The Annals of Statistics
2023, Vol. 51, No. 1, 179–210
https://doi.org/10.1214/22-AOS2246

SHARP GLOBAL CONVERGENCE GUARANTEES FOR ITERATIVE NONCONVEX OPTIMIZATION WITH RANDOM DATA

BY KABIR ALADIN CHANDRASEKHER[1,a], ASHWIN PANANJADY[2,b] AND CHRISTOS THRAMPOULIDIS[3,c]

[1]*Department of Electrical Engineering, Stanford University,* [a]*kabirc@stanford.edu*
[2]*Schools of Industrial & Systems Engineering and Electrical & Computer Engineering, Georgia Tech,* [b]*ashwinpm@gatech.edu*
[3]*Department of Electrical & Computer Engineering, University of British Columbia,* [c]*cthrampo@ece.ubc.ca*

We consider a general class of regression models with normally distributed covariates, and the associated nonconvex problem of fitting these models from data. We develop a general recipe for analyzing the convergence of iterative algorithms for this task from a random initialization. In particular, provided each iteration can be written as the solution to a convex optimization problem satisfying some natural conditions, we leverage Gaussian comparison theorems to derive a deterministic sequence that provides sharp upper and lower bounds on the error of the algorithm with sample splitting. Crucially, this deterministic sequence accurately captures both the convergence rate of the algorithm and the eventual error floor in the finite-sample regime, and is distinct from the commonly used "population" sequence that results from taking the infinite-sample limit. We apply our general framework to derive several concrete consequences for parameter estimation in popular statistical models including phase retrieval and mixtures of regressions. Provided the sample size scales near linearly in the dimension, we show sharp global convergence rates for both higher-order algorithms based on alternating updates and first-order algorithms based on subgradient descent. These corollaries, in turn, reveal multiple nonstandard phenomena that are then corroborated by extensive numerical experiments.

1. Introduction. In many modern statistical estimation problems involving nonlinear observations, latent variables, or missing data, the log-likelihood—when viewed as a function of the parameters of interest—is nonconcave. Accordingly, even though the maximum likelihood estimator enjoys favorable statistical properties, the more practically relevant question is one at the intersection of statistics and optimization: Can we optimize the likelihood in a computationally efficient manner to produce statistically useful estimates? This question is particularly interesting in the statistically relevant setting—in which data are drawn i.i.d. from a suitably "nice" distribution—where the resulting random ensembles of optimization problems are often amenable to iterative algorithms. At the same time, iterative algorithms run on nonconvex model-fitting problems with random data often exhibit several behaviors that are distinct from those observed in standard convex programming (see the monographs [16, 18, 43, 100] for examples). Obtaining sharp upper and lower bounds on the error of such iterative algorithms is of fundamental interest, since this enables a rigorous comparison between families of procedures and guides algorithm and hyperparameter choices in practice.

Existing, general-purpose methods for assessing rates of convergence in nonconvex optimization involve either comparing upper bounds with upper bounds (e.g., [17, 63]) or using "population-based" analyses that consider the algorithm's behavior (or possibly the landscape of the random loss function) in the infinite-sample limit (e.g., [5, 58]). For example, analyses

Received December 2021; revised September 2022.
MSC2020 subject classifications. Primary 62J02, 90C06; secondary 90C26.
Key words and phrases. Nonconvex optimization, convergence rate, precise iterate-by-iterate prediction.

that proceed via the *population update* are based on the following intuition. The algorithm can be viewed as successively applying a (random) data-dependent operator $\mathcal{T}_n$ over iterations, where the point $\mathcal{T}_n(\boldsymbol{\theta})$ is obtained upon running one iteration of the algorithm from the "current" parameter $\boldsymbol{\theta}$. The limiting, deterministic object $\lim_{n\to\infty} \mathcal{T}_n$ is the population update, and its evolution over time ought to serve as a proxy for the random iterates. Indeed, the overall style of population-based analysis is appealing for several reasons: (a) It applies (in principle) to any iterative algorithm run on any model-fitting problem, and (b) In contrast to the direct sample-based approach of handling the algorithmic iterates directly (see, e.g., some of the early papers [44, 53]), it does not require the analysis of a complex recursion involving highly nonlinear functions of the random data. In addition, decomposing the analysis into a deterministic, optimization-theoretic component applied to the population version of the algorithm and a stochastic component that captures the eventual statistical neighborhood of convergence provides a natural two-step approach. But does the population-based analysis provide a reliable prediction of convergence behavior in modern high-dimensional settings in which the number of unknown parameters is typically comparable to the sample size?

1.1. *Motivation*: *Accurate predictions of convergence behavior.* Toward answering the question posed above, we run a simulation on what is arguably the simplest nonlinear statistical model resulting in a nonconvex fitting problem: *phase retrieval* with a real signal. This is a regression model in which a scalar response y is related to a d-dimensional covariate $\boldsymbol{x}$ via $\mathbb{E}[y|\boldsymbol{x}] = |\langle \boldsymbol{x}, \boldsymbol{\theta}^* \rangle|$, and the task is to estimate $\boldsymbol{\theta}^* \in \mathbb{R}^d$ from i.i.d. observations $(\boldsymbol{x}_i, y_i)$. Two popular algorithms to optimize the corresponding noncave log-likelihood are alternating minimization (AM) [30, 32] and subgradient descent (GD) [98] (see Section 3 for details).

Before running our simulation, we emphasize two key aspects of it that form recurrent themes throughout. First, as is common in the literature [17, 38, 40, 51, 66], we assume that the covariates are normally distributed, and additionally employ a sample-splitting device: each iteration of the algorithm is executed using n fresh observations of the model, drawn independently of past iterations. The sample-splitting device has been used extensively in the analysis of iterative algorithms as a simplifying assumption (e.g., [19, 40, 44, 51, 63, 66]), and forms a natural starting point for our investigations. Second, over and above tracking the ℓ_2 error of parameter estimation, we track a more expressive statistic over iterations. In particular, we associate each parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ with a two-dimensional *state*

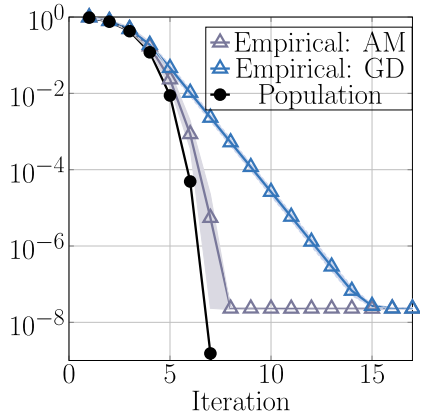
$$\alpha(\boldsymbol{\theta}) = \|\boldsymbol{P}_{\boldsymbol{\theta}^*}\boldsymbol{\theta}\|_2 \quad \text{and} \quad \beta(\boldsymbol{\theta}) = \left\|\boldsymbol{P}^{\perp}_{\boldsymbol{\theta}^*}\boldsymbol{\theta}\right\|_2, \tag{1}$$

where $\boldsymbol{P}_{\boldsymbol{\theta}^*}$ denotes the projection matrix onto the one-dimensional subspace spanned by $\boldsymbol{\theta}^*$ and $\boldsymbol{P}^{\perp}_{\boldsymbol{\theta}^*}$ denotes the projection matrix onto the orthogonal complement of this subspace. In words, these two scalars measure the component of $\boldsymbol{\theta}$ parallel to $\boldsymbol{\theta}^*$ and perpendicular to $\boldsymbol{\theta}^*$, respectively. Iterates $\boldsymbol{\theta}_t$ of the algorithm then give rise to a two-dimensional *state evolution*[1] (α_t, β_t), where $\alpha_t = \alpha(\boldsymbol{\theta}_t)$ and $\beta_t = \beta(\boldsymbol{\theta}_t)$. As several papers in this space have pointed out [17, 81, 92], tracking the state evolution instead of the evolving d-dimensional parameter provides a useful summary statistic of the algorithm's behavior, and natural losses such as the ℓ_2 or angular loss of parameter estimation can be expressed in terms of the state evolution.

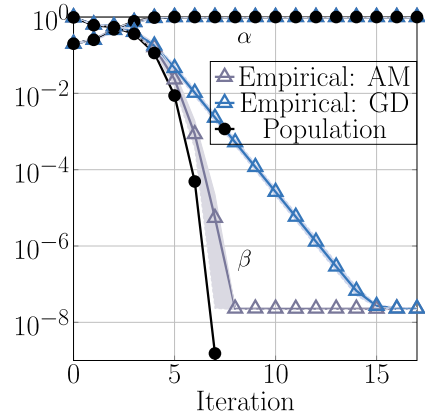
To run our first simulation in Figure 1, we choose the stepsize η in subgradient descent to ensure that its population update, that is, the population-based prediction of the next iteration from any current point coincides with that of alternating minimization. We plot the empirical error trajectories of both algorithms alongside that of the population.[2] Two conclusions are

[1]The state evolution terminology originated in the AMP literature [6, 25] and has been subsequently used more broadly in the analysis of nonconvex iterative algorithms (see, e.g., [17]). We adopt here the terminology in its broader context.

[2]See Sections 2 and 3 for the setting and Section 3 for the explicit population update.



(a) The ℓ_2 error of the population update vs. ℓ_2 error of the empirical trajectory.



(b) State evolution for the population update and empirical trajectory.

FIG. 1. *Estimation error* $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_2$ *over iterations along with population prediction for phase retrieval with* $n = 12{,}000$, $d = 600$ *and Gaussian noise with standard deviation* 10^{-8}. *Shaded envelopes around empirical curves denote* 95% *confidence bands over* 100 *independent trials*.

immediate from Figure 1. First, the population update is overly optimistic when predicting the convergence behavior of both algorithms. Second, algorithms with the same population update can exhibit very different convergence behaviors. As our simple experiment demonstrates, the population update is not, at least in general, a reliable predictor of convergence behavior. In Figure 6(a) in Section 4, we exhibit more drastic situations in which the population update can predict convergence when the empirical trajectory fails to converge. The underlying reason is simply that the problem is high-dimensional: it is too simplistic to hope for the algorithm's finite-sample behavior to resemble the case when the sample size goes to infinity. This observation leads to the principal question that we answer in this paper:

> *Is there a more faithful deterministic prediction for the empirical behavior of iterative algorithms in high dimensions?*

To be more specific, we would like such a deterministic update to satisfy two important desiderata. First and foremost, we should be able to accurately predict the error of parameter estimation after running one step of the algorithm from any point, allowing us to distinguish convergence from the lack thereof. Second, we desire *sharp* predictions of convergence behavior that differentiate, for instance, between linear and superlinear convergence. Such a sharp prediction for the iteration complexity can be used in conjunction with the per-step computational cost of the algorithm to rigorously guide the choice of the fastest procedure to implement for an optimization problem with random data.

1.2. *A glimpse of our contribution.* The principal contribution of this paper is to introduce a deterministic *Gordon state evolution update* that produces a sharp, deterministic prediction of the next state $(\alpha(\mathcal{T}_n(\theta)), \beta(\mathcal{T}_n(\theta)))$ as a function of the current state $(\alpha(\theta), \beta(\theta))$. This update satisfies the desiderata laid out above, and we develop a recipe that uses it to sharply analyze iterative algorithms. The Gordon update is derived using the machinery of Gaussian comparison inequalities—in particular, the convex Gaussian minmax theorem (CGMT) [86]. Despite the nonconvexity of the original problem,[3] this update applies provided each iteration of the algorithm can be written as the solution to a convex program

[3]In addition to nonconvex problems, our recipe can also provide sharp convergence guarantees for iterative convex optimization with random data (see, e.g., [2]).

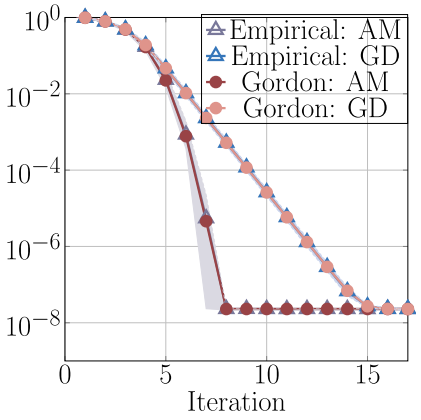
satisfying some mild assumptions. As a consequence of this generality, our recipe using the Gordon state evolution update yields several consequences for iterative algorithms when run on statistical models.

1. *One-step prediction for general nonlinear regression models with latent variables:* We consider a general class of regression models with Gaussian covariates (to be introduced precisely in Section 3), and derive a one-step prediction using the Gordon update for two general families of iterative algorithms. Let us highlight one consequence of this general characterization. The Gordon update is distinct from the population update in that it involves an additive correction term, which is nonzero in the high-dimensional setting. In particular, letting $\Lambda = n/d > 1$ denote the oversampling ratio used to implement each step of the algorithm, the perpendicular component of the Gordon update takes the form

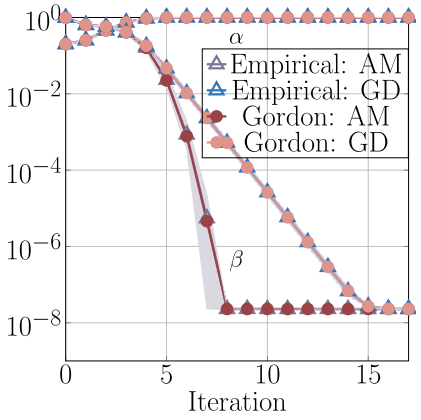
$$\overline{\beta}_{t+1} = \beta_{t+1}^{\mathsf{pop}} + \mathcal{O}(\Lambda^{-1/2}) \cdot \Delta_t, \tag{2}$$

where $\beta_{t+1}^{\mathsf{pop}}$ is the analogous prediction of the population update, and Δ_t is some nonnegative, algorithm-dependent scalar depending on the iterate at time t. Note that taking Λ to infinity, the predictions from the Gordon and population updates coincide. However, as we make clear shortly, the behavior of the Gordon update is dominated by the term $\mathcal{O}(\Lambda^{-1/2}) \cdot \Delta_t$ in high-dimensional settings, and in these scenarios the population update is a poor predictor of convergence behavior.

   Our recipe provides not only a deterministic update but also a finite-sample concentration bound, showing that the empirical error concentrates sharply around the prediction of the Gordon update. This is illustrated for the phase retrieval simulations in Figure 2. As is clear from these figures, the Gordon update does indeed satisfy the desiderata laid out above, providing both a sharp prediction of convergence behavior and near-exact predictions of the eventual error floor. In addition, our nonasymptotic characterization allows us to analyze algorithms from a random initialization. These results—when combined with a refined convergence analysis technique—allow us to uncover several nonstandard phenomena in popular statistical models, discussed next.
2. *Results for concrete models:* We use our one-step characterization to derive global convergence guarantees (i.e., from a random initialization) for both higher-order and first-order



(a) The ℓ_2 error of the Gordon update vs. ℓ_2 error of the empirical trajectory.



(b) State evolution: α and β components (1) for the Gordon update and the empirical trajectory.

FIG. 2. *Plots of behavior over iterations for both alternating minimization and subgradient descent with stepsize* $1/2$, *in a simulation identical to that of Figure* 1. *Overlaid is the prediction from their respective Gordon state evolution updates. As is evident from the occlusion of the triangle markers, the Gordon prediction exactly tracks behavior in both cases.*

TABLE 1

Summary of results for specific models and algorithms. In all cases, we provide global convergence guarantees showing that with high probability, convergence to the local neighborhood of the ground-truth parameter occurs after a number of iterations that is logarithmic in the dimension. Convergence rates within this neighborhood, as predicted by the Gordon update, are listed above. These exactly match empirical behavior in all cases

Algorithm	Model	Metric	Local convergence rate
Alternating minimization	Phase retrieval	ℓ_2	Superlinear, exponent $3/2$
Subgradient descent	Phase retrieval	ℓ_2	Linear
Alternating minimization	Mixture of regressions	Angular	Linear
Subgradient AM	Mixture of regressions	Angular	Linear

algorithms in two statistical models: phase retrieval and mixtures of regressions. Some salient takeaways are collected in Table 1.

To summarize, for the phase retrieval model, our primary contribution is to make quantitative the behavior observed in Figures 1 and 2. While the population update predicts quadratic convergence (i.e., superlinear convergence with exponent 2), we show that both alternating minimization and subgradient descent behave differently from this prediction. The former algorithm does converge superlinearly but with a nonstandard exponent $3/2$, while the latter converges linearly at best. For the mixture of regressions model, we propose a first-order method termed *subgradient AM*, which is inspired by the closely related gradient EM update [22, 62]. We study it alongside alternating minimization, and show that while both algorithms exhibit linear convergence in the angular metric, they are inconsistent in the ℓ_2 metric for any nonzero noise level. We exhibit regimes in which the first-order method is competitive (in terms of its iteration complexity) with its higher-order counterpart, suggesting that the first-order method should be preferred in these regimes given its lower per-iteration cost.

3. *Techniques of independent interest:* Over the course of proving our results, we develop some techniques that may be of broader interest, three of which we highlight below.

- In proving finite-sample concentration bounds around the deterministic Gordon updates, we handle a family of loss functions that is strictly more general than those used for proving analogous results in linear models [59, 67]. Our techniques are based on arguing about carefully chosen growth properties of these loss functions, and may prove useful in other nonasymptotic instantiations of the CGMT machinery.
- Characterizing algorithmic behavior near a random initialization requires a sharper bound on the deviation of the parallel component than what is provided by the general technique alluded to above. We develop a refined bound—applicable to higher-order updates that involve a matrix inversion in each iteration—by using a leave-one-out device. This characterization allows us to replace a polylogarithmic factor in the sample complexity bound with a doubly-iterated logarithm, and the technique may prove more broadly useful in analyzing other higher-order updates from a random initialization.
- Finally, our local convergence analysis for particular algorithms relies on a first-order expansion of the Gordon update. In particular, we show that the Gordon update is contractive in a local neighborhood of the ground truth $\boldsymbol{\theta}^*$, and combine this structural characterization with our refined concentration bounds on the sample state evolution to show deterministic upper *and* lower bounds on, that is, a high-probability envelope around, the error of the empirical trajectory of the algorithm. Such a technique may prove more broadly useful in producing sharp characterizations of convergence behavior for other classes of iterative algorithms and statistical models.

1.3. *Related work.* The literature on nonconvex optimization in statistical settings is vast, and we cannot hope to cover all of it here. We refer the reader to a few recent monographs [16, 18, 43, 100] for surveys, and the webpage [76] for an ever-expanding list of relevant references. We focus in this subsection on describing a few papers that are most closely related to our contributions, categorized for convenience under three broad headings.

Predictions in random optimization problems. As alluded to before, the population update has proven useful in analyzing many algorithms in a variety of settings including Gaussian mixture models [5, 19, 93], mixtures of regressions [5, 49, 51], phase retrieval [17], mixtures of experts [57] and neural networks [88]. In addition to providing local convergence guarantees, it has enabled researchers to study the more challenging setting with random initialization [17, 28, 92], and also revealed several surprising phenomena related to overparameterization and stability [41, 94]. The Gordon update that we derive is a much sharper deterministic predictor of convergence behavior than its population counterpart, and we hope that other surprising phenomena—over and above those that we present in the current paper—can be uncovered by making use of it.

In addition to papers that characterize the random loss landscape by utilizing properties of the population loss (e.g., [20, 39, 58]), we mention another line of inquiry—rooted in the literature on statistical physics—that leads to deterministic predictions. This framework is especially appealing when a prior on the underlying parameter is assumed, and employs the approximate message passing (AMP) algorithm [6, 25, 26, 60]. AMP is carefully designed to satisfy certain (approximate) independence properties across iterates and leads to a simple state evolution without sample splitting; see the recent tutorial [29] for an introduction. The analysis framework has recently been used to explore the (sub)optimality of first-order methods in terms of their eventual parameter estimation error [10], to predict computational barriers in a variety of problems including phase retrieval in high dimensions [56] and to demonstrate that logistic regression is biased in high dimensions [78]. We emphasize that analyses involving AMP do not require a sample-splitting assumption. In the context of first- order methods, a recent preprint [8] that was made available after our own exactly analyzes the dynamics of gradient flow—that is, a gradient descent algorithm with the stepsize tending to zero—for generalized linear models. In contrast to our motivation, predictions in this family are not designed with the dual goal of characterizing the (optimization-theoretic) rate of convergence of various algorithms as well as the statistical error of the eventual solution. Instead, they focus on producing a single algorithm that eventually attains statistical optimality—which is typically a member of the AMP family—or, in the case of the recent papers [8, 10], on writing down explicit state evolutions for the asymptotic correlation between iterates, from which optimization-theoretic rates of convergence may not be straightforward to derive.

Finally, we note that Oymak and Soltanolkotabi [65] focused on showing sharp time-data tradeofsf in linear inverse problems. In particular, they considered random design linear regression where the underlying parameter was constrained to an arbitrary (possibly nonconvex) set, and showed that employing projected gradient descent on the square loss with a particular choice of stepsize enjoys a linear rate of convergence to an order-optimal neighborhood of the true parameter. They also showed that a linear rate is the best achievable when the constraint set is convex. Follow-up work [66] obtained similar results for single-index model estimation, following the paradigm pioneered by Brillinger (see, e.g., [7, 69]). While these results are compelling, they are restricted to the analysis of a single algorithm, do not provide sharp iterate-by-iterate predictions and their primary focus is on exploiting structure in the underlying parameter. For comparison, and on the one hand, we do not explicitly model structure in the parameter of interest, and also require that each iteration of the algorithm solves a convex program. On the other hand, we allow for arbitrary nonlinear models,

and our machinery allows us to derive sharp tradeoffs applying to a broad class of iterative algorithms that go beyond first-order methods for linear regression.

Convergence guarantees for iterative algorithms beyond first-order updates. As made clear shortly, the Gordon state evolution recipe is particularly powerful when dealing with iterative algorithms that go beyond first-order updates, and consequently involve highly nonlinear functions of the random data. There are several "direct" analyses of such *higher-order* updates in the literature on matrix factorization, mixture models, neural networks, and index models, including for alternating projections [1, 33, 37, 40, 42, 44, 68, 77, 90, 97, 99], composite optimization [15, 27] and Gauss–Newton methods [31]. For the expectation maximization (EM) algorithm and its Newton (i.e., second-order) analog, the population update has been widely used to prove parameter estimation guarantees [5, 41, 93], although convergence in function value can be shown via other means [50, 95]. All of the analyses mentioned here are only able to provide upper bounds on the parameter estimation error over iterations, and we expect that employing our recipe in these settings would yield either matching lower bounds or sharper convergence rates.

Gordon's Gaussian comparison theorem in statistical models. Gordon proved his celebrated minmax theorem for doubly-indexed Gaussian processes in the 1980s [35, 36], which was later popularized in the statistical signal processing literature [11, 64, 70, 72]. Following a line of work [3, 67, 73–75], a sharp version of Gordon's result in the presence of convexity—providing both upper and lower bounds on the minmax value—was formalized in [86]; see [83] for broader historical context. Since then, the *convex* Gaussian minmax theorem (or CGMT for short) has been used to provide sharp performance guarantees for several convex programs with Gaussian data, including regularized M-estimators [85, 87], one-bit compressed sensing [84], regularized logistic regression [4, 24, 71, 79, 80], adversarial training for linear regression and classification [45, 46], max-margin linear classifiers [23, 48, 61], distributional characterization of minimum norm linear interpolators [14] and minimum ℓ_1 norm interpolation and boosting [52]. While this line of work typically uses the Gordon machinery to provide a one-step—and asymptotic—guarantee, the results of our paper are obtained by using the CGMT in each step of the iterative algorithm, which requires a nonasymptotic characterization. Having said that, we note that some nonasymptotic bounds have been obtained using the CGMT in the context of the LASSO [9, 59, 67], SLOPE [91] and a class of generalized linear models [54].

1.4. *General notation.* We use boldface small letters to denote vectors and boldface capital letters to denote matrices. We let $\mathsf{sgn}(v)$ denote the sign of a scalar v, with the convention that $\mathsf{sgn}(0) = 1$. We use $\mathsf{sgn}(\boldsymbol{v})$ to denote the sign function applied entrywise to a vector $\boldsymbol{v}$. Let $\mathbb{I}\{\cdot\}$ denote the indicator function. For $p \geq 1$, let $\mathbb{B}_p(\boldsymbol{v}; t) = \{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{v}\|_p \leq t\}$ denote the closed ℓ_p ball of radius t around a point $\boldsymbol{v}$, with the shorthand $\mathbb{B}_p(t) = \mathbb{B}_p(\boldsymbol{0}; t)$; the dimension will usually be clear from context. Analogously, let $\mathbb{B}_p(S; t) = \{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{v}\|_p \leq t \text{ for some } \boldsymbol{v} \in S\}$ denote the t-fattening of a set S in ℓ_p-norm. For an operator $\mathcal{A} : \mathbb{S} \to \mathbb{S}$, let $\mathcal{A}^t := \underbrace{\mathcal{A} \otimes \cdots \otimes \mathcal{A}}_{t \text{ times}}$ denote the operator obtained by t repeated applications of $\mathcal{A}$.

For two sequences of nonnegative reals $\{f_n\}_{n\geq 1}$ and $\{g_n\}_{n\geq 1}$, we use $f_n \lesssim g_n$ to indicate that there is a universal positive constant C such that $f_n \leq C g_n$ for all $n \geq 1$. The relation $f_n \gtrsim g_n$ indicates that $g_n \lesssim f_n$, and we say that $f_n \asymp g_n$ if both $f_n \lesssim g_n$ and $f_n \gtrsim g_n$ hold simultaneously. We also use standard-order notation $f_n = \mathcal{O}(g_n)$ to indicate that $f_n \lesssim g_n$ and $f_n = \widetilde{\mathcal{O}}(g_n)$ to indicate that $f_n \lesssim g_n \log^c n$, for a universal constant $c > 0$. We say that $f_n = \Omega(g_n)$ (resp., $f_n = \widetilde{\Omega}(g_n)$) if $g_n = \mathcal{O}(f_n)$ (resp., $g_n = \widetilde{\mathcal{O}}(f_n)$). The notation $f_n = o(g_n)$ is used when $\lim_{n\to\infty} f_n/g_n = 0$, and $f_n = \omega(g_n)$ when $g_n = o(f_n)$. Throughout, we use c,

C to denote universal positive constants, and their values may change from line to line. All logarithms are to the natural base unless otherwise stated.

We denote by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ a normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let $\mathsf{Unif}(S)$ denote the uniform distribution on a set S, where the distinction between a discrete and continuous distribution can be made from context. We say that $X \overset{(d)}{=} Y$ for two random variables X and Y that are equal in distribution. For $q \geq 1$ and a random variable X taking values in $\mathbb{R}^d$, we write $\|X\|_q = (\mathbb{E}[|X|^q])^{1/q}$ for its L^q norm. Finally, for a real valued random variable X and a strictly increasing convex function $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ satisfying $\psi(0) = 0$, we write $\|X\|_\psi = \inf\{t > 0 \mid \mathbb{E}[\psi(t^{-1}|X|)] \leq 1\}$ for its ψ-Orlicz norm. We make particular use of the ψ_q-Orlicz norm for $\psi_q(u) = \exp(|u|^q) - 1$. We say that X is sub-Gaussian if $\|X\|_{\psi_2}$ is finite and that X is subexponential if $\|X\|_{\psi_1}$ is finite.

2. General methodology. We begin with a high-level overview of the steps involved in our recipe. For concreteness, we focus on analyzing iterative algorithms on regression models—in which we observe covariate-response pairs $(\boldsymbol{x}_i, y_i)$ and the covariates $\boldsymbol{x}_i$ are drawn i.i.d. from a normal distribution. Our rigorous results in the next subsection—showing that the empirical iteration concentrates around an explicit Gordon prediction—are proved on a concrete class of regression models with latent variables.

2.1. *High-level sketch of the steps.* We begin with the ansatz—which will be proved rigorously when establishing the main results to follow—that it suffices to track the two-dimensional state evolution $(\alpha(\boldsymbol{\theta}), \beta(\boldsymbol{\theta}))$ defined in equation (1). In particular, when one step of the algorithm is run from the parameter $\boldsymbol{\theta}_t$ to obtain $\boldsymbol{\theta}_{t+1}$, we are interested in a *deterministic* prediction $(\overline{\alpha}_{t+1}, \overline{\beta}_{t+1})$ for the random pair $(\alpha(\boldsymbol{\theta}_{t+1}), \beta(\boldsymbol{\theta}_{t+1}))$ that is (a) a function only of the pair $(\alpha(\boldsymbol{\theta}_t), \beta(\boldsymbol{\theta}_t))$, and (b) accurate up to a small error. We use several steps to derive such a deterministic *state evolution update*. Let us begin by introducing the convex Gaussian minmax theorem, or CGMT, which forms the bedrock of our recipe.

PROPOSITION 1 (CGMT [86]). *Let $\boldsymbol{G}$ denote an $n \times d$ standard Gaussian random matrix, and let $\boldsymbol{\gamma}_d \in \mathbb{R}^d$ and $\boldsymbol{\gamma}_n \in \mathbb{R}^n$ denote standard Gaussian random vectors drawn independently of each other and of $\boldsymbol{G}$. Let $\boldsymbol{L} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ denote two fixed matrices. Also, let $\mathcal{U} \subseteq \mathbb{R}^d$ and $\mathcal{V} \subseteq \mathbb{R}^n$ denote compact sets, and let $Q : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$ denote a continuous function. Define*

$$P(\boldsymbol{G}) := \min_{\boldsymbol{u} \in \mathcal{U}} \max_{\boldsymbol{v} \in \mathcal{V}} \langle \boldsymbol{M}\boldsymbol{v}, \boldsymbol{G}\boldsymbol{L}\boldsymbol{u} \rangle + Q(\boldsymbol{u}, \boldsymbol{v}) \quad \textit{and} \tag{3a}$$

$$A(\boldsymbol{\gamma}_n, \boldsymbol{\gamma}_d) := \min_{\boldsymbol{u} \in \mathcal{U}} \max_{\boldsymbol{v} \in \mathcal{V}} \|\boldsymbol{M}\boldsymbol{v}\|_2 \cdot \langle \boldsymbol{\gamma}_d, \boldsymbol{L}\boldsymbol{u} \rangle + \|\boldsymbol{L}\boldsymbol{u}\|_2 \cdot \langle \boldsymbol{\gamma}_n, \boldsymbol{M}\boldsymbol{v} \rangle + Q(\boldsymbol{u}, \boldsymbol{v}). \tag{3b}$$

Then

(a) *For all $t \in \mathbb{R}$, we have*

$$\mathbb{P}\{P(\boldsymbol{G}) \leq t\} \leq 2\mathbb{P}\{A(\boldsymbol{\gamma}_n, \boldsymbol{\gamma}_d) \leq t\}.$$

(b) *If, in addition, the sets $\mathcal{U}$, $\mathcal{V}$ are convex and the function Q is convex–concave, then for all $t \in \mathbb{R}$, we have*

$$\mathbb{P}\{P(\boldsymbol{G}) \geq t\} \leq 2\mathbb{P}\{A(\boldsymbol{\gamma}_n, \boldsymbol{\gamma}_d) \geq t\}.$$

Strictly speaking, Proposition 1 is a generalization of the result appearing in [86], which is stated without the matrix pair $(\boldsymbol{L}, \boldsymbol{M})$. However, its proof follows identically, and we choose to state the more general result since it is most useful for our development. Following prior

terminology [86], we refer to equation (3a) as the *primary optimization problem*, and to equation (3b) as the *auxiliary optimization problem*. Having stated the CGMT, let us now provide a rough outline of the steps involved in deriving the Gordon state evolution update.

Step 1: Write one iteration of algorithm as solution to convex optimization problem. As alluded to in the Introduction, each iteration of most algorithms—even on nonconvex functions—can be written as the solution to a convex optimization problem. To make this explicit under our assumption of sample splitting, suppose that at each iteration, we form a *fresh* batch[4] of n observations by collecting the covariates in a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and the responses in a vector $\boldsymbol{y} \in \mathbb{R}^n$. By design, the pair $(\boldsymbol{X}, \boldsymbol{y})$ is statistically independent of the algorithmic iterates thus far. At iteration t, we update our current estimate of the parameter $\boldsymbol{\theta}_t$ to $\boldsymbol{\theta}_{t+1}$ by solving the optimization problem

$$\boldsymbol{\theta}_{t+1} \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\theta}_t, \boldsymbol{X}, \boldsymbol{y}), \tag{4}$$

for some loss function $\mathcal{L}$ that depends implicitly on the current point $\boldsymbol{\theta}_t$ and is formed using the data $(\boldsymbol{X}, \boldsymbol{y})$. In typical applications, $\mathcal{L}$ is convex in $\boldsymbol{\theta}$ for each fixed triple $(\boldsymbol{\theta}_t, \boldsymbol{X}, \boldsymbol{y})$.

Alternatively, and as alluded to in Section 1, each step of the algorithm can be viewed through the lens of a random, *empirical operator* $\mathcal{T}_n : \mathbb{R}^d \to \mathbb{R}^d$, with

$$\mathcal{T}_n(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta}'\in\mathbb{R}^d} \mathcal{L}\big(\boldsymbol{\theta}'; \boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}\big) \quad \text{for each } \boldsymbol{\theta} \in \mathbb{R}^d, \tag{5}$$

and $\boldsymbol{\theta}_{t+1} = \mathcal{T}_n(\boldsymbol{\theta}_t)$. The population update/operator alluded to before corresponds to the limiting object $\lim_{n\to\infty} \mathcal{T}_n$ (treating the dimension d as fixed), which is deterministic but an overly optimistic predictor of convergence behavior in high dimensions.

Step 2: Write equivalent auxiliary optimization problem. In this step, our goal is to write the minimization of the loss function $\mathcal{L}$—which is a function of the Gaussian design matrix $\boldsymbol{X}$—as a *simpler* minimization involving fewer Gaussian random variables. In particular, we would like to show that

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\theta}_t, \boldsymbol{X}, \boldsymbol{y}) \approx \min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathfrak{L}(\boldsymbol{\theta}; \boldsymbol{\theta}_t, \boldsymbol{\gamma}_d, \boldsymbol{\gamma}_n), \tag{6}$$

where $\boldsymbol{\gamma}_d, \boldsymbol{\gamma}_n$ denote (either d or n-dimensional) standard Gaussian *vectors* and the $\approx$ symbol denotes some form of approximate equality in distribution. The key workhorse in this step is the CGMT (Proposition 1).

Step 3: Scalarize to obtain deterministic Gordon state evolution update. Writing the optimization problem in terms of the objective $\mathfrak{L}$ is motivated by the fact that this objective can typically be *scalarized*, and equivalently written in terms of a small number of decision variables. In this step, our goal is to establish the low-dimensional representation

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathfrak{L}(\boldsymbol{\theta}; \boldsymbol{\theta}_t, \boldsymbol{\gamma}_d, \boldsymbol{\gamma}_n) \approx \min_{\boldsymbol{\xi}} \overline{L}_n(\boldsymbol{\xi}; \boldsymbol{\xi}_t), \tag{7}$$

as well as the approximate equivalence

$$\min_{\boldsymbol{\xi}} \overline{L}_n(\boldsymbol{\xi}; \boldsymbol{\xi}_t) \approx \min_{\boldsymbol{\xi}} \overline{L}(\boldsymbol{\xi}; \boldsymbol{\xi}_t, \Lambda), \tag{8}$$

where $\overline{L}_n(\cdot; \boldsymbol{\xi}_t) : \mathbb{R}^3 \to \mathbb{R}$ and $\overline{L}(\cdot; \boldsymbol{\xi}_t, \Lambda) : \mathbb{R}^3 \to \mathbb{R}$ are functions solely of a low (i.e., three) dimensional parameter, and moreover, depend on the previous iterate only through the three-dimensional parameter $\boldsymbol{\xi}_t$ that captures certain key properties of $\boldsymbol{\theta}_t$. We emphasize that the loss $\overline{L}_n$ is a random function, whereas $\overline{L}$ is a deterministic function, which depends explicitly

[4]Owing to sample splitting, the pair $(\boldsymbol{X}, \boldsymbol{y})$ can also be thought of as depending on the iteration number t, but we suppress this dependence and opt for more manageable notation.

on Λ. The minimizers of the RHS of (8)—along with some algebraic simplification—then yield the deterministic, two-dimensional Gordon state evolution update $(\overline{\alpha}_{t+1}, \overline{\beta}_{t+1})$.

Step 4: Argue that the empirical state evolution is tracked by the Gordon update. The final step is to use growth properties of the objective functions $\mathfrak{L}$ and $\overline{L}_n$ around their minima to show that if their optimum values coincide, then so must their optimizers, up to some negligible deviation. This is a technical step, and a large portion of our instantiation of the recipe is dedicated to establishing these properties.

2.2. *Model and main result.* Having described the recipe at a high level, we now formally derive and prove concentration of the one-step Gordon updates for higher-order and first-order methods run on a generic class of problems. To begin, let us set up a formal observation model, and provide a general form for the iterative algorithms that we study.

Suppose that we observe i.i.d. covariate-response pairs $(\boldsymbol{x}_i, y_i)$ generated according to the model[5]

$$y_i = f\big(\langle \boldsymbol{x}_i, \boldsymbol{\theta}^* \rangle; q_i\big) + \epsilon_i. \tag{9}$$

The covariates $\boldsymbol{x}_i$ are assumed to be d-dimensional and drawn i.i.d. from the standard normal distribution $\mathcal{N}(0, \boldsymbol{I}_d)$, and the function f is some known *link function*. The random variable $q_i \sim \mathbb{Q}$ represents a possible *latent variable*, that is, some source of auxiliary randomness that is unobserved, and ϵ_i represents additive noise drawn from the distribution $\mathcal{N}(0, \sigma^2)$; both of these are drawn i.i.d. Our goal is to use observations of pairs $(\boldsymbol{x}_i, y_i)_{i \geq 1}$ to estimate the unknown d-dimensional parameter $\boldsymbol{\theta}^*$. We assume that $\|\boldsymbol{\theta}^*\|_2 = 1$ in order to simplify statements of our theoretical results, but note that all our techniques extend to the more general case. We consider two classes of iterative algorithms run on observations from this model.

Higher-order methods: The class of higher-order methods that we consider typically involves running least squares in each iteration. These can be written in the form (4) by taking

$$\mathcal{L}(\boldsymbol{\theta}) := \sqrt{\frac{1}{n} \sum_{i=1}^{n} \big(\omega\big(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle, y_i\big) - \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle\big)^2}, \tag{10a}$$

where $\omega : \mathbb{R}^2 \to \mathbb{R}$ denotes a problem-dependent *weight* function and the square root is taken for convenience. The minimizer of the loss (10a) is given by

$$\boldsymbol{\theta}_{t+1} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1} \left(\sum_{i=1}^{n} \omega\big(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle, y_i\big) \cdot \boldsymbol{x}_i\right), \tag{10b}$$

and involves matrix inversion at each step. Several families of algorithms, including alternating projections and expectation maximization, can be written in this form [5, 97].

First-order methods: These take the form (4), with loss

$$\mathcal{L}(\boldsymbol{\theta}) := \frac{\|\boldsymbol{\theta}\|_2^2}{2} - \langle \boldsymbol{\theta}, \boldsymbol{\theta}_t \rangle + \frac{2\eta}{n} \sum_{i=1}^{n} \omega\big(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle, y_i\big) \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle. \tag{11a}$$

Here, η denotes a stepsize and ω is another problem-dependent weight function, typically distinct from the corresponding choice for higher-order updates in (10a). Note that the loss

[5]It is important to note that owing to our sample-splitting heuristic, the total sample size when the iterative algorithm is run for T iterations is given by $n \cdot T$. In the specific examples that we study, the number of iterations T required to obtain order-optimal parameter estimates will turn out to be at most logarithmic in the dimension, so that the total sample size nT also scales near linearly in the dimension d.

TABLE 2

Gordon state evolution updates for both classes of algorithms run on the model (9), where Ω is given by (13) and $\Lambda = n/d$. As justified in Section 3.1, setting $\eta = 1/2$ in the first-order updates and taking $\Lambda \to \infty$ yields the same update for both types of methods, which in turn coincides with the population prediction

	First-order	Higher-order
α^{gor}	$\alpha - 2\eta \cdot \mathbb{E}[Z_1\Omega]$	$\mathbb{E}[Z_1\Omega]$
β^{gor}	$\sqrt{(\beta - 2\eta \cdot \mathbb{E}[Z_2\Omega])^2 + \frac{4\eta^2}{\Lambda} \cdot \mathbb{E}[\Omega^2]}$	$\sqrt{(\mathbb{E}[Z_2\Omega])^2 + \frac{1}{\Lambda - 1}(\mathbb{E}[\Omega^2] - (\mathbb{E}[Z_1\Omega])^2 - (\mathbb{E}[Z_2\Omega])^2)}$

(11a) is clearly convex as required by Step 1 of our recipe. Further, its minimizers take the form

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \cdot \frac{2}{n} \sum_{i=1}^{n} \omega\big(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle, y_i\big) \cdot \boldsymbol{x}_i, \tag{11b}$$

which correspond to running a (sub)gradient algorithm.

Having described the model and class of algorithms that we study, we are now in a position to state our main result. Fix an *arbitrary* d-dimensional parameter $\boldsymbol{\theta}$ and consider the one-step update (5). For convenience, use the shorthand

$$(\alpha, \beta) = \big(\alpha(\boldsymbol{\theta}), \beta(\boldsymbol{\theta})\big) \quad \text{and} \quad \big(\alpha^+, \beta^+\big) = \big(\alpha\big(\mathcal{T}_n(\boldsymbol{\theta})\big), \beta\big(\mathcal{T}_n(\boldsymbol{\theta})\big)\big). \tag{12}$$

The main result of this section shows that for algorithms whose one-step updates take the form (10) or (11), the pair (α^+, β^+) concentrates around the deterministic Gordon state evolution update run from (α, β), denoted by $(\alpha^{\mathsf{gor}}, \beta^{\mathsf{gor}})$ and provided explicitly in Table 2. This result holds under mild assumptions on the weight functions $\omega : \mathbb{R}^2 \to \mathbb{R}$ used to define the algorithms. Recalling the model in (9), let $Q \sim \mathbb{Q}$ denote the latent variable. Let (Z_1, Z_2, Z_3) denote a triple of i.i.d. standard Gaussians, and define the random variable

$$\Omega = \omega\big(\alpha Z_1 + \beta Z_2, f(Z_1; Q) + \sigma Z_3\big). \tag{13}$$

The first assumption requires that this random variable is light-tailed (see, e.g., [89], Chapter 2). The second assumption requires a lower bound on a particular functional of Ω.

ASSUMPTION 1. The random variable Ω is sub-Gaussian with bounded Orlicz norm $\|\Omega\|_{\psi_2} \le K_1$, for some $K_1 > 0$.

ASSUMPTION 2. For a parameter $K_2 > 0$, we have $\mathbb{E}[\Omega^2] - (\mathbb{E}[Z_1\Omega])^2 - (\mathbb{E}[Z_2\Omega])^2 \ge K_2$.

We next state our main results characterizing the concentration of the random pair (α^+, β^+) around $(\alpha^{\mathsf{gor}}, \beta^{\mathsf{gor}})$. We state two very similar theorems for convenience since they apply under a slightly different set of assumptions. The first theorem applies to higher-order updates under both Assumptions 1 and 2, and the second theorem applies to first-order updates but requires only Assumption 1 to hold.

THEOREM 1 (Higher-order deterministic prediction). *Consider the general model (9) for the data, and procedures that obey the general one-step update (10). Recall the shorthand $(\alpha, \beta, \alpha^+, \beta^+)$ from equation (12). Suppose that Assumptions 1 and 2 hold on the associated weight function ω with parameters K_1 and K_2, respectively. Consider the pair of scalars $(\alpha^{\mathsf{gor}}, \beta^{\mathsf{gor}})$ for higher-order updates from Table 2. There exists a universal positive constant C_1 as well as a pair of positive constants (C_K, C'_K) depending only on the pair (K_1, K_2) such that the following is true. If $\Lambda \ge C_1$, then*

(a) *Provided we further have* $n \geq C'_K \cdot \log(1/\delta)$, *the perpendicular component satisfies*

$$\mathbb{P}\left\{|\beta^+ - \beta^{\mathsf{gor}}| \geq C_K\left(\frac{\log(1/\delta)}{n}\right)^{1/4}\right\} \leq \delta, \quad \textit{and} \tag{14a}$$

(b) *The parallel component satisfies*

$$\mathbb{P}\left\{|\alpha^+ - \alpha^{\mathsf{gor}}| \geq C_K\left(\frac{\log^7(1/\delta)}{n}\right)^{1/2}\right\} \leq \delta. \tag{14b}$$

For first-order methods, we make the additional assumption[6] $\alpha \vee \beta \leq 3/2$ and obtain a sharper rate via a more direct analysis. We also state the theorem for stepsize $\eta \leq 1/2$ for convenience.

THEOREM 2 (First-order deterministic prediction). *Consider the general model* (9) *for the data, and procedures that obey the general one-step update* (11) *for some* $\eta \leq 1/2$. *Recall the shorthand* $(\alpha, \beta, \alpha^+, \beta^+)$ *from equation* (12) *and assume that* $\alpha \vee \beta \leq 3/2$. *Suppose that Assumption* 1 *holds on the associated weight function* ω *with parameter* K_1. *Consider the pair of scalars* $(\alpha^{\mathsf{gor}}, \beta^{\mathsf{gor}})$ *for first-order updates from Table* 2. *There exists a universal positive constant* C_1 *as well as a pair of positive constants* (C_K, C'_K), *depending only on* K_1 *such that the following is true. If* $\Lambda \geq C_1$, *then*

(a) *Provided we further have* $n \geq C'_K \cdot \log(1/\delta)$, *the perpendicular component satisfies*

$$\mathbb{P}\left\{|\beta^+ - \beta^{\mathsf{gor}}| \geq C_K\left(\frac{\log(1/\delta)}{n}\right)^{1/2}\right\} \leq \delta, \quad \textit{and} \tag{15a}$$

(b) *The parallel component satisfies*

$$\mathbb{P}\left\{|\alpha^+ - \alpha^{\mathsf{gor}}| \geq C_K\left(\frac{\log(1/\delta)}{n}\right)^{1/2}\right\} \leq \delta. \tag{15b}$$

A few comments are in order. First, our formulas for β^{gor} in Table 2 make transparent the $\mathcal{O}(\Lambda^{-1/2}) \cdot \Delta_t$ term alluded to in equation (2). Indeed, the population update can be derived from these formulas by taking $\Lambda \to \infty$ (see Section 3 for explicit evaluations of these quantities). Second, we emphasize that Theorems 1 and 2 provide a *nonasymptotic* concentration result of the random state evolution around its deterministic counterpart, in contrast to results typically derived using the CGMT machinery, for example, [4, 24, 46, 48, 61, 71, 79]. While some recent nonasymptotic studies have been carried out for sparse linear regression [9, 59, 67, 91], our bounds are more general in that they apply under the general observation model (9), and not just to the overall minimizer of the empirical risk. A nonasymptotic characterization is essential for our purposes because we intend to apply these results iteratively, once per step of the algorithm. As mentioned earlier, this becomes particularly important near a random initialization of the algorithm, for which we have $\alpha \asymp d^{-1/2}$ (see, e.g., the Supplementary Material [13], Lemma 24). In this case, the predictions from Table 2 show that $\alpha^{\mathsf{gor}} \asymp d^{-1/2}$, and the deviation bound (14b) for α^+ of $\mathcal{O}(n^{-1/2})$ is crucial. In particular, even when the sample size scales linearly in the dimension (i.e., $\Lambda = \mathcal{O}(1)$), we can then show that the next iterate still retains nontrivial correlation with the ground truth with high probability, so that $\alpha^+ \geq \alpha$. The derivation of this sharp bound requires significant technical

[6]This assumption is not required for higher-order methods because the sub-Gaussianity of the ω function suffices to ensure that the pair $(\alpha^{\mathsf{gor}}, \beta^{\mathsf{gor}})$ remains bounded (see Table 2). The same is not true for first-order methods; as is evident from Table 2, we also require the pair (α, β) to be bounded.

effort over and above the general recipe presented in the previous section, which produces $\mathcal{O}(n^{-1/4})$ deviation bounds and would only be useful near a random initialization provided $n \gtrsim d^2$. Note that since our results only assume a constant lower bound on the oversampling ratio Λ, they hold in this regime as well. Theorems 1 and 2 are proved in the Supplementary Material [13], Sections 1 and 2.

As briefly alluded to before, we note that Theorem 2 does not require the Gordon state-evolution machinery and can instead be proved from first principles by directly using the update (11b). For completeness, we provide a proof with suboptimal deviation bounds for first-order methods using the Gordon state evolution machinery in [13], Section 1.3. The higher-order updates considered in Theorem 1, on the other hand, seem out of reach of such a direct method, and the machinery developed here allows to analyze both first as well as higher-order methods under a common framework. In addition, the Gordon state evolution machinery can—in principle—also be used to understand updates which do not even admit closed-form solutions such as the prox-linear method [15, 27] when applied on related statistical models.

3. Consequences for some concrete statistical models. In this section, we state consequences of our main results for two specific models and algorithms, although it is important to note that the Gordon recipe itself—as sketched in the previous section—is more broadly applicable. In particular, we will consider phase retrieval and a symmetric mixture of linear regressions, as well as the algorithms covered in Section 2. It is important to note that in both these models, the global sign of the parameter $\boldsymbol{\theta}^*$ is not identifiable from observations, and so parameter estimates should be assessed in terms of their "distance" to the set $\{-\boldsymbol{\theta}^*, \boldsymbol{\theta}^*\}$.

As mentioned before, we track the two-dimensional state $(\alpha(\boldsymbol{\theta}), \beta(\boldsymbol{\theta}))$ of each parameter $\boldsymbol{\theta} \in \mathbb{R}^d$, with $\alpha(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \boldsymbol{\theta}^* \rangle$ and $\beta(\boldsymbol{\theta}) = \|\boldsymbol{P}^{\perp}_{\boldsymbol{\theta}^*}\boldsymbol{\theta}\|_2$. The sign ambiguity will be resolved by the initialization, so we assume throughout that $\alpha(\boldsymbol{\theta}) \geq 0$ for parameters $\boldsymbol{\theta}$ that we consider. For any two-dimensional state evolution element $\boldsymbol{\zeta} = (\alpha, \beta)$, define two metrics

$$\mathsf{d}_{\ell_2}(\boldsymbol{\zeta}) := \sqrt{(1-\alpha)^2 + \beta^2} \quad \text{and} \quad \mathsf{d}_{\angle}(\boldsymbol{\zeta}) := \tan^{-1}(\beta/\alpha). \tag{16}$$

When $\alpha = \alpha(\boldsymbol{\theta})$ and $\beta = \beta(\boldsymbol{\theta})$, the quantity $\mathsf{d}_{\ell_2}(\alpha, \beta)$ measures the ℓ_2 distance between $\boldsymbol{\theta}$ and the set $\{-\boldsymbol{\theta}^*, \boldsymbol{\theta}^*\}$, that is, we have $\mathsf{d}_{\ell_2}(\alpha, \beta) = \min\{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2, \|\boldsymbol{\theta} + \boldsymbol{\theta}^*\|_2\}$. Similarly, the angular metric satisfies $\mathsf{d}_{\angle}(\alpha, \beta) = \min\{\angle(\boldsymbol{\theta}, \boldsymbol{\theta}^*), \angle(\boldsymbol{\theta}, -\boldsymbol{\theta}^*)\}$.

As alluded to in the previous sections. a *state evolution operator*, or update, is a function from state to state, thereby mapping $\mathbb{R}^2$ to itself. We begin with a few useful definitions for such operators. First, for any state evolution operator $\mathcal{S}$, recall that $\mathcal{S}^t$ denotes the operator formed by t iterated applications of $\mathcal{S}$. Next, we define an $\mathbb{S}$-faithful state evolution operator.

DEFINITION 1 ($\mathbb{S}$-faithful operator). For a set $\mathbb{S} \subseteq \mathbb{R}^2$, a state evolution operator $\mathcal{S} : \mathbb{R}^2 \to \mathbb{R}^2$ is said to be $\mathbb{S}$-faithful if $\mathcal{S}(\boldsymbol{\zeta}) \in \mathbb{S}$ for all $\boldsymbol{\zeta} \in \mathbb{S}$.

Next, we present two formal definitions of convergence rates, measuring linear (geometric) and faster-than-linear convergence.

DEFINITION 2 (Linear convergence of state evolution). For parameters $0 < c \leq C < 1$, a state evolution operator $\mathcal{S} : \mathbb{R}^2 \to \mathbb{R}^2$ is said to exhibit (c, C, t_0)-linear convergence in the metric d within the set $\mathbb{S}$ to level ε if $\mathcal{S}$ is $\mathbb{S}$-faithful, and for all $\boldsymbol{\zeta} \in \mathbb{S}$, we have

$$c \cdot \mathsf{d}\big(\mathcal{S}^t(\boldsymbol{\zeta})\big) + \frac{\varepsilon}{2} \leq \mathsf{d}\big(\mathcal{S}^{t+1}(\boldsymbol{\zeta})\big) \leq C \cdot \mathsf{d}\big(\mathcal{S}^t(\boldsymbol{\zeta})\big) + \varepsilon \quad \text{for all } t \geq t_0. \tag{17}$$

DEFINITION 3 (Superlinear convergence). Set parameters $0 < c \le C$ and $\lambda > 1$, and suppose that $\mathbb{S} \subseteq \{\boldsymbol{\zeta} : \mathsf{d}(\boldsymbol{\zeta}) \le C^{1-\lambda}\}$. A state evolution operator $\mathcal{S} : \mathbb{R}^2 \to \mathbb{R}^2$ is said to exhibit (c, C, λ, t_0)-superlinear convergence in the metric d within the set $\mathbb{S}$ to level ε if $\mathcal{S}$ is $\mathbb{S}$-faithful, and for all $\boldsymbol{\zeta} \in \mathbb{S}$, we have

$$c \cdot \big[\mathsf{d}\big(\mathcal{S}^t(\boldsymbol{\zeta})\big)\big]^{\lambda} + \frac{\varepsilon}{2} \le \mathsf{d}\big(\mathcal{S}^{t+1}(\boldsymbol{\zeta})\big) \le C \cdot \big[\mathsf{d}\big(\mathcal{S}^t(\boldsymbol{\zeta})\big)\big]^{\lambda} + \varepsilon \quad \text{for all } t \ge t_0. \tag{18}$$

A few comments on our definitions are worth making. First, note that both definitions require both upper and lower bounds on the per-step behavior of the algorithm, where the bounds apply after a "transient" period of t_0 iterations. This is a key feature of our framework, in that we are able to exactly characterize the convergence behavior as opposed to solely providing upper bounds. Both upper and lower bounds are characterized both by a rate of decrease of the error (linear in the case of equation (17) and superlinear in the case of equation (18)) and the eventual statistical neighborhood ε. Second, our choice of defining the lower bounds in equations (17) and (18) with $\varepsilon/2$ is arbitrary; any absolute constant in the denominator other than 2 preserves the qualitative convergence behavior.

As is common in the analysis of nonconvex optimization problems, our convergence guarantee will be established in two stages. In the first stage, we will show that the algorithm converges (typically slowly) to a "good region" around the optimal solution; once in the good region, the algorithm converges much faster. For both of the models that we consider, the following definition of the good region suffices. It is important to note that the numerical constants in this definition have not been optimized to be sharp. We note that the good region should be thought of as stronger than a "locally converging region" as it additionally allows us to show that the empirical iterates remain trapped in a small envelope around the deterministic predictions (see Theorems 3–6(b) to follow). This should be thought of as a sufficient definition of region in which such an envelope-type behavior can be established—it is an interesting technical question to pin down a necessary and sufficient definition.

DEFINITION 4 (Good region). Define the region

$$\mathbb{G} = \big\{(\alpha, \beta) \mid 0.55 \le \alpha \le 1.05, \text{ and } \alpha/\beta \ge 5\big\}.$$

With slight abuse of terminology, we say that $\boldsymbol{\theta} \in \mathbb{G}$ if $(\alpha(\boldsymbol{\theta}), \beta(\boldsymbol{\theta})) \in \mathbb{G}$.

Recall that $\Lambda = n/d$ denotes the per-step oversampling ratio. Also recall our notation for the iterated empirical operator, whereby the tth iterate after initialization at $\boldsymbol{\theta}$ is given by $\boldsymbol{\theta}_t = \mathcal{T}_n^t(\boldsymbol{\theta})$. We are now in a position to present our global convergence guarantees.

3.1. *Phase retrieval.* Our first example is phase retrieval, where equation (9) takes the form

$$y_i = \big|\langle \boldsymbol{x}_i, \boldsymbol{\theta}^* \rangle\big| + \epsilon_i, \tag{19}$$

that is, there is no auxiliary latent variable, and the function $f(t; q) = |t|$ depends solely on its first argument. We characterize the convergence behavior of both the alternating minimization algorithm and the subgradient descent method for this model.

3.1.1. *Alternating minimization.* The update here takes the general form (10) with $\omega(x, y) = \mathsf{sgn}(x) \cdot y$. That is, the empirical update run from the point θ is given by

$$\mathcal{T}_n(\boldsymbol{\theta}) = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^{\top}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \mathsf{sgn}\big(\langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle\big) \cdot y_i \cdot \boldsymbol{x}_i\right). \tag{20}$$

The following corollary follows from Theorem 1; in it, we state both the explicit Gordon state evolution and the concentration of the empirical iterates assuming that the update is run from some arbitrary "current" point $\boldsymbol{\theta}$. Its proof is proved in the Supplementary Material [13], Section C.1.

COROLLARY 1. *Let* $\alpha = \alpha(\boldsymbol{\theta})$ *and* $\beta = \beta(\boldsymbol{\theta})$ *with* $\boldsymbol{\zeta} = (\alpha, \beta)$ *and* $\phi = \tan^{-1}(\frac{\beta}{\alpha})$. *Let* $(\alpha^{\mathsf{gor}}, \beta^{\mathsf{gor}}) = \mathcal{S}_{\mathsf{gor}}(\boldsymbol{\zeta})$ *denote the Gordon state evolution from Table 2.*

(a) *We have*

$$\alpha^{\mathsf{gor}} = 1 - \frac{1}{\pi}\big(2\phi - \sin(2\phi)\big) \quad \textit{and} \tag{21a}$$

$$\beta^{\mathsf{gor}} = \sqrt{\frac{4}{\pi^2}\sin^4(\phi) + \frac{1}{\Lambda - 1}\left(1 - \left(1 - \frac{1}{\pi}\big(2\phi - \sin(2\phi)\big)\right)^2 - \frac{4}{\pi^2}\sin^4(\phi) + \sigma^2\right)}. \tag{21b}$$

(b) *Suppose* $\sigma > 0$. *There is a constant* $C_\sigma > 0$ *depending only on* σ *such that the following holds. With* $\mathcal{T}_n$ *as defined in equation* (20), *the empirical state evolution* $(\alpha^+, \beta^+) = (\alpha(\mathcal{T}_n(\boldsymbol{\theta})), \beta(\mathcal{T}_n(\boldsymbol{\theta})))$ *satisfies*

$$\mathbb{P}\left\{|\alpha^+ - \alpha^{\mathsf{gor}}| \le C_\sigma\left(\frac{\log^7(1/\delta)}{n}\right)^{1/2}\right\} \le \delta \quad \textit{and}$$

$$\mathbb{P}\left\{|\beta^+ - \beta^{\mathsf{gor}}| \le C_\sigma\left(\frac{\log(1/\delta)}{n}\right)^{1/4}\right\} \le \delta.$$

Note that the constant C_σ encodes the dependence on the parameters K_1 and K_2 from Assumptions 1 and 2. Indeed, in the Supplementary Material ([13], Section C.1.2), we compute $K_1 = 2(1 + \sigma^2)$ and $K_2 = \sigma^2$.

From equation (21), the following population update is obtained by letting $\Lambda \to \infty$:

$$\alpha^{\mathsf{pop}} = 1 - \frac{1}{\pi}\big(2\phi - \sin(2\phi)\big) \quad \text{and} \quad \beta^{\mathsf{pop}} = \frac{2}{\pi}\sin^2(\phi). \tag{22}$$

The population state evolution predicts superlinear convergence with exponent 2 in the good region, as shown in the following fact, proved in the Supplementary Material [13], Section C.5.

FACT 1. The population state evolution operator $\mathcal{S}_{\mathsf{pop}} = (\alpha^{\mathsf{pop}}, \beta^{\mathsf{pop}})$ is $(\frac{1}{20}, 1, \lambda, t_0)$-superlinearly convergent in the ℓ_2 metric[7] d_{ℓ_2} within the region $\mathbb{G}$ to level $\varepsilon = 0$, where $\lambda = 2$ and $t_0 = 1$.

However, the following theorem shows that the empirical quantities are instead tracked faithfully by the Gordon state evolution, which converges more slowly than the population state evolution. The proof of the theorem can be found in the Supplementary Material [13], Section 3.2.

THEOREM 3. *Consider the alternating minimization update* $\mathcal{T}_n$ *from equation* (20) *and the associated Gordon state evolution update* $\mathcal{S}_{\mathsf{gor}}$ *from equation* (21). *There is a universal positive constant* C *such that the following is true. If* $\Lambda \ge C(1 + \sigma^2)$, *then*:

(a) *The Gordon state evolution update*

$$\mathcal{S}_{\mathsf{gor}} \textit{ is } (c_\Lambda, C_\Lambda, \lambda, t_0)\textit{-superlinearly convergent in the } \ell_2 \textit{ metric } \mathsf{d}_{\ell_2} \textit{ within } \mathbb{G} \textit{ to level } \varepsilon_{n,d} = \tfrac{\sigma}{\sqrt{\Lambda}},$$

[7] In fact, the population state evolution (22) enjoys *global* quadratic convergence in the angular metric $\mathsf{d}_\angle$; see [13], Remark 2.

where $0 \le c_\Lambda \le C_\Lambda \le 1$ *are constants depending solely on* Λ, *and we have*

$$\lambda = 3/2 \quad \textit{and} \quad t_0 = 1.$$

(b) *If* $\sigma > 0$, *then there exist* $C_\sigma, C'_\sigma > 0$ *depending only on* σ *such that for all* $n \ge C'_\sigma$ *and for any* $\boldsymbol{\theta}$ *such that* $\boldsymbol{\zeta} = (\alpha(\boldsymbol{\theta}), \beta(\boldsymbol{\theta})) \in \mathbb{G}$, *we have*

$$\max_{1 \le t \le T} \left| \mathsf{d}_{\ell_2}\big(\mathcal{S}^t_{\mathsf{gor}}(\boldsymbol{\zeta})\big) - \big\|\mathcal{T}^t_n(\boldsymbol{\theta}) - \boldsymbol{\theta}^*\big\|_2 \right| \le C_\sigma \left(\frac{\log n}{n}\right)^{1/4}$$

with probability exceeding $1 - 2Tn^{-10}$.

(c) *Suppose* $\boldsymbol{\theta}_0$ *denotes a point such that* $\frac{\alpha(\boldsymbol{\theta}_0)}{\beta(\boldsymbol{\theta}_0)} \ge \frac{1}{50\sqrt{d}}$ *and further suppose that* $\Lambda \ge C''_\sigma \cdot \log^7\big(\frac{1+\log d}{\delta}\big)$ *for* C''_σ *depending solely on* σ. *Then for some* $t' \le C \log d$, *we have*

$$\mathcal{T}^{t'}_n(\boldsymbol{\theta}_0) \in \mathbb{G}$$

with probability exceeding $1 - \delta$.

Note that if $\boldsymbol{\theta}_0$ is chosen at random from the d-dimensional unit ball $\mathbb{B}_2(1)$ with $d \ge 130$, then we have $\frac{\alpha(\boldsymbol{\theta}_0)}{\beta(\boldsymbol{\theta}_0)} \ge \frac{1}{50\sqrt{d}}$ with probability at least 0.95 (see the Supplementary Material [13], Lemma 24(a)). Theorem 3 then shows that after $\tau = \mathcal{O}(\log d + \log\log(\Lambda/\sigma^2))$ iterations, we have

$$\big\|\mathcal{T}^\tau_n(\boldsymbol{\theta}) - \boldsymbol{\theta}^*\big\|_2 = \mathcal{O}\left(\sigma\sqrt{\frac{d}{n}}\right) + \widetilde{\mathcal{O}}\big(n^{-1/4}\big) \tag{23}$$

with high probability. Concretely, after taking $\mathcal{O}(\log d)$ steps to converge to the good region $\mathbb{G}$, the AM update converges *very* fast to within statistical error of the optimal parameter.

Some remarks on specific aspects of Theorem 3 are in order. First, note that this theorem predicts superlinear convergence with nonstandard exponent $3/2$ whenever Λ is bounded above. Comparing with Fact 1, we see that the population update is overly optimistic, and this corroborates what we saw in Figures 1 and 2 in the Introduction. Nonstandard superlinear convergence was recently observed in the noiseless case of this problem [34], but a larger exponent was conjectured. Theorem 3 shows that the exponent $3/2$ is indeed sharp, since we obtain both upper and lower bounds on the error of the algorithm. Furthermore, the convergence rate is superlinear with exponent $3/2$ for every value of the noise level. As we will see shortly, this is not the case for the closely related model of a symmetric mixture of regressions, in which the convergence rate is linear for any constant noise level.

Second, note that part (b) of the theorem shows that the (random) empirical state evolution is within ℓ_2 distance $n^{-1/4}$ of its (deterministic) Gordon counterpart once the iterates enter the good region. Consequently, the final result (23) on the empirical error has two terms. Note that this error is dominated by the $\sigma/\sqrt{\Lambda}$ term in modern high-dimensional problems.

Third, our convergence result is global, and holds from a random initialization. In particular, part (c) of the theorem guarantees that within $O(\log d)$ iterations, the iterations enter the good region $\mathbb{G}$, at which point parts (a) and (b) of the theorem become active. Convergence from a random initialization is also established by showing that the empirical state evolution tracks its Gordon counterpart closely. But rather than showing two deterministic envelopes around the empirical trajectory, we leverage closeness of the updates iterate-by-iterate. It is worth noting that this is the only step that requires the condition $n \asymp d \log^7(\log d)$; all other steps only require sample complexity that is linear in dimension.

Finally, we note that our assumption that σ^2/Λ be bounded above by a universal constant should not be viewed as restrictive. If this condition does not hold, then one can show using our analysis that running just one step of the algorithm from a random initialization already satisfies $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*\|^2 = \mathcal{O}(1) = \mathcal{O}(\sigma^2/\Lambda)$ and achieves order-optimal error.

3.1.2. *Subgradient descent.* To contrast with the superlinear convergence shown in the previous section, we now consider subgradient descent with stepsize $1/2$. As alluded to earlier and shown explicitly below, this update shares the same population update as AM, considered before. This update takes the form (11) with $\omega(x, y) = x - \mathsf{sgn}(x) \cdot y$. The general subgradient method for PR is thus given by the update

$$\mathcal{T}_n(\boldsymbol{\theta}) = \boldsymbol{\theta} - \frac{2\eta}{n} \cdot \sum_{i=1}^{n} \big(|\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle| - y_i\big) \cdot \mathsf{sgn}\big(\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle\big) \cdot \boldsymbol{x}_i, \tag{24}$$

where $\eta > 0$ denotes the stepsize. The Gordon state evolution update is given by the following corollary of Theorem 2, proved in the Supplementary Material [13], Section C.2.

COROLLARY 2. *Let $\alpha = \alpha(\boldsymbol{\theta})$ and $\beta = \beta(\boldsymbol{\theta})$ with $\phi = \tan^{-1}(\frac{\beta}{\alpha})$. Let $(\alpha^{\mathsf{gor}}, \beta^{\mathsf{gor}}) = \mathcal{S}_{\mathsf{gor}}(\alpha, \beta)$ denote the Gordon state evolution update for the subgradient descent operator (24), given by Table 2. Let $\eta \le 1/2$.*

(a) *We have*

$$\alpha^{\mathsf{gor}} = (1-2\eta)\alpha + 2\eta\bigg(1 - \frac{1}{\pi}\big(2\phi - \sin(2\phi)\big)\bigg), \quad \textit{and} \tag{25a}$$

$$\begin{aligned}\beta^{\mathsf{gor}} = \bigg(&\bigg\{(1-2\eta)\beta + 2\eta \cdot \frac{2}{\pi}\sin^2\phi\bigg\}^2 \\ &+ \frac{4\eta^2}{\Lambda}\bigg\{\alpha^2 + \beta^2 - 2\alpha\bigg(1 - \frac{1}{\pi}\big(2\phi - \sin(2\phi)\big)\bigg) \\ &- 2\beta \cdot \frac{2}{\pi}\sin^2\phi + 1 + \sigma^2\bigg\}\bigg)^{1/2}.\end{aligned} \tag{25b}$$

(b) *Suppose $\sigma > 0$ and $\alpha \vee \beta \le 3/2$. Then there is a positive constant C_σ depending only on σ such that with $\mathcal{T}_n$ as defined in equation (24), the empirical state evolution $(\alpha^+, \beta^+) = (\alpha(\mathcal{T}_n(\boldsymbol{\theta})), \beta(\mathcal{T}_n(\boldsymbol{\theta})))$ satisfies*

$$\mathbb{P}\bigg\{|\alpha^+ - \alpha^{\mathsf{gor}}| \le C_\sigma\bigg(\frac{\log(1/\delta)}{n}\bigg)^{1/2}\bigg\} \le \delta \quad \textit{and}$$

$$\mathbb{P}\bigg\{|\beta^+ - \beta^{\mathsf{gor}}| \le C_\sigma\bigg(\frac{\log(1/\delta)}{n}\bigg)^{1/4}\bigg\} \le \delta.$$

As in Corollary 1, C_σ encodes the dependence on the parameters $K_1 = 2(10 + \sigma^2)$ and $K_2 = \sigma^2$ (see [13], Section C.2.2). Sending $\Lambda \to \infty$ in equation (25) recovers the infinite-sample population update

$$\begin{aligned}\alpha^{\mathsf{pop}} &= (1-2\eta)\alpha + 2\eta\bigg(1 - \frac{1}{\pi}\big(2\phi - \sin(2\phi)\big)\bigg) \quad \text{and} \\ \beta^{\mathsf{pop}} &= (1-2\eta)\beta + 2\eta \cdot \frac{2}{\pi}\sin^2\phi.\end{aligned} \tag{26}$$

As previously noted, our interest[8] will be in analyzing the special case $\eta = 1/2$ so as to compare and contrast with the AM update. In this case, the population updates (22) and (26) coincide, and so Fact 1 suggests that subgradient descent ought to converge quadratically fast.

[8]Our techniques can also analyze the algorithm with general stepsize η, but we do not do so in this paper since a variety of other analysis methods tailored to first-order updates (e.g., [17, 82, 98]) also work in this case.

This would be quite surprising for a first-order method, and already suggests that the population update may be even more optimistic than before. However, the Gordon state evolution updates (21) and (25) are distinct even when $\eta = 1/2$, and as we saw before, these provide much more faithful predictions of convergence behavior. The proof of the following theorem can be found in the Supplementary Material [13], Section 3.3.

THEOREM 4. *Consider the subgradient descent update $\mathcal{T}_n$ (24) and the associated Gordon state evolution update $\mathcal{S}_{\mathsf{gor}}$ from equation (25), with stepsize $\eta = 1/2$. There is a universal positive constant C such that the following is true. If $\Lambda \geq C(1+\sigma^2)$, then*:

(a) *The Gordon state evolution update*

$$\mathcal{S}_{\mathsf{gor}} \text{ is } (c_\Lambda, C_\Lambda, 0)\text{-linearly convergent in the } \ell_2 \text{ metric } \mathsf{d}_{\ell_2} \text{ on } \mathbb{G} \text{ to level } \varepsilon_{n,d} = \tfrac{\sigma}{\sqrt{\Lambda}}.$$

Here, $0 \leq c_\Lambda \leq C_\Lambda < 1$ are constants depending solely on Λ.

(b) *Suppose $\sigma > 0$. Then there are positive constants C_σ, C'_σ depending only on σ such that for all $n \geq C'_\sigma$ and for any $\boldsymbol{\theta}$ such that $\boldsymbol{\zeta} = (\alpha(\boldsymbol{\theta}), \beta(\boldsymbol{\theta})) \in \mathbb{G}$, we have*

$$\max_{1\leq t\leq T} \left|\mathsf{d}_{\ell_2}\big(\mathcal{S}^t_{\mathsf{gor}}(\boldsymbol{\zeta})\big) - \big\|\mathcal{T}^t_n(\boldsymbol{\theta}) - \boldsymbol{\theta}^*\big\|_2\right| \leq C_\sigma \left(\frac{\log n}{n}\right)^{1/4}$$

with probability exceeding $1 - 2Tn^{-10}$.

(c) *Suppose $\boldsymbol{\theta}_0$ denotes a point such that $\frac{\alpha(\boldsymbol{\theta}_0)}{\beta(\boldsymbol{\theta}_0)} \geq \frac{1}{50\sqrt{d}}$ and $\alpha(\boldsymbol{\theta}_0) \vee \beta(\boldsymbol{\theta}_0) \leq 3/2$, and further suppose that $\Lambda \geq C''_\sigma \cdot \log(\frac{1+\log d}{\delta})$ for C''_σ depending solely on σ. Then for some $t' \leq C \log d$, we have*

$$\mathcal{T}^{t'}_n(\boldsymbol{\theta}_0) \in \mathbb{G}$$

with probability exceeding $1 - \delta$.

To be concrete once again, suppose $d \geq 130$. Then using $n \geq d$ observations $(\boldsymbol{x}_i, y_i)_{i=1}^n$ from the model (19) and setting $\boldsymbol{\theta}_0 = \sqrt{\frac{1}{n}\sum_{i=1}^n y_i^2} \cdot \boldsymbol{u}$ with the vector $\boldsymbol{u}$ chosen uniformly at random from the unit ball, we obtain the required initialization condition with probability greater than 0.95 (see the Supplementary Material [13], Lemma 24(b)). The theorem then guarantees that for some $\tau = \mathcal{O}(\log d + \log(\Lambda/\sigma^2))$, we have

$$\big\|\mathcal{T}^\tau_n(\boldsymbol{\theta}_0) - \boldsymbol{\theta}^*\big\| = \mathcal{O}\left(\sigma\sqrt{\frac{d}{n}}\right) + \widetilde{\mathcal{O}}\big(n^{-1/4}\big) \tag{27}$$

with high probability. Given our extensive discussion of Theorem 3 and that most of these comments also apply here, we make just one remark in passing that focuses on the difference. Note that as expected, Theorem 4 shows that subgradient descent only converges linearly in the good region. This corroborates what we saw in Figures 1 and 2, and shows once again—and more dramatically than before—that the (quadratically convergent) population update can be significantly optimistic in predicting convergence behavior.

3.2. *Mixture of regressions.* Our second specialization of (9) is the symmetric mixture of linear regressions model [21, 47]

$$y_i = q_i \cdot \langle \boldsymbol{x}_i, \boldsymbol{\theta}^* \rangle + \epsilon_i. \tag{28}$$

Here, the latent variables q_i are chosen i.i.d. from a Rademacher distribution $\mathsf{Unif}(\{\pm 1\})$, and we have taken $f(t;q) = q \cdot t$. Note that this model is statistically equivalent (for parameter estimation) to the phase retrieval model (19) in the absence of additive noise (i.e., $\sigma = 0$). On

the other hand, we will see that the models and their associated algorithms exhibit distinct behaviors for any nonzero noise level. We note in addition that Theorems 5 and 6 to follow impose an additional assumption that $\sigma \le c$. This assumption is technical in nature and allows to prove sharp "envelope" results (see Theorems 5, 6(b)). It is an interesting open question to determine whether such "envelope" results can be shown without this assumption on the noise level, and in particular whether such results hold when the σ scales all the way up to $\mathcal{O}(\sqrt{n/d})$.

3.2.1. *Alternating minimization.* The update here takes the general form (10) with $\omega(x, y) = \mathsf{sgn}(xy) \cdot y$. That is, the empirical update applied at $\boldsymbol{\theta}$ is given by

$$\mathcal{T}_n(\boldsymbol{\theta}) = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \mathsf{sgn}\big(y_i \cdot \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle\big) \cdot y_i \cdot \boldsymbol{x}_i\right). \tag{29}$$

The Gordon updates are given by the following corollary of Theorem 1, proved in the Supplementary Material [13], Section C.3. Before stating it, we define the convenient shorthand

$$A_\sigma(\rho) := \frac{2}{\pi}\tan^{-1}\big(\sqrt{\rho^2+\sigma^2+\sigma^2\rho^2}\big) \quad \text{and} \quad B_\sigma(\rho) := \frac{2}{\pi}\frac{\sqrt{\rho^2+\sigma^2+\sigma^2\rho^2}}{1+\rho^2}. \tag{30}$$

COROLLARY 3. *Let $\alpha = \alpha(\boldsymbol{\theta})$ and $\beta = \beta(\boldsymbol{\theta})$ with $\boldsymbol{\zeta} = (\alpha, \beta)$ and $\rho = \frac{\beta}{\alpha}$. Let $(\alpha^{\mathsf{gor}}, \beta^{\mathsf{gor}}) = \mathcal{S}_{\mathsf{gor}}(\boldsymbol{\zeta})$ denote the Gordon state evolution update in this case, given by Table 2.*

(a) *Using the shorthand (30), we have*

$$\alpha^{\mathsf{gor}} = 1 - A_\sigma(\rho) + B_\sigma(\rho), \quad \textit{and} \tag{31a}$$

$$\beta^{\mathsf{gor}} = \sqrt{\rho^2 B_\sigma(\rho)^2 + \frac{1}{\Lambda-1}\big(1+\sigma^2 - \big(1 - A_\sigma(\rho) + B_\sigma(\rho)\big)^2 - \rho^2 B_\sigma(\rho)^2\big)}. \tag{31b}$$

(b) *Suppose $\sigma > 0$. Then there is a positive constant C_σ depending only on σ such that with $\mathcal{T}_n$ as defined in equation (29), the empirical state evolution $(\alpha^+, \beta^+) = (\alpha(\mathcal{T}_n(\boldsymbol{\theta})), \beta(\mathcal{T}_n(\boldsymbol{\theta})))$ satisfies*

$$\mathbb{P}\left\{|\alpha^+ - \alpha^{\mathsf{gor}}| \le C_\sigma\left(\frac{\log^7(1/\delta)}{n}\right)^{1/2}\right\} \le \delta \quad \textit{and}$$

$$\mathbb{P}\left\{|\beta^+ - \beta^{\mathsf{gor}}| \le C_\sigma\left(\frac{\log(1/\delta)}{n}\right)^{1/4}\right\} \le \delta.$$

Note that C_σ is completely determined by the parameters $K_1 = 2(1+\sigma^2)$ and $K_2 = \sigma^2$ as computed in the Supplementary Material [13], Section C.3.2. By taking $\Lambda \to \infty$ in equation (31), we recover the population update, given by

$$\alpha^{\mathsf{pop}} = 1 - A_\sigma(\rho) + B_\sigma(\rho) \quad \text{and} \quad \beta^{\mathsf{pop}} = \rho B_\sigma(\rho). \tag{32}$$

The update (32) has no dependence on Λ and thus cannot recover the noise floor of the problem. On the other hand, and similar to before, the following theorem shows that the empirical quantities are tracked instead by the Gordon update (31). The proof of the following theorem can be found in the Supplementary Material [13], Section 3.4.

THEOREM 5. *Consider the alternating minimization update $\mathcal{T}_n$ given in equation (29) and the associated Gordon state evolution update $\mathcal{S}_{\mathsf{gor}}$ (31). There are universal positive constants (c, C) such that the following is true. If $\Lambda \ge C$ and $0 < \sigma \le c$, then:*

(a) *The Gordon state evolution update*

$\mathcal{S}_{\mathsf{gor}}$ *is* $(c_{\Lambda,\sigma}, C_{\Lambda,\sigma}, 0)$*-linearly convergent in the angular metric* $\mathsf{d}_\angle$ *on* $\mathbb{G}$ *to level* $\varepsilon_{n,d} = \frac{\sigma}{\sqrt{\Lambda}}$,

where $0 \le c_{\Lambda,\sigma} \le C_{\Lambda,\sigma} \le 1$ *are constants depending solely on the pair* (Λ, σ).

(b) *If* $n \ge C'_\sigma$, *then for any* $\boldsymbol{\theta}$ *such that* $\boldsymbol{\zeta} = (\alpha(\boldsymbol{\theta}), \beta(\boldsymbol{\theta})) \in \mathbb{G}$, *we have*

$$\max_{1\le t\le T} \left|\mathsf{d}_\angle\big(\mathcal{S}^t_{\mathsf{gor}}(\boldsymbol{\zeta})\big) - \angle(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\right| \le C_\sigma \left(\frac{\log n}{n}\right)^{1/4}$$

with probability exceeding $1 - 2Tn^{-10}$. *Here,* C_σ *and* C'_σ *are positive constants depending solely on* σ.

(c) *Suppose* $\boldsymbol{\theta}_0$ *denotes a point such that* $\frac{\alpha(\boldsymbol{\theta}_0)}{\beta(\boldsymbol{\theta}_0)} \ge \frac{1}{50\sqrt{d}}$ *and further suppose that* $\Lambda \ge C''_\sigma \cdot \log^7(\frac{1+\log d}{\delta})$ *for* C''_σ *depending solely on* σ. *Then for some* $t' \le C \log d$, *we have*

$$\mathcal{T}_n^{t'}(\boldsymbol{\theta}_0) \in \mathbb{G}$$

with probability exceeding $1 - \delta$.

Owing to the discussion following Theorem 3 (see [13], Lemma 24), we deduce that with a random initialization $\boldsymbol{\theta}_0$ and after $\tau = \mathcal{O}(\log d + \log\log(\Lambda/\sigma^2))$ iterations, we have

$$\angle(\mathcal{T}_n^\tau(\boldsymbol{\theta}_0), \boldsymbol{\theta}^*) = \mathcal{O}\left(\sigma\sqrt{\frac{d}{n}}\right) + \widetilde{\mathcal{O}}(n^{-1/4}) \tag{33}$$

with high probability.

Let us make a few remarks to compare and contrast Theorem 5 with our previous results. First, note that the convergence result proved here is in the angular metric $\mathsf{d}_\angle$ and not in the (stronger) ℓ_2 metric d_{ℓ_2}. This is a crucial difference between the phase retrieval and mixture of regressions models. Indeed, the parameter estimate for AM in mixtures of regressions can be shown to be inconsistent in the ℓ_2 distance; to see this, note that when $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, we have $\alpha^{\mathsf{gor}} = 1 + \Theta(\sigma^3)$. Combining this estimate with part (b) of Corollary 3, we see that $\mathsf{d}_{\ell_2}(\alpha^+, \beta^+) = \Theta(\sigma^3) + o(1)$, and so for any constant noise level, the algorithm is not consistent. Inconsistency of parameter estimation is a known phenomenon for alternating minimization algorithms in mixture models with noise (for instance, a similar conclusion follows from recent results [55] on label recovery for Lloyd's algorithm in a Gaussian mixture).

Second, note that when $\sigma = 0$, the mixture of regressions and phase retrieval models coincide. However, when there is noise, the convergence behavior predicted by Theorem 5 changes drastically to a linear rate, while in phase retrieval, superlinear convergence is preserved even when the noise level is nonzero (cf. Theorem 3). The Gordon update—and the ensuing sharpness of our characterization—enable us to make this distinction.

Finally, note that our assumption on the noise level in this case is that σ (as opposed to $\sigma/\sqrt{\Lambda}$) be bounded above by a universal constant, resulting in a more stringent condition than what we required in phase retrieval. While we make this assumption for convenience in our proof, we conjecture that it can be weakened to the optimal condition $\sigma/\sqrt{\Lambda} \le c$.

3.2.2. *Subgradient AM.* The update here is given by the general form (11) with $\omega(x, y) = x - \mathsf{sgn}(xy) \cdot y$. That is, the empirical update with stepsize η is given by

$$\mathcal{T}_n(\boldsymbol{\theta}) = \boldsymbol{\theta} - \frac{2\eta}{n} \cdot \sum_{i=1}^n \big(\mathsf{sgn}\big(y_i \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle\big) \cdot \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle - y_i\big) \cdot \mathsf{sgn}\big(y_i \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle\big) \cdot \boldsymbol{x}_i. \tag{34}$$

While we are not aware of the subgradient AM algorithm having been considered in the literature, it is natural for us to study it since when $\eta = 1/2$, it shares the same population

update as AM. Given that AM converges linearly for a mixture of regressions, it is natural to ask if the first-order method also does—while enjoying a much lower per-iteration cost.

The Gordon state evolution update in this case is given by the following corollary of Theorem 2, proved in the Supplementary Material [13], Section C.4.

COROLLARY 4. *Let* $\alpha = \alpha(\boldsymbol{\theta})$ *and* $\beta = \beta(\boldsymbol{\theta})$ *with* $\boldsymbol{\zeta} = (\alpha, \beta)$ *and* $\rho = \frac{\beta}{\alpha}$. *Let* $(\alpha^{\mathsf{gor}}, \beta^{\mathsf{gor}}) = \mathcal{S}_{\mathsf{gor}}(\boldsymbol{\zeta})$ *denote the Gordon state evolution corresponding to the update* (34), *given by Table* 2. *Let* $\eta \le 1/2$.

(a) *Using the shorthand* (30), *we have*

$$\alpha^{\mathsf{gor}} = (1-2\eta)\alpha + 2\eta \cdot \big(1 - A_\sigma(\rho) + B_\sigma(\rho)\big), \quad \textit{and} \tag{35a}$$

$$\beta^{\mathsf{gor}} = \Bigg(\big\{(1-2\eta)\beta + 2\eta \cdot \rho B_\sigma(\rho)\big\}^2 + \frac{4\eta^2}{\Lambda}\big\{\alpha^2 + \beta^2 - 2\alpha\big(1 - A_\sigma(\rho) + B_\sigma(\rho)\big) - 2\beta\rho B_\sigma(\rho) + 1 + \sigma^2\big\}\Bigg)^{1/2}. \tag{35b}$$

(b) *Suppose* $\sigma > 0$. *Then there is a positive constant* C_σ *depending solely on* σ *such that with* $\mathcal{T}_n$ *as defined in equation* (24), *the empirical state evolution* $(\alpha^+, \beta^+) = (\alpha(\mathcal{T}_n(\boldsymbol{\theta})), \beta(\mathcal{T}_n(\boldsymbol{\theta})))$ *satisfies*

$$\mathbb{P}\left\{\big|\alpha^+ - \alpha^{\mathsf{gor}}\big| \le C_\sigma\left(\frac{\log(1/\delta)}{n}\right)^{1/2}\right\} \le \delta \quad \textit{and}$$

$$\mathbb{P}\left\{\big|\beta^+ - \beta^{\mathsf{gor}}\big| \le C_\sigma\left(\frac{\log(1/\delta)}{n}\right)^{1/4}\right\} \le \delta.$$

As computed in the Supplementary Material ([13], Section C.4.2), we have $K_1 = 2(6+\sigma^2)$ and $K_2 = \sigma^2$, and C_σ is a function of only these constants. Sending $\Lambda \to \infty$ recovers the population update

$$\begin{aligned} \alpha^{\mathsf{pop}} &= (1-2\eta)\alpha + 2\eta \cdot \big(1 - A_\sigma(\rho) + B_\sigma(\rho)\big) \quad \text{and} \\ \beta^{\mathsf{pop}} &= (1-2\eta)\beta + 2\eta \cdot \rho B_\sigma(\rho). \end{aligned} \tag{36}$$

Once again, our interest will be in analyzing the special case $\eta = 1/2$, in which case the population updates (36) and (32) of both the first-order and higher-order algorithm coincide. The following theorem establishes a sharp characterization of the convergence behavior of the subgradient method. Its proof can be found in the Supplementary Material ([13], Section 3.5).

THEOREM 6. *Let the stepsize* $\eta = 1/2$ *and consider the subgradient update* $\mathcal{T}_n$ (29) *and the associated Gordon state evolution update* $\mathcal{S}_{\mathsf{gor}}$ (35). *There are universal positive constants* (c, C) *such that the following is true. If* $\Lambda \ge C$ *and* $\sigma \le c$, *then*:

(a) *The Gordon state evolution update*

$\mathcal{S}_{\mathsf{gor}}$ *is* $(c_{\Lambda,\sigma}, C_{\Lambda,\sigma}, 1)$*-linearly convergent in the angular metric* $\mathsf{d}_\angle$ *on* $\mathbb{G}$ *to level* $\varepsilon_{n,d} = \frac{\sigma}{\sqrt{\Lambda}}$,

where $0 \le c_{\Lambda,\sigma} \le C_{\Lambda,\sigma} \le 1$ *are constants depending solely on the pair* (Λ, σ).

(b) *If* $n \ge C'_\sigma$, *then for any* $\boldsymbol{\theta}$ *such that* $\boldsymbol{\zeta} = (\alpha(\boldsymbol{\theta}), \beta(\boldsymbol{\theta})) \in \mathbb{G}$, *we have*

$$\max_{1 \le t \le T} \big|\mathsf{d}_\angle\big(\mathcal{S}^t_{\mathsf{gor}}(\boldsymbol{\zeta})\big) - \angle(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\big| \le C_\sigma\left(\frac{\log n}{n}\right)^{1/4}$$

with probability exceeding $1 - 2Tn^{-10}$. *Here,* C'_σ *and* C_σ *are positive constants depending solely on* σ.

(c) *Suppose* $\boldsymbol{\theta}_0$ *denotes a point such that* $\frac{\alpha(\boldsymbol{\theta}_0)}{\beta(\boldsymbol{\theta}_0)} \geq \frac{1}{50\sqrt{d}}$ *and* $\alpha(\boldsymbol{\theta}_0) \vee \beta(\boldsymbol{\theta}_0) \leq 3/2$, *and further suppose that* $\Lambda \geq C''_\sigma \cdot \log(\frac{1+\log d}{\delta})$ *for* C''_σ *depending solely on* σ. *Then for some* $t' \leq C \log d$, *we have*

$$\mathcal{T}_n^{t'}(\boldsymbol{\theta}_0) \in \mathbb{G}$$

with probability exceeding $1 - \delta$.

As in the case of subgradient descent for phase retrieval (see the Supplementary Material [13], Lemma 24), we see that if $\boldsymbol{\theta}_0 = \sqrt{\frac{1}{n}\sum_{i=1}^n y_i^2} \cdot \boldsymbol{u}$ for a random vector $\boldsymbol{u}$ chosen from the unit sphere, then after $\tau = \mathcal{O}(\log d + \log\log(\Lambda/\sigma^2))$ iterations, we have with high probability, that

$$\angle(\mathcal{T}_n^\tau(\boldsymbol{\theta}_0), \boldsymbol{\theta}^*) = \mathcal{O}\left(\sigma\sqrt{\frac{d}{n}}\right) + \widetilde{\mathcal{O}}(n^{-1/4}). \tag{37}$$

The fact that both subgradient AM and AM (cf. Theorem 5) converge linearly in the good region suggests that the first-order method, which has smaller per-iteration cost, may be a more appropriate choice computationally. A closer look at the proof suggests that the corresponding coefficients of contraction $C_{\Lambda,\sigma}$ may be comparable for even moderately large Λ. Indeed, this is illustrated in Figure 7(b), where we see two settings of the pair (Λ, σ) in which both algorithms exhibit nearly identical behavior. This observation provides further evidence that our proposed subgradient AM method can be a compelling choice in such scenarios.

3.3. *A glimpse of the convergence proof mechanism.* To conclude this section, we provide a high level overview of our convergence proof technique, aspects of which may be of independent interest. A schematic of the proof mechanism is presented in Figure 3. The blue curve in the panel (top) represents the empirical state evolution (α_t, β_t), and our proof technique relies on tracking the transitions of this curve across three phases. Points (α, β) in Phase I are such that the ratio β/α is greater than some threshold. Phase II is characterized by β/α being between two distinct thresholds. Phase III corresponds to being in the good region $\mathbb{G}$, in which the ratio β/α is smaller than some small threshold and the parallel component α is larger than a threshold (see Definition 4). In each phase, depicted in detail in the (left), (right), and (bottom) plots of Figure 3, we track particular Gordon state evolution updates using red dots. The shaded light blue regions schematically depict confidence sets that show how each empirical iterate is "trapped" around its Gordon counterpart with high probability. In Phases I and II, we track Gordon state evolution updates when run from the "worst possible" empirical iterate in the previous confidence set, depicted in the figure using light blue triangles. In Phase III, on the other hand, we track the full Gordon *trajectory*, that is, the deterministic sequence of points that results from iteratively running the Gordon update from the initial dark blue triangle. The behavior of the Gordon update itself is model-dependent and governed by specific structural properties of the corresponding state evolution maps. We establish these properties in the Supplementary Material ([13], Section 3.1), and use them to establish part (a) of all our theorems in this section. For now, let us sketch the key ideas underlying our treatment of the empirical iterates in each phase.

Phase I: Immediately after initialization, the parallel component α is very small, of the order $d^{-1/2}$. To show that the empirical iterates proceed favorably through Phase I, we use the fact that the Gordon state evolution satisfies $\alpha^{\mathsf{gor}} \geq (1 + c)\alpha$ whenever β/α is large, thereby increasing the parallel component *exponentially* within this phase. The $\mathcal{O}(n^{-1/2})$ concentration of the empirical α_t update around its Gordon prediction traps each empirical

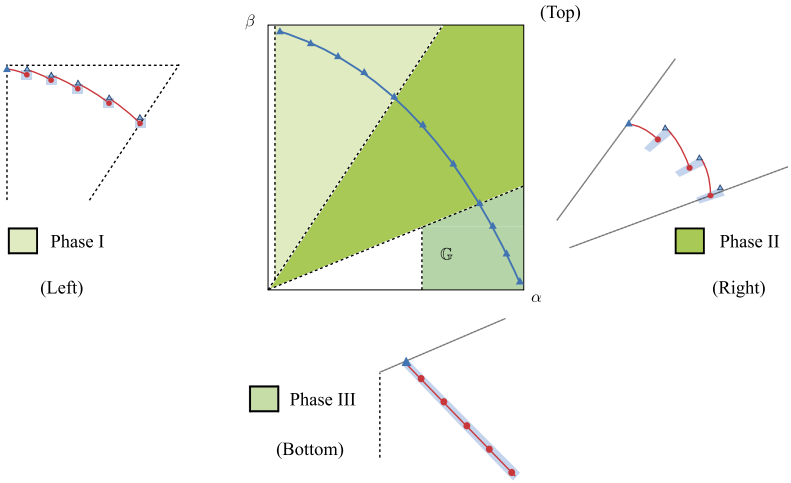


FIG. 3. *A schematic showing convergence of the algorithm in terms of its state space representation* (α, β), *in three distinct phases* (*top*). *The triangles in dark blue* (*and the corresponding curved line*) *denote the empirical iterates as they proceed through three* phases. *The panels* (*left*), (*right*) *and* (*bottom*) *are zoomed-in versions of Phases I*, *II and III*, *respectively*, *where the dark blue triangle at the start of the phase depicts the point of the trajectory within that phase. Each red circle in these subfigures denotes an iterate of the deterministic Gordon state evolution update when run from the point that it is connected to. The shaded blue regions in all three phases represent high-probability confidence sets for the empirical iterates. In panel* (*left*), *we leverage the fact that the* β*-component of each iterate is trapped around that of its Gordon counterpart. In panel* (*right*), *each such region is an angular "wedge" around the corresponding Gordon iterate*, *and in panel* (*bottom*), *the entire region* (*across iterations*) *is determined by a small envelope around the full Gordon trajectory. The light blue triangles in Phases I and II denote "worst-case" instances of the empirical updates within the corresponding confidence set. See the accompanying text for a more detailed discussion.*

iterate α_t within a small interval—as depicted in Figure 3 (left)—and allows us to argue when $n \gtrsim d$ that α_t also increases exponentially with t in Phase I. At the same time, the β_t iterates also remain bounded, so that β_t/α_t decreases below a threshold and enters Phase II. Phase I takes at most $\mathcal{O}(\log d)$ iterations with high probability.

Phase II: Next, we show that the ratio $\beta^{\mathsf{gor}}/\alpha^{\mathsf{gor}}$ of the Gordon state evolution decreases exponentially, and we translate this convergence to the empirical ratio β_t/α_t by using the relations (14) and (15). This traps each empirical iterate within a small *angular* neighborhood of its Gordon counterpart, and is depicted in Figure 3 (right). Together with the aforementioned convergence of the Gordon ratio $\beta^{\mathsf{gor}}/\alpha^{\mathsf{gor}}$, this ensures that we enter the good region $\mathbb{G}$. We show that with high probability, the iterates stay within Phase II for at most $\mathcal{O}(1)$ iterations. Along with the previously established convergence in Phase I, this establishes part (c) of all our model-specific theorems, showing that our iterates enter the good region, that is, Phase III, after at most $\mathcal{O}(\log d)$ steps after random initialization.

Phase III: In this final phase, we show a property that, to the best of our knowledge, is absent from local convergence guarantees in prior work. This is collected in part (b) of our individual theorems, and shows that a small *envelope* around the Gordon state evolution trajectory, as depicted in Figure 3 (bottom), fully traps the random iterates with high probability. The key property that we use to show this is in fact what guides our choice of the good region: The derivatives of the α^{gor} and β^{gor} maps when evaluated for any element in this region are both bounded above by $1 - c$ for some universal constant $c > 0$, so that small deviations of the empirical updates from these maps are not amplified over iterations.

4. Numerical illustrations. We conclude by providing several numerical simulations to illustrate the sharpness of our results. For each of the two models and two algorithms we

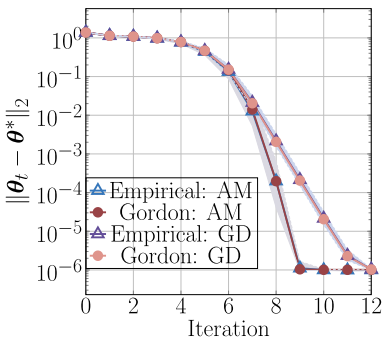
consider, we demonstrate both global convergence as well as local convergence. In particular, for each of the two models, we perform three families of experiments. The first explores convergence from a random initialization for both the higher-order and first-order method. These experiments are performed in dimension $d = 800$ with $n = 80{,}000$ samples per iteration (that is, $\Lambda = 100$) and noise standard deviation $\sigma = 10^{-6}$. First, a true parameter vector $\boldsymbol{\theta}^*$ is drawn uniformly at random from the unit sphere. Subsequently, an initialization $\boldsymbol{\theta}_0$ is drawn (independently of $\boldsymbol{\theta}^*$) uniformly at random on the unit sphere. Then, from this vector, we simulate 12 independent trials of the algorithm for 12 iterations. In the second family of experiments, we explore *local* convergence—from an initialization, which has constant correlation with the ground truth $\boldsymbol{\theta}^*$—for three different settings of noise standard deviation σ and oversampling ratio Λ. Each experiment is performed in dimension $d = 500$ for various values of n. Each simulation is done by first drawing the ground-truth vector $\boldsymbol{\theta}^*$ uniformly at random on the unit sphere and subsequently generating an initialization

$$\boldsymbol{\theta}_0 = 0.8 \cdot \boldsymbol{\theta}^* + \sqrt{1 - 0.8^2}\, \boldsymbol{P}^{\perp}_{\boldsymbol{\theta}^*} \boldsymbol{\gamma},$$

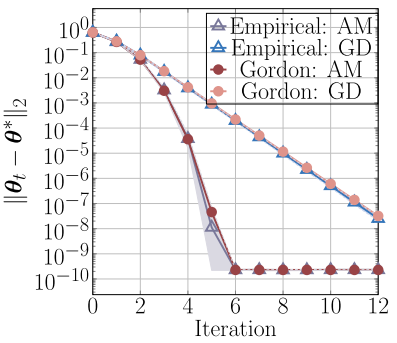
where $\boldsymbol{\gamma}$ is uniformly distributed on the unit sphere and is independent of all other randomness. Next, we run 100 independent trials of both algorithms for 12 iterations. Finally, in the third family of experiments, we demonstrate the sharpness of our results in the presence of constant noise. In particular, we consider dimension $d = 100$ with $n = 80{,}000$ samples per iteration ($\Lambda = 100$) and noise standard deviation $\sigma = 0.1$. As before, we draw a true parameter vector $\boldsymbol{\theta}^*$ uniformly at random on the unit sphere and an independent initialization uniformly at random on the unit sphere. We subsequently run 12 independent trials of alternating minimization.

4.1. *Phase retrieval.* We first consider phase retrieval. Figure 4(a) illustrates the global convergence of both alternating minimization and subgradient descent. Figure 4(a) plots (i) filled in circular marks denoting the Gordon state evolution started at the state $(\alpha(\boldsymbol{\theta}_0), \beta(\boldsymbol{\theta}_0))$; (ii) hollow triangular marks denoting the average of the empirical performance of AM over the 12 independent trials and (iii) a shaded region denoting the region between the minimum and maximum values taken in the empirical trajectory. The same three items are also plotted for subgradient descent.

Recall that part (c) of Theorems 3 and 4 states that each algorithm—when started from a random initialization—first consists of a transient phase which takes $\mathcal{O}(\log d)$ iterations



(a) Global convergence, $\sigma = 10^{-6}, \Lambda = 100$



(b) Local convergence, $\sigma = 10^{-10}, \Lambda = 20$

FIG. 4. *Convergence plots for phase retrieval. Each subplot shows*: (*in purple*) *the empirical trajectory of alternating minimization*, (*in red*) *the Gordon updates for alternating minimization*, (*in blue*) *the empirical trajectory for subgradient descent and* (*in orange*) *the Gordon updates for subgradient descent. Hollow triangles denote the average of the empirical iterates and the shaded regions denote the range of values taken by the empirical iterates over* 100 *trials.*

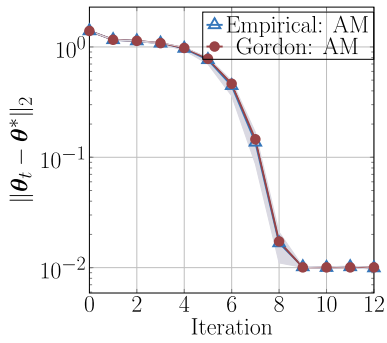


FIG. 5. *Global convergence of alternating minimization for phase retrieval with constant noise. The noise level is set as* $\sigma = 0.1$ *and* $\Lambda = 100$. *We plot the empirical trajectory of AM* (*blue triangular marks*) *as well as the Gordon prediction of the AM iterates* (*red circular marks*). *Each triangular mark is the average over* 12 *independent trials, and the shaded region denotes the range of values taken by the empirical iterates over the* 12 *trials.*

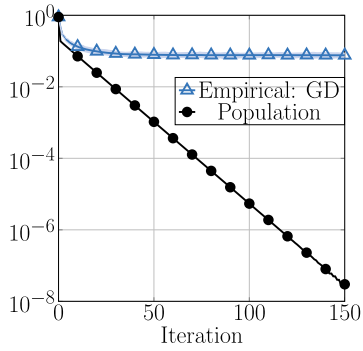
to reach a "good" region. This transient phase is witnessed by the first 5 iterations of each algorithm, which make very little progress in the ℓ_2 distance. Subsequently, parts (a) and (b) of each theorem state that in the "good" region, the Gordon state evolution converges at a specified rate and the empirical trajectory is trapped in a small envelope around this state evolution. Iterations $5 - 9$ illustrate the superlinear convergence of alternating minimization and iterations $5 - 12$ illustrate the linear convergence of subgradient descent. We remark that whereas the theorems show the empirical trajectory to be trapped in a small envelope surrounding the Gordon state evolution in the "good" region, the simulations suggest that this may hold even from random initialization—that is, even the transient phase may consist of an empirical trajectory trapped in an envelope around the Gordon state evolution.

Figure 4(b) zooms in and demonstrates local convergence. The Gordon prediction is distinct for the two algorithms, as predicted by part (a) of Theorems 3 and 4. Both plots also corroborate part (b) of these theorems: the empirical trajectory is trapped in a small envelope around the Gordon prediction.

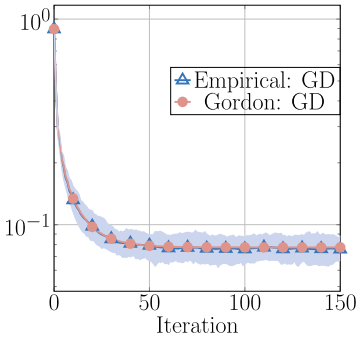
We note that while Figure 4 uses small noise levels $\sigma = 10^{-6}$ and $\sigma = 10^{-10}$, our theorems do not have such stringent noise requirements. Figure 5 illustrates global convergence in the presence of high noise, with $\sigma = 0.1$. Notice that our predictions remain exact.

In addition to plots showcasing global and local convergence, we provide another experiment in noiseless phase retrieval to illustrate the effect of stepsize in subgradient descent. Here, we take dimension $d = 250$, oversampling ratio $\Lambda = 10$ and start from an initial correlation $\alpha_0 = 0.6$. We take a large stepsize $\eta = 0.95$ and run 140 iterations of subgradient descent over 10 independent trials. As evident from Figure 6, this is a situation in which the population update predicts convergence, yet the empirical trajectory fails to converge. On the other hand, the Gordon updates continue to sharply characterize the empirical curve, and predict the lack of convergence to the ground truth parameter.

4.2. *Mixture of linear regressions.* The two sets of simulations performed in this subsection (Figure 7) follow the same dichotomy as the those performed in Figure 4 of the previous subsection. An important distinction is that the error metric used is the *angular* metric rather than the ℓ_2 distance. Figure 7(a) plots the trajectory of both AM and subgradient AM when started from a random initialization. As before, the simulations suggest that the empirical trajectory is trapped around the Gordon state evolution trajectory even from random initialization. Next, local convergence is illustrated in Figure 7(b), where the details of the setup are identical to the corresponding experiment in phase retrieval. Note the linearly convergent behavior evident in Figure 7(b), as well as the similarity in performance of the two algorithms



(a) Empirical curve alongside population curve



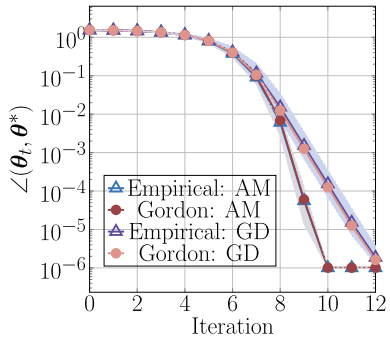
(b) Empirical curve alongside Gordon curve

FIG. 6. *Subgradient descent for stepsize* $\eta = 0.95$. *Markers are placed once every* 5 *iterations*.

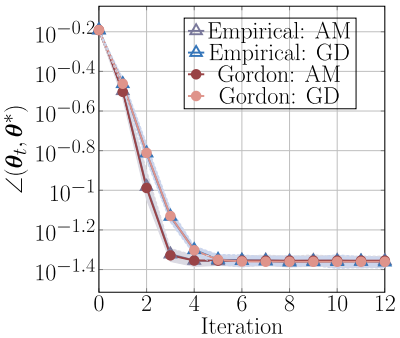
in the same simulation. As alluded to before, the competitiveness of the first-order method is an important feature of the mixtures of linear regressions model.

Finally, Figure 8 illustrates global convergence in the presence of constant noise. To this end, we consider the noise level $\sigma = 0.1$ and plot the error of alternating minimization over iterations. As before, we note that our predictions remain sharp even in this setting.

4.3. *Revisiting the sample-splitting assumption*. We conclude this section by performing an experiment to further examine our sample-splitting assumption. Recall that at every iteration, we assume the algorithm to use $n = \Lambda d$ samples. Consequently, after T iterations, a total of ΛdT samples are used. We consider three settings: (i) a total of ΛdT samples are used, but only Λd samples are used per iteration; (ii) total of Λd samples are used, and every sample is used at each iteration; (iii) a total of ΛdT samples are used, and every sample is used at each iteration. That is, setting (i) operates under our sample-splitting assumption. Setting (ii) does not employ sample splitting and is designed to match the per iteration sample complexity of setting (i), whereas setting (iii) does not employ sample splitting and is designed to match the total sample complexity of setting (i). Figure 9 plots the result of this experiment where the dimension is set as $d = 800$, the noise is set as $\sigma = 10^{-6}$, the per iteration oversampling ratio is set as $\Lambda = 20$ and we have run $T = 8$ iterations of the algorithm. We note that the optimal statistical performance is obtained by using all available samples at every iteration. Conversely, when the per iteration sample complexity is matched in settings (i) and (ii), both sets of iterates follow similar trajectories and eventually reach the same error floor.



(a) Global convergence, $\sigma = 10^{-6}, \Lambda = 100$



(b) Local convergence, $\sigma = 10^{-2}, \Lambda = 6$

FIG. 7. *Convergence in the mixtures of linear regressions model. For both noise settings, the local convergence is linear, even for the higher-order update (AM).*

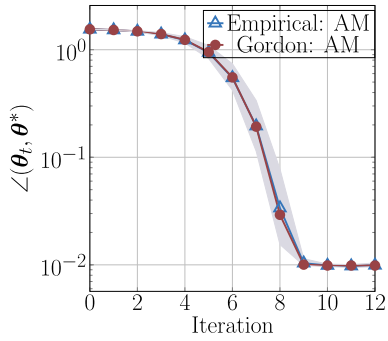


FIG. 8. *Global convergence of alternating minimization for mixtures of linear regression with constant noise. The noise level is set as* $\sigma = 0.1$ *and* $\Lambda = 100$. *We plot the empirical trajectory of AM (blue triangular marks) as well as the Gordon prediction of the AM iterates (red circular marks). Each triangular mark is the average over* 12 *independent trials, and the shaded region denotes the range of values taken by the empirical iterates over the* 12 *trials.*

5. Discussion. We presented the Gordon state evolution recipe for deriving sharp, deterministic predictions for the behavior of iterative algorithms in nonconvex statistical models with random data, which applies provided each iteration can be written as a convex program satisfying mild decomposability conditions and the data in the problem is normally distributed conditioned on the past. The key takeaway is that this enables a sharp characterization of convergence behavior, which we hope will prove useful in rigorously comparing algorithms in a wide range of problems. Using this recipe, we derived explicit Gordon updates and deviation bounds for a broad class of regression models with latent variables (9), and used this one-step characterization to establish several sharp, global convergence guarantees for both higher-order and first-order algorithms in two canonical statistical models; these global convergence results all appear to be novel contributions in themselves.

Our work opens the door to several interesting research directions, and we conclude by highlighting a few of them. The first question is technical. For higher-order methods, we showed that our deterministic predictions of the perpendicular component β were within $\mathcal{O}(n^{-1/4})$ of their empirical counterparts. While a direct analysis reveals an $\mathcal{O}(n^{-1/2})$ rate for first-order methods as well as an $\mathcal{O}(n^{-1/2})$ rate for the parallel component, we conjecture that a similar improvement can be carried out for the β component in higher-order meth-

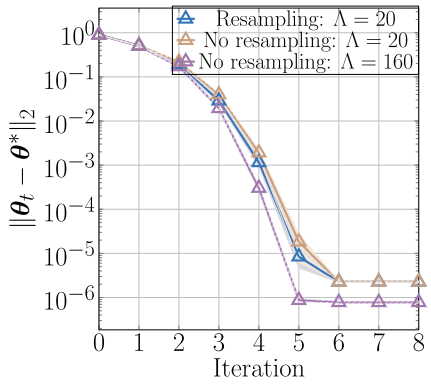


FIG. 9. *Comparison of trajectories with and without sample splitting. This simulation considers dimension* $d = 800$, *noise* $\sigma = 10^{-6}$, *and oversampling ratio* $\Lambda = 20$. *Blue triangular marks denote the trajectory of the resampled iterates, tan triangular marks denote the trajectory of the unresampled iterates keeping* $\Lambda = 20$, *and pink triangular marks denote the unresampled iterates when the total sample complexity (over* $T = 8$ *iterations) matches that of the sample-split algorithm.*

ods.[9] The second question has to do with our assumptions. Our analysis—which relied on Gaussianity of the data independent of the current iterate—used a sample-splitting device to partition the data into disjoint batches. While this is a reasonable method to obtain a practical algorithm—indeed, all the algorithms we analyzed converge very fast, so that at most a logarithmic number of batches suffices—how can the Gordon recipe be extended to analyze the case without sample splitting, or, more broadly, without Gaussianity? A final direction would be to broaden the scope of problems to which our analysis applies, for instance, by considering "weak" signal-to-noise regimes or settings with dependent data [96]. Weak signal-to-noise regimes have been the subject of recent work [28, 41, 92], and it is known that the optimal statistical rates of convergence are different from those in the strong signal-to-noise regimes that we consider in this paper. Deriving sharp rates of convergence of optimization algorithms in these settings should allow rigorous distinctions to be made among algorithmic behavior (using matching upper and lower bounds) in over, under and correctly specified situations.

Acknowledgments. We thank the program on Probability, Geometry and Computation in High Dimensions at the Simons Institute for the Theory of Computing for hosting us when part of this work was performed. We also thank the anonymous referees whose input improved the scope, clarity and presentation of this manuscript. In particular, we are grateful to an anonymous reviewer for a suggestion that led to the improved rate of $\widetilde{\mathcal{O}}(n^{-1/2})$ for first-order methods in Theorem 2.

Funding. KAC was supported in part by a National Science Foundation Graduate Research Fellowship and the Sony Stanford Graduate Fellowship. AP was supported in part by a research fellowship from the Simons Institute and National Science Foundation Grant CCF-2107455. CT was supported in part by the National Science Foundation Grant CCF-2009030, by an NSERC Discovery Grant and by a research grant from KAUST.

SUPPLEMENTARY MATERIAL

Supplement to: "Sharp global convergence guarantees for iterative nonconvex optimization with random data" (DOI: 10.1214/22-AOS2246SUPP; .pdf). Provides full proofs of all theorems and corollaries stated in the paper.

REFERENCES

[1] AGARWAL, A., ANANDKUMAR, A., JAIN, P. and NETRAPALLI, P. (2016). Learning sparsely used overcomplete dictionaries via alternating minimization. *SIAM J. Optim.* **26** 2775–2799. MR3580820 https://doi.org/10.1137/140979861

[2] AGARWAL, A., NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Statist.* **40** 2452–2482. MR3097609 https://doi.org/10.1214/12-AOS1032

[3] AMELUNXEN, D., LOTZ, M., MCCOY, M. B. and TROPP, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. *Inf. Inference* **3** 224–294. MR3311453 https://doi.org/10.1093/imaiai/iau005

[4] AUBIN, B., LU, Y., KRZAKALA, F. and ZDEBOROVA, L. (2020). Generalization error in high-dimensional perceptrons: Approaching Bayes error with convex optimization. In *Advances in Neural Information Processing Systems*.

[5] BALAKRISHNAN, S., WAINWRIGHT, M. J. and YU, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Ann. Statist.* **45** 77–120. MR3611487 https://doi.org/10.1214/16-AOS1435

[9] Indeed, a subset of the authors have shown this to be the case for a particular higher-order method in a distinct statistical model [12], but by using another set of technical tools.

[6] BAYATI, M. and MONTANARI, A. (2012). The LASSO risk for Gaussian matrices. *IEEE Trans. Inf. Theory* **58** 1997–2017. MR2951312 https://doi.org/10.1109/TIT.2011.2174612

[7] BRILLINGER, D. R. (2012). A generalized linear model with "Gaussian" regressor variables. In *Selected Works of David Brillinger* 589–606. Springer, Berlin.

[8] CELENTANO, M., CHENG, C. and MONTANARI, A. (2021). The high-dimensional asymptotics of first order methods with random data. ArXiv preprint. Available at arXiv:2112.07572.

[9] CELENTANO, M., MONTANARI, A. and WEI, Y. (2020). The Lasso with general Gaussian designs with applications to hypothesis testing. ArXiv preprint. Available at arXiv:2007.13716.

[10] CELENTANO, M., MONTANARI, A. and WU, Y. (2020). The estimation error of general first order methods. In *Conference on Learning Theory* 1078–1141. PMLR.

[11] CHANDRASEKARAN, V., RECHT, B., PARRILO, P. A. and WILLSKY, A. S. (2012). The convex geometry of linear inverse problems. *Found. Comput. Math.* **12** 805–849. MR2989474 https://doi.org/10.1007/s10208-012-9135-7

[12] CHANDRASEKHER, K. A., LOU, M. and PANANJADY, A. (2022). Alternating minimization for generalized rank one matrix sensing: Sharp predictions from a random initialization. ArXiv preprint. Available at arXiv:2207.09660.

[13] CHANDRASEKHER, K. A., PANANJADY, A. and THRAMPOULIDIS, C. (2023). Supplement to "Sharp global convergence guarantees for iterative nonconvex optimization with random data." https://doi.org/10.1214/22-AOS2246SUPP

[14] CHANG, X., LI, Y., OYMAK, S. and THRAMPOULIDIS, C. (2021). Provable benefits of overparameterization in model compression: From double descent to pruning neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* **35** 6974–6983.

[15] CHARISOPOULOS, V., CHEN, Y., DAVIS, D., DÍAZ, M., DING, L. and DRUSVYATSKIY, D. (2021). Low-rank matrix recovery with composite optimization: Good conditioning and rapid convergence. *Found. Comput. Math.* **21** 1505–1593. MR4343018 https://doi.org/10.1007/s10208-020-09490-9

[16] CHEN, Y. and CHI, Y. (2018). Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization. *IEEE Signal Process. Mag.* **35** 14–31.

[17] CHEN, Y., CHI, Y., FAN, J. and MA, C. (2019). Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Math. Program.* **176** 5–37. MR3960803 https://doi.org/10.1007/s10107-019-01363-6

[18] CHI, Y., LU, Y. M. and CHEN, Y. (2019). Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Trans. Signal Process.* **67** 5239–5269. MR4016283 https://doi.org/10.1109/TSP.2019.2937282

[19] DASKALAKIS, C., TZAMOS, C. and ZAMPETAKIS, M. (2017). Ten steps of EM suffice for mixtures of two Gaussians. In *Conference on Learning Theory* 704–710. PMLR.

[20] DAVIS, D., DRUSVYATSKIY, D. and PAQUETTE, C. (2020). The nonsmooth landscape of phase retrieval. *IMA J. Numer. Anal.* **40** 2652–2695. MR4167058 https://doi.org/10.1093/imanum/drz031

[21] DE VEAUX, R. D. (1989). Mixtures of linear regressions. *Comput. Statist. Data Anal.* **8** 227–245. MR1028403 https://doi.org/10.1016/0167-9473(89)90043-1

[22] DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

[23] DENG, Z., KAMMOUN, A. and THRAMPOULIDIS, C. (2022). A model of double descent for high-dimensional binary linear classification. *Inf. Inference* **11** 435–495. MR4474343 https://doi.org/10.1093/imaiai/iaab002

[24] DHIFALLAH, O. and LU, Y. M. (2020). A precise performance analysis of learning with random features. ArXiv preprint. Available at arXiv:2008.11904.

[25] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2009). Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* **106** 18914–18919.

[26] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2011). The noise-sensitivity phase transition in compressed sensing. *IEEE Trans. Inf. Theory* **57** 6920–6941. MR2882271 https://doi.org/10.1109/TIT.2011.2165823

[27] DUCHI, J. C. and RUAN, F. (2019). Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Inf. Inference* **8** 471–529. MR3994397 https://doi.org/10.1093/imaiai/iay015

[28] DWIVEDI, R., HO, N., KHAMARU, K., WAINWRIGHT, M. J., JORDAN, M. I. and YU, B. (2020). Singularity, misspecification and the convergence rate of EM. *Ann. Statist.* **48** 3161–3182. MR4185804 https://doi.org/10.1214/19-AOS1924

[29] FENG, O. Y., VENKATARAMANAN, R., RUSH, C. and SAMWORTH, R. J. (2021). A unifying tutorial on approximate message passing. ArXiv preprint. Available at arXiv:2105.02180.

[30] FIENUP, J. R. (1982). Phase retrieval algorithms: A comparison. *Appl. Opt.* **21** 2758–2769. https://doi.org/10.1364/AO.21.002758

[31] GAO, B. and XU, Z. (2017). Phaseless recovery using the Gauss–Newton method. *IEEE Trans. Signal Process.* **65** 5885–5896. MR3722968 https://doi.org/10.1109/TSP.2017.2742981

[32] GERCHBERG, R. W. (1972). A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik* **35** 237–246.

[33] GHOSH, A., PANANJADY, A., GUNTUBOYINA, A. and RAMCHANDRAN, K. (2022). Max-affine regression: Parameter estimation for Gaussian designs. *IEEE Trans. Inf. Theory* **68** 1851–1885. MR4395504 https://doi.org/10.1109/TIT.2021.3130717

[34] GHOSH, A. and RAMCHANDRAN, K. (2020). Alternating minimization converges super-linearly for mixed linear regression. In *International Conference on Artificial Intelligence and Statistics* 1093–1103. PMLR.

[35] GORDON, Y. (1985). Some inequalities for Gaussian processes and applications. *Israel J. Math.* **50** 265–289. MR0800188 https://doi.org/10.1007/BF02759761

[36] GORDON, Y. (1988). On Milman's inequality and random subspaces which escape through a mesh in $\mathbf{R}^n$. In *Geometric Aspects of Functional Analysis* (1986/87). *Lecture Notes in Math.* **1317** 84–106. Springer, Berlin. MR0950977 https://doi.org/10.1007/BFb0081737

[37] GUNASEKAR, S., ACHARYA, A., GAUR, N. and GHOSH, J. (2013). Noisy matrix completion using alternating minimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 194–209. Springer, Berlin.

[38] HAND, P., LEONG, O. and VORONINSKI, V. (2018). Phase retrieval under a generative prior. In *Advances in Neural Information Processing Systems* 9154–9164.

[39] HAND, P. and VORONINSKI, V. (2020). Global guarantees for enforcing deep generative priors by empirical risk. *IEEE Trans. Inf. Theory* **66** 401–418. MR4053402 https://doi.org/10.1109/TIT.2019.2935447

[40] HARDT, M. and WOOTTERS, M. (2014). Fast matrix completion without the condition number. In *Conference on Learning Theory* 638–678. PMLR.

[41] HO, N., KHAMARU, K., DWIVEDI, R., WAINWRIGHT, M. J., JORDAN, M. I. and YU, B. (2020). Instability, computational efficiency and statistical accuracy. ArXiv preprint. Available at arXiv:2005.11411.

[42] JAGATAP, G. and HEGDE, C. (2017). Fast, sample-efficient algorithms for structured phase retrieval. In *Advances in Neural Information Processing Systems* 4924–4934.

[43] JAIN, P. and KAR, P. (2017). Non-convex optimization for machine learning. *Found. Trends Mach. Learn.* **10** 142–363.

[44] JAIN, P., NETRAPALLI, P. and SANGHAVI, S. (2013). Low-rank matrix completion using alternating minimization (extended abstract). In *STOC'*13*—Proceedings of the* 2013 *ACM Symposium on Theory of Computing* 665–674. ACM, New York. MR3210828 https://doi.org/10.1145/2488608.2488693

[45] JAVANMARD, A. and SOLTANOLKOTABI, M. (2022). Precise statistical analysis of classification accuracies for adversarial training. *Ann. Statist.* **50** 2127–2156. MR4474485 https://doi.org/10.1214/22-aos2180

[46] JAVANMARD, A., SOLTANOLKOTABI, M. and HASSANI, H. (2020). Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory* 2034–2078. PMLR.

[47] JORDAN, M. I. and JACOBS, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Comput.* **6** 181–214.

[48] KAMMOUN, A. and ALOUINI, M.-S. (2021). On the precise error analysis of support vector machines. *IEEE Open J. Signal Process.* **2** 99–118.

[49] KLUSOWSKI, J. M., YANG, D. and BRINDA, W. D. (2019). Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *IEEE Trans. Inf. Theory* **65** 3515–3524. MR3959002 https://doi.org/10.1109/TIT.2019.2891628

[50] KUNSTNER, F., KUMAR, R. and SCHMIDT, M. (2021). Homeomorphic-invariance of EM: Non-asymptotic convergence in KL divergence for exponential families via mirror descent. In *International Conference on Artificial Intelligence and Statistics* 3295–3303. PMLR.

[51] KWON, J., QIAN, W., CARAMANIS, C., CHEN, Y. and DAVIS, D. (2019). Global convergence of the EM algorithm for mixtures of two component linear regression. In *Conference on Learning Theory* 2055–2110. PMLR.

[52] LIANG, T. and SUR, P. (2022). A precise high-dimensional asymptotic theory for boosting and minimum-ℓ1-norm interpolated classifiers. *Ann. Statist.* **50** 1669–1695. MR4441136 https://doi.org/10.1214/22-aos2170

[53] LOH, P.-L. and WAINWRIGHT, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.* **40** 1637–1664. MR3015038 https://doi.org/10.1214/12-AOS1018

[54] Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mézard, M. and Zdeborová, L. (2021). Capturing the learning curves of generic features maps for realistic data sets with a teacher-student model. In *Conference on Neural Information Processing Systems* (*NeurIPS*).

[55] Lu, Y. and Zhou, H. H. (2016). Statistical and computational guarantees of Lloyd's algorithm and its variants. ArXiv preprint. Available at arXiv:1612.02099.

[56] Maillard, A., Loureiro, B., Krzakala, F. and Zdeborová, L. (2020). Phase retrieval in high dimensions: Statistical and computational phase transitions. In *Advances in Neural Information Processing Systems* **33** 11071–11082.

[57] Makkuva, A., Viswanath, P., Kannan, S. and Oh, S. (2019). Breaking the gridlock in mixture-of-experts: Consistent and efficient algorithms. In *International Conference on Machine Learning* 4304–4313. PMLR.

[58] Mei, S., Bai, Y. and Montanari, A. (2018). The landscape of empirical risk for nonconvex losses. *Ann. Statist.* **46** 2747–2774. MR3851754 https://doi.org/10.1214/17-AOS1637

[59] Miolane, L. and Montanari, A. (2021). The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *Ann. Statist.* **49** 2313–2335. MR4319252 https://doi.org/10.1214/20-aos2038

[60] Montanari, A. (2013). *Statistical Estimation*: *From Denoising to Sparse Regression and Hidden Cliques*. *Statistical Physics*, *Optimization*, *Inference*, *and Message-Passing Algorithms*. *Lecture Notes of the Les Houches School of Physics*: *Special Issue*.

[61] Montanari, A., Ruan, F., Sohn, Y. and Yan, J. (2019). The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. ArXiv preprint. Available at arXiv:1911.01544.

[62] Neal, R. M. and Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models* 355–368. Springer, Berlin.

[63] Netrapalli, P., Jain, P. and Sanghavi, S. (2015). Phase retrieval using alternating minimization. *IEEE Trans. Signal Process.* **63** 4814–4826. MR3385838 https://doi.org/10.1109/TSP.2015.2448516

[64] Oymak, S. and Hassibi, B. (2010). New null space results and recovery thresholds for matrix rank minimization. ArXiv preprint. Available at arXiv:1011.6326.

[65] Oymak, S., Recht, B. and Soltanolkotabi, M. (2018). Sharp time-data tradeoffs for linear inverse problems. *IEEE Trans. Inf. Theory* **64** 4129–4158. MR3809731 https://doi.org/10.1109/TIT.2017.2773497

[66] Oymak, S. and Soltanolkotabi, M. (2017). Fast and reliable parameter estimation from nonlinear observations. *SIAM J. Optim.* **27** 2276–2300. MR3716592 https://doi.org/10.1137/17M1113874

[67] Oymak, S., Thrampoulidis, C. and Hassibi, B. (2013). The squared-error of generalized Lasso: A precise analysis. In 2013 51*st Annual Allerton Conference on Communication*, *Control*, *and Computing* 1002–1009. IEEE, Los Alamitos.

[68] Pananjady, A. and Foster, D. P. (2021). Single-index models in the high signal regime. *IEEE Trans. Inf. Theory* **67** 4092–4124. MR4289367 https://doi.org/10.1109/TIT.2021.3075142

[69] Plan, Y. and Vershynin, R. (2016). The generalized Lasso with non-linear observations. *IEEE Trans. Inf. Theory* **62** 1528–1537. MR3472264 https://doi.org/10.1109/TIT.2016.2517008

[70] Rudelson, M. and Vershynin, R. (2006). Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements. In 2006 40*th Annual Conference on Information Sciences and Systems* 207–212. IEEE, Los Alamitos.

[71] Salehi, F., Abbasi, E. and Hassibi, B. (2019). The impact of regularization on high-dimensional logistic regression. ArXiv preprint. Available at arXiv:1906.03761.

[72] Stojnic, M. (2009). Various thresholds for ℓ_1-optimization in compressed sensing. ArXiv preprint. Available at arXiv:0907.3666.

[73] Stojnic, M. (2013). A framework to characterize performance of Lasso algorithms. ArXiv preprint. Available at arXiv:1303.7291.

[74] Stojnic, M. (2013). Upper-bounding ℓ_1-optimization weak thresholds. ArXiv preprint. Available at arXiv:1303.7289.

[75] Stojnic, M. (2013). Regularly random duality. ArXiv preprint. Available at arXiv:1303.7295.

[76] Sun, J. (2021). Provable nonconvex Methods/Algorithms.

[77] Sun, R. and Luo, Z.-Q. (2016). Guaranteed matrix completion via non-convex factorization. *IEEE Trans. Inf. Theory* **62** 6535–6579. MR3565131 https://doi.org/10.1109/TIT.2016.2598574

[78] Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Natl. Acad. Sci. USA* **116** 14516–14525. MR3984492 https://doi.org/10.1073/pnas.1810420116

[79] TAHERI, H., PEDARSANI, R. and THRAMPOULIDIS, C. (2020). Sharp asymptotics and optimal performance for inference in binary models. In *International Conference on Artificial Intelligence and Statistics* 3739–3749. PMLR.

[80] TAHERI, H., PEDARSANI, R. and THRAMPOULIDIS, C. (2021). Fundamental limits of ridge-regularized empirical risk minimization in high dimensions. In *International Conference on Artificial Intelligence and Statistics* 2773–2781. PMLR.

[81] TAN, Y. S. and VERSHYNIN, R. (2019). Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval. ArXiv preprint. Available at arXiv:1910.12837.

[82] TAN, Y. S. and VERSHYNIN, R. (2019). Phase retrieval via randomized Kaczmarz: Theoretical guarantees. *Inf. Inference* **8** 97–123. MR3922404 https://doi.org/10.1093/imaiai/iay005

[83] THRAMPOULIDIS, C. (2016). Recovering structured signals in high dimensions via non-smooth convex optimization: Precise performance analysis Ph.D. thesis California Institute of Technology.

[84] THRAMPOULIDIS, C., ABBASI, E. and HASSIBI, B. (2015). Lasso with non-linear measurements is equivalent to one with linear measurements. In *Advances in Neural Information Processing Systems* 3420–3428.

[85] THRAMPOULIDIS, C., ABBASI, E. and HASSIBI, B. (2018). Precise error analysis of regularized M-estimators in high dimensions. *IEEE Trans. Inf. Theory* **64** 5592–5628. MR3832326 https://doi.org/10.1109/TIT.2018.2840720

[86] THRAMPOULIDIS, C., OYMAK, S. and HASSIBI, B. (2015). Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory* 1683–1709. PMLR.

[87] THRAMPOULIDIS, C., XU, W. and HASSIBI, B. (2018). Symbol error rate performance of box-relaxation decoders in massive MIMO. *IEEE Trans. Signal Process.* **66** 3377–3392. MR3832377 https://doi.org/10.1109/TSP.2018.2831622

[88] TIAN, Y. (2017). An analytical formula of population gradient for two-layered ReLu network and its applications in convergence and critical point analysis. In *International Conference on Machine Learning* 3404–3413. PMLR.

[89] VERSHYNIN, R. (2018). *High-Dimensional Probability*: *An Introduction with Applications in Data Science*. *Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge Univ. Press, Cambridge. MR3837109 https://doi.org/10.1017/9781108231596

[90] WALDSPURGER, I. (2018). Phase retrieval with random Gaussian sensing vectors by alternating projections. *IEEE Trans. Inf. Theory* **64** 3301–3312. MR3798378 https://doi.org/10.1109/TIT.2018.2800663

[91] WANG, S., WENG, H. and MALEKI, A. (2022). Does SLOPE outperform bridge regression? *Inf. Inference* **11** 1–54. MR4409197 https://doi.org/10.1093/imaiai/iaab025

[92] WU, Y. and ZHOU, H. H. (2021). Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations. *Math. Stat. Learn.* **4** 143–220. MR4383733 https://doi.org/10.4171/msl/29

[93] XU, J., HSU, D. J. and MALEKI, A. (2016). Global analysis of expectation maximization for mixtures of two Gaussians. *Adv. Neural Inf. Process. Syst.* **29**.

[94] XU, J., HSU, D. J. and MALEKI, A. (2018). Benefits of over-parameterization with EM. In *Advances in Neural Information Processing Systems* **31**.

[95] XU, L. and JORDAN, M. I. (1996). On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Comput.* **8** 129–151.

[96] YANG, F., BALAKRISHNAN, S. and WAINWRIGHT, M. J. (2017). Statistical and computational guarantees for the Baum–Welch algorithm. *J. Mach. Learn. Res.* **18** 125. MR3763759

[97] YI, X., CARAMANIS, C. and SANGHAVI, S. (2014). Alternating minimization for mixed linear regression. In *International Conference on Machine Learning* 613–621. PMLR.

[98] ZHANG, H., ZHOU, Y., LIANG, Y. and CHI, Y. (2017). A nonconvex approach for phase retrieval: Reshaped Wirtinger flow and incremental algorithms. *J. Mach. Learn. Res.* **18** 141. MR3763775

[99] ZHANG, T. (2020). Phase retrieval using alternating minimization in a batch setting. *Appl. Comput. Harmon. Anal.* **49** 279–295. MR4091199 https://doi.org/10.1016/j.acha.2019.02.001

[100] ZHANG, Y., QU, Q. and WRIGHT, J. (2020). From symmetry to geometry: Tractable nonconvex problems. ArXiv preprint. Available at arXiv:2007.06753.