D and Wang C C IEEE J. Select. Areas Commun., Special Issue on Multiuser Detection for Advanced Communication Systems and Networks Accepted
[2] Verdú S 1998 Multuser Detection (Cambridge University Press)
[3] Tanaka T 2002 IEEE Trans. Inform. Theory 48 2888–2910
[4] Guo D and Verdú S 2005 IEEE Trans. Inform. Theory 51 1982–2010
[5] Guo D and Tanaka T Advances in Multiuser Detection ed Honig M (Wiley) to be published
[6] Guo D, Shamai S and Verdú S 2005 IEEE Trans. Inform. Theory 51 1261–1282
[7] Yoshida M and Tanaka T 2006 Proc. IEEE Int. Symp. Inform. Theory (Seattle, WA, USA) pp 2378–2382
[8] Montanari A and Tse D 2006 Proc. IEEE Inform. Theory Workshop (Punta del Este, Uruguay) pp 122–126
[9] Raymond J and Saad D 2007 J. Phys. A: Math. Theor. 40 12315–12333
[10] Litsyn S and Shevelev V 2002 IEEE Trans. Inform. Theory 48 887–908
[11] Wang C C and Guo D 2006 Proc. 44th Annual Allerton Conference on Communication, Control, and Computing (Monticello, IL, USA) pp 926–935
[12] Pearl J 1988 Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (Morgan Kaufmann) revised 2nd printing