Information-Theoretically Optimal Compressed Sensing via Spatial Coupling and Approximate Message Passing

David L. Donoho[†], Adel Javanmard[*] and Andrea Montanari [*†]

December 3, 2011

Abstract

We study the compressed sensing reconstruction problem for a broad class of random, band-diagonal sensing matrices. This construction is inspired by the idea of spatial coupling in coding theory. As demonstrated heuristically and numerically by Krzakala et al. [KMS+11], message passing algorithms can effectively solve the reconstruction problem for spatially coupled measurements with undersampling rates close to the fraction of non-zero coordinates.

We use an approximate message passing (AMP) algorithm and analyze it through the state evolution method. We give a rigorous proof that this approach is successful as soon as the undersampling rate δ exceeds the (upper) Rényi information dimension of the signal, $\overline{d}(p_X)$. More precisely, for a sequence of signals of diverging dimension n whose empirical distribution converges to p_X, reconstruction is with high probability successful from $\overline{d}(p_X)\, n + o(n)$ measurements taken according to a band diagonal matrix.

For sparse signals, i.e. sequences of dimension n and $k(n)$ non-zero entries, this implies reconstruction from $k(n)+o(n)$ measurements. For 'discrete' signals, i.e. signals whose coordinates take a fixed finite set of values, this implies reconstruction from $o(n)$ measurements. The result is robust with respect to noise, does not apply uniquely to random signals, but requires the knowledge of the empirical distribution of the signal p_X.

1 Introduction and main results

1.1 Background and contributions

Assume that m linear measurements are taken of an unknown n-dimensional signal $x \in \mathbb{R}^n$, according to the model

$$y = Ax\,. \tag{1}$$

The reconstruction problem requires to reconstruct x from the measured vector $y \in \mathbb{R}^m$, and the measurement matrix $A \in \mathbb{R}^{m\times n}$.

It is an elementary fact of linear algebra that the reconstruction problem will not have a unique solution unless $m \geq n$. This observation is however challenged within compressed sensing. A

[*]Department of Electrical Engineering, Stanford University

[†]Department of Statistics, Stanford University

large corpus of research shows that, under the assumption that x is sparse, a dramatically smaller number of measurements is sufficient [Don06a, CRT06a, Don06b]. Namely, if only k entries of x are non-vanishing, then roughly $m \gtrsim 2k \log(n/k)$ measurements are sufficient for A random, and reconstruction can be solved efficiently by convex programming. Deterministic sensing matrices achieve similar performances, provided they satisfy a suitable restricted isometry condition [CT05]. On top of this, reconstruction is robust with respect to the addition of noise [CRT06b, DMM11], i.e. under the model

$$y = Ax + w\,, \tag{2}$$

with –say– $w \in \mathbb{R}^m$ a random vector with i.i.d. components $w_i \sim \mathsf{N}(0, \sigma^2)$. In this context, the notions of 'robustness' or 'stability' refers to the existence of universal constants C such that the per-coordinate mean square error in reconstructing x from noisy observation y is upper bounded by $C\,\sigma^2$.

From an information-theoretic point of view it remains however unclear why we cannot achieve the same goal with far fewer than $2\,k\log(n/k)$ measurements. Indeed, we can interpret Eq. (1) as describing an analog data compression process, with y a compressed version of x. From this point of view, we can encode all the information about x in a single real number $y \in \mathbb{R}$ (i.e. use $m = 1$), because the cardinality of $\mathbb{R}$ is the same as the one of $\mathbb{R}^n$. Motivated by this puzzling remark, Wu and Verdù [WV10] introduced a Shannon-theoretic analogue of compressed sensing, whereby the vector x has i.i.d. components $x_i \sim p_X$. Crucially, the distribution p_X is available to, and may be used by the reconstruction algorithm. Under the mild assumptions that sensing is linear (as per Eq. (1)), and that the reconstruction mapping is Lipschitz continuous, they proved that compression is asymptotically lossless if and only if

$$m \geq n\,\overline{d}(p_X) + o(n)\,. \tag{3}$$

Here $\overline{d}(p_X)$ is the (upper) Rényi information dimension of the distribution p_X. We refer to Section 1.2 for a precise definition of this quantity. Suffices to say that, if p_X is ε-sparse (i.e. if it puts mass at most ε on nonzeros) then $\overline{d}(p_X) \leq \varepsilon$. Also, if p_X is the convex combination of a discrete part (sum of Dirac's delta) and an absolutely continuous part (with a density), then $\overline{d}(p_X)$ is equal to the weight of the absolutely continuous part.

This result is quite striking. For instance, it implies that, for random k-sparse vectors, $m \geq k + o(n)$ measurements are sufficient. Also, if the entries of x are random and take values in – say– $\{-10, -9, \dots, -9, +10\}$, then a sublinear number of measurements $m = o(n)$, is sufficient! At the same time, the result of Wu and Verdù presents two important limitations. First, it does not provide robustness guarantees[1] of the type described above. It therefore leaves open the possibility that reconstruction is highly sensitive to noise when m is significantly smaller than the number of measurements required in classical compressed sensing, namely $\Theta(k\log(n/k))$ for k-sparse vectors. Second, it does not provide any computationally practical algorithms for reconstructing x from measurements y.

[1]While this paper was about to be posted, we became aware of a paper by Wu and Verdù [WV11b] claiming that the boundary $\delta = \overline{D}(p_X)$(see below for a definition of $\overline{D}(p_X)$) is achievable in principle by the Bayes minimum mean square error rule. Their result seems to be conditional on the validity of the replica method in this setting, which is not yet proved.

In an independent line of work, Krzakala et al. [KMS+11] developed an approach that leverages on the idea of *spatial coupling*. This idea was introduced for the compressed sensing literature by Kudekar and Pfister [KP10] (see [KRU11] and Section 1.5 for a discussion of earlier work on this topic). Spatially coupled matrices are –roughly speaking– random sensing matrices with a band-diagonal structure. The analogy is, this time, with channel coding.[2] In this context, spatial coupling, in conjunction with message-passing decoding, allows to achieve Shannon capacity on memoryless communication channels. By analogy, it is reasonable to hope that a similar approach might enable to sense random vectors x at an undersampling rate m/n close to the Rényi information dimension of the coordinates of x, $\overline{d}(p_X)$. Indeed, the authors of [KMS+11] evaluate this approach numerically on a few classes of random vectors and demonstrate that it indeed achieves rates close to the fraction of non-zero entries. They also support this claim by insightful statistical physics arguments.

In this paper, we fill the gap between the above works, and present the following contributions:

Construction. We describe a construction for spatially coupled sensing matrices A that is somewhat broader than the one of [KMS+11] and give precise prescriptions for the asymptotics of various parameters. We also use a somewhat different reconstruction algorithm from the one in [KMS+11], by building on the approximate message passing (AMP) approach of [DMM09, DMM10]. AMP algorithms have the advantage of smaller memory complexity with respect to standard message passing, and of smaller computational complexity whenever fast multiplication procedures are available for A.

Rigorous proof of convergence. Our main contribution is a rigorous proof that the above approach indeed achieves the information-theoretic limits set out by Wu and Verdù [WV10]. Indeed, we prove that, for sequences of spatially coupled sensing matrices $\{A(n)\}_{n\in\mathbb{N}}$, $A(n) \in \mathbb{R}^{m(n)\times n}$ with asymptotic undersampling rate $\delta = \lim_{n\to\infty} m(n)/n$, AMP reconstruction is with high probability successful in recovering the signal x, provided $\delta > \overline{d}(p_X)$.

Robustness to noise. We prove that the present approach is robust[3] to noise in the following sense. For any signal distribution p_X and undersampling rate δ, there exists a constant C such that the output $\widehat{x}(y)$ of the reconstruction algorithm achieves a mean square error per coordinate $n^{-1}\mathbb{E}\{\|\widehat{x}(y) - x\|_2^2\} \le C\,\sigma^2$. This result holds under the noisy measurement model (2) for a broad class of noise models for w, including i.i.d. noise coordinates w_i with $\mathbb{E}\{w_i^2\} = \sigma^2 < \infty$.

Non-random signals. Our proof does not apply uniquely to random signals x with i.i.d. components, but indeed to more general sequences of signals $\{x(n)\}_{n\in\mathbb{N}}$, $x(n) \in \mathbb{R}^n$ indexed by their dimension n. The conditions required are: (1) that the empirical distribution of the coordinates of $x(n)$ converges (weakly) to p_X; and (2) that $\|x(n)\|_2^2$ converges to the second moment of the asymptotic law p_X.

Interestingly, the present framework changes the notion of 'structure' that is relevant for reconstructing the signal x. Indeed, the focus is shifted from the *sparsity* of x to the *information dimension* $\overline{d}(p_X)$.

---

[2] Unlike [KMS+11], we follow here the terminology developed within coding theory.

[3] This robustness bound holds for all $\delta > \overline{D}(p_X)$, where $\overline{D}(p_X) = \overline{d}(p_X)$ for a broad class of distributions p_X (including distributions without *singular continuous component*). When $\overline{d}(p_X) < \overline{D}(p_X)$, a somewhat weaker robustness bound holds for $\overline{d}(p_X) < \delta \le \overline{D}(p_X)$.

In the rest of this section we state formally our results, and discuss their implications and limitations, as well as relations with earlier work. Section 2.3 provides a precise description of the matrix construction and reconstruction algorithm. Section 3 reduces the proof of our main results to two key lemmas. One of these lemmas is a (quite straightforward) generalization of the state evolution technique of [DMM09, BM11a]. The second lemma characterizes the behavior of the state evolution recursion, and is proved in Section 5. The proof of a number of intermediate technical steps is deferred to the appendices.

1.2 Formal statement of the results

We consider the noisy model (2). An instance of the problem is therefore completely specified by the triple (x, w, A). We will be interested in the asymptotic properties of sequence of instances indexed by the problem dimensions $\mathcal{S} = \{(x(n), w(n), A(n))\}_{n\in\mathbb{N}}$. We recall a definition from [BM11b]. (More precisely, [BM11b] introduces the $B = 1$ case of this definition.)

Definition 1.1. *The sequence of instances $\mathcal{S} = \{x(n), w(n), A(n)\}_{n\in\mathbb{N}}$ indexed by n is said to be a* B-converging sequence *if $x(n) \in \mathbb{R}^n$, $w(n) \in \mathbb{R}^m$, $A(n) \in \mathbb{R}^{m\times n}$ with $m = m(n)$ is such that $m/n \to \delta \in (0, \infty)$, and in addition the following conditions hold:*

(a) *The empirical distribution of the entries of $x(n)$ converges weakly to a probability measure p_X on $\mathbb{R}$ with bounded second moment. Further $n^{-1}\sum_{i=1}^n x_i(n)^2 \to \mathbb{E}_{p_X}\{X^2\}$.*

(b) *The empirical distribution of the entries of $w(n)$ converges weakly to a probability measure p_W on $\mathbb{R}$ with bounded second moment. Further $m^{-1}\sum_{i=1}^m w_i(n)^2 \to \mathbb{E}_{p_W}\{W^2\} \equiv \sigma^2$.*

(c) *If $\{e_i\}_{1\le i\le n}$, $e_i \in \mathbb{R}^n$ denotes the canonical basis, then $\limsup_{n\to\infty} \max_{i\in[n]} \|A(n)e_i\|_2 \le B$, $\liminf_{n\to\infty} \min_{i\in[n]} \|A(n)e_i\|_2 \ge 1/B$.*

We further say that $\{(x(n), w(n))\}_{n\ge 0}$ is a converging sequence of instances*, if they satisfy conditions (a) and (b). We say that $\{A(n)\}_{n\ge 0}$ is a* B-converging sequence *of sensing matrices if they satisfy condition (c) above. We say $\mathcal{S}$ is a* converging sequence *if it is B-converging for some B.*

Finally, if the sequence $\{(x(n), w(n), A(n))\}_{n\ge 0}$ is random, the above conditions are required to hold almost surely.

Notice that standard normalizations of the sensing matrix correspond to $\|A(n)e_i\|_2^2 = 1$ (and hence $B = 1$) or to $\|A(n)e_i\|_2^2 = m(n)/n$. Since throughout we assume $m(n)/n \to \delta \in (0, \infty)$, these conventions only differ by a rescaling of the noise variance. In order to simplify the proofs, we allow ourselves somewhat more freedom by taking B a fixed constant.

Given a sensing matrix A, and a vector of measurements y, a reconstruction algorithm produces an estimate $\widehat{x}(A; y) \in \mathbb{R}^n$ of x. In this paper we assume that the empirical distribution p_X, and the noise level σ^2 are known to the estimator, and hence the mapping $\widehat{x} : (A, y) \mapsto \widehat{x}(A; y)$ implicitly depends on p_X and σ^2. Since however p_X, σ^2 are fixed throughout, we avoid the cumbersome notation $\widehat{x}(A, y, p_X, \sigma^2)$.

Given a converging sequence of instances $\mathcal{S} = \{x(n), w(n), A(n)\}_{n\in\mathbb{N}}$, and an estimator $\widehat{x}$, we define the asymptotic per-coordinate reconstruction mean square error as

$$\mathsf{MSE}(\mathcal{S}; \widehat{x}) = \limsup_{n\to\infty} \frac{1}{n} \left\|\widehat{x}\big(A(n); y(n)\big) - x(n)\right\|^2. \tag{4}$$

Notice that the quantity on the right hand side depends on the matrix $A(n)$, which will be random, and on the signal and noise vectors $x(n)$, $w(n)$ which can themselves be random. Our results hold almost surely with respect to these random variables. In some applications it is more customary to take the expectation with respect to the noise and signal distribution, i.e. to consider the quantity

$$\overline{\mathsf{MSE}}(\mathcal{S};\widehat{x}) = \limsup_{n\to\infty} \frac{1}{n}\,\mathbb{E}\big\|\widehat{x}\big(A(n);y(n)\big) - x(n)\big\|^2\,. \tag{5}$$

It turns out that the almost sure bounds imply, in the present setting, bounds on the expected mean square error $\overline{\mathsf{MSE}}$, as well.

In this paper we study a specific low-complexity estimator, based on the AMP algorithm first proposed in [DMM09]. This proceed by the following iteration (initialized with $x_i^1 = \mathbb{E}_{p_X} X$ for all $i \in [n]$).

$$x^{t+1} = \eta_t(x^t + (Q_t \odot A)^* r^t)\,, \tag{6}$$

$$r^t = y - Ax^t + \mathsf{b}_t \odot r^{t-1}\,. \tag{7}$$

Here, for each t, $\eta_t : \mathbb{R}^n \to \mathbb{R}^n$ is a differentiable non-linear function that depends on the input distribution p_X. Further, η_t is separable[4], namely, for a vector $v \in \mathbb{R}^n$, we have $\eta_t(v) = (\eta_{1,t}(v_1),\dots,\eta_{n,t}(v_n))$. The matrix $Q^t \in \mathbb{R}^{m\times n}$ and the vector $\mathsf{b}_t \in \mathbb{R}^m$ can be efficiently computed from the current state x^t of the algorithm, $\odot$ indicates Hadamard (entrywise) product and X^* denotes the transpose of matrix X. Further Q^t does not depend on the problem instance and hence can be precomputed. Both Q_t and b_t are block-constants. This property makes their evaluation, storage and manipulation particularly convenient. We refer to the next section for explicit definitions of these quantities. In particular, the specific choice of $\eta_{i,t}$ is dictated by the objective of minimizing the mean square error at iteration $t+1$, and hence takes the form of a Bayes optimal estimator for the prior p_X. In order to stress this point, we will occasionally refer to this as to the Bayes optimal AMP algorithm.

We denote by $\mathsf{MSE}_{\rm AMP}(\mathcal{S};\sigma^2)$ the mean square error achieved by the Bayes optimal AMP algorithm, where we made explicit the dependence on σ^2. Since the AMP estimate depends on the iteration number t, the definition of $\mathsf{MSE}_{\rm AMP}(\mathcal{S};\sigma^2)$ requires some care. The basic point is that we need to iterate the algorithm only for a constant number of iterations, as n gets large. Formally, we let

$$\mathsf{MSE}_{\rm AMP}(\mathcal{S};\sigma^2) \equiv \lim_{t\to\infty}\limsup_{n\to\infty} \frac{1}{n}\big\|x^t\big(A(n);y(n)\big) - x(n)\big\|^2\,. \tag{8}$$

As discussed above, limits will be shown to exist almost surely, when the instances $(x(n),w(n),A(n))$ are random, and almost sure upper bounds on $\mathsf{MSE}_{\rm AMP}(\mathcal{S};\sigma^2)$ will be proved. (Indeed $\mathsf{MSE}_{\rm AMP}(\mathcal{S};\sigma^2)$ turns out to be deterministic.) On the other hand, one might be interested in the expected error

$$\overline{\mathsf{MSE}}_{\rm AMP}(\mathcal{S};\sigma^2) \equiv \lim_{t\to\infty}\limsup_{n\to\infty} \frac{1}{n}\mathbb{E}\big\{\big\|x^t\big(A(n);y(n)\big) - x(n)\big\|^2\big\}\,. \tag{9}$$

We will tie the success of our compressed sensing scheme to the fundamental information-theoretic limit established in [WV10]. The latter is expressed in terms of the Rényi information dimension of the probability measure p_X.

[4]We refer to [DJM11] for a study of non-separables denoisers in AMP algorithms.

Definition 1.2. *Let p_X be a probability measure over $\mathbb{R}$, and $X \sim p_X$. The* upper *and* lower information dimension *of p_X are defined as*

$$\overline{d}(p_X) = \limsup_{\ell\to\infty} \frac{H([X]_\ell)}{\log \ell}\,. \tag{10}$$

$$\underline{d}(p_X) = \liminf_{\ell\to\infty} \frac{H([X]_\ell)}{\log \ell}\,. \tag{11}$$

Here $H(\,\cdot\,)$ denotes Shannon entropy and, for $x \in \mathbb{R}$, $[x]_\ell \equiv \lfloor \ell x\rfloor/\ell$, and $\lfloor x\rfloor \equiv \max\{k \in \mathbb{Z} \,:\quad k \le x\}$. If the lim sup *and* lim inf *coincide, then we let $d(p_X) = \overline{d}(p_X) = \underline{d}(p_X)$.*

Whenever the limit of $H([X]_\ell)/\log \ell$ exists and is finite, the Rényi information dimension can also be characterized as follows. Write the binary expansion of X, $X = D_0.D_1D_2D_3\dots$ with $D_i \in \{0,1\}$ for $i \ge 1$. Then $\overline{d}(p_X)$ is the entropy rate of the stochastic process $\{D_1, D_2, D_3, \dots\}$. It is also convenient to recall the following result form [Rén59, WV10].

Proposition 1.3 ([Rén59, WV10])**.** *Let p_X be a probability measure over $\mathbb{R}$, and $X \sim p_X$. Assume $H(\lfloor X\rfloor)$ to be finite. If $p_X = (1-\varepsilon)\nu_{\rm d} + \varepsilon\widetilde{\nu}$ with $\nu_{\rm d}$ a discrete distribution (i.e. with countable support), then $\overline{d}(p_X) \le \varepsilon$. Further, if $\widetilde{\nu}$ has a density with respect to Lebesgue measure, then $d(p_X) = \overline{d}(p_X) = \underline{d}(p_X) = \varepsilon$. In particular, if $\mathbb{P}\{X \neq 0\} \le \varepsilon$ then $\overline{d}(p_X) \le \varepsilon$.*

In order to present our result concerning the robust reconstruction, we need the definition of *MMSE dimension* of the probability measure p_X.

Given the signal distribution p_X, we let $\mathsf{mmse}(s)$ denote the minimum mean square error in estimating $X \sim p_X$ from a noisy observation in gaussian noise, at signal-to-noise ratio s. Formally

$$\mathsf{mmse}(s) \equiv \inf_{\eta:\mathbb{R}\to\mathbb{R}} \mathbb{E}\big\{\big[X - \eta(\sqrt{s}\,X + Z)\big]^2\big\}\,, \tag{12}$$

where $Z \sim \mathsf{N}(0,1)$. Since the minimum mean square error estimator is just the conditional expectation, this is given by

$$\mathsf{mmse}(s) = \mathbb{E}\big\{\big[X - \mathbb{E}[X|Y]\big]^2\big\}\,, \qquad Y = \sqrt{s}\,X + Z\,. \tag{13}$$

Notice that $\mathsf{mmse}(s)$ is naturally well defined for $s = \infty$, with $\mathsf{mmse}(\infty) = 0$. We will therefore interpret it as a function $\mathsf{mmse} : \overline{\mathbb{R}}_+ \to \overline{\mathbb{R}}_+$ where $\overline{\mathbb{R}}_+ \equiv [0,\infty]$ is the completed non-negative real line.

We recall the inequality

$$0 \le \mathsf{mmse}(s) \le \frac{1}{s}, \tag{14}$$

obtained by the estimator $\eta(y) = y/\sqrt{s}$. A finer characterization of the scaling of $\mathsf{mmse}(s)$ is provided by the following definition.

Definition 1.4 ([WV11a])**.** *The* upper *and* lower MMSE dimension *of the probability measure p_X over $\mathbb{R}$ are defined as*

$$\overline{D}(p_X) = \limsup_{s\to\infty} s\cdot\mathsf{mmse}(s)\,, \tag{15}$$

$$\underline{D}(p_X) = \liminf_{s\to\infty} s\cdot\mathsf{mmse}(s)\,. \tag{16}$$

If the lim sup *and* lim inf *coincide, then we let $D(p_X) = \overline{D}(p_X) = \underline{D}(p_X)$.*

It is also convenient to recall the following result from [WV11a].

Proposition 1.5 ([WV11a]). *If $\mathbb{E}\{X^2\} < \infty$, then*

$$\underline{D}(p_X) \le \underline{d}(p_X) \le \overline{d}(p_X) \le \overline{D}(p_X). \tag{17}$$

Hence, if $D(p_X)$ exists, then $d(p_X)$ exists and $D(p_X) = d(p_X)$. In particular, this is the case if $p_X = (1-\varepsilon)\nu_{\rm d} + \varepsilon\widetilde{\nu}$ with $\nu_{\rm d}$ a discrete distribution (i.e. with countable support), and $\widetilde{\nu}$ has a density with respect to Lebesgue measure.

We are now in position to state our main results.

Theorem 1.6. *Let p_X be a probability measure on the real line and assume*

$$\delta > \overline{d}(p_X). \tag{18}$$

Then there exists a random converging sequence of sensing matrices $\{A(n)\}_{n\ge 0}$, $A(n) \in \mathbb{R}^{m\times n}$, $m(n)/n \to \delta$ (with distribution depending only on δ), for which the following holds. For any $\varepsilon > 0$, there exists σ_0 such that for any converging sequence of instances $\{(x(n), w(n))\}_{n\ge 0}$ with parameters (p_X, σ^2, δ) and $\sigma \in [0, \sigma_0]$, we have, almost surely

$$\mathsf{MSE}_{\rm AMP}(\mathcal{S}; \sigma^2) \le \varepsilon\,. \tag{19}$$

Further, under the same assumptions, we have $\overline{\mathsf{MSE}}_{\rm AMP}(\mathcal{S}; \sigma^2) \le \varepsilon$.

Theorem 1.7. *Let p_X be a probability measure on the real line and assume*

$$\delta > \overline{D}(p_X). \tag{20}$$

Then there exists a random converging sequence of sensing matrices $\{A(n)\}_{n\ge 0}$, $A(n) \in \mathbb{R}^{m\times n}$, $m(n)/n \to \delta$ (with distribution depending only on δ) and a finite stability constant $C = C(p_X, \delta)$, such that the following is true. For any converging sequence of instances $\{(x(n), w(n))\}_{n\ge 0}$ with parameters (p_X, σ^2, δ), we have, almost surely

$$\mathsf{MSE}_{\rm AMP}(\mathcal{S}; \sigma^2) \le C\,\sigma^2\,. \tag{21}$$

Further, under the same assumptions, we have $\overline{\mathsf{MSE}}_{\rm AMP}(\mathcal{S}; \sigma^2) \le C\,\sigma^2$.

Notice that, by Proposition 1.5, $\overline{D}(p_X) \ge \overline{d}(p_X)$, and $\overline{D}(p_X) = \overline{d}(p_X)$ for a broad class of probability measures p_X, including all measures that do not have a singular continuous component (i.e. decomposes into a pure point mass component and an absolutely continuous component).

The noiseless model (1) is covered as a special case of Theorem 1.6 by taking $\sigma^2 \downarrow 0$. For the reader's convenience, we state the result explicitly as a corollary.

Corollary 1.8. *Let p_X be a probability measure on the real line. Then, for any $\delta > \overline{d}(p_X)$ there exists a random converging sequence of sensing matrices $\{A(n)\}_{n\ge 0}$, $A(n) \in \mathbb{R}^{m\times n}$, $m(n)/n \to \delta$ (with distribution depending only on δ) such that, for any sequence of vectors $\{x(n)\}_{n\ge 0}$ whose empirical distribution converges to p_X, the Bayes optimal AMP asymptotically almost surely recovers $x(n)$ from $m(n)$ measurements $y = A(n)x(n) \in \mathbb{R}^{m(n)}$. (Namely, $\mathsf{MSE}_{\rm AMP}(\mathcal{S}; 0) = 0$ almost surely, and $\overline{\mathsf{MSE}}_{\rm AMP}(\mathcal{S}; 0) = 0$.)*

1.3 Discussion

Theorem 1.6 and Corollary 1.8 are, in many ways, puzzling. It is instructive to spell out in detail a few specific examples, and discuss interesting features.

Example 1 (Bernoulli-Gaussian signal). Consider a Bernoulli-Gaussian distribution

$$p_X = (1-\varepsilon)\,\delta_0 + \varepsilon\,\gamma_{\mu,\sigma} \tag{22}$$

where $\gamma_{\mu,\sigma}(\mathrm{d}x) = (2\pi\sigma^2)^{-1/2}\exp\{-(x-\mu)^2/(2\sigma^2)\}\mathrm{d}x$ is the Gaussian measure with mean μ and variance σ^2. This model has been studied numerically in a number of papers, including [BSB19, KMS+11]. By Proposition 1.3, we have $\overline{d}(p_X) = \varepsilon$, and by Proposition 1.5, $\overline{D}(p_X) = \underline{D}(p_X) = \varepsilon$ as well.

Construct random signals $x(n) \in \mathbb{R}^n$ by sampling i.i.d. coordinates $x(n)_i \sim p_X$. Glivenko-Cantelli's theorem implies that the empirical distribution of the coordinates of $x(n)$ converges almost surely to p_X, hence we can apply Corollary 1.8 to recover $x(n)$ from $m(n) = n\varepsilon + o(n)$ measurements $y(n) \in \mathbb{R}^{m(n)}$. Notice that the number of non-zero entries in $x(n)$ is, almost surely, $k(n) = n\varepsilon + o(n)$. Hence, we can restate the implication of Corollary 1.8 as follows. A sequence of vectors $x(n)$ with Bernoulli-Gaussian distribution and $k(n)$ nonzero entries can almost surely recovered by $m(n) = k(n) + o(n)$ measurements.

Example 2 (Mixture signal with a point mass). The above remarks generalize immediately to arbitrary mixture distributions of the form

$$p_X = (1-\varepsilon)\,\delta_0 + \varepsilon\,q\,, \tag{23}$$

where q is a measure that is absolutely continuous with respect to Lebesgue measure, i.e. $q(\mathrm{d}x) = f(x)\,\mathrm{d}x$ for some measurable function f. Then, by Proposition 1.3, we have $\overline{d}(p_X) = \varepsilon$, and by Proposition 1.5, $\overline{D}(p_X) = \underline{D}(p_X) = \varepsilon$ as well. Arguing as above we have the following.

Consequence 1.9. *Let $\{x(n)\}_{n\geq 0}$ be a sequence of vectors with i.i.d. components $x(n)_i \sim p_X$ where p_X is a mixture distribution as per Eq. (23). Denote by $k(n)$ the number of nonzero entries in $x(n)$. Then, almost surely as $n \to \infty$, Bayes optimal AMP recovers the signal $x(n)$ from $m(n) = k(n)+o(n)$ spatially coupled measurements.*

Under the regularity hypotheses of [WV10], no scheme can do substantially better, i.e. reconstruct $x(n)$ from $m(n)$ measurements if $\limsup_{n\to\infty} m(n)/k(n) < 1$.

One way to think about this result is the following. If an oracle gave us the support of $x(n)$, we would still need $m(n) \geq k(n) - o(n)$ measurements to reconstruct the signal. Indeed, the entries in the support have distribution q, and $\overline{d}(q) = 1$. Corollary 1.8 implies that the measurements overhead for estimating the support of $x(n)$ is sublinear, $o(n)$, even when the support is of order n.

It is sometimes informally argued that compressed sensing requires at least $\Theta(k\log(n/k))$ for 'information-theoretic reasons', namely that specifying the support requires about $nH(k/n) \approx k\log(n/k)$ bits. This argument is of course incomplete because it assumes that each measurement y_i is described by a bounded number of bits. Lower bounds of the form $m \geq C\,k\log(n/k)$ are proved in the literature but they do not contradict our results. Specifically, [Wai09, ASZ10] prove information-theoretic lower bounds on the required number of measurements, under specific constructions for the random sensing matrix A. Further, these papers focus on the specific problem of exact support recovery. The paper [RWY09] proves minimax bounds for reconstructing vectors belonging to ℓ_p-balls.

However, as the noise variance tends to zero, these bounds depend on the sensing matrix in a way that is difficult to quantify. In particular, they provide no explicit lower bound on the number of measurements required for exact recovery in the noiseless limit. Similar bounds were obtained for arbitrary measurement matrices in [CD11]. Again, these lower bounds vanish as noise tends to zero as soon as $m(n) \geq k(n)$.

A different line of work derives lower bounds from Gelfand' width arguments [Don06a, KT07]. These lower bounds are only proved to be a necessary condition for a stronger reconstruction guarantees. Namely, these works require the vector of measurements $y = Ax$ to enable recovery for *all* k-sparse vectors $x \in \mathbb{R}^n$. This corresponds to the 'strong' phase transition of [DT05, Don06b], and is also referred to as the 'for all' guarantee in the computer science literature [BGI$^+$08].

The lower bound that comes closest to the present setting is the 'randomized' lower bound [BIPW10]. The authors use an elegant communication complexity argument to show that $m(n) = \Omega(k(n)\log(n/k(n)))$ is necessary for achieving stable recovery with an $\ell_1 - \ell_1$ error guarantee. This is a stronger stability condition than what is achieved in Theorem 1.7, allowing for a more powerful noise process. Indeed the same paper also proves that recovery is possible from $m(n) = O(k(n))$ measurements under stronger conditions.

Example 3 (Discrete signal). Let K be a fixed integer, $a_1, \dots, a_K \in \mathbb{R}$, and $(p_1, p_2, \dots, p_K)$ be a collection of non-negative numbers that add up to one. Consider the probability distribution that puts mass p_i on each a_i

$$p_X = \sum_{i=1}^{K} p_i \, \delta_{a_i} \,, \tag{24}$$

and let $x(n)$ be a signal with i.i.d. coordinates $x(n)_i \sim p_X$. By Proposition 1.3, we have $\overline{d}(p_X) = 0$. As above, the empirical distribution of the coordinates of the vectors $x(n)$ converges to p_X. By applying Corollary 1.8 we obtain the following

Consequence 1.10. *Let $\{x(n)\}_{n\geq 0}$ be a sequence of vectors with i.i.d. components $x(n)_i \sim p_X$ where p_X is a discrete distribution as per Eq. (24). Then, almost surely as $n \to \infty$, Bayes optimal AMP recovers the signal $x(n)$ from $m(n) = o(n)$ spatially coupled measurements.*

It is important to further discuss the last statement because the reader might be misled into too optimistic a conclusion. Consider any signal $x \in \mathbb{R}^n$. For practical purposes, this will be represented with finite precision, say as a vector of ℓ-bit numbers. Hence, in practice, the distribution p_X is always discrete, with $K = 2^\ell$. One might conclude from the above that a sublinear number of measurements $m(n) = o(n)$ is sufficient for any signal.

This is of course too optimistic. The key point is that Theorem 1.6 and Corollary 1.8 are asymptotic statements. As demonstrated in [KMS$^+$11], for some classes of signals this asymptotic behavior is already relevant when n is of the order of a few thousands. On the other hand, the same will not be true for a discrete signal with a large number of levels K (which is the case for an ℓ-bit representation as in the above example, with ℓ moderately large). In particular, a necessary condition in that case is $n \gg K$. It would of course be important to substantiate/refine such a rule of thumb by numerical simulations or non-asymptotic bounds.

Example 4 (A discrete-continuous mixture). Consider the probability distribution

$$p_X = \varepsilon_+ \, \delta_{+1} + \varepsilon_- \delta_{-1} + \varepsilon \, q \,, \tag{25}$$

where $\varepsilon_+ + \varepsilon_- + \varepsilon = 1$ and the probability measure q has a density with respect to Lebesgue measure. Again, let $x(n)$ be a vector with i.i.d. components $x(n)_i \sim p_X$. We can apply Corollary 1.8 to conclude that $m(n) = n\varepsilon + o(n)$ spatially coupled measurements are sufficient. This should be contrasted with the case of sensing matrices with i.i.d. entries studied in [DT10] under convex reconstruction methods. In this case $m(n) = n(1+\varepsilon)/2 + o(n)$ measurements are necessary.

In the next section we describe the basic intuition behind the surprising phenomenon in Theorems 1.6 and 1.7, and why are spatially-coupled sensing matrices so useful. We conclude by stressing once more the limitations of these results:

- The Bayes-optimal AMP algorithm requires knowledge of the signal distribution p_X. Notice however that only a good approximation of p_X (call it $p_{\widetilde{X}}$, and denote by $\widetilde{X}$ the corresponding random variable) is sufficient. Assume indeed that p_X and $p_{\widetilde{X}}$ can be coupled in such a way that $\mathbb{E}\{(X - \widetilde{X})^2\} \le \tilde{\sigma}^2$. Then

$$x = \widetilde{x} + u \tag{26}$$

  where $\|u\|_2^2 \lesssim n\tilde{\sigma}^2$. This is roughly equivalent to adding to the noise vector z further 'noise' $\widetilde{z}$ with variance $\tilde{\sigma}^2/\delta$. Indeed, it can be shown that the guarantee in Theorem 1.7 degrades gracefully as $p_{\widetilde{X}}$ gets different from p_X.

  Finally, it was demonstrated numerically in [VS11, KMS+11] that, in some cases, a good 'proxy' for p_X can be learnt through an EM-style iteration.

- As mentioned above, the guarantees in Theorems 1.6 and 1.7 are only asymptotic. It would be important to develop analogous non-asymptotic results.

- The stability bound (21) is non-uniform, in that the proportionality constant C depends on the signal distribution. It would be important to establish analogous bounds that are uniform over suitable classes of distributions. (We do not expect Eq. (21) to hold uniformly over *all* distributions.)

1.4 How does spatial coupling work?

Spatially-coupled sensing matrices A are –roughly speaking– band diagonal matrices. It is convenient to think of the graph structure that they induce on the reconstruction problem. Associate one node (a *variable node* in the language of factor graphs) to each coordinate i in the unknown signal x. Order these nodes on the real line $\mathbb{R}$, putting the i-th node at location $i \in \mathbb{R}$. Analogously, associate a node (a *factor node*) to each coordinate a in the measurement vector y, and place the node a at position a/δ on the same line. Connect this node to all the variable nodes i such that $A_{ai} \neq 0$. If A is band diagonal, only nodes that are placed close enough will be connected by an edge. See Figure 1 for an illustration.

In a spatially coupled matrix, additional measurements are associated to the first few coordinates of x, say coordinates $x_1, \dots, x_{n_0}$ with n_0 much smaller than n. This has a negligible impact on the overall undersampling ratio as $n/n_0 \to \infty$. Although the overall undersampling remains $\delta < 1$, the coordinates $x_1, \dots, x_{n_0}$ are oversampled. This ensures that these first coordinates are recovered correctly (up to a mean square error of order σ^2). As the algorithm is iterated, the contribution of these first few coordinates is correctly subtracted from all the measurements, and hence we can

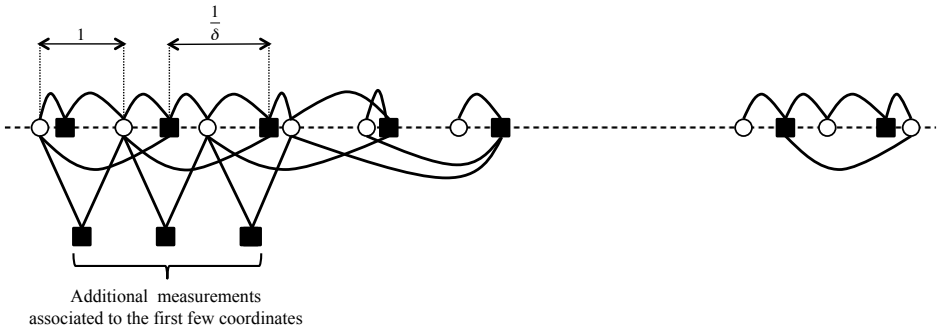

Figure 1: Graph structure of a spatially coupled matrix. Variable nodes are shown as circle and check nodes are represented by square.

effectively eliminate those nodes from the graph. In the resulting graph, the first few variables are effectively oversampled and hence the algorithm will reconstruct their values, up to a mean square error of order σ^2. As the process is iterated, variables are progressively reconstructed, proceeding from left to right along the node layout.

While the above explains the basic dynamics of AMP reconstruction algorithms under spatial coupling, a careful consideration reveals that this picture leaves open several challenging questions. In particular, why does the overall undersampling factor δ have to exceed $\overline{d}(p_X)$ for reconstruction to be successful? Our proof is based on a potential function argument. We will prove that there exists a potential function for the AMP algorithm, such that, when $\delta > \overline{d}(p_X)$, this function has its global minimum close to exact reconstruction. Further, we will prove that, unless this minimum is essentially achieved, AMP can always decrease the function. This technique is different from the one followed in [KRU11] for the LDPC codes over the binary erasure channel, and we think it is of independent interest.

1.5 Further related work

The most closely related earlier work was already discussed above.

More broadly, message passing algorithms for compressed sensing where the object of a number of studies studies, starting with [BSB19]. As mentioned, we will focus on approximate message passing (AMP) as introduced in [DMM09, DMM10]. As shown in [DJM11] these algorithms can be used in conjunction with a rich class of denoisers $\eta(\,\cdot\,)$. A subset of these denoisers arise as posterior mean associated to a prior p_X. Several interesting examples were studied by Schniter and collaborators [Sch10, Sch11, SPS10], and by Rangan and collaborators [Ran11, KGR11].

Spatial coupling has been the object of growing interest within coding theory over the last few years. The first instance of spatially coupled code ensembles were the convolutional LDPC codes of Felström and Zigangirov [FZ99]. While the excellent performances of such codes had been known for quite some time [SLJZ04], the fundamental reason was not elucidated until recently [KRU11] (see also [LF10]). In particular [KRU11] proved –for communication over the binary erasure channel (BEC)– that the thresholds of spatially coupled ensembles under message passing decoding coincide with the

thresholds of the base LDPC code under MAP decoding. In particular, this implies that spatially coupled ensembles achieve capacity over the BEC. The analogous statement for general memoryless symmetric channels remains open, but substantial evidence was put forward in [KMRU10]. The paper [HMU10] discusses similar ideas in a number of graphical models.

The first application of spatial coupling ideas to compressed sensing is due to Kudekar and Pfister [KP10]. Their proposed message passing algorithms do not make use of the signal distribution p_X, and do not fully exploit the potential of spatially coupled matrices. The message passing algorithm used here belongs to the general class introduced in [DMM09]. The specific use of the minimum-mean square error denoiser was suggested in [DMM10]. The same choice is made in [KMS$^+$11].

Finally, let us mention that robust sparse recovery of k-sparse vectors from $m = O(k \log\log(n/k))$ measurement is possible, using suitable 'adaptive' sensing schemes [IPW11].

2 Matrix and algorithm construction

In this section, we define an ensemble of random matrices, and the corresponding choices of Q_t, b_t, η_t that achieve the reconstruction guarantees in Theorems 1.6 and 1.7. We proceed by first introducing a general ensemble of random matrices. Correspondingly, we define a deterministic recursion named state evolution, that plays a crucial role in the algorithm analysis. In Section 2.3, we define the algorithm parameters and construct specific choices of Q_t, b_t, η_t. The last section also contains a restatement of Theorems 1.6 and 1.7, in which this construction is made explicit.

2.1 General matrix ensemble

The sensing matrix A will be constructed randomly, from an ensemble denoted by $\mathcal{M}(W, M, N)$. The ensemble depends on two integers $M, N \in \mathbb{N}$, and on a matrix with non-negative entries $W \in \mathbb{R}_+^{\mathsf{R}\times\mathsf{C}}$, whose rows and columns are indexed by the finite sets R, C (respectively 'rows' and 'columns'). The matrix is *roughly row-stochastic*, i.e.

$$\frac{1}{2} \le \sum_{c\in\mathsf{C}} W_{r,c} \le 2\,, \qquad \text{for all } r \in \mathsf{R}\,. \tag{27}$$

We will let $|\mathsf{R}| \equiv L_r$ and $|\mathsf{C}| = L_c$ denote the matrix dimensions. The ensemble parameters are related to the sensing matrix dimensions by $n = NL_c$ and $m = ML_r$.

In order to describe a random matrix $A \sim \mathcal{M}(W, M, N)$ from this ensemble, partition the columns and row indices in –respectively– L_c and L_r groups of equal size. Explicitly

$$[n] = \cup_{s\in\mathsf{C}} C(s)\,, \quad |C(s)| = N\,,$$
$$[m] = \cup_{r\in\mathsf{R}} R(r)\,, \quad |R(r)| = M\,.$$

Here and below we use $[k]$ to denote the set of first k integers $[k] \equiv \{1, 2, \dots, k\}$. Further, if $i \in R(r)$ or $j \in C(s)$ we will write, respectively, $r = \mathsf{g}(i)$ or $s = \mathsf{g}(j)$. In other words $\mathsf{g}(\,\cdot\,)$ is the operator determining the group index of a given row or column.

With this notation we have the following concise definition of the ensemble.

Definition 2.1. *A random sensing matrix A is distributed according to the ensemble $\mathcal{M}(W, M, N)$ (and we write $A \sim \mathcal{M}(W, M, N)$) if the entries $\{A_{ij},\ \ i \in [m], j \in [n]\}$ are independent Gaussian*

random variables with [5]

$$A_{ij} \sim \mathsf{N}\Big(0, \frac{1}{M}\, W_{\mathsf{g}(i),\mathsf{g}(j)}\Big)\,. \tag{28}$$

2.2 State evolution

State evolution allows an exact asymptotic analysis of AMP algorithms in the limit of a large number of dimensions. As indicated by the name, it bears close resemblance to the density evolution method in iterative coding theory [RU08]. Somewhat surprisingly, this analysis approach is asymptotically exact despite the underlying factor graph being far from locally tree-like.

State evolution was first developed in [DMM09] on the basis of heuristic arguments, and substantial numerical evidence. Subsequently, it was proved to hold for Gaussian sensing matrices with i.i.d. entries, and a broad class of iterative algorithm in [BM11a]. These proofs were further generalized in [Ran11], to cover 'generalized' AMP algorithms.

In the present case, state evolution takes the following form. [6]

Definition 2.2. *Given $W \in \mathbb{R}_+^{\mathsf{R}\times\mathsf{C}}$ roughly row-stochastic, and $\delta > 0$, the corresponding* state evolution maps $\mathsf{T}'_W : \mathbb{R}_+^{\mathsf{R}} \to \mathbb{R}_+^{\mathsf{C}}$, $\mathsf{T}''_W : \mathbb{R}_+^{\mathsf{C}} \to \mathbb{R}_+^{\mathsf{R}}$, *are defined as follows. For $\phi = (\phi_a)_{a\in\mathsf{R}} \in \mathbb{R}_+^{\mathsf{R}}$, $\psi = (\psi_i)_{i\in\mathsf{C}} \in \mathbb{R}_+^{\mathsf{C}}$, we let:*

$$\mathsf{T}'_W(\phi)_i \;=\; \mathsf{mmse}\Big(\sum_{b\in\mathsf{R}} W_{b,i}\phi_b^{-1}\Big)\,, \tag{29}$$

$$\mathsf{T}''_W(\psi)_a \;=\; \sigma^2 + \frac{1}{\delta}\sum_{i\in\mathsf{C}} W_{a,i}\,\psi_i\,. \tag{30}$$

We finally define $\mathsf{T}_W = \mathsf{T}'_W \circ \mathsf{T}''_W$.

In the following, we shall omit the subscripts from T_W whenever clear from the context.

Definition 2.3. *Given $W \in \mathbb{R}_+^{L_r\times L_c}$ roughly row-stochastic, the corresponding* state evolution sequence *is the sequence of vectors $\{\phi(t), \psi(t)\}_{t\ge 0}$, $\phi(t) = (\phi_a(t))_{a\in\mathsf{R}} \in \mathbb{R}_+^{\mathsf{R}}$, $\psi(t) = (\psi_i(t))_{i\in\mathsf{C}} \in \mathbb{R}_+^{\mathsf{C}}$, defined recursively by $\phi(t) = \mathsf{T}''_W(\psi(t))$, $\psi(t+1) = \mathsf{T}'_W(\phi(t))$, with initial condition*

$$\psi_i(0) = \infty \text{ for all } i \in \mathsf{C}\,. \tag{31}$$

Hence, for all $t \ge 0$,

$$\begin{aligned} \phi_a(t) &= \quad \sigma^2 + \frac{1}{\delta}\sum_{i\in\mathsf{C}} W_{a,i}\,\psi_i(t)\,,\\ \psi_i(t+1) &= \mathsf{mmse}\Big(\sum_{b\in\mathsf{R}} W_{b,i}\phi_b^{-1}(t)\Big)\,. \end{aligned} \tag{32}$$

[5] As in many papers on compressed sensing, the matrix here has independent Gaussian entries; however, unlike standard practice, here the entries are of widely different variances.

[6] In previous work, the state variable concerned a single scalar, representing the mean-squared error in the current reconstruction, averaged across all coordinates. In this paper, the dimensionality of the state variable is much larger, because it contains ψ an individualized MSE for each coordinate of the reconstruction and also ϕ a pseudo-data MSE for each measurement coordinate.

In words, and as implicit in Definition 2.3, $\psi = T'_W(\phi)$ computes the formal MSE ψ of coordinates of x, for a specified level of formal MSE ϕ of coordinates of pseudo-data r. Similarly $\phi = T''_W(\psi)$ computes the formal MSE in components of r given the formal MSE of coordinates of x. Section 4 below gives an heuristic derivation of state evolution.

2.3 General algorithm definition

In order to fully define the AMP algorithm (6), (7), we need to provide constructions for the matrix Q^t, the nonlinearities η_t, and the vector b_t. In doing this, we exploit the fact that the state evolution sequence $\{\phi(t)\}_{t\ge 0}$ can be precomputed.

We define the matrix Q^t by

$$Q^t_{ij} \equiv \frac{\phi_{\mathsf{g}(i)}(t)^{-1}}{\sum_{k=1}^{L_r} W_{k,\mathsf{g}(j)}\phi_k(t)^{-1}}\,. \tag{33}$$

Notice that Q^t is block-constant: for any $r, s \in [L]$, the block $Q^t_{R(r),C(s)}$ has all its entries equal.

As mentioned in Section 1, the function $\eta_t : \mathbb{R}^n \to \mathbb{R}^n$ is chosen to be separable, i.e. for $v \in \mathbb{R}^N$:

$$\eta_t(v) = \big(\eta_{t,1}(v_1), \eta_{t,2}(v_2),\ \dots\ , \eta_{t,N}(v_N)\big)\,. \tag{34}$$

We take $\eta_{t,i}$ to be a conditional expectation estimator for $X \sim p_X$ in gaussian noise:

$$\eta_{t,i}(v_i) = \mathbb{E}\big\{X \,\big|\, X + s_{\mathsf{g}(i)}(t)^{-1/2}Z = v_i\,\big\}\,, \quad s_r(t) \equiv \sum_{u\in\mathsf{R}} W_{u,r}\phi_u(t)^{-1}\,. \tag{35}$$

Notice that the function $\eta_{t,i}(\,\cdot\,)$ depends on i only through the group index $\mathsf{g}(i)$, and in fact only parametrically through $s_{\mathsf{g}(i)}(t)$.

Finally, in order to define the vector b^t_i, let us introduce the quantity

$$\langle \eta'_t\rangle_u = \frac{1}{N}\sum_{i\in C(u)} \eta'_{t,i}\big(x^t_i + ((Q^t\odot A)^* r^t)_i\big)\,. \tag{36}$$

The vector b^t is then defined by

$$\mathsf{b}^t_i \equiv \frac{1}{\delta}\sum_{u\in\mathsf{C}} W_{\mathsf{g}(i),u}\widetilde{Q}^{t-1}_{\mathsf{g}(i),u}\,\langle \eta'_{t-1}\rangle_u\,, \tag{37}$$

where we defined $Q^t_{i,j} = \widetilde{Q}^t_{r,u}$ for $i \in R(r)$, $j \in C(u)$. Again b^t_i is block-constant: the vector $\mathsf{b}^t_{C(u)}$ has all its entries equal.

This completes our definition of the AMP algorithm. Let us conclude with a few computational remarks:

1. The quantities $\tilde{Q}^t$, $\phi(t)$ can be precomputed efficiently iteration by iteration, because they are –respectively– $L_r \times L_c$ and L_r-dimensional, and, as discussed further below, L_r, L_c are much smaller than m, n. The most complex part of this computation is implementing the iteration (32), which has complexity $O((L_r+L_c)^3)$, plus the complexity of evaluating the mmse function, which is a one-dimensional integral.

2. The vector b^t is also block-constant, so can be efficiently computed using Eq. (37).

3. Instead of computing $\phi(t)$ analytically by iteration (32), $\phi(t)$ can also be estimated from data x^t, r^t. In particular, by generalizing the methods introduced in [DMM09, Mon12], we get the estimator

$$\widehat{\phi}_a(t) = \frac{1}{M} \|r^t_{R(a)}\|_2^2 \,, \tag{38}$$

where $r^t_{R(a)} = (r^t_j)_{j\in R(a)}$ is the restriction of r^t to the indices in $R(a)$. An alternative more robust estimator, would be

$$\widehat{\phi}_a(t)^{1/2} = \frac{1}{\Phi^{-1}(3/4)} \, |r^t_{R(a)}|_{(M/2)} \,, \tag{39}$$

where $\Phi(z)$ is the Gaussian distribution function, and, for $v \in \mathbb{R}^K$, $|v|_{(\ell)}$ is the ℓ-th largest entry in the vector $(|v_1|, |v_2|, \dots, |v_K|)$. The idea underlying both of the above estimator is that the components of $r^t_{R(a)}$ are asymptotically i.i.d. with mean zero and variance $\phi_a(t)$

2.4 Choices of parameters

In order to prove our main Theorem 1.6, we use a sensing matrix from the ensemble $\mathcal{M}(W, M, N)$ for a suitable choice of the matrix $W \in \mathbb{R}^{\mathsf{R}\times\mathsf{C}}$. Our construction depends on parameters $\rho \in \mathbb{R}_+$, $L, L_0 \in \mathbb{N}$, and on the 'shape function' $\mathcal{W}$. As explained below, ρ will be taken to be small, and hence we will treat $1/\rho$ as an integer to avoid rounding (which introduces in any case a negligible error).

Definition 2.4. *A* shape function *is a function $\mathcal{W} : \mathbb{R} \to \mathbb{R}_+$ continuously differentiable, with support in $[-1, 1]$ and such that $\int_{\mathbb{R}} \mathcal{W}(u)\,\mathrm{d}u = 1$, and $\mathcal{W}(-u) = \mathcal{W}(u)$.*

We let $\mathsf{C} \cong \{-2\rho^{-1}, \dots, 0, 1, \dots, L-1\}$, so that $L_c = L + 2\rho^{-1}$. The rows are partitioned as follows:

$$\mathsf{R} = \mathsf{R}_0 \cup \Big\{ \cup_{i=-2\rho^{-1}}^{-1} \mathsf{R}_i \Big\} \,,$$

where $\mathsf{R}_0 \cong \{-\rho^{-1}, \dots, 0, 1, \dots, L-1+\rho^{-1}\}$, and $|\mathsf{R}_i| = L_0$. Hence $L_r = L_c + 2\rho^{-1}L_0$.

Finally, we take N so that $n = NL_c$, and let $M = N\delta$ so that $m = ML_r = N(L_c + 2\rho^{-1}L_0)\delta$. Notice that $m/n = \delta(L_c + 2\rho^{-1}L_0)/L_c$. Since we will take L_c much larger than L_0/ρ, we in fact have m/n arbitrarily close to δ.

Given these inputs, we construct the corresponding matrix $W = W(L, L_0, \mathcal{W}, \rho)$ as follows

1. For $i \in \{-2\rho^{-1}, \dots, -1\}$, and each $a \in \mathsf{R}_i$, we let $W_{a,i} = 1$. Further, $W_{a,j} = 0$ for all $j \in \mathsf{C}\setminus\{i\}$.

2. For all $a \in \mathsf{R}_0 \cong \{-\rho^{-1}, \dots, 0, \dots, L-1+\rho^{-1}\}$, we let

$$W_{a,i} = \rho\, \mathcal{W}\big(\rho\,(a-i)\big) \quad i \in \{-2\rho^{-1}, \dots, L-1\}. \tag{40}$$

See Fig. 2 for an illustration of the matrix W. In the following we occasionally use the shorthand $W_{a-i} \equiv \rho\, \mathcal{W}\big(\rho\,(a-i)\big)$.

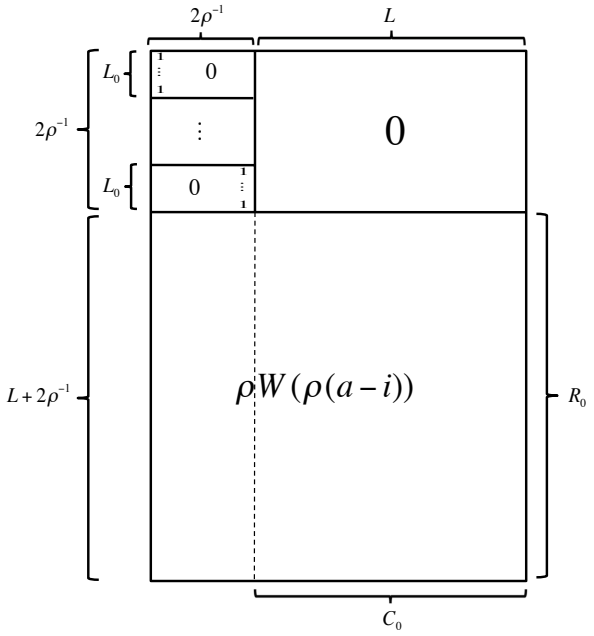

Figure 2: Matrix W

It is not hard to check that W is roughly row-stochastic. Also, the restriction of W to columns in C_0 is roughly column-stochastic.

We are now in position to restate Theorem 1.6 in a more explicit form.

Theorem 2.5. *Let p_X be a probability measure on the real line with $\delta > \overline{d}(p_X)$, and let $\mathcal{W} : \mathbb{R} \to \mathbb{R}_+$ be a shape function. For any $\varepsilon > 0$, there exist L_0, L, ρ, t_0, $\sigma_0^2 > 0$ such that $L_0/(L\rho) \le \varepsilon$, and further the following holds true for $W = W(L, L_0, \mathcal{W}, \rho)$.*

For $N \ge 0$, and $A(n) \sim \mathcal{M}(W, M, N)$ with $M = N\delta$, and for all $\sigma^2 \le \sigma_0^2$, $t \ge t_0$, we almost surely have

$$\limsup_{N\to\infty} \frac{1}{n} \big\| x^t\big(A(n); y(n)\big) - x(n) \big\|^2 \le \varepsilon\,. \tag{41}$$

Further, under the same assumptions, we have

$$\limsup_{N\to\infty} \frac{1}{n} \mathbb{E}\big\{ \big\| x^t\big(A(n); y(n)\big) - x(n) \big\|^2 \big\} \le \varepsilon\,. \tag{42}$$

In order to obtain a stronger form of robustness, as per Theorem 1.7, we slightly modify the sensing scheme. We construct the sensing matrix $\tilde{A}$ from A by appending $2\rho^{-1}L_0$ rows in the bottom.

$$\tilde{A} = \left(\begin{array}{c|c} A & \\ \hline 0 & I \end{array}\right), \tag{43}$$

where I is the identity matrix of dimensions $2\rho^{-1}L_0$. Note that this corresponds to increasing the number of measurements; however, the asymptotic undersampling rate remains δ, provided that $L_0/(L\rho) \to 0$, as $n \to \infty$.

The reconstruction scheme is modified as follows. Let x_1 be the vector obtained by restricting x to entries in $\cup_i C(i)$, where $i \in \{-2\rho^{-1}, \cdots, L - 2\rho^{-1} - 1\}$. Also, let x_2 be the vector obtained by restricting x to entries in $\cup_i C(i)$, where $i \in \{L - 2\rho^{-1}, \cdots, L - 1\}$. Therefore, $x = (x_1, x_2)^T$. Analogously, let $y = (y_1, y_2)^T$ where y_1 is given by the restriction of y to $\cup_{i\in \mathsf{R}} R(i)$ and y_2 corresponds to the additional $2\rho^{-1}L_0$ rows. Define w_1 and w_2 from the noise vector w, analogously. Hence,

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \left(\begin{array}{c|c} A & \\ \hline 0 & I \end{array}\right) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}. \tag{44}$$

Note that the sampling rate for vector x_2 is one, i.e., y_2 and x_2 are of the same length and are related to each other through the identity matrix I. Hence, we have a fairly good approximation of these entries. We use the AMP algorithm as described in the previous section to obtain an estimation of x_1. Formally, let x^t be the estimation at iteration t obtained by applying the AMP algorithm. The modified estimation is then $\tilde{x}^t = (x_1^t, y_2)^T$.

As we will see later, this modification in the sensing matrix and algorithm, while not necessary, simplifies some technical steps in the proof.

Theorem 2.6. *Let p_X be a probability measure on the real line with $\delta > \overline{D}(p_X)$, and let $\mathcal{W} : \mathbb{R} \to \mathbb{R}_+$ be a shape function. There exist L_0, L, ρ, t_0 and a finite stability constant $C = C(p_X, \delta)$, such that $L_0/(L\rho) < \varepsilon$, for any given $\varepsilon > 0$, and the following holds true for the modified reconstruction scheme.*

For $t \geq t_0$, we almost surely have,

$$\limsup_{N\to\infty} \frac{1}{n} \big\| \tilde{x}^t\big(A(n); y(n)\big) - x(n) \big\|^2 \leq C\sigma^2. \tag{45}$$

Further, under the same assumptions, we have

$$\limsup_{N\to\infty} \frac{1}{n} \mathbb{E}\big\{ \big\| \tilde{x}^t\big(A(n); y(n)\big) - x(n) \big\|^2 \big\} \leq C\sigma^2. \tag{46}$$

It is obvious that Theorems 2.5 and 2.6 respectively imply Theorems 1.6 and 1.7. We shall therefore focus on the proofs of Theorems 2.5 and 2.6 in the rest of the paper.

3 Key lemmas and proof of the main theorems

Our proof is based in a crucial way on state evolution. This effectively reduces the analysis of the algorithm (6), (7) to the analysis of the deterministic recursion (32).

Lemma 3.1. *Let $W \in \mathbb{R}_+^{\mathsf{R}\times\mathsf{C}}$ be a roughly row-stochastic matrix (see Eq. (27))and $\phi(t)$, Q^t, b_t be defined as in Section 2.3. Let $M = M(N)$ be such that $M/N \to \delta$, as $N \to \infty$. Define $m = ML_r$, $n = NL_c$, and for each $N \geq 1$, let $A(n) \sim \mathcal{M}(W, M, N)$. Let $\{(x(n), w(n))\}_{n\geq 0}$ be a converging sequence of instances with parameters (p_X, σ^2). Then, for all $t \geq 1$, almost surely we have*

$$\limsup_{N\to\infty} \frac{1}{N} \|x^t_{C(i)}(A(n); y(n)) - x_{C(i)}\|_2^2 = \mathsf{mmse}\Big(\sum_{a\in\mathsf{R}} W_{a,i} \phi_a^{-1}(t-1) \Big). \tag{47}$$

for all $i \in \mathsf{C}$.

This lemma is a straightforward generalization of [BM11a]. Since a formal proof does not require new ideas, but a significant amount of new notations, it is presented in a separate forthcoming publication [BLM12] which covers an even more general setting. In the interest of self-containedness, and to develop useful intuition on state evolution, we present an heuristic derivation of the state evolution equations (32) in Section 4.

The next Lemma provides the needed analysis of the recursion (32).

Lemma 3.2. *Let $\delta > 0$, and p_X be a probability measure on the real line. Let $\mathcal{W} : \mathbb{R} \to \mathbb{R}_+$ be a shape function.*

(a) If $\delta > \overline{d}(p_X)$, then for any $\varepsilon > 0$, there exist $\sigma_0, \rho, L_* > 0$, such that for any $\sigma^2 \in [0, \sigma_0^2]$, $L_0 > 3/\delta$, and $L > L_*$, the following holds for $W = W(L, L_0, \mathcal{W}, \rho)$:

$$\lim_{t\to\infty} \frac{1}{L} \sum_{a=-\rho^{-1}}^{L+\rho^{-1}-1} \phi_a(t) \le \varepsilon. \tag{48}$$

(b) If $\delta > \overline{D}(p_X)$, then there exist $\rho, L_* > 0$, and a finite stability constant $C = C(p_X, \delta)$, such that for $L_0 > 3/\delta$, and $L > L_*$, the following holds for $W = W(L, L_0, \mathcal{W}, \rho)$.

$$\lim_{t\to\infty} \frac{1}{L} \sum_{a=-\rho^{-1}}^{L-\rho^{-1}-1} \phi_a(t) \le C\sigma^2. \tag{49}$$

The proof of this lemma is deferred to Section 5 and is indeed the technical core of the paper.

Now, we have in place all we need to prove our main results.

Proof (Theorem 2.5). Recall that $\mathsf{C} \cong \{-2\rho^{-1} \cdots, L-1\}$. Therefore,

$$\begin{aligned}
\limsup_{N\to\infty} \frac{1}{n} \big\| x^t\big(A(n); y(n)\big) - x(n) \big\|^2 &= \frac{1}{L_c} \sum_{i\in\mathsf{C}} \limsup_{N\to\infty} \frac{1}{N} \big\| x^t_{C(i)}\big(A(n); y(n)\big) - x_{C(i)}(n) \big\|^2 \\
&\overset{(a)}{\le} \frac{1}{L_c} \sum_{i=-2\rho^{-1}}^{L-1} \mathsf{mmse}\left(\sum_{a\in\mathsf{R}} W_{a,i} \phi_a^{-1}(t-1) \right) \\
&\overset{(b)}{\le} \frac{1}{L_c} \sum_{i=-2\rho^{-1}}^{L-1} \mathsf{mmse}\left(\frac{1}{2} \phi^{-1}_{i+1/\rho}(t-1) \right) \\
&\overset{(c)}{\le} \frac{1}{L_c} \sum_{a=-\rho^{-1}}^{L+\rho^{-1}-1} 2\phi_a(t-1).
\end{aligned} \tag{50}$$

Here, (a) follows from Lemma 3.1; (b) follows from the fact that $\phi_a(t)$ is nondecreasing in a for every t (see Lemma 5.9 below) and from the fact that W is roughly column-stochastic; (c) follows from the inequality $\mathsf{mmse}(s) \le 1/s$. The result is immediate due to Lemma 3.2, Part (a).

The claim regarding the expected error follows readily since X has bounded second moment. □

Proof (Theorem 2.6). The proof proceeds in a similar manner to the proof of Theorem 1.6.

$$\begin{aligned}
&\limsup_{N\to\infty}\frac{1}{n}\big\|\tilde{x}^t\big(A(n);y(n)\big)-x(n)\big\|^2 \\
&\qquad=\frac{1}{L_c}\Big\{\sum_{i=-2\rho^{-1}}^{L-2\rho^{-1}-1}\limsup_{N\to\infty}\frac{1}{N}\big\|x^t_{C(i)}\big(A(n);y(n)\big)-x_{C(i)}(n)\big\|^2+\lim_{N\to\infty}\frac{1}{N}\big\|w_2(n)\big\|^2\Big\} \\
&\qquad\le\frac{1}{L_c}\Big\{\sum_{i=-2\rho^{-1}}^{L-2\rho^{-1}-1}\mathsf{mmse}\left(\sum_{a\in\mathsf{R}}W_{a,i}\phi_a^{-1}(t-1)\right)+\lim_{N\to\infty}\frac{1}{N}\big\|w_2(n)\big\|^2\Big\} \qquad (51)\\
&\qquad\le\frac{1}{L_c}\Big\{\sum_{i=-2\rho^{-1}}^{L-2\rho^{-1}-1}\mathsf{mmse}\left(\frac{1}{2}\phi^{-1}_{i+1/\rho}(t-1)\right)+\lim_{N\to\infty}\frac{1}{N}\big\|w_2(n)\big\|^2\Big\} \\
&\qquad\le\frac{1}{L_c}\Big\{\sum_{a=-\rho^{-1}}^{L-\rho^{-1}-1}2\phi_a(t-1)+\lim_{N\to\infty}\frac{1}{N}\big\|w_2(n)\big\|^2\Big\}\le C\,\sigma^2,
\end{aligned}$$

where the last step follows from Part (b) in Lemma 3.1, and Part (b) in Definition 1.1.

Again, the claim regarding the expected error is immediate since X has bounded second moment. □

4 State evolution: an heuristic derivation

This section presents an heuristic derivation of the state evolution equations (32). Our objective is to provide some basic intuition: a proof in a more general setting will appear in a separate publication [BLM12]. An heuristic derivation similar to the present one, for the special cases of sensing matrices with i.i.d. entries was presented in [BM11a].

Consider the recursion (32), and introduce the following modifications: (i) At each iteration, replace the random matrix A with a new independent copy $A(t)$; (ii) Replace the observation vector y with $y^t=A(t)x_0+w$; (iii) Eliminate the last term in the update equation for r^t. Then, we have the following update rules:

$$x^{t+1} = \eta_t(x^t+(Q_t\odot A(t))^*r^t)\,, \qquad (52)$$
$$r^t = y^t-A(t)x^t\,, \qquad (53)$$

where $A(0),A(1),A(2),\cdots$ are i.i.d. random matrices distributed according to the ensemble $\mathcal{M}(W,M,N)$, i.e.,

$$A_{ij}(t)\sim\mathsf{N}\Big(0,\frac{1}{M}W_{\mathsf{g}(i),\mathsf{g}(j)}\Big). \qquad (54)$$

Rewriting the recursion by eliminating r^t, we obtain:

$$\begin{aligned}
x^{t+1}&=\eta_t((Q_t\odot A(t))^*y^t+(I-(Q_t\odot A(t))^*A(t))x^t)\\
&=\eta_t(x_0+(Q_t\odot A(t))^*w+B(t)(x^t-x_0))\,,
\end{aligned} \qquad (55)$$

where $B(t)=I-(Q_t\odot A(t))^*A(t)\in\mathbb{R}^{n\times n}$. Note that the recursion (55) does not correspond to the AMP update rules defined per Eqs. (6) and (7). In particular, it does not correspond to

any practical algorithm. However, it is much easier to analyze, as it allows to neglect completely correlations induced by the fact that we should use the same sensing matrix A across different iterations. Also, it is useful for presenting the intuition behind the AMP algorithm and to emphasize the role of the term $b_t \odot r^{t-1}$ in the update rule for r^t. As it emerges from the proof of [BM11a], this term does asymptotically cancels dependencies across iterations.

By virtue of the central limit theorem, each entry of $B(t)$ is approximately normal. More specifically, $B_{ij}(t)$ is approximately normal with mean zero and variance $(1/M)\sum_{r\in\mathsf{R}} W_{r,\mathsf{g}(i)}W_{r,\mathsf{g}(j)}Q_{r,\mathsf{g}(i)}$, for $i,j\in[n]$. Define $\hat{\tau}_t(s) = \lim_{N\to\infty}\|x^t_{C(s)} - x_{C(s)}\|^2/N$, for $s\in\mathsf{C}$. It is easy to show that distinct entries in $B(t)$ are approximately independent. Also, $B(t)$ is independent of $\{B(s)\}_{1\le s\le t-1}$, and in particular, of $x^t - x_0$. Hence, $B(t)(x^t - x_0)$ converges to a vector, say v, with i.i.d. normal entries, and for $i\in[n]$,

$$\mathbb{E}\{v_i\} = 0, \quad \mathbb{E}\{v_i^2\} = \frac{N}{M}\sum_{s\in\mathsf{C}}\sum_{r\in\mathsf{R}} W_{r,\mathsf{g}(i)}W_{r,s}Q_{r,\mathsf{g}(i)}\,\hat{\tau}_t^2(s). \tag{56}$$

Conditional on w, $(Q_t \odot A(t))^* w$ is a vector of i.i.d. normal entries with mean 0. Also, the variance of its i^{th} entry, for $i\in[n]$, is

$$\frac{1}{M}\sum_{r\in\mathsf{R}} W_{r,\mathsf{g}(i)}Q_{r,\mathsf{g}(i)}\|w_{R(r)}\|^2 = \frac{1}{M}\sum_{r\in\mathsf{R}}\|w_{R(r)}\|^2, \tag{57}$$

which converges to σ^2, by the law of large numbers. With slightly more work, it can be shown that these entries are approximately independent of the ones of $B(t)(x^t - x_0)$.

Summarizing, the i^{th} entry of the vector in the argument of η_t in Eq. (55) converges to $X + \tau_t(\mathsf{g}(i))Z$ with $Z\sim\mathsf{N}(0,1)$ independent of X, and

$$\begin{aligned}\tau_t^2(s) &= \sigma^2 + \frac{1}{\delta}\sum_{u\in\mathsf{C}}\sum_{r\in\mathsf{R}} W_{r,s}W_{r,u}Q_{r,s}\,\hat{\tau}_t^2(u),\\ \hat{\tau}_t^2(s) &= \lim_{N\to\infty}\frac{1}{N}\|x^t_{C(s)} - x_{C(s)}\|^2,\end{aligned} \tag{58}$$

for $s\in\mathsf{C}$. In addition, using Eq. (55) and invoking Eqs. (34), (35), each entry of $x^{t+1}_{C(s)} - x_{C(s)}$ converges to $\eta_{t,s}(X + \tau_t(s)Z) - X$, for $s\in\mathsf{C}$. Therefore,

$$\begin{aligned}\hat{\tau}_{t+1}^2(s) &= \lim_{N\to\infty}\frac{1}{N}\|x^{t+1}_{C(s)} - x_{C(s)}\|^2\\ &= \mathbb{E}\{[\eta_{t,s}(X + \tau_t(s)Z) - X]^2\} = \mathsf{mmse}(\tau_t^{-2}(s)).\end{aligned} \tag{59}$$

Using Eqs. (58) and (59), we obtain:

$$\tau_{t+1}(s)^2 = \sigma^2 + \frac{1}{\delta}\sum_{r\in\mathsf{R}} W_{r,s}Q_{r,s}\Big(\sum_{u\in\mathsf{C}} W_{r,u}\,\mathsf{mmse}(\tau_t^{-2}(u))\Big). \tag{60}$$

Applying the change of variable $\tau_t^{-2}(u) = \sum_{b\in\mathsf{R}} W_{b,u}\phi_b^{-1}(t)$, and substituting for $Q_{r,s}$, from Eq. (33), we obtain the state evolution recursion, Eq. (32).

In conclusion, we showed that the state evolution recursion would hold if the matrix A was resampled independently from the ensemble $\mathcal{M}(W,M,N)$, at each iteration. However, in our proposed

AMP algorithm, the matrix A is constant across iterations, and the above argument is not valid since x^t and A are dependent. This dependency cannot be neglected even in the large system limit $N \to \infty$. Indeed, the term $b_t \odot r^{t-1}$ in the update rule for r^t (which was removed in our argument above) leads to an asymptotic cancellation of these dependencies as in [BM11a].

5 Analysis of state evolution

Throughout this section p_X is a given probability distribution over the real line, and $X \sim p_X$. Also, we will take $\sigma > 0$. The result for the noiseless model (Corollary 1.8) follows by letting $\sigma \downarrow 0$. Recall the inequality

$$\mathsf{mmse}(s) \le \min(\mathrm{Var}(X), \frac{1}{s}) \,. \tag{61}$$

Definition 5.1. *For two vectors $\phi, \tilde{\phi} \in \mathbb{R}^K$, we write $\phi \succeq \tilde{\phi}$ if all $\phi_r \ge \tilde{\phi}_r$ for $r \in \{1, \dots, K\}$.*

Proposition 5.2. *For any $W \in \mathbb{R}_+^{\mathsf{R}\times\mathsf{C}}$, the map $\mathsf{T}_W : \mathbb{R}_+^{\mathsf{R}} \to \mathbb{R}_+^{\mathsf{R}}$ is monotone; i.e., if $\phi \succeq \tilde{\phi}$ then $\mathsf{T}_W(\phi) \succeq \mathsf{T}_W(\tilde{\phi})$. Analogously, T'_W and T''_W are also monotone.*

Proof. It follows immediately from the fact that $s \mapsto \mathsf{mmse}(s)$ is a monotone decreasing function. □

Proposition 5.3. *The state evolution sequence $\{\phi(t), \psi(t)\}_{t\ge 0}$ with initial condition $\psi_i(0) = \infty$, for $i \in \mathsf{C}$, is monotone decreasing, in the sense that $\phi(0) \succeq \phi(1) \succeq \phi(2) \succeq \dots$ and $\psi(0) \succeq \psi(1) \succeq \psi(2) \succeq \dots$.*

Proof. Since $\psi_i(0) = \infty$ for all i, we have $\psi(0) \succeq \psi(1)$. The thesis follows from the monotonicity of the state evolution map. □

Lemma 5.4. *Assume $\delta L_0 > 3$. Then there exists t_0 (depending only on p_X), such that, for all $t \ge t_0$ and all $i \in \{-2\rho^{-1}, \dots, -1\}$, $a \in \mathsf{R}_i$, we have*

$$\psi_i(t) \;\le\; \mathsf{mmse}\Big(\frac{L_0}{2\sigma^2}\Big) \le \frac{2\sigma^2}{L_0} \,, \tag{62}$$

$$\phi_a(t) \;\le\; \sigma^2 + \frac{1}{\delta}\mathsf{mmse}\Big(\frac{L_0}{2\sigma^2}\Big) \le \Big(1 + \frac{2}{\delta L_0}\Big)\sigma^2 \,. \tag{63}$$

Proof. Take $i \in \{-2\rho^{-1}, \cdots, -1\}$. For $a \in \mathsf{R}_i$, we have $\phi_a(t) = \sigma^2 + (1/\delta)\psi_i(t)$. Further from $\mathsf{mmse}(s) \le 1/s$, we deduce that

$$\begin{aligned} \psi_i(t+1) = \mathsf{mmse}\Big(\sum_{b\in\mathsf{R}} W_{b,i}\phi_b^{-1}(t)\Big) &\le \Big(\sum_{b\in\mathsf{R}} W_{b,i}\phi_b^{-1}(t)\Big)^{-1} \\ &\le \Big(\sum_{a\in\mathsf{R}_i} W_{a,i}\phi_a^{-1}(t)\Big)^{-1} = \Big(L_0\phi_a^{-1}(t)\Big)^{-1} = \frac{\phi_a(t)}{L_0}. \end{aligned} \tag{64}$$

Substituting in the earlier relation, we get $\psi_i(t+1) \le (1/L_0)(\sigma^2 + (1/\delta)\psi_i(t))$. Recalling that $\delta L_0 > 3$, we have $\psi_i(t) \le 2\sigma^2/L_0$, for all t sufficiently large. Now, using this in the equation for $\phi_a(t)$, $a \in \mathsf{R}_i$, we obtain

$$\phi_a(t) = \sigma^2 + \frac{1}{\delta}\psi_i(t) \le \Big(1 + \frac{2}{\delta L_0}\Big)\sigma^2. \tag{65}$$

We prove the other claims by repeatedly substituting in the previous bounds. In particular,

$$\begin{aligned}\psi_i(t) &= \mathsf{mmse}\Big(\sum_{b\in\mathsf{R}} W_{b,i}\phi_b^{-1}(t-1)\Big) \le \mathsf{mmse}\Big(\sum_{a\in\mathsf{R}_i} W_{a,i}\phi_a^{-1}(t)\Big)\\ &= \mathsf{mmse}(L_0\phi_a^{-1}(t)) \le \mathsf{mmse}\Big(\frac{L_0}{(1+\frac{2}{\delta L_0})\sigma^2}\Big) \le \mathsf{mmse}\Big(\frac{L_0}{2\sigma^2}\Big),\end{aligned} \tag{66}$$

where we used Eq. (65) in the penultimate inequality. Finally,

$$\phi_a(t) \le \sigma^2 + \frac{1}{\delta}\psi_i(t) \le \sigma^2 + \frac{1}{\delta}\mathsf{mmse}\Big(\frac{L_0}{2\sigma^2}\Big), \tag{67}$$

where the inequality follows from Eq. (66). □

Next we prove a lower bound on the state evolution sequence. Here and below $\mathsf{C}_0 \equiv \mathsf{C}\setminus\{-2\rho^{-1},\dots,-1\} \cong \{0,\dots,L-1\}$. Also, recall that $\mathsf{R}_0 \equiv \{-\rho^{-1},\dots,0,\dots,L-1+\rho^{-1}\}$. (See Fig. 2).

Lemma 5.5. *For any $t \ge 0$, and any $i \in \mathsf{C}_0$, $\psi_i(t) \ge \mathsf{mmse}(2\sigma^{-2})$. Further, for any $a \in \mathsf{R}_0$ and any $t \ge 0$ we have $\phi_a(t) \ge \sigma^2 + (2\delta)^{-1}\mathsf{mmse}(2\sigma^2)$.*

Proof. Since $\phi_a(t) \ge \sigma^2$ by definition, we have, for $i \ge 0$, $\psi_i(t) \ge \mathsf{mmse}(\sigma^{-2}\sum_b W_{bi}) \ge \mathsf{mmse}(2\sigma^{-2})$, where we used the fact that the restriction of W to columns in C_0 is roughly column-stochastic. Plugging this into the expression for ϕ_a, we get

$$\phi_a(t) \ge \sigma^2 + \frac{1}{\delta}\sum_{i\in\mathsf{C}} W_{a,i}\,\mathsf{mmse}(2\sigma^{-2}) \ge \sigma^2 + \frac{1}{2\delta}\mathsf{mmse}(2\sigma^{-2})\,. \tag{68}$$

□

Notice that for $L_{0,*} \ge 4$ and for all $L_0 > L_{0,*}$, the upper bound for $\psi_i(t)$, $i \in \{-2\rho^{-1},\cdots,-1\}$, given in Lemma 5.4 is below the lower bound for $\psi_i(t)$, with $i \in \mathsf{C}_0$, given in Lemma 5.5; i.e. for all σ,

$$\mathsf{mmse}\Big(\frac{L_0}{2\sigma^2}\Big) \le \mathsf{mmse}\Big(\frac{2}{\sigma^2}\Big)\,. \tag{69}$$

Motivated by the above, we introduce modified state evolution maps $\mathsf{F}'_W : \mathbb{R}_+^{\mathsf{R}_0} \to \mathbb{R}_+^{\mathsf{C}_0}$, $\mathsf{F}''_W : \mathbb{R}_+^{\mathsf{C}_0} \to \mathbb{R}_+^{\mathsf{R}_0}$, by letting, for $\phi = (\phi_a)_{a\in\mathsf{R}_0} \in \mathbb{R}_+^{\mathsf{R}_0}$, $\psi = (\psi_i)_{i\in\mathsf{C}_0} \in \mathbb{R}_+^{\mathsf{C}_0}$, and for all $i \in \mathsf{C}_0$, $a \in \mathsf{R}_0$:

$$\mathsf{F}'_W(\phi)_i = \mathsf{mmse}\Big(\sum_{b\in\mathsf{R}_0} W_{b-i}\phi_b^{-1}\Big)\,, \tag{70}$$

$$\mathsf{F}''_W(\psi)_a = \sigma^2 + \frac{1}{\delta}\sum_{i\in\mathbb{Z}} W_{a-i}\,\psi_i\,. \tag{71}$$

where, in the last equation we set by convention, $\psi_i(t) = \mathsf{mmse}(L_0/(2\sigma^2))$ for $i \le -1$, and $\psi_i = \infty$ for $i \ge L$. We also let $\mathsf{F}_W = \mathsf{F}'_W \circ \mathsf{F}''_W$.

Definition 5.6. *The* modified state evolution sequence *is the sequence* $\{\phi(t), \psi(t)\}_{t\geq 0}$ *with* $\phi(t) = \mathsf{F}'_W(\psi(t))$ *and* $\psi(t+1) = \mathsf{F}'_W(\phi(t))$ *for all* $t \geq 0$, *and* $\psi_i(0) = \infty$ *for all* $i \in \mathsf{C}_0$. *We also adopt the convention that, for* $i \geq L$, $\psi_i(t) = +\infty$ *and for* $i \leq -1$, $\psi_i(t) = \mathsf{mmse}(L_0/(2\sigma^2))$, *for all* t.

Lemma 5.4 then implies the following.

Lemma 5.7. *Let* $\{\phi(t), \psi(t)\}_{t\geq 0}$ *denote the state evolution sequence as per Definition 2.3, and* $\{\phi^{\mathrm{mod}}(t), \psi^{\mathrm{mod}}(t)\}_{t\geq 0}$ *denote the modified state evolution sequence as per Definition 5.6. Then, there exists* t_0 *(depending only on* p_X*), such that, for all* $t \geq t_0$, $\phi(t) \preceq \phi^{\mathrm{mod}}(t-t_0)$ *and* $\psi(t) \preceq \psi^{\mathrm{mod}}(t-t_0)$.

Proof. Choose $t_0 = t(L_0, \delta)$ as given by Lemma 5.4. We prove the claims by induction on t. For the induction basis ($t = t_0$), we have from Lemma 5.4, $\psi_i(t_0) \leq \mathsf{mmse}(L_0/(2\sigma^2)) = \psi_i^{\mathrm{mod}}(0)$, for $i \leq -1$. Also, we have $\psi_i^{\mathrm{mod}}(0) = \infty \geq \psi_i(t_0)$, for $i \geq 0$. Further,

$$\phi_a^{\mathrm{mod}}(0) = \mathsf{F}''_W(\psi^{\mathrm{mod}}(0))_a \geq \mathsf{T}''_W(\psi^{\mathrm{mod}}(0))_a \geq \mathsf{T}''_W(\psi(t_0))_a = \phi_a(t_0), \tag{72}$$

for $a \in \mathsf{R}_0$. Here, the last inequality follows from monotonicity of T''_W (Proposition 5.2). Now, assume that the claim holds for t; we prove it for $t+1$. For $i \in \mathsf{C}_0$, we have

$$\begin{aligned}\psi_i^{\mathrm{mod}}(t+1-t_0) &= \mathsf{F}'_W(\phi^{\mathrm{mod}}(t-t_0))_i = \mathsf{T}'_W(\phi^{\mathrm{mod}}(t-t_0))_i \\ &\geq \mathsf{T}'_W(\phi(t))_i = \psi_i(t+1),\end{aligned} \tag{73}$$

where the inequality follows from monotonicity of T'_W (Proposition 5.2) and the induction hypothesis. In addition, for $a \in \mathsf{R}_0$,

$$\begin{aligned}\phi_a^{\mathrm{mod}}(t+1-t_0) &= \mathsf{F}''_W(\psi^{\mathrm{mod}}(t+1-t_0))_a \geq \mathsf{T}''_W(\psi^{\mathrm{mod}}(t+1-t_0))_a \\ &\geq \mathsf{T}''_W(\psi(t+1))_a = \phi_a(t+1).\end{aligned} \tag{74}$$

Here, the last inequality follows from monotonicity of T''_W and Eq. (73). □

By Lemma 5.7, we can now focus on the modified state evolution sequence in order to prove Lemma 3.2. Notice that the mapping F_W has a particularly simple description in terms of a shift-invariant state evolution mapping. Explicitly, define $\mathsf{T}'_{W,\infty} : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{Z}}$, $\mathsf{T}''_{W,\infty} : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{Z}}$, by letting, for $\phi, \psi \in \mathbb{R}^{\mathbb{Z}}$ and all $i, a \in \mathbb{Z}$:

$$\mathsf{T}'_{W,\infty}(\phi)_i = \mathsf{mmse}\Big(\sum_{b\in\mathbb{Z}} W_{b-i}\phi_b^{-1}\Big), \tag{75}$$

$$\mathsf{T}''_{W,\infty}(\psi)_a = \sigma^2 + \frac{1}{\delta}\sum_{i\in\mathbb{Z}} W_{a-i}\,\psi_i\,. \tag{76}$$

Further, define the embedding $\mathsf{H} : \mathbb{R}^{\mathsf{C}_0} \to \mathbb{R}^{\mathbb{Z}}$ by letting

$$(\mathsf{H}\psi)_i = \begin{cases} \mathsf{mmse}(L_0/(2\sigma^2)) & \text{if } i < 0, \\ \psi_i & \text{if } 0 \leq i \leq L-1, \\ +\infty & \text{if } i \geq L, \end{cases} \tag{77}$$

And the restriction mapping $\mathsf{H}'_{a,b} : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^{b-a+1}$ by $\mathsf{H}'_{a,b}\psi = (\psi_a, \dots, \psi_b)$.

Lemma 5.8. *With the above definitions,* $\mathsf{F}_W = \mathsf{H}'_{0,L-1} \circ \mathsf{T}_{W,\infty} \circ \mathsf{H}$.

Proof. Clearly, for any $\psi = (\psi_i)_{i\in\mathsf{C}_0}$, we have $\mathsf{T}''_W \circ \mathsf{H}(\psi)_a = \mathsf{F}''_W \circ \mathsf{H}(\psi)_a$ for $a \in \mathsf{R}_0$, since the definition of the embedding H is consistent with the convention adopted in defining the modified state evolution. Moreover, for $i \in \mathsf{C}_0 \cong \{0, \dots, L-1\}$, we have

$$\begin{aligned}\mathsf{T}'_{W,\infty}(\phi)_i &= \mathsf{mmse}\Big(\sum_{b\in\mathbb{Z}} W_{b-i}\phi_b^{-1}\Big) = \mathsf{mmse}\Big(\sum_{-\rho^{-1}\le b\le L-1+\rho^{-1}} W_{b-i}\phi_b^{-1}\Big)\\ &= \mathsf{mmse}\Big(\sum_{b\in\mathsf{R}_0} W_{b-i}\phi_b^{-1}\Big) = \mathsf{F}'_W(\phi)_i.\end{aligned} \tag{78}$$

Hence, $\mathsf{T}'_{W,\infty}\circ\mathsf{T}''_{W,\infty}\circ\mathsf{H}(\psi)_i = \mathsf{F}'_W\circ\mathsf{F}''_W\circ\mathsf{H}(\psi)_i$, for $i \in \mathsf{C}_0$. Therefore, $\mathsf{H}'_{0,L-1}\circ\mathsf{T}_{W,\infty}\circ\mathsf{H}(\psi) = \mathsf{F}_W\circ\mathsf{H}(\psi)$, for any $\psi \in \mathbb{R}_+^{\mathsf{C}_0}$, which completes the proof. □

We will say that $\psi \in \mathbb{R}^K$ is *nondecreasing* if, for every $1 \le i < j \le K$, $\psi_i \le \psi_j$.

Lemma 5.9. *If* $\psi \in \mathbb{R}^{\mathsf{C}_0}$ *is nondecreasing, with* $\psi_i \ge \mathsf{mmse}(L_0/(2\sigma^2))$ *for all* i*, then* $\mathsf{F}_W(\psi)$ *is nondecreasing as well. In particular, if* $\{\phi(t), \psi(t)\}_{t\ge 0}$ *is the modified state evolution sequence, then* $\phi(t)$ *and* $\psi(t)$ *are nondecreasing for all* t.

Proof. By Lemma 5.8, we know that $\mathsf{F}_W = \mathsf{H}'_{0,L-1}\circ\mathsf{T}_{W,\infty}\circ\mathsf{H}$. We first notice that, by the assumption $\psi_i \ge \mathsf{mmse}(L_0/(2\sigma^2))$, we have that $\mathsf{H}(\psi)$ is nondecreasing.

Next, if $\psi \in \mathbb{R}^{\mathbb{Z}}$ is nondecreasing, $\mathsf{T}_{W,\infty}(\psi)$ is nondecreasing as well. In fact, the mappings $\mathsf{T}'_{W,\infty}$ and $\mathsf{T}''_{W,\infty}$ both preserve the nondecreasing property, since both are shift invariant, and $\mathsf{mmse}(\,\cdot\,)$ is a decreasing function. Finally, the restriction of a nondecreasing vector is obviously nondecreasing.

This proves that F_W preserves the nondecreasing property. To conclude that $\psi(t)$ is nondecreasing for all t, notice that the condition $\psi_i(t) \ge \mathsf{mmse}(L_0/(2\sigma^2))$ is satisfied at all t by Lemma 5.5 and condition (69). The claim for $\psi(t)$ follows by induction.

Now, since F''_W preserves the nondecreasing property, we have $\phi(t) = \mathsf{F}''_W(\psi(t))$ is nondecreasing for all t, as well. □

5.1 Continuum state evolution

We start by defining the continuum state evolution mappings. For $\Omega \subseteq \mathbb{R}$, let $\mathscr{M}(\Omega)$ be the space of non-negative measurable functions on Ω (up to measure-zero redefinitions). Define $\mathcal{F}'_{\mathcal{W}} : \mathscr{M}([-1, \ell+1]) \to \mathscr{M}([0,\ell])$ and $\mathcal{F}''_{\mathcal{W}} : \mathscr{M}([0,\ell]) \to \mathscr{M}([-1,\ell+1])$ as follows. For $\phi \in \mathscr{M}([-1,\ell+1]), \psi \in \mathscr{M}([0,\ell])$, and for all $x \in [0,\ell], y \in [-1,\ell+1]$, we let

$$\mathcal{F}'_{\mathcal{W}}(\phi)(x) = \mathsf{mmse}\Big(\int_{-1}^{\ell+1} \mathcal{W}(x-z)\phi^{-1}(z)\mathrm{d}z\Big), \tag{79}$$

$$\mathcal{F}''_{\mathcal{W}}(\psi)(y) = \sigma^2 + \frac{1}{\delta}\int_{\mathbb{R}} \mathcal{W}(y-x)\psi(x)\mathrm{d}x\,, \tag{80}$$

where we adopt the convention that $\psi(x) = \mathsf{mmse}(L_0/(2\sigma^2))$ for $x < 0$, and $\psi(x) = \infty$ for $x > \ell$.

Definition 5.10. *The continuum state evolution sequence is the sequence* $\{\phi(\,\cdot\,;t), \psi(\,\cdot\,;t)\}_{t\ge 0}$*, with* $\phi(t) = \mathcal{F}''_{\mathcal{W}}(\psi(t))$ *and* $\psi(t+1) = \mathcal{F}'_{\mathcal{W}}(\phi(t))$ *for all* $t \ge 0$*, and* $\psi(x;0) = \infty$ *for all* $x \in [0,\ell]$.

Recalling Eq. (61), we have $\psi(x;t) = \mathcal{F}'_{\mathcal{W}}(\phi(t-1))(x) \le \mathrm{Var}(X)$, for $t \ge 1$. Also, $\phi(x;t) = \mathcal{F}''_{\mathcal{W}}(\psi(t))(x) \le \sigma^2 + (1/\delta)\mathrm{Var}(X)$, for $t \ge 1$. Define,

$$\Phi_M = 1 + \frac{1}{\delta}\,\mathrm{Var}(X). \tag{81}$$

Assuming $\sigma < 1$, we have $\phi(x;t) < \Phi_M$, for all $t \ge 1$.

Lemma 5.11. *Let $\{\phi(\,\cdot\,;t), \psi(\,\cdot\,;t)\}_{t\ge 0}$ be the continuum state evolution sequence and $\{\phi(t), \psi(t)\}_{t\ge 0}$ be the modified discrete state evolution sequence, with parameters ρ and $L = \ell/\rho$. Then for any $t \ge 0$*

$$\lim_{\rho\to 0} \frac{1}{L} \sum_{i=0}^{L-1} \big|\psi_i(t) - \psi(\rho i; t)\big| = 0\,, \tag{82}$$

$$\lim_{\rho\to 0} \frac{1}{L} \sum_{a=-\rho^{-1}}^{L-\rho^{-1}-1} \big|\phi_a(t) - \phi(\rho a; t)\big| = 0\,. \tag{83}$$

Lemma 5.11 is proved in Appendix A.

Corollary 5.12. *The continuum state evolution sequence $\{\phi(\,\cdot\,;t), \psi(\,\cdot\,;t)\}_{t\ge 0}$, with initial condition $\psi(x) = \mathsf{mmse}(L_0/(2\sigma^2))$ for $x < 0$, and $\psi(x) = \infty$ for $x > \ell$, is monotone decreasing, in the sense that $\phi(x;0) \ge \phi(x;1) \ge \phi(x;2) \ge \cdots$ and $\psi(x;0) \ge \psi(x;1) \ge \psi(x;2) \ge \cdots$, for all $x \in [0,\ell]$.*

Proof. Follows immediately from Lemmas 5.3 and 5.11. □

Corollary 5.13. *Let $\{\phi(\,\cdot\,;t), \psi(\,\cdot\,;t)\}_{t\ge 2}$ be the continuum state evolution sequence. Then for any t, $x \mapsto \psi(x;t)$ and $x \mapsto \phi(x;t)$ are nondecreasing Lipschitz continuos functions.*

Proof. Nondecreasing property of functions $x \mapsto \psi(x;t)$, and $x \mapsto \phi(x;t)$ follows immediately from Lemmas 5.9 and 5.11. Further, since $\psi(x;t)$ is bounded for $t \ge 1$, and $\mathcal{W}(\,\cdot\,)$ is Lipschitz continuos, recalling Eq. (80), the function $x \mapsto \phi(x;t)$ is Lipschitz continuos as well, for $t \ge 1$. Similarly, since $\sigma^2 < \phi(x;t) < \Phi_M$, invoking Eq. (79), the function $x \mapsto \psi(x;t)$ is Lipschitz continuos for $t \ge 2$. □

Free Energy. We define the mutual information between X and a noisy observation of X at signal-to-noise ratio s by

$$\mathsf{I}(s) \equiv I(X; \sqrt{s}X + Z)\,, \tag{84}$$

with $Z \sim \mathsf{N}(0,1)$ independent of $X \sim p_X$. Recall the relation [GSV05]

$$\frac{\mathrm{d}}{\mathrm{d}s}\,\mathsf{I}(s) = \frac{1}{2}\,\mathsf{mmse}(s)\,. \tag{85}$$

Furthermore, the following identities relate the scaling law of mutual information under weak noise to Rényi information dimension [WV11a].

Proposition 5.14. *Assume $H(\lfloor X \rfloor) < \infty$. Then*

$$\begin{aligned} \liminf_{s\to\infty} \frac{\mathsf{I}(s)}{\frac{1}{2}\log s} &= \underline{d}(p_X), \\ \limsup_{s\to\infty} \frac{\mathsf{I}(s)}{\frac{1}{2}\log s} &= \overline{d}(p_X). \end{aligned} \tag{86}$$

A key role in our analysis is played by the free energy functional.

Definition 5.15. *Let $\mathcal{W}(.)$ be a shape function, and $\sigma, \delta > 0$ be given. The corresponding* free energy *is the functional $\mathsf{E}_W : \mathscr{M}([-1, \ell+1]) \to \overline{\mathbb{R}}$ defined as follows for $\phi \in \mathscr{M}([-1, \ell+1])$:*

$$\mathsf{E}_{\mathcal{W}}(\phi) = \frac{\delta}{2}\int_{-1}^{\ell-1}\Big\{\frac{\sigma^2(x)}{\phi(x)} + \log\phi(x)\Big\}\mathrm{d}x + \int_0^{\ell}\mathsf{I}\Big(\int \mathcal{W}(x-z)\phi^{-1}(z)\mathrm{d}z\Big)\mathrm{d}x, \tag{87}$$

where

$$\sigma^2(x) = \sigma^2 + \frac{1}{\delta}\Big(\int_{y\le 0}\mathcal{W}(y-x)\mathrm{d}y\Big)\mathsf{mmse}\Big(\frac{L_0}{2\sigma^2}\Big). \tag{88}$$

Viewing $\mathsf{E}_{\mathcal{W}}$ as a function defined on the Banach space $L_2([-1,\ell])$, we will denote by $\nabla E_{\mathcal{W}}(\phi)$ its Fréchet derivative at ϕ. This will be identified, via standard duality, with a function in $L_2([-1,\ell])$. It is not hard to show that the Fréchet derivative exists on $\{\phi:\ \phi(x) \ge \sigma^2\}$ and is such that

$$\nabla\mathsf{E}_{\mathcal{W}}(\phi)(y) = \frac{\delta}{2\phi^2(y)}\Big\{\phi(y) - \sigma^2(y) - \frac{1}{\delta}\int_0^{\ell}\mathcal{W}(x-y)\mathsf{mmse}\Big(\int\mathcal{W}(x-z)\phi^{-1}(z)\mathrm{d}z\Big)\mathrm{d}x\Big\}, \tag{89}$$

for $-1 \le y \le \ell - 1$.

Corollary 5.16. *If $\{\phi, \psi\}$ is the fixed point of the continuum state evolution, then $\nabla\mathsf{E}_W(\phi)(y) = 0$, for $-1 \le y \le \ell - 1$.*

Proof. We have $\phi = \mathcal{F}''_{\mathcal{W}}(\psi)$ and $\psi = \mathcal{F}'_{\mathcal{W}}(\phi)$, whereby for $-1 \le y \le \ell - 1$,

$$\begin{aligned}
\phi(y) &= \sigma^2 + \frac{1}{\delta}\int\mathcal{W}(y-x)\psi(x)\mathrm{d}x \\
&= \sigma^2 + \frac{1}{\delta}\Big(\int_{x\le 0}\mathcal{W}(y-x)\mathrm{d}x\Big)\mathsf{mmse}\left(\frac{L_0}{2\sigma^2}\right) + \\
&\quad \frac{1}{\delta}\int_0^{\ell}\mathcal{W}(y-x)\mathsf{mmse}\Big(\int_{-1}^{\ell+1}\mathcal{W}(x-z)\phi^{-1}(z)\mathrm{d}z\Big)\mathrm{d}x \\
&= \sigma^2(y) + \frac{1}{\delta}\int_0^{\ell}\mathcal{W}(y-x)\mathsf{mmse}\Big(\int_{-1}^{\ell+1}\mathcal{W}(x-z)\phi^{-1}(z)\mathrm{d}z\Big)\mathrm{d}x.
\end{aligned} \tag{90}$$

The result follows immediately from Eq. (89). □

Definition 5.17. *Define the* potential *function $V : \mathbb{R}_+ \to \mathbb{R}_+$ as follows.*

$$V(\phi) = \frac{\delta}{2}\Big(\frac{\sigma^2}{\phi} + \log\phi\Big) + \mathsf{I}(\phi^{-1}). \tag{91}$$

Using Eq. (86), we have for $\phi \ll 1$,

$$\begin{aligned}
V(\phi) &\lesssim \frac{\delta}{2}(\frac{\sigma^2}{\phi} + \log\phi) + \frac{1}{2}\overline{d}(p_X)\log(\phi^{-1}) \\
&= \frac{\delta\sigma^2}{2\phi} + \frac{1}{2}[\delta - \overline{d}(p_X)]\log(\phi).
\end{aligned} \tag{92}$$

Define

$$\phi^* = \sigma^2 + \frac{1}{\delta}\mathsf{mmse}\Big(\frac{L_0}{2\sigma^2}\Big). \tag{93}$$

Notice that $\sigma^2 < \phi^* \le (1 + 2/(\delta L_0))\sigma^2 < 2\sigma^2$, given that $\delta L_0 > 3$. The following proposition upper bounds $V(\phi^*)$ and its proof is deferred to Appendix B.

Proposition 5.18. *There exists $\sigma_2 > 0$, such that, for $\sigma \in (0, \sigma_2]$, we have*

$$V(\phi^*) \le \frac{\delta}{2} + \frac{\delta - \overline{d}(p_X)}{4}\log(2\sigma^2). \tag{94}$$

Now, we write the energy functional in term of the potential function.

$$\mathsf{E}_{\mathcal{W}}(\phi) = \int_{-1}^{\ell-1} V(\phi(x))\,\mathrm{d}x + \frac{\delta}{2}\int_{-1}^{\ell-1}\frac{\sigma^2(x) - \sigma^2}{\phi(x)}\,\mathrm{d}x + \tilde{\mathsf{E}}_{\mathcal{W}}(\phi), \tag{95}$$

with,

$$\tilde{\mathsf{E}}_{\mathcal{W}}(\phi) = \int_0^{\ell}\big\{\mathsf{I}(\mathcal{W} * \phi^{-1}(y)) - \mathsf{I}(\phi^{-1}(y-1))\big\}\mathrm{d}y. \tag{96}$$

Lemma 5.19. *Let $\delta > 0$, and p_X be a probability measure on the real line with $\delta > \bar{d}(p_X)$. For any $\kappa > 0$, there exist ℓ_0, σ_0^2, such that, for any $\ell > \ell_0$ and $\sigma \in (0, \sigma_0]$, and any fixed point of continuum state evolution,$\{\phi, \psi\}$, with ψ and ϕ nondecreasing Lipschitz functions and $\psi(x) \ge \mathsf{mmse}(L_0/(2\sigma^2))$, the following holds.*

$$\int_{-1}^{\ell-1} |\phi(x) - \phi^*|\,\mathrm{d}x \le \kappa\ell. \tag{97}$$

Proof. The claim is trivial for $\kappa \ge \Phi_M$, since $\phi(x) \le \Phi_M$. Fix $\kappa < \Phi_M$, and choose σ_1, such that $\phi^* < \kappa/2$, for $\sigma \in (0, \sigma_1]$. Since ϕ is a fixed point of continuum state evolution, we have $\nabla\mathsf{E}_{\mathcal{W}}(\phi) = 0$, on the interval $[-1, \ell - 1]$ by Corollary 5.16. Now, assume that $\int_{-1}^{\ell-1}|\phi(x) - \phi^*| > \kappa\ell$. We introduce an infinitesimal perturbation of ϕ that decreases the energy in the first order; this contradicts the fact $\nabla\mathsf{E}_{\mathcal{W}}(\phi) = 0$ on the interval $[-1, \ell - 1]$.

Claim 5.20. *For each fixed point of continuum state evolution that satisfies the hypothesis of Lemma 5.19, the following holds. For any $K > 0$, there exists ℓ_0, such that, for $\ell > \ell_0$ there exist $x_1 < x_2 \in [0, \ell - 1)$, with $x_2 - x_1 = K$ and $\kappa/2 + \phi^* < \phi(x)$, for $x \in [x_1, x_2]$.*

Claim 5.20 is proved in Appendix C.

Fix $K > 2$ and let $x_0 = (x_1 + x_2)/2$. Thus, $x_0 \ge 1$. For $a \in (0, 1]$, define

$$\phi_a(x) = \begin{cases} \phi(x), & \text{for } x_2 \le x, \\ \phi(\frac{x_2 - x_0}{x_2 - x_0 - a}\,x - \frac{a x_2}{x_2 - x_0 - a}), & \text{for } x \in [x_0 + a, x_2), \\ \phi(x - a), & \text{for } x \in [-1 + a, x_0 + a), \\ \phi^*, & \text{for } x \in [-1, -1 + a). \end{cases} \tag{98}$$

See Fig. 3 for an illustration. (Note that from Eq. (80), $\phi(-1) = \phi^*$). In the following, we bound the difference of the free energies of functions ϕ and ϕ_a.

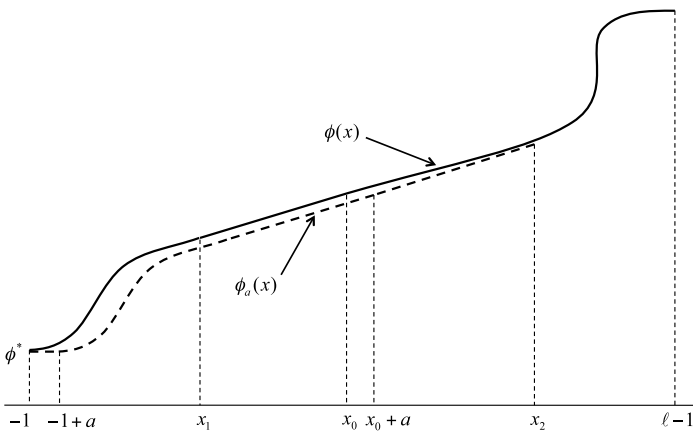

Figure 3: An illustration of function $\phi(x)$ and its perturbation $\phi_a(x)$.

Proposition 5.21. *For each fixed point of continuum state evolution, satisfying the hypothesis of Lemma 5.19, there exists a constant $C(K)$, such that*

$$\int_{-1}^{\ell-1} \Big\{ \frac{\sigma^2(x)-\sigma^2}{\phi_a(x)} - \frac{\sigma^2(x)-\sigma^2}{\phi(x)} \Big\} \,\mathrm{d}x \le C(K)a.$$

We refer to Appendix D for the proof of Proposition 5.21.

Proposition 5.22. *For each fixed point of continuum state evolution, satisfying the hypothesis of Lemma 5.19, there exists a constant $C(\kappa, K)$, such that,*

$$\tilde{\mathsf{E}}_{\mathcal{W}}(\phi_a) - \tilde{\mathsf{E}}_{\mathcal{W}}(\phi) \le C(\kappa, K)a.$$

Proof of Proposition 5.22 is deferred to Appendix E.

Using Eq. (95) and Proposition 5.22, we have

$$\mathsf{E}_{\mathcal{W}}(\phi_a) - \mathsf{E}_{\mathcal{W}}(\phi) \le \int_{-1}^{\ell-1} \big\{ V(\phi_a(x)) - V(\phi(x)) \big\} \mathrm{d}x + C(\kappa, K)a, \tag{99}$$

where the constants $(\delta/2)C(K)$ and $C(\kappa, K)$ are absorbed in $C(\kappa, K)$.

In addition,

$$\begin{aligned} \int_{-1}^{\ell-1} \big\{ V(\phi_a(x)) - V(\phi(x)) \big\} \mathrm{d}x = & \int_{x_2}^{\ell-1} \big\{ V(\phi_a(x)) - V(\phi(x)) \big\} \mathrm{d}x \\ & + \Big(\int_{x_0+a}^{x_2} V(\phi_a(x)) \mathrm{d}x - \int_{x_0}^{x_2} V(\phi(x)) \mathrm{d}x \Big) \\ & + \Big(\int_{-1+a}^{x_0+a} V(\phi_a(x)) \mathrm{d}x - \int_{-1}^{x_0} V(\phi(x)) \mathrm{d}x \Big) \\ & + \int_{-1}^{-1+a} V(\phi_a(x)) \mathrm{d}x. \end{aligned} \tag{100}$$

Notice that the first and the third terms on the right hand side are zero. Also,

$$\begin{aligned}\int_{x_0+a}^{x_2} V(\phi_a(x))\mathrm{d}x - \int_{x_0}^{x_2} V(\phi(x))\mathrm{d}x = -\frac{a}{x_2-x_0}\int_{x_0}^{x_2} V(\phi(x))\mathrm{d}x,\\ \int_{-1}^{-1+a} V(\phi_a(x))\mathrm{d}x = aV(\phi^*).\end{aligned} \tag{101}$$

Substituting Eq. (101) in Eq. (100), we get

$$\int_{-1}^{\ell-1} \big\{V(\phi_a(x)) - V(\phi(x))\big\}\mathrm{d}x = \frac{a}{x_2-x_0}\int_{x_0}^{x_2} \big\{V(\phi^*) - V(\phi(x))\big\}\mathrm{d}x. \tag{102}$$

We proceed by proving the following claim.

Claim 5.23. *For any $C = C(\kappa, K) \geq 0$, there exists σ_3, such that for $\sigma \in (0, \sigma_3]$, the following holds.*

$$\int_{-1}^{\ell-1} \big\{V(\phi_a(x)) - V(\phi(x))\big\}\mathrm{d}x < -2C(\kappa, K)a. \tag{103}$$

Proof. By Proposition 5.18, we have

$$V(\phi^*) \leq \frac{\delta}{2} + \frac{\delta - \overline{d}(p_X)}{4}\log(2\sigma^2), \tag{104}$$

for $\sigma \in (0, \sigma_2]$. Also, since $\phi(x) > \kappa/2$ for $x \in [x_0, x_2]$, we have $V(\phi(x)) \geq (\delta/2)\log\phi > (\delta/2)\log(\kappa/2)$. Therefore,

$$\begin{aligned}\frac{1}{2}\int_{-1}^{\ell-1} \big\{V(\phi_a(x)) - V(\phi(x))\big\}\mathrm{d}x &= \frac{a}{2(x_2-x_0)}\int_{x_0}^{x_2} \big\{V(\phi^*) - V(\phi(x))\big\}\mathrm{d}x\\ &< \frac{a}{2}\Big[\frac{\delta}{2} + \frac{\delta - \overline{d}(p_X)}{4}\log(2\sigma^2) - \frac{\delta}{2}\log(\frac{\kappa}{2})\Big].\end{aligned} \tag{105}$$

It is now obvious that by choosing $\sigma_3 > 0$ small enough, we can ensure that for values $\sigma \in (0, \sigma_3]$,

$$\frac{a}{2}\Big[\frac{\delta}{2} + \frac{\delta - \overline{d}(p_X)}{4}\log(2\sigma^2) - \frac{\delta}{2}\log(\frac{\kappa}{2})\Big] < -2C(\kappa, K)a. \tag{106}$$

(Notice that the right hand side of Eq. (106) does not depend on σ). □

Let $\sigma_0 = \min\{\sigma_1, \sigma_2, \sigma_3\}$. As a result of Eq. (99) and Claim 5.23,

$$\begin{aligned}\mathsf{E}_{\mathcal{W}}(\phi_a) - \mathsf{E}_{\mathcal{W}}(\phi) &< \int_{-1}^{\ell-1} \big\{V(\phi_a(x)) - V(\phi(x))\big\}\mathrm{d}x + C(\kappa, K)a\\ &\leq -C(\kappa, K)a\,.\end{aligned} \tag{107}$$

Since ϕ is a Lipschitz function by assumption, it is easy to see that $\|\phi_a - \phi\|_2 \leq C\,a$, for some constant C. By Taylor expansion of the free energy functional around function ϕ, we have

$$\begin{aligned}\langle \nabla\mathsf{E}_{\mathcal{W}}(\phi), \phi_a - \phi\rangle &= \mathsf{E}_{\mathcal{W}}(\phi_a) - \mathsf{E}_{\mathcal{W}}(\phi) + o(\|\phi_a - \phi\|_2)\\ &\leq -C(\kappa, K)a + o(a).\end{aligned} \tag{108}$$

However, since $\{\phi, \psi\}$ is a fixed point of the continuum state evolution, we have $\nabla \mathsf{E}_{\mathcal{W}}(\phi) = 0$ on the interval $[-1, \ell - 1]$ (cf. Corollary 5.16). Also, $\phi_a - \phi$ is zero out of $[-1, \ell - 1]$. Therefore, $\langle \nabla \mathsf{E}_{\mathcal{W}}(\phi), \phi_a - \phi \rangle = 0$, which leads to a contradiction in Eq (108). This implies that our first assumption $\int_{-1}^{\ell-1} |\phi(x) - \phi^*| \, \mathrm{d}x > \kappa \ell$ is false. The result follows. □

Next lemma pertains to the robust reconstruction of the signal. Prior to stating the lemma, we need to establish some definitions. Due to technical reasons in the proof, we consider an alternative decomposition of $\mathsf{E}_{\mathcal{W}}(\phi)$ to Eq. (95).

Define the potential function $V_{\mathrm{rob}} : \mathbb{R}_+ \to \mathbb{R}_+$ as follows.

$$V_{\mathrm{rob}}(\phi) = \frac{\delta}{2}\Big(\frac{\sigma^2}{\phi} + \log \phi\Big), \tag{109}$$

and decompose the Energy functional as:

$$\mathsf{E}_{\mathcal{W}}(\phi) = \int_{-1}^{\ell-1} V_{\mathrm{rob}}(\phi(x)) \, \mathrm{d}x + \frac{\delta}{2} \int_{-1}^{\ell-1} \frac{\sigma^2(x) - \sigma^2}{\phi(x)} \, \mathrm{d}x + \tilde{\mathsf{E}}_{\mathcal{W},\mathrm{rob}}(\phi), \tag{110}$$

with,

$$\tilde{\mathsf{E}}_{\mathcal{W},\mathrm{rob}}(\phi) = \int_0^{\ell} \mathsf{I}(\mathcal{W} * \phi^{-1}(y)) \mathrm{d}y. \tag{111}$$

Lemma 5.24. *Let $\delta > 0$, and p_X be a probability measure on the real line with $\delta > \overline{D}(p_X)$. There exist ℓ_0, σ_0^2, and C, such that , for any $\ell > \ell_0$ and $\sigma \in (0, \sigma_0]$, and for any fixed point of continuum state evolution, $\{\phi, \psi\}$, with ψ and ϕ nondecreasing Lipschitz functions and $\psi(x) \geq \mathsf{mmse}(L_0/(2\sigma^2))$, the following holds.*

$$\int_{-1}^{\ell-1} |\phi(x) - \phi^*| \, \mathrm{d}x \leq C\sigma^2 \ell. \tag{112}$$

Proof. Suppose $\int_{-1}^{\ell-1} |\phi(x) - \phi^*| \mathrm{d}x > C\sigma^2\ell$, for any constant C. Similar to the proof of Lemma 5.19, we obtain an infinitesimal perturbation of ϕ which decreases the free energy in the first order, contradicting the fact $\nabla \mathsf{E}_{\mathcal{W}}(\phi) = 0$ on the interval $[-1, \ell - 1]$.

By definition of upper MMSE dimension (Eq. (15)), for any $\varepsilon > 0$, there exists ϕ_1, such that, for $\phi \in [0, \phi_1]$,

$$\mathsf{mmse}(\phi^{-1}) \leq (\overline{D}(p_X) + \varepsilon)\phi. \tag{113}$$

Claim 5.25. *For each fixed point of continuum state evolution that satisfies the hypothesis of Lemma 5.24, the following holds. For any $K > 0$, there exists ℓ_0, such that, for $\ell > \ell_0$ there exist $x_1 < x_2 \in [0, \ell - 1)$, with $x_2 - x_1 = K$ and $C\sigma^2/2 \leq \phi(x) \leq \phi_1$, for $x \in [x_1, x_2]$.*

Claim 5.25 is proved in Appendix F. For positive values of a, define

$$\phi_a(x) = \begin{cases} \phi(x), & \text{for } x \leq x_1, x_2 \leq x, \\ (1-a)\phi(x) & \text{for } x \in (x_1, x_2). \end{cases} \tag{114}$$

Our aim is to show that $\mathsf{E}_{\mathcal{W}}(\phi_a) - \mathsf{E}_{\mathcal{W}}(\phi) \leq -c\, a$, for some constant $c > 0$.

Invoking Eq. (95), we have

$$\begin{aligned}\mathsf{E}_{\mathcal{W}}(\phi_a) - \mathsf{E}_{\mathcal{W}}(\phi) = &\int_{-1}^{\ell-1} \{V_{\rm rob}\phi_a(x)) - V_{\rm rob}(\phi(x))\}\, \mathrm{d}x \\ &+ \frac{\delta}{2}\int_{-1}^{\ell-1} (\sigma^2(x) - \sigma^2)\left(\frac{1}{\phi_a(x)} - \frac{1}{\phi(x)}\right)\, \mathrm{d}x + \tilde{\mathsf{E}}_{\mathcal{W},\rm rob}(\phi_a) - \tilde{\mathsf{E}}_{\mathcal{W},\rm rob}(\phi).\end{aligned} \tag{115}$$

The following proposition bounds each term on the right hand side separately.

Proposition 5.26. *For the function $\phi(x)$ and its perturbation $\phi_a(x)$, we have*

$$\int_{-1}^{\ell-1} \{V_{\rm rob}(\phi_a(x)) - V_{\rm rob}(\phi(x))\}\, \mathrm{d}x \le \frac{\delta}{2} K\log(1-a) + K\frac{\delta a}{C(1-a)}, \tag{116}$$

$$\int_{-1}^{\ell-1} (\sigma^2(x) - \sigma^2)\left(\frac{1}{\phi_a(x)} - \frac{1}{\phi(x)}\right)\, \mathrm{d}x \le K\frac{2a}{C(1-a)}, \tag{117}$$

$$\tilde{\mathsf{E}}_{\mathcal{W},\rm rob}(\phi_a) - \tilde{\mathsf{E}}_{\mathcal{W},\rm rob}(\phi) \le -\frac{\overline{D}(p_X)+\varepsilon}{2}(K+2)\log(1-a). \tag{118}$$

We refer to Appendix G for the proof of Proposition 5.26.

Combining the bounds given by Proposition 5.26, we obtain

$$\mathsf{E}_{\mathcal{W}}(\phi_a) - \mathsf{E}_{\mathcal{W}}(\phi) \le \frac{K}{2}\log(1-a)\Big\{\delta - (\overline{D}(p_X)+\varepsilon)(1+\frac{2}{K})\Big\} + K\frac{2\delta a}{C(1-a)}. \tag{119}$$

Since $\delta > \overline{D}(p_X)$ by our assumption, there exist ε, K, C such that

$$c = \delta - (\overline{D}(p_X)+\varepsilon)(1+\frac{2}{K}) - \frac{4\delta}{C(1-a)} > 0.$$

Using Eq. (119), we get

$$\mathsf{E}_{\mathcal{W}}(\phi_a) - \mathsf{E}_{\mathcal{W}}(\phi) \le -\frac{cK}{2}a. \tag{120}$$

By an argument analogous to the one in the proof of Lemma 5.19, this is in contradiction with $\nabla\mathsf{E}_{\mathcal{W}}(\phi) = 0$. The result follows. □

5.2 Proof of Lemma 3.2

By Lemma 5.7, $\phi_a(t) \le \phi_a^{mod}(t-t_0)$, for $a \in \mathsf{R}_0 \cong \{\rho^{-1}, \cdots, L-1+\rho^{-1}\}$ and $t \ge t_1(L_0,\delta)$. Therefore, we only need to prove the claim for the modified state evolution. The idea of the proof is as follows. In the previous section, we analyzed the continuum state evolution and showed that at the fixed point, the function $\phi(x)$ is close to the constant ϕ^*. Also, in Lemma 5.11, we proved that the modified state evolution is essentially approximated by the continuum state evolution as $\rho \to 0$. Combining these results implies the thesis.

Proof (Part(a)). By monotonicity of continuum state evolution (cf. Corollary 5.12), $\lim_{t\to\infty}\phi(x;t) = \phi(x)$ exists. Further, by continuity of state evolution recursions, $\phi(x)$ is a fixed point. Finally, $\phi(x)$

is a nondecreasing Lipschitz function (cf. Corollary 5.13). Using Lemma 5.19 in conjunction with the Dominated Convergence theorem, we have, for any $\varepsilon > 0$

$$\lim_{t\to\infty} \frac{1}{\ell} \int_{-1}^{\ell-1} |\phi(x;t) - \phi^*| \mathrm{d}x \leq \frac{\varepsilon}{4}, \tag{121}$$

for $\sigma \in (0, \sigma_0^2]$ and $\ell > \ell_0$. Therefore, there exists $t_2 > 0$ such that $\frac{1}{\ell}\int_{-1}^{\ell-1} |\phi(x;t_2) - \phi^*| \mathrm{d}x \leq \varepsilon/2$. Moreover, for any $t \geq 0$,

$$\frac{1}{\ell} \int_{-1}^{\ell-1} |\phi(x;t) - \phi^*| \mathrm{d}x = \lim_{\rho\to 0} \frac{\rho}{\ell} \sum_{a=-\rho^{-1}}^{L-\rho^{-1}-1} |\phi(\rho a;t) - \phi^*| = \lim_{\rho\to 0} \frac{1}{L} \sum_{a=-\rho^{-1}}^{L-\rho^{-1}-1} |\phi(\rho a;t) - \phi^*|. \tag{122}$$

By triangle inequality, for any $t \geq 0$,

$$\begin{aligned} \lim_{\rho\to 0} \frac{1}{L} \sum_{a=-\rho^{-1}}^{L-\rho^{-1}-1} |\phi_a(t) - \phi^*| &\leq \lim_{\rho\to 0} \frac{1}{L} \sum_{a=-\rho^{-1}}^{L-\rho^{-1}-1} |\phi_a(t) - \phi(\rho a;t)| + \lim_{\rho\to 0} \frac{1}{L} \sum_{a=-\rho^{-1}}^{L-\rho^{-1}-1} |\phi(\rho a;t) - \phi^*| \\ &= \frac{1}{\ell} \int_{-1}^{\ell-1} |\phi(x;t) - \phi^*| \mathrm{d}x, \end{aligned} \tag{123}$$

where the last step follows from Lemma 5.11 and Eq. (122). Since the sequence $\{\phi(t)\}$ is monotone decreasing in t, we have

$$\begin{aligned} \lim_{\rho\to 0} \lim_{t\to\infty} \frac{1}{L} \sum_{a=-\rho^{-1}}^{L-\rho^{-1}-1} \phi_a(t) &\leq \lim_{\rho\to 0} \frac{1}{L} \sum_{a=-\rho^{-1}}^{L-\rho^{-1}-1} \phi_a(t_2) \\ &\leq \lim_{\rho\to 0} \frac{1}{L} \sum_{a=-\rho^{-1}}^{L-\rho^{-1}-1} (|\phi_a(t_2) - \phi^*| + \phi^*) \\ &\leq \frac{1}{\ell} \int_{-1}^{\ell-1} |\phi(x;t_2) - \phi^*| \mathrm{d}x + \phi^* \\ &\leq \frac{\varepsilon}{2} + \phi^*. \end{aligned} \tag{124}$$

Finally,

$$\begin{aligned} \lim_{t\to\infty} \sum_{a=-\rho^{-1}}^{L+\rho^{-1}-1} \phi_a(t) &\leq \frac{2\rho^{-1}}{L} \Phi_M + \frac{\varepsilon}{2} + \phi^* \\ &\leq \frac{2\rho^{-1}}{L_*} \Phi_M + \frac{\varepsilon}{2} + 2\sigma_0. \end{aligned} \tag{125}$$

Clearly, by choosing L_* large enough and σ_0 sufficiently small, we can ensure that the right hand side of Eq. (125) is less than ε. □

Proof (Part(b)). Consider the following two cases.

- $\sigma \leq \sigma_0$: In this case, proceeding along the same lines as the proof of Part (a), and using Lemma 5.24 in lieu of Lemma 5.19, we have

$$\lim_{t\to\infty} \frac{1}{L} \sum_{a=-\rho^{-1}}^{L-\rho^{-1}-1} \phi_a(t) \leq C_1 \sigma^2, \tag{126}$$

  for some constant C_1.

- $\sigma > \sigma_0$: Since $\phi_a(t) \leq \Phi_M$ for any $t > 0$, we have

$$\lim_{t\to\infty} \frac{1}{L} \sum_{a=-\rho^{-1}}^{L-\rho^{-1}-1} \phi_a(t) \leq \Phi_M. \tag{127}$$

Choosing $C = \max\{C_1, \Phi_M/\sigma_0^2\}$ proves the claim in both cases. □

Acknowledgements

A.M. would like to thank Florent Krzakala, Marc Mézard, François Sausset, Yifan Sun and Lenka Zdeborová for a stimulating exchange about their results. A.J. is supported by a Caroline and Fabian Pease Stanford Graduate Fellowship. This work was partially supported by the NSF CAREER award CCF- 0743978, the NSF grant DMS-0806211, and the AFOSR grant FA9550-10-1-0360.

A Proof of Lemma 5.11

We prove the first claim, Eq. (82). The second one follows by a similar argument. The proof uses induction on t. It is a simple exercise to show that the induction basis ($t = 1$) holds (the calculation follows the same lines as the induction step). Assuming the claim for t, we write, for $i \in \{0, 1, \dots, L-1\}$

$$
\begin{aligned}
|\psi_i(t+1) - \psi(\rho i; t+1)| &= \Big|\mathsf{mmse}\Big(\sum_{b\in\mathsf{R}_0} W_{b-i}\,[\sigma^2 + \frac{1}{\delta}\sum_{j\in\mathbb{Z}} W_{b-j}\psi_j(t)]^{-1}\Big) \\
&\quad - \mathsf{mmse}\Big(\int_{-1}^{\ell+1} \mathcal{W}(z-\rho i)\,[\sigma^2 + \frac{1}{\delta}\int_{\mathbb{R}} \mathcal{W}(z-y)\psi(y;t)\mathrm{d}y]^{-1}\mathrm{d}z\Big)\Big| \\
&\le \Big|\mathsf{mmse}\Big(\sum_{b\in\mathsf{R}_0} W_{b-i}\,[\sigma^2 + \frac{1}{\delta}\sum_{j\in\mathbb{Z}} W_{b-j}\psi_j(t)]^{-1}\Big) \\
&\quad - \mathsf{mmse}\Big(\sum_{b\in\mathsf{R}_0} W_{b-i}\,[\sigma^2 + \frac{1}{\delta}\sum_{j\in\mathbb{Z}} W_{b-j}\psi(\rho j;t)]^{-1}\Big)\Big| \\
&\quad + \Big|\mathsf{mmse}\Big(\sum_{b\in\mathsf{R}_0} \rho\mathcal{W}(\rho(b-i))\,[\sigma^2 + \frac{1}{\delta}\sum_{j\in\mathbb{Z}} \rho\mathcal{W}(\rho(b-j))\psi(\rho j;t)]^{-1}\Big) \\
&\quad - \mathsf{mmse}\Big(\int_{-1}^{\ell+1} \mathcal{W}(z-\rho i)\,[\sigma^2 + \frac{1}{\delta}\int_{\mathbb{R}} \mathcal{W}(z-y)\psi(y;t)\mathrm{d}y]^{-1}\mathrm{d}z\Big)\Big|.
\end{aligned}
\tag{128}
$$

Now, we bound the two terms on the right hand side separately. Note that the arguments of $\mathsf{mmse}(\,\cdot\,)$ in the above terms are at most $2/\sigma^2$. Since mmse has a continuous derivative, there exists a constant C such that $|d/ds\ \mathsf{mmse}(s)| \le C$, for $s \in [0, 2/\sigma^2]$. Then, considering the first term in the upper bound (128), we have

$$
\begin{aligned}
&\Big|\mathsf{mmse}\Big(\sum_{b\in\mathsf{R}_0} W_{b-i}\,[\sigma^2 + \frac{1}{\delta}\sum_{j\in\mathbb{Z}} W_{b-j}\psi_j(t)]^{-1}\Big) - \mathsf{mmse}\Big(\sum_{b\in\mathsf{R}_0} W_{b-i}\,[\sigma^2 + \frac{1}{\delta}\sum_{j\in\mathbb{Z}} W_{b-j}\psi(\rho j;t)]^{-1}\Big)\Big| \\
&\qquad \le C\Big|\sum_{b\in\mathsf{R}_0} W_{b-i}\,\Big([\sigma^2 + \frac{1}{\delta}\sum_{j\in\mathbb{Z}} W_{b-j}\psi_j(t)]^{-1} - [\sigma^2 + \frac{1}{\delta}\sum_{j\in\mathbb{Z}} W_{b-j}\psi(\rho j;t)]^{-1}\Big)\Big| \\
&\qquad \le \frac{C}{\sigma^4}\sum_{b\in\mathsf{R}_0} W_{b-i}\,\frac{1}{\delta}\Big|\sum_{j=-\infty}^{L-1} W_{b-j}(\psi(\rho j;t) - \psi_j(t))\Big| \\
&\qquad \le \frac{C}{\delta\sigma^4}\sum_{b\in\mathsf{R}_0} W_{b-i}\sum_{j=-\infty}^{L-1} W_{b-j}|\psi(\rho j;t) - \psi_j(t)| \\
&\qquad = \frac{C}{\delta\sigma^4}\sum_{j=0}^{L-1}\Big(\sum_{b\in\mathsf{R}_0} W_{b-i}W_{b-j}\Big)\,|\psi(\rho j;t) - \psi_j(t)| \\
&\qquad \le \frac{C}{\delta\sigma^4}\Big(\sum_{i\in\mathbb{Z}} W_i^2\Big)\sum_{j=0}^{L-1}|\psi(\rho j;t) - \psi_j(t)| \\
&\qquad \le \frac{C'\rho}{\delta\sigma^4}\sum_{j=0}^{L-1}|\psi(\rho j;t) - \psi_j(t)|.
\end{aligned}
\tag{129}
$$

Here we used $\sum_{i\in\mathbb{Z}} W_i^2 = \sum_{i\in\mathbb{Z}} \rho^2 \mathcal{W}(\rho i)^2 \le C \sum_{|i|\le \rho^{-1}} \rho^2 \le C\rho$ (where the first inequality follows from the fact that $\mathcal{W}$ is bounded).

To bound the second term in Eq. (128), note that

$$
\begin{aligned}
&\Big|\mathsf{mmse}\Big(\sum_{b\in\mathsf{R}_0} \rho\mathcal{W}(\rho(b-i))\,[\sigma^2 + \frac{1}{\delta}\sum_{j\in\mathbb{Z}} \rho\mathcal{W}(\rho(b-j))\psi(\rho j;t)]^{-1}\Big) \\
&\qquad\qquad - \mathsf{mmse}\Big(\int_{-1}^{\ell+1} \mathcal{W}(z-\rho i)\,[\sigma^2 + \frac{1}{\delta}\int_{\mathbb{R}} \mathcal{W}(z-y)\psi(y;t)\mathrm{d}y]^{-1}\mathrm{d}z\Big)\Big| \\
&\le C\Big|\sum_{b\in\mathsf{R}_0} \rho\mathcal{W}(\rho(b-i))\,[\sigma^2 + \frac{1}{\delta}\sum_{j\in\mathbb{Z}} \rho\mathcal{W}(\rho(b-j))\psi(\rho j;t)]^{-1} \\
&\qquad\qquad - \int_{-1}^{\ell+1} \mathcal{W}(z-\rho i)\,[\sigma^2 + \frac{1}{\delta}\int_{\mathbb{R}} \mathcal{W}(z-y)\psi(y;t)\mathrm{d}y]^{-1}\mathrm{d}z\Big| \\
&\le C\Big|\sum_{b\in\mathsf{R}_0} \rho\mathcal{W}(\rho(b-i))\,[\sigma^2 + \frac{1}{\delta}\sum_{j\in\mathbb{Z}} \rho\mathcal{W}(\rho(b-j))\psi(\rho j;t)]^{-1} \\
&\qquad\qquad - \sum_{b\in\mathsf{R}_0} \rho\mathcal{W}(\rho(b-i))\,[\sigma^2 + \frac{1}{\delta}\int_{\mathbb{R}} \mathcal{W}(\rho b-y)\psi(y;t)\mathrm{d}y]^{-1}\Big| \\
&+ C\Big|\sum_{b\in\mathsf{R}_0} \rho\mathcal{W}(\rho(b-i))\,[\sigma^2 + \frac{1}{\delta}\int_{\mathbb{R}} \mathcal{W}(\rho b-y)\psi(y;t)\mathrm{d}y]^{-1}\mathrm{d}z \\
&\qquad\qquad - \int_{-1}^{\ell+1} \mathcal{W}(z-\rho i)\,[\sigma^2 + \frac{1}{\delta}\int_{\mathbb{R}} \mathcal{W}(z-y)\psi(y;t)\mathrm{d}y]^{-1}\mathrm{d}z\Big| \\
&\le \frac{C}{\delta\sigma^4}\sum_{b\in\mathsf{R}_0} \rho\mathcal{W}(\rho(b-i))\Big|\sum_{j\in\mathbb{Z}} \rho F_1(\rho b;\rho j) - \int_{\mathbb{R}} F_1(\rho b;y)\mathrm{d}y\Big| \\
&+ C\Big|\sum_{b\in\mathsf{R}_0} \rho F_2(\rho b) - \int_{-1}^{\ell+1} F_2(z)\mathrm{d}z\Big|
\end{aligned}
\tag{130}
$$

where $F_1(x;y) = \mathcal{W}(x-y)\psi(y;t)$ and $F_2(z) = \mathcal{W}(z-\rho i)\,[\sigma^2 + \frac{1}{\delta}\int_{\mathbb{R}} \mathcal{W}(z-y)\psi(y;t)\mathrm{d}y]^{-1}$. Since the functions $\mathcal{W}(\,\cdot\,)$ and $\psi(\,\cdot\,)$ have continuos (and thus bounded) derivative on compact interval $[0,\ell]$, the same is true for F_1 and F_2. Using the standard convergence of Riemann sums to Riemann integrals, right hand side of Eq. (130) can be bounded by $C_3\rho/\delta\sigma^4$, for some constant C_3. Let $\epsilon_i(t) = |\psi_i(t) - \psi(\rho i;t)|$. Combining Eqs. (129) and (130), we get

$$
\epsilon_i(t+1) \le \frac{\rho}{\delta\sigma^4}\left(C'\sum_{j=0}^{L-1}\epsilon_j(t) + C_3\right). \tag{131}
$$

Therefore,

$$
\frac{1}{L}\sum_{i=0}^{L-1}\epsilon_i(t+1) \le \frac{\ell}{\delta\sigma^4}\left(\frac{C'}{L}\sum_{j=0}^{L-1}\epsilon_j(t)\right) + \frac{C_3\rho}{\delta\sigma^4}. \tag{132}
$$

The claims follows from the induction hypothesis.

B Proof of Proposition 5.18

By Eq. (86), for any $\varepsilon > 0$, there exists ϕ_0, such that for $0 \le \phi \le \phi_0$,

$$\mathsf{I}(\phi^{-1}) \le \frac{\overline{d}(p_X) + \varepsilon}{2} \log(\phi^{-1}). \tag{133}$$

Therefore,

$$V(\phi) \le \frac{\delta\sigma^2}{2\phi} + \frac{\delta - \overline{d}(p_X) - \varepsilon}{2} \log \phi. \tag{134}$$

Now let $\varepsilon = (\delta - \overline{d}(p_X))/2$ and $\sigma_2 = \sqrt{\phi_0/2}$. Hence, for $\sigma \in (0, \sigma_2]$, we get $\phi^* < 2\sigma^2 \le \phi_0$. Plugging in ϕ^* for ϕ in the above equation, we get

$$\begin{aligned} V(\phi^*) &\le \frac{\delta\sigma^2}{2\phi^*} + \frac{\delta - \overline{d}(p_X)}{4} \log \phi^* \\ &< \frac{\delta}{2} + \frac{\delta - \overline{d}(p_X)}{4} \log(2\sigma^2)\,. \end{aligned} \tag{135}$$

C Proof of Claim 5.20

Recall that $\kappa < \Phi_M$ and $\phi(x)$ is nondecreasing. Let

$$0 < \theta = \frac{\Phi_M - \kappa}{\Phi_M - \frac{\kappa}{2}} < 1.$$

We show that $\phi(\theta\ell - 1) \ge \kappa/2 + \phi^*$. If this is not true, using the nondecreasing property of $\phi(x)$, we obtain

$$\begin{aligned} \int_{-1}^{\ell-1} |\phi(x) - \phi^*| \, \mathrm{d}x &= \int_{-1}^{\theta\ell-1} |\phi(x) - \phi^*| \, \mathrm{d}x + \int_{\theta\ell-1}^{\ell-1} |\phi(x) - \phi^*| \, \mathrm{d}x \\ &< \frac{\kappa}{2}\theta\ell + \Phi_M(1-\theta)\ell \\ &= \kappa\ell, \end{aligned} \tag{136}$$

contradicting our assumption. Therefore, $\phi(x) \ge \kappa/2 + \phi^*$, for $\theta\ell - 1 \le x \le \ell - 1$. For given K, choose $\ell_0 = K/(1-\theta)$. Hence, for $\ell > \ell_0$, interval $[\theta\ell - 1, \ell - 1)$ has length at least K. The result follows.

D Proof of Proposition 5.21

We first establish some properties of function $\sigma^2(x)$.

Remark D.1. *The function $\sigma^2(x)$ as defined in Eq. (88), is non increasing in x. Also, $\sigma^2(x) = \sigma^2 + (1/\delta)\ \mathsf{mmse}(L_0/(2\sigma^2))$, for $x \le -1$ and $\sigma^2(x) = \sigma^2$, for $x \ge 1$. For $\delta L_0 > 3$, we have $\sigma^2 \le \sigma^2(x) < 2\sigma^2$.*

Remark D.2. *The function $\sigma^2(x)/\sigma^2$ is Lipschitz continuous. More specifically, there exists a constant C, such that, $|\sigma^2(\alpha_1) - \sigma^2(\alpha_2)| < C\sigma^2|\alpha_2 - \alpha_1|$, for any two values α_1, α_2. Further, if $L_0\delta > 3$ we can take $C < 1$.*

The proof of Remarks D.1 and D.2 are immediate from Eq. (88).

To prove the proposition, we split the integral over the intervals $[-1, -1+a), [-1+a, x_0+a), [x_0+a, x_2), [x_2, \ell-1)$, and bound each one separately. Firstly, note that

$$\int_{x_2}^{\ell-1} \left\{ \frac{\sigma^2(x) - \sigma^2}{\phi_a(x)} - \frac{\sigma^2(x) - \sigma^2}{\phi(x)} \right\} \mathrm{d}x = 0, \tag{137}$$

since $\phi_a(x)$ and $\phi(x)$ are identical for $x \geq x_2$.

Secondly, let $\alpha = (x_2 - x_0)/(x_2 - x_0 - a)$, and $\beta = (ax_2)/(x_2 - x_0 - a)$. Then,

$$\begin{aligned}
&\int_{x_0+a}^{x_2} \left\{ \frac{\sigma^2(x) - \sigma^2}{\phi_a(x)} - \frac{\sigma^2(x) - \sigma^2}{\phi(x)} \right\} \mathrm{d}x \\
&= \int_{x_0}^{x_2} \frac{\sigma^2(\frac{x+\beta}{\alpha}) - \sigma^2}{\phi(x)} \frac{\mathrm{d}x}{\alpha} - \int_{x_0+a}^{x_2} \frac{\sigma^2(x) - \sigma^2}{\phi(x)} \mathrm{d}x \\
&= \int_{x_0}^{x_2} \left\{ \frac{1}{\alpha} \frac{\sigma^2(\frac{x+\beta}{\alpha}) - \sigma^2}{\phi(x)} - \frac{\sigma^2(x) - \sigma^2}{\phi(x)} \right\} \mathrm{d}x + \int_{x_0}^{x_0+a} \frac{\sigma^2(x) - \sigma^2}{\phi(x)} \, \mathrm{d}x \\
&\overset{(a)}{\leq} \frac{1}{\sigma^2} \int_{x_0}^{x_2} \left| \frac{1}{\alpha} \sigma^2\big(\frac{x+\beta}{\alpha}\big) - \sigma^2(x) \right| \mathrm{d}x + \left(1 - \frac{1}{\alpha}\right) \int_{x_0}^{x_2} \frac{\sigma^2}{\phi(x)} \, \mathrm{d}x + \int_{x_0}^{x_0+a} \frac{\sigma^2}{\phi(x)} \, \mathrm{d}x \\
&\leq \frac{1}{\sigma^2} \int_{x_0}^{x_2} \left(1 - \frac{1}{\alpha}\right) \sigma^2\big(\frac{x+\beta}{\alpha}\big) \, \mathrm{d}x + \frac{1}{\sigma^2} \int_{x_0}^{x_2} \left| \sigma^2\big(\frac{x+\beta}{\alpha}\big) - \sigma^2(x) \right| \mathrm{d}x + \frac{K}{2} \left(1 - \frac{1}{\alpha}\right) + a \\
&\leq \left(1 - \frac{1}{\alpha}\right) K + \frac{1}{\sigma^2} \int_{x_0}^{x_2} \left| \sigma^2\big(\frac{x+\beta}{\alpha}\big) - \sigma^2(x) \right| \mathrm{d}x + \frac{K}{2} \left(1 - \frac{1}{\alpha}\right) + a \\
&\overset{(b)}{\leq} \left(1 - \frac{1}{\alpha}\right) K + CK^2 \left(1 - \frac{1}{\alpha}\right) + CK\, a + \frac{K}{2} \left(1 - \frac{1}{\alpha}\right) + a \\
&\leq C(K)a,
\end{aligned} \tag{138}$$

where (a) follows from the fact $\sigma^2 \leq \phi(x)$ and Remark D.1; (b) follows from Remark D.2.

Thirdly, recall that $\phi_a(x) = \phi(x - a)$, for $x \in [-1 + a, x_0 + a)$. Therefore,

$$\begin{aligned}
&\int_{-1+a}^{x_0+a} \left\{ \frac{\sigma^2(x) - \sigma^2}{\phi_a(x)} - \frac{\sigma^2(x) - \sigma^2}{\phi(x)} \right\} \mathrm{d}x \\
&= \int_{-1}^{x_0} \frac{\sigma^2(x+a) - \sigma^2}{\phi(x)} \, \mathrm{d}x - \int_{-1+a}^{x_0+a} \frac{\sigma^2(x) - \sigma^2}{\phi(x)} \, \mathrm{d}x \\
&= \int_{-1}^{x_0} \frac{\sigma^2(x+a) - \sigma^2(x)}{\phi(x)} \, \mathrm{d}x - \int_{x_0}^{x_0+a} \frac{\sigma^2(x) - \sigma^2}{\phi(x)} \, \mathrm{d}x + \int_{-1}^{-1+a} \frac{\sigma^2(x) - \sigma^2}{\phi(x)} \, \mathrm{d}x \\
&\leq 0 + 0 + \int_{-1}^{-1+a} \frac{\sigma^2}{\phi(x)} \, \mathrm{d}x \\
&\leq a,
\end{aligned} \tag{139}$$

where the first inequality follows from Remark D.1 and the second follows from $\phi(x) \geq \sigma^2$.

Finally, using the facts $\sigma^2 \le \sigma^2(x) \le 2\sigma^2$, and $\sigma^2 \le \phi(x)$, we have

$$\int_{-1}^{-1+a} \left\{ \frac{\sigma^2(x) - \sigma^2}{\phi_a(x)} - \frac{\sigma^2(x) - \sigma^2}{\phi(x)} \right\} \mathrm{d}x \le a. \tag{140}$$

Combining Eqs. (137), (138), (139),and (140) implies the desired result.

E Proof of Proposition 5.22

Proof. Let $\tilde{\mathsf{E}}_{\mathcal{W}}(\phi_a) = \tilde{\mathsf{E}}_{\mathcal{W},1}(\phi_a) + \tilde{\mathsf{E}}_{\mathcal{W},2}(\phi_a) + \tilde{\mathsf{E}}_{\mathcal{W},3}(\phi_a)$, where

$$\begin{aligned}
\tilde{\mathsf{E}}_{\mathcal{W},1}(\phi_a) &= \int_{x_0+a}^{\ell-1} \{\mathsf{I}(\mathcal{W} * \phi_a^{-1}(y)) - \mathsf{I}(\phi_a^{-1}(y-1))\} \mathrm{d}y, \\
\tilde{\mathsf{E}}_{\mathcal{W},2}(\phi_a) &= \int_{a}^{x_0+a} \{\mathsf{I}(\mathcal{W} * \phi_a^{-1}(y)) - \mathsf{I}(\phi_a^{-1}(y-1))\} \mathrm{d}y, \\
\tilde{\mathsf{E}}_{\mathcal{W},3}(\phi_a) &= \int_{0}^{a} \{\mathsf{I}(\mathcal{W} * \phi_a^{-1}(y)) - \mathsf{I}(\phi_a^{-1}(y-1))\} \mathrm{d}y.
\end{aligned} \tag{141}$$

Also let $\tilde{\mathsf{E}}_{\mathcal{W}}(\phi) = \tilde{\mathsf{E}}_{\mathcal{W},1}(\phi) + \tilde{\mathsf{E}}_{\mathcal{W},2,3}(\phi)$, where

$$\begin{aligned}
\tilde{\mathsf{E}}_{\mathcal{W},1}(\phi) &= \int_{x_0+a}^{\ell-1} \{\mathsf{I}(\mathcal{W} * \phi^{-1}(y)) - \mathsf{I}(\phi^{-1}(y-1))\} \mathrm{d}y, \\
\tilde{\mathsf{E}}_{\mathcal{W},2,3}(\phi) &= \int_{0}^{x_0+a} \{\mathsf{I}(\mathcal{W} * \phi^{-1}(y)) - \mathsf{I}(\phi^{-1}(y-1))\} \mathrm{d}y.
\end{aligned} \tag{142}$$

The following remark is used several times in the proof.

Remark E.1. *For any two values* $0 \le \alpha_1 < \alpha_2$*,*

$$\mathsf{I}(\alpha_2) - \mathsf{I}(\alpha_1) = \int_{\alpha_1}^{\alpha_2} \frac{1}{2} \mathsf{mmse}(z) \mathrm{d}z \le \int_{\alpha_1}^{\alpha_2} \frac{1}{2z} \mathrm{d}z = \frac{1}{2} \log\left(\frac{\alpha_2}{\alpha_1}\right) \le \frac{1}{2}\left(\frac{\alpha_2}{\alpha_1} - 1\right). \tag{143}$$

• Bounding $\tilde{\mathsf{E}}_{\mathcal{W},1}(\phi_a) - \tilde{\mathsf{E}}_{\mathcal{W},1}(\phi)$.
Notice that the functions $\phi(x) = \phi_a(x)$, for $x_2 \le x$. Also $\kappa/2 < \phi_a(x) \le \phi(x) \le \Phi_M$, for $x_1 < x < x_2$. Let $\alpha = (x_2 - x_1)/(x_2 - x_1 - a)$, and $\beta = (a x_2)/(x_2 - x_1 - a)$. Then, $\phi_a(x) = \phi(\alpha x - \beta)$ for

$x \in [x_0 + a, x_2)$. Hence,

$$
\begin{aligned}
&\tilde{\mathsf{E}}_{\mathcal{W},1}(\phi_a) - \tilde{\mathsf{E}}_{\mathcal{W},1}(\phi) \\
&= \int_{x_0+a}^{x_2+1} \mathsf{I}(\mathcal{W} * \phi_a^{-1}(y)) - \mathsf{I}(\mathcal{W} * \phi^{-1}(y))\, \mathrm{d}y + \int_{x_0+a}^{x_2+1} \mathsf{I}(\phi^{-1}(y-1)) - \mathsf{I}(\phi_a^{-1}(y-1))\, \mathrm{d}y \\
&\le \frac{1}{2} \int_{x_0+a}^{x_2+1} \frac{1}{\mathcal{W} * \phi^{-1}(y)} (\mathcal{W} * \phi_a^{-1}(y) - \mathcal{W} * \phi^{-1}(y))\, \mathrm{d}y \\
&\le \frac{\Phi_M}{2} \int_{x_0+a}^{x_2+1} \Big(\int_{x_0+a-1}^{x_2} \mathcal{W}(y-z) \phi_a^{-1}(z)\, \mathrm{d}z - \int_{x_0+a-1}^{x_2} \mathcal{W}(y-z) \phi^{-1}(z)\, \mathrm{d}z \Big) \mathrm{d}y \\
&= \frac{\Phi_M}{2} \int_{x_0+a}^{x_2+1} \Big(\int_{x_0+a}^{x_2} \mathcal{W}(y-z) \phi^{-1}(\alpha z - \beta)\, \mathrm{d}z + \int_{x_0+a-1}^{x_0+a} \mathcal{W}(y-z) \phi^{-1}(z-a)\, \mathrm{d}z \\
&\qquad\qquad - \int_{x_0+a-1}^{x_2} \mathcal{W}(y-z) \phi^{-1}(z)\, \mathrm{d}z \Big) \mathrm{d}y \\
&\le \frac{\Phi_M}{2} \int_{x_0+a}^{x_2+1} \Big\{ \int_{x_0}^{x_2} \Big(\frac{1}{\alpha} \mathcal{W}(y - \frac{z+\beta}{\alpha}) - \mathcal{W}(y-z) \Big) \phi^{-1}(z)\, \mathrm{d}z \\
&\qquad\qquad + \int_{x_0-1}^{x_0} \Big(\mathcal{W}(y-z-a) - \mathcal{W}(y-z) \Big) \phi^{-1}(z)\, \mathrm{d}z \\
&\qquad\qquad + \int_{x_0-1}^{x_0+a-1} \mathcal{W}(y-z) \phi^{-1}(z)\, \mathrm{d}z \Big\} \mathrm{d}y \\
&\le \frac{\Phi_M}{2} \int_{x_0+a}^{x_2+1} \Big\{ \int_{x_0}^{x_2} \Big(\mathcal{W}(y - \frac{z+\beta}{\alpha}) - \mathcal{W}(y-z) \Big) \phi^{-1}(z)\, \mathrm{d}z \\
&\qquad\qquad + \int_{x_0-1}^{x_0} \Big(\mathcal{W}(y-z-a) - \mathcal{W}(y-z) \Big) \phi^{-1}(z)\, \mathrm{d}z \\
&\qquad\qquad + \int_{x_0-1}^{x_0+a-1} \mathcal{W}(y-z) \phi^{-1}(z)\, \mathrm{d}z \Big\} \mathrm{d}y \\
&\le C_1 (1 - \frac{1}{\alpha}) + C_2 \frac{\beta}{\alpha} + C_3\, a \le C_4\, a.
\end{aligned} \tag{144}
$$

Here C_1, C_2, C_3, C_4 are some constants that depend only on K and κ. The last step follows from the facts that $\mathcal{W}(\,\cdot\,)$ is a bounded Lipschitz function and $\phi^{-1}(z) \le 2/\kappa$ for $z \in [x_1, x_2]$. Also, note that in the first inequality, $\mathsf{I}(\phi^{-1}(y-1)) - \mathsf{I}(\phi_a^{-1}(y-1)) \le 0$, since $\phi^{-1}(y-1) \le \phi_a^{-1}(y-1)$, and $\mathsf{I}(\,\cdot\,)$ is nondecreasing.

• Bounding $\tilde{\mathsf{E}}_{\mathcal{W},2}(\phi_a) - \tilde{\mathsf{E}}_{\mathcal{W},2,3}(\phi)$.
We have

$$
\begin{aligned}
\tilde{\mathsf{E}}_{\mathcal{W},2}(\phi_a) = &\int_{x_0+a-1}^{x_0+a} \{\mathsf{I}(\mathcal{W} * \phi_a^{-1}(y)) - \mathsf{I}(\phi_a^{-1}(y-1))\} \mathrm{d}y \\
&+ \int_{a}^{x_0+a-1} \{\mathsf{I}(\mathcal{W} * \phi_a^{-1}(y)) - \mathsf{I}(\phi_a^{-1}(y-1))\} \mathrm{d}y.
\end{aligned} \tag{145}
$$

We treat each term separately. For the first term,

$$
\begin{aligned}
&\int_{x_0+a-1}^{x_0+a} \{\mathsf{I}(\mathcal{W} * \phi_a^{-1}(y)) - \mathsf{I}(\phi_a^{-1}(y-1))\}\mathrm{d}y \\
&= \int_{x_0+a-1}^{x_0+a} \Big\{ \mathsf{I}\left(\int_{x_0+a}^{x_0+a+1} \mathcal{W}(y-z)\phi_a^{-1}(z)\ \mathrm{d}z + \int_{x_0+a-2}^{x_0+a} \mathcal{W}(y-z)\phi_a^{-1}(z)\ \mathrm{d}z \right) - \mathsf{I}(\phi_a^{-1}(y-1)) \Big\} \mathrm{d}y \\
&= \int_{x_0+a-1}^{x_0+a} \mathsf{I}\left(\int_{x_0}^{x_0+\alpha} \mathcal{W}(y - \frac{z+\beta}{\alpha})\phi^{-1}(z)\ \frac{\mathrm{d}z}{\alpha} + \int_{x_0-2}^{x_0} \mathcal{W}(y-a-z)\phi^{-1}(z)\ \mathrm{d}z \right) \mathrm{d}y \\
&\quad - \int_{x_0-1}^{x_0} \mathsf{I}(\phi^{-1}(y-1))\mathrm{d}y \\
&= \int_{x_0-1}^{x_0} \mathsf{I}\left(\int_{x_0}^{x_0+\alpha} \mathcal{W}(y+a - \frac{z+\beta}{\alpha})\phi^{-1}(z)\ \frac{\mathrm{d}z}{\alpha} + \int_{x_0-2}^{x_0} \mathcal{W}(y-z)\phi^{-1}(z)\ \mathrm{d}z \right) \mathrm{d}y \\
&\quad - \int_{x_0-1}^{x_0} \mathsf{I}(\phi^{-1}(y-1))\mathrm{d}y \\
&\leq C_5\ a + \int_{x_0-1}^{x_0} \Big\{ \mathsf{I}\left(\int_{x_0-2}^{x_0+1} \mathcal{W}(y-z)\phi^{-1}(z)\ \mathrm{d}z \right) - \mathsf{I}(\phi^{-1}(y-1)) \Big\} \mathrm{d}y \\
&= C_5\ a + \int_{x_0-1}^{x_0} \Big\{ \mathsf{I}(\mathcal{W} * \phi^{-1}(y)) - \mathsf{I}(\phi^{-1}(y-1)) \Big\} \mathrm{d}y, \qquad (146)
\end{aligned}
$$

where the last inequality is an application of remark E.1. More specifically,

$$
\begin{aligned}
&\mathsf{I}\left(\int_{x_0}^{x_0+\alpha} \mathcal{W}(y+a - \frac{z+\beta}{\alpha})\phi^{-1}(z)\ \frac{\mathrm{d}z}{\alpha} + \int_{x_0-2}^{x_0} \mathcal{W}(y-z)\phi^{-1}(z)\ \mathrm{d}z \right) \\
&\quad - \mathsf{I}\left(\int_{x_0-2}^{x_0+1} \mathcal{W}(y-z)\phi^{-1}(z)\ \mathrm{d}z \right) \\
&\leq \frac{\Phi_M}{2} \left(\int_{x_0}^{x_0+\alpha} \mathcal{W}(y+a - \frac{z+\beta}{\alpha})\phi^{-1}(z)\ \frac{\mathrm{d}z}{\alpha} - \int_{x_0}^{x_0+1} \mathcal{W}(y-z)\phi^{-1}(z)\ \mathrm{d}z \right) \\
&\leq \frac{\Phi_M}{2} \int_{x_0+1}^{x_0+\alpha} \mathcal{W}(y+a - \frac{z+\beta}{\alpha})\phi^{-1}(z)\ \mathrm{d}z \\
&\quad + \frac{\Phi_M}{2} \int_{x_0}^{x_0+1} \left(\mathcal{W}(y+a - \frac{z+\beta}{\alpha}) - \mathcal{W}(y-z) \right) \phi^{-1}(z)\mathrm{d}z \\
&\leq C_1'(1 - \frac{1}{\alpha}) + C_2' \frac{\beta}{\alpha} + C_3'\ a \leq C_5\ a,
\end{aligned}
$$

where C_1', C_2', C_3', C_5 are constants that depend only on κ. Here, the penultimate inequality follows from $\alpha > 1$, and the last one follows from the fact that $\mathcal{W}(\,\cdot\,)$ is a bounded Lipschitz function and that $\phi^{-1}(z) \leq 2/\kappa$, for $z \in [x_1, x_2]$.

To bound the second term on the right hand side of Eq. (146), notice that $\phi_a(z) = \phi(z-a)$, for $z \in [-1+a, x_0+a)$, whereby

$$
\int_{a}^{x_0+a-1} \{\mathsf{I}(\mathcal{W} * \phi_a^{-1}(y)) - \mathsf{I}(\phi_a^{-1}(y-1))\}\mathrm{d}y = \int_{0}^{x_0-1} \{\mathsf{I}(\mathcal{W} * \phi^{-1}(y)) - \mathsf{I}(\phi^{-1}(y-1))\}\mathrm{d}y. \qquad (147)
$$

Now, using Eqs. (142), (145) and (147), we obtain

$$\begin{aligned}\tilde{\mathsf{E}}_{\mathcal{W},2}(\phi_a) - \tilde{\mathsf{E}}_{\mathcal{W},2,3}(\phi) &\le C_5\, a - \int_{x_0}^{x_0+a} \{\mathsf{I}(\mathcal{W} * \phi^{-1}(y)) - \mathsf{I}(\phi^{-1}(y-1))\}\mathrm{d}y \\ &\le C_5\, a + \int_{x_0}^{x_0+a} \log\left(\frac{\phi^{-1}(y-1)}{\mathcal{W} * \phi^{-1}(y)}\right) \\ &\le C_5\, a + a \log(\frac{\Phi_M}{\kappa}) = C_6\, a,\end{aligned} \tag{148}$$

where C_6 is a constant that depends only on κ.

• Bounding $\tilde{\mathsf{E}}_{\mathcal{W},3}(\phi_a)$.
Notice that $\phi_a(y) \ge \sigma^2$. Therefore, $\mathsf{I}(\mathcal{W} * \phi_a^{-1}(y)) \le \mathsf{I}(\sigma^{-2})$, since $\mathsf{I}(\,\cdot\,)$ is nondecreasing. Recall that $\phi_a(y) = \phi^* < 2\sigma^2$, for $y \in [-1, -1+a)$. Consequently,

$$\tilde{\mathsf{E}}_{\mathcal{W},3}(\phi_a) \le \int_0^a \{\mathsf{I}(\sigma^{-2}) - \mathsf{I}(\phi^{*-1})\}\mathrm{d}y \le \frac{a}{2}\log\left(\frac{\phi^*}{\sigma^2}\right) < \frac{a}{2}\log 2, \tag{149}$$

where the first inequality follows from Remark E.1.

Finally, we are in position to prove the proposition. Using Eqs. (144), (148) and (149), we get

$$\tilde{\mathsf{E}}_{\mathcal{W}}(\phi_a) - \tilde{\mathsf{E}}_{\mathcal{W}}(\phi) \le C_4\, a + C_6\, a + \frac{a}{2}\log 2 = C(\kappa, K)\, a. \tag{150}$$

□

F Proof of Claim 5.25

Similar to the proof of Claim 5.20, the assumption $\int_{-1}^{\ell-1} |\phi(x) - \phi^*|\mathrm{d}x > C\sigma^2\ell$ implies $\phi(\theta\ell - 1) > C\sigma^2/2$, where

$$0 < \theta = \frac{\Phi_M - C\sigma^2}{\Phi_M - \frac{C\sigma^2}{2}} < 1.$$

Choose σ small enough such that $\phi^* < \phi_1$. Let $\kappa = (\phi_1 - \phi^*)(1-\theta)/2$. Applying Lemma 5.11, there exists ℓ_0, and σ_0, such that, $\int_{-1}^{\ell-1} |\phi(x) - \phi^*|\,\mathrm{d}x \le \kappa\ell$, for $\ell > \ell_0$ and $\sigma \in (0, \sigma_0]$. We claim that $\phi(\mu\ell - 1) < \phi_1$, with

$$\mu = 1 - \frac{\kappa}{\phi_1 - \phi^*} = \frac{1+\theta}{2}.$$

Otherwise, by monotonicity of $\phi(x)$,

$$(\phi_1 - \phi^*)(1-\mu)\ell \le \int_{\mu\ell-1}^{\ell-1} |\phi(x) - \phi^*|\,\mathrm{d}x < \int_{-1}^{\ell-1} |\phi(x) - \phi^*|\,\mathrm{d}x \le \kappa\ell. \tag{151}$$

Plugging in for μ yield a contradiction.

Therefore, $C\sigma^2/2 < \phi(x) < \phi_1$, for $x \in [\theta\ell - 1, \mu\ell - 1]$, and $(\mu - \theta)\ell = (1-\theta)\ell/2$. Choosing $\ell > \max\{\ell_0, 2K/(1-\theta)\}$ gives the result.

G Proof of Proposition 5.26

To prove Eq. (116), we write

$$\begin{aligned}\int_{-1}^{\ell-1}\{V_{\rm rob}(\phi_a(x))-V_{\rm rob}(\phi(x))\}\,{\rm d}x &= -\int_{x_1}^{x_2}\int_{\phi_a(x)}^{\phi(x)} V'(s)\,{\rm d}s\,{\rm d}x\\ &\le -\int_{x_1}^{x_2}\int_{\phi_a(x)}^{\phi(x)}\frac{\delta}{2s^2}\left(s-\sigma^2\right)\,{\rm d}s\,{\rm d}x\\ &= -\frac{\delta}{2}\int_{x_1}^{x_2}\left\{\log\left(\frac{\phi(x)}{\phi_a(x)}\right)+\frac{\sigma^2}{\phi(x)}-\frac{\sigma^2}{\phi_a(x)}\right\}{\rm d}x\\ &\le \frac{\delta}{2}K\log(1-a)+K\frac{\delta a}{C(1-a)},\end{aligned} \tag{152}$$

where the second inequality follows from the fact $C\sigma^2/2<\phi(x)$, for $x\in[x_1,x_2]$.

Next, we pass to prove Eq. (117).

$$\begin{aligned}\int_{-1}^{\ell-1}(\sigma^2(x)-\sigma^2)\left(\frac{1}{\phi_a(x)}-\frac{1}{\phi(x)}\right)\,{\rm d}x &= \int_{x_1}^{x_2}\frac{\sigma^2(x)-\sigma^2}{\phi(x)}\left(\frac{1}{1-a}-1\right)\\ &\le \frac{a}{1-a}\int_{x_1}^{x_2}\frac{\sigma^2}{\phi(x)}{\rm d}x \le K\frac{2a}{C(1-a)},\end{aligned} \tag{153}$$

where the first inequality follows from Remark D.1.

Finally, we have

$$\begin{aligned}\tilde{\mathsf{E}}_{\mathcal{W},{\rm rob}}(\phi_a)-\tilde{\mathsf{E}}_{\mathcal{W},{\rm rob}}(\phi) &= \int_0^{\ell}\{\mathsf{I}(\mathcal{W}*\phi_a^{-1}(y))-\mathsf{I}(\mathcal{W}*\phi^{-1}(y))\}{\rm d}y\\ &= \int_0^{\ell}\int_{\mathcal{W}*\phi^{-1}(y)}^{\mathcal{W}*\phi_a^{-1}(y)}\frac{1}{2}\mathsf{mmse}(s)\,{\rm d}s\,{\rm d}y\\ &\le \frac{\overline{D}(p_X)+\varepsilon}{2}\int_0^{\ell}\int_{\mathcal{W}*\phi^{-1}(y)}^{\mathcal{W}*\phi_a^{-1}(y)}s^{-1}{\rm d}s\,{\rm d}y\\ &\le \frac{\overline{D}(p_X)+\varepsilon}{2}\int_0^{\ell}\log\left(\frac{\mathcal{W}*\phi_a^{-1}(y)}{\mathcal{W}*\phi^{-1}(y)}\right)\,{\rm d}y\\ &\le -\frac{\overline{D}(p_X)+\varepsilon}{2}(K+2)\log(1-a),\end{aligned} \tag{154}$$

where the first inequality follows from Eq. (113) and Claim 5.25.

References

[ASZ10] S. Aeron, V. Saligrama, and Manqi Zhao, *Information theoretic bounds for compressed sensing*, IEEE Trans. on Inform. Theory **56** (2010), 5111 – 5130.

[BGI+08] R. Berinde, A.C. Gilbert, P. Indyk, H. Karloff, and M.J. Strauss, *Combining geometry and combinatorics: A unified approach to sparse signal recovery*, 47th Annual Allerton Conference (Monticello, IL), September 2008, pp. 798 – 805.

[BIPW10] K. Do Ba, P. Indyk, E. Price, and D. P. Woodruff, *Lower bounds for sparse recovery*, Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '10, 2010, pp. 1190–1197.

[BLM12] M. Bayati, M. Lelarge, and A. Montanari, *Universality in message passing algorithms*, In preparation, 2012.

[BM11a] M. Bayati and A. Montanari, *The dynamics of message passing on dense graphs, with applications to compressed sensing*, IEEE Trans. on Inform. Theory **57** (2011), 764–785.

[BM11b] ———, *The LASSO risk for gaussian matrices*, IEEE Trans. on Inform. Theory (2011), arXiv:1008.2581.

[BSB19] D. Baron, S. Sarvotham, and R. Baraniuk, *Bayesian Compressive Sensing Via Belief Propagation*, IEEE Trans. on Signal Proc. **58** (2019), 269–280.

[CD11] E. Candés and M. Davenport, *How well can we estimate a sparse vector?*, arXiv:1104.5246v3, 2011.

[CRT06a] E. Candes, J. K. Romberg, and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. on Inform. Theory **52** (2006), 489 – 509.

[CRT06b] ———, *Stable signal recovery from incomplete and inaccurate measurements*, Communications on Pure and Applied Mathematics **59** (2006), 1207–1223.

[CT05] E. J. Candés and T. Tao, *Decoding by linear programming*, IEEE Trans. on Inform. Theory **51** (2005), 4203–4215.

[DJM11] D. Donoho, I. Johnstone, and A. Montanari, *Accurate Prediction of Phase Transitions in Compressed Sensing via a Connection to Minimax Denoising*, arXiv:1111.1041, 2011.

[DMM09] D. L. Donoho, A. Maleki, and A. Montanari, *Message Passing Algorithms for Compressed Sensing*, Proceedings of the National Academy of Sciences **106** (2009), 18914–18919.

[DMM10] ———, *Message Passing Algorithms for Compressed Sensing: I. Motivation and Construction*, Proceedings of IEEE Inform. Theory Workshop (Cairo), 2010.

[DMM11] D.L. Donoho, A. Maleki, and A. Montanari, *The Noise Sensitivity Phase Transition in Compressed Sensing*, IEEE Trans. on Inform. Theory **57** (2011), 6920–6941.

[Don06a] D. L. Donoho, *Compressed sensing*, IEEE Trans. on Inform. Theory **52** (2006), 489–509.

[Don06b] D. L. Donoho, *High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension*, Discrete Comput. Geometry **35** (2006), 617–652.

[DT05] D. L. Donoho and J. Tanner, *Neighborliness of randomly-projected simplices in high dimensions*, Proceedings of the National Academy of Sciences **102** (2005), no. 27, 9452–9457.

[DT10] D. L. Donoho and J. Tanner, *Counting the faces of randomly-projected hypercubes and orthants, with applications*, Discrete & Computational Geometry **43** (2010), no. 3, 522–541.

[FZ99] A.J. Felstrom and K.S. Zigangirov, *Time-varying periodic convolutional codes with low-density parity-check matrix*, IEEE Trans. on Inform. Theory **45** (1999), 2181–2190.

[GSV05] D. Guo, S. Shamai, and S. Verdú, *Mutual information and minimum mean-square error in gaussian channels*, IEEE Trans. Inform. Theory **51** (2005), 1261–1282.

[HMU10] S.H. Hassani, N. Macris, and R. Urbanke, *Coupled graphical models and their thresholds*, Proceedings of IEEE Inform. Theory Workshop (Dublin), 2010.

[IPW11] P. Indyk, E. Price, and D.P. Woodruff, *On the Power of Adaptivity in Sparse Recovery*, IEEE Symposium on the Foundations of Computer Science, FOCS, October 2011.

[KGR11] U.S. Kamilov, V.K. Goyal, and S. Rangan, *Message-Passing Estimation from Quantized Samples*, arXiv:1105.6368, 2011.

[KMRU10] S. Kudekar, C. Measson, T. Richardson, and R. Urbanke, *Threshold Saturation on BMS Channels via Spatial Coupling*, Proceedings of the International Symposium on Turbo Codes and Iterative Information Processing (Brest), 2010.

[KMS+11] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborova, *Statistical physics-based reconstruction in compressed sensing*, arXiv:1109.4424, 2011.

[KP10] S. Kudekar and H.D. Pfister, *The effect of spatial coupling on compressive sensing*, 48th Annual Allerton Conference, 2010, pp. 347 –353.

[KRU11] S. Kudekar, T. Richardson, and R. Urbanke, *Threshold Saturation via Spatial Coupling: Why Convolutional LDPC Ensembles Perform So Well over the BEC*, IEEE Trans. on Inform. Theory **57** (2011), 803–834.

[KT07] B.S. Kashin and V.N. Temlyakov, *A remark on compressed sensing*, Mathematical Notes **82** (2007), 748–755.

[LF10] M. Lentmaier and G. P. Fettweis, *On the thresholds of generalized LDPC convolutional codes based on protographs*, IEEE Intl. Symp. on Inform. Theory (Austin, Texas), August 2010.

[Mon12] A. Montanari, *Graphical models concepts in compressed sensing*, Compressed Sensing (Y.C. Eldar and G. Kutyniok, eds.), Cambridge University Press, 2012.

[Ran11] S. Rangan, *Generalized Approximate Message Passing for Estimation with Random Linear Mixing*, IEEE Intl. Symp. on Inform. Theory (St. Perersbourg), August 2011.

[Rén59] A. Rényi, *On the dimension and entropy of probability distributions*, Acta Mathematica Hungarica **10** (1959), 193–215.

[RU08] T.J. Richardson and R. Urbanke, *Modern Coding Theory*, Cambridge University Press, Cambridge, 2008.

[RWY09] G. Raskutti, M. J. Wainwright, and B. Yu, *Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls*, 47th Annual Allerton Conference (Monticello, IL), September 2009.

[Sch10] P. Schniter, *Turbo Reconstruction of Structured Sparse Signals*, Proceedings of the Conference on Information Sciences and Systems (Princeton), 2010.

[Sch11] ———, *A message-passing receiver for BICM-OFDM over unknown clustered-sparse channels*, arXiv:1101.4724, 2011.

[SLJZ04] A. Sridharan, M. Lentmaier, D. J. Costello Jr, and K. S. Zigangirov, *Convergence analysis of a class of LDPC convolutional codes for the erasure channel*, 43rd Annual Allerton Conference (Monticello, IL), September 2004.

[SPS10] S. Som, L.C. Potter, and P. Schniter, *Compressive Imaging using Approximate Message Passing and a Markov-Tree Prior*, Proc. Asilomar Conf. on Signals, Systems, and Computers, November 2010.

[VS11] J. Vila and P. Schniter, *Expectation-maximization bernoulli-gaussian approximate message passing*, Proc. Asilomar Conf. on Signals, Systems, and Computers (Pacific Grove, CA), 2011.

[Wai09] M.J. Wainwright, *Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting*, IEEE Trans. on Inform. Theory **55** (2009), 5728–5741.

[WV10] Y. Wu and S. Verdú, *Rényi Information Dimension: Fundamental Limits of Almost Lossless Analog Compression*, IEEE Trans. on Inform. Theory **56** (2010), 3721–3748.

[WV11a] ———, *MMSE dimension*, IEEE Trans. on Inform. Theory **57** (2011), no. 8, 4857–4879.

[WV11b] ———, *Optimal Phase Transitions in Compressed Sensing*, arXiv:1111.6822, 2011.