Agnostic Learning of Mixed Linear Regressions with EM and AM Algorithms

Avishek Ghosh [1] **Arya Mazumdar** [2]

Abstract

Mixed linear regression is a well-studied problem in parametric statistics and machine learning. Given a set of samples, tuples of covariates and labels, the task of mixed linear regression is to find a small list of linear relationships that best fit the samples. Usually it is assumed that the label is generated stochastically by randomly selecting one of two or more linear functions, applying this chosen function to the covariates, and potentially introducing noise to the result. In that situation, the objective is to estimate the ground-truth linear functions up to some parameter error. The popular expectation maximization (EM) and alternating minimization (AM) algorithms have been previously analyzed for this.

In this paper, we consider the more general problem of agnostic learning of mixed linear regression from samples, without such generative models. In particular, we show that the AM and EM algorithms, under standard conditions of separability and good initialization, lead to agnostic learning in mixed linear regression by converging to the population loss minimizers, for suitably defined loss functions. In some sense, this shows the strength of AM and EM algorithms that converges to "optimal solutions" even in the absence of realizable generative models.

1. Introduction

Suppose we obtain samples from a data distribution $\mathcal{D}$ on $\mathbb{R}^{d+1}$, i.e., $\{x_i,y_i\}\sim\mathcal{D}$, $x_i\in\mathbb{R}^d,y_i\in\mathbb{R},i=1,...,n$. We consider the problem of learning a list of k $\mathbb{R}^d\to\mathbb{R}$ linear functions $y=\theta_j^T x,\theta_j\in\mathbb{R}^d,j=1,...,k$, that best fits the samples.

This problem is well-studies as the *mixed linear regression*, when there are ground-truth $\tilde{\theta}_j,j=1,...,k$, that generate the samples. For example, the setting where

$$x_i\sim\mathcal{N}(0,I_d),\theta\sim\text{Unif}\{\tilde{\theta}_1,...,\tilde{\theta}_k\},y_i|\theta\sim\mathcal{N}(x^T\theta,\sigma^2), \quad (1)$$

for $i=1,\dots,n$ has been analyzed thoroughly. Bounds on sample complexity are provided in terms of d,σ^2 and error in estimating parameters $\tilde{\theta}_j,j=1,...,k$ ((Chaganty & Liang, 2013; Faria & Soromenho, 2010; Städler et al., 2010; Li & Liang, 2018; Kwon & Caramanis, 2018; Viele & Tong, 2002; Yi et al., 2014; 2016; Balakrishnan et al., 2017; Klusowski et al., 2019)).

In this paper, we consider an agnostic and general learning theoretic setup to study the mixed linear regression problem first studied in (Pal et al., 2022). In particular, we do not assume a generative model on the samples. Instead we focus on finding the optimal set of lines that minimize a certain loss.

Suppose, we denote a loss function $\ell:\mathbb{R}^{d\times k}\to\mathbb{R}$ evaluated on a sample as $\ell(\theta_1,\theta_2,...,\theta_k;x,y)$. The population loss is

$$\mathcal{L}(\theta_1,\theta_2,...,\theta_k)\equiv\mathbb{E}_{(x,y)\sim\mathcal{D}}\ell(\theta_1,\theta_2,...,\theta_k;x,y),$$

and the population loss minimizers

$$(\theta_1^*,...,\theta_k^*)\equiv\text{argmin }\mathcal{L}(\theta_1,\theta_2,...,\theta_k).$$

Learning in this setting makes sense if we are allowed to predict a *list* (of size k) of labels for an input, as pointed out in (Pal et al., 2022). We may set some goodness criteria, such as an weighted average of prediction error over all elements in the list. In (Pal et al., 2022), it was called a 'good' prediction if at least one of the labels in the list is good, in particular, the following loss function was proposed, that we will call *min-loss*:

$$\ell_{\min}(\theta_1,\theta_2,...,\theta_k;x,y)=\min_{j\in[k]}\left\{(y-\langle x,\theta_j\rangle)^2\right\}. \quad (2)$$

The intuition behind min-loss is simple. Each sample is assigned to a best-fit line, which define a partition of the samples. This is analogous to the popular k-means clustering objective. In addition to the min-loss function, we will also

[1]Systems and Control Engg. and Centre for Machine Intelligence for Data Sciences, Indian Institute of Technology, Bombay, India. [2]Halıcıoğlu Data Science Institute, University of California, San Diego. Correspondence to: Avishek Ghosh <avishek.ghosh38@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

consider the following *soft-min* loss function:

$$\ell_{\text{softmin}}(\theta_1,\theta_2,...,\theta_k;x,y)=\sum_{j=1}^{k}p_{\theta_1,..,\theta_k}(x,y;\theta_j)[y-\langle x,\theta_j\rangle]^2, \tag{3}$$

$$\text{where} \quad p_{\theta_1,..,\theta_k}(x,y;\theta_j)=\frac{e^{-\beta(y-\langle x,\theta_j\rangle)^2}}{\sum_{l=1}^{k}e^{-\beta(y-\langle x,\theta_l\rangle)^2}}$$

with $\beta\geq 0$ as the inverse temperature parameter. Note that, at $\beta\to\infty$, this loss function correspond to the min-loss defined above. On the other hand, at $\beta=0$, this is simply an average of the squared errors, if a label is uniformly chosen from the list. Depending on how the prediction would occur, the loss function, and therefore the best-fit lines $\theta_1^*,...,\theta_k^*$ will change.

As is the usual case in machine learning, a learner has access to the distribution $\mathcal{D}$ only through the samples $\{x_i,y_i\},i=1,...,n$. Therefore instead of the population loss, one may attempt to minimize the empirical loss:

$$L(\theta_1,...,\theta_k)\equiv\frac{1}{n}\sum_{i=1}^{n}\ell(\theta_1,\theta_2,...,\theta_k;x_i,y_i).$$

Usual learning theoretic generalization bounds on excess risk should hold provided the loss function satisfies some properties[1]. However, there are certain caveats in solving the empirical loss minimization problem. For example, even the presumably simple case of squared error (Eq.(2)), the minimization problem is NP-hard, by reduction to the subset sum problem (Yi et al., 2014).

An intuitive and generic iterative method that is widely-applicable for problems with latent variables (in our case, which line is best fit for a sample) is the *alternating minimization* (AM) algorithm. At a very high level, starting from some initial estimate of the parameters, the AM algorithm first tries to find a partition of samples according to the current estimate, and then finds the best fit lines within each part. Again under the generative model of (1), AM can approach the original parameters assuming suitable initialization (Yi et al., 2014).

Another popular method of solving mixed regression problems (or in general mixture models) is the well-known *expectation maximization* (EM) algorithm. EM is an iterative algorithm that, starting from an initial estimate of parameters, iteratively update the estimates based on data, by taking an expectation-step and maximization-step repeatedly. For example, it was shown in (Balakrishnan et al., 2017) that, under the assumption of the generative model that was defined in Eq. (1), one can give guarantees on recovering the ground-truth parameters $\tilde{\theta}_1,...,\tilde{\theta}_k$ assuming a suitable initialization.

[1]Some discussions on generalization with soft-min loss can be found in Section 5.

In this paper, we show that the AM and the EM algorithms are in fact more powerful in the sense that even in the absence of a generative model, they lead to agnostic learning of parameters. It turns out, under standard assumptions on data-samples and $\mathcal{D}$, these iterative methods can output the minimizers of the population loss θ_1^*, ... , θ_k^* with appropriately defined loss functions. In particular, starting from reasonable initial points, the estimates of the AM algorithm approach $\theta_1^*,...,\theta_k^*$ under the min-loss (Eq. 2), and the estimates of the EM algorithm approach the minimizers of the population loss under the soft-min loss (Eq. 3).

Instead of the standard AM (or EM), a version that has been referred to as *gradient EM* (and *gradient AM*) is also popular and has been analyzed in (Balakrishnan et al., 2017; Zhu et al., 2017; Wang et al., 2020; Pal et al., 2022) to name a few. Here, in lieu of the maximization step involved in EM (minimization for AM), a gradient step with appropriately chosen step size is taken. This version is amenable to analysis and is strictly worse than the actual EM (or AM) in their generative setting. In this paper as well, we analyze the gradient EM algorithm, and the analogous gradient AM algorithm.

Recently (Pal et al., 2022) proposed a gradient AM algorithm for the agnostic mixed linear regression problem. However, they require a strong assumption on initialization of $\{\theta_i\}_{i=1}^k$ within a radius of $\mathcal{O}(\frac{1}{\sqrt{d}})$ of the corresponding $\{\theta_i^*\}_{i=1}^k$. As we can see, in high dimension, the initialization condition is prohibitive. The dimension dependence initialization in (Pal et al., 2022) comes from a discretization (ϵ-net) argument, which was crucially used to remove inter-iteration dependence of the gradient AM algorithm.

In this paper, we show that a dimension independent initialization is sufficient for gradient AM. In particular, we showed that the initialization needed for $\{\theta_i\}_{i=1}^k$ is $\Theta(1)$, which is a significant improvement over the past work (Pal et al., 2022). Instead of an ϵ-net argument, we use fresh samples every round. Moreover, we thoroughly analyze the behavior of restricted covariates on a (problem defined) set, in the agnostic setup, which turns out to be non-trivial. In particular, we observe that the restricted covariates are sub Gaussian with a *shifted mean* and variance, and we need to control the minimum singular value of the covariance matrix of such restricted covariates (which dictates the convergence rate). We leverage some properties of restricted distributions (Tallis, 1961), and were able to analyze such covariates rigorously, obtain bounds and show convergence of AM.

In this paper we also propose and analyze the soft variant of gradient AM, namely gradient EM. As discussed above, the associated loss function is the *soft-min* loss. We show that gradient EM also requires dimension independent $\mathcal{O}(1)$ initialization, and also converges in an exponential rate.

While the performance of both the gradient AM and gradient

EM algorithms are similar, AM minimizes a min-loss whereas EM minimizes the optimal soft-min loss (maximum likelihood loss in the generative setup). As shown in the subsequent sections, AM requires a separation condition (appropriately defined in Theorem 2.1) whereas EM does not. On the other hand, EM requires the initialization parameter to satisfy certain condition, albeit mild (exact condition in Theorem 3.1).

1.1. Setup and Geometric Parameters

Recall that the parameters $\theta_1^*,\dots,\theta_k^*$ are the minimizers of the population loss function, and we consider both *min-loss* ($\ell_{\min}(.)$) as well as *soft-min* loss ($\ell_{\text{softmin}}(.)$) as defined in the previous section. We define

$$S_j^* = \{(x\in\mathbb{R}^d, y\in\mathbb{R}) : (y-\langle x,\theta_j^*\rangle)^2 < (y-\langle x,\theta_l^*\rangle)^2,$$

for all $l\in[k]\setminus j\}$ as the possible set of observations where θ_j^* is a better (linear) predictor (in ℓ_2 norm) compared to $\theta_1^*,\dots,\theta_k^*$. Furthermore, in order to avoid degeneracy, we assume, for any $j\in[k]$

$$\Pr_{\mathcal{D}}(x:(x,y)\in S_j^*)\geq \pi_{\min},$$

for some $\pi_{\min} > 0$. We are interested in the probability measure corresponding to the random vector x only, and we integrate (average-out) with respect to y to achieve this. We emphasize that, in the realizable setup, the distribution of y is governed by that of x (and possibly some noise independent of x), and in that setting our definition of S_j^* and $\pi_{\min}$ becomes analogous to that of (Yi et al., 2014; 2016)[2].

Since we are interested in recovering $\theta_j^*, j=1,\dots,k$, a few geometric quantities naturally arises in our setup. We define the *misspecification* parameter λ as a smallest non-negative number satisfying

$$|y_i-\langle x_i,\theta_j^*\rangle|\leq\lambda \quad \text{for all } (x_i,y_i)\in S_j^* \quad \text{and } j\in[k].$$

Moreover, we also define the *separation* parameter Δ as the largest non-negative number satisfying

$$\min_{l\in[k]\setminus j}|y_i-\langle x_i,\theta_l^*\rangle|\geq\Delta \quad \text{for all } (x_i,y_i)\in S_j^*.$$

Let us comment on these geometric quantities. Note that in the case of a realizable setup, the parameter $\lambda=0$ in the noiseless case or proportional to the noise in the noisy case. In words, λ captures the level of misspecification from the linear model. On the other hand, the parameter Δ denotes the separation or margin in the problem. In classical mixture of linear regression framework, with realizable structure, similar assumptions are present in terms of the (generative) parameters. Moreover, with the realizable setup, our assumption can be shown to be *exactly* same as the usual separation assumption.

[2] In (Yi et al., 2014; 2016), the authors denote $\{S_j^*\}_{j=1}^k$ as set of indices, but that can be thought of as an analogue to a subset of $\mathbb{R}^{d+1}$ as shown above.

1.2. Summary of Contributions

Let us now describe the main results of the paper. To simplify exposition, we state the results here informally and the rigorous statements may be found in Sections 3 and 2.

Our main contribution is analysis of the gradient AM and gradient EM algorithms. The gradient AM algorithm works in the following way. At iteration t, based on the current parameter estimates $\{\theta_j^{(t)}\}_{j=1}^k$, the gradient AM algorithm constructs estimates of $\{S_j^*\}_{j=1}^k$, namely $\{S_j^{(t)}\}_{j=1}^k$. The next iteration is then obtained by taking a gradient (with γ as step size) over the quadratic loss over all such data points $\{i:(x_i,y_i)\in S_j^{(t)}\}$ for all $j\in[k]$.

On the other hand, in the t-th iteration, the gradient EM algorithm uses the current estimate of $\{\theta_j^*\}_{j=1}^k$, namely $\{\theta_j^{(t)}\}_{j=1}^k$ to compute the *soft-min* probabilities $p_{\theta_1^{(t)},\dots,\theta_k^{(t)}}(x_i,y_i;\theta_j^{(t)})$ for all $j\in[k]$ and $i\in[n]$. Then, using these probabilities, the algorithm takes a gradient of the *soft-min* loss function with step size γ to obtain the next iteration.

We begin by assuming the covariates $x_i \overset{i.i.d}{\sim} \mathcal{N}(0,I_d)$. Note that this assumption serves as a natural starting point of analyzing several EM and AM algorithms ((Balakrishnan et al., 2017; Yi et al., 2014; 2016; Netrapalli et al., 2015; Ghosh & Kannan, 2020)). Furthermore, as stated earlier, we emphasize that in order to obtain convergence, we need to understand the behavior of restricted covariates in the agnostic setting. We require Gaussians, because the behavior of restricted Gaussians are well studied in statistics (Tallis, 1961) and we use several such classical results.

We first consider the min-loss and employ the gradient AM algorithm, similar to (Pal et al., 2022). In particular, we show that the iterates returned by the gradient AM algorithm after T iterations, $\{\theta_j^{(T)}\}_{j=1}^k$ satisfy

$$\|\theta_j^{(T)}-\theta_j^*\|\leq\rho^T\|\theta_j^{(0)}-\theta_j^*\|+\delta,$$

with high probability (where $\rho<1$) provided n is large enough and $\|\theta_j^{(0)}-\theta_j^*\|\leq c_{\text{ini}}\|\theta_j^*\|$. Here c_{ini} is the initialization parameter and δ is the error floor that stems from the agnostic setting and the gradient AM update (see (Balakrishnan et al., 2017) where, even with generative setup, an error floor is shown to be unavoidable). Here δ depends on the step size of the gradient AM algorithm as well as the several geometric properties of the problem like misspecification and separation. However, the result of (Pal et al., 2022) in this regard requires an initialization of $\{\theta_i\}_{i=1}^k$ within a radius of $\mathcal{O}(\frac{1}{\sqrt{d}})$ of the corresponding $\{\theta_i^*\}_{i=1}^k$ which we improve on.

In this paper, we show that it suffices for the initial parameters to be within a (constant) $\Theta(1)$ radius for convergence, provided the geometric parameter $\Delta-\lambda$ is large enough. The $\Theta(1)$ initialization matches the standard (non agnostic,

generative) initialization for mixed linear regression (see (Yi et al., 2014; 2016)). In order to analyze the gradient AM algorithm we need to characterize the behavior of covariates $\{x_i\}_{i=1}^n$ restricted to sets $\{S_j^*\}_{j=1}^k$. In particular we need to control the norm of such restricted Gaussians as well as control the minimum singular value of a random matrix whose rows are made of such random variables. Specifically, we require (i) a lower bound on the minimum singular value of $\frac{1}{n}\sum_{x_i \in S} x_i x_i^T$, where the set S is problem dependent, (ii) an upper bound on $\|x_i\|$ where $x_i \in S$ and (iii) a concentration on $\langle x_i, u\rangle$ where u is some vector and $x_i \in S$.

In order to obtain the above, we leverage the properties of restricted Gaussians ((Tallis, 1961; Ghosh et al., 2019)) on a (generic) set with Gaussian volume bounded away from zero and show that the resulting distribution of the covariates is sub Gaussian with non-zero mean and constant parameter. We obtain upper bounds on the shift and the sub Gaussian parameter. We would like to emphasize that in the realizable setup of mixed linear regressions, as shown in (Yi et al., 2014; 2016) such a characterization may be obtained with lesser complication. However, in the agnostic setup, it turns out to be quite non-trivial.

Moreover, in gradient AM, the setup is complex since the sets are formed by the current iterates of the algorithm (and hence random), unlike $\{S_j^*\}_{j=1}^k$, which are fixed. In order to handle this, we employ re-sampling in each iteration to remove the inter-iteration dependency. We would like to emphasize that sample splitting is a standard technique in the analysis of AM type algorithms and several papers (e.g. (Yi et al., 2014; 2016; Ghosh & Kannan, 2020) for mixed linear regression, (Netrapalli et al., 2015) for phase retrieval and (Ghosh et al., 2020) for distributed optimization) employ such a technique. While this is not desirable, this is a way to remove the inter iteration dependence that comes through data points. Finer techniques like leave-one-out analysis (LOO) is also used ((Chen et al., 2019)) but for simpler problems (like phase retrieval) since the LOO updates are quite non-trivial. This problem exaggerates further in the agnostic setup. Hence, as a first step, in this paper we assume a simpler sample split based framework and keep finer techniques like LOO as future direction.

We would also like to take this opportunity to correct an error in (Pal et al., 2022, Theorem 4.2). In particular, that theorem should hold only for Gaussian covariates, not for general bounded covariates as stated. It was incorrectly assumed in that paper that the lower bound on the singular value mentioned above holds for general covariates.

We then move on to analyze the *soft-min* loss and analyze the gradient EM algorithm. Here, we show similar contraction guarantees in the parameter space as in gradient EM. There are several technical difficulties that arise in the analysis of the gradient EM algorithm for agnostic mixed linear regressions– (i) First, we show that if $(x_i, y_i) \in S_j^*$, then the soft-min probability $p_{\theta_1^*,\ldots,\theta_k^*}(x_i, y_i; \theta_j^*) \geq 1-\eta$, where η is small. (ii) Moreover, using the initialization condition, and the properties of the soft-max function ((Gao & Pavel, 2017)) we argue that $p_{\theta_1^{(t)},\ldots,\theta_k^{(t)}}(x_i, y_i; \theta_j^{(t)})$ is close to $p_{\theta_1^*,\ldots,\theta_k^*}(x_i, y_i; \theta_j^*)$, where $\{\theta_j^{(t)}\}_{t=1}^T$ are the updated of the gradient EM algorithm.

Our results for agnostic gradient AM and EM consist some extra challenge over the existing results in literature ((Balakrishnan et al., 2017; Waldspurger, 2018)). Usually, the population operator with Gaussian covariates are analyzed (mainly in EM, see (Balakrishnan et al., 2017)), and then a finite sample guarantee is obtained using concentration arguments. However, in our setup, with the soft-min probabilities and the $\min$ function, it is not immediately clear how to analyze the population operator. Second, in the gradient EM algorithm, we do not split the samples over iterations, and necessarily handle the inter-iteration dependency of covariates.

Furthermore, to understand the *soft-min* and *min* loss better, in Section 5, we obtain generalization guarantees that involve computing the Rademacher complexity of such function classes. Agreeing with intuition, the complexity of *soft-min* and *min* loss class is at most k times the complexity of the learning problem of simple linear regression with quadratic loss.

1.3. Related works

As discussed earlier, most works on the mixture of linear regressions are in the realizable setting, and aim to do parameter estimation. Algorithms like EM and AM are most popularly used to achieve this task. For instance, in (Balakrishnan et al., 2017), it was proved that a *suitable initialized* EM algorithm is able to find the correct parameters of the mixed linear regressions. Although (Balakrishnan et al., 2017) obtains the convergence results within an ℓ_2 ball, it is then extended to an appropriately defined cone by (Klusowski et al., 2019). On the AM side, (Yi et al., 2014) introduced the AM algorithm for the mixture of 2 regressions, where the initialization is done by the spectral methods. Then, (Yi et al., 2016) extends that to a mixture of k linear regressions. Perhaps surprisingly, for the case of 2 lines, (Kwon & Caramanis, 2018) shows that any random initialization suffices for EM algorithm to converge. In the above mentioned works, the covariates are assumed to be standard Gaussians, which was relaxed in (Li & Liang, 2018), allowing Gaussian covariates to have different covariances. Here, near optimal sample as well as computational complexities were achieved albeit not via EM or AM type algorithm.

In another line of work, the convergence rates of AM or its close variants are investigated. In particular, in (Ghosh & Kannan, 2020; Shen & Sanghavi, 2019), it is shown that AM (or its variants) converge at a double-exponential

(super-linear) rate. Recent work, (Chandrasekher et al., 2021) shows similar results for larger class of problems.

We emphasize that apart from mixture of linear regressions, EM or AM type algorithms are used to address other problems as well. Classically parameter estimation in the mixture of Gaussians is done by EM mixture of Gaussians (see (Balakrishnan et al., 2017; Daskalakis & Kamath, 2014) and the references therein). The seminal paper by (Balakrishnan et al., 2017) addresses the problem of Gaussian mean estimation as well as linear regression with missing covariates. Moreover, AM type algorithms are used in phase retrieval ((Netrapalli et al., 2015; Waldspurger, 2018)), parameter estimation in max-affine regression ((Ghosh et al., 2019)), clustering in distributed optimization ((Ghosh et al., 2020)).

In all of the above mentioned works, the covariates are given to the learner. However, there is another line of research that focuses on analyzing AM type algorithms when the learner has the freedom to design the covariates ((Yin et al., 2019; Krishnamurthy et al., 2019; Mazumdar & Pal, 2020; 2022; Pal et al., 2021)).

However, none of these works is directly comparable to our setting. All these works assume a realizable model where the parameters come with the problem setup. However, ours is an agnostic setup, and here there are no optimal parameters associated with the setup, rather solutions of (naturally emerging) loss functions.

Our work is a direct follow up of (Pal et al., 2022), who introduced the agnostic learning framework for mixed linear regression, and also used the AM algorithm in lieu of empirical risk minimization. Also, (Pal et al., 2022) only considered the min-loss, and neither the soft-min loss nor the EM algorithm, whereas we consider both EM and AM. Moreover, the AM guarantees we obtain are sharper than that of (Pal et al., 2022).

1.4. Organization

We start with the *soft-min* loss function and the gradient EM algorithm in Section 3. In Section 3.2, we obtain the theoretical results of gradient EM. We then move to *min* loss function in Section 2, where we analyze the gradient AM algorithm, with theoretical guarantees given in Section 2.2. We present a rough overview of the proof techniques in Section 4. Finally, in Section 5, we provide some generalization guarantees using Rademacher complexity. We conclude in Section 6 with a few open problems and future direction. We collection all the proofs (both EM and AM) in Appendix B and A.

1.5. Notation

Throughout this paper, we use $\|.\|$ to denote the ℓ_2 norm of a d dimensional vector unless otherwise specified. Also for a positive integer r, we use $[r]$ to denote the set $\{1,...,r\}$. We use $C,C_1,C_2,...,c,c_1,c_2...$ to denote positive universal constants, the value of which may differ from instance to instance.

2. Agnostic Mixed Linear Regression-Min-Loss

In this section, we analyze the min-loss function and analyze gradient AM algorithm. First, recall the definition of $\ell_{\min}(.)$ from Eq. 2. Similar to the section above, we are given a set of n data-points $\{x_i,y_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ drawn from an unknown distribution $\mathcal{D}$. We want to obtain

$$(\theta_1^*,...,\theta_k^*) = \text{argmin}\mathbb{E}_{(x,y)\sim\mathcal{D}}\ell_{\min}(\theta_1,...,\theta_k;x,y).$$

With the given n datapoints, we aim to learn these k hyperplanes via the AM algorithm (Algorithm 1), which tries to minimize the empirical optimization version instead.

2.1. Gradient AM Algorithm

In this section we use the gradient AM algorithm for minimizing $L(\theta_1,\dots,\theta_k)$. The details of our algorithm is given in Algorithm 1.

First note that here, we split the n samples $\{x_i, y_i\}_{i=1}^n$ into $2T$ disjoint samples where we run Algorithm 1 for T iterations. We would like to remind that sample splitting is a standard in AM type algorithms ((Yi et al., 2014; 2016; Ghosh & Kannan, 2020; Netrapalli et al., 2015; Ghosh et al., 2020)). While this is not desirable, this is a way to remove the inter iteration dependence that comes through data points.

Hence, at each iteration of gradient AM we are given $n' = n/2T$ samples. Each iteration consists of 2 stages (see Algorithm 1). In the first stage of the t-th iteration, we use n' samples to construct the index sets $I_j^{(t)}$ in the following way

$$I_j^{(t)} = \{i \in [n'] : (y_i^{(t)} - \langle x_i^{(t)}, \theta_j^{(t)}\rangle)^2 < (y_i^{(t)} - \langle x_i^{(t)}, \theta_{j'}^{(t)}\rangle)^2\}$$

$\forall j' \in [k] \setminus j$. Here, we collect the data points for which the current estimate of θ_j^*, namely $\theta_j^{(t)}$ is a better (linear) estimator than $\{\theta_{j'}^{(t)}\}$ where $j' \neq j$. Notw that $\{I_j^{(t)}\}_{j=1}^k$ partitions $[n']$.

At the second stage of gradient AM, we use another set of fresh n' data points to run the gradient update on the set $\{I_j^{(t)}\}_{j=1}^k$ with step size γ to obtain the next iterate $\{\theta_j^{(t+1)}\}_{j=1}^k$. The details is given in Algorithm 1.

2.2. Theoretical Guarantees

In this section, we obtain theoretical guarantees for Algorithm 1. Similar to the previous section, we assume $|y_i| \leq b$ for all $i \in [n]$. In the following, we consider one iteration of Algorithm 1, and show a contraction in parameter space. Let the current parameter estimates are $\{\theta_j\}_{j=1}^k$ and the corresponding to the index $\{I_j\}_{j=1}^k$. Moreover, let the next

Algorithm 1 Gradient AM for Mixture of Linear Regressions

Input: $\{x_i, y_i\}_{i=1}^n$, Step size γ
Initialization: Initial iterate $\{\theta_j^{(0)}\}_{j=1}^k$
Split all samples into $2T$ disjoint datasets $\{x_i^{(t)}, y_i^{(t)}\}_{i=1}^{n'}$ with $n' = n/2T$ for all $t = 0,1,\ldots,T-1$
for $t = 0,1,\ldots,T-1$ **do**
  Partition:
  For all $j \in [k]$, use n' samples to construct index sets $\{I_j^{(t)}\}_{j=1}^k$ such that $\forall j' \in [k] \setminus j$,
  $$I_j^{(t)} = \{i : (y_i^{(t)} - \langle x_i^{(t)}, \theta_j^{(t)} \rangle)^2 < (y_i^{(t)} - \langle x_i^{(t)}, \theta_{j'}^{(t)} \rangle)^2\}$$
  Gradient Step:
  Use fresh set of n' samples to run gradient update
  $$\theta_j^{(t+1)} = \theta_j^{(t)} - \frac{\gamma}{n} \sum_{i \in [n']} \nabla F_i(\theta_j^{(t)}) \mathbf{1}\{i \in I_j^{(t)}\}, \forall j \in [k]$$
  where $F_i(\theta_j^{(t)}) = (y_i^{(t)} - \langle x_i^{(t)}, \theta_j^{(t)} \rangle)^2$
end for
Output: $\{\theta_j^{(T)}\}_{j=1}^k$

iterates are $\{\theta_j^+\}_{j=1}^k$. Unpacking, the next iterate is given by

$$\theta_j^+ = \theta_j - \frac{2\gamma}{n} \sum_{i \in I_j} [x_i x_i^T \theta_j - y_i x_i] \tag{4}$$

for all $j \in [k]$. We now present our main results of this section.

Theorem 2.1 (Gradient AM). *Suppose $x_i \overset{i.i.d}{\sim} \mathcal{N}(0, I_d)$ and that $n' \geq C \frac{d\log(1/\pi_{\min})}{\pi_{\min}^3}$. Furthermore,*

$$\|\theta_j - \theta_j^*\| \leq c_{\mathsf{ini}} \|\theta_j^*\|$$

for all $j \in [k]$ where c_{ini} is a small positive constant (initialization parameter). Moreover, let the separation parameter satisfy

$$\Delta > \lambda + C_1 [c_{\mathsf{ini}} \sqrt{\log(1/\pi_{\min})} \max_{j \in [k]} \|\theta_j^*\| + \sqrt{1 + \log(1/\pi_{\min})}].$$

Then, running one iteration of Gradient AM with step size γ, yields $\{\theta_j^+\}_{j=1}^k$ satisfying

$$\|\theta_j^+ - \theta_j^*\| \leq \rho \|\theta_j - \theta_j^*\| + \varepsilon, \quad \textit{with probability exceeding}$$

$1 - C_1 \exp(-C_2 \pi_{\min}^4 n') - c_1 \exp(-P_e n') - \frac{n'}{\mathsf{poly}(d)}$, *where $\rho = (1 - c\gamma\pi_{\min}^3)$, and the error floor*

$$\varepsilon \leq C\gamma\lambda\sqrt{d\log d\log(1/\pi_{\min})} + C_1\gamma(k-1)P_e \times \Big[d\log d\log(1/\pi_{\min})\|\theta_1^*\| + Cb\sqrt{d\log d\log(1/\pi_{\min})}\Big],$$

$$\textit{and } P_e \leq 4\exp\Big(-\frac{1}{c_{\mathsf{ini}^2} \max_{j \in [k]} \|\theta_j^*\|^2} [\frac{\Delta - \lambda}{2}]^2\Big).$$

The proof of Theorem 2.1 is deferred to Appendix A. We make a few remarks here.

Remark 2.2 (Contraction factor ρ). We observe that if $\rho < 1$, the above result implies a contraction in parameter space with a slack of ε, which we call the error-floor. Note that by choosing $\gamma < \frac{c_0}{(1-\eta)\pi_{\min}^3}$, where c_0 is a small constant, we can always make $\rho < 1$.

Remark 2.3 (Error floor ε). Observe that the error floor ε depends linearly on the step size γ, similar to any standard stochastic optimization problem. The error floor also decays linearly with the misspecification parameter λ, which may be thought as an agnostic bias. In previous works (Yi et al., 2016; 2014), even in the realizable setting, either the authors assume $\lambda = 0$ or very small. In a related field of online learning (multi armed bandits and reinforcement learning in linear framework), this model misspecification also impacts the regret in a linear fashion as seen by (Jin et al., 2020, Theorem 5). Even in these realizable setting, is it unknown how to tackle large λ.

Remark 2.4 (Re-sampling). Note that the gradient AM algorithm of ours requires re-sampling fresh data points in every iteration. Similar to the analysis of the gradient EM, here also we need to control the lower spectrum of a random matrix consisting Gaussians restricted to a set. From the structure of gradient AM, this set here is given by $S_j^{(t)} = \{(x_i, y_i) : i \in I_j^{(t)}\}$. Note that without re-sampling of data points, analyzing the behavior of Gaussians on the sets $\{S_j^{(t)}\}_{j=1}^k$ turns out to be quite non-trivial since $\{S_j^{(t)}\}_{j=1}^k$ depends on $\{\theta_j^{(t)}\}_{j=1}^k$ which depends on all the data point $\{x_i, y_i\}_{i=1}^n$.

Remark 2.5 (Probability of error P_e). One major part in showing the convergence guarantee is to show that provided good initialization, the probability of a datapoint lying in an incorrect index set is at most P_e. With a closer look, it turns out that if the problem is separated enough (Δ large) and the initialization is suitable (c_{ini} is small), P_e decays exponentially fast. Hence, in such a setup, the second term in ε is quite small.

Remark 2.6 (Sample complexity). Note that we require the number of samples satisfying the following: $n \geq C \frac{d\log(1/\pi_{\min})}{\pi_{\min}^3}$, where the dependence on k comes through $\pi_{\min}$ (and from definition, we have $\pi_{\min} \leq 1/k$). Note that information theoretically, we only require $\Omega(kd)$ samples, since there are kd unknown parameters to learn. Hence, our sample complexity is optimal in d. However, it is sub-optimal in k compared to the standard (non-agnostic) AM guarantees ((Yi et al., 2014; 2016)). The sub-optimality comes from the proof techniques we use for the agnostic setting. In particular, we use spectral properties of a restricted Gaussian vectors on a set with (Gaussian) volume at least $\pi_{\min}$. As shown in (Ghosh et al., 2019), this gives rise to a dependence of $1/\pi_{\min}^3$ in sample complexity. Moreover, in (Ghosh et al., 2019), it is argued (albeit in a different problem), that when spectral properties of such restricted Gaussians are employed, a $1/\pi_{\min}^3$ dependency is in general unavoidable.

Algorithm 2 Gradient EM for Mixture of Linear Regressions

Input: $\{x_i,y_i\}_{i=1}^n$, Step size γ
Initialization: Initial iterate $\{\theta_j^{(0)}\}_{j=1}^k$
for $t=0,1,...,T-1$ **do**
  Compute Probabilities:
  Compute $p_{\theta_1^{(t)},..,\theta_k^{(t)}}(x_i,y_i;\theta_j^{(t)})$ for all $j\in[k]$ and $i\in[n]$
  Gradient Step: (for all $j\in[k]$)

$$\theta_j^{(t+1)}=\theta_j^{(t)}-\frac{\gamma}{n}\sum_{i=1}^n p_{\theta_1^{(t)},..,\theta_k^{(t)}}(x_i,y_i;\theta_j^{(t)})\nabla F_i(\theta_j^{(t)}),$$

  where $F_i(\theta_j^{(t)})=(y_i-\langle x_i,\theta_j^{(t)}\rangle)^2$
end for
Output: $\{\theta_j^{(T)}\}_{j=1}^k$

3. EM algorithm for Soft-Min Loss

In this section we analyze the *soft-min* loss function and propose gradient EM algorithm to address this. Recall the definition of $\ell_{\text{softmin}}(.)$ from Eq. 3. Moreover, recall that we are given a set of n data-points $\{x_i,y_i\}_{i=1}^n$, where $x_i\in\mathbb{R}^d$ and $y_i\in\mathbb{R}$ drawn from an unknown distribution $\mathcal{D}$. Our goal here is to obtain

$$(\theta_1^*,...,\theta_k^*)=\text{argmin}\mathbb{E}_{(x,y)\sim\mathcal{D}}\ell_{\text{softmin}}(\theta_1,...,\theta_k;x,y).$$

We aim to learn these k hyperplanes through the given data. The EM algorithm (Algorithm 2) tries to minimize the empirical version of the problem.

3.1. Gradient EM Algorithm

We propose EM based algorithm for minimizing the empirical loss function $L(\theta_1,..,\theta_k)$. In particular we propose a variant of EM, popularly known as gradient EM for this. The steps are given in Algorithm 2. Each iteration of gradient EM consists of two steps. First, in the compute probability step, based on the current estimates of $\{\theta_j^*\}_{j=1}^k$, namely $\{\theta^{(t)}\}_{j=1}^k$, Algorithm 2 computes the soft-min probabilities computed using the current iterates $\{\theta^{(t)}\}_{j=1}^k$, which is $p_{\theta_1^{(t)},...,\theta_k^{(t)}}(x_i,y_i;\theta_j^{(t)})$ for all $j\in[k]$ and $i\in[n]$. In the subsequent step, using these probabilities, the algorithm takes a gradient step with step size γ. In particular, for the j-th iterate $\theta_j^{(t)}$, gradient EM weights the standard quadratic loss computed on the i-th data point, given by $(y_i-\langle x_i,\theta_j^{(t)}\rangle)^2$ and takes the gradient to obtain the next iterate $\{\theta_j^{(t+1)}\}_{j=1}^k$. We truncate Algorithm 2 after T steps.

We split the n samples $\{x_i,y_i\}_{i=1}^n$ into $2T$ disjoint samples where we run Algorithm 2 for T iterations. Again sample splitting is a standard in EM type algorithms ((Balakrishnan et al., 2017; Kwon & Caramanis, 2018)). Hence, at each iteration of gradient EM we are given $n'=n/2T$ samples. Each iteration consists of 2 stages (see Algorithm 2). The first n' samples are used to compute the probabilities, and the next set of samples are used to take the gradient step.

3.2. Theoretical Guarantees

We now look at the convergence guarantees of Algorithm 2. In particular, here we consider one iterate of the gradient EM algorithm with current estimate $(\theta_1,...,\theta_k)$. Also, assume that the next iterate with these current estimates is given by $(\theta_1^+,...,\theta_k^+)$. Unrolling the iterate, we have

$$\theta_j^+=\theta_j-\frac{2\gamma}{n'}\sum_{i=1}^{n'}p_{\theta_1,...,\theta_k}(x_i,y_i;\theta_j)\big(x_ix_i^T\theta_j-y_ix_i\big). \quad (5)$$

for all $j\in[k]$. Furthermore, we assume $|y_i|\le b$ for all $i\in[n']$ for a non-negative b. With this, we are now ready to present the main result of this section.

Theorem 3.1 (Gradient EM). *Suppose that $x_i\overset{i.i.d}{\sim}\mathcal{N}(0,I_d)$ and that $n'\ge C\frac{d\log(1/\pi_{\min})}{\pi_{\min}^3}$. Moreover,*

$$\|\theta_j-\theta_j^*\|\le c_{\text{ini}}\|\theta_j^*\|$$

for all $j\in[k]$, where c_{ini} is a small positive constant (initialization parameter) satisfying $c_{\text{ini}}<c_2\frac{\lambda}{\sqrt{\log(1/\pi_{\min})}\|\theta_1^*\|}$. Then running one iteration of gradient EM algorithm with step size γ yields $\{\theta_j^+\}_{j=1}^k$ satisfying

$$\|\theta_j^+-\theta_j^*\|\le\rho\|\theta_j-\theta_j^*\|+\varepsilon,$$

with probability at least $1-C_1\exp(-c_1\pi_{\min}^4n')-C_2\exp(-c_2d)-n'/\text{poly}(d)-n'C_3\exp(-\frac{\lambda^2}{c_{\text{ini}}^2\|\theta_1^*\|^2})$, where

$$\varepsilon\le C\gamma\lambda\sqrt{d\log d\log(1/\pi_{\min})}$$
$$+C_1\gamma\eta'(b+\sqrt{d\log d\log(1/\pi_{\min})})^2(c_{\text{ini}}+1))\|\theta_1^*\|,$$

$\rho=(1-2\gamma c(1-\eta)\pi_{\min}^3)$, $\eta'=e^{-((\Delta-C\lambda)^2-C_2\lambda^2)}$ and $\eta=\left(\frac{1-e^{-C_2\lambda^2}+(k-1)e^{-(\Delta-C\lambda)^2}}{1+(k-1)e^{-(\Delta-C\lambda)^2}}\right)$, with $C,C_1,..,c,c_1,..$ as universal positive constants.

We defer the proof of the theorem in Appendix B. The remarks we made after the AM algorithm continues to hold here as well.

Remark 3.2 (Error floor ε). Observe that the error floor ε depends linearly on the step size γ. The error floor also decays linearly with the misspecification parameter λ and an exponentially decaying term dependent on the gap.

Discussion and Comparison between gradient EM and AM: Note that both the algorithms require initialization and provides exponential convergence with error

floor. However, gradient AM minimizes an intuitive min-loss while gradient EM minimizes *optimal* (maximum likelihood in the generative setup) soft-min loss. Moreover, the gradient AM algorithm requires the separation $\Delta = \Omega(\lambda + \sqrt{\log k}(1 + c_{\text{ini}}))$ (exact condition in Theorem 2.1), whereas we do not have any such requirement for gradient EM. On the flip side, the convergence of gradient EM requires a condition on the initialization parameter c_{ini} that depends on misspecification λ, whereas for gradient AM algorithm, no such restriction is imposed.

4. Proof Sketches

In this section, we present a rough sketch of the proof of Theorems 2.1 and 3.1.

4.1. Gradient AM (Theorem 2.1)

For gradient AM algorithm, based on the current iterates $\{\theta_j\}_{j=1}^k$, we first construct the index sets $\{I_j\}_{j=1}^k$ using n' fresh samples, where I_j consists of all such indices such that θ_j is a better predictor compared to the other parameters. Similarly, one can construct $\{I_j^*\}_{j=1}^k$ based on $\{\theta_j^*\}_{j=1}^k$. Unrolling gradient AM update (Eq. 4), using another set of n' samples we have

$$\|\theta_1^+ - \theta_1^*\| = \|\theta_1 - \theta_1^* - \frac{2\gamma}{n'}\sum_{i\in I_1}\left(x_i x_i^T\theta_1 - y_i x_i\right)\|.$$

Similar to the gradient EM setup, it turns out that we need to lower bound $\sigma_{\min}(\frac{1}{n'}\sum_{i\in I_j} x_i x_i^T)$. Note that since we use n' fresh samples to construct I_j, the set can be considered fixed with respect to the samples used in the gradient step and we can leverage Lemma B.2. We use $\sigma_{\min}(\frac{1}{n'}\sum_{i\in I_1} x_i x_i^T) \geq \sigma_{\min}(\frac{1}{n'}\sum_{i\in I_1\cap I_1^*} x_i x_i^T)$. Thanks to the suitable initialization and Lemma A.1, we show that $|I_1 \cap I_1^*|$ is big enough, yielding a singular value lower bound of $\approx \pi_{\min}^3$. The control of other terms are done similar to the gradient EM setup, and upon combining, we get the final theorem.

4.2. Gradient EM (Theorem 3.1)

Recall that we consider one iteration of Algorithm 2 with current and next iterates as $\{\theta_j\}_{j=1}^k$ and $\{\theta_j^+\}_{j=1}^k$ respectively. Recall the update given by Eq. 5. Without loss of generality, we focus on $j=1$ and use shorthand $p(\theta_1)$ to denote $p_{\theta_1,\ldots,\theta_k}(x_i,y_i;\theta_1)$. With this we have

$$\|\theta_1^+ - \theta_1^*\| = \|\theta_1 - \theta_1^* - \frac{2\gamma}{n'}\sum_{i=1}^{n'} p(\theta_1)\left(x_i x_i^T\theta_1 - y_i x_i\right)\|.$$

We now break the sum to indices $i : (x_i, y_i) \in S_1^*$ and otherwise. When we look at indices such that $(x_i,y_i) \in S_1^*$, after a few algebraic manipulation, it turns out we need to lower bound $\sigma_{\min}[\frac{1}{n'}\sum_{i:(x_i,y_i)\in S_1^*} x_i x_i^T]$. Since $\Pr(x_i : (x_i, y_i) \in S_1^*) \geq \pi_{\min}$ by definition, leveraging properties of restricted Gaussians (Lemma B.2), we obtain $\sigma_{\min}[\frac{1}{n'}\sum_{i:(x_i,y_i)\in S_1^*}(1-\eta)x_i x_i^T] \geq (1-\eta)\pi_{\min}^3$. Furthermore, leveraging the fact that if $(x_i,y_i) \in S_1^*$, we have $p(\theta_1^*) \geq 1-\eta$ (Lemma B.1), and using the norm upper bound on restricted Gaussians (Lemma B.3) we control such indices. Finally, combining all the terms and using the geometric parameters succinctly, we obtain the desired result.

5. Generalization Guarantees

In this section, we obtain generalization guarantees for the *soft-min* loss functions. Note that similar generalization guarantee for the *min* loss function has appeared in (Pal et al., 2022).

We learn a mixture of functions from $\mathcal{X} \to \mathcal{Y}$ for $\mathcal{X} \subseteq \mathbb{R}^d$ fitting data distribution $\mathcal{D}$ over $(\mathcal{X},\mathcal{Y})$. A learner has access to samples $\{x_i,y_i\}_{i=1}^n$. There is a base class $\mathcal{H} : \mathcal{X} \to \mathcal{Y}$. Here, we work with the setup of list decoding where the learner outputs a list while testing. In (Pal et al., 2022) the list decodable function class has been defined. We rewrite here for completeness.

Definition 5.1. Let $\mathcal{H}$ be the base function class $\mathcal{H}$. We construct a vector valued k-list-decodable function class, namely $\bar{\mathcal{H}}_k$ such that any $\bar{h} \in \bar{\mathcal{H}}_k$ is defined as

$$\bar{h} = (h_1(\cdot),\cdots,h_k(\cdot))$$

such that $h_j \in \mathcal{H}_j$ for all $j \in [k]$. Thus $\bar{h}$'s map $\mathcal{X} \to \mathcal{Y}^k$ and form the new function class $\bar{\mathcal{H}}_k$.

To ease notation, we omit the k in $\bar{\mathcal{H}}$ when clear from context.

In our setting, the base function class is linear, i.e., for all $j \in [k]$

$$\mathcal{H}_j = \mathcal{H} = \{\langle\theta,\cdot\rangle : \forall\theta \in \mathbb{R}^d \text{ s.t } \|\theta\|_2 \leq R\},$$

and the base loss function $\ell : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}^+$ is given by

$$\ell(h_j(x),y)) = (y - \langle x,\theta_j\rangle)^2.$$

In what follows, we obtain generalization guarantees for bounded covariates and response, i.e., $|y| \leq 1$ and $\|x\| \leq 1$.

Claim 5.2. For bounded regression problem, the loss function $\ell(h_j(x),y))$ is Lipschitz with parameter $2(1+R)$ with respect to the first argument.

The proof is deferred to Appendix C. We are interested in the soft loss function, which is a function of the k-base loss

functions:

$$\begin{aligned}\mathcal{L}(\bar{h}(x),y) &= \mathcal{L}(x,y;\theta_1,...,\theta_k)\\ &= \sum_{j=1}^{k} p_{\theta_1,..,\theta_k}(x,y;\theta_j)[y-\langle x,\theta_j\rangle]^2\\ &= \sum_{j=1}^{k} p_{\theta_1,..,\theta_k}(x,y;\theta_j)\ell(h_j(x),y),\end{aligned}$$

where

$$p_{\theta_1,..,\theta_k}(x,y;\theta_j) = \frac{e^{-(y-\langle x,\theta_j\rangle)^2}}{\sum_{\ell=1}^{k} e^{-(y-\langle x,\theta_\ell\rangle)^2}}.$$

We have n datapoints $\{x_i,y_i\}_{i=1}^n$ drawn from $\mathcal{D}$ and we want to understand how well this *soft-min* loss generalizes. In order to do that, a standard metric one studies in statistical learning theory is (emprirical) Rademacher Complexity ((Mohri et al., 2018)). In our setup, the loss class is defined by

$$\{(x,y)\mapsto \sum_{j=1}^{k} p_{\theta_1,..,\theta_k}(x,y;\theta_j)\ell(h_j(x),y);\{\theta_j:\|\theta_j\|\le R\}_{j=1}^k\}.$$

Let us define this class as Φ. The Rademacher complexity of the loss class is given by

$$\begin{aligned}\hat{\mathfrak{R}}_n(\Phi) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}_k}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\mathcal{L}(\bar{h}(x_i),y_i)\right|\right]\\ &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\{\theta_j:\|\theta_j\|\le R\}_{j=1}^k}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\sum_{j=1}^{k}p_{\theta_1,..,\theta_k}(x,y;\theta_j)\ell(h_j(x),y)\right|\right],\end{aligned}$$

where $\boldsymbol{\sigma}$ is a set of Rademacher RV's $\{\sigma_i\}_{i=1}^n$. We have the following result:

Lemma 5.3. *The Rademacher complexity of* Φ *satisfies*

$$\hat{\mathfrak{R}}(\Phi) \le 4k(1+R)\hat{\mathfrak{R}}(\mathcal{H}) \le \frac{4kR(1+R)}{\sqrt{n}}.$$

We observe that the (empirical) Rademacher complexity of the *soft-min* loss class does not blow-up provided the complexity of the base class $\mathcal{H}$ is controlled. Moreover, since the base class is a linear hypothesis class (with bounded ℓ_2 norm), the Rademacher complexity scales as $\mathcal{O}(1/\sqrt{n})$, resulting in the above bound. The proof is deferred in Appendix C. In a nutshell, we consider a bigger class of all possible convex combination of the base losses, and connect Φ to that bigger function class.

6. Conclusion and Open Problems

In this work, we have studied the agnostic setup for mixed linear regression, and show that EM and AM algorithms are strong enough to provide provable guarantees even in this setup. However we believe such algorithms may be used in a broader context of agnostic learning. We conclude the paper with a few interesting problems. Beyond mixture of linear regressions, can this agnostic setup be used for other problems such as mixture of classifiers, mixture of experts, to name a few? What is the role of Gaussian covariates in such an agnostic setting? Can we relax this to some extent? In (Ghosh et al., 2019) it is explained how restricted Gaussian analysis can be extended to sub-Gaussians satisfying a *small ball* condition for the particular problem of max-affine regression. Another interesting direction is to analyze the AM based algorithms without resampling in the agnostic setup, leveraging techniques like Leave One Out (LOO) as an example. We keep these as our future endevors.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements. This research is supported in part by NSF awards 2133484, 2217058, 2112665.

References

Balakrishnan, S., Wainwright, M. J., and Yu, B. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1): 77–120, 2017.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Chaganty, A. T. and Liang, P. Spectral experts for estimating mixtures of linear regressions. In *International Conference on Machine Learning*, pp. 1040–1048. PMLR, 2013.

Chandrasekher, K. A., Pananjady, A., and Thrampoulidis, C. Sharp global convergence guarantees for iterative nonconvex optimization: A gaussian process perspective. *arXiv preprint arXiv:2109.09859*, 2021.

Chen, Y., Chi, Y., Fan, J., and Ma, C. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176:5–37, 2019.

Daskalakis, C. and Kamath, G. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Conference on Learning Theory*, 2014.

Faria, S. and Soromenho, G. Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, 80(2):201–225, 2010.

Gao, B. and Pavel, L. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

Ghosh, A. and Kannan, R. Alternating minimization converges super-linearly for mixed linear regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 1093–1103. PMLR, 2020.

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

Ghosh, A., Chung, J., Yin, D., and Ramchandran, K. An efficient framework for clustered federated learning. *arXiv preprint arXiv:2006.04088*, 2020.

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/jin20a.html.

Klusowski, J. M., Yang, D., and Brinda, W. Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *IEEE Transactions on Information Theory*, 65(6):3515–3524, 2019.

Krishnamurthy, A., Mazumdar, A., McGregor, A., and Pal, S. Sample complexity of learning mixture of sparse linear regressions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Kwon, J. and Caramanis, C. Global convergence of em algorithm for mixtures of two component linear regression. *arXiv preprint arXiv:1810.05752*, 2018.

Li, Y. and Liang, Y. Learning mixtures of linear regressions with nearly optimal complexity. In *Conference On Learning Theory*, pp. 1125–1144. PMLR, 2018.

Mazumdar, A. and Pal, S. Recovery of sparse signals from a mixture of linear samples. In *International Conference on Machine Learning (ICML)*, 2020.

Mazumdar, A. and Pal, S. On learning mixture models with sparse parameters. *arXiv preprint arXiv:2202.11940*, 2022.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Netrapalli, P., Jain, P., and Sanghavi, S. Phase retrieval using alternating minimization. *IEEE Transactions on Signal Processing*, 63(18):4814–4826, 2015.

Pal, S., Mazumdar, A., and Gandikota, V. Support recovery of sparse signals from a mixture of linear measurements. *Advances in Neural Information Processing Systems*, 34, 2021.

Pal, S., Mazumdar, A., Sen, R., and Ghosh, A. On learning mixture of linear regressions in the non-realizable setting. In *International Conference on Machine Learning*, pp. 17202–17220. PMLR, 2022.

Shen, Y. and Sanghavi, S. Iterative least trimmed squares for mixed linear regression. *arXiv preprint arXiv:1902.03653*, 2019.

Städler, N., Bühlmann, P., and Van De Geer, S. l1-penalization for mixture regression models. *Test*, 19(2): 209–256, 2010.

Tallis, G. M. The moment generating function of the truncated multi-normal distribution. *Journal of the Royal Statistical Society. Series B (Methodological)*, 23(1):223–229, 1961. ISSN 00359246. URL http://www.jstor.org/stable/2983860.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Viele, K. and Tong, B. Modeling with mixtures of linear regressions. *Statistics and Computing*, 12(4):315–330, 2002.

Waldspurger, I. Phase retrieval with random gaussian sensing vectors by alternating projections. *IEEE Transactions on Information Theory*, 64(5):3301–3312, 2018.

Wang, D., Ding, J., Hu, L., Xie, Z., Pan, M., and Xu, J. Differentially private (gradient) expectation maximization algorithm with statistical guarantees. *arXiv preprint arXiv:2010.13520*, 2020.

Yi, X., Caramanis, C., and Sanghavi, S. Alternating minimization for mixed linear regression. In *International Conference on Machine Learning*, pp. 613–621. PMLR, 2014.

Yi, X., Caramanis, C., and Sanghavi, S. Solving a mixture of many random linear equations by tensor decomposition and alternating minimization. *arXiv preprint arXiv:1608.05749*, 2016.

Yin, D., Pedarsani, R., Chen, Y., and Ramchandran, K. Learning mixtures of sparse linear regressions using sparse graph codes. *IEEE Transactions on Information Theory*, 65(3):1430–1451, 2019.

Zhu, R., Wang, L., Zhai, C., and Gu, Q. High-dimensional variance-reduced stochastic gradient expectation-maximization algorithm. In *International Conference on Machine Learning*, pp. 4180–4188. PMLR, 2017.

A. Proof of Theorem 2.1

Without loss of generality, let us focus on θ_1^+. We have

$$\begin{aligned}\|\theta_1^+-\theta_1^*\| &= \|\theta_1-\theta_1^*-\frac{\gamma}{n'}\sum_{i\in I_1}\nabla F_i(\theta_1)\| \\ &= \|(\theta_1-\theta_1^*)-\frac{\gamma}{n'}\sum_{i\in I_1}(\nabla F_i(\theta_1)-\nabla F_i(\theta_1^*))-\frac{\gamma}{n'}\sum_{i\in I_1}\nabla F_i(\theta_1^*)\| \\ &\le \underbrace{\|(\theta_1-\theta_1^*)-\frac{\gamma}{n'}\sum_{i\in I_1}(\nabla F_i(\theta_1)-\nabla F_i(\theta_1^*))\|}_{T_1}+\underbrace{\frac{\gamma}{n'}\|\sum_{i\in I_1}\nabla F_i(\theta_1^*)\|}_{T_2}.\end{aligned}$$

Let us first consider T_1. Substituting the gradients, we obtain

$$T_1=\|(I-\frac{2\gamma}{n}\sum_{i\in I_1}x_ix_i^\top)(\theta_1-\theta_1^*)\|=\|(I-\frac{2\gamma}{n'}\sum_{i:(x_i,y_i)\in S_1}x_ix_i^\top)(\theta_1-\theta_1^*)\|.$$

We require a lower bound on

$$\sigma_{\min}(\frac{1}{n}\sum_{i\in I_1}x_ix_i^\top)\ge\sigma_{\min}(\frac{1}{n'}\sum_{i:(x_i,y_i)\in S_1\cap S_1^*}x_ix_i^\top)$$

Similar to the EM framework, in order to bound the above, we need to look at the behavior of the covariates (which are standard Gaussian) over the restricted set given by $S_1\cap S_1^*$. Note that since we are resampling at each step, and using fresh set of samples to construct S_j and another fresh set of samples to run the Gradient AM algorithm, we can directly use Lemma B.2 here. Moreover, we use the fact that $|i:(x_i,y_i)\in S_1\cap S_1^*|\ge C|i:(x_i,y_i)\in S_1^*|\ge C'\pi_{\min}n$ with probability at least $1-C\exp(-\pi_{\min}n)$ where we use the initialization Lemma A.1. Thus, we have

$$\sigma_{\min}(\frac{1}{n'}\sum_{i:(x_i,y_i)\in S_1}x_ix_i^\top)\ge c\pi_{\min}^3$$

with probability at least $1-C_1\exp(-C_2\pi_{\min}^4n')-C_3\exp(-\pi_{\min}n')$ provided $n'\ge C\frac{d\log(1/\pi_{\min})}{\pi_{\min}^3}$. As a result,

$$T_1\le(1-c\gamma\pi_{\min}^3)\|\theta_1-\theta_1^*\|,$$

with probability at least $1-C_1\exp(-C_2\pi_{\min}^4n')$.

Let us now consider the term T_2. We have

$$\begin{aligned}T_2 &= \frac{\gamma}{n}\|\sum_{i:(x_i,y_i)\in S_1}\nabla F_i(\theta_1^*)\| \\ &\le \frac{\gamma}{n}\sum_{i:(x_i,y_i)\in S_1}\|\nabla F_i(\theta_1^*)\| \\ &= \frac{\gamma}{n}\sum_{i:(x_i,y_i)\in S_1\cap S_1^*}\|\nabla F_i(\theta_1^*)\|+\frac{\gamma}{n}\sum_{j=2}^{k}\sum_{i:(x_i,y_i)\in S_1\cap S_j^*}\|\nabla F_i(\theta_1^*)\|\end{aligned}$$

When $\{i:(x_i,y_i)\in S_1^*\}$, we have

$$\begin{aligned}\|\nabla F_i(\theta_1^*)\| &= 2|y_i-\langle x_i,\theta_1^*\rangle|\|x_i\| \\ &\le 2\lambda\|x_i\|\le C\lambda\sqrt{d\log d\log(1/\pi_{\min})}\end{aligned}$$

with probability at least $1-n'/\mathsf{poly}(d)$, where in the first inequality, we have used the misspecification assumption, and in the second inequality, we use Lemma B.3. Let us now compute an upper bound on $\|\nabla F_i(\theta_1^*)\|$, which we use to bound the second part. We have

$$\begin{aligned}\|\nabla F_i(\theta_1^*)\| &\leq \|x_i\|^2\|\theta_1^*\|+\|x_i\||y_i| \\ &\leq C_1 d\log d\log(1/\pi_{\min})\|\theta_1^*\|+Cb\sqrt{d\log d\log(1/\pi_{\min})}\end{aligned}$$

with probability at least $1-1/\mathsf{poly}(d)$.

With this, we have

$$\begin{aligned}T_2 &\leq \frac{\gamma}{n}|I_1\cap I_1^*|C\lambda\sqrt{d\log d\log(1/\pi_{\min})}+\frac{\gamma}{n}\sum_{j=2}^{k}|I_1\cap I_j^*|\Bigg(C_1 d\log d\log(1/\pi_{\min})\|\theta_1^*\| \\ &\qquad +Cb\sqrt{d\log d\log(1/\pi_{\min})}\Bigg) \\ &\leq \gamma C\lambda\sqrt{d\log d\log(1/\pi_{\min})}+C_1\gamma(k-1)P_e\Big[d\log d\log(1/\pi_{\min})\|\theta_1^*\|+Cb\sqrt{d\log d\log(1/\pi_{\min})}\Big],\end{aligned}$$

with probability at least $1-\exp(-cP_e n)-\frac{n'}{\mathsf{poly}(d)}-\frac{P_e n}{\mathsf{poly}(d)}$, where P_e is defined in Lemma A.1. In this case, we use $|I_1\cap I_1^*|\leq n'$ (trivially holds) as well as the standard binomial concentration on $|I_1\cap I_j^*|$ with mean at most $n'P_e$ with probability at least $1-\exp(-cP_e n')$. Moreover we take the union bound. Here, we use Lemma B.3 along with the fact that $|y_i|\leq b$.

Combining T_1 and T_2, we have

$$\begin{aligned}\|\theta_1^+-\theta_1^*\| &\leq (1-c\gamma\pi_{\min}^3)\|\theta_1-\theta_1^*\|+C\gamma\lambda\sqrt{d\log d\log(1/\pi_{\min})} \\ &\qquad +C_1\gamma(k-1)P_e\Big[d\log d\log(1/\pi_{\min})\|\theta_1^*\|+Cb\sqrt{d\log d\log(1/\pi_{\min})}\Big],\end{aligned}$$

with probability at least $1-C_1\exp(-C_2\pi_{\min}^4 n')-\exp(-cP_e n')-\frac{n'}{\mathsf{poly}(d)}$.

A.1. Good Initialization

We stick to analyzing θ_1^+. In the following lemma, we only consider θ_2. In general, the same argument holds for $\{\theta_3,\ldots,\theta_k\}$.

Lemma A.1. *We have*

$$\begin{aligned}P_e &= \mathbb{P}\Big(F_i(\theta_1)>F_i(\theta_2)|i\in I_1^*\Big) \\ &\leq 4\exp\Bigg(-\frac{1}{c_{\mathsf{ini}^2}\max_{j\in[k]}\|\theta_j^*\|^2}\left[\frac{\Delta-\lambda}{2}\right]^2\Bigg)\end{aligned}$$

Let us consider the event

$$F_i(\theta_1)>F_i(\theta_2),$$

which is equivalent to

$$|y_i-\langle x_i,\theta_1\rangle|>|y_i-\langle x_i,\theta_2\rangle|.$$

Let us look at the left hand side of the above inequality. We have

$$\begin{aligned}&|y_i-\langle x_i,\theta_1^*\rangle+\langle x_i,\theta_1-\theta_1^*\rangle| \\ &\leq |y_i-\langle x_i,\theta_1^*\rangle|+|\langle x_i,\theta_1-\theta_1^*\rangle| \\ &\leq \lambda+|\langle x_i,\theta_1-\theta_1^*\rangle|,\end{aligned}$$

where we have used the fact that if $i\in I_1^*$, the first term is at most λ.

Similarly, for the right hand side, we have

$$\begin{aligned}&|y_i-\langle x_i,\theta_2^*\rangle-\langle x_i,\theta_2-\theta_2^*\rangle|\\&\geq|y_i-\langle x_i,\theta_2^*\rangle|-|\langle x_i,\theta_2-\theta_2^*\rangle|\\&\geq\Delta-|\langle x_i,\theta_2-\theta_2^*\rangle|\end{aligned}$$

where we use the fact that if $i\in I_1^*$, the first term is lower bounded by Δ.

Combining these, we have

$$\begin{aligned}\mathbb{P}\Big(F_i(\theta_1)>F_i(\theta_2)|i\in I_1^*\Big)&\leq\mathbb{P}\Big(|\langle x_i,\theta_1-\theta_1^*\rangle|+|\langle x_i,\theta_2-\theta_2^*\rangle|\geq\Delta-\lambda\Big)\\&\leq\mathbb{P}\Big(|\langle x_i,\theta_1-\theta_1^*\rangle|\geq\frac{\Delta-\lambda}{2}\Big)+\mathbb{P}\Big(|\langle x_i,\theta_2-\theta_2^*\rangle|\geq\frac{\Delta-\lambda}{2}\Big)\end{aligned}$$

Let us look at the first term. Lemma B.2 shows that if $i\in I_1^*$ (accordingly $(x_i,y_i)\in S_1^*$), the distribution of $x_i-\mu_\tau$ is subGaussian with (squared) parameter at most $C(1+\log(1/\pi_{\min}))$, where μ_τ is the mean of x_i (under the restriction $(x_i,y_i)\in S_1^*$). With this we have

$$\begin{aligned}\mathbb{P}\Big(|\langle x_i,\theta_1-\theta_1^*\rangle|\geq\frac{\Delta-\lambda}{2}\Big)&\leq\mathbb{P}\Big(|\langle x_i-\mu_\tau,\theta_1-\theta_1^*\rangle|+\|\mu_\tau\|\|\theta_1-\theta_1^*\|\geq\frac{\Delta-\lambda}{2}\Big)\\&\leq\mathbb{P}\Big(|\langle x_i-\mu_\tau,\theta_1-\theta_1^*\rangle|\geq\frac{\Delta-\lambda}{2}-c_{\mathsf{ini}}C\sqrt{\log(1/\pi_{\min})}\|\theta_1^*\|\Big)\end{aligned}$$

where we use the initialization condition $\|\theta_1-\theta_1^*\|\leq c_{\mathsf{ini}}\|\theta_1^*\|$, and from Lemma B.2, we have $\|\mu_\tau\|^2\leq C\log(1/\pi_{\min})$. Now, provided $\Delta-\lambda>C(c_{\mathsf{ini}}\sqrt{\log(1/\pi_{\min})}\|\theta_1^*\|)+C_1\sqrt{1+\log(1/\pi_{\min})}$, using sub-Gaussian concentration, we obtain

$$\mathbb{P}\Big(|\langle x_i,\theta_1-\theta_1^*\rangle|\geq\frac{\Delta-\lambda}{2}\Big)\leq2\exp\Big(-\frac{1}{c_{\mathsf{ini}^2}\|\theta_1^*\|^2}\Big[\frac{\Delta-\lambda}{2}\Big]^2\Big).$$

Similarly, for the second term, similar calculation yields

$$\mathbb{P}\Big(|\langle x_i,\theta_2-\theta_2^*\rangle|\geq\frac{\Delta-\lambda}{2}\Big)\leq2\exp\Big(-\frac{1}{c_{\mathsf{ini}^2}\|\theta_2^*\|^2}\Big[\frac{\Delta-\lambda}{2}\Big]^2\Big),$$

and hence

$$\mathbb{P}\Big(F_i(\theta_1)>F_i(\theta_2)|i\in I_1^*\Big)\leq4\exp\Big(-\frac{1}{c_{\mathsf{ini}^2}\max_{j\in[k]}\|\theta_j^*\|^2}\Big[\frac{\Delta-\lambda}{2}\Big]^2\Big)$$

which proves the lemma.

B. Proof of Theorem 3.1

Let us look at the iterate of gradient EM after one step and without loss of generality, we focus on recovering θ_1^*. We have

$$\|\theta_1^+-\theta_1^*\|=\|\theta_1-\theta_1^*-\frac{2\gamma}{n'}\sum_{i=1}^{n'}p_{\theta_1,\ldots,\theta_k}(x_i,y_i;\theta_1)\big(x_ix_i^T\theta_1-y_ix_i\big)\|$$

Let us use the shorthand $p(\theta_1)$ to denote $p_{\theta_1,\ldots,\theta_k}(x_i,y_i;\theta_1)$ and $p(\theta_1^*)$ to denote $p_{\theta_1^*,\ldots,\theta_k^*}(x_i,y_i;\theta_1^*)$ respectively. We have

$$\begin{aligned}\|\theta_1^+-\theta_1^*\|&=\|\theta_1-\theta_1^*-\frac{2\gamma}{n'}\sum_{i:(x_i,y_i)\in S_1^*}p(\theta_1)\big(x_ix_i^T\theta_1-y_ix_i\big)-\frac{2\gamma}{n'}\sum_{i:(x_i,y_i)\notin S_1^*}p(\theta_1)\big(x_ix_i^T\theta_1-y_ix_i\big)\|\\&\leq\underbrace{\|\theta_1-\theta_1^*-\frac{2\gamma}{n'}\sum_{i:(x_i,y_i)\in S_1^*}p(\theta_1)\big(x_ix_i^T\theta_1-y_ix_i\big)-\frac{2\gamma}{n'}\sum_{i:(x_i,y_i)\notin S_1^*}p(\theta_1)\big(x_ix_i^T\theta_1-y_ix_i\big)\|}_{T_1}\end{aligned}$$

First we argue from the separability and the closeness condition that, if $(x_i,y_i)\in S_1^*$, the probability $p(\theta_1)$ is bounded away from 0. Lemma B.1 shows that conditioned on $(x_i,y_i)\in S_j^*$, we have $p_{\theta_1,\ldots,\theta_k}(x_i,y_i;\theta_j)\geq 1-\eta$, where

$$\eta=\left(\frac{1-e^{-C_2\lambda^2}+(k-1)e^{-(\Delta-C\lambda)^2}}{1+(k-1)e^{-(\Delta-C\lambda)^2}}\right).$$

with probability at least $1-C_3\exp\left(-C_1\frac{\lambda^2}{c_{\text{ini}2}\|\theta_1^*\|^2}\right)$. With this, let us look at T_1. We have

$$T_1\leq \underbrace{\|\theta_1-\theta_1^*-\frac{2\gamma}{n'}\sum_{i:(x_i,y_i)\in S_1^*}p(\theta_1)\big(x_ix_i^T\theta_1-y_ix_i\big)\|}_{T_{11}}+\underbrace{\frac{2\gamma}{n'}\|\sum_{i:(x_i,y_i)\notin S_1^*}p(\theta_1)\big(x_ix_i^T-y_ix_i\big)\|}_{T_{12}}.$$

We continue to upper bound T_{11}:

$$\begin{aligned}
T_{11}&\leq \|\theta_1-\theta_1^*-\frac{2\gamma}{n'}\sum_{i:(x_i,y_i)\in S_1^*}p(\theta_1)\big(x_ix_i^T\theta_1-y_ix_i\big)\|\\
&\leq \|\theta_1-\theta_1^*-\frac{2\gamma}{n'}\sum_{i:(x_i,y_i)\in S_1^*}p(\theta_1)\big(x_ix_i^T\theta_1-x_ix_i^T\theta_1^*\big)\|+\frac{2\gamma}{n'}\|\sum_{i:(x_i,y_i)\in S_1^*}p(\theta_1)\big(x_ix_i^T\theta_1^*-y_ix_i\big)\|\\
&\leq \|\left[I-\frac{2\gamma}{n'}\sum_{i:(x_i,y_i)\in S_1^*}p(\theta_1)x_ix_i^T\right](\theta_1-\theta_1^*)\|+\frac{2\gamma}{n'}\sum_{i:(x_i,y_i)\in S_1^*}p(\theta_1)|y_i-\langle x_i,\theta_1^*\rangle|\|x_i\|\\
&\leq \|\left[I-\frac{2\gamma}{n'}\sum_{i:(x_i,y_i)\in S_1^*}p(\theta_1)x_ix_i^T\right](\theta_1-\theta_1^*)\|+C\lambda\gamma\sqrt{d\log d\log(1/\pi_{\min})},
\end{aligned}$$

with probability at least $1-C_3n'\exp\left(-C_1\frac{\lambda^2}{c_{\text{ini}2}\|\theta_1^*\|^2}\right)-n'/\text{poly}(d)$, where we use the misspecification condition, $|y_i-\langle x_i,\theta_1^*\rangle|\leq\lambda$ for all $(x_i,y_i)\in S_1^*$, along with the fact that the number of such indices is trivially upper bounded by the total number of observations, n. Moreover, we also use Lemma B.3 to bound $\|x_i\|$.

Note that since $(x_i,y_i)\in S_1^*$, we have $p(\theta_1)\geq 1-\eta$. We need to look at $\sigma_{\min}\left(\frac{1}{n'}\sum_{i:(x_i,y_i)\in S_1^*}p(\theta_1)x_ix_i^T\right)$, where $p(\theta_1)\geq 1-\eta$. We use the fact that

$$\sigma_{\min}\left(\frac{1}{n'}\sum_{i:(x_i,y_i)\in S_1^*}p(\theta_1)x_ix_i^T\right)\geq\sigma_{\min}\left(\frac{1}{n'}\sum_{i:(x_i,y_i)\in S_1^*}(1-\eta)x_ix_i^T\right).$$

Note that we need to analyze the behavior of the data restricted on the set S_1^*. In particular we are interested in the second moment estimation of such restricted Gaussian random variable. We show that, conditioned on S_1^*, the distribution of x_i changes to a sub-Gaussian with a shifted mean. Lemma B.2 characterizes the behavior as well as the second moment estimation for such variables.

We invoke the Lemma B.2 and use the standard binomial concentration to obtain $|i:(x_i,y_i)\in S_1^*|\geq C\pi_{\min}n$ with probability at least $1-\exp(-c\pi_{\min}n)$. With this, we obtain

$$\sigma_{\min}\left(\frac{1}{n'}\sum_{i:(x_i,y_i)\in S_1^*}(1-\eta)x_ix_i^T\right)\geq c(1-\eta)\pi_{\min}^3$$

with probability at least $1-C_1\exp(-C_2\pi_{\min}^4n')$, provided $n'\geq C\frac{d\log(1/\pi_{\min})}{\pi_{\min}^3}$.

Using this, we obtain

$$T_{11}\leq(1-2\gamma c(1-\eta)\pi_{\min}^3)\|\theta_1-\theta_1^*\|+C\gamma\lambda\sqrt{d\log d\log(1/\pi_{\min})}.$$

with high probability. Let us now look at T_{12}. We have

$$\begin{aligned}
T_{12} &= \frac{2\gamma}{n'}\Big\|\sum_{i:(x_i,y_i)\notin S_1^*} p(\theta_1)\big(x_i x_i^T\theta_1 - y_i x_i\big)\Big\| \\
&\le \frac{2\gamma}{n'}\sum_{i:(x_i,y_i)\notin S_1^*} p(\theta_1)\|x_i x_i^T\theta_1 - y_i x_i\| \\
&\overset{(i)}{\le} \frac{2\gamma\eta'}{n'}\sum_{i:(x_i,y_i)\notin S_1^*} |y_i - x_i^T\theta_1|\|x_i\| \\
&\le \frac{2\gamma\eta'}{n'}\sum_{i:(x_i,y_i)\notin S_1^*} (|y_i| + \|x_i\|\|\theta_1\|)\|x_i\| \\
&\overset{(ii)}{\le} \frac{2\gamma\eta'}{n'}\sum_{i:(x_i,y_i)\notin S_1^*} (b + C\sqrt{d\log d\log(1/\pi_{\min})})[\|\theta_1-\theta_1^*\| + \|\theta_1^*\|])\sqrt{d\log d\log(1/\pi_{\min})} \\
&\le 2\gamma\eta'(b + C\sqrt{d\log d\log(1/\pi_{\min})})^2(c_{\mathsf{ini}}+1))\|\theta_1^*\|.
\end{aligned}$$

with probability at least $1 - n'/\mathsf{poly}(d) - C_3 n'\exp\Big(-C_1\frac{\lambda^2}{c_{\mathsf{ini}^2}\|\theta_1^*\|^2}\Big)$ (using union bound). Here (i) follows from the fact that $p(\theta_1^*) \le \eta'$ where $\eta' = e^{-((\Delta - C\lambda)^2 - C_2\lambda^2)}$. (since $(x_i,y_i)\notin S_1^*$, which follows from Lemma B.1), (ii) follows from the fact that $|y_i| \le b$ for all i. Moreover, since $\{S_j^*\}_{j=1}^d$ partitions $\mathbb{R}^d$, $(x_i,y_i)\notin S_1^*$ implies that $(x_i,y_i)\in S_\ell^*$ where $\ell\in[k]\setminus\{1\}$, and we can invoke Lemma B.3.

Collecting all the terms: We now collect the terms and combine them to obtain

$$\begin{aligned}
\|\theta_1^+ - \theta_1^*\| &\le T_{11} + T_{12} \\
&\le (1 - 2\gamma c(1-\eta)\pi_{\min}^3)\|\theta_1 - \theta_1^*\| + C\gamma\lambda\sqrt{d\log d\log(1/\pi_{\min})} \\
&\quad + 2\gamma\eta'(b + C\sqrt{d\log d\log(1/\pi_{\min})})^2(c_{\mathsf{ini}}+1))\|\theta_1^*\|.
\end{aligned}$$

with probability at least $1 - C_1\exp(-c_1\pi_{\min}^4 n') - C_2\exp(-c_2 d) - n'/\mathsf{poly}(d) - n'C_3\exp\Big(-\frac{\lambda^2}{c_{\mathsf{ini}^2}\|\theta_1^*\|^2}\Big)$.

Let $\rho = (1 - 2\gamma c(1-\eta)\pi_{\min}^3)$ and we choose γ such that $\rho < 1$. We obtain

$$\|\theta_1^+ - \theta_1^*\| \le \rho\|\theta_1 - \theta_1^*\| + \varepsilon,$$

where

$$\varepsilon \le C\gamma\lambda\sqrt{d\log d\log(1/\pi_{\min})} + 2\gamma\eta'(b + C\sqrt{d\log d\log(1/\pi_{\min})})^2(c_{\mathsf{ini}}+1))\|\theta_1^*\|,$$

with probability at least $1 - C_1\exp(-c_1\pi_{\min}^4 n') - C_2\exp(-c_2 d) - n'/\mathsf{poly}(d) - n'C_3\exp\Big(-\frac{\lambda^2}{c_{\mathsf{ini}^2}\|\theta_1^*\|^2}\Big)$.

B.1. Proofs of Auxiliary Lemmas:

Lemma B.1. *For any $(x_i,y_i)\in S_j^*$, we have $p_{\theta_1,\ldots,\theta_k}(x_i,y_i;\theta_j) \ge 1-\eta$, where*

$$\eta = \left(\frac{1 - e^{-C_2\lambda^2} + (k-1)e^{-(\Delta - C\lambda)^2}}{1 + (k-1)e^{-(\Delta - C\lambda)^2}}\right).$$

Moreover, for $(x_i,y_i)\notin S_j^*$ we have

$$p_{\theta_1,\ldots,\theta_k}(x_i,y_i;\theta_j) \le e^{-((\Delta - C\lambda)^2 - C_2\lambda^2)}.$$

Proof. Consider any $(x_i,y_i)\in S_j^*$ and use the definition of $p_{\theta_1,\ldots,\theta_k}(x_i,y_i;\theta_j)$. We obtain

$$p_{\theta_1,\ldots,\theta_k}(x_i,y_i;\theta_j)=\frac{e^{-(y_i-\langle x_i,\theta_j\rangle)^2}}{\sum_{\ell=1}^k e^{-(y_i-\langle x_i,\theta_\ell\rangle)^2}}$$

Note that

$$\begin{aligned}|y_i-\langle x_i,\theta_j\rangle| &= |y_i-\langle x_i,\theta_j^*\rangle+\langle x_i,\theta_j^*-\theta_j\rangle| \\ &\le |y_i-\langle x_i,\theta_j^*\rangle|+|\langle x_i,\theta_j^*-\theta_j\rangle|\end{aligned}$$

Furthermore, using reverse triangle inequality, we also have

$$|y_i-\langle x_i,\theta_j\rangle|\ge |y_i-\langle x_i,\theta_j^*\rangle|-|\langle x_i,\theta_j^*-\theta_j\rangle|.$$

Since we are re-sampling at every step, and from the initialization condition, we handle the random variable $\langle x_i,\theta_j^*-\theta_j\rangle$.

Using Lemma B.2 shows that if $(x_i,y_i)\in S_1^*$, the distribution of $x_i-\mu_\tau$ is subGaussian with (squared) parameter at most $C(1+\log(1/\pi_{\min}))$, where μ_τ is the mean of x_i (under the restriction $(x_i,y_i)\in S_1^*$). With this we have

$$\begin{aligned}\mathbb{P}\Big(|\langle x_i,\theta_1-\theta_1^*\rangle|\ge C\lambda\Big) &\le \mathbb{P}\Big(|\langle x_i-\mu_\tau,\theta_1-\theta_1^*\rangle|+\|\mu_\tau\|\|\theta_1-\theta_1^*\|\ge C\lambda\Big) \\ &\le \mathbb{P}\Big(|\langle x_i-\mu_\tau,\theta_1-\theta_1^*\rangle|\ge C\lambda-c_{\mathsf{ini}}C_1\sqrt{\log(1/\pi_{\min})}\|\theta_1^*\|\Big)\end{aligned}$$

where we use the initialization condition $\|\theta_1-\theta_1^*\|\le c_{\mathsf{ini}}\|\theta_1^*\|$, and from Lemma B.2, we have $\|\mu_\tau\|^2\le C\log(1/\pi_{\min})$.

Now, provided $c_{\mathsf{ini}}<C_2\frac{\lambda}{\sqrt{\log(1/\pi_{\min})}\|\theta_1^*\|}$, using sub-Gaussian concentration, we obtain

$$\Big(|\langle x_i,\theta_1-\theta_1^*\rangle|\ge C\lambda\Big)\le 2\exp\Big(-C_1\frac{1}{c_{\mathsf{ini}^2}\|\theta_1^*\|^2}\lambda^2\Big).$$

Using the assumption, i,.e., the separability and the misspecification condition, we obtain

$$\begin{aligned}p_{\theta_1,\ldots,\theta_k}(x_i,y_i;\theta_j) &\ge \frac{e^{-C_2\lambda^2}}{e^{-(y_i-\langle x_i,\theta_j\rangle)^2}+\sum_{\ell\neq j}e^{-(y_i-\langle x_i,\theta_\ell\rangle)^2}} \\ &\ge \frac{e^{-C_2\lambda^2}}{e^{-(y_i-\langle x_i,\theta_j\rangle)^2}+(k-1)e^{-(\Delta-C\lambda)^2}} \\ &\ge \frac{e^{-C_2\lambda^2}}{1+(k-1)e^{-(\Delta-C\lambda)^2}} \\ &= 1-\left(\frac{1-e^{-C_2\lambda^2}+(k-1)e^{-(\Delta-C\lambda)^2}}{1+(k-1)e^{-(\Delta-C\lambda)^2}}\right).\end{aligned}$$

Let us look at the condition $(x_i,y_i)\notin S_j^*$. Since $\{S_j^*\}_{j=1}^k$ partitions $\mathbb{R}^d$, $(x_i,y_i)\in S_{j'}^*$ for $j'\in[k]$. With this,

$$\begin{aligned}p_{\theta_1,\ldots,\theta_k}(x_i,y_i;\theta_j) &\le \frac{e^{-(\Delta-C\lambda)^2}}{e^{-(y_i-\langle x_i,\theta_{j'}\rangle)^2}+\sum_{\ell\neq j'}e^{-(y_i-\langle x_i,\theta_\ell\rangle)^2}} \\ &\le \frac{e^{-(\Delta-C\lambda)^2}}{e^{-C_2\lambda^2}+0}=e^{-((\Delta-C\lambda)^2-C_2\lambda^2)}.\end{aligned}$$

The above events occur with probability at least $1-C_3\exp\Big(-C_1\frac{\lambda^2}{c_{\mathsf{ini}^2}\|\theta_1^*\|^2}\Big)$.

□

Lemma B.2. *Suppose $x \sim \mathcal{N}(0, I_d)$ and a fixed set S such that $\mathbb{P}(x \in S) \geq \nu$. Let τ denote the restriction of x onto S. Moreover, suppose we have n draws from a standard Gaussian and m of them falls in S. Provided $n \geq \frac{C\log(1/\nu)}{\nu^3} d$, we have*

$$\sigma_{\min}\left(\frac{1}{m}\sum_{i=1}^{m}\tau_i\tau_i^T\right) \geq \frac{C}{2}\nu^2,$$

with probability at least $1-2\exp(-c_1\nu^4 n)$.

Proof. Consider a random vector τ drawn from such restricted Gaussian distribution, and let μ_τ and Σ_τ be the first and second moment respectively. Using (Ghosh et al., 2019, Equation 38 (a-c)), we have

$$\|\mu_\tau\|^2 \leq C\log(1/\nu),$$

$$C\nu^2 I_d \preccurlyeq \Sigma_\tau,$$

Moreover (Yi et al., 2016, Lemma 15 (a)) shows that τ is subGaussian with ψ_2 norm at most $\zeta^2 \leq C(1+\log(1/\pi_{\min})$. Coupled with the definition of ψ_2 norm, (Vershynin, 2018), we obtain that the centered random variable $\tau - \mu_\tau$ admits a ψ_2 norm squared of at most $C_1(1+\log(1/\pi_{\min})$.

With m draws of such random variables, from (Ghosh et al., 2019, Equation 39), we have

$$\sigma_{\min}\left(\frac{1}{m}\sum_{i=1}^{m}\tau_i\tau_i^T\right) \geq C\nu^2 - \zeta^2\left(\frac{d}{m}+\sqrt{\frac{d}{m}}+\delta\right),$$

with probability at least $1-2\exp(-c_1 m\min\{\delta,\delta^2\})$

If there are n samples from the unrestricted Gaussian distribution, the number of samples, m that fall in S is given by $m \geq \frac{1}{2}\nu n$ with high proibability. This can be seen directly from the binomial tail bounds. We have

$$\mathbb{P}(m \leq \frac{\nu n}{2}) \leq \exp(-c\nu n)$$

Combining the above, with $\nu \geq c$ where c is a constant as well as $n \geq \frac{C\log(1/\nu)}{\nu^3} d$, we have

$$\sigma_{\min}\left(\frac{1}{m}\sum_{i=1}^{m}\tau_i\tau_i^T\right) \geq \frac{C}{2}\nu^2,$$

with probability at least $1-2\exp(-c_1 m\min\{\delta,\delta^2\})$. Substituting $\delta = C\nu^2$ yields the result. □

Lemma B.3. *Suppose $(x_i, y_i) \in S_j^*$ for some $j \in [k]$. We have*

$$\|x_i\| \leq C(\sqrt{d\log d\log(1/\pi_{\min})}+\sqrt{\log(1/\pi_{\min})}) \leq C_1\sqrt{d\log d\log(1/\pi_{\min})},$$

with probability at least $1-1/\mathsf{poly}(d)$, where the degree of the polynomial depends on the constant C.

Proof. Note that Lemma B.2 shows that under $(x_i, y_i) \in S_j^*$ for some $j \in [k]$, the centered random variable $\tau_i - \mu_\tau$ is sub-Gaussian with ψ_2 norm squared of at most $C(1+\log(1/\pi_{\min}))$. Note that since, $\tau_i - \mu_\tau$ is centered, the ψ_2 norm is (orderwise) same as the sub-Gaussian parameter.

We now use the standard norm concentration for sub-Gaussian random variables (Jin et al., 2019). We have, for a sub-Gaussian random vector with parameter at most $C(1+\log(1/\pi_{\min}))$, we have

$$\mathbb{P}\Big(\|X-\mathbb{E}X\| \geq t\sqrt{d}\sqrt{(1+\log(1/\pi_{\min})}\Big) \leq 2\exp(-c_1 t^2).$$

Using this with $t = C\sqrt{\log d}$ along with the fact that $\|\mu_\tau\|^2 \leq C\log(1/\pi_{\min})$, we obtain the lemma. □

C. Proof of Generalization

C.1. Proof of Claim 5.2

In order to see this, suppose $h_j^{(1)} \in \mathcal{H}_j$ and $h_j^{(2)} \in \mathcal{H}_j$, and so we have $h_j^{(1)}(x) = \left\langle x, \theta_j^{(1)} \right\rangle$ and $h_j^{(2)}(x) = \left\langle x, \theta_j^{(2)} \right\rangle$ with $\|\theta_j^{(1)}\| \leq R$ as well as $\|\theta_j^{(2)}\| \leq R$. With this, we have

$$\begin{aligned}
|\ell(h_j^{(1)}(x),y) - \ell(h_j^{(2)}(x),y)| &= \left| \left\langle x_i, \theta_j^{(2)} - \theta_j^{(1)} \right\rangle [2y - \left\langle x, \theta_j^{(2)} + \theta_j^{(1)} \right\rangle] \right| \\
&\leq |h_j^{(1)}(x) - h_j^{(2)}(x)| \left[2|y| + \|x\| (\|\theta_j^{(1)}\| + \|\theta_j^{(2)}\|) \right] \\
&\leq 2(1+R)|h_j^{(1)}(x) - h_j^{(2)}(x)|,
\end{aligned}$$

which proves the claim.

C.2. Proof of Lemma 5.3

Proof. Note that the soft-min loss is a convex combination of the base losses, and the probabilities are computed by $p_{\theta_1,..,\theta_k}(x,y;\theta_j)$. Instead, if we consider the loss class with *all possible* convex combinations of the base losses, the corresponding loss class will be a superset of the current loss class. From the definition of Rademacher complexity, if $F_1 \subseteq F_2$ for any two sets F_1 and F_2, we have $\hat{\mathfrak{R}}_n(F_1) \leq \hat{\mathfrak{R}}_n(F_2)$. We define the following loss class

$$\bar{\Phi} = \left\{ (x,y) \mapsto \sum_{j=1}^{k} \alpha_j \ell(h_j(x),y); \theta_j \in \mathbb{R}^d, \|\theta_j\| \leq R, \alpha_j \geq 0 \forall j \in [k], \sum_{j=1}^{k} \alpha_j = 1 \right\},$$

and hence from the definition of Rademacher complexity, we have $\hat{\mathfrak{R}}(\Phi) \leq \hat{\mathfrak{R}}(\bar{\Phi})$. Continuing we have

$$\begin{aligned}
\hat{\mathfrak{R}}(\bar{\Phi}) &= \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\{\theta_j : \|\theta_j\| \leq R, \alpha_j \geq 0\}_{j=1}^k, \sum_{j=1}^k \alpha_j = 1} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \sum_{j=1}^{k} \alpha_j \ell(h_j(x),y) \right| \right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\{\theta_j : \|\theta_j\| \leq R, \alpha_j \geq 0\}_{j=1}^k, \sum_{j=1}^k \alpha_j = 1} \left| \sum_{j=1}^{k} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \alpha_j \ell(h_j(x),y) \right| \right] \\
&\leq \sum_{j=1}^{k} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\theta_j : \|\theta_j\| \leq R, \alpha_j \geq 0, |\alpha_j| \leq 1} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \alpha_j \ell(h_j(x),y) \right| \right] \\
&\leq \sum_{j=1}^{k} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\theta_j : \|\theta_j\| \leq R, \alpha_j \geq 0, |\alpha_j| \leq 1} |\alpha_j| \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(h_j(x),y) \right| \right] \\
&\leq \sum_{j=1}^{k} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\theta_j : \|\theta_j\| \leq R, \alpha_j \geq 0, |\alpha_j| \leq 1} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(h_j(x),y) \right| \right] \\
&\leq \sum_{j=1}^{k} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\theta_j : \|\theta_j\| \leq R} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(h_j(x),y) \right| \right] \\
&= k \hat{\mathfrak{R}}(\ell \circ \mathcal{H}) \\
&\leq 4k(1+R) \hat{\mathfrak{R}}(\mathcal{H}) \\
&\leq \frac{4kR(1+R)}{\sqrt{n}}
\end{aligned}$$

where in the third line, we have used the sub-additivity property of the supremum function as well as the triangle inequality. We also used the above claim regarding the Lipschitz constant of the loss function $\ell(.,.)$ and invoked the contraction result for Rademacher averages by (Bartlett & Mendelson, 2002). Finally, for linear hypothesis class, we use (Mohri et al., 2018)

to obtain the final result. Hence, we obtain

$$\hat{\mathfrak{R}}(\Phi) \leq \frac{4kR(1+R)}{\sqrt{n}},$$

which proves the result. □