On Linear Model with Markov Signal Priors

Lan V. Truong
The University of Cambridge

Abstract

In this paper, we estimate free energy, average mutual information, and minimum mean square error (MMSE) of a linear model under the assumption that the source is generated by a Markov chain. Our estimates are based on the replica method in statistical physics. We show that under the MMSE estimator, the linear model with Markov sources is decoupled into single-input AWGN channels with state information available at both encoder and decoder where the state distribution follows the stationary distribution of the stochastic matrix of Markov chains. Numerical results show that the free energies and MSEs obtained via the replica method are closely approximate to their counterparts via MCMC simulations.

1 Introduction

In the canonical compressed sensing problem, the primary goal is to reconstruct an n-dimensional vector $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$ with independent and identical prior from an m-dimensional vector of noisy linear observations $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_m)$ of the form $Y_k = \langle \boldsymbol{\Phi}_k, \boldsymbol{X} \rangle + W_k, k = 1, 2, \cdots, m$, where $\{\boldsymbol{\Phi}_k\}$ is a sequence of n-dimensional measurement vectors, $\{W_k\}$ is a sequence of standard Gaussian random variables, and $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product between vectors. In this paper, under the assumption that $\boldsymbol{X}$ has a Markov or hidden Markov prior, we wish to estimate the asymptotic mutual information $\lim_{n\to\infty} \frac{1}{n} I(\boldsymbol{X}; \boldsymbol{Y})$ and the MMSE $\lim_{n\to\infty} \frac{1}{n} \mathbb{E}[\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{\Phi}]\|^2]$. Our estimates are based on the replica method which was developed originally to study mean field approximations in spin glasses (Edwards and Anderson, 1975). Although this method lacks of rigorous mathematical proof in some particular parts, it has been widely accepted as an analytic tool and utilized to investigate a variety of problems in statistics, information processing, and coding (Bereyhi et al., 2019).

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

1.1 Related Work

In recent years, there has been the progress on a coherent mathematical theory of the replica and interpolation method in statistical physics of spin glasses (Edwards and Anderson, 1975). These methods have been fruitfully extended and adapted to the problems of interest in a wide range of applications in Bayesian inferences, multiuser communications, and theoretical computer science (Tanaka, 2002; Guo et al., 2005). The results of replica method have been rigorously in a number of settings in compressed sensing. One example is given by message passing on matrices with special structure, such as sparsity (Guo and Wang, 2006; Montanari and Tse, 2006; Baron et al., 2010; Korada and Macris, 2010; Barbier et al., 2020) or spatial coupling (Kudekar and Pfister, 2010; Krzakala et al., 2012; Donoho et al., 2011). In (Rangan et al., 2012), Rangan et al. studied the asymptotic performance of a class of Maximize-A-Posterior (MAP) estimators. Using standard large deviation techniques, the authors represented the MAP estimator as the limit of an indexed MMSE estimator's sequence. Consequently, they determined the estimator's asymptotics employing the results from (Guo and Verdu, 2005) and justified the decoupling property of MAP estimators under Replica Symmetry (RS) assumption for an i.i.d. measurement matrix $\boldsymbol{\Phi}$. The asymptotic performance for the MAP estimator where the RS assumption does not hold but satisfies some looser symmetric assumptions, called Replica Symmetry Breaking (RSB) is considered in (Bereyhi et al., 2019). Under the RSB assumption with b steps of breaking (bRSB), the equivalent noisy single-user channel is given in form of an input term added by an impairment term. The impairment term, moreover, is expressed as a sum of an independent Gaussian random variable and b correlated non-Gaussian interference terms.

Recently, there have been some works which aim to close the gap between mathematically rigorous proof and results from the replica method. Reeves and Pfister considered the fundamental limit of compressed sensing for i.i.d. signal distributions and i.i.d. Gaussian measurement matrices (Reeves and Pfister, 2019). Under some mild technical conditions, their results show that the limiting mutual information and Minimum Mean Square Error (MMSE) are equal to the values predicted by the replica method. Their proof techniques are based on establishing relationships between mutual information and MMSE at finite n, m, and extending obtained results in large system limits. In (Barbier et al., 2019), Barbier et al. showed that the results for Generalized Linear Models (GLM) and i.i.d. sources stemming from the replica method are indeed correct and imply the optimal value of both estimation and generalization error. The proof is based on the adaptive interpolation method (Barbier and Macris, 2017) which is an extension of interpolation method developed by Guerra and Toninelli (Guerra and Toninelli, 2002) in the context of spin glasses, with an adaptive interpolation path. Recently, the exact asymptotic expressions for the normalized mutual information and minimum mean-square-error (MMSE) of sparse linear regression in the sub-linear sparsity regime have been established by using the same approach (Truong, 2021b).

In all above research literature, the authors assume that the source is independently and identically distributed (i.i.d.). In many practical applications, samples of input data may be dependent on each other, e.g., Markov chains or hidden Markov models. There were a few non-rigorous literatures handling Markov chains using the replica method (Skantzos et al., 1999; Takeda and Kabashima, 2011, 2010). However, to the best of our knowledge, there exists no rigorously analytic result which was developed based on replica-related methods for these models. It looks hard to apply the adaptive interpolation method looks for the linear model with Markov sources since this method requires that $X_1, X_2, \dots, X_n$ are i.i.d. (or at least i.i.d. block-by-block) to guarantee a fixed interpolating free energy at the final (k, t)-interpolation model for each finite value of n (Barbier and Macris, 2017). There exist some other works which proposed MSE fundamental limits which can be achieved by practical Approximate Message Passing algorithms (AMP) for the linear model with Markov or hidden Markov sources (Schniter, 2010; Ma et al., 2019; Berthier et al., 2020). AMP is initially proposed for sparse signal recovery and compressed sensing (Kabashima, 2003; Donoho, 2006; Candès and Wakin, 2008; Metzler et al., 2016). AMP algorithms achieve state-of-the-art performance for several high-dimensional statistical estimation problems, including compressed sensing and low-rank matrix estimation (Bayati and Montanari, 2011; Montanari and Venkataramanan, 2020).

1.2 Main Contributions

In this paper, based on the same replica assumptions as (Guo and Verdu, 2005), we establish free energy, mutual information, and MMSE for the linear model with Markov sources. The same fundamental limits for the linear model with hidden Markov signal priors were characterized in (Truong, 2021a). When limiting to the linear model with i.i.d. sources as case, we recover Guo and Verdú's results (Guo and Verdu, 2005), which extends Tanaka work (Tanaka, 2002) to more general alphabets. More specially, our main contributions are as follows:

- Using the replica method, we estimate the free energy, the normalized mutual information in the large system limit for the linear model with Markov sources (cf. Claim 1).
- Using the replica method, we characterize MMSE in the large system limit for the linear model with Markov signal prior (cf. Claim 2). We show that under the posterior mean estimator, the linear model with Markov sources is decoupled into single-input AWGN channels with state information available at both encoder and decoder where the state distribution follows the left Perron-Frobenius eigenvector with unit Manhattan norm of the stochastic matrix of Markov chains[1].
- We show that the free energies and MSEs obtained via the replica method are closely approximate to their counterparts achieved by the well-known MCMC algorithm called Metropolis–Hastings algorithm (cf. Section 4).

Compared with the linear model with i.i.d. sources (Guo and Verdu, 2005), we need to deal with some new technical challenges related to the estimation of the derivative of Perron-Frobenius eigenvalue of nonnegative matrices. In this work, we develop a new technique to estimate this derivative in the large system limit.

1.3 Paper Organization

The problem setting is placed in Section 2, where we introduce the system model, MMSE estimation,

[1]For any irreducible Markov process $\{Z_n\}_{n=1}^{\infty}$, the left Perron-Frobenius eigenvector with unit Manhattan norm is the stationary distribution of this Markov process, and the Perron-Frobenius eigenvalue is equal to 1 (Lancaster and Tismenetsky, 1985).

free energy and replica method in statistical physics. We also introduce some new concepts such as single-symbol MMSE channel with state information, free energy functions, and other related notations in this section. Our main results are stated and proved in Section 3. We apply our main results to estimate free energy, mutual information, and MMSE for some specific Markov chains in Section 4, where we also compare our obtained MMSEs with achievable MSEs by the classical Metropolis–Hastings algorithm in research literature. An outline of proof is given in Section 5 and more detailed proofs can be found in (Truong, 2020).

1.4 Notation

Use $[n]$ to denote the set $\{1,\dots,n\}$. Random vectors and matrices are in bold letters. Expectations with respect to "quenched" random variables (i.e., the variables that are fixed by the realization of the problem) are denoted by $\mathbb{E}$ and those with respect to "annealed" random variables (i.e., dynamical variables) are denoted by Gibbs bracket $\langle - \rangle$ possibly with appropriate subscripts. This choice follows the stardards of statistical physics.

As standard literature, we define $x^n = (x_1, x_2, \cdots, x_n)^T$ to denote a vector of length n. However, if the dimension of a vector x is clear from context, we omit it for simplicity. Define two loss functions $l_1 : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ and $l_2 : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ as $l_1(x,y) = |x-y|$ and $l_2(x,y) = (x-y)^2$. Let $\log x := \log_2 x$ and $\ln x$ be the natural logarithm of x for all $x \in \mathbb{R}^+$. Manhattan and Euclidean norms of a vector $x \in \mathbb{R}^n$ are defined as $\|x\|_1 := \sum_{i=1}^n |x_i|, \|x\|_2 := \sqrt{\sum_{i=1}^n |x_i|^2}$, respectively. In addition, $\mathrm{vec}(\cdot)$ denotes the vectorization operator.

The moment generating function of a random vector $\boldsymbol{X} \in \mathbb{R}^n$ is defined as $\mathcal{M}(\lambda) := \mathbb{E}[\exp(\lambda^T \boldsymbol{X})]$ for all $\lambda \in \mathbb{R}^n$. Let $\mathcal{M}(\tilde{Q}) := \mathbb{E}[\exp(\mathrm{tr}(\tilde{Q}\boldsymbol{Q}))]$ be the moment generating function of a random matrix $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ for all matrix $\tilde{Q} \in \mathbb{R}^{n\times n}$.

Denote by

$$\mathcal{Q} := \left\{ sxx^T \ \text{ for some } \ (s,x) \in \mathcal{S}\times\mathcal{X}^{\nu+1} \right\}. \tag{1}$$

For simplicity of presentation, we enumerate all matrices in $\mathcal{Q}$ as $\bar{Q}_0, \bar{Q}_1, \cdots, \bar{Q}_M$ where $M := |\mathcal{Q}| - 1$.

2 Problem Setting

We consider the linear model

$$\boldsymbol{Y} = \boldsymbol{\Phi}\boldsymbol{X} + \boldsymbol{W} = \boldsymbol{A}\boldsymbol{S}^{1/2}\boldsymbol{X} + \boldsymbol{W}. \tag{2}$$

Here $\boldsymbol{Y} \in \mathbb{R}^m$ is a vector of observations, $\boldsymbol{X} \in \mathbb{R}^n$ is the signal vector, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is a measurement matrix, $\boldsymbol{S}$ is diagonal matrix of positive scale factors:

$$\boldsymbol{S} = \mathrm{diag}(S_1, S_2, \cdots, S_n), \quad S_j \in \mathbb{R}^+, \tag{3}$$

and $\boldsymbol{W} \in \mathbb{R}^m$ is a noise vector. We consider a sequence of problems indexed by n, and make the following assumptions on the model. These assumptions are identical to those in earlier works (Guo and Verdu, 2005; Rangan et al., 2012) except for the signal prior, which we allow to be Markov or hidden Markov in contrast to the i.i.d. priors considered in earlier works.

1. We assume that the number of measurements m scales linearly with n, and $\lim_{n\to\infty} \frac{n}{m} = \beta$, for some $\beta > 0$.

2. The elements $\{A_{ij}\}_{i\in[m],j\in[n]}$ of the matrix $\boldsymbol{A}$ are i.i.d. and distributed as $A_{ij} \overset{\mathrm{d}}{=} \frac{1}{\sqrt{m}} A$, where A is a random variable with zero mean, unit variance and all moments finite.

3. The scale factors $(S_1, \dots, S_n)$ are i.i.d. according to P_S, which is supported on a set $\mathcal{S} \subset \mathbb{R}^+$. The scale factors $(S_1,\dots,S_n)$ are independent of $\boldsymbol{A}$, $\boldsymbol{X}$, and $\boldsymbol{W}$.

4. The noise vector $\boldsymbol{W}$ is standard normal, i.e., $W_j \sim_{\text{i.i.d.}} \mathcal{N}(0,1)$ for $j \in [m]$.

5. *Signal prior*: We assume that the components of $\boldsymbol{X}$ take values on a Polish space on $\mathbb{R}$, and are distributed according to a Markov prior, i.e.,

$$\begin{aligned} p_{\boldsymbol{X}}&(x_1,\dots,x_n) \\ &= p(x_1)\pi(x_1,x_2)\cdots\pi(x_{n-1},x_n) \end{aligned} \tag{4}$$

   for some initial probability distribution $p(\cdot)$ on $\mathcal{X}$, where $\pi(\cdot,\cdot)$ is the transition probability of a time-homogeneous, irreducible Markov chain on $\mathcal{X}$.

For simplicity of presentation, we assume that Markov chains $\{X_n\}_{n=1}^\infty$ has finite state spaces and $\mathcal{S}$ has a finite number of elements in our proofs. However, it is not hard to extend these proofs to Markov chains on Polish spaces in $\mathbb{R}$ and an infinite set $\mathcal{S}$ by referring to a more general definition of Markov chain in (Tuominen and Tweedie, 1979). An *irreducible* and *recurrent* Markov chain on an infinite state-space is called a Harris chain (Tuominen and Tweedie, 1979), which owns many similar properties to the finite state-space version such as the existence of an unique stationary distribution. For both models, we denote the joint probability mass distribution (pmf) of the signal by $p(x_1,\dots,x_n)$. For general proofs, we use Radon–Nikodym derivatives with respect to corresponding measures (Royden and Fitzpatrick, 2010).

2.1 MMSE Estimation

The problem setting described above induces a posterior distribution $p_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\Phi}}$, given by

$$p_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\Phi}}(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\phi}) = \frac{p_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\Phi}}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\phi}) p_{\boldsymbol{X}}(\boldsymbol{x})}{p_{\boldsymbol{Y}|\boldsymbol{\Phi}}(\boldsymbol{y} \mid \boldsymbol{\phi})}, \quad (5)$$

where

$$p_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\Phi}}(\boldsymbol{y} \mid \boldsymbol{x}, \Phi) = (2\pi)^{-m/2} \exp\left[- \frac{\|\boldsymbol{y} - \boldsymbol{\phi}\boldsymbol{x}\|^2}{2} \right], \quad (6)$$

and

$$p_{\boldsymbol{Y}|\boldsymbol{\Phi}}(\boldsymbol{y} \mid \boldsymbol{\phi}) = \mathbb{E}_p[p_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\Phi}}(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\phi})] \quad (7)$$

$$= \sum_{\boldsymbol{x}} p_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\Phi}}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\phi}) p_{\boldsymbol{X}}(\boldsymbol{x}). \quad (8)$$

The (canonical) MMSE estimator, which computes the mean value of the posterior distribution $p_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\Phi}}$ is given by,

$$\langle \boldsymbol{X} \rangle = \mathbb{E}_p\big[\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{\Phi}\big]. \quad (9)$$

This estimator achieves the minimum Mean-Square Error (MSE) between the estimated and the original signal.

2.2 Free Energy and Replica Method

Let

$$Z(\boldsymbol{Y}, \boldsymbol{\Phi}) := p_{\boldsymbol{Y}|\boldsymbol{\Phi}}(\boldsymbol{Y}|\boldsymbol{\Phi}). \quad (10)$$

The free energy of the model in (2) is defined as

$$\mathcal{F}_n := -\frac{1}{n} \log Z(\boldsymbol{Y}, \boldsymbol{\Phi}). \quad (11)$$

The expectation of the free energy (with respect to $p_{\boldsymbol{Y}|\boldsymbol{\Phi}}(\boldsymbol{Y}|\boldsymbol{\Phi})$) is equal to the conditional entropy of the observation $\frac{1}{n} H_p(\boldsymbol{Y}|\boldsymbol{\Phi})$ as well as (up to an additive constant) to the mutual information density between the signal and the observations $\frac{1}{n} I_p(\boldsymbol{X}, \boldsymbol{Y})$.

The asymptotic free energy is the limit of the sequence $\{\mathcal{F}_n\}_{n=1}^{\infty}$, i.e.,

$$\mathcal{F} := \lim_{n\to\infty} \mathcal{F}_n. \quad (12)$$

In general, it is very challenging to prove the existence and estimate the limit in (12). Replica method, originally developed in statistical physics, is usually used to evaluate this limit (Tanaka, 2002; Guo and Verdu, 2005) because the linear model is similar to the thermodynamic system. For this model, replica method is based on the following assumptions (A) and facts (F):

- (A1) The free energy $\mathcal{F}_n$ has the self-averaging property as $n \to \infty$. This means that

$$\mathcal{F} := \lim_{n\to\infty} \mathbb{E}[\mathcal{F}_n]. \quad (13)$$

  The self-averaging property essentially assumes that the variations of $Z(\boldsymbol{Y}, \boldsymbol{\Phi})$ due to the randomness of the measurement matrix Φ vanish in the limit $n \to \infty$. Although a large number of statistical physics quantities exhibit such self-averaging, the self-averaging of the relevant quantities for the MMSE and MAP analyses has not been rigorously established (Rangan et al., 2012). For the purpose of estimating the average mutual information of the Markov model only, we don't need to make use of this assumption.

- (F1) The following identity holds:

$$\mathbb{E}[\log Z(\boldsymbol{Y}, \boldsymbol{\Phi})] = \lim_{\nu\to 0} \frac{\partial}{\partial \nu} \log \mathbb{E}[Z^{\nu}(\boldsymbol{Y}, \boldsymbol{\Phi})]. \quad (14)$$

- (A2) Estimation of $\mathbb{E}[Z(\boldsymbol{Y}, \boldsymbol{\Phi})^{\nu}]$ for a positive real number ν in the neighbourhood of 0 can be done by two steps: (1) Estimate $\mathbb{E}[Z^{\nu}(\boldsymbol{Y}, \boldsymbol{\Phi})]$ for a general positive integer ν (2) Take the limit of the obtained result as $\nu \to 0$. This is called "replica trick" in statistical physics.

- (F2) For any positive integer ν and a realization $(\boldsymbol{y}, \Phi)$ of $(\boldsymbol{Y}, \boldsymbol{\Phi})$, the quantity $Z^{\nu}(\boldsymbol{y}, \boldsymbol{\Phi})$ can be written as

$$Z^{\nu}(\boldsymbol{y}, \Phi) = \big\{p_{\boldsymbol{Y}|\boldsymbol{\Phi}}(\boldsymbol{y}|\Phi)\big\}^{\nu} \quad (15)$$

$$= \left\{ \mathbb{E}_{p_{\boldsymbol{X}}} \Big[p_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\Phi}}(\boldsymbol{y}|\boldsymbol{X}, \Phi) \Big] \right\}^{\nu} \quad (16)$$

$$= \mathbb{E}_{p_{\boldsymbol{X}}} \left\{ \prod_{a=1}^{\nu} p_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\Phi}}(\boldsymbol{y}|\boldsymbol{X}^{(a)}, \Phi) \right\}. \quad (17)$$

  where the last expectation is taken over relicated vectors $\boldsymbol{X}^{(a)}, a = 1, 2, \cdots, \nu$ which are independent copies of a random vector with distribution $p_{\boldsymbol{X}}$.

- (A3) The order of limit $n \to \infty$ and $\nu \to 0$ can be interchanged. Mathematically, under some conditions such as Theorem Moore-Osgood (Stewart, 2008), the interchange between limits work. This theorem is used in (Barbier et al., 2016) for a similar purpose.

- (A4) Usually, the free energy can be expressed an optimal value of an optimization problem over the space of covariance matrices of replica samples, say $\mathcal{Q}$. This optimization is general difficult to perform. To overcome this, the replica method also makes an additional assumption that the optimizer Q^* is symmetric with respect to permutations of ν replica indices. This assumption is called Replica Symmetry (RS) in statistical physics. See Definition 9 for our assumption about RS in this paper.

(A1)-(A4) are assumed in various research literature on replica method such as (Tanaka, 2002; Guo and Verdu, 2005; Rangan et al., 2012).

3 Main results

3.1 Results for Markov Priors

Our results on the free energy and MMSE will be stated in terms of a *single-symbol* channel, similar to the equivalent single-user Gaussian channel which is obtained via decoupling as in (Guo and Verdu, 2005, Section D). Let $\lambda^{(\pi)}$ be the left Perron-Frobenius with unit Manhattan norm[2] of $P_\pi = \{\pi(x,y)\}_{x\in\mathcal{X},y\in\mathcal{X}}$ which is the stochastic matrix of the Markov chain $\{X_n\}_{n=1}^\infty$, and let $\lambda_{x_0}^{(\pi)}$ be the component of $\lambda^{(\pi)}$ associated with the x_0-th row of P_π. Let us consider the composition of a Gaussian channel with one state X_0 available at both encoder and decoder such that $\mathsf{X}_0 \sim \lambda^{(\pi)}$, a one-state MMSE, and a companion retrochannel in the single-symbol setting depicted in Fig. 1. Given the state information $\mathsf{X}_0 = x_0$, the input-output relationship of this single-symbol channel is given by

$$U = \sqrt{S}\,\mathsf{X}_1 + \frac{1}{\sqrt{\eta}}W, \tag{18}$$

where the input $\mathsf{X}_1 \sim p_{\mathsf{X}_1|\mathsf{X}_0}(\cdot|x_0) := \pi(x_0,\cdot)$, $S \sim P_S$ is the input Signal-to-Noise Ratio (SNR) which is independent X_0 and X_1, $W \sim \mathcal{N}(0,1)$ the noise independent of X_0 and X_1, and $\eta > 0$ the inverse noise variance.

The conditional distribution associated with the channel is

$$p_{U|\mathsf{X}_0,\mathsf{X}_1,S;\eta}(u \mid x_0,x_1,s;\eta) = \sqrt{\frac{\eta}{2\pi}}\exp\left[-\frac{\eta}{2}(u-\sqrt{s}x_1)^2\right]. \tag{19}$$

By setting the input distribution to be $p_{\mathsf{X}_1|\mathsf{X}_0}(\cdot|x_0) = \pi(x_0,\cdot)$, a posterior probability distribution $p_{\mathsf{X}_1|\mathsf{X}_0,U,S;\eta}$ is induced by $p_{\mathsf{X}_1|\mathsf{X}_0}$ and $p_{U|\mathsf{X}_0,\mathsf{X}_1,S;\eta}$ using the Bayes rule, i.e.,

$$p_{\mathsf{X}_1|\mathsf{X}_0,S,U;\eta}(x \mid x_0,s,u;\eta) = \frac{p_{\mathsf{X}_1|\mathsf{X}_0}(x \mid x_0)p_{U|\mathsf{X}_0,\mathsf{X}_1,S;\eta}(u \mid x_0,x_1,s;\eta)}{p_{U|\mathsf{X}_0,S;\eta}(u \mid x_0,s;\eta)}. \tag{20}$$

This induces a single-use retrochannel with random transformation $p_{\mathsf{X}_1|\mathsf{X}_0,U,S;\eta}$, which outputs a random variable X given the channel output U and the channel state X_0 (Fig. 1). An single-symbol MMSE estimator with state available $\mathsf{X}_0 = x_0$ is defined naturally as

$$\langle \mathsf{X}|\mathsf{X}_0 = x_0\rangle_p = \mathbb{E}_p\big[\mathsf{X}|\mathsf{X}_0 = x_0, U, S;\eta\big], \tag{21}$$

where the expectation is taken over the (conditionally) distribution in (20).

The single-symbol MMSE estimator (21) is merely a decision function applied to the Gaussian channel output with state $\mathsf{X}_0 = x_0$ available at both encoder and decoder (or input and output), which can expressed explicitly as

$$\mathbb{E}_p\big[\mathsf{X}|U,\mathsf{X}_0 = x_0, S;\eta\big] = \frac{p_1(U,x_0,S;\eta)}{p_0(U,x_0,S;\eta)}, \tag{22}$$

where

$$p_0(u,x_0,S;\eta) := p_{U|\mathsf{X}_0,S;\eta}(u \mid x_0,s;\eta) \tag{23}$$

$$= \mathbb{E}_{\pi(x_0,\cdot)}\left[p_{U|\mathsf{X}_0,\mathsf{X}_1,S;\eta}(u \mid x_0,\mathsf{X},S;\eta)\middle|S\right], \tag{24}$$

$$p_1(z,x_0,S;\eta) = \mathbb{E}_{\pi(x_0,\cdot)}\left[\mathsf{X}\,p_{U|\mathsf{X}_0,\mathsf{X}_1,S;\eta}(z \mid x_0,\mathsf{X},S;\eta)\middle|S\right]. \tag{25}$$

The probability law of the (composite) single-symbol channel depicted by Fig. 1 is determined by S and parameter η given state X_0. We define the conditional mean-square error of the MMSE estimator as

$$\mathcal{E}(S;\eta|x_0) = \mathbb{E}[(\mathsf{X}_1 - \langle \mathsf{X}\,|\,\mathsf{X}_0 = x_0\rangle_p)^2 \mid \mathsf{X}_0 = x_0, S;\eta]. \tag{26}$$

Define

$$\mathcal{G} := \sum_{x_0\in\mathcal{X}} \lambda_{x_0}^{(\pi)}\mathcal{G}(x_0), \tag{27}$$

where $\mathcal{G}(x_0)$ is defined in (28), and η is the solution of the following equation

$$\eta^{-1} = 1 + \beta\sum_{x_0\in\mathcal{X}}\lambda_{x_0}^{(\pi)}\mathbb{E}[S\mathcal{E}(S;\eta \mid x_0)], \tag{29}$$

[2]Since there exists a unique left Perron-Frobenius eigenvector up to a positive scaling factor (Lancaster and Tismenetsky, 1985), $\lambda^{(\pi)}$ exists uniquely, which is the stationary distribution of the Markov chain.

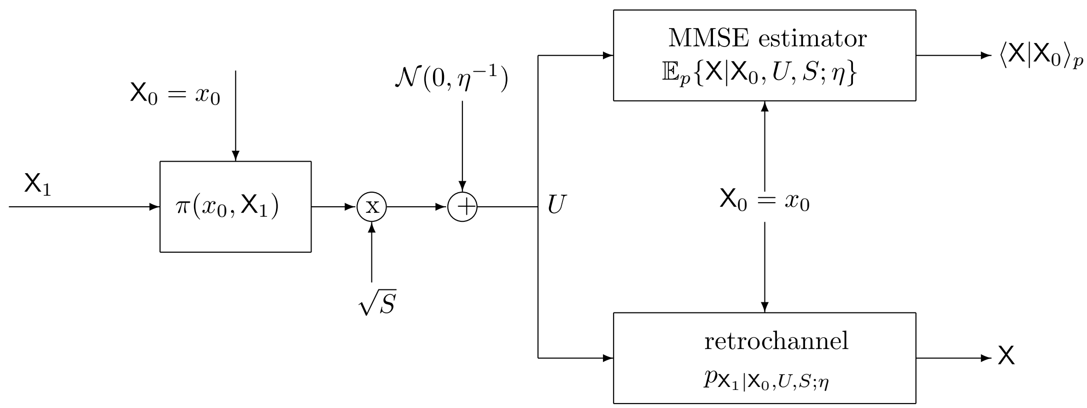

Figure 1: The equivalent single-symbol Gaussian channel with state available at both encoder and decoder, MMSE estimator, and retrochannel.

$$\mathcal{G}(x_0) := -\mathbb{E}\bigg\{\int p_{U|\mathsf{X}_0,\mathsf{X}_1,S;\eta}(u \mid x_0, \mathsf{X}_1, S; \eta) \log p_{U|\mathsf{X}_0,\mathsf{X}_1,S;\eta}(u \mid x_0, \mathsf{X}_1, S; \eta) du\bigg\} + \frac{1}{2\beta}\bigg[(\eta - 1)\log e - \log \eta\bigg] - \frac{1}{2}\log\frac{2\pi}{\eta} - \frac{\eta}{2\eta}\log e + \frac{1}{2\beta}\log(2\pi) + \frac{\eta}{2\beta\eta}\log e. \quad (28)$$

such that they minimize $\mathcal{G}$. Observe that for the case $\mathsf{X}_0, \mathsf{X}_1, \cdots, \mathsf{X}_n$ are i.i.d., $\mathcal{G}(x_0)$ does not depend on x_0 and we recovers the result in (Guo and Verdu, 2005, Eq. (22)) for this special case.

Claim 1. *The free energy of the linear model with Markov sources in Section 2 satisfies*

$$\mathcal{F} = \mathcal{G}, \quad (30)$$

where $\mathcal{G}$ is defined in (27). In addition, the average mutual information of this model satisfies:

$$C = \lim_{n\to\infty} \frac{1}{n} I(\boldsymbol{X}^n; \boldsymbol{Y}^m) = \mathcal{F} - \frac{1}{2\beta}. \quad (31)$$

Claim 2. *Recall the definition of $\{\lambda_{x_0}^{(\pi)}\}_{x_0\in\mathcal{X}}$ in Section 3.1. Assume that the MMSE estimator defined in (9) is used for estimation. Then, for all $k \in \{1, 2, \cdots, n\}$, the joint moments satisfy:*

$$\lim_{n\to\infty} \mathbb{E}\bigg[X_k^{i_0} \tilde{X}_k^{j_0} \langle X_k\rangle_p^{l_0}\bigg] = \sum_{x_0\in\mathcal{X}} \lambda_{x_0}^{(\pi)} \mathbb{E}\bigg[\mathsf{X}_1^{i_0} \mathsf{X}^{j_0} \langle \mathsf{X}|\mathsf{X}_0\rangle_p^{l_0} \big| \mathsf{X}_0 = x_0\bigg] \quad (32)$$

for all $i_0, j_0, l_0 \in \mathbb{Z}_+$, where $(\mathsf{X}_1, \mathsf{X}, \langle\mathsf{X}|\mathsf{X}_0\rangle_p)$ is the input and outputs defined in the (composite) single-symbol MMSE channel in Fig. 1, and $(X_k, \tilde{X}_k, \langle X_k\rangle)$ is the k-th symbol in the vector $\boldsymbol{X} \in \mathcal{X}^n$, the k-th output of the vector retrochanel defined in (5), and its corresponding estimated symbol by using the MMSE estimator in (9), respectively.

In addition, the average MMSE satisfies:

$$\frac{1}{n}\mathbb{E}\big[\|\boldsymbol{X} - \langle\boldsymbol{X}\rangle\|_2^2\big] = \mathbb{E}\big[\mathsf{X}_1^2\big] - \sum_{x_0\in\mathcal{X}} \lambda_{x_0}^{(\pi)} \mathbb{E}\big[\langle\mathsf{X}|\mathsf{X}_0\rangle_p^2 \big| \mathsf{X}_0 = x_0\big], \quad (33)$$

where $\mathsf{X}_1 \sim \sum_{x_0\in\mathcal{X}} \pi(x_0, \cdot)p(x_0)$.

4 Numerical examples and comparison with algorithmic performance

4.1 Binary-valued Markov Prior

In this experiment, we consider a homogeneous Markov chain on the alphabet $\mathcal{X} = \{-1, 1\}$ with the stochastic matrix

$$P_\pi = \begin{bmatrix} \pi(-1,-1) & \pi(-1,1) \\ \pi(1,-1) & \pi(1,1) \end{bmatrix} = \begin{bmatrix} 1-\alpha & \alpha \\ \delta & 1-\delta \end{bmatrix} \quad (34)$$

for some α and δ in $(0, 1)$.

We compare the free energies achieved by the replica prediction (cf. Claim 1) and MCMC (Metropolis–Hastings algorithm) for the linear model with binary-valued Markov prior defined in (34). More specifically, we use the Markov Chain Monte-Carlo (MCMC) simulation method to estimate the density function $\boldsymbol{P}_{\boldsymbol{y}|\boldsymbol{\Phi}}(\boldsymbol{y}|\boldsymbol{\Phi})$ and verify our replica predictions in Claim 1. Our results show that the free energy curves by the replica method and MCMC nearly coincide to each other for all three cases: (1) i.i.d. prior ($\alpha = \delta = 0.5$), (2) symmetric Markov prior $\alpha = \delta = 0.3$, (3) asymmetric Markov prior ($\alpha = 0.2, \delta = 0.5$) (cf. Figs. 2, 3, and 4). In our MCMC simulations, the Metropolis–Hastings algorithm is used where the state $\boldsymbol{x}_t := \mathrm{vec}(\mathrm{vec}(\Phi_{t+1}), \mathrm{y}_{t+1})$ and the probability transition $g(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t) \sim \mathcal{N}(\boldsymbol{x}_t, \boldsymbol{I}_{mn+n})$.

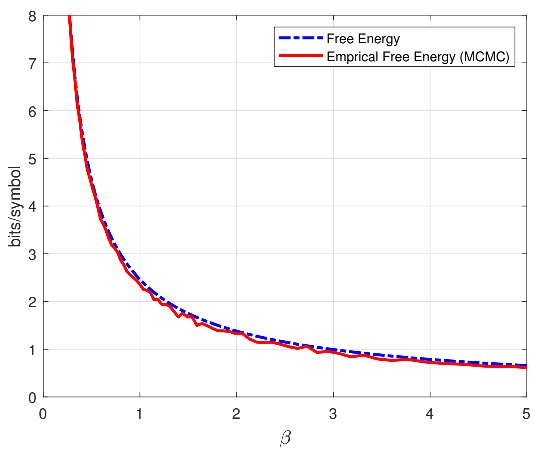

Figure 2: Free energy by Replica Method and MCMC as functions for the i.i.d. prior $\alpha = \delta = 0.5$.

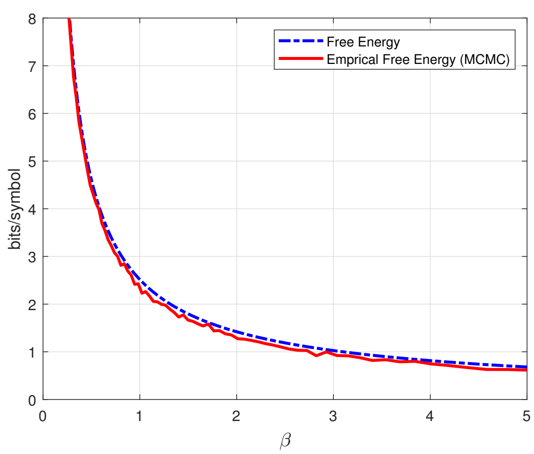

Figure 3: Free energy by Replica Method and MCMC as functions of β for the symmetric case $\alpha = \delta = 0.3$.

Since MMSE is fixed function of the free energy (or mutual information) (Guo et al., 2005), these simulation results also indicate that our replica prediction in Claim 2 closely approximates the MMSE of the model. Designing practical AMP algorithms to achieve these fundamental limits is a future research direction of interest.

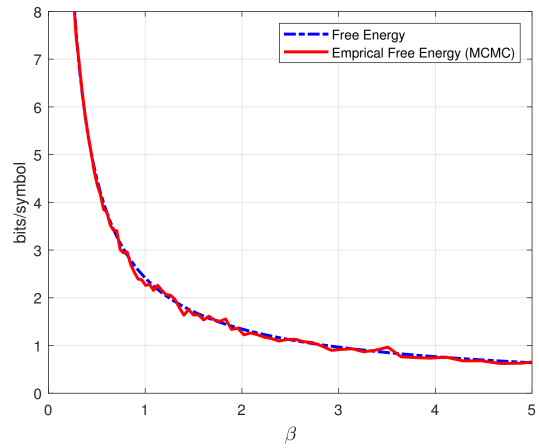

Figure 4: Free energy by Replica Method and MCMC as functions of β for the non-symmetric case $\alpha = 0.2$ and $\delta = 0.5$.

4.2 Gauss-Markov Prior

In this experiment, we consider a Gauss-Markov prior $\{X_n\}_{n=1}^{\infty}$ on $\mathcal{X} = \mathbb{R}$, i.e., $X_n = \nu X_{n-1} + Z_n$, where $Z_n \sim \mathcal{N}(0, \sigma_0^2)$ and $\nu \in (0,1)$. Then, the transition probability is

$$\pi(x_0, x) := \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_0^2}(x - \nu x_0)^2\right]. \quad (35)$$

In this subsection, we use the same MCMC algorithm as Subsection 4.1, i.e., the Metropolis–Hastings algorithm. In Fig. 5, we plot the free energy curves for the linear model with Markov prior in (35) for three cases $\nu = 0.1$, $\nu = 0.5$, and $\nu = 0.8$. In these plots, we set $X_1 \sim \mathcal{N}(0, \frac{\sigma_0^2}{1-\nu^2})$ to force the state distribution of the Markov (Harris) chain $X_n \sim \mathcal{N}(0, \frac{\sigma_0^2}{1-\nu^2})$ for all $n \geq 1$. The plot shows that the replica prediction for the free energy (cf. Claim 12) is very closed to the MCMC simulation result. As in the first experiment, since the MMSE is a fixed function of the free energy (or mutual information) (Guo et al., 2005), this also means that the MMSE curve by replica method closely approaches the MMSE of the model.

5 Sketch of Proofs of Claims 1 and 2

The proofs of Claims 1 and 2 are based on (Guo and Verdu, 2005) with some important changes to account for the Markov prior assumptions. The following are main steps in the proof.

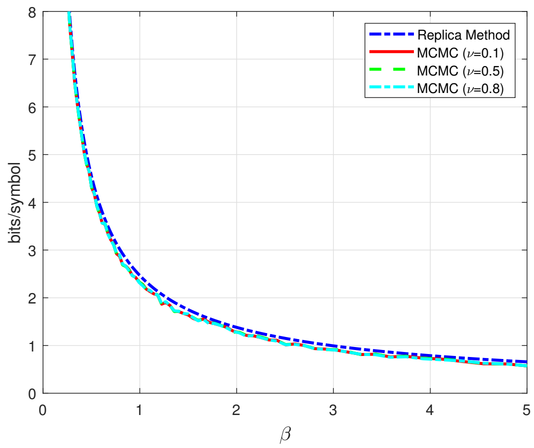

Figure 5: Free energy by replica method and empirical MCMC as functions of β for $\sigma_0^2 = 1$ and $s_0 = 1$.

- Let $\{S_n\}_{n=1}^{\infty}$ be an i.i.d. sequence of random variable on a finite set $\mathcal{S} \subset \mathbb{R}^+$. Let $\boldsymbol{X} := \{X_n\}_{n=1}^{\infty}$ be a Markov chain with states on a Polish space $\mathcal{X}$ with the transition matrix $P = \{\pi(x,x')\}_{x,x' \in \mathcal{X}}$. Assume this Markov chain is irreducible. Set $\boldsymbol{X}^{(0)} = \boldsymbol{X}$. Let $\boldsymbol{X}^{(a)} := \{X_n^{(a)}\}_{n=1}^{\infty}$ be a set of ν replica sequences with distribution $p_{\boldsymbol{X}}$ defined in (4) for each $a = 1, 2, \cdots, \nu$. Define a new sequence of $(\nu+1)\times(\nu+1)$ random matrices $\{\boldsymbol{Q}_n\}_{n=1}^{\infty}$ such that

$$Q_n^{(a,b)} = S_n X_n^{(a)} X_n^{(b)} \tag{36}$$

for all $a \in [\nu]$ and $b \in [\nu]$ and for all $n = 1, 2, \cdots$. Then, $\{\boldsymbol{Q}_n\}_{n=1}^{\infty}$ is also an irreducible Markov chain with states on $\mathcal{Q}$, where $\mathcal{Q}$ is defined in (1).

- Let $\boldsymbol{T}_n := \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{Q}_k$. Prove that

$$\begin{aligned} &\frac{1}{n}\log \mathbb{E}[Z^{\nu}(\boldsymbol{Y}, \boldsymbol{\Phi})] \\ &\quad = \frac{1}{n}\log \mathbb{E}\bigg\{\exp\bigg[m\Big(G^{(\nu)}(\boldsymbol{T}_n) + O(n^{-1})\Big)\bigg]\bigg\}, \end{aligned} \tag{37}$$

where $G^{(\nu)} : \mathcal{Q} \to \mathbb{R}$ such that

$$\begin{aligned} G^{(\nu)}(Q) &:= -\frac{1}{2}\log\det(I + \Sigma Q) \\ &\quad - \frac{1}{2}\log\left(1 + \frac{\nu}{\sigma^2}\right) - \frac{\nu}{2}\log(2\pi\sigma^2), \end{aligned} \tag{38}$$

and Σ is a $(\nu+1)\times(\nu+1)$ matrix

$$\Sigma = \frac{\beta}{\sigma^2 + \nu}\begin{bmatrix} \nu & -e^T \\ -e & (1 + \frac{\nu}{\sigma^2})I - \frac{1}{\sigma^2}ee^T \end{bmatrix}, \tag{39}$$

where e is a $\nu \times 1$ column vector whose entries are all 1.

- Based on (F1)-(F2), (A1)-(A3), and the Large Deviations Principle for the distribution of $\boldsymbol{T}_n$, we can show that

$$\begin{aligned} \mathcal{F} &= \lim_{n\to\infty} -\frac{1}{n}\log \mathbb{E}[Z^{\nu}(\boldsymbol{Y}, \boldsymbol{\Phi})] && (40) \\ &= -\lim_{\nu\to 0}\frac{\partial}{\partial\nu}\sup_{Q}\Big[\beta^{-1}G^{(\nu)}(Q) - I^{(\nu)}(Q)\Big], && (41) \end{aligned}$$

where

$$I^{(\nu)}(Q) := \sup_{\tilde{Q}}\Big[\mathrm{tr}(\tilde{Q}Q) - \log\rho(P_{\tilde{Q}})\Big], \tag{42}$$

and $\rho(P_{\tilde{Q}})$ is the Perron-Frobenius eigenvalue of the matrix $P_{\tilde{Q}} = \{e^{\mathrm{tr}(\tilde{Q}\bar{Q}_j)} P_{\bar{Q}_j|\bar{Q}_i}\}_{0\le i,j\le M}$ and $M = |\mathcal{Q}| - 1$ where $\mathcal{Q} := \{sxx^T$ for some $s \in \mathcal{S},\ x \in \mathcal{X}^{\nu+1}\}$.

- From (42), it is easy to see that the optimizer $(Q^*, \tilde{Q}^*)$ of (41) and the large deviation rate function $I^{(\nu)}(Q)$, respectively, must satisfy

$$Q^* = \frac{\partial \log\rho(P_{\tilde{Q}})}{\partial\tilde{Q}}(\tilde{Q}^*). \tag{43}$$

Then, by adapting a result in (Deutsch and Neumann, 1984) related to the derivatives of the Perron-eigenvalue of a non-negative matrix and (43), it can be shown that under the replica assumptions (A1)-(A4), Q^* is a convex combination of moment generating functions which can be estimated by using the methods in (Guo and Verdu, 2005).

- Furthermore, by using (A1)-(A4) and the fact that

$$\rho(P_{\tilde{Q}^*}) = \sum_{i=1}^{M}\lambda_i(\tilde{Q}^*)\mathbb{E}[e^{\mathrm{tr}(\tilde{Q}^* Q_1)}|\boldsymbol{Q}_0 = \bar{Q}_i], \tag{44}$$

we can show that $\frac{\partial\log\rho(P_{\tilde{Q}})}{\partial\nu}(\tilde{Q}^*)$ is also a convex combination of some conditional moment generating functions which can be estimated based on (Guo and Verdu, 2005).

- Finally, by using (41) and results in bullet points 4 and 5, we can prove that the free energy of the linear model with Markov signal prior is a convex combination of the other conditional free energies where each conditional free energy corresponds to the free energy of a linear model with i. i. d. signal prior with known expression in (Guo and Verdu, 2005). Here, the convex combination coefficients follow the stationary distribution of the stochastic matrix of Markov chains $\{X_n\}_{n=1}^{\infty}$.

- To prove Claim (107), we first recall that the Large Deviations Principle for probability measures (distribution of $\boldsymbol{T}_n$) also holds for any finite Borel measures on compact metric space (e.g.(Young, 1990)) or on Polish space (Swart, 2012). Then, the proof follows the same ideas as the proof of Claim 12.

Acknowledgements

The author is grateful to Prof. Ramji Venkataramanan, the University of Cambridge, for many suggestions to improve the paper.

References

J. Barbier, M. Dia, N. Macris, and F. Krzakala. The mutual information in random linear estimation. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 625–632, 2016.

Jean Barbier and Nicolas Macris. The adaptive interpolation method: a simple scheme to prove replica formulas in bayesian inference. *Probability Theory and Related Fields*, 174:1133–1185, 2017.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

Jean Barbier, Nicolas Macris, Mohamad Dia, and Florent Krzakala. Mutual information and optimality of approximate message-passing in random linear estimation. *IEEE Transactions on Information Theory*, 66(7):4270–4303, 2020.

D. Baron, S. Sarvotham, and R. G. Baraniuk. Bayesian compressive sensing via belief propagation. *IEEE Transactions on Signal Processing*, 58:269–280, 2010.

M. Bayati and A. Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. on Inform. Th.*, 57(2):764–785, Feb 2011.

A. Bereyhi, R. Müller, and H. S. Baldes. Statistical mechanics of MAP estimation: General replica ansatz. *IEEE Trans. on Inform. Th.*, 65(12):7896–7934, Dec. 2019.

R. Berthier, A. Montanari, and P. M. Nguyen. State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of IMA*, 9(1), 2020.

P. Billingsley. *Probability and Measure*. Wiley-Interscience, 3rd edition, 1995.

E. Candès and M. Wakin. An introduction to compressive sampling. *IEEE Signal Process. Mag.*, 25 (2):21–30, 2008.

I. Chalendar and J. R. Partington. Multivariable approximate Carleman-type theorems for complex measures. *Annals of Probability*, 35(1):384–396, 2007.

E. Deutsch and M. Neumann. Derivatives of the perron root at an essential nonnegative matrix and the group inverse of an M-matrix. *Journal of Mathematical Analysis and Applications*, 102:1–29, 1984.

D. Donoho. Compressed sensing. *IEEE Trans. on Inform. Th.*, 52(1):1289–1306, 2006.

D. L. Donoho, A. Javanmard, and A. Montanari. Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Transactions on Information Theory*, 59:7434–7464, 2011.

S. F. Edwards and P. W. Anderson. Theory of spin glasses. *J. Phys. F: Metal Physics*, 5:965–974, 1975.

F. Guerra and F. L. Toninelli. The thermodynamic limit in mean field spin glass models. *Communications in Mathematical Physics*, 230:71–79, 2002.

D. Guo and S. Verdu. Randomly spread CDMA: asymptotics via statistical physics. *IEEE Transactions on Information Theory*, 51(6):1983–2010, June 2005.

D. Guo and C.-C. Wang. Asymptotic mean-square optimality of belief propagation for sparse linear systems. *2006 IEEE Information Theory Workshop - ITW '06 Chengdu*, pages 194–198, 2006.

D. Guo, S. Shamai, and S. Verdu. Mutual information and minimum mean-square error in gaussian channels. *IEEE Transactions on Information Theory*, 51 (4):1261–1282, 2005.

Yoshiyuki Kabashima. A CDMA multiuser detection algorithm on the basis of belief propagation. *Journal of Physics A: Mathematical and General*, 36(43): 11111–11121, oct 2003.

Satish Babu Korada and Nicolas Macris. Tight bounds on the capacity of binary input random cdma systems. *IEEE Transactions on Information Theory*, 56(11):5590–5613, 2010.

F. Krzakala, M. Mézard, F. Sausset, Y. F. Sun, and L. Zdeborová. Statistical-physics-based reconstruction in compressed sensing. *Phys. Rev. X*, 2:021005, May 2012.

S. Kudekar and H. D. Pfister. The effect of spatial coupling on compressive sensing. In *The 48th Annual Allerton Conference on Communications and Computation*, pages 347–353, 2010.

P. Lancaster and M. Tismenetsky. *The Theory of Matrices*. Academic Press, 2nd edition, 1985.

Y. Ma, C. Rush, and D. Baron. Analysis of approximate message passing with a class of non-separable denoiser and markov random field priors. *IEEE Trans. on Inform. Th.*, 65(11):7367–7389, 2019.

C. A. Metzler, A. Maleki, and R. G. Baraniuk. From denoising to compressed sensing. *IEEE Trans. on Inform. Th.*, 62(9):5117–514, 2016.

A. Montanari and D. Tse. Analysis of belief propagation for non-linear problems: The example of CDMA (or: How to prove Tanaka's formula). *2006 IEEE Information Theory Workshop - ITW '06 Punta del Este*, pages 160–164, 2006.

A. Montanari and R. Venkataramanan. Estimation of low-rank matrices via approximate message passing. *Annals of Mathematical Statistics*, 2020.

S. Rangan, A. K. Fletcher, and V. K. Goyal. Asymptotic analysis of MAP estimation via the replica method and applications to compressed sensing. *IEEE Trans. on Inform. Th.*, 58(3):1902–1923, March 2012.

G. Reeves and H. D. Pfister. The replica-symmetric prediction for random linear estimation with gaussian matrices is exact. *IEEE Trans. on Inform. Th.*, 65(4):2252–2283, April 2019.

H. Royden and P. Fitzpatrick. *Real Analysis*. Pearson, 4th edition, 2010.

P. Schniter. Turbo reconstruction of structured sparse signals. In *Proceedings of Conference on Information Sciences and Systems*, 2010.

N. Skantzos, C. Beckmann, and Anthony Coolen. Discontinuous recall transitions induced by competition between short- and long-range interactions in recurrent networks. In M. Kearns, S. Solla, and D. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11. MIT Press, 1999.

J. Stewart. *Multivariate Calculus*. Cengate Learning, 6th edition, 2008.

J. M. Swart. Large deviation theory. The Czech Academy of Sciences Institute of Information Theory and Automation, 2012.

K. Takeda and Y. Kabashima. Statistical mechanical assessment of a reconstruction limit of compressed sensing: Toward theoretical analysis of correlated signals. *EPL (Europhysics Letters)*, 95(1):18006, jun 2011.

Koujin Takeda and Yoshiyuki Kabashima. Statistical mechanical analysis of compressed sensing utilizing correlated compression matrix. In *2010 IEEE International Symposium on Information Theory*, pages 1538–1542, 2010.

T. Tanaka. A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors. *IEEE Trans. on Inform. Th.*, 48(11):2888–2909, 2002.

Lan V. Truong. Replica analysis of the linear model with markov or hidden markov signal priors. `arXiv:2009.13370 [cs.IT]`, 2020.

Lan V. Truong. Linear models with hidden markov sources via replica method. In *IEEE International Symposium on Information Theory (ISIT)*, pages 396–401, 2021a.

Lan V. Truong. Fundamental limits and algorithms for sparse linear regression with sublinear sparsity. `arXiv:2101.11156 [cs.IT]`, 2021b.

P. Tuominen and R. L. Tweedie. Markov chains with continuous components. *Proceedings of the London Mathematical Society*, s3-38(1):18914–18919, 1979.

L. S. Young. Some large deviations results for dynamical systems. *Trans. of. The American Mathematical Society*, 318(2):525–543, 1990.

Suplementary Material

This section provides a sketch of proofs for Claims 1–2 using the replica method. We first state some related lemmas which are required to estimate the free energy of the linear model with Markov sources. Then, we obtain the joint moments for the linear model with Markov sources.

Lemma 3. *(Guo and Verdu, 2005, p. 1998) Let $X_n^{(a)}$ be replicated vectors with distribution $p_{\boldsymbol{X}}$. Define a sequence of $(\nu+1)\times(\nu+1)$ random matrices $\{\boldsymbol{Q}_n\}_{n=1}^{\infty}$ such that*

$$Q_n^{(a,b)} = S_n X_n^{(a)} X_n^{(b)} \tag{45}$$

for all $a \in [\nu]$ and $b \in [\nu]$ for all $n = 1, 2, \cdots$. Let

$$\boldsymbol{T}_n = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{Q}_k, \quad n = 1, 2, \cdots. \tag{46}$$

Then, the following holds:

$$\frac{1}{n}\log \mathbb{E}[Z^{\nu}(\boldsymbol{Y},\boldsymbol{\Phi})] = \frac{1}{n}\log \mathbb{E}\bigg\{\exp\bigg[m\Big(G^{(\nu)}(\boldsymbol{T}_n) + O(n^{-1})\Big)\bigg]\bigg\}, \tag{47}$$

where

$$G^{(\nu)}(Q) := -\frac{1}{2}\log\det(I+\Sigma Q) - \frac{1}{2}\log\left(1+\frac{\nu}{\sigma^2}\right) - \frac{\nu}{2}\log(2\pi\sigma^2), \tag{48}$$

and Σ is a $(\nu+1)\times(\nu+1)$ matrix

$$\Sigma = \frac{\beta}{\sigma^2+\nu}\begin{bmatrix} \nu & -e^T \\ -e & (1+\frac{\nu}{\sigma^2})I - \frac{1}{\sigma^2}ee^T \end{bmatrix}, \tag{49}$$

where e is a $\nu \times 1$ column vector whose entries are all 1.

The following two lemmas state some new results on large deviations for Markov chains induced by the channel setting.

Lemma 4. *Let $\{S_n\}_{n=1}^{\infty}$ be an i.i.d. sequence of random variable on a finite set $\mathcal{S} \subset \mathbb{R}^+$. Let $\boldsymbol{X} := \{X_n\}_{n=1}^{\infty}$ be a Markov chain with states on a Polish space $\mathcal{X}$ with the transition matrix $P = \{\pi(x,x')\}_{x,x'\in\mathcal{X}}$. Assume this Markov chain is irreducible. Set $\boldsymbol{X}^{(0)} = \boldsymbol{X}$. Let $\boldsymbol{X}^{(a)} := \{X_n^{(a)}\}_{n=1}^{\infty}$ be a set of ν replica sequences with (postulated) distribution $q_{\boldsymbol{X}}$ for each $a = 1, 2, \cdots, \nu$. This means that*

$$p_{\boldsymbol{X}^{(0)}\boldsymbol{X}^{(1)}\boldsymbol{X}^{(2)}\cdots\boldsymbol{X}^{(\nu)}}(x^{(0)}, x^{(1)}, x^{(2)}, \cdots, x^{(\nu)}) \sim \prod_{i=0}^{\nu} p_{\boldsymbol{X}}(x^{(i)}), \tag{50}$$

where

$$p_{\boldsymbol{X}}(x^{(a)}) = \prod_{i=1}^{\infty} \pi(x_i^{(a)}, x_{i+1}^{(a)}), \qquad \forall a \in \{0, 1, \cdots, \nu\}. \tag{51}$$

Define a new sequence of $(\nu+1)\times(\nu+1)$ random matrices $\{\boldsymbol{Q}_n\}_{n=1}^{\infty}$ such that

$$Q_n^{(a,b)} = S_n X_n^{(a)} X_n^{(b)} \tag{52}$$

for all $a \in [\nu]$ and $b \in [\nu]$ and for all $n = 1, 2, \cdots$. Then, $\{\boldsymbol{Q}_n\}_{n=1}^{\infty}$ is also an irreducible Markov chain with states on $\mathcal{Q}$, where $\mathcal{Q}$ is defined in (1). In addition, the transition probability, namely $P(Q|Q')$, of this Markov chain satisfies (53) where $p_{X_{n-1}}(\cdot)$ is the state distribution at time $n-1$ of the Markov chain $\{X_n\}_{n=1}^{\infty}$ with the transition probability π defined in (4) and

$$\mathcal{A}_Q := \big\{(s,x) \in \mathcal{S}\times\mathcal{X}^{\nu+1} : sxx^T = Q\big\}, \quad \forall Q \in \mathcal{Q}. \tag{54}$$

Lemma 5. *Let $\mathcal{X}$ be a Polish space with finite cardinality and a irreducible Markov chain $\boldsymbol{X} := \{X_n\}_{n=1}^{\infty}$ defined on $\mathcal{X}$ and ν be a positive integer number. Let $X_n^{(a)}$ for $a \in \{1, 2, \cdots, \nu\}$ be replicas of the Markov process $\boldsymbol{X}$. Recall the definition of the sequence $\boldsymbol{Q}_n$ in Lemma 4 and $\boldsymbol{T}_n = \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{Q}_j$. Let $P_n(U) := \mathbb{P}(\boldsymbol{T}_n \in U)$ for any measurable set U on the σ-algebra generated by $\{\boldsymbol{Q}_n\}_{n=1}^{\infty}$. Then, for and bounded and continuous function $F : \mathcal{Q} \to \mathbb{R}$*

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}\big[e^{nF(\boldsymbol{T}_n)}\big] = \lim_{n\to\infty}\frac{1}{n}\log\int e^{nF(Q)}dP_n(Q) \tag{55}$$

$$= \sup_{Q}\Big[F(Q) - I(Q)\Big] \tag{56}$$

where $I(Q) = \sup_{\tilde{Q}}(\mathrm{tr}(\tilde{Q}Q) - \log\rho(P_{\tilde{Q}}))$ and $\rho(P_{\tilde{Q}})$ is the Perron-Frobenius eigenvalue of the matrix $P_{\tilde{Q}} = \{e^{\mathrm{tr}(\tilde{Q}\bar{Q}_j)}P_{\bar{Q}_j|\bar{Q}_i}\}_{0\le i,j\le M}$ and $M = |\mathcal{Q}|-1$, where $\mathcal{Q}$ and $\{\bar{Q}_i\}_{i=0}^{M}$ are defined in Subsection 1.4.

$$P(Q|Q') = \frac{\sum_{(s,x_0,x_1,\cdots,x_\nu,s',x'_0,x'_1,\cdots,x'_\nu)\in\mathcal{A}_Q\times\mathcal{A}_{Q'}} P_S(s')P_S(s)p_{X_{n-1}}(x'_0)\pi(x'_0,x_0)\prod_{i=1}^{\nu} p_{X_{n-1}}(x'_i)\pi(x'_i,x_i)}{\sum_{(s',x'_0,x'_1,\cdots,x'_\nu)\in\mathcal{A}_{Q'}} P_S(s')p_{X_{n-1}}(x'_0)\prod_{i=1}^{\nu} p_{X_{n-1}}(x'_i)} \quad (53)$$

Lemma 6. *Let $\{\bar{Q}_i\}_{i=0}^M$ be states of the Markov chain $\{\boldsymbol{Q}_n\}_{n=1}^\infty$ in Lemma 4. Then, the following holds:*

$$\frac{\partial \log\rho(P_{\tilde{Q}})}{\partial\tilde{Q}}(\tilde{Q}) = \frac{1}{\rho(P_{\tilde{Q}})}\sum_{i=0}^M \lambda_i(\tilde{Q})\sum_{j=0}^M \psi_j(\tilde{Q})\bar{Q}_j P(\bar{Q}_j|\bar{Q}_i)e^{\mathrm{tr}(\tilde{Q}\bar{Q}_j)}, \quad (57)$$

where $\lambda(\tilde{Q})$ and $\psi(\tilde{Q})$ are left and right eigenvectors associated with the Perron-Frobenius eigenvalue $\rho(P_{\tilde{Q}})$ which are normalized such that $\lambda(\tilde{Q})^T\psi(\tilde{Q}) = 1$.

Theorem 7. *Recall the definition of $G^{(\nu)}(Q)$ in Lemma 3. In the large system limit, the free energy satisfies:*

$$\mathcal{F} = -\lim_{\nu\to 0}\frac{\partial}{\partial\nu}\sup_Q\Big[\beta^{-1}G^{(\nu)}(Q) - I^{(\nu)}(Q)\Big], \quad (58)$$

where

$$I^{(\nu)}(Q) := \sup_{\tilde{Q}}\Big[\mathrm{tr}(\tilde{Q}Q) - \log\rho(P_{\tilde{Q}})\Big], \quad (59)$$

and $\rho(P_{\tilde{Q}})$ is the Perron-Frobenius eigenvalue of the matrix $P_{\tilde{Q}} = \{e^{\mathrm{tr}(\tilde{Q}\bar{Q}_j)}P_{\bar{Q}_j|\bar{Q}_i}\}_{0\le i,j\le M}$ and $M = |\mathcal{Q}| - 1$ where $\mathcal{Q} := \{sxx^T \text{ for some } s\in\mathcal{S},\ x\in\mathcal{X}^{\nu+1}\}$.

Proof. The proof follows the same idea as (Guo and Verdu, 2005, Part A, Sect. IV) with some important changes to account for the Markov setting.

1. By applying Lemma 5, from (47), we obtain

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}[Z^\nu(\boldsymbol{Y},\boldsymbol{\Phi})] = \lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}\Big\{\exp\Big[\frac{n}{\beta}\Big(G^{(\nu)}(\boldsymbol{T}_n) + O(n^{-1})\Big)\Big]\Big\} \quad (60)$$

$$= \sup_Q\Big[\frac{1}{\beta}G^{(\nu)}(Q) - I^{(\nu)}(Q)\Big]. \quad (61)$$

2. Estimate the free energy.

Now, observe that

$$\mathcal{F} = -\lim_{n\to\infty}\frac{1}{n}\lim_{\nu\to 0}\frac{\partial}{\partial\nu}\log\mathbb{E}[Z^\nu(\boldsymbol{Y},\boldsymbol{\Phi})] \quad (62)$$

$$= -\lim_{\nu\to 0}\frac{\partial}{\partial\nu}\lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}[Z^\nu(\boldsymbol{Y},\boldsymbol{\Phi})] \quad (63)$$

$$= -\lim_{\nu\to 0}\frac{\partial}{\partial\nu}\sup_Q\Big[\frac{1}{\beta}G^{(\nu)}(Q) - I^{(\nu)}(Q)\Big], \quad (64)$$

where (62) follows from the assumption (A1), (A2), and the fact (F1), (63) follows from the assumption (A3), and (64) follows from (61).

□

Theorem 8. *Recall the definitions of Σ and the matrix $P_{\tilde{Q}}$ in Theorem 7 and the definitions of $\{\bar{Q}_i\}_{i=0}^M$ in Lemma 6. The optimal matrix Q^* of equation (58) in Theorem 7 must satisfy the following constraints:*

$$Q^* = \frac{\partial\log\rho(P_{\tilde{Q}})}{\partial\tilde{Q}}(\tilde{Q}^*), \quad (65)$$

$$\tilde{Q}^* = -(2\beta)^{-1}(I+\Sigma Q^*)^{-1}\Sigma, \quad (66)$$

$$\frac{\partial\log\rho(P_{\tilde{Q}})}{\partial\tilde{Q}}(\tilde{Q}^*) = \frac{1}{\rho(P_{\tilde{Q}^*})}\sum_{i=0}^M\lambda_i(\tilde{Q}^*)\sum_{j=0}^M\psi_j(\tilde{Q}^*)\bar{Q}_jP(\bar{Q}_j|\bar{Q}_i)e^{\mathrm{tr}(\tilde{Q}^*\bar{Q}_j)}, \quad (67)$$

where $\lambda(\tilde{Q}^*)$ and $\psi(\tilde{Q}^*)$ are left and right eigenvectors associated with the Perron-Frobenius eigenvalue $\rho(P_{\tilde{Q}^*})$ which are normalized such that $\lambda(\tilde{Q}^*)^T\psi(\tilde{Q}^*) = 1$.

Proof. Recall the definition of $I^{(\nu)}$ in Theorem 7. It is easy to see that the optimization problem in (61) is equivalent to the following optimization problem:

$$\sup_Q\inf_{\tilde{Q}} T^{(\nu)}(Q,\tilde{Q}) \quad (68)$$

where

$$T^{(\nu)}(Q,\tilde{Q}) := -\frac{1}{2\beta}\log\det(I+\Sigma Q) - \mathrm{tr}(\tilde{Q}Q) + \log\rho(P_{\tilde{Q}}) - \frac{1}{2\beta}\log\big(1+\frac{\nu}{\sigma^2}\big) - \frac{\nu}{2\beta}\log(2\pi\sigma^2). \quad (69)$$

For an arbitrary Q, we first seek critical points with respect to $\tilde{Q}$ and find that for any given Q, the extremum in $\tilde{Q}$ satisfies

$$Q = \frac{\partial \log \rho(P_{\tilde{Q}})}{\partial \tilde{Q}} \tag{70}$$

Let $\tilde{Q}(Q)$ be a solution to (70). We then seek the critical point of $T^{(\nu)}(Q, \tilde{Q}(Q))$ with respect to Q.

Let

$$K_{Q,\tilde{Q}} := \left[\frac{\partial \tilde{Q}_{a,b}}{\partial Q_{a,b}}\right]_{a,b=0}^{\nu} \in \mathbb{R}^{\nu+1\times\nu+1}. \tag{71}$$

Observe that

$$\frac{\partial \mathrm{tr}(\tilde{Q}Q)}{\partial Q} = \frac{\partial \mathrm{tr}(Q\tilde{Q})}{\partial Q} \tag{72}$$

$$= \tilde{Q} + Q \odot K_{Q,\tilde{Q}}, \tag{73}$$

where $\odot$ is the Hadamard product.

It follows that

$$\frac{\partial T^{(\nu)}(Q,\tilde{Q})}{\partial Q}$$
$$= -\frac{1}{2\beta}(I+\Sigma Q)^{-1}\Sigma - \left(\tilde{Q} + Q\odot K_{Q,\tilde{Q}}\right) + \frac{\partial \log\rho(P_{\tilde{Q}})}{\partial Q} \tag{74}$$
$$= -\frac{1}{2\beta}(I+\Sigma Q)^{-1}\Sigma - \left(\tilde{Q} + Q\odot K_{Q,\tilde{Q}}\right) + \frac{\partial \log\rho(P_{\tilde{Q}})}{\partial \tilde{Q}} \odot K_{Q,\tilde{Q}} \tag{75}$$
$$= -\frac{1}{2\beta}(I+\Sigma Q)^{-1}\Sigma - \tilde{Q} - \left[Q - \frac{\partial \log\rho(P_{\tilde{Q}})}{\partial \tilde{Q}}\right]\odot K_{Q,\tilde{Q}} \tag{76}$$
$$= -\frac{1}{2\beta}(I+\Sigma Q)^{-1}\Sigma - \tilde{Q}, \tag{77}$$

where (74) follows from (73), and (77) follows from (70). Hence, the optimal value of the Theorem 7 is the solution of the following equation systems:

$$Q = \frac{\partial \log \rho(P_{\tilde{Q}})}{\partial \tilde{Q}}, \tag{78}$$

$$\tilde{Q} = -(2\beta)^{-1}(I+\Sigma Q)^{-1}\Sigma. \tag{79}$$

Finally, from Lemma 6, we also obtain an additional constraint in (67). $\square$

Observe that the matrix Σ defined in Theorem 7 is invariant if two non-zero indices are interchanged, i.e., Σ is symmetric in replicas. Now, we use the RS assumption (A4) to simplify the result in Theorem 7. More specifically, we use the following RS assumption:

Definition 9. *(Guo and Verdu, 2005, p. 1999) An solution $(\tilde{Q}^*, Q^*)$ of the optimization problem in Theorem 7, i.e.,*

$$\sup_Q \left[\beta^{-1}G^{(\nu)}(Q) - I^{(\nu)}(Q)\right]$$
$$= \sup_Q \inf_{\tilde{Q}} \bigg[-\frac{1}{2\beta}\log\det(I+\Sigma Q) - \mathrm{tr}(\tilde{Q}Q) + \log\rho(P_{\tilde{Q}}) - \frac{1}{2\beta}\log\left(1+\frac{\nu}{\sigma^2}\right) - \frac{\nu}{2\beta}\log(2\pi\sigma^2)\bigg] \tag{80}$$

is called to satisfy the Replica Symmetry (RS) if both Q^* and $\tilde{Q}^*$ are invariant if two (nonzero) replica indices are interchanged. In other words, the extremum can be written as

$$Q^* = \begin{bmatrix} r & m & m & \cdots & m \\ m & p & q & \cdots & q \\ m & q & p & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & q \\ m & q & \cdots & q & p \end{bmatrix}, \tag{81}$$

$$\tilde{Q}^* = \begin{bmatrix} c & d & d & \cdots & d \\ d & g & f & \cdots & f \\ d & f & g & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & q \\ d & f & \cdots & f & g \end{bmatrix}, \tag{82}$$

where r, m, p, q, c, d, f, g are some real numbers which are not dependent on ν.

First, we show the following auxiliary results:

Lemma 10. *Let $\{\bar{Q}_i\}_{i=0}^M$ be states of the Markov chain $\{\boldsymbol{Q}_n\}_{n=1}^\infty$ in Lemma 4. Assume that*

$$\rho(P_{\tilde{Q}^*}) \to 1 \quad \text{and} \quad \sum_{j=0}^M P(\bar{Q}_j|\bar{Q}_i)e^{\mathrm{tr}(\tilde{Q}^*\bar{Q}_j)} \to 1 \tag{83}$$

for all $i \in [M]$ as $\nu \to 0$. Then, under the replica symmetry assumptions in Definition 9, the following holds:

$$Q^* = \lim_{\nu\to 0}\sum_{i=0}^M \lambda_i(\tilde{Q}^*)\mathbb{E}[\boldsymbol{Q}_1 e^{\mathrm{tr}(\tilde{Q}^*\boldsymbol{Q}_1)}\big|\boldsymbol{Q}_0 = \bar{Q}_i] \tag{84}$$

where Q^* is defined in Theorem 8 and $\lambda(\tilde{Q}^*)$ is a left (positive) eigenvector associated with the Perron-Frobenius eigenvalue $\rho(P_{\tilde{Q}^*})$ such that $\|\lambda(\tilde{Q}^*)\|_1 = 1$. In addition, we have

$$\rho(P_{\tilde{Q}^*}) = \sum_{i=1}^M \lambda_i(\tilde{Q}^*)\mathbb{E}[e^{\mathrm{tr}(\tilde{Q}^*\mathrm{Q}_1)}|\boldsymbol{Q}_0 = \bar{Q}_i]. \tag{85}$$

Proof. Since $\psi(\tilde{Q}^*)$ is the right eigenvector associated with the Perron-Frobenius eigenvalue of the matrix $P_{\tilde{Q}^*}$, it holds that

$$\sum_{j=0}^{M} P(\bar{Q}_j|\bar{Q}_i)e^{\mathrm{tr}(\tilde{Q}^*\bar{Q}_j)}\psi_j(\tilde{Q}^*) = \rho(P_{\tilde{Q}^*})\psi_i(\tilde{Q}^*) \quad (86)$$

for all $i \in [M]$. From (86) and (83), we can set $\psi(\tilde{Q}^*) = (1, 1, \cdots, 1)^T$ is a right eigenvector associated with the eigenvalue $\rho(P_{\tilde{Q}^*})$ as $\nu \to 0$.

Hence, from Theorem 8, we have

$$Q^* = \lim_{\nu\to 0} \sum_{i=0}^{M} \lambda_i(\tilde{Q}^*)\mathbb{E}[\boldsymbol{Q}_1 e^{\mathrm{tr}(\tilde{Q}^*\boldsymbol{Q}_1)}\big|\boldsymbol{Q}_0 = \bar{Q}_i]. \quad (87)$$

Now, since by Theorem 8, it holds that

$$\sum_{j=0}^{M} \psi_j(\tilde{Q}^*)\lambda_j(\tilde{Q}^*) = 1, \quad (88)$$

so we have

$$\|\lambda(\tilde{Q}^*)\|_1 = 1. \quad (89)$$

Now, since $\lambda(\tilde{Q}^*) := (\lambda_1(\tilde{Q}^*), \lambda_2(\tilde{Q}^*), \cdots, \lambda_M(\tilde{Q}^*))$ is the left (positive) eigenvector associated with $\rho(P_{\tilde{Q}^*})$, it holds that

$$\lambda_j(\tilde{Q}^*)\rho(P_{\tilde{Q}^*}) = \sum_{i=1}^{M} \lambda_i(\tilde{Q}^*)e^{\mathrm{tr}(\tilde{Q}^*\bar{Q}_j)}P(\bar{Q}_j|\bar{Q}_i). \quad (90)$$

Then, it follows that

$$\begin{aligned}
\rho(P_{\tilde{Q}^*}) &= \sum_{j=1}^{M} \lambda_j(\tilde{Q}^*)\rho(P_{\tilde{Q}^*}) && (91)\\
&= \sum_{j=1}^{M}\sum_{i=1}^{M} \lambda_i(\tilde{Q}^*)e^{\mathrm{tr}(\tilde{Q}^*\bar{Q}_j)}P(\bar{Q}_j|\bar{Q}_i) && (92)\\
&= \sum_{i=1}^{M} \lambda_i(\tilde{Q}^*)\sum_{j=1}^{M} e^{\mathrm{tr}(\tilde{Q}^*\bar{Q}_j)}P(\bar{Q}_j|\bar{Q}_i) && (93)\\
&= \sum_{i=1}^{M} \lambda_i(\tilde{Q}^*)\mathbb{E}[e^{\mathrm{tr}(\tilde{Q}^*Q_1)}|\boldsymbol{Q}_0 = \bar{Q}_i], && (94)
\end{aligned}$$

where (91) follows from (89), (92) follows from (90). ☐

Lemma 11. *Under the RS assumption in Definition 9, as $\nu \to 0$, the following hold:*

$$\rho(P_{\tilde{Q}^*}) \to 1 \quad \textit{and} \quad \sum_{j=0}^{M} P(\bar{Q}_j|\bar{Q}_i)e^{\mathrm{tr}(\tilde{Q}^*\bar{Q}_j)} \to 1. \quad (95)$$

Furthermore, it holds that

$$\begin{aligned}
&\frac{\partial \log\rho(P_{\tilde{Q}^*})}{\partial\nu}\Big|_{\nu=0}\\
&= -\frac{\xi}{2}\left(\mathbb{E}[S]\mathbb{E}_{\mathsf{X}_0\sim\lambda^{(\pi)}}\Big[\mathbb{E}[\mathsf{X}_1^2|\mathsf{X}_0]\Big] + \frac{1}{\eta}\right)\log e + \frac{1}{2}\log\frac{2\pi}{\xi}\\
&\quad + \mathbb{E}_{\mathsf{X}_0\sim\lambda^{(\pi)}}\Big[\mathbb{E}_S\Big\{\int_{\mathbb{R}} p_{U|\mathsf{X}_0,S;\eta}(u|x_0,S;\eta)\\
&\quad \times \log q_{U|\mathsf{X}_0,S;\eta}(u|x_0,S;\eta)du\Big\}\Big]. \quad (96)
\end{aligned}$$

Then, we obtain our first main result as follows.

Theorem 12. *The free energy of the linear model with Markov sources in Section 2 satisfies*

$$\mathcal{F} = \mathcal{G}, \quad (97)$$

where $\mathcal{G}$ is defined in (27). In addition, the average mutual information of this model satisfies:

$$C = \lim_{n\to\infty}\frac{1}{n}I(\boldsymbol{X}^n;\boldsymbol{Y}^m) = \mathcal{F} - \frac{1}{2\beta}. \quad (98)$$

The following corollary recovers (Guo and Verdu, 2005, Sect. II-D):

Corollary 13. *For any i.i.d. sequence $\{X_n\}_{n=1}^\infty$ on the Polish space $\mathcal{X}$ defined in Section 2, the free energy satisfies*

$$\mathcal{F} = \mathcal{G}(\emptyset), \quad (99)$$

where $\mathcal{G}(\emptyset)$ is the free-energy function estimated in Section 3.1 when no state information appears in the corresponding single-symbol MMSE channel.

In addition, the average mutual information of this model satisfies

$$C = \lim_{n\to\infty}\frac{1}{n}I(\boldsymbol{X}^n;\boldsymbol{Y}^m) = \mathcal{F} - \frac{1}{2\beta}. \quad (100)$$

Proof. Observe that an i.i.d. sequence $\{X_n\}_{n=1}^\infty$ can be considered as a Markov sequence with transition probability (function) $\pi(x,y) = p(y)$ for all $x, y \in \mathcal{X}$. Hence, $\mathcal{G}(x_0)$ is a constant, say $\mathcal{G}(\emptyset)$, for all $x_0 \in \mathcal{X}$. Here, $\mathcal{G}(\emptyset)$ is the free energy function estimated in Section 3.1 when there is no state information appeared in the correponding single-symbol MMSE channel, i.e. $X_0 = \emptyset$. In addition, the left Perron-Frobenius with unit Manhattan norm for this special case is $\{P_{X_1}(x)\}_{x\in\mathcal{X}}$.

Hence, by Theorem 12, we have

$$\begin{aligned}
\mathcal{F} &= \sum_{x_0\in\mathcal{X}} \lambda_{x_0}^{(\pi)}\mathcal{G}(x_0) && (101)\\
&= \Big(\sum_{x_0\in\mathcal{X}} P_{X_1}(x_0)\Big)\mathcal{G}(\emptyset) && (102)\\
&= \mathcal{G}(\emptyset), && (103)
\end{aligned}$$

where the last equation follows from $\|\lambda^{(\pi)}\|_1 = 1$. Hence, we obtain (99). Finally, (100) is an direct application of (98) in Theorem 12. ☐

To state our next main result, we recall Carleman theorem.

Lemma 14. *(Chalendar and Partington, 2007, Theorem 3.1) Denote $\mathcal{M}(\mathbb{R}^n)$ be the set of all positive Borel measures μ on $\mathbb{R}^n$ such that*

$$\int_{\mathbb{R}^n} \|x\|_2^d d\mu(x) < \infty \quad \forall d \geq 0. \tag{104}$$

Suppose that $\mu_1, \mu_2 \in \mathcal{M}(\mathbb{R}^n)$ satisfy

$$s(\alpha) := \int_{\mathbb{R}^n} x^\alpha d\mu_1(x) = \int_{\mathbb{R}^n} x^\alpha d\mu_2(x) \quad \text{for all} \quad \alpha \in \mathbb{N}^n \tag{105}$$

and that the conditions

$$\sum_{m=1}^{\infty} s(2me_j)^{-1/(2m)} = \infty, \quad j = 1, 2, \cdots, n, \tag{106}$$

hold, where e_j is the jth canonical basis vector of $\mathbb{R}^n$. Then $\mu_1 = \mu_2$.

Claim 15. *Recall the definition of $\{\lambda_{x_0}^{(\pi)}\}_{x_0 \in \mathcal{X}}$ in Section 3.1. Assume that the MMSE estimator defined in (9) is used for estimation. Then, for all $k \in \{1, 2, \cdots, n\}$, the joint moments satisfy:*

$$\lim_{n \to \infty} \mathbb{E}\Big[X_k^{i_0} \tilde{X}_k^{j_0} \langle X_k \rangle_p^{l_0}\Big] = \sum_{x_0 \in \mathcal{X}} \lambda_{x_0}^{(\pi)} \mathbb{E}\Big[\mathsf{X}_1^{i_0} \mathsf{X}^{j_0} \langle \mathsf{X} | \mathsf{X}_0 \rangle_p^{l_0} \big| \mathsf{X}_0 = x_0\Big], \tag{107}$$

for all $i_0, j_0, l_0 \in \mathbb{Z}_+$, where $(\mathsf{X}_1, \mathsf{X}, \langle \mathsf{X} | \mathsf{X}_0 = x_0 \rangle_p)$ is the input and outputs defined in the (composite) single-symbol MMSE channel in Fig. 1, and $(X_k, \tilde{X}_k, \langle X_k \rangle)$ is the k-th symbol in the vector $\boldsymbol{X} \in \mathcal{X}^n$, the k-th output of the vector retrochanel defined in (5), and its corresponding estimated symbol by using the MMSE estimator in (9), respectively.

In addition, the average MMSE satisfies:

$$\frac{1}{n}\mathbb{E}\big[\|\boldsymbol{X} - \langle \boldsymbol{X} \rangle\|_2^2\big] = \mathbb{E}\big[\mathsf{X}_1^2\big] - \sum_{x_0 \in \mathcal{X}} \lambda_{x_0}^{(\pi)} \mathbb{E}\big[\langle \mathsf{X}_1 | \mathsf{X}_0 \rangle^2 \big| \mathsf{X}_0 = \mathsf{x}_0\big], \tag{108}$$

where $\mathsf{X}_1 \sim \sum_{x_0 \in \mathcal{X}} \pi(x_0, \cdot) p(x_0)$.

Remark 16. *Some remarks are in order.*

- *For the i.i.d. sequence $\{X_n\}_{n=1}^\infty$, we have a tight bound in (107). It is not hard to check that the Carleman condition (106) holds for the joint Gaussian distribution on the composite single-symbol Gaussian channel in Fig. 1. Hence, from Carleman Theorem in Lemma 14, in the large system limit, the channel between the input X_k and $\langle X_k \rangle_p$ for each symbol k is equivalent to the Gaussian channel $p_{U|\mathsf{X},\mathsf{X}_0,S;\eta}$ with available state $\mathsf{X}_0 = \emptyset$ at both encoder and decoder concatenated with the one-to-one decision function with $S = S_k$. This result recovers (Guo and Verdu, 2005, Corrolary 1) as a special case for the i.i.d. sequence $\{X_n\}_{n=1}^\infty$.*
- *From Theorem 15, it can be inferred that under the MMSE estimator, the channel (model) has been decoupled into AWGN channels with state information at both transmitters and receivers, where state vector distribution follows the left Perron-Frobenius eigenvalue $\lambda^{(\pi)}$ of the stochastic matrix P_π.*

Proof. Observe that by using the MMSE estimator defined in Section 2.1, we have

$$\mathbb{E}\big[\|\boldsymbol{X} - \langle \boldsymbol{X} \rangle\|_2^2\big] = \sum_{k=1}^n \mathbb{E}\big[\big|X_k - \langle \boldsymbol{X} \rangle_k\big|^2\big] \tag{109}$$

$$= \sum_{k=1}^n \mathbb{E}\big[\big|X_k - \langle X_k \rangle\big|^2\big] \tag{110}$$

$$= \sum_{k=1}^n \mathbb{E}\Big[\mathbb{E}\big[\big|X_k - \langle X_k \rangle\big|^2 \big| \boldsymbol{Y}, \boldsymbol{\Phi}\big]\Big] \tag{111}$$

$$= \sum_{k=1}^n \mathbb{E}\Big[\mathbb{E}\big[X_k^2\big] - \langle X_k \rangle^2 \big| \boldsymbol{Y}, \boldsymbol{\Phi}\Big] \tag{112}$$

$$= \sum_{k=1}^n \mathbb{E}\big[X_k^2\big] - \sum_{k=1}^n \mathbb{E}\big[\langle X_k \rangle^2\big], \tag{113}$$

where (110) follows from (9), (111) follows from the tower property (Billingsley, 1995), and (112) follows from the fact that

$$\langle X_k \rangle = \mathbb{E}\big[X_k \big| \boldsymbol{Y}, \boldsymbol{\Phi}\big] \tag{114}$$

which is drawn from (9).

Now, by (107), we have as $n \to \infty$,

$$\mathbb{E}\big[\langle X_k \rangle^2\big] = \sum_{x_0 \in \mathcal{X}} \lambda_{x_0}^{(\pi)} \mathbb{E}\big[\langle \mathsf{X}_1 | \mathsf{X}_0 = x_0 \rangle^2\big] \tag{115}$$

for all $k \in \{1, 2, \cdots, n\}$.

In addition, for all $k \in \{1, 2, \cdots, n\}$, we also have

$$\begin{aligned} \mathbb{E}\big[X_k^2\big] &= \mathbb{E}\big[\mathbb{E}\big[X_k^2 \big| X_{k-1}\big]\big] && (116) \\ &= \mathbb{E}\big[\mathbb{E}\big[X_1^2 \big| X_0\big]\big] && (117) \\ &= \mathbb{E}[X_1^2] && (118) \\ &= \mathbb{E}[\mathsf{X}_1^2], && (119) \end{aligned}$$

where (116) follows from the tower property (Billingsley, 1995), and (117) follows from the time-homogeneity of Markov process $\{X_n\}_{n=1}^{\infty}$.

From (113), (115), and (118), as $n \to \infty$, we have

$$\mathbb{E}\big[\|\boldsymbol{X} - \langle \boldsymbol{X} \rangle\|_2^2\big] = n\left(\mathbb{E}[\mathsf{X}_1^2] - \sum_{x_0 \in \mathcal{X}} \lambda_{x_0}^{(\pi)} \mathbb{E}\big[\langle \mathsf{X}_1 | \mathsf{X}_0 = x_0 \rangle^2\big]\right), \tag{120}$$

which leads to (108). □