Proceedings of Machine Learning Research vol 99:1–16, 2019 32nd Annual Conference on Learning Theory

Accuracy-Memory Tradeoffs and Phase Transitions in Belief Propagation

Vishesh Jain VISHESHJ@MIT.EDU
Frederic Koehler FKOEHLER@MIT.EDU
Massachusetts Institute of Technology. Department of Mathematics.

Jingbo Liu JINGBO@MIT.EDU
Massachusetts Institute of Technology. IDSS.

Elchanan Mossel ELMOS@MIT.EDU
Massachusetts Institute of Technology. Department of Mathematics and IDSS.

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

The analysis of Belief Propagation and other algorithms for the *reconstruction problem* plays a key role in the analysis of community detection in inference on graphs, phylogenetic reconstruction in bioinformatics, and the cavity method in statistical physics. We prove a conjecture of Evans, Kenyon, Peres, and Schulman (2000) which states that any bounded memory message passing algorithm is statistically much weaker than Belief Propagation for the reconstruction problem. More formally, any recursive algorithm with bounded memory for the reconstruction problem on the trees with the binary symmetric channel has a phase transition strictly below the Belief Propagation threshold, also known as the Kesten-Stigum bound. The proof combines in novel fashion tools from recursive reconstruction, information theory, and optimal transport, and also establishes an asymptotic normality result for BP and other message-passing algorithms near the critical threshold.

1. Introduction

Belief Propagation is one of the most popular algorithms in graphical models. The main result of this paper (Theorem 3) shows that bounded memory variants of Belief Propagation have no asymptotic statistical power in regimes where Belief Propagation does. This proves a long-standing conjecture (Conjecture 2) (Evans et al., 2000).

Belief Propagation: Belief Propagation (BP) is one of most popular algorithms in machine learning and probabilistic inference (Pearl, 1988). It is also a key algorithm and a key analytic tool in statistical physics with applications to inference problems and coding where it is an important ingredient of replica analysis (e.g. Mézard and Montanari (2009)), and in probability theory, where it is studied under the names of "broadcasting on trees" and the "reconstruction problem on trees" (e.g. Mossel (2004b)). The analysis of BP on trees plays a crucial role in many inference problems arising in different fields. In the problem of phylogenetic reconstruction arising in biology, both belief propagation and bounded memory algorithms like recursive majority have been extensively studied in theory and practice, and

. Authors are sorted alphabetically. Extended abstract. Full version appears as [arXiv:1905.10031v1].

they have also played an important role in works on learning phylogenies (the underlying tree structure): see e.g. Mossel (2004a); Daskalakis et al. (2006). In the analysis of community detection in block models, BP on trees and related message-passing algorithms (e.g. linearizations of BP) play a fundamental role in predicting and rigorously analyzing the recoverability of community structure, as the sparse SBM is locally tree-like: see e.g. Decelle et al. (2011); Krzakala et al. (2013); Mossel et al. (2015, 2014). Finally, there is a long history of BP being studied and used in the theory of error correcting codes (e.g. Richardson and Urbanke (2001); Montanari (2005)).

One of the reasons for the popularity of BP is the fact that its time complexity is linear in the number of nodes in the factor graph (assuming real-number operations count as one operation), while a brute force algorithm generally has exponential complexity. Given that the algorithm is a simple recursive algorithm that is easy to implement and runs in linear time, it is natural to ask how fragile it is. In particular, are there bounded (bit) memory variants of the algorithm that are as statistically efficient as the algorithm itself? Is the algorithm robust to a small amount of noise during it execution? These natural problems, which were open for almost two decades, intimately relate to a recent impressive body of work in machine learning which tries to understand the statistical implications of computationally limited algorithms.

Statistical efficiency of computationally efficient algorithms: Understanding the power of computationally limited algorithms for inference tasks is a major (or perhaps the) task of computational learning theory. A recent trend in this area deals with reductions between inference problems on graphs that are known to be informational theoretically solvable but are assumed to be unsolvable in polynomial time. Thus a recent line of work including Berthet et al. (2013); Ma and Wu (2015); Brennan et al. (2018) aims to prove that for certain inference problems a *computational-statistical gap* exists: more data is needed to infer in polynomial time than is needed information theoretically. For many other problems, no reductions are known though it is believed that similar phenomena occur. Some notable recent examples include the multi-community stochastic block model (Decelle et al., 2011; Zdeborová and Krzakala, 2016) and sparse linear regression (see e.g. Zhang et al. (2017)). Interestingly, in most of the problems discussed above, BP and other message passing algorithms are either known or conjectured to be the optimal algorithms among all computationally efficient algorithms.

Can statements proving computational-statistical gaps be proved unconditionally (i.e. not by reduction)? Very impressive results were recently proven by Raz (2018) and follow up work (Raz, 2017; Kol et al., 2017; Garg et al., 2018; Moshkovitz and Moshkovitz, 2017) for learning tasks with bounded memory, where it is shown that unless the memory is quadratic in the instance size, the running time of the algorithm has to be exponential.

Communication complexity of distributed estimation: In many problems concerning the trade-offs between the communication complexity and the statistical risk (or other system performance measures), a useful tool for establishing lower bounds is the *strong data processing inequality*, which dictates how fast the mutual information must decay along a Markov chain (Ahlswede and Csiszár (1986); Zhang et al. (2013); Liu et al. (2014); Liu et al. (2015); Braverman et al. (2016); Liu et al. (2017); Xu and Raginsky (2017); Hadar et al. (2019)). Yet, sometimes, strictly better lower bounds may be obtained by replacing

the strong data processing argument with a careful analysis of the contraction of the Fisher information or χ^2-information due to compression (Han et al. (2018); Barnes et al. (2018); Acharya et al. (2018)). For example, Barnes et al. (2018) proved the contraction of the Fisher information in the Gaussian location model via a geometric analysis of the quantization of the score function (which is a high dimensional Gaussian vector). The χ^2-contraction idea is relevant to the BP with bounded memory problem considered in the current paper, but a key difficulty (which we resolved using novel optimal transportation methods) is to show a Gaussian approximation result for "good" reconstruction algorithms.

Our results: Our results show that any message-passing algorithm using finite memory is much weaker than BP in the following sense: there is a range of parameters of the model for which BP is a good estimator while any bounded message-passing algorithm has no statistical power. This has immediate implications for applications of BP in phylogeny, for the block model and for population dynamics, as this imply that in these applications too, the algorithms used (BP or others) cannot be replaced by bounded memory message-passing algorithms. We proceed with definitions and formal statements of the main results.

1.1. Broadcasting on trees

On a rooted tree T with root ρ, the *broadcast process* with error probability $\varepsilon \in (0, 1/2)$ is defined as follows, generating labels $X_v \in \{\pm 1\}$ for every vertex $v \in T$. The label X_ρ of the root ρ is assigned either $+1$ or -1 with equal probability, and then for each edge $e = (v_1, v_2)$, the probability that the labels assigned to v_1 and v_2 are different is ε, independent of all other edges. This model has several interpretations. In communication theory, one may consider each of the edges as an independent binary symmetric channel. In biology, one may say that there is some binary property each child inherits from its parent independently for all children. This model is also an example of an Ising model on T with constant interaction strength Lyons (1989). We refer the reader to Evans et al. (2000) and Mossel (2004b) for a detailed account of the history of this model, as well as classical and modern results.

One of the most fundamental questions about this process is the following: for a given set of vertices V (e.g. the leaves of a finite tree), can we typically infer the original label correctly from the labels at V? More precisely, let $p(T, V, \varepsilon)$ denote the probability of reconstructing the label at the root given the labels at V. Since random guessing succeeds in reconstructing the original label with probability $1/2$, it is natural to write $p(T, V, \varepsilon) := (1/2) + s(T, V, \varepsilon)$, so that $s(T, V, \varepsilon)$ represents the advantage over random guessing gained by having access to the labels at V. For an infinite tree T, a natural question to ask is whether there is a *uniform* advantage over random guessing provided by knowing the labels at the vertices at *any* level of the tree. Formally, we let V_n denote the set of all vertices at distance n from the root ρ and ask whether $\lim_{n\to\infty} s(T, V_n, \varepsilon) = \inf_{n\geq 1} s(T, V_n, \varepsilon) > 0$ (the first equality follows by the data-processing inequality); if so, we say that the *reconstruction problem* for T at ε is *solvable*. The solvability threshold is determined by the branching number of T:

Theorem 1 (Evans et al. (2000)) *Consider the broadcasting process with parameter ε on an infinite tree T with branching number $br(T)$. Then, with $\varepsilon_c := (1 - brT^{-1/2})/2$,*

$$\lim_{n\to\infty} s(T, V_n, \varepsilon) \begin{cases} > 0, & \text{if } \varepsilon < \varepsilon_c \\ = 0, & \text{if } \varepsilon > \varepsilon_c. \end{cases} \tag{1}$$

(Note that for d-ary trees, $br(T) = d$.)

Remarkably, the exact same threshold also dictates the limit of *weak recovery* for the 2-community stochastic block model (Mossel et al., 2015; Massoulie, 2013; Mossel et al., 2018). Intuitively, this is because locally around a typical node, the SBM graph looks like a Galton-Watson tree, and the community assignment of nodes can be coupled to the aforementioned broadcast process on the tree. The same threshold also dictates the phase transition for the phylogenetic reconstruction problem, where the goal is to reconstruct the underlying tree (Mossel, 2004a; Daskalakis et al., 2006; Steel, 2001). The intuition here is that information about the deep part of the structure of the tree is transmitted via the broadcast channel.

The proof that the above limit is positive for $\varepsilon < \varepsilon_c$ first appears in Kesten and Stigum (1966) The proof that the limit is 0 when $\varepsilon > \varepsilon_c$ is harder, and only appeared around two decades later, first for regular trees (in which case $\mathrm{br}(T)$ coincides with the arity of T) in Bleher et al. (1995), and then for general trees in Evans et al. (2000). Subsequently, different proofs appeared in Ioffe (1996) and Berger et al. (2005).

1.2. Message-passing algorithms for reconstruction

For simplicity of notation, we focus on the setting where V, the set of revealed nodes, are the leaves of some finite-depth tree. Then a *message-passing* algorithm for reconstructing the label at the root, given the labels X_V at the set of leaves V of the tree is specified by the following data: (i) a *message space* Σ (possibly infinite) to which messages belong; (ii) initial messages $(Y_v)_{v\in V}$ in Σ^V, where Y_v is a (possibly random) function of X_v for all $v \in V$; and (iii) for each vertex $u \in T$, a fixed *reconstruction function* $f_u : \Sigma^{C(u)} \to \Sigma$, where $C(u)$ denotes the children of u and f_u is allowed to be a randomized function (i.e. a channel; The randomness of this channel is independent of the randomness of the tree broadcast process). Reconstruction proceeds recursively – the message output by node u (to its parent) is

$$Y_u = f_u(Y_{C(u)}). \tag{2}$$

To visualize, we can think of the X as living on a "broadcasting tree" and the Y as living on a mirrored "reconstruction tree" (see Figure 1). Letting $Y_\pm$ denote the random variables corresponding to the output Y_ρ of f_ρ under $T_V^\pm$ (the distribution of labels at the leaves V, conditioned on the label at ρ being $\pm$), the probability of correct reconstruction of the root given Y_ρ is $0.5(1 + \mathbf{TV}(Y_+, Y_-))$. Thus, a natural measure of the power of the message passing algorithm is $\mathbf{TV}(Y_+, Y_-)$. The size of the message space plays a crucial role in this paper; we call $\log_2 |\Sigma|$ the number of bits of memory used by the algorithm.

In this context BP is just the usual recursive scheme for computing exactly the marginal distribution of the label at ρ, given the labels at V. Explicitly, $\Sigma = [-1, 1]$, $Y_v = X_v$ for leaf nodes v, and the reconstruction function at every node u is (by Bayes rule, with $\theta = 1 - 2\epsilon$)

$$f_u^{(BP)}\left(y_1, \ldots, y_{|C(u)|}\right) := \frac{\prod_i (1 + \theta y_i) - \prod_i (1 - \theta y_i)}{\prod_i (1 + \theta y_i) + \prod_i (1 - \theta y_i)}$$

Note that, by definition, outputting the more likely label under the marginal is the Bayes optimal classifier in this setting i.e. it achieves the maximum probability of reconstruction among *all* algorithms. In particular, the advantage of belief propagation over random guessing enjoys the limiting behavior in (1).

1.3. Limitations of bounded memory algorithms

Another natural algorithm for the reconstruction problem is to output the label present at a majority of the vertices in V. While this does not achieve the same probability of success as belief propagation, it is known that it *does* achieve the limiting behavior in (1) (Kesten and Stigum, 1967). Note that this is a message-passing algorithm with $\Sigma = \mathbb{Z}$ and f which sums its inputs.

What if Σ is a bounded size alphabet? In the case $|\Sigma| = 2$, the natural message-passing variant of this algorithm is to estimate the label at ρ via *recursive majority* i.e. $\Sigma = \{\pm 1\}$, and $f_u \colon \Sigma^{C(u)} \to \Sigma$ is the majority function (we assume for convenience that $|C(u)|$ is odd for each u). Note that recursive majority is easier to implement than belief propagation, not requiring access to ε; for this reason among others, it is quite popular in practice, in particular in biological applications.

The following striking conjecture states that reconstruction down to the KS threshold requires unbounded memory, i.e. there is no bounded-memory analogue of BP:

Conjecture 2 (Evans et al. (2000)) *For any fixed $L > 0$, no message-passing algorithm on an alphabet of size L can achieve the guarantee of (1). In other words, there exists a fixed noise level $\varepsilon(L)$ such that reconstruction is information-theoretically possible but no such message-passing algorithm is asymptotically better than a random guess.*

The conjecture is also discussed in (Mossel, 2004b). Mossel (1998) verified this conjecture on periodic trees in the special case when $\Sigma = \{\pm 1\}$, the reconstruction function $f_u = f$ is the same for all nodes u, and $f(-x) = -f(x)$ for all inputs x; in particular, this includes the important case of recursive majority.

The main result of this paper is to verify Conjecture 2 on the d-ary tree[1] *for all $L > 0$.* The proof will rely upon a careful analysis of the distributional recursion induced by combining the recursive broadcast and reconstruction (message-passing) steps. In fact, we present two approaches along these lines: a relatively elementary approach which proves the result for one bit memory ($L = 2$) and illustrates many of the main difficulties in this problem, and a higher-powered method (which crucially builds upon Wasserstein estimates from optimal transport theory) which proves the result for all L and even pins down the correct quantitative dependence on the distance to the threshold:

Theorem 3 *For any integer $d \geq 2$, there exist positive real numbers c_d and C_d such that the following holds: for any fixed L, the maximum error probability $\varepsilon(L)$ for which there exists a message-passing algorithm with alphabet size L guaranteeing asymptotic reconstruction on the infinite d-ary tree satisfies*

$$L^{-C_d} \leq \varepsilon_c - \varepsilon(L) \leq L^{-c_d}. \tag{3}$$

As a by product, we remark that the (renormalized) BP message distribution at the root of the infinite d-ary tree approaches a Gaussian as ε approaches criticality. More precisely:

Corollary 4 *Fix $d \geq 2$ and broadcasting parameter $\varepsilon \in (0, \varepsilon_c)$. Let ρ denote the root of a d-ary tree of depth n, V_n denote the set of leaves, and let $Y_n := \mathbf{E}[X_\rho | X_{V_n}]$ under the*

1. We will also assume for convenience that f_u is constant at each level of the tree.

broadcast process. Then there exists a limit r.v. Y such that $Y_n \to Y$ in distribution and

$$W_2\left(\frac{Y}{\sqrt{\mathbf{Var}(Y)}}, G\right) \leq (\varepsilon_c - \varepsilon)^{C_d},$$

for some $C_d > 0$ independent of ε, where $G \sim N(0,1)$ and W_2 denotes the 2-Wasserstein distance (see e.g. the definition in Villani (2008)).

Let us emphasize that the lower bound in Theorem 3 applies to general reconstruction schemes (not necessarily discretized BP), even though Corollary 4 only concerns the specific BP algorithm. We also note that Gaussian approximation of BP is widely used in *density evolution* analysis. This is a different setup where the number of iterations is bounded but the degree goes to infinity, see e.g. Bayati and Montanari (2011). Note in particular, that in the density evolution setting, normal approximation is used both above and below the reconstruction threshold.

A subsequent work to ours, Moitra et al. (2019), studies the complexity of Belief Propagation from the point of view of circuit complexity. Most relevant to us is the result of Moitra et al. (2019) showing that BP can be computed in $\mathbf{NC}^1$. Thus, there exist a circuit of depth $O(n)$ with binary AND and OR gates and NOT gates that computes Belief Propagation in the following sense. The input to gates are the leaves values (repeated many times). The circuit returns a bit that agrees with the more likely posterior according to BP, whenever the BP posterior has a bias of more than $1/d^n$. These results hold independently of broadcast parameter ε. The results of Moitra et al. (2019) do not contradict the results of the current paper as the circuit constructed does not confirm to a tree topology with each input bits appearing only once. Rather, in the circuit constructed each input bit is repeated $d^{O(n)}$ times. In other results, Moitra et al. (2019) show that bounded depth circuits with AND and OR gates (the class $\mathbf{AC^0}$) cannot compute a nontrivial approximation to BP even in an average sense above the KS bound.

Organization: As described above, we first sketch a more elementary proof of the lower bound (impossibility result) in the $L = 2$ case of Theorem 3 in Section 2, then prove it completely (along with Corollary 4) with a more powerful approach in Section 3. Missing proofs and the matching upper bound via a quantization of BP can be found in the appendices of the arXiv version.

2. Impossibility of 1-bit reconstruction near criticality

In this section, we will sketch the main ideas behind a relatively simple and self-contained information theoretic proof of the impossibility of 1-bit message-passing algorithms solving the reconstruction problem all the way to the Kesten-Stigum (KS) threshold. Our analysis of this case will also serve to illustrate the challenges encountered towards resolving Conjecture 2, and shed additional light on its ultimate resolution in the next section. Complete statements and proofs for this section can be found in the appendices of the arXiv version.

Throughout this section, we will adopt the convenient reparameterization $\varepsilon = 1/2 - \nu$. Note that with this reparameterization, the KS threshold corresponds to $4d\nu^2 = 1$ i.e. the reconstruction problem is solvable if $4d\nu^2 > 1$ and unsolvable if $4d\nu^2 < 1$.

2.1. A direct proof of the Kesten-Stigum bound

Here, for simplicity, we will only discuss the KS bound in the case of the infinite d-ary tree. Denoting by $T_n^{\pm}$ the distributions on the labels $(X_v)_{v\in V}$ for the leaves V of the depth n tree, conditioned on the root being $\pm$. Recall that $2s(T, V_n, \varepsilon) = \mathbf{TV}(T_n^+, T_n^-)$. Note also that by considering the labels at the vertices one level below the root, it is easily seen that

$$T_n^{\pm} \sim \left(\left(\frac{1}{2}+\nu\right) T_{n-1}^{\pm} + \left(\frac{1}{2}-\nu\right) T_{n-1}^{\mp}\right)^{\otimes d}.$$

Owing to this recursive structure of the problem, it is much more convenient to switch to an information measure which tensorizes well (and which is also 'stronger' than TV). Here, we make the choice of working with the symmetrized version of KL-divergence (also known as Jeffrey's divergence), defined by $\mathbf{SKL}(P,Q) := \mathbf{KL}(P,Q) + \mathbf{KL}(Q,P)$; by Pinsker's inequality, $\mathbf{SKL}(T_n^+, T_n^-) \to 0$ shows that $\mathbf{TV}(T_n^+, T_n^-) \to 0$ as well.

Of key importance to us is the fact that SKL behaves very well under 'symmetric mixtures'; we show using a short direct computation (see arXiv version) that

$$\mathbf{SKL}\left(\left(\frac{1}{2}+\nu\right)P + \left(\frac{1}{2}-\nu\right)Q, \left(\frac{1}{2}-\nu\right)P + \left(\frac{1}{2}+\nu\right)Q\right) \le 4\nu^2\,\mathbf{SKL}(P,Q).$$

Given this 'mixing inequality', the proof of the KS bound is now immediate. Indeed,

$$\begin{aligned}\mathbf{SKL}(T_n^+, T_n^-) &= d\,\mathbf{SKL}\left(\left(\frac{1}{2}+\nu\right)T_{n-1}^+ + \left(\frac{1}{2}-\nu\right)T_{n-1}^-, \left(\frac{1}{2}-\nu\right)T_{n-1}^+ + \left(\frac{1}{2}+\nu\right)T_{n-1}^-\right)\\ &\le 4\nu^2 d\,\mathbf{SKL}(T_{n-1}^+, T_{n-1}^-) \le (4\nu^2 d)^{n-1}\,\mathbf{SKL}(T_1^+, T_1^-) = O\left((4\nu^2 d)^{n-1}\right),\end{aligned}$$

so we see that $\lim_{n\to\infty} \mathbf{SKL}(T_n^+, T_n^-) = 0$ if $4\nu^2 d < 1$.

2.2. Interlude: noisy message-passing algorithms fail near criticality

Showing that 'noisy' message-passing algorithms fail near criticality requires only a slight extension of the above discussion, and is a natural segue into our discussion of 1-bit message-passing algorithms. Here, by a noisy message-passing algorithm, we mean that for every node u in the tree, messages from the children of u to u are processed through independent copies of a noisy channel $P_{Y|X} : \Sigma \to \Sigma$. In this setting, instead of $T_n^{\pm}$, the natural choice of distributions to look at are $P_n^{\pm}$, where $P_n^{\pm}$ denotes the distribution on Σ (which we interpret as the final message from the root) obtained by broadcasting n levels down, conditioned on the root being $\pm$, and reconstructing using our (noisy) message-passing algorithm. Once again, by considering the labels at the vertices one level below the root, it is immediate that

$$P_n^{\pm} = f_*\left(\left(P_{Y|X} \circ \left(\left(\frac{1}{2}+\nu\right)P_{n-1}^{\pm} + \left(\frac{1}{2}-\nu\right)P_{n-1}^{\mp}\right)\right)^{\otimes d}\right), \tag{4}$$

where $f : \Sigma^d \to \Sigma$ denotes the reconstruction function at the root, and $f_*(\mu)$ denotes the pushforward of the measure μ by the function f. In particular, it follows that if the channel $P_{Y|X}$ satisfies a strong data-processing inequality (SDPI) i.e. there exists some constant

$\eta \in [0,1)$ such that for all distributions P, Q on Σ, $\mathbf{SKL}(P_{Y|X} \circ P, P_{Y|X} \circ Q) \leq \eta\, \mathbf{SKL}(P, Q)$ then it follows by a similar computation as above that

$$\mathbf{SKL}(P_n^+, P_n^-) \leq 4\nu^2 \eta d\, \mathbf{SKL}\left(P_{n-1}^+, P_{n-1}^-\right) = O\left((4\nu^2\eta d)^{n-1}\right),$$

so we see that such algorithms can solve the reconstruction problem only if $4\nu^2 d \geq \eta^{-1}$ i.e. they do not work all the way to the KS threshold.

2.3. 1-bit message-passing algorithms fail near criticality

Motivated by the observation that for a fixed finite alphabet Σ, any function $f : \Sigma^d \to \Sigma$ cannot be injective (for d large enough) on the support of any non-trivial product distribution on Σ^d, and therefore, must 'lose' information, it is tempting to think that the above strategy for noisy message-passing algorithms can be adapted directly to show that finite-bit message-passing algorithms fail near criticality as well. However, it is not the case that a general function $f : \Sigma^d \to \Sigma$ satisfies an SDPI, even when $\Sigma = \{0,1\}$:

Example 1 (No SDPI for general f) *For $P = \mathbf{Ber}(p), Q = \mathbf{Ber}(q)$, and $f : \{0,1\}^d \to \{0,1\}$ equal to the OR-function,*

$$\lim_{p,q \to 0} \frac{\mathbf{SKL}(f_*(P^{\otimes d}), f_*(Q^{\otimes d}))}{\mathbf{SKL}(P^{\otimes d}, Q^{\otimes d})} = 1.$$

This is because in the limit we can disregard all events as negligible except that either all inputs are 0, or that a single input is 1, and the OR function memorizes which event occurred.

On the other hand, it turns out that our original intuition is 'mostly correct': more precisely, we will show that such functions do indeed satisfy an SDPI provided the input distributions under consideration have 'robust full support' i.e. they assume each symbol of the alphabet Σ with probability at least some uniform positive constant. In particular, this shows that any potential finite-bit message-passing algorithm which succeeds near criticality must get close to the boundary of the probability simplex in $\mathbb{R}^\Sigma$ infinitely often.

In the case when $\Sigma = \{0,1\}$, we can say even more, and prove an inverse theorem for the non-contraction of SKL: not only can SKL non-contraction only occur near the boundary of the probability simplex (in this case, identified with $[0,1]$), but also, the functions achieving such non-contraction are only those with behavior similar to the OR-function (for p_n^+, p_n^- close to 0), in that they are able to distinguish the all 0s input from inputs with a single 1, or symmetrically for AND (with p_n^+, p_n^- close to 1). From this we see that 1-bit algorithms with the same reconstruction function at every node fail near criticality (eliminating the symmetry assumption from Mossel (1998)), because the only mechanism which prevents contraction in SKL (behaving like OR or AND near appropriate boundaries) also 'repels' the distributional iterates away from the boundary.

We consider the 1-bit reconstruction problem when the reconstruction functions at different levels are allowed to be different. This larger class includes natural reconstruction schemes such as the so-called 'TRIBES' function from Boolean analysis, which uses either the AND-function or the OR-function depending on the level: the potential problem is that such an algorithm could alternate 'losing' steps in which the distributional dynamics go towards the boundary of the probability simplex $[0,1]$ with 'gaining' steps, where a function

like AND or OR (depending on the boundary) is applied to gain in SKL (note that we are assuming that $4d\nu^2 > 1$).

To overcome this obstacle, we introduce a Lyapunov function for our discrete time dynamical system; more precisely, we define a function ϕ such that $\phi \to -\infty$ implies that $\mathbf{SKL}(P_n^+, P_n^-) \to 0$, and for which we can show that ϕ decreases at every step. The essential idea is to define $\phi(P, Q)$ to be $\log \mathbf{SKL}(P, Q)$ plus some 'negative log-barrier' term, which penalizes $\phi(P, Q)$ for moving away from the boundary — by carefully balancing these terms, we can ensure ϕ indeed goes down at every step. Finally, since the log-barrier term is bounded from below, it follows that $\phi(P_n^+, P_n^-) \to -\infty$ implies that $\mathbf{SKL}(P_n^+, P_n^-) \to 0$.

3. Impossibility of multibit reconstruction near criticality

In the previous section, we saw (Example 1) that contrary to what may be the natural intuition, even restricting the messages to a single bit does not imply that a significant amount of information is destroyed at a particular level of reconstruction. In the 1-bit case, we overcame this obstacle using a multilevel analysis (of the iteration $(P_t^+, P_t^-) \mapsto (P_{t+1}^+, P_{t+1}^-)$) that treated the boundary of the 1-dimensional simplex specially. In the multibit case, the dynamics live in a higher-dimensional simplex and the boundary behavior appears very complicated to analyze.

In this section, we give a new lower bound argument which completely overcomes this difficulty, proving the lower bound in Theorem 3. This argument requires several significant innovations, which we briefly summarize:

Tracking only the law of the "score" $\mathbf{E}[X|Y]$: Let $X = X_\rho$ be the label of the root (of a depth n tree) and $Y = Y_\rho$ be the reconstructed data at the root (i.e. the message that would be passed to an imaginary parent of the root node). The previous analysis tracked the complete distribution of $Y|X$. The recursion from the law for a depth n tree to depth $n+1$ tree is very easy to describe, but the resulting dynamics may be very complex in the multibit case (where the distributional recursion lives in a high-dimensional probability simplex). The new analysis considers only the induced law of $\mathbf{E}[X|Y]$ (i.e. the distribution of a natural real-valued random variable) and studies a BP-style recursion to relate the law at depths n and $n+1$. We remark that such an approach has been successfully used in many of the previous works around the reconstruction problem (see, e.g., Bleher et al. (1995); Borgs et al. (2006); Pemantle et al. (2010)) The analysis of this nonlinear recursion is tamed by an approximate linearization and decoupling argument (Lemma 5).

Identifying attraction towards Gaussianity (in the natural Wasserstein metric): Recall that in the 1-bit case we were able to prove that "good" reconstruction functions (those that do not destroy much information) must push the law of $Y|X$ towards the middle of the 1-dimensional probability simplex. Analogously, we ultimately show (Lemma 8) that good reconstruction functions push the law of $\mathbf{E}[X|Y]$ towards Gaussianity (measured in W_2 distance, see equation (7)).

Multilevel analysis of kurtosis, variance, and Gaussianity: The proof of the key Gaussian attraction result (Lemma 8) would be much simpler if, for X_1, X_2 i.i.d., the sum $\frac{1}{\sqrt{2}}(X_1 + X_2)$ were significantly closer to Gaussian (in W_2) than X_1 itself. Unfortunately, this is only true in the case of bounded kurtosis. Instead, we make a more complex argument

with two main steps: (1) we first argue (Lemma 7) that the fourth moment is reasonably bounded after a sufficient number of reconstruction steps, and (2) give a multilevel tradeoff analysis showing that the kurtosis becomes large only when the variance of $\mathbf{E}[X|Y]$ shrinks significantly (i.e. information is destroyed). Combining these ideas, we are able to show that any reconstruction algorithm on an alphabet of size L which reconstructs all the way to the critical threshold would have to induce a distribution on $\mathbf{E}[X|Y]$ which is arbitrarily close to Gaussian — but this is impossible for a distribution supported on L atoms.

Preliminaries: Given an equiprobable $X \in \{\pm 1\}$ and an arbitrary random variable Y, denote the posterior mean by

$$S_X(Y) := \mathbf{E}[X|Y] = \mathbf{Pr}(X = 1|Y) - \mathbf{Pr}(X = -1|Y).$$

We remark that $S_X(Y)$ can be viewed as a discrete analogue of the score function in the estimation literature. Note that the probability of correctly reconstructing X based on Y equals $\frac{1}{2}\,\mathbf{E}\left[|S_X(Y)|\right] + \frac{1}{2}$. The χ^2-mutual information between X and Y equals $I_2(X;Y) := \mathbf{E}[S_X^2(Y)]$. Since $S_X(Y)$ is bounded in $[-1,1]$, we have $\mathbf{E}[S_X^2(Y)] \le \mathbf{E}[|S_X(Y)|] \le \mathbf{E}^{1/2}[S_X^2(Y)]$, so that as in Bleher et al. (1995); Evans et al. (2000), the problem of solvability is reduced to bounding the χ^2-mutual information.

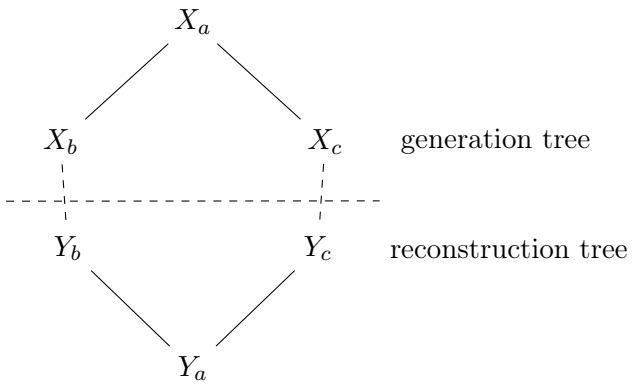

Figure 1: In this example, $S_{n-1} = S_{X_b}(Y_b)$, $S_n = S_{X_a}(Y_a)$, $\hat{S}_n = S_{X_a}(Y_b, Y_c)$, and $\bar{S}_n$ is an approximation of $\hat{S}_n$.

Next, we define several key quantities in the proof of Theorem 3. For notational simplicity, we assume the tree is 2-ary ($d = 2$) throughout the proofs, noting that the result for general d follows from the same argument. Figure 1 depicts parts of the generation tree and the mirroring reconstruction tree for node a with children b and c. Note that X_a, X_b, X_c denote binary labels and Y_a, Y_b, Y_c denote reconstruction messages with values in Σ. Within any height h tree (that is, from the root to a leaf there are h edges) and for any $n \le h$, consider the following random variables:

- S_n: defined as $S_{X_a}(Y_a)$ where X_a is the height n node in the generation tree and Y_a is the mirroring height n node in the reconstruction tree (recall (2)).

- $\hat{S}_n$: defined as $S_{X_a}(Y_b, Y_c)$ where Y_b and Y_c are the children of Y_a in the reconstruction tree. Note that S_n is a conditional expectation of $\hat{S}_n$ (induced by applying f_a):

$$S_{X_a}(Y_a) = \mathbf{E}[X_a|Y_a] = \mathbf{E}[\mathbf{E}[X_a|Y_{a,b,c}]|Y_a] = \mathbf{E}[\mathbf{E}[X|Y_{b,c}]|Y_a] = \mathbf{E}[S_X(Y_b, Y_c)|Y_a].$$

- $\bar{S}_n$: defined to be equal in distribution to $(1-2\varepsilon)(S_{n-1}+S'_{n-1})$, where S'_{n-1} is an independent copy of S_{n-1}. Since $S_{X_a}(Y_b) = (1-\varepsilon)S_{X_b}(Y_b) - \varepsilon S_{X_b}(Y_b) = (1-2\varepsilon)S_{X_b}(Y_b)$, one may view $\bar{S}_n$ as an idealized, decoupled version of $\hat{S}_n$. As will be shown in Lemma 5, $\hat{S}_n$ and $\bar{S}_n$ are close in Wasserstein distance, which will allow us to carry out our analysis with the simpler quantity $\bar{S}_n$.

Note that the subscript n in all the random variables above denotes the height of the node in the generative tree (not the reconstruction tree). Moreover, let σ_n^2, $\hat{\sigma}_n^2$, $\bar{\sigma}_n^2$, μ_n, $\hat{\mu}_n$, $\bar{\mu}_n$ be the second and the fourth moments of these random variables. The key of the proof is to study the evolution of the sequence $S_0 \to \bar{S}_1 \to \hat{S}_1 \to S_1 \to \dots$.

For any generative tree with height h and any integer[2] L, define

$$\xi_h := \sup \mathbf{E}[I_2(X;Y)], \tag{5}$$

where $X = X_\rho$ is the binary label on the root, $Y = Y_\rho$ is the final reconstruction, and the supremum is over all (possibly randomized) recursive reconstruction algorithms with memory $\log L$. Here, we assume that the reconstruction functions at the same level are the same, but they are allowed to vary across levels. Similarly, for randomized algorithms, we assume that the distribution of the random reconstruction function at a node depends only on its level.

Since any randomized reconstruction algorithm for an $h+1$-level tree can be simulated by one for an h-level tree with the same memory, it follows that ξ_h monotonically decreases in h. Let $\xi := \xi(\varepsilon, L)$ be the limit (which exists by monotonicity). Let $\varepsilon_c \in (0, 1/2)$ be the supremum ε for which $\xi(\varepsilon, \infty) > 0$. As mentioned before, $d(1-2\varepsilon_c)^2 = 1$ is the KS bound. We now proceed to detail the steps in the proof of our main result:

Wasserstein approximation lemmas: Recall that the idea of the converse proof is to show that if ε is close to ε_c, then S_n must converge to Gaussian for good algorithms. This requires showing that S_n is close to an i.i.d. sum. While $\bar{S}_n$ is a bona fide i.i.d. sum (of d independent copies of S_{n-1}, suitably scaled), the distribution of $\hat{S}_n$ in relation to S_{n-1} is more complicated. However, it is possible to show that $\bar{S}_n$ and $\hat{S}_n$ are close near the critical threshold. The idea is that the Wasserstein distance between $\bar{S}$ and $\hat{S}$ is small compared to their moments. A more general version of the following lemma is stated and proved in the appendices of the arXiv version.

Lemma 5 *With notation as above, we have that for any $n \in \{1, \dots, h-1\}$,*

$$W_2^2(\hat{S}_{n+1}, \bar{S}_{n+1}) \le \sigma_n^2 \alpha_2(\sigma_n^2),$$

where $\alpha_2(\cdot)$ is a function satisfying $\lim_{x\to 0^+} \alpha_2(x) = 0$. A similar statement holds for W_4^4 with a suitable function α_4, and with μ_n in place of σ_n^2.

2. We will also consider $L = \infty$, by which we mean there is no constraint on the alphabet size.

To use Lemma 5, we need to upper-bound the moments of the score. For the second moment, we begin with the useful observation that in a tree of height h, $\sigma_n^2 \le \xi_n$ for any $n \in \{1, 2, \dots, h\}$. The following result shows that, luckily, there is a simple upper bound on $\xi(\varepsilon, L)$ which does not depend on L, which in particular shows that the moments vanish when the noise is near the critical threshold.

Proposition 6 *Recall that we assumed that the degree $d = 2$. For any $\varepsilon \in [0,1]$, and $L \in \mathbb{N} \cup \{\infty\}$, either $\xi(\varepsilon, L) = 0$ or $\xi(\varepsilon, L) \le \omega(\epsilon)$, where we defined*

$$\omega(\varepsilon) := 4 - \frac{2}{(1-2\varepsilon)^2}. \tag{6}$$

In particular, Proposition 6 implies the KS bound for reconstruction (including that the reconstruction problem is not solvable *at* the threshold). The proof follows from analysis of the standard BP recursion and can be found in the appendices of the arXiv version; we note that a similar proof of the KS bound was previously discovered in Borgs et al. (2006).

Bounding the fourth moment: Wasserstein CLT results for i.i.d. sums are known under bounded fourth moment assumptions. The following bound on the fourth moment of S_n is derived from a recursive analysis which can be found in the appendices of the arXiv version.

Lemma 7 *There exists $\varepsilon_1 \in (0, \varepsilon_c)$ such that for any $\varepsilon \in (\varepsilon_1, \varepsilon_c)$, there exist $h_1 = h_1(\varepsilon, L), h_2 = h_2(\varepsilon, L)$ for which the following holds. For any tree of height $h \ge h_2$ and any reconstruction algorithm, we have either $\xi(\varepsilon, L) = 0$ or $\mu_n \le 13\xi^2(\varepsilon, L)$ at any level $n \in \{h_1, \dots, h\}$.*

Normality for good algorithms near the threshold: Given a real-valued random variable Z, let us define its Wasserstein non-Gaussianness by

$$\mathcal{E}(Z) := \inf_{\sigma > 0} W_2(Z, \mathbf{E}[Z] + \sigma G) \tag{7}$$

where G is the standard Gaussian random variable. The following lemma, exploiting the Wasserstein CLT, is proved in the appendices of the arXiv version.

Lemma 8 *There exists $\varepsilon_2 \in (0, \varepsilon_c)$ such that for any $\varepsilon \in (\varepsilon_2, \varepsilon_c)$, $L \in \{1, 2, \dots\} \cup \{\infty\}$, and $\delta \in (0, 1/2)$, either $\xi(\varepsilon, L) = 0$ or*

$$\lim_{h \to \infty} \sup_{\text{algorithms: } \sigma_h^2 \ge (1-\delta)\xi(\varepsilon, L)} \mathcal{E}(S_h) \le c_4 \sqrt{\xi(\varepsilon, L)} \left((\varepsilon_c - \varepsilon)^{1/13} + \sqrt{\delta \log \frac{1}{\varepsilon_c - \varepsilon}} \right),$$

where c_4 is an absolute constant.

We are finally in the position of proving the lower bound:

Proof [Proof of the lower bound in Theorem 3] Consider any $\varepsilon \in (\varepsilon_2, \varepsilon_c)$ and $L \in \{1, 2, \dots\}$, where ε_2 is as defined in Lemma 8. Choose any $\delta \in (0, 1/2)$. Note that for any h, there exists an algorithm such that $\sigma_h^2 \ge (1-\delta)\xi$. In the appendices of the arXiv version we show that $\mathcal{E}(S_h) \ge \frac{1}{2L}$. Comparing with this result we have $c_4 \left((\varepsilon_c - \varepsilon)^{1/13} + \sqrt{\delta \log \frac{1}{\varepsilon_c - \varepsilon}} \right) \ge \sqrt{1-\delta}/2L$. Taking $\delta \downarrow 0$ establishes that there exists an absolute constant c such that if $\xi(\varepsilon, L) > 0$ then $L \ge c(\varepsilon_c - \varepsilon)^{-1/13}$. ■

We conclude this section by noting that Corollary 4 is proved in the arXiv version, by combining the above with a simple argument to prove distributional convergence of BP.

References

Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Inference under information constraints I: Lower bounds from chi-square contraction. *arXiv preprint arXiv:1812.11476*, 2018.

Rudolf Ahlswede and Imre Csiszár. Hypothesis testing with communication constraints. *IEEE Transactions on Information Theory*, 32(4), 1986.

Leighton Pate Barnes, Yanjun Han, and Ayfer Özgür. A geometric characterization of fisher information from quantized samples with applications to distributed statistical estimation. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 16–23, 2018.

Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57 (2):764–785, 2011.

Noam Berger, Claire Kenyon, Elchanan Mossel, and Yuval Peres. Glauber dynamics on trees and hyperbolic graphs. *Probability Theory and Related Fields*, 131(3):311–340, 2005.

Quentin Berthet, Philippe Rigollet, et al. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013.

Pavel M. Bleher, Jean Ruiz, and Valentin A. Zagrebnov. On the purity of the limiting Gibbs state for the Ising model on the Bethe lattice. *J. Statist. Phys.*, 79(1-2):473–482, 1995.

Christian Borgs, Jennifer Chayes, Elchanan Mossel, and Sébastien Roch. The kesten-stigum reconstruction bound is tight for roughly symmetric binary channels. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 518–530. IEEE, 2006.

Mark Braverman, Ankit Garg, Tengyu Ma, Huy L. Nguyen, and David P Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1011–1020. ACM, 2016.

Matthew Brennan, Guy Bresler, and Wasim Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. In *Conference on Learning Theory*, pages 48–166, 2018.

Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. Optimal phylogenetic reconstruction. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 159–168. ACM, 2006.

Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Inference and phase transitions in the detection of modules in sparse networks. *Physical Review Letters*, 107(6):065701, 2011.

William Evans, Claire Kenyon, Yuval Peres, and Leonard J. Schulman. Broadcasting on trees and the ising model. *Annals of Applied Probability*, pages 410–433, 2000.

Sumegha Garg, Ran Raz, and Avishay Tal. Extractor-based time-space lower bounds for learning. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 990–1002. ACM, 2018.

Uri Hadar, Jingbo Liu, Yury Polyanskiy, and Ofer Shayevitz. Communication complexity of estimating correlations. In *Proceedings of the 51st ACM Symp. on Theory of Comp. (STOC)*, 2019.

Yanjun Han, Ayfer Özgür, and Tsachy Weissman. Geometric lower bounds for distributed parameter estimation under communication constraints. In *Conference On Learning Theory*, pages 3163–3188, 2018.

Dmitry Ioffe. Extremality of the disordered state for the Ising model on general trees. In *Trees (Versailles, 1995)*, volume 40 of *Progr. Probab.*, pages 3–14. Birkhäuser, Basel, 1996.

Harry Kesten and Bernt P. Stigum. Additional limit theorems for indecomposable multidimensional Galton-Watson processes. *Ann. Math. Statist.*, 37:1463–1481, 1966.

Harry Kesten and Bernt P. Stigum. Limit theorems for decomposable multi-dimensional Galton-Watson processes. *J. Math. Anal. Appl.*, 17:309–338, 1967.

Gillat Kol, Ran Raz, and Avishay Tal. Time-space hardness of learning sparse parities. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1067–1080. ACM, 2017.

Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

Jingbo Liu, Paul Cuff, and Sergio Verdú. Key capacity with limited one-way communication for product sources. In *Proceedings of the 2014 IEEE International Symposium on Information Theory*, pages 1146–1150, 2014.

Jingbo Liu, Paul Cuff, and Sergio Verdú. Secret key generation with one communicator and a one-shot converse via hypercontractivity. In *Proceedings of the 2015 IEEE International Symposium on Information Theory (ISIT)*, pages 710–714, 2015.

Jingbo Liu, Paul Cuff, and Sergio Verdú. Secret key generation with limited interaction. *IEEE Transactions on Information Theory*, 63(11):7358–7381, 2017.

Russell Lyons. The Ising model and percolation on trees and tree-like graphs. *Communications in Mathematical Physics*, 125(2):337–353, 1989.

Zongming Ma and Yihong Wu. Computational barriers in minimax submatrix detection. *The Annals of Statistics*, 43(3):1089–1116, 2015.

Laurent Massoulie. Community detection thresholds and the weak ramanujan property. arXiv preprint arXiv:1311.3085, 2013.

Marc Mézard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

Ankur Moitra, Elchanan Mossel, and Colin Sandon. The circuit complexity of inference. *arXiv preprint arXiv:1904.05483*, 2019.

Andrea Montanari. Tight bounds for ldpc and ldgm codes under map decoding. *IEEE Transactions on Information Theory*, 51(9):3221–3246, 2005.

Dana Moshkovitz and Michal Moshkovitz. Mixing implies lower bounds for space bounded learning. In *Conference on Learning Theory*, pages 1516–1566, 2017.

Elchanan Mossel. Recursive reconstruction on periodic trees. *Random Structures & Algorithms*, 13(1):81–97, 1998.

Elchanan Mossel. Phase transitions in phylogeny. *Transactions of the American Mathematical Society*, 356(6):2379–2404, 2004a.

Elchanan Mossel. Survey-information flow on trees. *DIMACS series in discrete mathematics and theoretical computer science*, 63:155–170, 2004b.

Elchanan Mossel, Joe Neeman, and Allan Sly. Belief propagation, robust reconstruction and optimal recovery of block models. In *Conference on Learning Theory*, pages 356–370, 2014.

Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3-4):431–461, 2015.

Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *Combinatorica*, 38(3):665–708, 2018.

Judea Pearl. *Probabilistic reasoning in intelligent systems*. Morgan Kaufman, San Mateo, 1988.

Robin Pemantle, Yuval Peres, et al. The critical ising model on trees, concave recursions and nonlinear capacity. *The Annals of Probability*, 38(1):184–206, 2010.

Ran Raz. A time-space lower bound for a large class of learning problems. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, pages 732–742. IEEE, 2017.

Ran Raz. Fast learning requires good memory: A time-space lower bound for parity learning. *Journal of the ACM (JACM)*, 66(1):3, 2018.

Thomas J. Richardson and Rüdiger L Urbanke. The capacity of low-density parity-check codes under message-passing decoding. *IEEETIT: IEEE Transactions on Information Theory*, 47, 2001. URL citeseer.ist.psu.edu/richardson98capacity.html.

Mike Steel. My Favourite Conjecture. http://www.math.canterbury.ac.nz/∼mathmas/conjecture.pdf, 2001.

Cédric Villani. *Topics in optimal transportation*, volume 338. Springer Science & Business Media, 2008.

Aolin Xu and Maxim Raginsky. Information-theoretic lower bounds for distributed function computation. *IEEE Transactions on Information Theory*, 63(4):2314–2337, 2017.

Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

Yuchen Zhang, John Duchi, Michael I. Jordan, and Martin J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *In Advances in Neural Information Processing Systems*, pages 2328–2336, 2013.

Yuchen Zhang, Martin J Wainwright, and Michael I. Jordan. Optimal prediction for sparse linear models? lower bounds for coordinate-separable m-estimators. *Electronic Journal of Statistics*, 11(1):752–799, 2017.