# New approach to Bayesian high-dimensional linear regression

Shirin Jalali*

*Math and Algorithms Group, Nokia - Bell Labs, 600 Mountain Ave, Murray Hill, NJ 07974, USA*
*Corresponding author: Shirin.Jalali@nokia-bell-labs.com

and

Arian Maleki

*Department of Statistics, Columbia University, 1255 Amsterdam Ave, New York, NY 10027, USA*
arian@stat.columbia.edu

Consider the problem of estimating parameters $X^n \in \mathbb{R}^n$, from $m$ response variables $Y^m = AX^n + Z^m$, under the assumption that the distribution of $X^n$ is known. Lack of computationally feasible algorithms that employ generic prior distributions and provide a good estimate of $X^n$ has limited the set of distributions researchers use to model the data. To address this challenge, in this article, a new estimation scheme named quantized maximum *a posteriori* (Q-MAP) is proposed. The new method has the following properties: (i) In the noiseless setting, it has similarities to maximum *a posteriori* (MAP) estimation. (ii) In the noiseless setting, when $X_1, \ldots, X_n$ are independent and identically distributed, asymptotically, as $n$ grows to infinity, its required sampling rate ($m/n$) for an almost zero-distortion recovery approaches the fundamental limits. (iii) It scales favorably with the dimensions of the problem and therefore is applicable to high-dimensional setups. (iv) The solution of the Q-MAP optimization can be found via a proposed iterative algorithm that is provably robust to error (noise) in response variables.

*Keywords*: linear regression, Bayesian estimation, information dimension, compressed sensing.

## 1. Introduction

Consider the problem of Bayesian linear regression defined as follows; $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$ denotes a stationary random process whose distribution is known. The goal is to estimate $X^n$ from $m$ response variables of the form $Y^m = AX^n + Z^m$, where $A \in \mathbb{R}^{m \times n}$ and $Z^m \in \mathbb{R}^m$ denote the design matrix and the noise vector, respectively. (Both cases of $m < n$ and $m \geq n$ are of interest and valid under our model.) To solve this problem, there are two fundamental questions that can be raised:

1. How should we use the distribution of $X^n$ to obtain an efficient estimator? To answer this question, there are two main criteria that should be taken into account: (i) computational complexity: how efficiently can the estimate be computed? (ii) accuracy: how well does the estimator perform? If we ignore the computational complexity constraint, then the answer to our first question is simple. An optimal Bayes estimator seeks to minimize the Bayes risk defined as $\mathrm{E}[\ell(\hat{X}^n, X^n)]$, where $\ell : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$ denotes the considered loss function. For instance, $\ell(x^n, \hat{x}^n) = \left\| x^n - \hat{x}^n \right\|_2^2$ leads to the minimum mean square error (MMSE) estimator. However, the computational complexity of MMSE estimator for generic distributions is very high.

2. Can the performance of the estimator be analyzed in high-dimensional settings? The answer to this question is also complicated. Even the performance analysis of standard estimators such as

MMSE estimator is challenging. In fact, even if we assume that **X** is an independent and identically distributed (i.i.d.) process, the analysis is still complicated and heuristic tools such as replica method from statistical physics have been employed to achieve this goal.

In response to the above two questions, we propose an optimization problem, which we refer to as quantized maximum *a posteriori* (Q-MAP). We then show how this optimization problem can be analyzed and solved. Before presenting the Q-MAP optimization, we introduce some notation.

### 1.1 *Notation*

Calligraphic letters such as $\mathcal{X}$ and $\mathcal{Y}$ denote sets. The size of a set $\mathcal{X}$ is denoted by $|\mathcal{X}|$. Given vector $(v_1, v_2, \ldots, v_n) \in \mathbb{R}^n$ and integers $i, j \in \{1, \ldots, n\}$, where $i \leq j$, $v_i^j \triangleq (v_i, v_{i+1}, \ldots, v_j)$. For simplicity $v_1^j$ and $v_j^j$ are denoted by $v^j$ and $v_j$, respectively. For two vectors $u^n$ and $v^n$, both in $\mathbb{R}^n$, let $\langle u^n, v^n \rangle$ denote their inner product defined as $\langle u^n, v^n \rangle \triangleq \sum_{i=1}^n u_i v_i$. The all-zero and all-one vectors of length $n$ are denoted by $0^n$ and $1^n$, respectively.

Uppercase letters such as $X$ and $Y$ denote random variables. The alphabet of a random variable $X$ is denoted by $\mathcal{X}$. The entropy of a finite-alphabet random variable $U$ with probability mass function (pmf) $p(u)$, $u \in \mathcal{U}$, is defined as $H(U) = -\sum_{u \in \mathcal{U}} p(u) \log p(u)$ [6]. Given finite-alphabet random variables $U$ and $V$ with joint pmf $p(u, v)$, $(u, v) \in \mathcal{U} \times \mathcal{V}$, the conditional entropy of $U$ given $V$ is defined as $H(U|V) = -\sum_{(u,v) \in \mathcal{U} \times \mathcal{V}} p(u, v) \log p(u|v)$.

Matrices are also denoted by uppercase letters such as $A$ and $B$ and are differentiated from random variables by context. Throughout the article log and ln refer to logarithm in base 2 and the natural logarithm, respectively.

For $x \in \mathbb{R}$, $\lfloor x \rfloor$ denotes the largest integer smaller than or equal to $x$. Therefore, $0 \leq x - \lfloor x \rfloor < 1$, for all $x$. The *b*-bit quantized version of $x$ is denoted by $[x]_b$ and is defined as

$$[x]_b \triangleq \lfloor x \rfloor + \sum_{i=1}^b 2^{-i} a_i, \tag{1.1}$$

where for all $i$, $a_i \in \{0, 1\}$, and $0.a_1 a_2 \ldots$ denotes the binary representation of $x - \lfloor x \rfloor$. When $x - \lfloor x \rfloor$ is a dyadic real number, which have two possible binary representations, let $0.a_1 a_2 \ldots$ denote the representation which has a finite number of ones. For a vector $x^n \in \mathbb{R}^n$, $[x^n]_b \triangleq ([x_1]_b, \ldots, [x_n]_b)$.

Consider a vector $u^n \in \mathcal{U}^n$, where $|\mathcal{U}| < \infty$. The $(k + 1)$th order empirical distribution of $u^n$ is denoted by $\hat{p}^{(k+1)}$ and is defined as follows: for any $a^{k+1} \in \mathcal{U}^{k+1}$,

$$\hat{p}^{(k+1)}(a^{k+1}|u^n) \triangleq \frac{|\{i : u_{i-k}^i = a^{k+1}, k + 1 \leq i \leq n\}|}{n - k}$$

$$= \frac{1}{n - k} \sum_{i=k+1}^n 1_{u_{i-k}^i = a^{k+1}}, \tag{1.2}$$

where $1_{\mathcal{E}}$ denotes the indicator function of event $\mathcal{E}$. In other words, $\hat{p}^{(k+1)}(.|u^n)$ denotes the distribution of a randomly selected length-$(k + 1)$ substring of $u^n$.

### 1.2 *Contributions*

Consider the stochastic process $\mathbf{X} = \{X_i\}_{i=1}^\infty$ discussed earlier. Let $X_i \in \mathcal{X}$, for all $i$. Assume that $\mathcal{X}$ is a bounded subset of $\mathbb{R}$. Define the *b*-bit quantized version of $\mathcal{X}$ as

$$\mathcal{X}_b \triangleq \{[x]_b : x \in \mathcal{X}\}. \tag{1.3}$$

Note that since $\mathcal{X}$ is a bounded set, $\mathcal{X}_b$ is a finite set, i.e. $|\mathcal{X}_b| < \infty$. Given $k \in \mathbb{N}^+$ and a set of non-negative weights $\mathbf{w} = (w_{a^{k+1}} : a^{k+1} \in \mathcal{X}_b^{k+1})$, define function $c_{\mathbf{w}} : \mathbf{X}_b^n \to \mathbb{R}$ as follows. For $u^n \in \mathcal{X}_b^n$,

$$c_{\mathbf{w}}(u^n) \triangleq \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} \hat{p}^{(k+1)}(a^{k+1}|u^n). \tag{1.4}$$

As it will be described later, for a proper choice of weights $\mathbf{w}$, this function measures both the level of 'structuredness' of sequences in $\mathcal{X}_b^n$ and also how closely they match the $(k+1)^{\text{th}}$ order distribution of the process $\mathbf{X}$. For instance, as will be shown later in Section 3.1, for an i.i.d. process $\mathbf{X}$ with $X_i \sim (1-p)\delta_0 + pf_c$, where $p \in (0,1)$, $\delta_0$ denotes a point mass at zero, and $f_c$ denotes an absolutely continuous distribution over a bounded set, a bound of the form $c_{\mathbf{w}}(u^n) \leq \gamma$, with $k = 0$, simplifies to a bound on the $\ell_0$-norm of sequence $u^n$. Hence, intuitively speaking, the constraint $c_{\mathbf{w}}(u^n) \leq \gamma$ holds for 'structured' sequences that comply with the known source model.

For a stationary process $\mathbf{X}$, the Q-MAP optimization estimates $X^n$ from noisy linear measurements $Y^m = AX^n + Z^m$, by solving the following optimization:

$$\hat{X}^n = \underset{u^n \in \mathcal{X}_b^n}{\arg\min} \quad \|Au^n - Y^m\|^2$$

$$\text{subject to} \quad c_{\mathbf{w}}(u^n) \leq \gamma_n, \tag{1.5}$$

where $\gamma_n$ is a number that may depend on $n$ and the distribution of $\mathbf{X}$, $b$ and $k$ are parameters that need to be set properly, and the non-negative weights $\mathbf{w} = (w_{a^{k+1}} : a^{k+1} \in \mathcal{X}_b^{k+1})$ are defined as a function of the $(k+1)$th order marginal distribution of stationary process $\mathbf{X}$ as:

$$w_{a^{k+1}} \triangleq \log \frac{1}{\mathrm{P}([X_{k+1}]_b = a_{k+1}|[X^k]_b = a^k)}. \tag{1.6}$$

For this specific choice of $\mathbf{w}$, the function $c_{\mathbf{w}}$ can also be written as

$$c_{\mathbf{w}}(u^n) = \sum_{a^k} \hat{q}_k(a^k) D_{\mathrm{KL}}(\hat{q}_{k+1}(\cdot|a^k)\|q_{k+1}(\cdot|a^k)) + \hat{H}_k(u^n), \tag{1.7}$$

where $\hat{q}_k$, $\hat{q}_{k+1}$ and $q_{k+1}$ denote the $k$th order empirical distribution induced by $u^n$, the $(k+1)$th order empirical distribution induced by $u^n$ and the distribution of $[X^{k+1}]_b$, respectively. (The derivation of (1.7) can be found in (7.92), as part of the proof of Theorem 5.1.) This alternative expression reveals how $c_{\mathbf{w}}(u^n)$ captures both the structuredness of $u^n$ through $\hat{H}_k(u^n)$ and the similarity between the empirical distribution (type) of $u^n$ and the distribution of the quantized source process $\mathbf{X}$ through $\sum_{a^k} \hat{q}_k(a^k) D_{\mathrm{KL}}(\hat{q}_{k+1}(\cdot|a^k)\|q_{k+1}(\cdot|a^k))$.

REMARK 1.1 To better understand the weights $\mathbf{w}$ specified in (1.6), consider an i.i.d. sparse process $\mathbf{X}$ distributed as $pf_c + (1-p)\delta_0$. Here, $p \in [0,1]$ and $f_c$ denotes an absolutely continuous distribution over a bounded interval $(x_1, x_2)$, where $x_1 < x_2$. In that case, since the process is memoryless, it is enough to study the weights for $k = 0$. (Note that $\mathrm{P}([X_{k+1}]_b = a_{k+1}|[X^k]_b = a^k) = \mathrm{P}([X_{k+1}]_b = a_{k+1})$, for every

$a^{k+1} \in \mathcal{X}_b^{k+1}$.) For $k = 0$, for $a \in \mathcal{X}_b$, $w_a = \log \frac{1}{P([X_1]_b = a)}$. On the other hand, it is straightforward to check that for $b$ large, $P([X_1]_b = 0) \approx 1 - p$. For $a \in \mathcal{X}_b$, $a \neq 0$, by the mean value theorem,

$$P([X_1]_b = a) = 2^{-b} f_c(x_a),$$

for some $x_a \in (a, a + 2^{-b}]$. Therefore, $w_0 \approx -\log(1 - p)$, and, for $a \in \mathcal{X}_b$, $a \neq 0$,

$$w_a = b - \log f_c(x_a).$$

It can be observed that, as $b$ grows, all weights, except $w_0$, grow to infinity, at a rate linear in $b$. (This issue and its implications are discussed in further details in Section 3.)

Note that minimizing $\|Au^n - Y^m\|^2$ is natural since we would like to obtain a parameter vector that matches the response variables. However, with no constraint on potential solution vectors, the estimate will suffer from overfitting (unless $m$ is much larger than $n$). Hence, some constraints should be imposed on the set of potential solutions. In Q-MAP optimization, this constraint requires a potential solution $u^n \in \mathcal{X}_n^n$ to satisfy

$$c_{\mathbf{w}}(u^n) \leq \gamma_n. \tag{1.8}$$

There are two other features of the above optimization that are worth emphasis and clarification at this point:

1. Quantized reconstructions: While the parameter vector $X^n$ and the response variables $Y^m$ are typically real valued, the estimate produced by the Q-MAP optimization lies in the quantized space $\mathcal{X}_b^n$. The motivation for this quantization will be explained later, but in a nutshell, this step helps both the theoretical analysis and also the implementation of the optimization.

2. Memory parameter ($k$): again both for the convenience of the theoretical analysis and also for the ease of implementation, only dependencies captured by the $(k + 1)$th order probability distribution of the process $\mathbf{X}$ are taken into account in the Q-MAP optimization. This memory parameter is an free parameter that can be selected based on the source distribution. As shown later, for instance, in the noiseless setting, for an i.i.d. process, $k = 0$ is enough to achieve the fundamental limits in terms of sampling rate ($m/n$).

Although the Q-MAP optimization provides a new approach to Bayesian compressed sensing, it is still not an easy optimization problem. For instance, for the i.i.d. distribution mentioned earlier, $(1 - p)\delta_0 + p f_c$, the constraint becomes equivalent to having an upper bound on $\|u^n\|_0$. This is similar to the notoriously difficult optimal variable selection problem. Hence, despite the fact that the Q-MAP optimization problems might appear simpler than other estimators such as MMSE, in fact it can still be computationally infeasible. However, inspired by the projected gradient descent (PGD) method in convex optimization, we propose the following algorithm to solve the Q-MAP optimization. Define

$$\mathcal{F}_o \triangleq \left\{ u^n \in \mathcal{X}_b^n : c_{\mathbf{w}}(u^n) \leq \gamma_n \right\}, \tag{1.9}$$

where function $c_{\mathbf{w}}$ is defined in (1.4). Note that the set $\mathcal{F}_o$ depends on quantization level $b$, memory parameter $k$, weights $\mathbf{w} = \{w_{a^{k+1}}\}_{a^{k+1} \in \mathcal{X}_b^{k+1}}$ and also parameter $\gamma_n$.

The PGD algorithm generates a sequence of estimates $\hat{X}^n(t)$, $t = 0, 1, \ldots$ of the sequence $X^n$. It starts by setting $\hat{X}^n(0) = 0^n$, and proceeds by updating $\hat{X}^n(t)$, its estimate at time $t$, as follows

$$S^n(t+1) = \hat{X}^n(t) + \mu A^\top (Y^m - A\hat{X}^n(t))$$
$$\hat{X}^n(t+1) = \arg\min_{u^n \in \mathcal{F}_o} \|u^n - S^n(t+1)\|, \tag{1.10}$$

where $\mu$ denotes the step-size and ensures that the algorithm does *not* diverge to infinity. Intuitively, the above procedure, at each step, moves the current estimate toward the $Ax^n = Y^m$ hyperplane and then projects the new estimate to the set of structured vectors. As will be proved later, when $m$ is large enough, in the noiseless setting, the estimates provided by the PGD algorithm converge to $X^n$, with high probability.

The challenging step in running the PGD method is the projection step, which requires finding the closest point to $S^n(t+1)$ in $\mathcal{F}_o$. For some special distributions such as sparse or piecewise-constant discussed in Section 3, the corresponding set $\mathcal{F}_o$ has a special form that simplifies the projection considerably. In general, while projection to a non-convex discrete set can be complicated, as we will discuss in Section 6.2, we believe that because of the special structure of the set $\mathcal{F}_o$, a dynamic programming approach can be used for performing this projection. More specifically, we will explain how a Viterbi algorithm [31] with $2^{bk}$ states and $n$ stages can be used for this purpose. Hence, the complexity of the proposed method for doing the projection task required by the PGD grows linearly in $n$, but exponentially in $kb$. We expect that for 'structured distributions' the scaling with $b$ and $k$ can be improved much beyond this. We will describe our intuition in Section 6.2 but leave the formal exploration of this direction to future research.

The main question we have not addressed yet is how well the Q-MAP optimization and the proposed PGD method recover $X^n$ from $Y^m$. In the next few paragraphs, we informally state our main results regarding the performance of the QMAP optimization and the PGD-based algorithm. Before that, note that, to recover $X^n$ from $m < n$ response variables, intuitively, the process should be of structured. Hence, first, we briefly review a measure of structuredness developed for real-valued stochastic processes. The $k$th order upper information dimension of stationary process $\mathbf{X}$ is defined as

$$\bar{d}_k(\mathbf{X}) \triangleq \limsup_{b \to \infty} \frac{H([X_{k+1}]_b | [X^k]_b)}{b}, \tag{1.11}$$

where $H([X_{k+1}]_b | [X^k]_b)$ denotes the conditional entropy of $[X_{k+1}]_b$ given $[X^k]_b$. Similarly, the lower $k$th order upper information dimension of $\mathbf{X}$ is defined as

$$\underline{d}_k(\mathbf{X}) \triangleq \liminf_{b \to \infty} \frac{H([X_{k+1}]_b | [X^k]_b)}{b}. \tag{1.12}$$

If $\bar{d}_k(\mathbf{X}) = \underline{d}_k(\mathbf{X})$, then the $k$th order information dimension of process $\mathbf{X}$ is defined as $d_k(\mathbf{X}) = \bar{d}_k(\mathbf{X}) = \underline{d}_k(\mathbf{X})$ [15]. For $k = 0$, $\bar{d}_k(\mathbf{X})$ ($\underline{d}_k(\mathbf{X})$) is equal to the upper (lower) Rényi information dimension of $X_1$ [23], which is a well-known measure of structuredness for real-valued random variables or random vectors. It can be proved that for all stationary sources with $H(\lfloor X_1 \rfloor)$, $\bar{d}_k(\mathbf{X}) \leq 1$ and $\underline{d}_k(\mathbf{X}) \leq 1$ [15].

To gain some insight on these definitions, consider an i.i.d. process $\mathbf{X}$ with $X_1 \sim (1 - p)\delta_0(x) + p\mathrm{Unif}(0, 1)$, where Unif denotes a uniform distribution. This is called the spike and slab prior [19]. It can be proved that for this process $d_0(\mathbf{X}) = \bar{d}_0(\mathbf{X}) = \underline{d}_0(\mathbf{X}) = p$ [23]. For general stationary sources with

infinite memory, the limit of $\bar{d}_k(\mathbf{X})$ as $k$ grows to infinity is defined as the upper information dimension of the process $\mathbf{X}$ and is denoted by $\bar{d}_o(\mathbf{X})$. As argued in [15], the information dimension of a process measures its level of structuredness, and is related to the number of response variables required for its accurate recovery. On the basis of these definitions and concepts, we state our results in the following. The exposition of our results is informal and lacks many details. All the details will be clarified later in the article.

**Informal Result 1.** Consider the noiseless setting ($Z^m = 0^m$), and assume that the elements of the design matrix $A$ are i.i.d. Gaussian. Further assume that the process $\mathbf{X}$ satisfies certain mixing conditions, and for a fixed $k$, $\frac{m}{n} > \bar{d}_k(\mathbf{X})$. Then, asymptotically, for a proper quantization level which grows with $n$, the Q-MAP optimization recovers $X^n$ with high probability.

There is an interesting feature of this result that we would like to emphasize here: (i) if $\bar{d}_k(\mathbf{X})$ is strictly smaller than 1, then we can estimate $X^n$ accurately, from $m < n$ response variables. In fact the smaller $\bar{d}_k(\mathbf{X})$, the less response variables are required. In particular, we can consider the spike and slab prior we discussed before that corresponds to sparse parameter vectors that are studied in the literature [1,4]. For this prior $\bar{d}_0(\mathbf{X}) = p$. Hence, as long as $m > np$, asymptotically, the estimate of Q-MAP with $k = 0$ will be accurate. Note that $np$ is in fact the expected number of non-zero elements in $\beta$.

We believe that even an MMSE estimator that employs only the $k$th order distribution of the source cannot recover with a smaller number of response variables. We present some examples that confirm our claim; however, the optimality of the result we obtain above is an open question that we leave for future research. The above result is for Q-MAP that is still computationally complicated. Our next result is about our proposed PGD-based algorithm.

**Informal Result 2.** Consider again the noiseless setting, and assume that the elements of $A$ are i.i.d. Gaussian. If the process $\mathbf{X}$ satisfies certain mixing conditions, and $\frac{m}{n} > 80b\bar{d}_k(\mathbf{X})$, where $k$ is a fixed parameter, then the estimates derived by the PGD algorithm, with high probability, converge to $X^n$. We will also characterize the convergence rate of the PGD-based algorithm and its performance in the presence of an additive white Gaussian noise (AWGN).

Compared with Informal Result 1, the number of response variables required in Informal Result 2 is a factor $80b$ higher. As we will discuss later we let $b$ grow as in $O(\log \log n)$, and hence the difference between Informal Result 1 and Informal Result 2 is not substantial.

### 1.3 *Related work and discussion*

Bayesian linear regression has been the topic of extensive research in the past 50 years [**10–12**,**16**,**17**,**19–21**,**26**,**27**,**30**,**32**]. In all these papers, $X^n$ is considered as a random vector whose distribution is known. However, often simple models are considered for the distribution of $X^n$ to simplify either the posterior calculations or to apply Markov chain Monte Carlo methods such as Gibbs sampling. This article considers a different scenario. We ignore the computational issues at first and consider an arbitrary distribution for $X^n$. This is in particular useful for applications in which complicated prior can be learned. (For instance, one might have access to a large database that has many different draws of process $\mathbf{X}$.) We then present an optimization for estimating $X^n$ and prove the optimality of this approach under some conditions. This approach let us avoid the limitations that are imposed by posterior calculations. On the other hand, one main advantage of having posterior distributions is that they can be used in calculating confidence intervals. Exploring confidence intervals and related topics remains an open question for future research.

Our theoretical analyses are inspired by the recent surge of interest toward understanding the high-dimensional linear regression problem [1,4,5,8,24,28]. In this front, there has been very limited work on the theoretical analysis of the Bayesian recovery algorithms, especially beyond memoryless sources. Two of the main tools that have been used for this purpose in the literature are the replica method [9] and state evolution [18]. Both methods have been employed to analyze the asymptotic performance of MMSE and MAP estimators under the asymptotic setting where $m, n \to \infty$, while $m/n$ is fixed. They both have some limitations. For instance, replica-based methods are not fully rigorous. Moreover, while they work well for i.i.d. sequences, it is not clear how they can be applied to sources with memory. The state evolution framework suffers from similar issues. Our article presents the first result in this direction for processes with memory.

### 1.4 *Organization of the article*

The organization of the article is as follows. The Q-MAP estimator is developed in Section 2, and in Section 3, it is simplified for some simple distributions and shown to have connections to some well-known algorithms. In Section 4, two classes of stochastic processes are studied. The empirical distributions of the quantized versions of processes in each class have exponential convergence rates. The performance of the Q-MAP estimator is studied in Section 5. An iterative method based on PGD is proposed and studied in Section 6. Section 7 presents the proofs of the main results of the article and finally Section 8 concludes the article.

## 2. Quantized MAP estimator

Consider the problem of estimating $X^n$ from noise-free response variables $Y^m = AX^n$, where $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$ is a stationary process. Since there is no noise, one can employ a MAP estimator and find the most likely parameter vector given response variables $Y^m$. Instead of solving the original MAP estimator, we consider finding the most probable sequence in the quantized space $\mathcal{X}_b^n$. That is,

$$
\begin{aligned}
\text{maximize} \quad & \mathrm{P}([X^n]_b = u^n) \\
\text{subject to} \quad & u^n \in \mathcal{X}_b^n, \\
& [x^n]_b = u^n, \\
& Ax^n = Y^m,
\end{aligned}
$$ (2.1)

where P denotes the law of process $\mathbf{X}$. The optimization described in (2.1) can be further simplified and made more amenable to both analysis and implementation. Note that if $Ax^n = Y^m$, and $|x_i - u_i| \le 2^{-b}$, for all $i$, then

$$
\begin{aligned}
\|Au^n - Y^m\|_2^2 &\le (\sigma_{\max}(A))^2 \|x^n - u^n\|^2 \\
&\le 2^{-2b} n^2 (\sigma_{\max}(A))^2,
\end{aligned}
$$ (2.2)

where $\sigma_{\max}(A)$ denotes the maximum singular value of the design matrix $A$. This provides an upper bound on $\frac{1}{n^2}\|A[x^n]_b - Y^m\|_2^2$ in terms of $\sigma_{\max}(A)$ and $b$. In other words, since $u^n$ is a quantized version of $x^n$, and $Ax^n = Y^m$, $\|Au^n - Y^m\|_2^2$ is also expected to be small.

To further simplify (2.2), we focus on the other term in (2.2), i.e. $-\log P([X^n]_b = u^n)$. Assume that the process $\mathbf{X}$ is such that $P([X^n]_b = u^n)$ can be factored as

$$P([X^n]_b = u^n) = P([X^k]_b = u^k) \prod_{i=k+1}^{n} P([X_i]_b = u_i | [X_{i-k}^{i-1}]_b = u_{i-k}^{i-1}), \tag{2.3}$$

for some finite $k$. In other words, the $b$-bit quantized version of $\mathbf{X}$ is a Markov process of order $k$. Then define coefficients $(w_{a^{k+1}} : a^{k+1} \in \mathcal{X}_b^{k+1})$ according to (1.6). This assumption simplifies the term $-\log P([X^n]_b = u^n)$ in the following way:

$$-\log P([X^n]_b = u^n)$$

$$= -\log P([X^k]_b = u^k) - \sum_{i=k+1}^{n} \log P([X_i]_b = u_i | [X_{i-k}^{i-1}]_b = u_{i-k}^{i-1})$$

$$= -\log P([X^k]_b = u^k)$$
$$\quad - \sum_{i=k+1}^{n} \log P([X_i]_b = u_i | [X_{i-k}^{i-1}]_b = u_{i-k}^{i-1}) \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} 1_{u_{i-k}^i = a^{k+1}}$$

$$= -\log P([X^k]_b = u^k)$$
$$\quad - \sum_{i=k+1}^{n} \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} 1_{u_{i-k}^i = a^{k+1}} \log P([X_i]_b = a_i | [X_{i-k}^{i-1}]_b = a_{i-k}^{i-1})$$

$$= -\log P([X^k]_b = u^k) + \sum_{i=k+1}^{n} \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} 1_{u_{i-k}^i = a^{k+1}}$$

$$= -\log P([X^k]_b = u^k) + \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} \sum_{i=k+1}^{n} 1_{u_{i-k}^i = a^{k+1}}$$

$$= -\log P([X^k]_b = u^k) + (n-k) \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} \hat{p}^{(k+1)}(a^{k+1}|u^n)$$

$$= -\log P([X^k]_b = u^k) + (n-k) c_{\mathbf{w}}(u^n), \tag{2.4}$$

where $\hat{p}^{(k+1)}$ denotes the $(k+1)$th order empirical distribution of $u^n$ as defined in (1.2), and $c_{\mathbf{w}}(u^n)$ is defined in (1.4). Assuming $k$ is much smaller than $n$, and ignoring the negligible term of $-\log P([X^k]_b = u^k)/(n-k)$, instead of minimizing $-\log P([X^n]_b = u^n)$, subject to an upper bound on $\|Au^n - Y^m\|_2^2$, we consider the following optimization where the roles of the cost and constraint functions are flipped

$$\hat{X}^n = \underset{u^n \in \mathcal{X}_b^n}{\arg\min} \quad \|Au^n - Y^m\|_2^2$$

$$\text{subject to} \quad c_{\mathbf{w}}(u^n) \leq \gamma, \tag{2.5}$$

or its Lagrangian form

$$\hat{X}^n = \underset{u^n \in \mathcal{X}_b^n}{\arg\min} \left[ c_{\mathbf{w}}(u^n) + \frac{\lambda}{n^2} \|Au^n - Y^m\|^2 \right].$$ (2.6)

The choice of parameters $\lambda > 0$ and $\gamma$ is discussed later in our analysis. We refer to both (2.5) and (2.6) as quantized MAP (Q-MAP) estimators.

Obtaining the Q-MAP estimator involved several approximation and relaxation steps. It is not clear how accurate these approximations are, and what the performance of the ultimate algorithm is. Also, solving Q-MAP estimator requires specifying parameters $b$ and $\lambda$, which significantly affect the performance of the estimator. These questions are all answered in Section 5. Before that, in the following section, we focus on two specific processes, which are well studied in the literature and derive the Q-MAP formulation in each case. This will clarify some of the properties of our Q-MAP formulation.

## 3. Special distributions

To get a better understanding of Q-MAP optimization described in (2.5) and (2.6) and especially the term

$$c_{\mathbf{w}}(u^n) = \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} \hat{p}^{(k+1)}(a^{k+1}|u^n),$$

in this section we study two special distributions and derive simpler statement for the Q-MAP optimization in each case.

### 3.1 Independent and identically distributed sparse processes

One of the most popular models for sparse parameter vectors is the spike and slab prior [19]. Consider an i.i.d. process $\mathbf{X}$, such that $X_i \sim (1-p)\delta_0 + pU_{[0,1]}$. Since the process is i.i.d., by setting $k = 0$ the optimization stated in (2.5) can be simplified as

$$\hat{X}^n = \underset{u^n \in \mathcal{X}_b^n : \sum_{a \in \mathcal{X}_b} w_a \hat{p}^{(1)}(a|u^n) \leq \gamma}{\arg\min} \|Au^n - Y^m\|^2,$$ (3.1)

where

$$w_a = \log \frac{1}{P([X_1]_b = a_1)} = \log \frac{1}{(1-p)1_{a_1=0} + p2^{-b}}.$$ (3.2)

Therefore, $\sum_{a \in \mathcal{X}_b} w_a \hat{p}^{(1)}(a|u^n)$ in (3.1) can be written as

$$\sum_{a \in \mathcal{X}_b} w_a \hat{p}^{(1)}(a|u^n) = \sum_{a \in \mathcal{X}_b} w_a \left( \frac{1}{n} \sum_{i=1}^n 1_{u_i=a} \right)$$

$$= \frac{1}{n} \sum_{i=1}^n \sum_{a \in \mathcal{X}_b} w_a 1_{u_i=a}$$

$$\stackrel{(a)}{=} \frac{1}{n} \sum_{i=1}^{n} w_{u_i}$$

$$= -\frac{1}{n} \sum_{i=1}^{n} \log((1-p)1_{u_i=0} + p2^{-b})$$

$$= -\hat{p}(0|u^n) \log((1-p) + p2^{-b}) - (1 - \hat{p}(0|u^n)) \log(p2^{-b})$$

$$= \hat{p}(0|u^n) \log \frac{p2^{-b}}{(1-p) + p2^{-b}} - \log(p2^{-b}), \tag{3.3}$$

where (a) holds because $\sum_{a \in \mathcal{X}_b} w_a 1_{u_i=a} = w_{u_i}$. Since $-\log(p2^{-b})$ is constant, and

$$\log \frac{(1-p) + p2^{-b}}{p2^{-b}}$$

is positive, from (3.3), an upper bound on $\sum_{a \in \mathcal{X}_b} w_a \hat{p}^{(1)}(a|u^n)$ is in fact an upper bound on the $\ell_0$-norm of $u^n$ defined as

$$\|u^n\|_0 \triangleq |\{i : u_i \neq 0\}|.$$

(Note that $\|u^n\|_0 = (1 - \hat{p}(0|u^n))n$.) Therefore, given these simplifications, (3.1) can be written as

$$\hat{X}^n = \underset{u^n \in \mathcal{X}_b^n : \|u^n\|_0 \leq \gamma'}{\arg \min} \|Au^n - Y^m\|^2, \tag{3.4}$$

where $\gamma'$ is a function of $\gamma$, $b$ and $p$.

### 3.2 Piecewise-constant processes

Another popular example is a piecewise-constant process. (Refer to [**29**] for some applications of this model.) As our second example, we introduce a first-order Markov process that can model piecewise-constant functions. Conditioned on $X_i = x_i$, $X_{i+1}$ is distributed as $(1-p)\delta_{x_i} + p\text{Unif}_{[0,1]}$. In other words, at each time step, the Markov chain either stays at its previous value or jumps to a new value, which is drawn from a uniform distribution over $[0, 1]$, independent of the past values. The jump process can be modeled as a $\text{Bern}(p)$ process which is independent of the past values of the Markov chain. Then, since the process has a memory of order one, (2.6) can be written as

$$\hat{X}^n = \underset{u^n \in \mathcal{X}_b^n}{\arg \min} \left[ \sum_{a^2 \in \mathcal{X}_b^2} w_{a^2} \hat{p}^{(2)}(a^2|u^n) + \frac{\lambda}{n^2} \|Au^n - Y^m\|^2 \right], \tag{3.5}$$

where for $a^2 \in \mathcal{X}_b^2$

$$w_{a^2} = \log \frac{1}{\text{P}([X_2]_b = a_2 | [X_1]_b = a_1)}. \tag{3.6}$$

Given the Kernel of the Markov chain, we have

$$P([X_2]_b = a_2 | [X_1]_b = a_1) = (1 - p)1_{a_2=a_1} + p2^{-b}. \tag{3.7}$$

Let $N_J(u^n)$ denote the number of jumps in sequence $u^n$, i.e. $N_J(u^n) = \sum_{i=2}^{n} 1_{u_i \neq u_{i-1}}$. Then, the first term in the cost function in (2.5) can be rewritten as

$$\sum_{a^2 \in \mathcal{X}_b^2} w_{a^2} \hat{p}^{(2)}(a^2 | u^n) = \frac{1}{n-1} \sum_{i=2}^{n} \sum_{a^2 \in \mathcal{X}_b} w_{a^2} 1_{u_{i-1}^i = a^2} = \frac{1}{n-1} \sum_{i=2}^{n} w_{u_{i-1}^i}$$

$$= \frac{1}{n-1} \sum_{i=2}^{n} \log\left((1 - p)1_{u_i=u_{i-1}} + p2^{-b}\right)$$

$$= -\frac{N_J(u^n)}{n-1} \log(p2^{-b}) - \left(1 - \frac{N_J(u^n)}{n-1}\right) \log(1 - p + p2^{-b})$$

$$= \frac{N_J(u^n)}{n-1} \log \frac{1 - p + p2^{-b}}{p2^{-b}} - \log(1 - p + p2^{-b}). \tag{3.8}$$

Inserting (3.8) in (3.9), it follows that

$$\hat{X}^n = \underset{u^n \in \mathcal{X}_b^n}{\arg\min} \left[\alpha_b \left(\frac{N_J(u^n)}{n-1}\right) + \frac{\lambda}{n^2} \|Au^n - Y^m\|^2\right], \tag{3.9}$$

where $\alpha_b = \log \frac{(1-p)+p2^{-b}}{p2^{-b}}$. Note that the term $\frac{N_J(u^n)}{n-1}$ is counting the number of jumps in $u^n$ which seems to be a natural regularizer here.

## 4. Exponential convergence rates

In our theoretical analysis, one of the main features required from the source process $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$ is that the empirical statistics derived from $[X^n]_b$ to converge, asymptotically, to their expected values. In all our analysis, we require this to hold even when $b$ grows with $n$. Intuitively, if this is not the case, we do not expect the Q-MAP estimator to be able to obtain a good estimate of $X^n$. In the following two sections, we study two important classes of stochastic processes which satisfy this property.

### 4.1 $\Psi^*$-mixing processes

The first class of processes that satisfy our requirements are $\Psi^*$-mixing processes. Consider a stationary process $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$. Let $\mathcal{F}_j^k$ denote the $\sigma$-field of events generated by random variables $X_j^k$, where $j \leq k$. Define

$$\psi^*(g) = \sup \frac{P(\mathcal{A} \cap \mathcal{B})}{P(\mathcal{A}) P(\mathcal{B})}, \tag{4.1}$$

where the supremum is taken over all events $\mathcal{A} \in \mathcal{F}_0^j$ and $\mathcal{B} \in \mathcal{F}_{j+g}^{\infty}$, where $P(\mathcal{A}) > 0$ and $P(\mathcal{B}) > 0$.

DEFINITION 4.1  A stationary process **X** is called $\psi^*$-mixing, if $\psi^*(g) \to 1$, as $g$ grows to infinity.

There are many examples of $\Psi^*$-mixing processes. For instance, it is straightforward to check that any i.i.d. sequence is $\Psi^*$-mixing. Also, every finite-state Markov chain is $\Psi^*$-mixing [25]. (For further information on $\Psi^*$-mixing processes, the reader may refer to [3].) Another example of $\Psi^*$-mixing processes are processes built by taking the moving average of i.i.d. processes. More specifically, consider an i.i.d. process **Y** and let process **X** denote the moving average of process **Y** defined as $X_i = \frac{1}{r} \sum_{j=0}^{r-1} Y_{i-j}$, for all $i$. Then, process **Y** is $\Psi^*$-mixing.

As mentioned earlier, the advantage of $\Psi^*$-mixing processes is the fast convergence of their empirical distributions to their expected values. This is captured by the following result from [15], which is a straightforward extension of a similar result in [25] for finite alphabet.

THEOREM 4.2  Consider a $\Psi^*$-mixing process $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$, and its $b$-bit quantized version $\mathbf{Z} = \{Z_i\}_{i=1}^{\infty}$, where $Z_i = [X_i]_b$ and $\mathcal{Z} = \mathcal{X}_b$. Define measure $\mu_k$, such that for $a^k \in \mathcal{Z}^k$, $\mu_k(a^k) = P(Z^k = a^k)$. Then, for any $\epsilon > 0$, and any $b$ large enough, there exists $g \in \mathbb{N}$, depending only on $\epsilon$ and function $\Psi^*(g)$ defined in (4.1), such that for any $n > k + 6(k+g)/\epsilon$,

$$P(\|\hat{p}^{(k)}(\cdot|Z^n) - \mu_k\|_1 \geq \epsilon) \leq 2^{c\epsilon^2/8}(k+g)n^{|\mathcal{Z}|^k}2^{-\frac{nc\epsilon^2}{8(k+g)}}, \tag{4.2}$$

where $c = 1/(2 \ln 2)$.

Note that the upper bound in (4.2) only depends on $b$ through $|\mathcal{Z}| = |\mathcal{X}_b|$. Hence, if $b = b_n$ grows with $n$, it should grow slowly enough, such that overall $2^{c\epsilon^2/8}(k+g)n^{|\mathcal{Z}|^k}2^{-\frac{nc\epsilon^2}{8(k+g)}}$ still converges to zero, as $n$ grows to infinity. One such example, used in our results, is $b = b_n = \lceil r \log \log n \rceil$, $r \geq 1$. For this choice of $b_n$, Theorem 4.2 guarantees that the empirical distribution derived from the quantized sequence remains close to its expected value, with high probability.

## 4.2  Weak $\Psi_q^*$-mixing Markov processes

Finite-alphabet Markov chains are known to be $\Psi^*$-mixing, and therefore their empirical distributions have exponential convergence rates [25]. Continuous space Markov processes on the other hand are not $\Psi^*$-mixing in general, and hence the results of the previous section may not hold. However, for many such Markov processes, it is still possible to show that the empirical distributions of their quantized versions converge to their expected values, even if the quantization level $b$ grows with $n$. In this section, we show how such results can be proved by extending the definition of $\Psi^*$-mixing processes to define weak $\Psi_q^*$-mixing *Markov* processes. As shown at the end of this section, one important example of continuous-space Markov processes which is provably weak $\Psi_q^*$-mixing is the piecewise-constant process discussed in Section 3.2.

Consider a real-valued stationary stochastic process $\mathbf{X} = \{X_i\}$, with alphabet $\mathcal{X} = [l, u]$, where $l, u \in \mathbb{R}$. Let process $\mathbf{Z} = \{Z_i\}$ denote the $b$-bit quantized version of process **X**. That is, $Z_i = [X_i]_b$, for all $i$. The alphabet of process **Z** is clearly $\mathcal{Z} = \mathcal{X}_b = \{[x]_b : x \in \mathcal{X}_b\}$. Let $\mu_k^{(b)}$ denote the distribution of $Z^k$. That is, for any $z^k \in \mathcal{Z}^k$,

$$\mu_k^{(b)}(z^k) = P(Z^k = z^k) = P([X^k]_b = z^k). \tag{4.3}$$

The following lemma proves that if the process $\mathbf{X}$ has a property analogous to being $\Psi^*$-mixing, then potentially it has exponential convergence rates.

LEMMA 4.1 Suppose that the stationary process $\mathbf{X}$ is such that there exists a function $\Psi : \mathbb{N} \times \mathbb{N} \to \mathbb{R}^+$, which satisfies the following. For any $(b, g, \ell_1, \ell_2) \in \mathbb{N}^4$, $u^{\ell_1} \in \mathcal{Z}^{\ell_1}$ and $v^{\ell_2} \in \mathcal{Z}^{\ell_2}$:

$$P\left(Z^{\ell_1} = u^{\ell_1}, Z_{\ell_1+g+1}^{\ell_1+g+\ell_2} = v^{\ell_2}\right) \leq P\left(Z^{\ell_1} = u^{\ell_1}\right) P\left(Z_{\ell_1+g+1}^{\ell_1+g+\ell_2} = v^{\ell_2}\right) \Psi(b, g), \tag{4.4}$$

where $b$ denotes the quantization level of process $\mathbf{Z}$. Then for any given $\epsilon > 0$, for any positive integers $k$ and $g$ such that $4(k + g)/(n - k) < \epsilon$,

$$P\left(\|\hat{p}^{(k)}(\cdot|Z^n) - \mu_k^{(b)}\|_1 \geq \epsilon\right) \leq (k + g)\Psi^t(b, g)(t + 1)^{|\mathcal{Z}|^k} 2^{-c\epsilon^2 t/4}, \tag{4.5}$$

where $t = \lfloor \frac{n-k+1}{k+g} \rfloor$ and $c = 1/(2 \ln 2)$.

The proof is presented in Section 7.3. Note that if a process is $\Psi^*$-mixing, then it is straightforward to confirm the existence of $\tilde{\Psi}(g)$ that satisfies

$$P\left(Z^{\ell_1} = u^{\ell_1}, Z_{\ell_1+g+1}^{\ell_1+g+\ell_2} = v^{\ell_2}\right) \leq P\left(Z^{\ell_1} = u^{\ell_1}\right) P\left(Z_{\ell_1+g+1}^{\ell_1+g+\ell_2} = v^{\ell_2}\right) \tilde{\Psi}(g). \tag{4.6}$$

However, in this section, we are interested in processes that are not necessarily $\Psi^*$-mixing. Lemma 4.1 allows us to prove the convergence of the empirical distributions for many such processes. To justify our claims, we focus on the class of Markov processes. For notational simplicity, we focus on first-order Markov processes. It is straightforward to extend these results to higher order Markov processes as well.

Let $\mathbf{X}$ denote a first-order stationary Markov process with Kernel function $K : (\mathcal{X}, 2^{\mathcal{X}}) \to \mathbb{R}^+$ and first-order stationary distribution $\pi : 2^{\mathcal{X}} \to \mathbb{R}^+$. (Here $2^{\mathcal{X}}$ denotes the set of subsets of $\mathcal{X}$.) In other words, for any $x \in \mathcal{X}$ and any measurable subset of $\mathcal{X}$,

$$K(x, \mathcal{A}) = P(X_2 \in \mathcal{A}|X_1 = x) \tag{4.7}$$

and

$$\pi(\mathcal{A}) = P(X_1 \in \mathcal{A}). \tag{4.8}$$

Also, for $g \in \mathbb{N}^+$,

$$K^g(x, \mathcal{A}) = P(X_{1+g} \in \mathcal{A}|X_1 = x). \tag{4.9}$$

Clearly $k^g$ can be evaluated from function $K$. Finally, with a slight overloading of notation, for $x \in \mathcal{X}$ and $z \in \mathcal{X}_b$,

$$\pi(z) = P([X_1]_b = z) \tag{4.10}$$

and

$$K(x, z) = P([X_2]_b = z|X_1 = x). \tag{4.11}$$

Similarly, for $g \in \mathbb{N}^+$, $K^g(x, z) = \mathrm{P}([X_{1+g}]_b = z | X_1 = x)$. Again, with another slight overloading of notation, for $x \in \mathcal{X}$, and $w^{l+1} \in \mathcal{Z}^{l+1}$,

$$\pi\left(w_2^{l+1} | x\right) = \mathrm{P}\left([X_2^{l+1}]_b = w_2^{l+1} | X_1 = x\right) \qquad (4.12)$$

and

$$\pi\left(w_2^{l+1} | w_1\right) = \mathrm{P}\left([X_2^{l+1}]_b = w_2^{l+1} | [X_1]_b = w_1\right). \qquad (4.13)$$

Define functions $\Psi_1 : \mathbb{N} \times \mathbb{N} \to \mathbb{R}^+$ and $\Psi_2 : \mathbb{N} \to \mathbb{R}^+$ as

$$\Psi_1(b, g) \triangleq \sup_{(x,z) \in \mathcal{X} \times \mathcal{X}_b} \frac{K^g(x, z)}{\pi(z)} \qquad (4.14)$$

and

$$\Psi_2(b) \triangleq \sup_{(x, \ell_2, w^{\ell_2}) \in \mathcal{X} \times \mathbb{N} \times \mathcal{Z}^{\ell_2} : [x]_b = w_1} \frac{\pi(w_2^{\ell_2} | x)}{\pi(w_2^{\ell_2} | w_1)}. \qquad (4.15)$$

Our next lemma shows how $\Psi(b, g)$ in Lemma 4.1 can be calculated using $\Psi_1$ and $\Psi_2$.

LEMMA 4.2 Consider a first-order aperiodic Markov process $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$. Let $\mathbf{Z} = \{Z_i\}$ denote the $b$-bit quantized version of process $\mathbf{X}$. That is, $Z_i = [X_i]_b$, and $\mathcal{Z} = \mathcal{X}_b = \{[x]_b : x \in \mathcal{X}\}$. Also, let $\mu_b$ denote the distribution associated with the finite-alphabet process $\mathbf{Z}$. Then, for all $(\ell_1, g, \ell_2) \in \mathbb{N}^3$, $u^{\ell_1} \in \mathcal{Z}^{\ell_1}$, $v^g \in \mathcal{Z}^g$ and $w^{\ell_2} \in \mathcal{Z}^{\ell_2}$, we have

$$\mu_b\left(u^{\ell_1} v^g w^{\ell_2}\right) \le \mu_b\left(u^{\ell_1}\right) \mu_b\left(w^{\ell_2}\right) \Psi_1(b, g) \Psi_2(b), \qquad (4.16)$$

where by definition $\mu_b\left(u^{\ell_1} v^g w^{\ell_2}\right) = \mathrm{P}\left(Z^{\ell_1 + g + \ell_2} = [u^{\ell_1}, v^g, w^{\ell_2}]\right)$, $\mu_b\left(u^{\ell_1}\right) = \mathrm{P}\left(Z^{\ell_1} = u^{\ell_1}\right)$ and $\mu_b\left(w^{\ell_2}\right) = \mathrm{P}\left(Z^{\ell_2} = w^{\ell_2}\right)$. Furthermore, for any fixed $b$, $\Psi_1(b, g)$ is a non-increasing function of $g$ that converges to 1, as $g$ grows to infinity.

The proof is presented in Section 7.4.

Combining Lemmas 4.1 and 4.2, we obtain an upper bound of the form

$$(k + g) \Psi^t(b, g)(t + 1)^{|\mathcal{Z}|^k} 2^{-c\epsilon^2 t/4}$$

on $\mathrm{P}(\|\hat{p}^{(k)}(\cdot | Z^n) - \mu_k^{(b)}\|_1 \ge \epsilon)$. To prove our desired convergence results, we need to ensure that this upper bound goes to zero as $n$ grows to infinity. It is straightforward to note that as $n \to \infty$, $t \to \infty$, and hence the term $2^{-c\epsilon^2 t/4}$ converges to zero. However, if $\Psi^t(b, g)(t + 1)^{|\mathcal{Z}|^k}$ grows faster than $2^{-c\epsilon^2 t/4}$, then we do not reach the desired goal. Our next theorems prove that under some mild conditions on the Markov process, for slow enough growth of $b = b_n$ with $n$, $\Psi^t(b, g)$ does not grow too fast. First note that since for any $z^n \in \mathcal{Z}^n$, $\|\mu_{k_1} - \hat{p}^{(k_1)}(\cdot | z^n)\|_1 \le \|\mu_{k_2} - \hat{p}^{(k_2)}(\cdot | z^n)\|_1$, for all $k_1 \le k_2$, to prove fast convergence of $\hat{p}^{(k)}(\cdot | Z^n)$ it is enough to prove this statement for $k$ large.

THEOREM 4.3 Consider an aperiodic stationary first-order Markov chain $\mathbf{X}$, and its $b$-bit quantized version $\mathbf{Z}$, where $Z_i = [X_i]_b$ and $\mathcal{Z} = \mathcal{X}_b$. Let $\mu_k^{(b)}$ denote the $k$th order probability distribution of process $\mathbf{Z}$, i.e. for any $z^k \in \mathcal{Z}^k$,

$$\mu_k^{(b)}(z^k) = P(Z^k = z^k). \tag{4.17}$$

Let $b = b_n = \lceil r \log \log n \rceil$, where $r \geq 1$. Assume that there exists a sequence $g = g_n$, such that $g_n = o(n)$, and process $\mathbf{X}$ satisfies the following conditions:

1. $\lim_{n \to \infty} \Psi_1(b_n, g_n) = 1$ and
2. $\lim_{b \to \infty} \Psi_2(b) = 1$,

where functions $\Psi_1$ and $\Psi_2$ are defined in (4.14) and (4.15), respectively. Then, given $\epsilon > 0$ and positive integer $k$, for $n$ large enough,

$$P\left( \|\hat{p}^{(k)}(\cdot | Z^n) - \mu_k^{(b)}\|_1 \geq \epsilon \right) \leq 2^{c\epsilon^2/4}(k + g_n)n^{|\mathcal{Z}|^k} 2^{-\frac{c\epsilon^2 n}{8(k+g_n)}}, \tag{4.18}$$

where $c = 1/(2 \ln 2)$.

The proof is presented in Section 7.5.

REMARK 4.1 Lemma 4.2 proves that for any fixed $b$, $\Psi_1(b, g)$ converges to one, as $g$ grows without bound. However, in this article we are mainly interested in the cases where $b_n = \lceil r \log \log n \rceil$. The condition on $\Psi_1$ specified in Theorem 4.3 ensures that even if $b$ also grows to infinity, there exists a proper choice of sequence $g_n$ as a function of $n$, for which $\Psi_1(b_n, g_n)$ still converges to one, as $n$ grows without bound.

Theorem 4.3 proves that if $\lim_{n \to \infty} \Psi_1(b_n, g_n) = 1$ and $\lim_{b \to \infty} \Psi_2(b) = 1$, then the quantized version of an analog Markov process also has fast convergence rates. We refer to a Markov process that satisfies these two conditions with $b_n = \lceil r \log \log n \rceil$ as a *weak $\Psi_q^*$-mixing Markov* process. To better understand these conditions, we next consider the piecewise-constant source studied in Section 3.2 and prove that it is a weak $\Psi_q^*$-mixing Markov process.

THEOREM 4.4 Consider a first-order stationary Markov process $\mathbf{X}$, such that conditioned on $X_i = x_i$, $X_{i+1}$ is distributed as $(1 - p)\delta_{x_i} + pf_c$, where $f_c$ denotes an absolutely continuous distribution over $\mathcal{X} = [0, 1]$. Further assume that there exists $f_{\min} > 0$, such that $f_c(x) \geq f_{\min}$, for all $x \in (0, 1)$. Then,

1. for $b = b_n = \lceil r \log \log n \rceil$ and $g = g_n = \lfloor \gamma r \log \log n \rfloor$, where $\gamma > -\frac{1}{\log(1-p)}$, $\lim_{n \to \infty} \Psi_1(b_n, g_n) = 1$, and
2. $\Psi_2(b) = 1$, for all $b$.

The proof is presented in Section 7.6.

## 5. Theoretical analysis of Q-MAP

In this section, we formalize Informal Result 1 presented in Section 1. The following theorem provides conditions for the success of the Q-MAP estimator, for the case where the response variables are noise-free. We state all the results for $\Psi^*$-mixing processes, but they also apply to weak $\Psi_q^*$-mixing Markov processes.

THEOREM 5.1 Consider a $\Psi^*$-mixing stationary process $\mathbf{X}$, and let $Y^m = AX^n$. Assume that the entries of the design matrix $A$ are i.i.d. $\mathcal{N}(0,1)$. Choose $k$, $r > 1$ and $\delta > 0$, and let $b = b_n = \lceil r \log \log n \rceil$, $\gamma = \gamma_n = b_n(\bar{d}_k(\mathbf{X}) + \delta)$ and $m = m_n \geq (1 + \delta)n\bar{d}_k(\mathbf{X})$. Assume that there exists a constant $f_{k+1} > 0$, such that for any quantization level $b$ and any $u^{k+1} \in \mathcal{X}_b^{k+1}$ with $\mathrm{P}([X^{k+1}]_b = u^{k+1}) \neq 0$,

$$\mathrm{P}\left([X^{k+1}]_b = u^{k+1}\right) \geq f_{k+1}|\mathcal{X}_b|^{-(k+1)}. \tag{5.1}$$

Further, assume that $\hat{X}^n$ denotes the solution of (2.5), where the coefficients are computed according to (1.6). Then, for any $\epsilon > 0$,

$$\lim_{n \to \infty} \mathrm{P}\left(\frac{1}{\sqrt{n}}\|X^n - \hat{X}^n\|_2 > \epsilon\right) = 0. \tag{5.2}$$

The proof is presented in Section 7.7.

REMARK 5.1 A technical condition required by Theorem 5.1 is existence of a constant $f_{k+1}$, for which (5.1) holds. It is straightforward to confirm that this condition holds for many distributions of interest. For instance, consider an i.i.d. process $X_i \overset{\text{i.i.d.}}{\sim} pf_c + (1-p)f_d$, where $f_c$ and $f_d$ denote an absolutely continuous distribution and a discrete distribution, respectively. Then, if $f_c$ has a bounded support and $\inf f_c$ over its support is non-zero, then (5.1) holds. Intuitively, this condition guarantees that the probability of every quantized sequence with non-zero probability cannot get too small. The reason this condition is required is that it simplifies bounding the Kullback–Leibler distance between empirical distribution and the underlying distribution. We believe that the theorem holds in a more general setting, but its proof is left for future work.

We remind the reader that we also introduced a Lagrangian version of Q-MAP in (2.6). It turns out that we can derive the same performance guarantees for the Lagrangian Q-MAP as well.

THEOREM 5.2 Consider a $\Psi^*$-mixing stationery process $\mathbf{X}$. Let $Y^m = AX^n$, where the entries of $A$ are i.i.d. $\mathcal{N}(0,1)$. Choose $k$, $r > 1$ and $\delta > 0$, and let $b = b_n = \lceil r \log \log n \rceil$, $\lambda = \lambda_n = (\log n)^{2r}$ and $m = m_n \geq (1 + \delta)n\bar{d}_k(\mathbf{X})$. Assume that there exists a constant $f_{k+1} > 0$, such that for any quantization level $b$, and any $u^{k+1} \in \mathcal{X}_b^{k+1}$ with $\mathrm{P}([X^{k+1}]_b = u^{k+1}) \neq 0$, (5.1) holds. Further, assume that $\hat{X}^n$ denotes the solution of (2.6), where the coefficients are computed according to (1.6). Then, for any $\epsilon > 0$,

$$\lim_{n \to \infty} \mathrm{P}\left(\frac{1}{\sqrt{n}}\|X^n - \hat{X}^n\|_2 > \epsilon\right) = 0. \tag{5.3}$$

The proof is presented in Section 7.8.

To better understand the implications of Theorems 5.1 and 5.2, consider the case where the process **X** is stationary and memoryless. All such processes are $\Psi^*$-mixing, and satisfy $\bar{d}_k(\mathbf{X}) = \bar{d}_0(\mathbf{X})$, for all $k \geq 0$. Therefore, as long as $m_n \geq (1 + \delta)n\bar{d}_0(\mathbf{X})$, asymptotically, the Q-MAP algorithm provides an accurate estimate of the parameters vector. On the other hand, since the process is i.i.d., $\bar{d}_0(\mathbf{X}) = \bar{d}(X_1)$, where $\bar{d}(X_1)$ denotes the upper Rényi information dimension of $X_1$ [23]. For an i.i.d. process whose marginal distribution is a mixture of discrete and continuous distributions, asymptotically, the Rényi information dimension of the marginal distribution characterizes the minimum sampling rate $(m/n)$ required for an accurate recovery of the parameters vector [33]. Hence, for such i.i.d. sources, in a noiseless setting, the algorithm presented in (3.4) achieves the fundamental limits in terms of sampling rate.

Finally, another interesting implication of Theorem 5.2 is the following. The Q-MAP optimization mentioned in (2.5) is not a convex optimization. Hence, its solution does not necessarily coincide with the solution of (2.6). However, at least in the noiseless setting we can derive similar performance bounds for both.

## 6. Solving Q-MAP

### 6.1 *PGD*

The goal of this section is to analyze the performance of the PGD algorithm introduced in Section 1.2. The results are presented for $\Psi^*$-mixing processes, but they also apply to weak $\Psi_q^*$-mixing Markov processes with no change. Note that even though PGD algorithms have been studied extensively for convex optimization problems, since our optimization is discrete and consequently not convex, those analyses do not apply to our problem. Given $\delta > 0$, consider the Q-MAP optimization characterized as

$$\hat{X}^n = \underset{u^n \in \mathcal{X}_b^n}{\arg\min} \quad \|Y^m - Au^n\|^2$$

$$\text{subject to} \quad c_{\mathbf{w}}(u^n) \leq b(\bar{d}_k(\mathbf{X}) + \delta). \tag{6.1}$$

The corresponding PGD algorithm proceeds as follows. For $t = 1, 2, \ldots, t_n$,

$$S^n(t + 1) = \hat{X}^n(t) + \mu A^\top(Y^m - A\hat{X}^n(t))$$

$$\hat{X}^n(t + 1) = \underset{u^n \in \mathcal{F}_o}{\arg\min} \|u^n - S^n(t + 1)\|, \tag{6.2}$$

where $\mathcal{F}_o = \left\{ u^n \in \mathcal{X}_b^n : c_{\mathbf{w}}(u^n) \leq b(\bar{d}_k(\mathbf{X}) + \delta) \right\}$, defined earlier in (1.9).

Theorem 6.1 below proves that, having enough number of response variables, with probability approaching one, the PGD-based algorithm recovers parameters $X^n$ with high probability, even in the presence of measurement noise.

THEOREM 6.1 Consider a $\Psi^*$-mixing process **X**. Let $Y^m = AX^n + Z^m$, where the elements of matrix $A$ are i.i.d. $\mathcal{N}(0, 1)$ and $Z_i$, $i = 1, \ldots, m$, are i.i.d. $\mathcal{N}(0, \sigma^2)$. Choose $k, r > 1$ and $\delta > 0$, and let $b = b_n = \lceil r \log \log n \rceil$, $\gamma = \gamma_n = b_n(\bar{d}_k(\mathbf{X}) + \delta)$, and $m = m_n \geq (1+\delta)n\bar{d}_k(\mathbf{X})$. Assume that there exists a constant $f_{k+1} > 0$, such that for any quantization level $b$, and any $u^{k+1} \in \mathcal{X}_b^{k+1}$ with $\mathrm{P}([X^{k+1}]_b = u^{k+1}) \neq 0$, (5.1)

holds. For $r > 1$, let $b = b_n = \lceil r \log \log n \rceil$, and $m = m_n = 80nb(\bar{d}_k(\mathbf{X}) + \delta)$. Let $\mu = \frac{1}{m}$, and consider $\hat{X}_n(t)$, $t = 0, 1, \ldots, t_n$, generated according to (6.2). Define the error vector at iteration $t$ as

$$E^n(t) = \hat{X}^n(t) - [X^n]_b. \tag{6.3}$$

Then, with probability approaching one,

$$\frac{1}{\sqrt{n}} \|E^n(t+1)\| \leq \frac{0.9}{\sqrt{n}} \|E^n(t)\| + 2\left(2 + \sqrt{\frac{n}{m}}\right)^2 2^{-b} + \frac{\sigma}{2}\sqrt{\frac{b(\bar{d}_k(\mathbf{X}) + 3\delta)}{m}}, \tag{6.4}$$

for $t = 1, 2, \ldots$.

The proof is presented in Section 7.9.

Comparing this result with Theorem 5.2 reveals that the minimum value of $m$ required in this theorem is $80b_n = 80\lceil r \log \log n \rceil$ times higher than the number of response variables required in Theorem 5.2. One can also decrease (increase) the factor 80 and slow down (speed up) the convergence rate of the algorithm. At this point it is not clear to us whether the factor $r \log \log n$ in Theorem 6.1 (for the number of response variables) is necessary in general or is an artifact of our proof technique. For some specific priors such as the spike and slab distribution discussed earlier, with a slight modification of the algorithm, it is known that this factor can be improved. In that case, given the special form of the coefficients, we may let $b$ grow to infinity for a fixed $n$. Then the algorithm becomes equivalent to the iterative hard thresholding (IHT) algorithm introduced in [2]. The analysis in [2] shows that the number of response variables $m_n$ required by the IHT algorithm is proportional to $n$ and does not have the $\log \log n$ factor that appears in Theorem 6.1.

In Theorem 6.1 and all of the previous results, the elements of the design matrix $A$ were assumed to be generated according to $\mathcal{N}(0, 1)$ distribution. In the noisy setup, where the response variables are distorted by a noise of variance $\sigma^2$, this model implies having per response signal to noise ratio (SNR) that grows linearly with $n$. This has made the result of the previous theorem misleading. If we consider per element error, i.e. $\frac{1}{\sqrt{n}} \|E^n(t+1)\|$ then the error seems to go to zero. To fix this issue, we assume that the elements of $A$ are generated according to $\mathcal{N}(0, \frac{1}{n})$. The following corollary restates the result of Theorem 6.1 under this scaling and an appropriate adjustment of coefficient $\mu$.

COROLLARY 6.1 Consider the setup of Theorem 6.1, where the elements of $A$ are generated i.i.d. $\mathcal{N}(0, \frac{1}{n})$. Let

$$S^n(t+1) = \hat{X}^n(t) + \frac{n}{m} A^\top (Y^m - A\hat{X}^n(t))$$

$$\hat{X}^n(t+1) = \underset{u^n \in \mathcal{F}_o}{\arg \min} \|u^n - S^n(t+1)\|. \tag{6.5}$$

Then, with probability approaching one,

$$\frac{1}{\sqrt{n}} \|E^n(t+1)\| \leq \frac{0.9}{\sqrt{n}} \|E^n(t)\| + \frac{2(\sqrt{n} + 2\sqrt{m})^2}{m} 2^{-b} + \frac{\sigma}{2}\sqrt{\frac{nb(\bar{d}_k(\mathbf{X}) + 3\delta)}{m}}, \tag{6.6}$$

for $t = 1, 2, \ldots$.

Note that for $m = m_n = 80nb(\bar{d}_k(\mathbf{X}) + 3\delta)$,

$$\frac{\sigma}{2}\sqrt{\frac{nb(\bar{d}_k(\mathbf{X}) + 3\delta)}{m}} \leq \frac{\sigma}{12}.$$

### 6.2 Discussion of computational complexity of PGD

As explained earlier, at iteration $t + 1$, the PGD-based algorithm updates its estimate $\hat{X}^n(t)$ to $\hat{X}^n(t+1)$ by performing the following two steps:

1. $S^n(t+1) = \hat{X}^n(t) + \mu A^\top(Y^m - A\hat{X}^n(t))$,
2. $\hat{X}^n(t+1) = \arg\min_{u^n \in \mathcal{F}_o} \|u^n - S^n(t+1)\|$.

Clearly, the challenging part is performing the second step, which is projection on the set $\mathcal{F}_o$. For some special distributions, such as the spike and slab prior, discussed in Section 3.1, and piecewise-constant processes, discussed in Section 3.2, and their extensions this projection step is not complicated. For instance, for the aforementioned sparse vector, $\mathcal{F}_o$ contains sparse quantized vectors, and hence the projection step is just keeping the quantized versions of the largest components of $S^t$ and setting the rest to zero. This is very similar to the IHT algorithm [2]. However, for more general distributions this projection step may be challenging.

Hence, to make the PGD method efficient, we need to be able to solve the following optimization efficiently:

$$\hat{x}^n = \underset{u^n \in \mathcal{X}_b^n}{\arg\min} \quad \|u^n - x^n\|$$

$$\text{subject to} \quad c_{\mathbf{w}}(u^n) \leq \gamma, \tag{6.7}$$

where $x^n \in \mathbb{R}^n$, weights $\mathbf{w} = \{w_{a^{k+1}} : a^{k+1} \in \mathcal{X}_b^{k+1}\}$ and $\gamma \in \mathbb{R}^+$ are all given input parameters. Equation (6.7) can be stated in the Lagrangian form as

$$\hat{x}^n = \underset{u^n \in \mathcal{X}_b^n}{\arg\min} \left[\frac{1}{n^2}\|u^n - x^n\|^2 + \alpha c_{\mathbf{w}}(u^n)\right], \tag{6.8}$$

where $\alpha > 0$ is a parameter that depends on $\gamma$. Since $\|u^n - x^n\|^2 = \sum_{i=1}^n (u_i - x_i)^2$, the optimization stated in (6.8) is exactly the optimization studied in [14]. It has been proved in [14] that the solution of (6.8) can be found efficiently via the standard dynamic programming (Viterbi algorithm) [31]. (For further information, refer [14].)

The question is whether, for an appropriate choice of $\alpha$, the minimizers of (6.7) and (6.8) are the same. If the answer to this question is affirmative, it implies that both steps of the PGD method can be implemented efficiently. In the following, we intuitively argue why we believe that this might be the case. Making the argument rigorous and a deeper investigation of this connection is left to future research.

Consider partitioning the set of sequences in $\mathcal{X}_b^n$ based on their $(k+1)$th order empirical distributions, which are referred to as their $(k+1)$th order types. For a $(k+1)$th order type $q_{k+1}(\cdot) : \mathcal{X}_b^{k+1} \to \mathbb{R}^+$, let $\mathcal{T}_n(q_{k+1})$ denote the set of sequences in $\mathcal{X}_b^n$ whose $(k+1)$th order types agree with $q_{k+1}$. That is,

$$\mathcal{T}_n(q_{k+1}) \triangleq \left\{ u^n : \hat{p}^{(k+1)}(a^{k+1}|u^n) = q_{k+1}(a^{k+1}), \forall\, a^{k+1} \in \mathcal{U}_b^{k+1} \right\}. \tag{6.9}$$

Let $\mathcal{P}_n^{k+1}$ denote the set of all possible $(k+1)$th order types, for sequences in $\mathcal{X}_b^n$. In other words,

$$\mathcal{P}_n^{k+1} \triangleq \left\{ \hat{p}^{(k+1)}(\cdot|u^n) : u^n \in \mathcal{X}_b^n \right\}. \tag{6.10}$$

It can be proved that (Theorem I.6.14 in [**25**])

$$|\mathcal{P}_n^{k+1}| \leq (n+1)^{|\mathcal{X}_b|^{k+1}}. \tag{6.11}$$

Furthermore, we have

$$\mathcal{X}_b^n = \cup_{q_{k+1} \in \mathcal{P}_n^{k+1}} \mathcal{T}_n(q_{k+1}). \tag{6.12}$$

Therefore,

$$
\begin{aligned}
&\min_{u^n \in \mathcal{X}_b^n} \left[ \frac{1}{n} \|u^n - x^n\|^2 + \alpha \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} \hat{p}^{(k+1)}(a^{k+1}|u^n) \right] \\
&= \min_{q_{k+1} \in \mathcal{P}_n^{k+1}} \min_{u^n \in \mathcal{T}_n(q_{k+1})} \left[ \frac{1}{n} \|u^n - x^n\|^2 + \alpha \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} q_{k+1}(a^{k+1}) \right] \\
&= \min_{q_{k+1} \in \mathcal{P}_n^{k+1}} \left[ \left[ \min_{u^n \in \mathcal{T}_n(q_{k+1})} \frac{1}{n} \|u^n - x^n\|^2 \right] + \alpha \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} q_{k+1}(a^{k+1}) \right],
\end{aligned} \tag{6.13}
$$

where the last line follows because $\sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} \hat{p}^{(k+1)}(a^{k+1}|u^n)$ only depends on the $(k+1)$th order type of sequence $u^n$. For any type $q_{k+1} \in \mathcal{P}_n^{k+1}$, define the minimum distortion attainable by sequences of that type as $D(q_{k+1}, x^n)$, i.e.

$$D(q_{k+1}, x^n) = \min_{u^n \in \mathcal{T}_n(q_{k+1})} \frac{1}{n} \|u^n - x^n\|^2. \tag{6.14}$$

Then (6.13) and (6.7) can be written as

$$\min_{q_{k+1} \in \mathcal{P}_n^{k+1}} \left[ D(q_{k+1}, x^n) + \alpha \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} q_{k+1}(a^{k+1}) \right] \tag{6.15}$$

and

$$\min_{q_{k+1} \in \mathcal{P}_n^{k+1} : \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} q_{k+1}(a^{k+1}) \leq \gamma} D(q_{k+1}, x^n), \tag{6.16}$$

respectively. Both of these optimizations are discrete optimization. However, since

$$\sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} q_{k+1}(a^{k+1})$$

is a convex function of $q_{k+1}$, if, in the high-dimensional setting, for input sequences $x^n$ of interest, $D(q_{k+1})$ also behaves almost as a convex function, then we expect the two optimizations to be the same, for a proper choice of parameter $\alpha$. In the remainder of this section, we argue why, in a high-dimensional setting, we conjecture that $D(q_{k+1})$ satisfies the mentioned property. We leave further investigation of the subject to future research.

First, note that if $x^n$ is almost stationary, for instance it is generated by a Markov process with finite memory, then $D(q_{k+1}, x^n)$ only depends on $q_{k+1}$ and finite-order empirical distributions of $x^n$ and not on $n$ or $x^n$. Now assuming that this is true, consider $q_{k+1}^{(1)}$ and $q_{k+1}^{(2)}$ in $\mathcal{P}^{k+1}$. Also given $\theta \in (0, 1)$, let $n_1 = \lfloor \theta n \rfloor$ and $n_2 = n - n_1$. Also, let $\tilde{x}^{n_1}$ and $\bar{x}^{n_2}$ denote the minimizers of $D(q_{k+1}^{(1)}, x^{n_1})$ and $D(q_{k+1}^{(2)}, x^n_{n_1+1})$, respectively. Assume that $\theta q_{k+1}^{(1)} + (1 - \theta) q_{k+1}^{(2)} \in \mathcal{P}_n^{k+1}$ and let $\hat{x}^n = [\tilde{x}^{n_1}, \bar{x}^{n_2}]$. Then, for large $n$, it is straightforward to check that $\hat{p}^{(k+1)}(\cdot|\hat{x}^n) \approx \theta \hat{p}^{(k+1)}(\cdot|\tilde{x}^{n_1}) + (1 - \theta)\hat{p}^{(k+1)}(\cdot|\bar{x}^{n_2}) = \theta q_{k+1}^{(1)} + (1 - \theta) q_{k+1}^{(2)}$. Therefore,

$$
\begin{aligned}
nD(\theta p_1 + (1 - \theta)p_2, x^n) &\leq \|x^n - \hat{x}^n\|^2 \\
&= \|x^{n_1} - \tilde{x}^{n_1}\|^2 + \|x^n_{n_1+1} - \bar{x}^{n_2}\|^2 \\
&= n_1 D\left(q_{k+1}^{(2)}, x^n_{n_1+1}\right) + n_2 D\left(q_{k+1}^{(2)}, x^n_{n_1+1}\right).
\end{aligned}
\tag{6.17}
$$

Dividing both sides by $n$, it follows that

$$D(\theta p_1 + (1 - \theta)p_2, x^n) \leq \theta D\left(q_{k+1}^{(2)}, x^n_{n_1+1}\right) + (1 - \theta)D\left(q_{k+1}^{(2)}, x^n_{n_1+1}\right). \tag{6.18}$$

Therefore, as we expect, for large values of $n$ and stationary sequences $x^n$, $D(q_{k+1}, x^n)$ depends on $x^n$ only through its empirical distribution, then in (6.18), since $x^n_{n_1+1}$ and $x^n_{n_1+1}$ have almost the same empirical distribution as $x^n$, they can be replaced by $x^n$. This establishes our conjecture about almost convexity of function $D$.

## 7. Proofs

### 7.1 Preliminaries on information theory

Before presenting the proofs, in this section, we review some preliminary definitions and concepts that are used in some of the proofs.

Consider stationary process $\mathbf{U} = \{U_i\}_{i=1}^{\infty}$, with finite alphabet $\mathcal{U}$. The entropy rate of process $\mathbf{U}$ is defined as

$$\bar{H}(\mathbf{U}) \triangleq \lim_{k \to \infty} H\left(U_{k+1}|U^k\right). \tag{7.1}$$

Consider $u^n \in \mathcal{U}^n$, where $\mathcal{U}$ is a finite set. The $(k + 1)$th order empirical distribution of $u^n$ is defined in (1.2). The $k$th order conditional empirical entropy of $u^n$ is defined as $\hat{H}_k(u^n) = H(U_{k+1}|U^k)$, where

$U^{k+1}$ is distributed as $\hat{p}^{(k+1)}(\cdot|u^n)$. In other words,

$$\hat{H}_k(u^n) = -\sum_{a^{k+1} \in \mathcal{U}^{k+1}} \hat{p}^{(k+1)}(a^{k+1}|u^n) \log \frac{\hat{p}^{(k+1)}(a^{k+1}|u^n)}{\hat{p}^{(k)}(a^k|u^n)}. \tag{7.2}$$

In some of the proofs, we employ a compression scheme called *Lempel-Ziv*. Compression schemes aim to represent a sequence $u^n \in \mathcal{U}^n$ in as few bits as possible. It turns out that intuitively if $u^n$ is a sample of a finite-alphabet stationary ergodic process **U**, asymptotically, the smallest expected number of bits per symbol required to represent $u^n$ is $\bar{H}(\mathbf{U})$. Compression algorithms that achieve this bound are called optimal. One of the well-known examples of optimal compression schemes is Lempel–Ziv (LZ) [34] coding. (LZ is also a universal compression code, since it does not use any information regarding the distribution of the process.)

In summary, the LZ compression code operates at follows: it first incrementally parses the input sequence into unique phrases such that each phrase is the shortest phrase that is not seen earlier. Then, each phrase is encoded by (i) an index to the location of the phrase consisting of the current phrase except its last symbol, and (ii) last symbol of the phrase.

Given $u^n \in \mathcal{U}^n$, let $\ell_{\mathrm{LZ}}(u^n)$ denote the length of the binary coded sequence assigned to $u^n$ using the LZ compression code. Note that since LZ algorithm assigns a unique coded sequence to every input sequence, we have

$$|\{u^n : \ell_{\mathrm{LZ}}(u^n) \le r\}| \le \sum_{i=1}^{r} 2^i \le 2^{r+1}. \tag{7.3}$$

The LZ length function $\ell_{\mathrm{LZ}}(\cdot)$ is mentioned in some of the following proofs because of its connections with the conditional empirical entropy function $\hat{H}_k(\cdot)$. This connection established in [22] for binary sources and extended in [15] to general sources with alphabet $\mathcal{U}$ such that $|\mathcal{U}| = 2^b$ states that, for all $u^n \in \mathcal{U}^n$,

$$\frac{1}{n}\ell_{\mathrm{LZ}}(u^n) \le \hat{H}_k(u^n) + \frac{b(kb+b+3)}{(1-\epsilon_n)\log n - b} + \gamma_n, \tag{7.4}$$

where

$$\epsilon_n = \frac{2b + \log\left(2^b + \frac{2^b-1}{b}\log n - 2\right)}{\log n}, \tag{7.5}$$

and $\gamma_n = o(1)$ does not depend on sequence $u^n$ or $b$.

Finally, we finish this section, by two lemmas related to continuity properties of the entropy function and the Kullback–Leibler distance.

LEMMA 7.1 (Theorem 17.3.3 in [6]) Consider distributions $p$ and $q$ on finite alphabet $\mathcal{U}$ such that $\|p - q\|_1 \le \epsilon$. Then,

$$|H(p) - H(q)| \le -\epsilon \log \epsilon + \epsilon \log |\mathcal{U}|. \tag{7.6}$$

LEMMA 7.2 Consider distributions $p$ and $q$ over discrete set $\mathcal{U}$ such that $\|p - q\|_1 \leq \epsilon$. Further assume that $p \ll q$, and let $q_{\min} = \min_{u \in \mathcal{U}: q(u) \neq 0} q(u)$. Then,

$$D(p\|q) \leq -\epsilon \log \epsilon + \epsilon \log |\mathcal{U}| - \epsilon \log q_{\min}. \tag{7.7}$$

*Proof.* Let $\mathcal{U}^* \triangleq \{u \in \mathcal{U} : q(u) \neq 0\}$. Since $p \ll q$, if $q(u) = 0$, then $p(u) = 0$. Therefore, by definition

$$\begin{aligned}
D(p\|q) &= \sum_{u \in \mathcal{U}} p(u) \log \frac{p(u)}{q(u)} \\
&= \sum_{u \in \mathcal{U}^*} p(u) \log \frac{p(u)}{q(u)} \\
&= \sum_{u \in \mathcal{U}^*} p(u) \log p(u) - \sum_{u \in \mathcal{U}^*} p(u) \log q(u) \\
&= \sum_{u \in \mathcal{U}^*} p(u) \log p(u) - \sum_{u \in \mathcal{U}^*} (p(u) - q(u) + q(u)) \log q(u) \\
&= H(q) - H(p) - \sum_{u \in \mathcal{U}^*} (p(u) - q(u)) \log q(u). \tag{7.8}
\end{aligned}$$

Hence, by the triangle inequality,

$$\begin{aligned}
D(p\|q) &\leq |H(q) - H(p)| - \sum_{u \in \mathcal{U}^*} |p(u) - q(u)| \log q(u) \\
&\overset{(a)}{\leq} -\epsilon \log \epsilon + \epsilon \log |\mathcal{U}| - \sum_{u \in \mathcal{U}^*} |p(u) - q(u)| \log q(u) \\
&\leq -\epsilon \log \epsilon + \epsilon \log |\mathcal{U}| + \log \left( \frac{1}{q_{\min}} \right) \sum_{u \in \mathcal{U}^*} |p(u) - q(u)| \\
&\overset{(b)}{\leq} -\epsilon \log \epsilon + \epsilon \log |\mathcal{U}| - \epsilon \log q_{\min}, \tag{7.9}
\end{aligned}$$

where (*a*) and (*b*) follow from Theorem 17.3.3 in [6] and $\|p - q\|_1 \leq \epsilon$, respectively. □

### 7.2 *Useful concentration lemmas*

LEMMA 7.3 Consider $u^n \in \mathbb{R}^n$ and $v^n \in \mathbb{R}^n$ such that $\|u^n\| = \|v^n\| = 1$. Let $\alpha \triangleq \langle u^n, v^n \rangle$. Consider matrix $A \in \mathbb{R}^{m \times n}$ with i.i.d. standard normal entries. Then, for any $\tau > 0$,

$$\mathrm{P}\left( \frac{1}{m} \langle Au^n, Av^n \rangle - \langle u^n, v^n \rangle \leq -\tau \right) \leq \mathrm{e}^{m((\alpha - \tau)s) - \frac{m}{2} \ln((1 + s\alpha)^2 - s^2)}, \tag{7.10}$$

where $s > 0$ is a free parameter smaller than $\frac{1}{1 - \alpha}$.

*Proof.* Let $A_i^n$ denote the $i$th row of matrix $A$. Then,

$$
Au^n = \begin{bmatrix} \langle A_1^n, u^n \rangle \\ \langle A_2^n, u^n \rangle \\ \vdots \\ \langle A_m^n, u^n \rangle \end{bmatrix}, \quad Av^n = \begin{bmatrix} \langle A_1^n, v^n \rangle \\ \langle A_2^n, v^n \rangle \\ \vdots \\ \langle A_m^n, v^n \rangle \end{bmatrix}
\tag{7.11}
$$

and

$$
\frac{1}{m} \langle Au^n, Av^n \rangle = \frac{1}{m} \sum_{i=1}^{m} \langle A_i^n, u^n \rangle \langle A_i^n, v^n \rangle.
\tag{7.12}
$$

Let

$$
X_i = \langle A_i^n, u^n \rangle
\tag{7.13}
$$

and

$$
Y_i = \langle A_i^n, v^n \rangle.
\tag{7.14}
$$

Since $A$ is generated by drawing its entries from an i.i.d. standard normal distribution, and $\|u^n\| = \|v^n\| = 1$, $\{(X_i, Y_i)\}_{i=1}^{m}$ is a sequence of i.i.d. random vectors. To derive the joint distribution of $(X_i, Y_i)$, note that both $X_i$ and $Y_i$ are linear combination of Gaussian random variables. Therefore, they are also jointly distributed Gaussian random variables, and hence it suffices to find their first- and second-order moments. Note that

$$
\mathrm{E}[X_i] = \mathrm{E}[Y_i] = 0,
\tag{7.15}
$$

$$
\mathrm{E}[X_i^2] = \sum_{j,k} \mathrm{E}[A_{i,j} A_{i,k}] u_j u_k = \sum_{j} \mathrm{E}[A_{i,j}^2] u_j^2 = \sum_{j} u_j^2 = 1
\tag{7.16}
$$

and similarly $\mathrm{E}[Y_i^2] = 1$. Also,

$$
\begin{aligned}
\mathrm{E}[X_i Y_i] &= \mathrm{E}[\langle A_i^n, u^n \rangle \langle A_i^n, v^n \rangle] \\
&= \sum_{j,k} \mathrm{E}[A_{i,j} A_{i,k}] u_j v_k \\
&= \sum_{j} \mathrm{E}[A_{i,j}^2] u_j v_j \\
&= \langle u^n, v^n \rangle = \alpha.
\end{aligned}
\tag{7.17}
$$

Therefore, in summary,

$$
(X_i, Y_i) \sim \mathcal{N} \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix} \right).
\tag{7.18}
$$

For any $s' > 0$, by the Chernoff bounding method, we have

$$P\left(\frac{1}{m}\langle Au^n, Av^n\rangle - \langle u^n, v^n\rangle \le -\tau\right) = P\left(\frac{1}{m}\sum_{i=1}^{m}(X_iY_i - \alpha) \le -\tau\right)$$

$$= P\left(\frac{s'}{m}\sum_{i=1}^{m}(X_iY_i - \alpha) \le -s'\tau\right)$$

$$= P\left(e^{s'(\tau-\alpha)} \le e^{-\frac{s'}{m}\sum_{i=1}^{m}X_iY_i}\right)$$

$$\le e^{s'(\alpha-\tau)} E\left[e^{-\frac{s'}{m}\sum_{i=1}^{m}X_iY_i}\right]$$

$$= e^{s'(\alpha-\tau)}\left(E\left[e^{-\frac{s'}{m}X_1Y_1}\right]\right)^m, \tag{7.19}$$

where the last line follows because $(X_i, Y_i)$ is an i.i.d. sequence. In the following we compute $E[e^{\frac{s}{m}X_1Y_1}]$. Let $A = (X_1 + Y_1)/2$ and $B = (X_1 - Y_1)/2$. Then, $E[A] = E[B] = 0$ and

$$E[A^2] = \frac{1+\alpha}{2}, \tag{7.20}$$

$$E[B^2] = \frac{1-\alpha}{2} \tag{7.21}$$

and $E[AB] = E[(X_1^2 - Y_1^2)/4] = 0$. Therefore, $A$ and $B$ are independent Gaussian random variables. Therefore,

$$E[e^{-\frac{s'}{m}X_1Y_1}] = E\left[e^{-\frac{s'}{m}(A+B)(A-B)}\right]$$

$$= E\left[e^{-\frac{s'}{m}A^2}\right]E\left[e^{\frac{s'}{m}B^2}\right]. \tag{7.22}$$

Given $Z \sim \mathcal{N}(0, \sigma^2)$, it is straightforward to show that, for $\lambda > -1/(2\sigma^2)$,

$$E\left[e^{-\lambda Z^2}\right] = \frac{1}{\sqrt{1 + 2\lambda\sigma^2}}. \tag{7.23}$$

Therefore, for $\frac{s'}{m} < \frac{1}{1-\alpha}$,

$$E[e^{\frac{s'}{m}X_1Y_1}] = \frac{1}{\sqrt{(1 + \frac{s'}{m}(1+\alpha))(1 - \frac{s'}{m}(1-\alpha))}}$$

$$= \frac{1}{\sqrt{(1 + \frac{s'\alpha}{m})^2 - \frac{s'^2}{m^2}}}. \tag{7.24}$$

Therefore, combining (7.19) and (7.24), it follows that

$$P\left(\frac{1}{m}\langle Au^n, Av^n\rangle - \langle u^n, v^n\rangle \leq -\tau\right) = e^{s'(\alpha-\tau)}e^{-\frac{m}{2}\log\left(\left(1+\frac{s'\alpha}{m}\right)^2 - \left(\frac{s'}{m}\right)^2\right)}. \tag{7.25}$$

Replacing $s'/m$ with $s$ in (7.25) yields the desired result. $\qquad\square$

COROLLARY 7.1 Consider $u^n \in \mathbb{R}^n$ and $v^n \in \mathbb{R}^n$ such that $\|u^n\| = \|v^n\| = 1$. Also, consider matrix $A \in \mathbb{R}^{m\times n}$ with i.i.d. standard normal entries. Then,

$$P\left(\frac{1}{m}\langle Au^n, Av^n\rangle - \langle u^n, v^n\rangle \leq -0.45\right) \leq 2^{-0.05m}. \tag{7.26}$$

*Proof.* From Lemma 7.3, for $\alpha = \langle u^n, v^n\rangle$, and $s < 1/(1-\alpha)$,

$$P\left(\frac{1}{m}\langle Au^n, Av^n\rangle - \alpha \leq -0.45\right) \leq e^{m((\alpha-0.45)s) - \frac{m}{2}\ln((1+s\alpha)^2 - s^2)} = 2^{-mf(\alpha,s)}, \tag{7.27}$$

where

$$f(\alpha, s) = (\log e)\left(\frac{1}{2}\ln((1+s\alpha)^2 - s^2) - (\alpha - 0.45)s\right). \tag{7.28}$$

Figure 1 plots $\max_{s\in(0,\frac{1}{1-\alpha})} f(\alpha, s)$ and shows that

$$\min_{\alpha\in(-1,1)} \max_{s\in\left(0,\frac{1}{1-\alpha}\right)} f(\alpha, s) \geq 0.05. \tag{7.29}$$

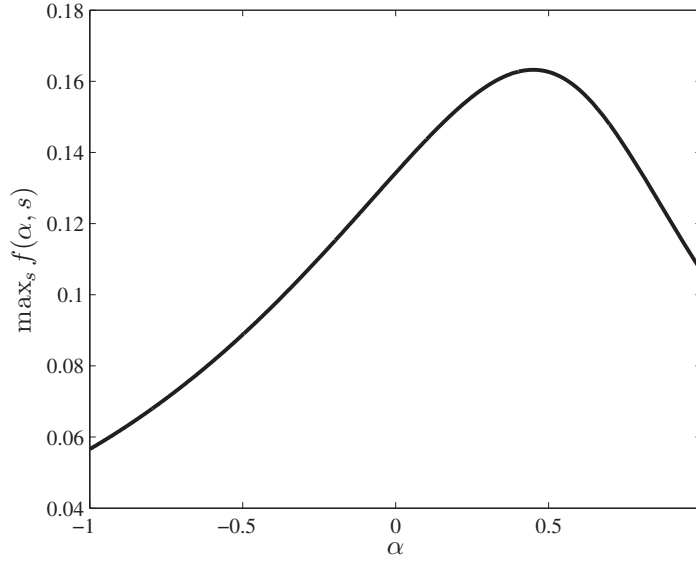$\qquad\square$

The following two lemmas are proved in [13].

LEMMA 7.4 ($\chi^2$ concentration) Fix $\tau > 0$, and let $U_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $i = 1, 2, \ldots, m$. Then,

$$P\left(\sum_{i=1}^m U_i^2 < m(1-\tau)\right) \leq e^{\frac{m}{2}(\tau + \ln(1-\tau))} \tag{7.30}$$

and

$$P\left(\sum_{i=1}^m U_i^2 > m(1+\tau)\right) \leq e^{-\frac{m}{2}(\tau - \ln(1+\tau))}. \tag{7.31}$$

LEMMA 7.5 Consider $U^n$ and $V^n$, where, for each $i$, $U_i$ and $V_i$ are two independent standard normal random variables. Then the distribution of $\langle U^n, V^n\rangle = \sum_{i=1}^n U_i V_i$ is the same as the distribution of $\|U^n\|G$, where $G \sim \mathcal{N}(0,1)$ is independent of $\|U^n\|_2$.

FIG. 1. $\max_{s \in (0, \frac{1}{1-\alpha})} f(\alpha, s)$.

### 7.3 *Proof of Lemma* 4.1

Before presenting the proof, we establish some preliminary results. Consider an analog process $\mathbf{X} = \{X_i\}$ with alphabet $\mathcal{X} = [l, u]$, where $l, u \in \mathbb{R}$. Let process $\mathbf{Z} = \{Z_i\}$ denote the $b$-bit quantized version of process $\mathbf{X}$. That is, $Z_i = [X_i]_b$, and the alphabet of process $\mathbf{Z}$ is $\mathcal{Z} = \mathcal{X}_b = \{[x]_b : x \in \mathcal{X}_b\}$. For $i = 1, \ldots, k + g$, define a sequence of length $t$, $\{S_j^{(i)}\}_{j=1}^t$ over super-alphabet $\mathcal{Z}^k$ as follows. For $i = 1$, $\{S_j^{(1)}\}_{j=1}^t$ is defined as

$$\underbrace{Z_1, \ldots, Z_k}_{S_1^{(1)}}, Z_{k+1}, \ldots, Z_{k+g}, \underbrace{Z_{k+g+1}, \ldots, Z_{2k+g+1}}_{S_2^{(1)}}, \tag{7.32}$$

$$Z_{2k+g+2}, \ldots \ldots, \underbrace{Z_{(t-1)(k+g)+1}, \ldots, Z_{(t-1)(k+g)+k}}_{S_t^{(1)}}, Z_{t(k+g)-g+1}, \ldots, Z_n. \tag{7.33}$$

Similarly, $\left\{ S_j^{(i)} \right\}_{j=1}^t$, $i = 1, \ldots, k + g$, is defined by starting the grouping of the symbols at $Z_i$. In other words,

$$S_j^{(i)} \triangleq Z_{(k+g)(j-1)+i}^{(k+g)(j-1)+i+k-1}. \tag{7.34}$$

For instance, the sequence $\left\{ S_j^{(k+g)} \right\}_{j=1}^t$, which corresponds to the largest shift at beginning, is defined as

$$Z_1, \ldots, Z_{k+g-1}, \underbrace{Z_{k+g}, \ldots, Z_{2k+g-1}}_{S_1^{(k)}}, Z_{2k+g}, \ldots, Z_{2k+2g-1}, \tag{7.35}$$

$$\underbrace{Z_{2k+2g}, \ldots, Z_{3k+2g-1}}_{S_2^{(k)}}, Z_{3k+2g} \ldots . \tag{7.36}$$

This definition implies that $n$ satisfies

$$(t-1)(k+g) + 2k + g - 1 \le n < (t-1)(k+g) + 2k + g - 1 + k + g. \tag{7.37}$$

That is, $t$ is the only integer in the $\left(\frac{n-2k-g+1}{k+g}, \frac{n-k+1}{k+g}\right]$ interval, or in other words, $t = \lfloor \frac{n-k+1}{k+g} \rfloor$.

Before we prove Lemma 4.1, we prove the following auxiliary lemma.

LEMMA 7.6 For any given $\epsilon > 0$ and any positive integers $g$ and $k$ such that $4(k+g)/(n-k) < \epsilon$, if $\|\hat{p}^{(k)}(\cdot|Z^n) - \mu_k^{(b)}\|_1 \ge \epsilon$, then that there exists $i = 1, \ldots, k+g$, such that

$$\|\hat{p}^{(1)}(\cdot|S^{(i),t}) - \mu_k^{(b)}(\cdot)\|_1 \ge \frac{\epsilon}{2}, \tag{7.38}$$

where $S^{(i),t}$ denotes the sequence $S_1^{(i)}, S_2^{(i)}, \ldots, S_t^{(i)}$.

Note that in Lemma 7.6, $\hat{p}^{(1)}(\cdot|S^{(i),t})$ denotes the standard first order empirical distribution of the sup-alphabet sequence $S^{(i),t}$, i.e. for $a^k \in \mathcal{X}^k$,

$$\hat{p}^{(1)}(a^k|S^{(i),t}) = \frac{|\{j : S_j^{(i)} = a^k, 1 \le j \le t\}|}{t}. \tag{7.39}$$

*Proof.* Note that by definition, for any $a^k \in \mathcal{Z}^k$,

$$\hat{p}^{(k)}(a^k|Z^n) = \frac{1}{n-k} \sum_{i=k}^{n} 1_{Z_{i-k+1}^i = a^k}$$

$$= \frac{1}{n-k} \sum_{i=k}^{n} 1_{Z_{i-k+1}^i = a^k}$$

$$= \frac{1}{n-k} \left( \sum_{i=1}^{k+g} \sum_{j=1}^{t} 1_{S_j^{(i)} = a^k} + \sum_{t(k+g)+k}^{n} 1_{Z_{i-k+1}^i = a^k} \right)$$

$$= \frac{t}{n-k} \sum_{i=1}^{k+g} \left( \frac{1}{t} \sum_{j=1}^{t} 1_{S_j^{(i)} = a^k} \right) + \frac{1}{n-k} \sum_{t(k+g)+k}^{n} 1_{Z_{i-k+1}^i = a^k}$$

$$= \frac{t}{n-k} \sum_{i=1}^{k+g} \hat{p}^{(1)}(a^k|S^{(i),t}) + \frac{1}{n-k} \sum_{t(k+g)+k}^{n} 1_{Z_{i-k+1}^i = a^k}. \tag{7.40}$$

Therefore,

$$\|\hat{p}^{(k)}(\cdot|Z^n) - \mu_k^{(b)}\|_1 = \sum_{a^k} |\hat{p}^{(k)}(a^k|Z^n) - \mu_k^{(b)}(a^k)|$$

$$\stackrel{(a)}{=} \sum_{a^k} \left| \frac{t}{n-k} \sum_{i=1}^{k+g} \hat{p}^{(1)}(a^k|S^{(i),t}) + \frac{1}{n-k} \sum_{t(k+g)+k}^{n} 1_{Z_{i-k+1}^i = a^k} - \mu_k^{(b)}(a^k) \right|$$

$$\stackrel{(b)}{=} \sum_{a^k} \left| \frac{t}{n-k} \sum_{i=1}^{k+g} \hat{p}^{(1)}(a^k|S^{(i),t}) - \mu_k^{(b)}(a^k) \right| + \frac{1}{n-k} \sum_{a^k} \sum_{t(k+g)+k}^{n} 1_{Z_{i-k+1}^i = a^k}$$

$$\stackrel{(c)}{=} \sum_{a^k} \left| \frac{t}{n-k} \sum_{i=1}^{k+g} \hat{p}^{(1)}(a^k|S^{(i),t}) - \mu_k^{(b)}(a^k) \right| + \frac{n - t(k+g) - k + 1}{n-k}, \qquad (7.41)$$

where (*a*) follows from (7.40), (*b*) follows from the triangle inequality and (*c*) holds because

$$\sum_{a^k} 1_{Z_{i-k+1}^i = a^k} = 1. \qquad (7.42)$$

But since $n$ satisfies the bounds of (7.37), the last term on the right-hand side of (7.40) can be upper-bounded as

$$\frac{n - t(k+g) - k + 1}{n-k} \leq \frac{k+g}{n-k}. \qquad (7.43)$$

Therefore, from (7.41) we have

$$\|\hat{p}^{(k)}(\cdot|Z^n) - \mu_k^{(b)}\|_1 \leq \sum_{a^k} \left| \frac{t}{n-k} \sum_{i=1}^{k+g} \hat{p}^{(1)}(a^k|S^{(i),t}) - \mu_k^{(b)}(a^k) \right| + \frac{k+g}{n-k}. \qquad (7.44)$$

On the other hand, again by the triangle inequality,

$$\sum_{a^k} \left| \frac{t}{n-k} \sum_{i=1}^{k+g} \hat{p}^{(1)}(a^k|S^{(i),t}) - \mu_k^{(b)}(a^k) \right|$$

$$= \sum_{a^k} \left| \frac{t}{n-k} \sum_{i=1}^{k+g} \left( \hat{p}^{(1)}(a^k|S^{(i),t}) - \mu_k^{(b)}(a^k) \right) + \left( \frac{t(k+g)}{n-k} - 1 \right) \mu_k^{(b)}(a^k) \right|$$

$$\leq \sum_{a^k} \left| \frac{t}{n-k} \sum_{i=1}^{k+g} \left( \hat{p}^{(1)}(a^k|S^{(i),t}) - \mu_k^{(b)}(a^k) \right) \right| + \left| \frac{t(k+g)}{n-k} - 1 \right| \sum_{a^k} \mu_k^{(b)}(a^k)$$

$$= \sum_{a^k} \left| \frac{t}{n-k} \sum_{i=1}^{k+g} \left( \hat{p}^{(1)}(a^k|S^{(i),t}) - \mu_k^{(b)}(a^k) \right) \right| + \left| \frac{t(k+g)}{n-k} - 1 \right|$$

$$\leq \frac{t}{n-k} \sum_{i=1}^{k+g} \sum_{a^k} \left| \hat{p}^{(1)}(a^k|S^{(i),t}) - \mu_k^{(b)}(a^k) \right| + \left| \frac{t(k+g)}{n-k} - 1 \right|$$

$$= \frac{t}{n-k} \sum_{i=1}^{k+g} \left\| \hat{p}^{(1)}(\cdot|S^{(i),t}) - \mu_k^{(b)}(\cdot) \right\|_1 + \left| \frac{t(k+g)}{n-k} - 1 \right|. \tag{7.45}$$

Therefore, combining (7.41) and (7.45) yields

$$\left\| \hat{p}^{(k)}(\cdot|Z^n) - \mu_k^{(b)} \right\|_1 \leq \frac{t}{n-k} \sum_{i=1}^{k+g} \left\| \hat{p}^{(1)}(\cdot|S^{(i),t}) - \mu_k^{(b)}(\cdot) \right\|_1 + \left| \frac{t(k+g)}{n-k} - 1 \right| + \frac{k+g}{n-k}. \tag{7.46}$$

Since by construction $t(k+g) \leq n-k$, we have $t/(n-k) \leq 1/(k+g)$. Therefore, it follows from (7.46) that:

$$\left\| \hat{p}^{(k)}(\cdot|Z^n) - \mu_k^{(b)} \right\|_1 \leq \frac{1}{k+g} \sum_{i=1}^{k+g} \left\| \hat{p}^{(1)}(\cdot|S^{(i),t}) - \mu_k^{(b)}(\cdot) \right\|_1 + 1 - \frac{t(k+g)}{n-k} + \frac{k+g}{n-k}$$

$$= \frac{1}{k+g} \sum_{i=1}^{k+g} \left\| \hat{p}^{(1)}(\cdot|S^{(i),t}) - \mu_k^{(b)}(\cdot) \right\|_1 + \frac{n - t(k+g) + g}{n-k}. \tag{7.47}$$

Notice that if

$$\frac{n - t(k+g) + g}{n-k} \leq \frac{\epsilon}{2}, \tag{7.48}$$

and

$$\left\| \hat{p}^{(1)}(\cdot|S^{(i),t}) - \mu_k^{(b)}(\cdot) \right\|_1 \leq \frac{\epsilon}{2}, \tag{7.49}$$

for all $i$, then, from (7.47), $\|\hat{p}^{(k)}(\cdot|Z^n) - \mu_k^{(b)}\|_1 \leq \epsilon$. But if $4(k+g)/(n-k) < \epsilon$, since $t = \lfloor \frac{n-k+1}{k+g} \rfloor$, then (7.48) holds. This means that, to have $\|\hat{p}^{(k)}(\cdot|Z^n) - \mu_k^{(b)}\|_1 > \epsilon$, we need $\|\hat{p}^{(1)}(\cdot|S^{(i),t}) - \mu_k^{(b)}(\cdot)\|_1 > \frac{\epsilon}{2}$, for at least one $i$ in $\{1, \dots, k+g\}$. $\qquad\square$

Now we can discuss the proof of Lemma 4.1. The proof is a straightforward extension of Lemma III.1.3 in [**25**]. However, we include a summary of the proof for completeness. By Lemma 7.6, if $4(k+g)/(n-k) < \epsilon$, then $\|\hat{p}^{(k)}(\cdot|Z^n) - \mu_k^{(b)}\|_1 \geq \epsilon$ implies that there exists $i \in \{1, \dots, k+g\}$ such that

$$\left\| \hat{p}^{(1)}(\cdot|S^{(i),t}) - \mu_k^{(b)}(\cdot) \right\|_1 \geq \frac{\epsilon}{2}. \tag{7.50}$$

We next bound the probability that the above event happens. For each $i \in \{1, \dots, k+g\}$, define event $\mathcal{E}^{(i)}$ as follows

$$\mathcal{E}^{(i)} \triangleq \left\{ D_{\mathrm{KL}}(\hat{p}^{(1)}(\cdot|S^{(i),t}), \mu_k^{(b)}) > \epsilon^2/2 \right\}. \tag{7.51}$$

By the Pinsker's inequality, for any $i$,

$$\left\| \hat{p}^{(1)}(\cdot|S^{(i),t}) - \mu_k^{(b)} \right\|_1 \leq \sqrt{(2\ln 2)D_{\mathrm{KL}}(\hat{p}^{(1)}(\cdot|S^{(i),t}), \mu_k^{(b)}(\cdot))}. \tag{7.52}$$

Therefore, if $D_{\mathrm{KL}}(\hat{p}^{(1)}(\cdot|S^{(i),t}), \mu_k^{(b)}(\cdot)) \leq c\epsilon^2/4$, where as defined earlier $c = 1/(2\ln 2)$, then $\|\hat{p}^{(1)}(\cdot|S^{(i),t}) - \mu_k^{(b)}(\cdot)\|_1 \leq \epsilon/2$. This implies that

$$\mathrm{P}\left(\left\| \hat{p}^{(1)}(\cdot|S^{(i),t}) - \mu_k^{(b)} \right\|_1 > \frac{\epsilon}{2}\right) \leq \mathrm{P}\left(D_{\mathrm{KL}}\left(\hat{p}^{(1)}\left(\cdot|S^{(i),t}\right), \mu_k^{(b)}\right) > \frac{c\epsilon^2}{4}\right). \tag{7.53}$$

On the other hand, for $S^{(i),t} = s^t$, where $s^t \in (\mathcal{Z}^k)^t$, we have

$$\mathrm{P}(S^{(i),t} = s^t) = \mathrm{P}\left(Z_{(k+g)(j-1)+i}^{(k+g)(j-1)+i+k-1} = s_j, \, j = 1, \ldots, t\right)$$

$$\overset{(a)}{\leq} \Psi^t(b,g) \prod_{j=1}^{t} \mathrm{P}\left(Z_{(k+g)(j-1)+i}^{(k+g)(j-1)+i+k-1} = s_j\right), \tag{7.54}$$

where $(a)$ follows from applying condition (4.4) $t$ times. However, by the standard method of types techniques [7], we have

$$\prod_{j=1}^{t} \mathrm{P}\left(Z_{(k+g)(j-1)+i}^{(k+g)(j-1)+i+k-1} = s_j\right) = 2^{-t\left(\hat{H}_1(s^t) + D_{\mathrm{KL}}\left(\hat{p}^{(1)}(\cdot|s^t), \mu_k^{(b)}\right)\right)}. \tag{7.55}$$

Therefore, if $D_{\mathrm{KL}}(\hat{p}^{(1)}(\cdot|s^t), \mu_k^{(b)}) > \frac{c\epsilon^2}{4}$, then

$$\prod_{j=1}^{t} \mathrm{P}\left(Z_{(k+g)(j-1)+i}^{(k+g)(j-1)+i+k-1} = s_j\right) \leq 2^{-t\hat{H}_1(s^t) - c\epsilon^2 t/4}. \tag{7.56}$$

Hence,

$$\mathrm{P}\left(D_{\mathrm{KL}}\left(\hat{p}^{(1)}\left(\cdot|S^{(i),t}\right), \mu_k^{(b)}\right) > \frac{c\epsilon^2}{4}\right) = \sum_{s^t: D_{\mathrm{KL}}\left(\hat{p}^{(1)}(\cdot|s^t), \mu_k^{(b)}\right) > \frac{c\epsilon^2}{4}} \mathrm{P}\left(S^{(i),t} = s^t\right)$$

$$\leq 2^{-c\epsilon^2 t/4} \Psi^t(b,g) \sum_{s^t: D_{\mathrm{KL}}\left(\hat{p}^{(1)}(\cdot|s^t), \mu_k^{(b)}\right) > \frac{c\epsilon^2}{4}} 2^{-t\hat{H}_1(s^t)}$$

$$\leq 2^{-c\epsilon^2 t/4} \Psi^t(b,g) \sum_{s^t} 2^{-t\hat{H}_1(s^t)}. \tag{7.57}$$

Since $\sum_{s^t} 2^{-t\hat{H}_1(s^t)}$ can be proven to be smaller than the total number of types of sequences $s^t \in (\mathcal{Z}^k)^t$, we have $\sum_{s^t} 2^{-t\hat{H}_1(s^t)} \leq (t+1)^{|\mathcal{Z}|^k}$. This upper bound combined by the union bound on $\mathcal{E}^{(i)}$, $i = 1, \ldots, k+g$, yields the desired result.

### 7.4 *Proof of Lemma* 4.2

For any $u^{\ell_1} \in \mathcal{Z}^{\ell_1}$, $v^g \in \mathcal{Z}^g$, and $w^{\ell_2} \in \mathcal{Z}^{\ell_2}$, we have

$$
\begin{aligned}
& \mathrm{P}\left(Z^{\ell_1+g+\ell_2} = [u^{\ell_1} v^g w^{\ell_2}]\right) \\
& \leq \sum_{v^g \in \mathcal{Z}^g} \mathrm{P}\left(Z^{\ell_1+g+\ell_2} = [u^{\ell_1} v^g w^{\ell_2}]\right) \\
& = \mathrm{P}\left(Z^{\ell_1} = u^{\ell_1}, Z^{\ell_1+g+\ell_2}_{\ell_1+g+1} = w^{\ell_2}\right) \\
& = \mathrm{P}\left(Z^{\ell_1} = u^{\ell_1}\right) \mathrm{P}\left(Z_{\ell_1+g+1} = w_1 | Z^{\ell_1} = u^{\ell_1}\right) \\
& \quad \times \mathrm{P}\left(Z^{\ell_1+g+\ell_2}_{\ell_1+g+2} = w_2^{\ell_2} | Z_{\ell_1+g+1} = w_1, Z^{\ell_1} = u^{\ell_1}\right) \\
& = \mathrm{P}\left(Z^{\ell_1} = u^{\ell_1}\right) \int_{\mathcal{X}} \mathrm{P}\left(Z_{\ell_1+g+1} = w_1 | X_{\ell_1} = x_{\ell_1}, Z^{\ell_1} = u^{\ell_1}\right) \mathrm{d}\mu\left(x_{\ell_1} | Z^{\ell_1} = u^{\ell_1}\right) \\
& \quad \times \mathrm{P}\left(Z^{\ell_1+g+\ell_2}_{\ell_1+g+2} = w_2^{\ell_2} | Z_{\ell_1+g+1} = w_1, Z^{\ell_1} = u^{\ell_1}\right) \\
& \stackrel{(a)}{=} \mathrm{P}\left(Z^{\ell_1} = u^{\ell_1}\right) \int_{\mathcal{X}} \mathrm{P}\left(Z_{\ell_1+g+1} = w_1 | X_{\ell_1} = x_{\ell_1}\right) \mathrm{d}\mu\left(x_{\ell_1} | Z^{\ell_1} = u^{\ell_1}\right) \\
& \quad \times \mathrm{P}\left(Z^{\ell_1+g+\ell_2}_{\ell_1+g+2} = w_2^{\ell_2} | Z_{\ell_1+g+1} = w_1, Z^{\ell_1} = u^{\ell_1}\right) \\
& \stackrel{(b)}{=} \mu_b(u^{\ell_1}) \left( \int_{\mathcal{X}} K^{g+1}(x_{\ell_1}, w_1) \mathrm{d}\mu\left(x_{\ell_1} | u^{\ell_1}\right) \right) \mu_b\left(w_2^{\ell_2} | w_1, u^{\ell_1}\right) \\
& = \mu_b(u^{\ell_1}) \left( \int_{\mathcal{X}^l} \frac{K^{g+1}(x_{\ell_1}, w_1)}{\pi(w_1)} \mathrm{d}\mu(x_{\ell_1} | u^{\ell_1}) \right) \pi(w_1) \mu_b\left(w_2^{\ell_2} | w_1, u^{\ell_1}\right) \\
& \leq \mu_b(u^{\ell_1}) \pi(w_1) \mu_b(w_2^{\ell_2} | w_1, u^{\ell_1}) \left( \sup_{x:[x]_b = u_{\ell_1}} \frac{K^{g+1}(x, w_1)}{\pi(w_1)} \right) \\
& \leq \mu_b(u^{\ell_1}) \pi(w_1) \mu_b(w_2^{\ell_2} | w_1, u^{\ell_1}) \left( \sup_{(x,z) \in \mathcal{X} \times \mathcal{Z}} \frac{K^{g+1}(x, z)}{\pi(z)} \right) \\
& = \mu_b(u^{\ell_1}) \pi(w_1) \mu_b(w_2^{\ell_2} | w_1, u^{\ell_1}) \Psi_1(b, g), \quad\quad (7.58)
\end{aligned}
$$

where (a) holds because **X** is a first-order Markov chain and in (b),

$$
\mu_b\left(u^{\ell_1}\right) = \mathrm{P}\left(Z^{\ell_1} = u^{\ell_1}\right), \quad\quad (7.59)
$$

$$
\mu_b\left(w_2^{\ell_2} | w_1, u^{\ell_1}\right) = \mathrm{P}\left(Z^{\ell_1+g+\ell_2}_{\ell_1+g+2} = w_2^{\ell_2} | Z_{\ell_1+g+1} = w_1, Z^{\ell_1} = u^{\ell_1}\right) \quad\quad (7.60)
$$

and $\mu(x_{\ell_1} | u^{\ell_1})$ denotes the probability measure of $X_{\ell_1}$ conditioned on $Z^{\ell_1} = u^{\ell_1}$. Also, since the Markov chain is a stationary process, we have

$$
K^{g+1}(x_{\ell_1}, w_1) = \mathrm{P}([X_{g+1}]_b = w_1 | X_0 = x_{\ell_1}). \qu\quad (7.61)
$$

Another term in (7.58) is $\pi(w_1)\mu_b(w_2^{\ell_2}|w_1, u^{\ell_1})$. Since $\pi(w^{\ell_2}) = \pi(w_1)\pi(w_2^{\ell_2}|w_1)$, we have

$$\pi(w^l)\mu_b(w_{l+1}^{\ell_2}|w^l, u^{\ell_1}) = \pi(w^{\ell_2})\frac{\mu_b(w_{l+1}^{\ell_2}|w^l, u^{\ell_1})}{\pi(w_{l+1}^{\ell_2}|w^l)}. \tag{7.62}$$

But

$$\mu_b(w_2^{\ell_2}|w_1, u^{\ell_1}) = \int \mu_b(w_2^{\ell_2}|x, w_1, u^{\ell_1}) \, d\mu(x|w_1, u^{\ell_1})$$

$$= \int \mu_b(w_2^{\ell_2}|x) \, d\mu(x|w_1, u^{\ell_1}), \tag{7.63}$$

where the second equality holds because $(Z^{\ell_1}, Z_{\ell_1+1}) \rightarrow X_{\ell_1+1} \rightarrow Z_{\ell_1+2}^{\ell_1+l+\ell_2}$ forms a Markov chain. Therefore,

$$\frac{\mu_b\left(w_2^{\ell_2}|w_1, u^{\ell_1}\right)}{\pi(w_2^{\ell_2}|w_1)} = \frac{\int \mu_b(w_2^{\ell_2}|x) \, d\mu(x|w_1, u^{\ell_1})}{\pi(w_2^{\ell_2}|w_1)},$$

$$\stackrel{(a)}{\leq} \frac{\left(\sup\limits_{[x]_b=w_1} \mu_b(w_2^{\ell_2}|x)\right)\int d\mu(x|w_1, u^{\ell_1})}{\pi(w_2^{\ell_2}|w_1)},$$

$$\stackrel{(b)}{=} \sup\limits_{[x]_b=w_1} \frac{\pi(w_2^{\ell_2}|x)}{\pi(w_2^{\ell_2}|w_1)}$$

$$\leq \sup\limits_{(x,w^{\ell_2}):[x]_b=w_1} \frac{\pi(w_2^{\ell_2}|x)}{\pi(w_2^{\ell_2}|w_1)}$$

$$\leq \Psi_2(b), \tag{7.64}$$

where (a) and (b) hold because $\mu(x|w_1, u^{\ell_1})$ is only non-zero when $x$ is such that $[x]_b = w_1$, and $\int d\mu(x|w_1, u^{\ell_1}) = 1$, respectively. Finally, combining (7.58), (7.62) and (7.64) yields the desired result.

We next prove that, for a fixed $b$, $\Psi_1(b, g)$ is non-increasing function of $g$. For any $x \in \mathcal{X}$ and $z \in \mathcal{X}_b$, we have

$$\frac{K^{g+1}(x, z)}{\pi(z)} = \frac{P([X_{g+1}]_b = z|X_0 = x)}{P([X_{g+1}]_b = z)}$$

$$= \frac{\int P([X_{g+1}]_b = z|X_1 = x', X_0 = x) \, d\mu(x'|X_0 = x)}{P([X_{g+1}]_b = z)}$$

$$\stackrel{(a)}{=} \frac{\int P([X_g]_b = z|X_0 = x') \, d\mu(x'|X_0 = x)}{P([X_g]_b = z)}$$

$$= \sup\limits_{x' \in \mathcal{X}} \frac{P([X_g]_b = z|X_0 = x')}{P([X_g]_b = z)} \int d\mu(x'|X_0 = x)$$

$$\stackrel{(b)}{\leq} \Psi_1(g, b), \tag{7.65}$$

where (*a*) follows because of the Markovity and stationarity assumptions and (*b*) follows because $\int d\mu(x'|X_0 = x) = 1$. Since the right-hand side of (7.65) only depends on $g$ and $b$, taking the supremum of the left-hand side over $(x, z) \in \mathcal{X} \times \mathcal{X}_b$ proves that

$$\Psi_1(g + 1) \leq \Psi_1(g). \tag{7.66}$$

Furthermore, since **X** is assumed to be an aperiodic Markov chain, $\lim_{g \to \infty} K^g(x, z) = \pi(z)$, for all $x$ and $z$. Therefore, $\Psi_1(g)$ converges to one, as $g \to \infty$.

### 7.5 *Proof of Theorem* 4.3

Define $\Psi(b, g) = \Psi_1(b, g)\Psi_2(b)$. Then it follows from Lemmas 4.1 and 4.2 that, given $\epsilon > 0$, for any positive integers $g$ and $k$ that satisfy $4(k + g)/(n - k) < \epsilon$,

$$P\left( \|\hat{p}^{(k)}(\cdot|Z^n) - \mu_k^{(b)}\|_1 \geq \epsilon \right) \leq (k + g)\Psi_1^t(b, g)\Psi_2^t(b)(t + 1)^{|\mathcal{Z}|^k} 2^{-c\epsilon^2 t/4}, \tag{7.67}$$

where $t = \lfloor \frac{n-k+1}{k+g} \rfloor$ and $c = 1/(2 \ln 2)$. Since by assumption $\lim_{b \to \infty} \Psi_2(b) = 1$, there exists $b_\epsilon$ such that for all $b \geq b_\epsilon$, $\Psi_2(b) \leq 2^{c\epsilon^2/16}$. But $b_n = \lceil r \log \log n \rceil$ is a diverging sequence of $n$. Therefore, there exists $n_\epsilon > 0$, such that for all $n \geq n_\epsilon$,

$$\Psi_2(b_n) \leq 2^{c\epsilon^2/16}. \tag{7.68}$$

On the other hand, by the theorem's assumption, there exists a sequence $g = g_n$, where $g = o(n)$, such that $\lim_{n \to \infty} \Psi_1(b_n, g_n) = 1$. Therefore, there exists $n'_\epsilon$ such that for all $n \geq n'_\epsilon$,

$$\Psi_1(b_n, g_n) \leq 2^{c\epsilon^2/16}. \tag{7.69}$$

Moreover, since $g = g_n = o(n)$ and $k$ is fixed, there exists $n''_\epsilon > 0$ such that for all $n \geq n''_\epsilon$,

$$\frac{4(k + g_n)}{n - k} < \epsilon. \tag{7.70}$$

Therefore, for $n > \max(n_\epsilon, n'_\epsilon, n''_\epsilon)$, from (7.67), (7.68), (7.69) and (7.70), we have

$$
\begin{aligned}
P(\|\hat{p}^{(k)}(\cdot|Z^n) - \mu_k^{(b)}\|_1 \geq \epsilon) &\leq (k + g)\Psi_1^t(b, g)\Psi_2^t(b)(t + 1)^{|\mathcal{Z}|^k} 2^{-c\epsilon^2 t/4} \\
&\leq (k + g)(t + 1)^{|\mathcal{Z}|^k} 2^{-tc\epsilon^2/8} \\
&\leq (k + g)n^{|\mathcal{Z}|^k} 2^{-tc\epsilon^2/8}, 
\end{aligned}
\tag{7.71}
$$

where the last line follows from the fact that $t + 1 \leq n$. But since $t = \lfloor \frac{n-k+1}{k+g} \rfloor$, $t \geq \frac{n-k}{k+g} - 1$. Hence,

$$
\begin{aligned}
P(\|\hat{p}^{(k)}(\cdot|Z^n) - \mu_k^{(b)}\|_1 \geq \epsilon) &\leq 2^{c\epsilon^2(1+k/(k+g))/8}(k + g)n^{|\mathcal{Z}|^k} 2^{-\frac{c\epsilon^2 n}{8(k+g)}} \\
&\leq 2^{c\epsilon^2/4}(k + g)n^{|\mathcal{Z}|^k} 2^{-\frac{c\epsilon^2 n}{8(k+g)}}, 
\end{aligned}
\tag{7.72}
$$

which is the desired result.

### 7.6 *Proof of Theorem* 4.4

For each $i$, let random variable $J_i$ be an indicator of a jump at time $i$. That is,

$$J_i = 1_{X_i \neq X_{i-1}}. \tag{7.73}$$

Consider $x \in \mathcal{X}$ and $z \in \mathcal{X}_b$. Then, by definition,

$$\frac{K^g(x,z)}{\pi(z)} = \frac{P([X_g]_b = z|X_0 = x)}{P([X_0]_b = z)}. \tag{7.74}$$

But,

$$P([X_g]_b = z|X_0 = x) = \sum_{d^g \in \{0,1\}^g} P([X_g]_b = z, J^g = d^g|X_0 = x)$$

$$\stackrel{(a)}{=} \sum_{d^g \in \{0,1\}^g} P([X_g]_b = z|J^g = d^g, X_0 = x) \, P(J^g = d^g), \tag{7.75}$$

where $(a)$ follows from the independence of the jump events and the value of the Markov process at each time. Now if there is a jump between time 1 and time $g$, then by definition of the transition probabilities the value of $[X_g]_b$ become independent of $[X_1]_b$ and also the jumps pattern. In other words, for any $J^g \neq (0, \ldots, 0)$,

$$P([X_g]_b = z|J^g = d^g, X_0 = x) = P([X_g]_b = z) = P([X_0]_b = z), \tag{7.76}$$

where the last equality follows from the stationarity of the Markov process. But $J^g \neq (0, \ldots, 0)$ means that there has been no jump from time 0 up to time $g$, and therefore $X_g = X_0$. This implies that

$$P([X_g]_b = z|J^g = 0^g, X_0 = x) = 1_{z=[x]_b}. \tag{7.77}$$

Since $J^g = (0, \ldots, 0) = (1-p)^g$, combining the intermediate steps, it follows that

$$P([X_g]_b = z|X_0 = x) = (1 - (1-p)^g) \, P([X_0]_b = z) + (1-p)^g 1_{z=[x]_b}, \tag{7.78}$$

and as a result

$$\frac{K^g(x,z)}{\pi(z)} = (1 - (1-p)^g) + \frac{(1-p)^g 1_{z=[x]_b}}{P([X_0]_b = z)}. \tag{7.79}$$

But given that by our assumption $f(x) \geq f_{\min}$, $P([X_0]_b = z) \geq f_{\min} 2^{-b}$. Therefore,

$$\Psi_1(b,g) = \sup_{(x,z) \in \mathcal{X} \times \mathcal{X}_b} \frac{K^g(x,z)}{\pi(z)}$$

$$\leq (1 - (1-p)^g) + \frac{(1-p)^g 2^b}{f_{\min}}. \tag{7.80}$$

For $b = b_n = \lceil r \log \log n \rceil$ and $g = g_n = \lfloor \gamma r \log \log n \rfloor$, we have

$$\log((1-p)^g 2^b) = g \log(1-p) + b \tag{7.81}$$

$$\leq r(\gamma \log(1-p) + 1) \log \log n. \tag{7.82}$$

But since $\gamma > -\frac{1}{\log(1-p)}$, $\gamma \log(1-p) + 1 < 0$, which from (7.80) proves the desired result, i.e. $\lim_{n \to \infty} \Psi_1(b_n, g_n) = 1$.

It is easy to check that due to its special distribution, the quantized version of process $\mathbf{X}$ is also a first-order Markov process. Therefore, from (4.15), we have

$$\Psi_2(b) = \sup_{(x,w_2) \in \mathcal{X} \times \mathcal{Z}^2 : [x]_b = w_1} \frac{\pi(w_2|x)}{\pi(w_2|w_1)}$$

$$= \sup_{(x,w_2) \in \mathcal{X} \times \mathcal{Z}} \frac{P([X_2]_b = w_2|X_1 = x)}{P([X_2]_b = w_2|[X_1]_b = [x]_b)}. \tag{7.83}$$

But

$$P([X_2]_b = w_2|X_1 = x) = P([X_2]_b = w_2, J_2 = 1|X_1 = x)p$$
$$+ P([X_2]_b = w_2, J_2 = 0|X_1 = x)(1-p)$$
$$= p\, P([X_2]_b = w_2) + (1-p)1_{w_2 = [x]_b}, \tag{7.84}$$

and similarly

$$P([X_2]_b = w_2|[X_1]_b = [x]_b) = P([X_2]_b = w_2, J_2 = 1|[X_1]_b = [x]_b)p$$
$$+ P([X_2]_b = w_2, J_2 = 0|[X_1]_b = [x]_b)(1-p)$$
$$= p\, P([X_2]_b = w_2) + (1-p)1_{w_2 = [x]_b}, \tag{7.85}$$

which proves that $\Psi_2(b) = 1$, for all $b$.

### 7.7 *Proof of Theorem* 5.1

By definition

$$\bar{d}_k(\mathbf{X}) = \limsup_{b \to \infty} \frac{H([X_{k+1}]_b|[X^k]_b)}{b}, \tag{7.86}$$

and therefore, for any $\delta_1 > 0$, there exists $b_{\delta_1}$ such that for all $b \geq b_{\delta_1}$, $\frac{H([X_{k+1}]_b|[X^k]_b)}{b} \leq \bar{d}_k(\mathbf{X}) + \delta_1$. Since $b = b_n = \lceil r \log \log n \rceil$ converges to infinity as $n \to \infty$, for all $n$ large enough, $b = b_n > b_{\delta_1}$, and as a result

$$\frac{H([X_{k+1}]_b|[X^k]_b)}{b} \leq \bar{d}_k(\mathbf{X}) + \delta_1. \tag{7.87}$$

For the rest of the proof, assume that $n$ is large enough such that $b_n > b_{\delta_1}$.

Define distribution $q_{k+1}$ over $\mathcal{X}_b^{k+1}$ as the $(k + 1)$th order distribution of the quantized process $[X_1]_b, \ldots, [X_n]_b$. That is, for $a^{k+1} \in \mathcal{X}_b^{k+1}$,

$$q_{k+1}(a_{k+1}|a^k) = \mathrm{P}([X_{k+1}]_b = a_{k+1}|[X^k]_b = a^k), \tag{7.88}$$

and

$$q_k(a^k) = \sum_{a_{k+1} \in \mathcal{X}_b} q_{k+1}(a^{k+1}) = \mathrm{P}([X^k]_b = a^k). \tag{7.89}$$

Also define distributions $\hat{q}_{k+1}^{(1)}$ and $\hat{q}_{k+1}^{(2)}$ as the empirical distributions induced by $\hat{X}^n$ and $[X^n]_b$, respectively. In other words, $\hat{q}_k^{(1)}(a^k) = \hat{p}^{(k)}(a^k|\hat{X}^n)$, $\hat{q}_k^{(2)}(a^k) = \hat{p}^{(k)}(a^k|[X^n]_b)$ and

$$\hat{q}_{k+1}^{(1)}(a_{k+1}|a^k) = \frac{\hat{q}_{k+1}^{(1)}(a^{k+1})}{\hat{q}_k^{(1)}(a^k)} = \frac{\hat{p}^{(k+1)}(a^{k+1}|\hat{X}^n)}{\hat{p}^{(k)}(a^k|\hat{X}^n)} \tag{7.90}$$

and

$$\hat{q}_{k+1}^{(2)}(a_{k+1}|a^k) = \frac{\hat{q}_{k+1}^{(2)}(a^{k+1})}{\hat{q}_k^{(2)}(a^k)} = \frac{\hat{p}^{(k+1)}(a^{k+1}|[X^n]_b)}{\hat{p}^{(k)}(a^k|[X^n]_b)}. \tag{7.91}$$

As the first step, we would like to prove that $\frac{1}{b}\hat{H}_k(\hat{X}^n) \leq \bar{d}_k(\mathbf{X}) + \delta$. Using the definitions above, we have

$$\sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} \hat{p}^{(k+1)}(a^{k+1}|\hat{X}^n) = \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} \hat{p}^{(k+1)}(a^{k+1}|\hat{X}^n) \log \frac{1}{q_{k+1}(a_{k+1}|a^k)}$$

$$= \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} \hat{p}^{(k+1)}(a^{k+1}|\hat{X}^n) \log \frac{\hat{q}_{k+1}^{(1)}(a_{k+1}|a^k)}{q_{k+1}(a_{k+1}|a^k)}$$

$$+ \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} \hat{p}^{(k+1)}(a^{k+1}|\hat{X}^n) \log \frac{1}{\hat{q}_{k+1}^{(1)}(a_{k+1}|a^k)}$$

$$= \sum_{a^k} \hat{q}_k^{(1)}(a^k) D_{\mathrm{KL}}(\hat{q}_{k+1}^{(1)}(\cdot|a^k)\|q_{k+1}(\cdot|a^k)) + \hat{H}_k(\hat{X}^n). \tag{7.92}$$

Since $\hat{X}^n$ is the minimizer of (2.5), we have

$$\sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} \hat{p}^{(k+1)}(a^{k+1}|\hat{X}^n) \leq (\bar{d}_k(\mathbf{X}) + \delta)b. \tag{7.93}$$

Combining this equation with (7.92) and the fact that $D_{\mathrm{KL}}$ is always positive, we obtain

$$\frac{1}{b}\hat{H}_k(\hat{X}^n) \leq \bar{d}_k(\mathbf{X}) + \delta. \tag{7.94}$$

As the second step of the proof, we show that with high probability

$$\frac{1}{b} \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} \hat{p}^{(k+1)}(a^{k+1}|[X^n]_b) \le \bar{d}_k(\mathbf{X}) + \delta. \tag{7.95}$$

In other words, we would like to show that the vector $[X^n]_b = ([X_1]_b, [X_2]_b, \dots, [X_n]_b)$ satisfies the constraint of the optimization (2.5). Following the same steps as those used in deriving (7.92), we get

$$\sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} \hat{p}^{(k+1)}(a^{k+1}|[X^n]_b)$$

$$= \sum_{a^k} \hat{q}_k^{(2)}(a^k) D_{\mathrm{KL}}(\hat{q}_{k+1}^{(2)}(\cdot|a^k) \| q_{k+1}(\cdot|a^k)) + \hat{H}_k([X^n]_b). \tag{7.96}$$

Also, note that

$$\sum_{a^k} \hat{q}_k^{(2)}(a^k) D_{\mathrm{KL}}(\hat{q}_{k+1}^{(2)}(\cdot|a^k) \| q_{k+1}(\cdot|a^k)) = \sum_{a^k} \hat{q}_k^{(2)}(a^k) \sum_{a_{k+1}} \hat{q}_{k+1}^{(2)}(a_{k+1}|a^k) \log \frac{\hat{q}_{k+1}^{(2)}(a_{k+1}|a^k)}{q_{k+1}(a_{k+1}|a^k)}$$

$$= \sum_{a^{k+1}} \hat{q}_{k+1}^{(2)}(a^{k+1}) \left( \log \frac{\hat{q}_{k+1}^{(2)}(a^{k+1})}{q_{k+1}(a^{k+1})} - \log \frac{\hat{q}_k^{(2)}(a^k)}{q_k(a^k)} \right)$$

$$= D_{\mathrm{KL}}(\hat{q}_{k+1}^{(2)} \| q_{k+1}) - D_{\mathrm{KL}}(\hat{q}_k^{(2)} \| q_k). \tag{7.97}$$

Therefore, since $0 \le D_{\mathrm{KL}}(\hat{q}_{k+1}^{(2)} \| q_{k+1}) - D_{\mathrm{KL}}(\hat{q}_k^{(2)} \| q_k) \le D_{\mathrm{KL}}(\hat{q}_{k+1}^{(2)} \| q_{k+1})$, from (7.96),

$$\sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} \hat{p}^{(k+1)}(a^{k+1}|[X^n]_b) \le \hat{H}_k([X^n]_b) + D_{\mathrm{KL}}(\hat{q}_{k+1}^{(2)} \| q_{k+1}). \tag{7.98}$$

Given $\delta_2 > 0$, define event $\mathcal{E}_1$ as

$$\mathcal{E}_1 \triangleq \{\|\hat{q}_{k+1}^{(2)} - q_{k+1}\|_1 < \delta_2\}. \tag{7.99}$$

Consider random vector $U^{k+1}$ distributed according to $\hat{q}_{k+1}^{(2)}$, which denotes the empirical distribution of $[X^n]_b$. Then, by definition, $\hat{H}_k([X^n]_b) = H(U_{k+1}|U^k) = H(U^{k+1}) - H(U^k)$. Therefore,

$$|\hat{H}_k([X^n]_b) - H([X_{k+1}]_b|[X^k]_b)| = |H(U^{k+1}) - H(U^k) - H([X^{k+1}]_b) + H([X^k]_b)|$$

$$\le |H(U^{k+1}) - H([X^{k+1}]_b)| + |H(U^k) - H([X^k]_b)|. \tag{7.100}$$

Conditioned on $\mathcal{E}_1$, $\|\hat{q}_k^{(2)} - q_k\|_1 \le \|\hat{q}_{k+1}^{(2)} - q_{k+1}\|_1 \le \delta_2$, and therefore, from Lemma 7.1,

$$|\hat{H}_k([X^n]_b) - H([X_{k+1}]_b|[X^k]_b)| \le -2\delta_2 \log \delta_2 + 2\delta_2(k+1) \log |\mathcal{X}_b| \tag{7.101}$$

or

$$\left| \frac{\hat{H}_k([X^n]_b)}{b} - \frac{H([X_{k+1}]_b|[X^k]_b)}{b} \right| \le -\frac{2\delta_2}{b} \log \delta_2 + \left( \frac{2(k+1)\log|\mathcal{X}_b|}{b} \right) \delta_2. \tag{7.102}$$

Moreover, conditioned on $\mathcal{E}_1$, since $\|\hat{q}_{k+1}^{(2)} - q_{k+1}\|_1 \le \delta_2$ and $\hat{q}_{k+1}^{(2)} \ll q_{k+1}$, from Lemma 7.2, we have

$$D(\hat{q}_{k+1}^{(2)} \| q_{k+1}) \le -\delta_2 \log \delta_2 + \delta_2(k+1) \log |\mathcal{X}_b| - \delta_2 \log q_{\min}, \tag{7.103}$$

where

$$q_{\min} = \min_{u^{k+1} \in \mathcal{X}_b^{k+1} : q_{k+1}(u^{k+1}) \ne 0} \mathrm{P}([X^{k+1}]_b = u^{k+1}) \ge f_{k+1} |\mathcal{X}_b|^{-(k+1)}, \tag{7.104}$$

where the last line follows from (5.1). Therefore,

$$\begin{aligned} D(\hat{q}_{k+1}^{(2)} \| q_{k+1}) &\le -\delta_2 \log \delta_2 + \delta_2(k+1) \log |\mathcal{X}_b| \\ &\quad - \delta_2 \log f_{k+1} + \delta_2(k+1) \log |\mathcal{X}_b| \end{aligned} \tag{7.105}$$

or

$$\frac{1}{b} D(\hat{q}_{k+1}^{(2)} \| q_{k+1}) \le -\frac{\delta_2}{b} (\log \delta_2 + \log f_{k+1}) + \left( \frac{2(k+1)\log|\mathcal{X}_b|}{b} \right) \delta_2. \tag{7.106}$$

Hence, combining (7.98), (7.102) and (7.106), it follows that, conditioned on $\mathcal{E}_1$,

$$\begin{aligned} &\frac{1}{b} \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} \hat{p}^{(k+1)}(a^{k+1}|[X^n]_b) \\ &\le \bar{d}_k(\mathbf{X}) + \delta_1 + \left( \frac{4(k+1)\log|\mathcal{X}_b|}{b} \right) \delta_2 - \frac{\delta_2}{b} (3 \log \delta_2 + \log f_{k+1}). \end{aligned} \tag{7.107}$$

Choosing $\delta_1 = \delta/2$ and $\delta_2$ small enough such that

$$\left( \frac{4(k+1)\log|\mathcal{X}_b|}{b} \right) \delta_2 - \frac{\delta_2}{b} (3 \log \delta_2 + \log f_{k+1}) \le \frac{\delta}{2}. \tag{7.108}$$

Note that while $|\mathcal{X}_b|$ grows exponentially in $b$, for all bounded sources, $\frac{1}{b} \log |\mathcal{X}_b| < 2$. Therefore, it is always possible to make sure that the above condition is satisfied for an appropriate choice of parameter $\delta_2$. For this choice of parameters, from (7.107), conditioned on $\mathcal{E}_1$

$$\frac{1}{b} \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} \hat{p}^{(k+1)}(a^{k+1}|[X^n]_b) \le \bar{d}_k(\mathbf{X}) + \delta, \tag{7.109}$$

and hence $[X^n]_b$ satisfies the constraint of the Q-MAP optimization described in (2.5). Hence, since $\hat{X}^n$ is the minimizer of $\|Au^n - Y^m\|^2$, among all sequences that satisfy this constraint, we conclude that, conditioned on $\mathcal{E}_1$,

$$
\begin{aligned}
\|A\hat{X}^n - Y^m\| &\leq \|A[X^n]_b - Y^m\| \\
&= \|A([X^n]_b - X^n)\| \\
&\leq \sigma_{\max}(A)\|X^n - [X^n]_b\| \\
&\leq \sigma_{\max}(A)2^{-b}\sqrt{n}.
\end{aligned}
\tag{7.110}
$$

Our goal is to use this equation to derive a bound for $\|\hat{X}^n - X^n\|$. The main challenge here is to find a lower bound for $\|A\hat{X}^n - Y^m\|$ in terms of $\|\hat{X}^n - X^n\|$. Given $\delta_3 > 0$ and $\tau > 0$, define set $\mathcal{C}_n$ and events $\mathcal{E}_2$ and $\mathcal{E}_3$ as

$$
\mathcal{C}_n \triangleq \left\{ u^n \in \mathcal{X}_b^n : \frac{1}{nb}\ell_{\mathrm{LZ}}(u^n) \leq \bar{d}_k(\mathbf{X}) + 2\delta \right\},
\tag{7.111}
$$

$$
\mathcal{E}_2 \triangleq \{\sigma_{\max}(A) < \sqrt{n} + 2\sqrt{m}\}
\tag{7.112}
$$

and

$$
\mathcal{E}_3 \triangleq \{\|A(u^n - X^n)\| \geq \|u^n - X^n\|\sqrt{(1-\tau)m} : \forall u^n \in \mathcal{C}_n\},
\tag{7.113}
$$

respectively. We will prove the following:

1. $\hat{X}^n \in \mathcal{C}_n$, for $n$ large enough.
2. $P(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3)$ converges to one as $n$ grows to infinity.

For the moment we assume that these two statements are true and complete the proof. Therefore, conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, it follows from (7.110) that

$$
\begin{aligned}
\|\hat{X}^n - X^n\|\sqrt{(1-\tau)m} &\leq n\left(1 + 2\sqrt{\frac{m}{n}}\right)2^{-b} \\
&\leq 3n2^{-b},
\end{aligned}
\tag{7.114}
$$

where the last line follows form the fact that $m \leq n$. Therefore, conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$,

$$
\frac{1}{\sqrt{n}}\|\hat{X}^n - X^n\| \leq \sqrt{\frac{9n}{(1-\tau)m2^{2b}}}.
\tag{7.115}
$$

To prove that $\hat{X}^n \in \mathcal{C}_n$, for $n$ large enough, note that, from (7.94), $\frac{1}{b}\hat{H}_k(\hat{X}^n) \leq \bar{d}_k(\mathbf{X}) + \delta$. On the other hand, from (7.4), for our choice of parameter $b = b_n$, for any given $\delta'' > 0$, for all $n$ large enough,

$$
\frac{1}{n}\ell_{\mathrm{LZ}}(\hat{X}^n) \leq \hat{H}_k(\hat{X}^n) + \delta''.
\tag{7.116}
$$

Therefore, for all $n$ large enough,

$$\frac{1}{nb}\ell_{\text{LZ}}(\hat{X}^n) \leq \bar{d}_k(\mathbf{X}) + \delta + \frac{\delta''}{b}. \tag{7.117}$$

Choosing $\delta''$ such that $\frac{\delta''}{b} \leq \delta$ proves the desired result, i.e. $\hat{X}^n \in \mathcal{C}_n$.

Let $\tau = 1 - (\log n)^{-2r/(1+f)}$, where $f > 0$ is a free parameter. For $b = b_n = \lceil r \log \log n \rceil$, $2^{2b} \geq (\log n)^{2r}$. Therefore, from (7.115),

$$\begin{aligned}
\frac{1}{\sqrt{n}}\|\hat{X}^n - X^n\| &\leq \sqrt{\frac{9(\log n)^{\frac{2r}{1+f}}}{(1+\delta)\bar{d}_k(\mathbf{X})(\log n)^{2r}}} \\
&= \sqrt{\frac{9}{(1+\delta)\bar{d}_k(\mathbf{X})(\log n)^{\frac{2rf}{1+f}}}}.
\end{aligned} \tag{7.118}$$

Therefore, for any $\epsilon > 0$, $n$ large enough, conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$,

$$\frac{1}{\sqrt{n}}\|\hat{X}^n - X^n\| \leq \epsilon. \tag{7.119}$$

To finish the proof we study the probability of $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$. By Theorem 4.2, there exists integer $g_{\delta_2}$, only depending on the source distribution and $\delta_2$ such that for $n > 6(k + g_{\delta_2})/\delta_2 + k$,

$$P(\mathcal{E}_1^c) \leq 2^{c\delta_2^2/8}(k + g_{\delta_2})n^{|\mathcal{X}_b|^k}2^{-\frac{n\delta_2^2}{8(k+g_{\delta_2})}}, \tag{7.120}$$

where $c = 1/(2\ln 2)$. Also, as proved in [5],

$$P(\mathcal{E}_2^c) \leq 2^{-m/2}. \tag{7.121}$$

Finally, from (7.3), the size of $\mathcal{C}_n$ can be upper bounded as

$$|\mathcal{C}_n| \leq 2^{nb(\bar{d}_k(\mathbf{X})+2\delta)+1}. \tag{7.122}$$

Now Lemma 7.4 combined with the union bound proves that, for a fixed vector $X^n$,

$$P_A(\mathcal{E}_3^c) \leq 2^{nb(\bar{d}_k(\mathbf{X})+2\delta)+1}e^{\frac{m}{2}(\tau+\ln(1-\tau))}, \tag{7.123}$$

where $P_A$ reflects the fact that $[X^n]_b$ is fixed, and the randomness is in the generation of matrix $A$. For our choice of parameter $\tau$ combined with the Fubini's Theorem and the Borel Cantelli Lemma proves that $P_{X^n}(\mathcal{E}_3^c) \to 0$, almost surely.

### 7.8  *Proof of Theorem* 5.2

The proof is very similar to the proof of Theorem 5.1 and follows the same logic. Similar to the proof of Theorem 5.1, for $\delta_1 > 0$, we assume that $n$ is large enough such that

$$\frac{H([X_{k+1}]_b|[X^k]_b)}{b} \leq \bar{d}_k(\mathbf{X}) + \delta_1. \tag{7.124}$$

Also, given $\delta_2 > 0$, $\delta_3 > 0$ and $\tau > 0$, consider set $\mathcal{C}_n$ and events $\mathcal{E}_2$ and $\mathcal{E}_3$, defined in (7.111), (7.112) and (7.113), respectively. Since $\hat{X}^n$ is a minimizer of $\sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} \hat{p}^{(k)}(a^{k+1}|u^n) + \frac{\lambda}{n^2}\|Au^n - Y^m\|^2$, we have

$$\sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} \hat{p}^{(k+1)}(a^{k+1}|\hat{X}^n) + \frac{\lambda}{n^2}\|A\hat{X}^n - Y^m\|^2$$

$$\leq \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} \hat{p}^{(k+1)}(a^{k+1}|[X^n]_b) + \frac{\lambda}{n^2}\|A[X^n]_b - Y^m\|^2$$

$$\leq \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} \hat{p}^{(k+1)}(a^{k+1}|[X^n]_b) + \frac{\lambda(\sigma_{\max}(A))^2}{n^2}\|[X^n]_b - X^n\|^2$$

$$\leq \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} \hat{p}^{(k+1)}(a^{k+1}|[X^n]_b) + \frac{\lambda(\sigma_{\max}(A))^2 2^{-2b}}{n}. \tag{7.125}$$

Define distribution $q_{k+1}$, $\hat{q}_{k+1}^{(1)}$ and $\hat{q}_{k+1}^{(2)}$ over $\mathcal{X}_b^{k+1}$ as in the proof of Theorem 5.1. Then, given $\delta > 0$, from (7.102) and (7.107), we set $\delta_1 = \delta/2$ and $\delta_2$ small enough such that (7.108) is satisfied. Then following the same steps as the ones that led to (7.109) we obtain that conditioned on $\mathcal{E}_1$ (defined in (7.99)),

$$\frac{1}{b}\hat{H}_k([X^n]_b) \leq \bar{d}_k(\mathbf{X}) + \delta. \tag{7.126}$$

Hence, we have

$$\hat{H}_k(\hat{X}^n) + \frac{\lambda}{n^2}\|A\hat{X}^n - Y^m\|^2 \leq b(\bar{d}_k(\mathbf{X}) + \delta) + \frac{\lambda(\sigma_{\max}(A))^2 2^{-2b}}{n}. \tag{7.127}$$

Since both terms on the left-hand side of (7.127) are positive, each of them should be smaller than the bound on the right-hand side, i.e.

$$\frac{1}{b}\hat{H}_k(\hat{X}^n) \leq \bar{d}_k(\mathbf{X}) + \delta + \frac{\lambda(\sigma_{\max}(A))^2 2^{-2b}}{bn} \tag{7.128}$$

and

$$\frac{\lambda}{bn^2}\|A\hat{X}^n - Y^m\|^2 \leq \bar{d}_k(\mathbf{X}) + \delta + \frac{\lambda(\sigma_{\max}(A))^2 2^{-2b}}{bn}. \tag{7.129}$$

Since $\lambda = \lambda_n = (\log n)^{2r}$ and $b = b_b = \lceil r \log \log n \rceil$, $\lambda 2^{-2b} \leq 1$, and hence, conditioned on $\mathcal{E}_2$,

$$\frac{\lambda(\sigma_{\max}(A))^2 2^{-2b}}{bn} \leq \frac{(\sqrt{n} + 2\sqrt{m})^2}{nb} \leq \frac{9}{b}. \tag{7.130}$$

Therefore, since $b_n \to \infty$, as $n \to \infty$, for all $n$ large enough, conditioned on $\mathcal{E}_2$,

$$\frac{\lambda(\sigma_{\max}(A))^2 2^{-2b}}{bn} \leq \delta, \tag{7.131}$$

and therefore, from (7.128) and (7.129),

$$\frac{1}{b}\hat{H}_k(\hat{X}^n) \leq \bar{d}_k(\mathbf{X}) + 2\delta \tag{7.132}$$

and

$$\frac{\lambda}{bn^2}\|A\hat{X}^n - Y^m\|^2 \leq \bar{d}_k(\mathbf{X}) + 2\delta. \tag{7.133}$$

Therefore, choosing $\delta_3 = 3\delta$, conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, $\hat{X}^n \in \mathcal{C}_n$. Finally, from (7.129) we have

$$\frac{\lambda(1-\tau)m}{n^2 b}\|\hat{X}^n - X^n\|^2 \leq \bar{d}_k(\mathbf{X}) + 2\delta, \tag{7.134}$$

or

$$\frac{1}{\sqrt{n}}\|\hat{X}^n - X^n\| \leq \sqrt{\frac{(\bar{d}_k(\mathbf{X}) + 2\delta)bn}{\lambda(1-\tau)m}}, \tag{7.135}$$

which proves that, for our set of parameters, conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, $\frac{1}{\sqrt{n}}\|X^n - \hat{X}^n\|$ can be made arbitrary small. Setting of parameter $\tau$ and proving that $P(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3)$ converges to one can be done exactly as it was done in the proof of Theorem 5.1.

### 7.9 *Proof of Theorem* 6.1

As argued in the proof of Theorem 5.1, given $\delta > 0$, there exists $b_\delta$, such that for $b > b_\delta$,

$$\frac{H([X_{k+1}]_b|[X^k]_b)}{b} \leq \bar{d}_k(\mathbf{X}) + \frac{\delta}{2}. \tag{7.136}$$

In the rest of the proof we assume that $n$ is large enough so that $b = b_n > b_\delta$. Define distributions $q_{k+1}$ and $\hat{q}_{k+1}$ over $\mathcal{X}_b^{k+1}$ as follows. Let $q_{k+1}$ and $\hat{q}_{k+1}$ denote the distribution of $[X^{k+1}]_b$ and the empirical distribution of $[X^n]_b$, respectively. From (7.96) and (7.97), it follows that

$$\sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}}\hat{p}^{(k+1)}(a^{k+1}|[X^n]_b) \leq \hat{H}_k([X^n]_b) + D_{\mathrm{KL}}(\hat{q}_{k+1}\|q_{k+1}). \tag{7.137}$$

Define events $\mathcal{E}_1$ and $\mathcal{E}_2$ as

$$\mathcal{E}_1 = \{\sigma_{\max}(A) < \sqrt{n} + 2\sqrt{m}\} \tag{7.138}$$

and

$$\mathcal{E}_2 = \{\|\hat{q}_{k+1} - q_{k+1}\|_1 < \delta'\}, \tag{7.139}$$

where $\delta' > 0$ is selected such that

$$\frac{1}{b}(-2\delta' \log \delta' + 2\delta'(k+1) \log |\mathcal{X}_b|) \leq \frac{\delta}{4} \tag{7.140}$$

and

$$\frac{1}{b}(-\delta' \log \delta' - \delta' \log f_{k+1} + 2\delta'(k+1) \log |\mathcal{X}_b|) \leq \frac{\delta}{4}, \tag{7.141}$$

for all $b$ large enough. This is always possible, since $(-2\delta' \log \delta')/b$ is a decreasing function of $b$ and $\log |\mathcal{X}_b|/b$ can be upper bounded by a constant not depending on $b$. Hence, by picking $\delta'$ small enough both (7.140) and (7.141) hold.

Given this choice of parameters, from (7.102), conditioned on $\mathcal{E}_2$, we have

$$\frac{1}{b} \left| \hat{H}_k([X^n]_b) - H([X_{k+1}]_b | [X^k]_b) \right| \leq \frac{\delta}{4}. \tag{7.142}$$

Also, from Lemma 7.2, and (7.106), conditioned on $\mathcal{E}_2$,

$$\frac{1}{b}D(\hat{q}_{k+1} \| q_{k+1}) \leq -\frac{\delta'}{b}(\log \delta' + \log f_{k+1}) + \left( \frac{2(k+1) \log |\mathcal{X}_b|}{b} \right) \delta'. \tag{7.143}$$

Therefore, from (7.141),

$$\frac{1}{b}D(\hat{q}_{k+1} \| q_{k+1}) \leq \frac{\delta}{4}. \tag{7.144}$$

Conditioned on $\mathcal{E}_2$, from (7.136) and (7.137), it follows that

$$\frac{1}{b} \sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} \hat{p}^{(k+1)}(a^{k+1} | [X^n]_b) \leq \frac{\hat{H}_k([X^n]_b)}{b} + \frac{1}{b}D_{\mathrm{KL}}(\hat{q}_{k+1} \| q_{k+1})$$

$$\leq \frac{H([X_{k+1}]_b | [X^k]_b)}{b} + \frac{\delta}{4} + \frac{\delta}{4}$$

$$\leq \bar{d}_k(\mathbf{X}) + \frac{\delta}{2} + \frac{\delta}{4} + \frac{\delta}{4} = \bar{d}_k(\mathbf{X}) + \delta, \tag{7.145}$$

which implies that $[X^n]_b \in \mathcal{F}_o$. Since $\hat{X}^n(t+1)$ is the solution of (6.2), automatically, $\hat{X}^n(t+1) \in \mathcal{F}_o$. Also, as we just proved, conditioned on $\mathcal{E}_2 \cap \mathcal{E}_3$, $[X^n]_b \in \mathcal{F}_o$ as well. Therefore, conditioned on $\mathcal{E}_2 \cap \mathcal{E}_3$,

$$\|\hat{X}^n(t+1) - S^n(t+1)\| \leq \|[X^n]_b - S^n(t+1)\|, \tag{7.146}$$

or equivalently

$$\|\hat{X}^n(t+1) - [X^n]_b + [X^n]_b - S^n(t+1)\| \leq \|[X^n]_b - S^n(t+1)\|. \tag{7.147}$$

Raising both sides of (7.147) to power two and canceling out the common term from both sides, we derive

$$\|\hat{X}^n(t+1) - [X^n]_b\|^2 + 2\langle \hat{X}^n(t+1) - [X^n]_b, [X^n]_b - S^n(t+1)\rangle \leq 0. \tag{7.148}$$

If we plug in the expression for $S^n(t+1)$ we obtain

$$\begin{aligned}
\|\hat{X}^n(t+1) - [X^n]_b\|^2 &\leq 2\langle \hat{X}^n(t+1) - [X^n]_b, -[X^n]_b + S^n(t+1)\rangle \\
&= 2\langle \hat{X}^n(t+1) - [X^n]_b, -[X^n]_b + \hat{X}^n(t) + \mu A^\top(Y^m - A\hat{X}^n(t))\rangle \\
&= 2\langle \hat{X}^n(t+1) - [X^n]_b, \hat{X}^n(t) - [X^n]_b\rangle \\
&\quad - 2\mu\langle \hat{X}^n(t+1) - [X^n]_b, A^\top A(\hat{X}^n(t) - X^n)\rangle \\
&\quad + 2\mu\langle \hat{X}^n(t+1) - [X^n]_b, A^\top Z^m\rangle \\
&= 2\langle \hat{X}^n(t+1) - [X^n]_b, \hat{X}^n(t) - [X^n]_b\rangle \\
&\quad - 2\mu\langle \hat{X}^n(t+1) - [X^n]_b, A^\top A(\hat{X}^n(t) - [X]_b^n)\rangle \\
&\quad + 2\mu\langle \hat{X}^n(t+1) - [X^n]_b, A^\top A(X^n - [X]_b^n)\rangle \\
&\quad + 2\mu\langle \hat{X}^n(t+1) - [X^n]_b, A^\top Z^m\rangle. \tag{7.149}
\end{aligned}$$

Define

$$E^n(t) \triangleq \|\hat{X}^n(t+1) - [X^n]_b\| \tag{7.150}$$

and

$$\tilde{E}^n(t) \triangleq \frac{E^n(t)}{\|E^n(t)\|}. \tag{7.151}$$

Then, it follows from (7.149) that

$$\begin{aligned}
\|E^n(t+1)\| &\leq 2\langle \tilde{E}^n(t+1), \tilde{E}^n(t)\rangle\|E^n(t)\| - 2\mu\langle \tilde{E}^n(t+1), A^\top A\tilde{E}^n(t)\rangle\|E^n(t)\| \\
&\quad + 2\mu\langle \tilde{E}^n(t+1), A^\top A(X^n - [X]_b^n)\rangle \\
&\quad + 2\mu\langle \tilde{E}^n(t+1), A^\top Z^m\rangle
\end{aligned}$$

$$
\begin{aligned}
= \ &2\Big(\langle \tilde{E}^n(t+1), \tilde{E}^n(t)\rangle - \mu\langle A\tilde{E}^n(t+1), A\tilde{E}^n(t)\rangle\Big)\|E^n(t)\| \\
&+ 2\mu\langle \tilde{E}^n(t+1), A^\top A(X^n - [X]_b^n)\rangle \\
&+ 2\mu\langle \tilde{E}^n(t+1), A^\top Z^m\rangle \\
\leq \ &2\Big(\langle \tilde{E}^n(t+1), \tilde{E}^n(t)\rangle - \mu\langle A\tilde{E}^n(t+1), A\tilde{E}^n(t)\rangle\Big)\|E^n(t)\| \\
&+ 2\mu\sigma_{\max}(A^\top A)\|X^n - [X]_b^n\| \\
&+ 2\mu\langle \tilde{E}^n(t+1), A^\top Z^m\rangle.
\end{aligned} \tag{7.152}
$$

Note that for a fixed $X^n$, $\tilde{E}^n(t)$ and $\tilde{E}^n(t+1)$ can only take a finite number of different values. Let $\mathcal{S}_e$ denote the set of all possible normalized error vectors. That is,

$$
\mathcal{S}_e \triangleq \left\{ \frac{u^n - v^n}{\|u^n - v^n\|} : u^n, v^n \in \mathcal{F}_o \right\}. \tag{7.153}
$$

Clearly, $\tilde{E}^n(t)$ and $\tilde{E}^n(t+1)$ are both members of $\mathcal{S}_e$. Define event $\mathcal{E}_3$ as follows

$$
\mathcal{E}_3 \triangleq \left\{ \langle u^n, v^n\rangle - \frac{1}{m}\langle Au^n, Av^n\rangle \leq 0.45 : \ \forall\, (u^n, v^n) \in \mathcal{S}_e^2 \right\}. \tag{7.154}
$$

Conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, it follows from (7.152) that

$$
\|E^n(t+1)\| \leq 0.9\|E^n(t)\| + \frac{2(\sqrt{n} + 2\sqrt{m})^2}{m}(2^{-b}\sqrt{n})
$$
$$
+ 2\mu\langle \tilde{E}^n(t+1), A^\top Z^m\rangle. \tag{7.155}
$$

The only remaining term on the right-hand side of (7.155) is $2\mu\langle \tilde{E}^n(t+1), A^\top Z^m\rangle$. To upper bound this term, we employ Lemma 7.5. Let $A_i^m$, $i = 1, \ldots, n$, denote the $i$th column of matrix $A$. Then,

$$
A^\top Z^m = \begin{bmatrix} \langle A_1^m, Z^m\rangle \\ \langle A_2^m, Z^m\rangle \\ \vdots \\ \langle A_n^m, Z^m\rangle \end{bmatrix}, \tag{7.156}
$$

and

$$
\langle \tilde{E}^n(t+1), A^\top Z^m\rangle = \sum_{i=1}^n \tilde{E}_i(t+1)\langle A_i^m, Z^m\rangle. \tag{7.157}
$$

By Lemma 7.5, for each $i$, $\langle A_i^m, Z^m\rangle$ is distributed as $\|Z^m\|G_i$, where $G_i$ is a standard normal distribution independent of $\|Z^m\|$. Therefore, since the columns of matrix $A$ are also independent, overall $\langle \tilde{E}^n(t+1), A^\top Z^m\rangle$ is distributed as

$$
\|Z^m\| \sum_{i=1}^n \tilde{E}_i(t+1)G_i, \tag{7.158}
$$

where $G_i$ are i.i.d. standard normal independent of $\|Z^m\|$. Given $\tau_1 > 0$ and $\tau_2 > 0$, define events $\mathcal{E}_4$ and $\mathcal{E}_5$ as follows:

$$\mathcal{E}_4 \triangleq \left\{ \frac{1}{m}\|Z^m\|^2 \le (1+\tau_1)\sigma^2 \right\} \tag{7.159}$$

and

$$\mathcal{E}_5 \triangleq \{ |\langle \tilde{e}^n, G^n \rangle|^2 \le 1+\tau_2 : \ \forall \tilde{e}^n \in \mathcal{S}_e \}. \tag{7.160}$$

By Lemma 7.4,

$$\begin{aligned} \mathrm{P}(\mathcal{E}_4^c) &= \mathrm{P}\left( \frac{1}{\sigma^2}\|Z^m\|^2 > (1+\tau_1)m \right) \\ &\le \mathrm{e}^{-\frac{m}{2}(\tau_1 - \ln(1+\tau_1))}. \end{aligned} \tag{7.161}$$

For a fixed vector $\tilde{e}^n$, $\langle \tilde{e}^n, G^n \rangle = \sum_{i=1}^{n} \tilde{e}_i G_i$ has a standard normal distribution and therefore, by letting $m = 1$ in Lemma 7.4, it follows that

$$\mathrm{P}(|\langle \tilde{e}^n, G^n \rangle|^2 > 1+\tau_2) \le \mathrm{e}^{-0.5(\tau_2 - \ln(1+\tau_2))}. \tag{7.162}$$

Hence, by the union bound,

$$\begin{aligned} \mathrm{P}(\mathcal{E}_4^c) &\le |\mathcal{S}_e| \, \mathrm{e}^{-0.5(\tau_2 - \ln(1+\tau_2))} \\ &\le 2^{nb(\bar{d}_k(\mathbf{X})+2\delta)} \mathrm{e}^{-0.5(\tau_2 - \ln(1+\tau_2))}, \end{aligned} \tag{7.163}$$

where the last inequality follows from (7.173) and (7.179). But, for $\tau_2 > 7$,

$$\mathrm{e}^{-0.5(\tau_2 - \ln(1+\tau_2))} \le 2^{-0.5\tau_2}, \tag{7.164}$$

which implies that for $\tau_2 > 7$,

$$\mathrm{P}(\mathcal{E}_5^c) \le 2^{nb(\bar{d}_k(\mathbf{X})+2\delta)-0.5\tau_2}. \tag{7.165}$$

Setting

$$\tau_2 = 2nb(\bar{d}_k(\mathbf{X})+3\delta) - 1 \tag{7.166}$$

ensures that

$$\mathrm{P}(\mathcal{E}_5^c) \le 2^{-\delta b n + 0.5}, \tag{7.167}$$

which converges to zero as $n$ grows to infinity. Finally, setting $\tau_1 = 1$, conditioned on $\mathcal{E}_4 \cap \mathcal{E}_5$, we have

$$
\begin{aligned}
2\mu\langle \tilde{E}^n(t+1), A^\top Z^m \rangle &= \frac{2}{m}\langle \tilde{E}^n(t+1), A^\top Z^m \rangle \\
&\leq \frac{1}{m}\sqrt{m(1+\tau_1)\sigma^2(1+\tau_2)} \\
&= \sqrt{\frac{(2m)\sigma^2(2nb(\bar{d}_k(\mathbf{X})+3\delta))}{m^2}} \\
&= \frac{\sigma}{2}\sqrt{\frac{nb(\bar{d}_k(\mathbf{X})+3\delta)}{m}}.
\end{aligned}
\tag{7.168}
$$

Combining (7.168) and (7.155) yields the desired upper bound on the last term in (7.155).

To finish the proof, we need to show that $P(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3)$ also approaches one, as $n$ grows without bound. Reference [5] proves that

$$
P(\mathcal{E}_1^c) \leq 2^{-m/2}.
\tag{7.169}
$$

By Theorem 4.2, there exists integer $g_{\delta'}$, depending only on $\delta'$ and the source distribution, such that for any $n > 6(k+g_{\delta'})/(b\delta) + k$,

$$
P(\mathcal{E}_2^c) \leq 2^{c\delta'^2/8}(k+g_{\delta'}+1)n^{|\mathcal{X}_b|^{k+1}}2^{-\frac{nc\delta'^2}{8(k+g_{\delta'}+1)}},
\tag{7.170}
$$

where $c = 1/(2\ln 2)$. This proves that for our choice of parameters, $P(\mathcal{E}_2^c)$ converges to zero.

In the rest of the proof we bound $P(\mathcal{E}_3^c)$. From Corollary 7.1, for any $u^n, v^n \in \mathcal{S}_e$,

$$
P\left(\frac{1}{m}\langle Au^n, Av^n\rangle - \langle u^n, v^n\rangle \leq -0.45\right) \leq 2^{-0.05m}.
\tag{7.171}
$$

Therefore, by the union bound,

$$
\begin{aligned}
P(\mathcal{E}_3^c) &= P\left(\frac{1}{m}\langle Au^n, A\tilde{e}^n(t)\rangle - \langle u^n, \tilde{e}^n(t)\rangle \leq -0.45 : \text{for some } (u^n, v^n) \in \mathcal{S}_e^2\right) \\
&\leq |\mathcal{S}_e|^2 2^{-0.05m}.
\end{aligned}
\tag{7.172}
$$

Note that

$$
|\mathcal{S}_e| \leq |\mathcal{F}_o|^2.
\tag{7.173}
$$

In the following, we derive an upper bound on the size of $\mathcal{F}_o$. For any $u^n \in \mathcal{F}_o$, by definition, we have

$$
\sum_{a^{k+1}\in\mathcal{X}_b^{k+1}} w_{a^{k+1}}\hat{p}^{(k+1)}(a^{k+1}|u^n) \leq b(\bar{d}_k(\mathbf{X})+\delta).
\tag{7.174}
$$

Let $\hat{q}_{k+1}^{(u)} = \hat{p}^{(k+1)}(\cdot|u^n)$ denote the $(k+1)$th order empirical distributions of $u^n$. Following the steps used in deriving (7.92), it follows that

$$\sum_{a^{k+1} \in \mathcal{X}_b^{k+1}} w_{a^{k+1}} \hat{p}^{(k+1)}(a^{k+1}|u^n)$$

$$= \sum_{a^k} \hat{q}_k^{(u)}(a^k) D_{\mathrm{KL}}(\hat{q}_{k+1}^{(u)}(\cdot|a^k)\|q_{k+1}(\cdot|a^k)) + \hat{H}_k(u^n). \tag{7.175}$$

Since all the terms on the right-hand side or (7.175) are positive, it follows from (7.174) that

$$\hat{H}_k(u^n) \leq b(\bar{d}_k(\mathbf{X}) + \delta). \tag{7.176}$$

On the other hand, given our choice of quantization level $b$, for $n$ large enough, for any $v^n \in \mathcal{X}_b^n$,

$$\frac{1}{nb}\ell_{\mathrm{LZ}}(v^n) \leq \frac{1}{b}\hat{H}_k(v^n) + \delta. \tag{7.177}$$

Therefore, for any $u^n \in \mathcal{F}_o$, from (7.176) and (7.177), it follows that

$$\frac{1}{nb}\ell_{\mathrm{LZ}}(u^n) \leq \frac{1}{b}\hat{H}_k(u^n) + \delta$$

$$\leq \bar{d}_k(\mathbf{X}) + 2\delta. \tag{7.178}$$

Note that, from (7.3), we have

$$|\mathcal{F}_o| \leq \left|\left\{v^n : \frac{1}{nb}\ell_{\mathrm{LZ}}(v^n) \leq \bar{d}_k(\mathbf{X}) + 2\delta\right\}\right| \leq 2^{nb(\bar{d}_k(\mathbf{X})+2\delta)}. \tag{7.179}$$

Hence, from (7.172),

$$\mathrm{P}(\mathcal{E}_3^c) \leq |\mathcal{F}_o|^4 2^{-0.05m} \leq 2^{nb(\bar{d}_k(\mathbf{X})+2\delta)-0.05m} \leq 2^{-2nb\delta}. \tag{7.180}$$

## 8. Conclusion

For a stationary process $\mathbf{X}$, we have studied the problem of estimating $X^n$ from $m$ response variables $Y^m = AX^n + Z^m$, under the assumption that the distribution of $\mathbf{X}$ is known. We have proposed the Q-MAP optimization, which estimates $X^n$ from $Y^m$. The new optimization satisfies the following properties: (i) It applies to generic classes of distributions, as long as they satisfy certain mixing conditions. (ii) Unlike other Bayesian approaches, in the high-dimensional settings, the performance of the Q-MAP optimization can be characterized for generic distributions. Our analyses show that, for certain distributions such as spike-and-slab, asymptotically, Q-MAP achieves the minimum required sampling rate. (Whether Q-MAP achieves the optimal sampling rate for general distributions is still an open question.) (iii) PGD can be applied to approximate the solution of Q-MAP. While the optimization involved in Q-MAP is non-convex, we have characterized the performance of the corresponding PGD algorithm, under both noiseless and noisy settings. Our analysis has revealed that with slightly more measurements than Q-MAP, the PGD-based method recovers $X^n$ accurately in the noiseless setting.

## Funding

## REFERENCES

1. BICKEL, P. J., RITOV, Y. & TSYBAKOV, A. B. (2009) Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **37**, 1705–1732.
2. BLUMENSATH, T. & DAVIES, M. E. (2009) Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.*, **27**, 265–274.
3. BRADLEY, R. C. (2005) Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.*, **2**, 107–144.
4. CANDES, E. & TAO, T. (2007) The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, **35**, 2313–2351.
5. CANDÈS, E., ROMBERG, J. & TAO, T. (2005) Decoding by linear programming. *IEEE Trans. Inform. Theory*, **51**, 4203–4215.
6. COVER, T. & THOMAS, J. (2006) *Elements of Information Theory*, 2nd edn. New York: Wiley.
7. CSISZAR, I. & KÖRNER, J. (2011) *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York, NY: Cambridge University Press.
8. DONOHO, D. & MONTANARI, A. (2016) High dimensional robust m-estimation: asymptotic variance via approximate message passing. *Probab. Theory Relat. Fields*, **166**, 935–969.
9. GUO, D. & VERDÚ, S. (2005) Randomly spread CDMA: Asymptotics via statistical physics. *IEEE Trans. Inform. Theory*, **51**, 1983–2010.
10. HANS, C. (2009) Bayesian Lasso regression. *Biometrika*, **96**, 835–845.
11. HANS, C. (2010) Model uncertainty and variable selection in Bayesian Lasso regression. *Statist. Comput.*, **20**, 221–229.
12. HOADLEY, B. (1970) A Bayesian look at inverse linear regression. *J. Amer. Statist. Assoc.*, **65**, 356–369.
13. JALALI, S., MALEKI, A. & BARANIUK, R. (2014) Minimum complexity pursuit for universal compressed sensing. *IEEE Trans. Inform. Theory*, **60**, 2253–2268.
14. JALALI, S., MONTANARI, A. & WEISSMAN, T. (2012) Lossy compression of discrete sources via the Viterbi algorithm. *IEEE Trans. Inform. Theory*, **58**, 2475–2489.
15. JALALI, S. & POOR, H. V. (2017) Universal compressed sensing for almost lossless recovery. *IEEE Trans. Inform. Theory*, **63**, 2933–2953.
16. LINDLEY, D. V. & SMITH, A. F. (1972) Bayes estimates for the linear model. *J. R. Statist. Soc. Ser. B*, **34**, 1–41.
17. LIU, C. (1996) Bayesian robust multivariate linear regression with incomplete data. *J. Amer. Statist. Assoc.*, **91**, 1219–1227.
18. MALEKI, A. (2010) Approximate message passing algorithm for compressed sensing. Stanford University. PhD Thesis.
19. MITCHELL, T. J. & BEAUCHAMP, J. J. (1988) Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.*, **83**, 1023–1032.
20. O'HARA, R. B. & SILLANPÄÄ, M. J. (2009) A review of Bayesian variable selection methods: what, how and which. *Bayes. Anal.*, **4**, 85–117.
21. PARK, T. & CASELLA, G. (2008) The Bayesian Lasso. *J. Amer. Statist. Assoc.*, **103**, 681–686.
22. PLOTNIK, E., WEINBERGER, M. J. & ZIV, J. (1992) Upper bounds on the probability of sequences emitted by finite-state sources and on the redundancy of the Lempel-Ziv algorithm. *IEEE Trans. Inform. Theory*, **38**, 66–72.
23. RÉNYI, A. (1959) On the dimension and entropy of probability distributions. *Acta Math. Acad. Sci. Hung.*, **10**, 193–215.
24. RIGOLLET, P. & TSYBAKOV, A. (2011) Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, **39**, 731–771.

25. SHIELDS, P. (1996a) *The Ergodic Theory of Discrete Sample Paths*. Providence, RI: American Mathematical Society.
26. SHIHAO, J., YA, X. & CARIN, L. (2008) Bayesian Compressive Sensing. *IEEE Trans. Signal. Process.*, **56**, 2346–2356.
27. SOM, S. & SCHNITER, P. (2012) Compressive imaging using approximate message passing and a Markov-tree prior. *IEEE Trans. Signal Process.*, **60**, 3439–3448.
28. SU, W. & CANDES, E. (2016) SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.*, **44**, 1038–1068.
29. TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. & KNIGHT, K. (2005) Sparsity and smoothness via the fused Lasso. *J. R. Statist. Soc.: Ser. B*, **67**, 91–108.
30. TIPPING, M. E. (2001) Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, **1**, 211–244.
31. VITERBI, A. J. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inform. Theory*, **13**, 260–269.
32. WEST, M. (1984) Outlier models and prior distributions in Bayesian linear regression. *J. R. Statist. Soc. Ser. B*, **46**, 431–439.
33. WU, Y. & VERDÚ, S. (2010) Rényi information dimension: fundamental limits of almost lossless analog compression. *IEEE Trans. Inform. Theory*, **56**, 3721–3748.
34. ZIV, J. & LEMPEL, A. (1978) Compression of individual sequences via variable-rate coding. *IEEE Trans. Inform. Theory*, **24**, 530–536.