Max-affine regression with universal parameter estimation for small-ball designs

Avishek Ghosh[†,*], Ashwin Pananjady[†,*], Aditya Guntuboyina[†] Kannan Ramchandran[‡]
[†]Department of Electrical Engineering and Computer Sciences, UC Berkeley,
avishek_ghosh@berkeley.edu,{ashwinpm,kannanr}@eecs.berkeley.edu
[‡]Department of Statistics, UC Berkeley
agun@berkeley.edu

***Abstract*—We study the max-affine regression model, where the unknown regression function is modeled as a maximum of a fixed number of affine functions. In recent work [1], we showed that end-to-end parameter estimates were obtainable using this model with an alternating minimization (AM) algorithm provided the covariates (or designs) were normally distributed, and chosen independently of the underlying parameters. In this paper, we show that AM is significantly more robust than the setting of [1]: It converges locally under small-ball design assumptions (which is a much broader class, including bounded log-concave distributions), and even when the underlying parameters are chosen with knowledge of the realized covariates. Once again, the final rate obtained by the procedure is near-parametric and minimax optimal (up to a polylogarithmic factor) as a function of the dimension, sample size, and noise variance. As a by-product of our analysis, we obtain convergence guarantees on a classical algorithm for the (real) phase retrieval problem in the presence of noise under considerably weaker assumptions on the design distribution than was previously known.**

I. Introduction

The max-affine regression model is given by

$$Y = \max_{1 \le j \le k} \left(\langle X, \theta_j^* \rangle + b_j^* \right) + \epsilon, \tag{1}$$

where Y is a scalar response, X is a d-dimensional covariate vector and the noise ϵ is drawn from a (univariate) distribution that is zero-mean and sub-Gaussian, with unknown sub-Gaussian parameter σ. Furthermore, the noise ϵ is independent of the random covariates X. Assuming $k \ge 1$ is a known integer, we study the problem of estimating the unknown parameters $\theta_1^*, \ldots, \theta_k^* \in \mathbb{R}^d$ and $b_1^*, \ldots, b_k^* \in \mathbb{R}$ from independent observations $(x_1, y_1), \ldots, (x_n, y_n)$ drawn according to the model (1). We often use the convenient notation $\beta_j^* := (\theta_j^*, b_j^*) \in \mathbb{R}^{d+1}$ and $\xi_i := (x_i, 1) \in \mathbb{R}^{d+1}$ to denote augmented parameters and covariates, respectively.

The model (1) appears in multiple contexts. The case $k = 1$ corresponds to linear regression, and setting $k = 2$ strictly generalizes the (real) phase retrieval problem: note that phase retrieval corresponds to the special case

$$Y = |\langle X, \theta^* \rangle| + \epsilon, \tag{2}$$

which can be obtained by setting $\theta_1^* = -\theta_2^*$ and $b_1 = b_2 = 0$. Also, since the function $x \mapsto \max_{1 \le j \le k} (\langle x, \theta_j^* \rangle + b_j^*)$ is always convex, estimation under the model (1) can be used to fit convex functions to data. Indeed, imposing such piece-wise affine structure is known to be an effective method by which to avoid the curse of dimensionality in applications of convex regression and its relatives [2]–[5].

[*]Avishek Ghosh and Ashwin Pananjady contributed equally to this work.

In the companion paper that forms the basis for this work [1], we studied the Gaussian design case $X \sim \mathcal{N}(0, I_d)$, and where the parameters $\theta_1^*, \ldots, \theta_k^* \in \mathbb{R}^d$ and $b_1^*, \ldots, b_k^* \in \mathbb{R}$ were fixed independently of the covariates. We proposed a computationally efficient, end-to-end algorithm for estimating the parameters $\{\beta_j^*\}_{j=1}^k$. This algorithm consisted of two steps: a spectral method to obtain an "initialization", and an alternating minimization (AM) algorithm [1]. The focus of the current paper is on the AM algorithm, and so in order to set the stage, let us describe iteration t of this algorithm in more detail. At the start of this iteration, we have the parameter estimates $\{\beta_j^{(t)}\}_{j=1}^k$. From these, we compute any partition of the n samples, indexed by the sets

$$S_j^{(t)} \equiv S_j(\beta_1^{(t)}, \ldots, \beta_k^{(t)}) \ \text{ for } 1 \le j \le k,$$

such that for each sample index $i \in S_j^{(t)}$, we have $\langle \xi_i, \beta_j^{(t)} \rangle \ge \max_{j \ne j'} \langle \xi_i, \beta_{j'}^{(t)} \rangle$. In order to obtain the next set of parameters, we compute a least squares solution within each partition, returning parameters $\{\beta_j^{(t+1)}\}_{j=1}^k$ such that

$$\beta_j^{(t+1)} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^{d+1}} \sum_{i \in S_j^{(t)}} (y_i - \langle \xi_i, \beta \rangle)^2 \quad \text{for each } 1 \le j \le k.$$

For a more formal description of the algorithm, see our companion paper [1]. The AM procedure is a heuristic to solve the non-convex optimization problem that arises from least squares estimation under the max-affine model, and resembles alternating update algorithms that have been designed in related contexts [7]–[9]. Recent theoretical investigations have shown that in spite of the non-convexity of many of these problems, such algorithms can exhibit favorable convergence properties in statistical settings [10]–[14]. Having said that, it is important to note that this is not always the case, and we often require the procedure to be initialized in a neighborhood of the optimal parameters defining the model [15]. Indeed, our own prior work showed that for the max-affine model (1), the AM algorithm converges linearly to the near-optimal statistical neighborhood of the true parameters for Gaussian ensembles, provided it is initialized with a spectral method.

In this paper, we show that a similar phenomenon occurs under significantly weaker statistical assumptions. In particular, we allow the distribution of the covariates to come from the larger class of sub-Gaussian distributions that satisfy a *small-ball* condition. In addition, we also consider the scenario of

[1]We note that the alternating minimization heuristic was proposed in the context of this problem by Magnani and Boyd [6]

universal parameter estimation, meaning that our guarantees hold uniformly over all $\beta_1^*, \ldots, \beta_k^*$. This allows the parameters to be chosen with knowledge of the realized covariates, a setting that is common in signal processing applications like phase retrieval [16]. In contrast, our prior work [1] only handled the case where the parameters are fixed.

More precisely, our covariate assumption relies on the following definition.

Definition 1. *(Small-ball) A distribution P_X satisfies a (ζ, c_s)-small-ball property if, for $X \sim P_X$ and each $\delta > 0$, we have*

$$\sup_{u \in \mathbb{S}^{d-1},\, w \in \mathbb{R}} \Pr\left\{(\langle X,\, u\rangle + w)^2 \leq \delta\right\} \leq (c_s \delta)^{\zeta}. \quad (3)$$

The small-ball properties of various classes of distributions have been studied extensively in the probability literature [17], [18]; for instance, a simple calculation yields that provided the density of $\langle X,\, u\rangle$ is bounded by $\sqrt{c}$ for each $u \in \mathbb{S}^{d-1}$, the distribution P_X satisfies the $(1/2, c)$-small ball property. We now present our assumption on the covariate distribution; recall that $X \in \mathbb{R}^d$ is said to be η-sub-Gaussian if

$$\sup_{u \in \mathbb{S}^{d-1}} \mathbb{E}[\exp(\lambda \langle X,\, u\rangle)] \leq \exp\left(\frac{\lambda^2 \eta^2}{2}\right) \text{ for each } \lambda \in \mathbb{R}.$$

Assumption 1. *The distribution P_X is isotropic, η-sub-Gaussian, and satisfies a (ζ, c_s) small-ball condition.*

Let us briefly state a few examples where Assumption 1 is satisfied with particular values of the tuple (η, ζ, c_s). The first is the class of compactly supported log-concave random vectors, which satisfy the the small ball conditions with $(\zeta, c_s) = (1/2, C)$ for an absolute constant C (see [19, Appendix G.1]). Boundedness further implies sub-Gaussianity. As a specific example, consider X with each entry drawn i.i.d. according to the distribution $\mathsf{Unif}[-\sqrt{3}, \sqrt{3}]$, which is commonly used as a random design in investigations of non-parametric regression problems [20]. The associated distribution P_X is isotropic by definition, and has $(\eta, \zeta, c_s) = (12, 1/2, C)$. Similarly, any other uniform distribution on a bounded, isotropic convex set would also satisfy Assumption 1. The second (canonical) example for which Assumption 1 is satisfied is the standard Gaussian distribution. As we verify in [19, Appendix G.2] with χ^2 tail bounds, the standard Gaussian satisfies $(\eta, \zeta, c_s) = (1, 1/2, e)$.

Thus, Assumption 1 is strictly more general than the Gaussian covariate assumption. It is also important to note that Assumption 1 allows a larger class of distributions than even log-concave distributions; heuristically speaking, the small-ball condition only disallows distributions that are significantly "peakier" than the Gaussian distribution, and the sub-Gaussian condition disallows heavy-tailed distributions. Also, while we have only presented examples in which $\zeta = 1/2$, there are distributions that satisfy the small ball condition for other values of ζ: for example, any random variable with density $f(x) \propto e^{-\|x\|^c}$ for a positive constant c.

In order to make our guarantees on universal parameter estimation more clear, let us define a few geometric quantities induced by the model (1). For $X \sim P_X$, let

$$\pi_j(\beta_1^*, ., \beta_k^*) := \Pr\{\langle X,\, \theta_j^*\rangle + b_j^* = \max_{j' \in [k]} (\langle X,\, \theta_{j'}^*\rangle + b_{j'}^*)\},$$

and define

$$\pi_{\min}(\beta_1^*, \ldots, \beta_k^*) := \min_{j \in [k]} \pi_j(\beta_1^*, \ldots, \beta_k^*).$$

For a fixed set of true parameters, the quantity $\pi_{\min}(\beta_1^*, \ldots, \beta_k^*) \cdot n$ is the expected number of samples that are noisy linear combinations of one of these parameters. Thus, even if the underlying parameters are fixed, we can only hope to estimate these parameters when $\pi_{\min}(\beta_1^*, \ldots, \beta_k^*)$ is sufficiently large.

The signal strength of the problem is the minimum separation

$$\Delta(\beta_1^*, \ldots, \beta_k^*) = \min_{j, j' : j \neq j'} \left\|\theta_j^* - \theta_{j'}^*\right\|^2.$$

We also define a notion of condition number, given by

$$\kappa(\beta_1^*, \ldots, \beta_k^*) = \max_{j \in [k]} \frac{\max_{j' \neq j} \left\|\theta_j^* - \theta_{j'}^*\right\|^2}{\min_{j' \neq j} \left\|\theta_j^* - \theta_{j'}^*\right\|^2}.$$

We often use the shorthand

$$\pi_{\min} = \pi_{\min}(\beta_1^*, \ldots, \beta_k^*), \;\; \Delta = \Delta(\beta_1^*, \ldots., \beta_k^*), \text{ and } \kappa = \kappa(\beta_1^*, \ldots, \beta_k^*).$$

when the true parameters $\beta_1^*, \ldots, \beta_k^*$ are clear from context.

Recall that our goal was to prove a result that holds uniformly for all true parameters $\{\beta_j^*\}_{j=1}^k$. However, this is clearly impossible in a general sense, since we cannot hope to obtain consistent estimates if some parameters are never observed in the sample. A workaround is to hold certain geometric quantities fixed while sweeping over all possible allowable parameters $\beta_j^*, j = 1, \ldots, k$. Accordingly, for each triple of positive scalars (π, Δ, κ), we define the set of "admissible" true parameters as

$$\begin{aligned}\mathsf{B}_{\mathsf{vol}}(\pi, \Delta, \kappa) = \{&\beta_1, \ldots, \beta_k : \pi_{\min}(\beta_1, \ldots, \beta_k) \geq \pi\,, \\ &\Delta(\beta_1, \ldots, \beta_k) \geq \Delta, \kappa(\beta_1, \ldots, \beta_k) \leq \kappa\}.\end{aligned}$$

We let the true parameters $\beta_1^*, \ldots, \beta_k^*$ take values in the set $\mathsf{B}_{\mathsf{vol}}(\pi, \Delta, \kappa)$, and prove guarantees uniformly over all such $\beta_1^*, \ldots, \beta_k^*$.

With these definitions at hand, we are now ready to discuss our contributions.

Contributions: Suppose Assumption 1 holds, and $\{\beta_j^{(t)}\}_{j=1}^k$ are the parameter estimates returned by the AM algorithm at the t-th iteration. In Theorem 1, we show that for any $\epsilon > 0$, there is a sufficiently large iteration t such that

$$\sum_{i=1}^k \|\beta_i^{(t)} - \beta_i^*\|^2 \leq \epsilon + C^{(1)}_{\eta, \zeta, c_s} \frac{\sigma^2 kd}{n \pi_{\min}^{1+2\zeta^{-1}}} \log(kd) \log\left(\frac{n}{kd}\right).$$

Such a result holds *simultaneously* for all $(\beta_1^*, \ldots, \beta_k^*) \in \mathsf{B}_{\mathsf{vol}}(\pi_{\min}, \Delta, \kappa)$ with high probability provided the sample size is large enough and the initialization is chosen close enough to the true parameters.

In Corollary 1, we specialize this result to the phase retrieval problem, showing that the AM algorithm exhibits linear convergence provided the covariates (called "measurements" in the signal processing literature) are drawn from an isotropic, sub-Gaussian distribution satisfying the small-ball condition. Since our result holds for universal parameter estimation, we allow the underlying "signal" to be chosen adversarially, with knowledge of the measurements. Such a robust setting is common for phase retrieval problems. However, to the best of our knowledge, all previous results on the AM algorithm for phase retrieval [21], [22] only held under the assumptions of Gaussian covariates and noiseless observations, and/or required resampling of the measurements [14]. Ours is thus the first work to handle non-Gaussian covariates in the presence of noise, while also analyzing the algorithm without resampling.

Notation: For a positive integer n, let $[n] := \{1, 2, \ldots, n\}$. For a finite set S, we use $|S|$ to denote its cardinality. All logarithms are to the natural base. For two sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n \lesssim b_n$ if there is a universal constant C such that $a_n \leq Cb_n$ for all $n \geq 1$. The relation $a_n \gtrsim b_n$ is defined analogously, and we use $a_n \sim b_n$ to indicate that both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold simultaneously. We use $c, C, c_1, c_2, \ldots$ to denote universal constants that may change from line to line, but do not depend on any of the problem parameters. We use $\|\cdot\|$ to denote the ℓ_2 norm. Denote by I_d the $d \times d$ identity matrix. Let $\mathsf{sgn}(t)$ denote the sign of a scalar t, with the convention that $\mathsf{sgn}(0) = 1$.

II. Main results

Let us now state and discuss our results in precise terms.

A. Local geometric convergence of alternating minimization

For each pair $1 \leq i \neq j \leq k$ and $t \geq 0$, we use the shorthand $v_{i,j}^* = \beta_i^* - \beta_j^*$ and $v_{i,j}^{(t)} = \beta_i^{(t)} - \beta_j^{(t)}$ to denote the pairwise differences between parameters.

Theorem 1. *Suppose that Assumption 1 holds. Then there exists a pair of universal constants (c_1, c_2) and constants $(C^{(1)}_{\eta,\zeta,c_s}, C^{(2)}_{\eta,\zeta,c_s})$ depending only on the triple (η, ζ, c_s) such that if the sample size satisfies the bound*

$$n \geq C^{(1)}_{\eta,\zeta,c_s} \max\{d, 10 \log n\} \cdot \frac{k\kappa}{\pi_{\min}^{1+2\zeta^{-1}}} \log(n/d) \quad (4)$$

then simultaneously for all true parameters $\beta_1^*, \ldots, \beta_k^* \in \mathsf{B}_{\mathsf{vol}}(\pi_{\min}, \Delta, \kappa)$ and all initializations satisfying

$$\min_{c>0} \max_{1 \leq j \neq j' \leq k} \frac{\left\| c v_{j,j'}^{(0)} - v_{j,j'}^* \right\|}{\|\theta_j^* - \theta_{j'}^*\|} \leq C^{(2)}_{\eta,\zeta,c_s} \left(\frac{\pi_{\min}^{1+2\zeta^{-1}}}{k\kappa}\right)^{\zeta^{-1}} \times \left[\log^{1+\zeta^{-1}}\left(\frac{k\kappa}{\pi_{\min}^{1+2\zeta^{-1}}}\right)\right], \quad (5a)$$

the estimation error at all iterations $t \geq 1$ satisfies

$$\sum_{j=1}^{k} \|\beta_j^{(t)} - \beta_j^*\|^2 \leq \left(\frac{3}{4}\right)^t \left(\sum_{j=1}^{k} \|c^* \beta_j^{(0)} - \beta_j^*\|^2\right) + C^{(1)}_{\eta,\zeta,c_s} \cdot \sigma^2 \frac{kd}{n\pi_{\min}^{1+2\zeta^{-1}}} \log(kd) \log(n/kd) \quad (5b)$$

with probability exceeding $1 - c_1 \left\{ \frac{k^2}{n^7} + \exp\left(-c_2 n \pi_{\min}^2\right) \right\}$. Here, $c^*(> 0)$ minimizes the LHS of inequality (5a)*.*

The proof of the theorem has been omitted due to space constraints, and can be found in [19, Appendix B]. Let us discuss the initialization conditions of the theorem in more detail. We require the initialization $\beta_1^{(0)}, \ldots, \beta_k^{(0)}$ to satisfy condition (5a). In the well-balanced case (with $\pi_{\min} \sim 1/k$) and treating k as a fixed constant, the initialization condition (5a) posits that the parameters are a constant "distance" from the true parameters. Closeness here is measured in a relative sense, i.e., between pairwise differences of the parameter estimates as opposed to the parameters themselves. The intuition for this is that $\beta_1^{(0)}, \ldots, \beta_k^{(0)}$ induces a partition of samples $S_1(\beta_1^{(0)}, \ldots, \beta_k^{(0)}), \ldots, S_1(\beta_1^{(0)}, \ldots, \beta_k^{(0)})$, and the closeness to this to the true partition depends only on the relative pairwise differences between parameters. Furthermore, the initialization condition is also invariant to a global scaling of the parameters, since scaling does not change the initial partition of samples. Also note that the geometric convergence guarantee (5b) holds uniformly *for all* initializations satisfying condition (5a). Hence, the initialization parameters are not additionally required to be independent of the covariates or noise. This allows us to use the same n samples for initialization of the parameters.

Let us now turn our attention to the bound (5b), which consists of two terms. When $t \to \infty$, the second term of the bound (5b) provides an estimate of the closeness of the final parameters to the true parameters. Up to a constant, this is the statistical error term

$$\delta^{\mathsf{sb}}_{n,\sigma}(d, k, \pi_{\min}) = \sigma^2 \frac{kd}{\pi_{\min}^{1+2\zeta^{-1}} n} \log(kd) \log(n/kd) \quad (6)$$

that converges to 0 as $n \to \infty$, thereby providing a consistent estimate in the large sample limit.

The first term of (5b) is an optimization error that is best interpreted in the noiseless case $\sigma = 0$, wherein the parameters $\beta_1^{(t)}, \ldots, \beta_k^{(t)}$ converge at a geometric rate to the true parameters $\beta_1^*, \ldots, \beta_k^*$. In the noiseless case, we obtain exact recovery of the parameters provided

$$n \geq C \frac{kd}{\pi_{\min}^{1+2\zeta^{-1}}} \log(n/d).$$

Thus, the "sample complexity" of parameter recovery is linear in the dimension d, which is optimal. In the well-balanced case, we require $n \sim k^{2+2\zeta^{-1}} d$, but lower bounds based on parameter counting suggest that the true dependence ought to be linear. We are not aware of whether the dependence on $\pi_{\min}$ in the noiseless case is optimal; our simulations in panel (a) suggest that the sample complexity depends inversely on $\pi_{\min}$, and so closing this gap is an interesting open problem.

In Figure 1, we verify that for independent, isotropic covariates chosen uniformly from a symmetric interval[2], intitializing the AM algorithm in a neighborhood of the true parameters suffices to ensure that it converges to the true parameters. Furthermore, both the sample size requirement and final error of the algorithm exhibit the behaviors predicted by Theorem 1.

[2]Such a distribution is compactly supported and log-concave, and therefore satisfies Assumption 1

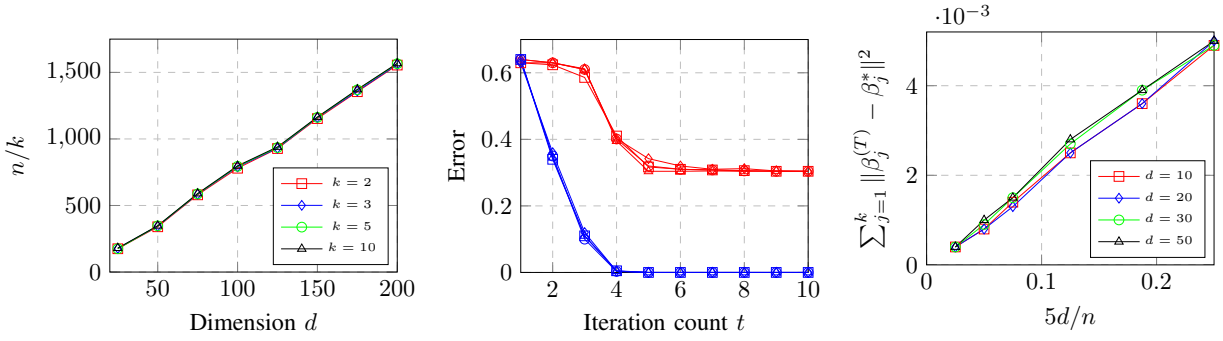

Fig. 1. Convergence of AM when the covariates are drawn i.i.d. from the distribution $\mathsf{Unif}[-\sqrt{3}, \sqrt{3}]^{\otimes d}$—in panel (a), we plot the noiseless sample complexity of AM; we fix $\|\beta_i^*\| = 1$ for all $i \in [k]$, $\sigma = 0$ and $\pi_{\min} = 1/k$. We say β_i^* is recovered if $\left\|\beta_i^{(t)} - \beta_i^*\right\| \leq 0.01$. For a fixed dimension d, we run a linear search on the number of samples n, such that the empirical probability of success over 100 trials is more than 0.95, and output the least such n. In panel (b), we plot the optimization error (in blue) $\sum_{j=1}^{k} \|\beta_j^{(t)} - \beta_j^{(T)}\|^2$ and the deviation from the true parameters (in red) $\sum_{j=1}^{k} \left\|\beta_j^{(t)} - \beta_j^*\right\|^2 / \sigma^2$ over iterations t for different σ $(0.15, 0.25, 0.4, 0.5)$, with $k = 5$, $d = 100$, $T = 50$ and $n = 5d$, and averaged over 50 trials. Panel (c) shows that the estimation error at $T = 50$ scales at the parametric rate d/n, where we have chosen a fixed $k = 5$ and $\sigma = 0.25$.

We now compare Theorem 1 with [1, Theorem 1] for the special case of Gaussian covariates, where $\eta = 1$ and $\zeta = 1/2$. In this case, all terms of the form $\pi_{\min}^3$ in [1, Theorem 1] are replaced by terms of the form $\pi_{\min}^5$. In particular, we see that the initialization condition (5a) is more stringent and the final statistical rate of the estimate (corresponding to the limit $t \to \infty$) now attains an estimation error that is a factor $\pi_{\min}^{-2}$ higher than the corresponding rate of [1, Theorem 1]. The sample size requirement is similarly affected. On the other hand, the geometric convergence result (5b) now holds uniformly for all true parameters $\beta_1^*, \ldots, \beta_k^* \in \mathsf{B}_{\mathsf{vol}}(\pi_{\min}, \Delta, \kappa)$, as opposed to [1, Theorem 1], which holds only when the true parameters are held fixed. The more stringent initialization condition and sample size requirements can be viewed as the price to pay for the more robust convergence of the AM algorithm. Notably, the dependence on all other parameters remains unchanged.

B. Consequences for phase retrieval

A notable consequence of Theorem 1 is that it can be applied to the phase retrieval model (2)—in which results are usually proved uniformly over all true parameters [16], [23]—to yield a convergence result under general distributional assumptions on the covariates. In particular, setting $\pi_{\min} = 1/2$ and $k = 2$ yields a local linear convergence result for the AM algorithm of the Gershberg-Saxton-Fienup type (presented for completeness as Algorithm 1) uniformly *for all* θ^* provided the covariates satisfy a small-ball condition. We note that other algorithms for phase retrieval have also been shown to succeed under such small-ball assumptions [24], [25].

Corollary 1. *Suppose that Assumption 1 holds. There exists a universal constant c_1 and a pair of constants $(C^{(1)}_{\eta,\zeta,c_s}, C^{(2)}_{\eta,\zeta,c_s})$ depending only on (η, ζ, c_s) such that if*

$$n \geq C^{(1)}_{\eta,\zeta,c_s} \max\{d, 10 \log n\} \log(n/d),$$

then simultaneously for all true parameters $\theta^* \in \mathbb{R}^d$ and all initializations $\theta^{(0)}$ satisfying

$$\min_{c>0} \min_{s \in \{-1,1\}} \frac{\left\|c\theta^{(0)} - s\theta^*\right\|}{\|\theta^*\|} \leq C^{(2)}_{\eta,\zeta,c_s}, \tag{7a}$$

the estimation error for all $t \geq 1$ satisfies

$$\min_{s \in \{-1,1\}} \|\theta^{(t)} - s\theta_j^*\|^2 \leq \left(\frac{3}{4}\right)^t \min_{s \in \{-1,1\}} \|\theta^{(0)} - s\theta^*\|^2 + \frac{c_1 \sigma^2 d \log(n/d)}{n}$$

with probability exceeding $1 - c_1 n^{-7}$.

The proof of the corollary follows almost immediately from Theorem 1, and a full argument can be found in [19, Appendix C]. Let us now compare this with the sharpest existing local convergence result of AM for phase retrieval due to Waldspurger [21], which holds for Gaussian covariates and in the noiseless setting[3]. Specializing Corollary 1 to the noiseless setting, we observe that provided the ratio n/d is larger than a fixed constant (that depends only on (η, ζ, c_s)), we obtain exact recovery of the underlying parameter, up to a global sign, with high probability provided the covariates (or "measurement vectors" as they are called in the signal processing literature) are sub-Gaussian and satisfy a small-ball condition. To the best of our knowledge, prior work on the AM algorithm had not established provable guarantees for non-Gaussian covariates even in the noiseless setting. In the noisy case, Corollary 1 guarantees convergence of the iterates to a small neighborhood around either θ^* or $-\theta^*$, and the size of this neighborhood is within a logarithmic factor of being minimax optimal [16], [26]. Once again, to the best of our knowledge, guarantees for the AM algorithm as applied to noisy phase retrieval did not exist in the literature.

C. Proof ideas and technical challenges

Let us sketch, at a high level, the main ideas required to establish guarantees on the AM algorithm. Note that we analyze AM without sample splitting across iterations, and hence the iterates depend on the data points $\{\xi_i, y_i\}_{i=1}^n$. One standard way to address this issue (see [13]) is to (a) first analyze the *population* updates assuming $n \to \infty$, and (b)

[3]Waldspurger [21] deals with the complex phase retrieval, whose analysis is significantly more complicated than real phase retrieval considered here.

Algorithm 1: Alternating minimization for real phase retrieval

Input: Data $\{x_i, y_i\}_{i=1}^n$; initial parameter estimate $\theta^{(0)} \in \mathbb{R}^d$; number of iterations T.

Output: Final estimator $\widehat{\theta}$.

Initialize $t \leftarrow 0$.

repeat

Compute sign vector $s^{(t)}$ with i-th entry as

$$s_i^{(t)} = \mathsf{sgn}(\langle x_i, \theta^{(t)} \rangle) \quad \text{for each } i \in [n]. \tag{8a}$$

Update

$$\theta^{(t+1)} = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} (y_i - s_i^{(t)} \langle x_i, \theta \rangle)^2. \tag{8b}$$

until $t = T$;

Return $\widehat{\theta} = \theta^{(T)}$.

use concentration inequalities to show that the updates based on the observed samples are close to the population updates. Unfortunately, in our setting, handling the population updates is quite non-trivial, since it requires understanding the geometry of the covariate distribution induced by the maxima of affine functions. Hence, we work with the random iterates directly.

We analyze the sample-based update of the AM algorithm by relating the error of the parameters generated by this update to the error of the parameters from which the update is run. This involves three steps. First, we control the behavior of the noise using standard concentration bounds for quadratic forms of sub-Gaussian random variables, along with bounds on the growth function of multi-class linear classifiers [27]. We then control the prediction error of the noiseless problem, and this step crucially uses the small-ball condition satisfied by the covariates along with the initialization condition. Finally, in order to translate our bounds on the prediction error into bounds on parameter estimation, we invert specifically chosen sub-matrices of the covariate matrix over the course of the algorithm. Our bounds naturally depend on how these sub-matrices are conditioned. A key technical difficulty of the proof is to control the spectrum of these random matrices, rows of which are drawn from (randomly) truncated variants of the covariate distribution. Our techniques for controlling the spectrum of these matrices is more broadly applicable, and we expect this result to be of broader interest.

III. Future Work

Owing to the local nature of our guarantees, a natural open question is how to obtain a good initialization for the AM algorithm. When the covariates are Gaussian, our prior work [1] showed that a natural spectral method combined with random search in lower-dimensional space was able to provide such an initialization. However, this relied heavily on Gaussianity; how can such ideas be extended to general small-ball distributions? In a complementary direction, understanding the behavior of the randomly initialized AM algorithm is a known open problem in the context of phase retrieval [21], [28]; is this convergence robust to distributional assumptions? Finally, can our estimation procedures for the max-affine model (1) be used to produce estimators for convex regression and its relatives?

References

[1] A. Ghosh, A. Pananjady, A. Guntuboyina, and K. Ramchandran, "Max-affine regression I: Parameter estimation for Gaussian designs," 2019. [Online]. Available: https://tinyurl.com/spyjmgv

[2] J. Gregor and F. R. Rannou, "Three-dimensional support function estimation and application for projection magnetic resonance imaging," *International journal of imaging systems and technology*, vol. 12, no. 1, pp. 43–50, 2002.

[3] R. J. Gardner, *Geometric Tomography*, 2nd ed., ser. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2006.

[4] Q. Han and J. A. Wellner, "Multivariate convex regression: global risk bounds and adaptation," *arXiv preprint arXiv:1601.06844*, 2016.

[5] Y. S. Soh and V. Chandrasekaran, "Fitting tractable convex sets to support function evaluations," *arXiv preprint arXiv:1903.04194*, 2019.

[6] A. Magnani and S. P. Boyd, "Convex piecewise-linear fitting," *Optimization and Engineering*, vol. 10, no. 1, pp. 1–17, 2009.

[7] E. M. Beale and R. J. Little, "Missing values in multivariate analysis," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 37, no. 1, pp. 129–145, 1975.

[8] R. Gerchberg and W. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," *Optik*, vol. 35, pp. 237–246, 1972.

[9] J. R. Fienup, "Phase retrieval algorithms: a comparison," *Applied optics*, vol. 21, no. 15, pp. 2758–2769, 1982.

[10] C. J. Wu, "On the convergence properties of the EM algorithm," *The Annals of statistics*, vol. 11, no. 1, pp. 95–103, 1983.

[11] P. Tseng, "An analysis of the EM algorithm and entropy-like proximal point methods," *Math. of Operations Res.*, vol. 29, pp. 27–44, 2004.

[12] S. Chrétien and A. O. Hero, "On EM algorithms and their proximal generalizations," *ESAIM: Prob. and Stats.*, vol. 12, pp. 308–326, 2008.

[13] S. Balakrishnan, M. J. Wainwright, and B. Yu, "Statistical guarantees for the EM algorithm: From population to sample-based analysis," *The Annals of Statistics*, vol. 45, no. 1, pp. 77–120, 2017.

[14] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Advances in Neural Information Processing Systems*, 2013, pp. 2796–2804.

[15] C. Jin, Y. Zhang, S. Balakrishnan, M. J. Wainwright, and M. I. Jordan, "Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences," in *Advances in neural information processing systems*, 2016, pp. 4116–4124.

[16] Y. Chen and E. Candes, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," in *Advances in Neural Information Processing Systems*, 2015, pp. 739–747.

[17] G. Paouris, "Small ball probability estimates for log-concave measures," *Transactions of the American Mathematical Society*, vol. 364, no. 1, pp. 287–308, 2012.

[18] M. Rudelson and R. Vershynin, "Small ball probabilities for linear images of high-dimensional distributions," *International Mathematics Research Notices*, vol. 2015, no. 19, pp. 9594–9617, 2014.

[19] A. Ghosh, A. Pananjady, K. Ramchandran, and A. Guntuboyina, "Max-affine regression: Provable, tractable, and near-optimal statistical estimation," *arXiv preprint arxiv:1906.09255*, 2019.

[20] L. Wasserman, *All of nonparametric statistics*. Springer Science & Business Media, 2006.

[21] I. Waldspurger, "Phase retrieval with random Gaussian sensing vectors by alternating projections," *IEEE Transactions on Information Theory*, vol. 64, no. 5, pp. 3301–3312, May 2018.

[22] T. Zhang, "Phase retrieval using alternating minimization in a batch setting," *Applied and Computational Harmonic Analysis*, 2019.

[23] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via wirtinger flow: Theory and algorithms," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1985–2007, 2015.

[24] Y. C. Eldar and S. Mendelson, "Phase retrieval: Stability and recovery guarantees," *Applied and Computational Harmonic Analysis*, vol. 36, no. 3, pp. 473–494, 2014.

[25] J. C. Duchi and F. Ruan, "Solving (most) of a set of quadratic equalities: composite optimization for robust phase retrieval," *Information and Inference: A Journal of the IMA*, vol. 8, no. 3, pp. 471–529, 09 2018. [Online]. Available: https://doi.org/10.1093/imaiai/iay015

[26] G. Lecué and S. Mendelson, "Minimax rate of convergence and the performance of empirical risk minimization in phase recovery," *Electron. J. Probab.*, vol. 20, p. 29 pp., 2015. [Online]. Available: https://doi.org/10.1214/EJP.v20-3525

[27] A. Daniely, S. Sabato, and S. S. Shwartz, "Multiclass learning approaches: A theoretical comparison with implications," in *Advances in Neural Information Processing Systems*, 2012, pp. 485–493.

[28] T. Zhang, "Phase retrieval by alternating minimization with random initialization," *IEEE Transactions on Information Theory*, 2020.