A Wasserstein Minimax Framework for Mixed Linear Regression

Theo Diamandis [*1] **Yonina C. Eldar** [2] **Alireza Fallah** [1] **Farzan Farnia** [1] **Asuman Ozdaglar** [1]

Abstract

Multi-modal distributions are commonly used to model clustered data in statistical learning tasks. In this paper, we consider the Mixed Linear Regression (MLR) problem. We propose an optimal transport-based framework for MLR problems, Wasserstein Mixed Linear Regression (WMLR), which minimizes the Wasserstein distance between the learned and target mixture regression models. Through a model-based duality analysis, WMLR reduces the underlying MLR task to a nonconvex-concave minimax optimization problem, which can be provably solved to find a minimax stationary point by the Gradient Descent Ascent (GDA) algorithm. In the special case of mixtures of two linear regression models, we show that WMLR enjoys global convergence and generalization guarantees. We prove that WMLR's sample complexity grows linearly with the dimension of data. Finally, we discuss the application of WMLR to the federated learning task where the training samples are collected by multiple agents in a network. Unlike the Expectation Maximization algorithm, WMLR directly extends to the distributed, federated learning setting. We support our theoretical results through several numerical experiments, which highlight our framework's ability to handle the federated learning setting with mixture models.

1. Introduction

Learning mixture models which describe data collected from multiple subpopulations has been a basic task in the machine learning literature. Multi-modal distributions typically emerge in distributed learning settings where the training data are gathered from a heterogeneous group of users. For example, speech data or genetic data may exhibit a clustered distribution based on language and ethnicity, respectively. Such settings require learning methods that can efficiently learn an underlying multi-modal distribution in both a centralized and a distributed setting.

In this paper, we specifically focus on Mixed Linear Regression (MLR) problems. In the MLR problem, the output variable for every user is a randomized linear function of the feature variables, generated according to one of k unknown linear regression models. This structured model provides a simple but expressive framework to analyze multimodal labeled data. The clustered structure of MLR appears in several supervised learning applications. For example, users of a recommendation engine usually have unknown yet clustered sets of preferences which leads to multiple regression models. In genetic datasets, the underlying cell-type of collected samples is a latent variable that can result in different linear regression models. Under such scenarios, the cluster identity is an unknown latent variable that should be estimated along with the linear regression models.

To address the MLR problem, we propose an optimal transport-based learning framework, which we refer to as *Wasserstein Mixed Linear Regression (WMLR)*. We revisit optimal transport theory to formulate the centralized MLR task as a minimax optimization problem solved by the WMLR algorithm. The formulated minimax problem is the dual problem of minimizing the Wasserstein distance between the target and learned mixture regression models. Because the original minimax problem formulated by applying the standard Kantorovich duality (Villani, 2008) incurs significant computational and statistical costs, we reduce the minimax learning task to a tractable problem by a model-based simplification of the dual maximization variables.

For a general MLR problem, we prove that the proposed minimax problem can be reduced to a nonconvex-concave optimization problem for which the gradient descent ascent (GDA) algorithm is guaranteed to converge to a stationary minimax solution. Furthermore, under the well-studied benchmark of a mixture of two symmetric linear regression models, we theoretically support our framework by providing global convergence and generalization guarantees. In particular, we show that our framework can provably converge to the global minimax solution and properly gen-

[*]The authors are in alphabetical order. [1]Department of Electrical Engineering & Computer Science, MIT, USA [2]Faculty of Math and Computer Science, Weizmann Institute of Science, Israel. Correspondence to: Theo Diamandis <tdiamand@mit.edu>, Alireza Fallah <afallah@mit.edu>, Farzan Farnia <farnia@mit.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

eralize from the empirical distribution of training samples to the underlying mixture regression model.

Next, we examine the WMLR algorithm for MLR tasks in the distributed federated learning setting (McMahan et al., 2017). In a federated learning task, a set of local users connected to a central server train a global model over the samples observed in the network. While the Expectation-Maximization (EM) algorithm is widely considered as the state-of-the-art approach for centralized MLR problems, in the federated learning setting, the maximization step of every iteration of the EM algorithm requires multiple gradient computation and communication steps to obtain an exact solution via an iterative method. As a result, the EM algorithm cannot be decomposed into an efficient distributed form.

On the other hand, we show that while the maximization step in the EM algorithm does not directly reduce to a distributed form, the gradient steps of WMLR extend to the federated learning setting. As a result, our theoretical guarantees in the centralized case also hold in the federated learning setting. Finally, we present the results of several numerical experiments which support the flexibility of our proposed minimax framework in both centralized and decentralized learning tasks.

Our main contributions are summarized as follows:

1. We propose a minimax framework, Wasserstein Mixed Linear Regression (WMLR), to solve the MLR problem using optimal transport theory.

2. We reduce WMLR to a tractable nonconvex-concave minimax optimization problem, which can be solved by the GDA algorithm.

3. We show that WMLR enjoys convergence and generalization guarantees in both centralized and federated learning settings in the symmetric MLR case.

4. We support WMLR's theoretical guarantees with numerical experiments for the centralized and federated learning settings.

1.1. Related Work

The MLR model, introduced in the statistics literature by De Veaux (1989) and later in the machine learning literature by Jordan & Jacobs (1994) as "hierarchical mixtures of experts", provides a simple but expressive framework to analyze multimodal data. However, despite the simplicity of the model, learning mixed regression models is computationally difficult; the maximum likelihood problem is intractable in the general case (Yi et al., 2014).

EM-based Algorithms for MLR Kwon et al. (2019) prove global convergence for balanced mixtures of symmetric two component linear regressions. Several other papers have extended (Kwon et al., 2019)'s results to unequally weighted components and K components in the noiseless setting (See (Kwon & Caramanis, 2020) and references therein). Furthermore, Kwon & Caramanis (2020) prove local convergence for k-MLR in the noisy case. However, the EM algorithm still requires "good" initialization for convergence to the optimal solution (Balakrishnan et al., 2017). For finding such a good initialization, several methods have been proposed in the EM literature, including methods based on PCA (Yi et al., 2014) and method of moments (Chaganty & Liang, 2013). Without proper initialization, the EM algorithm has been empirically shown to find poor estimations due to EM's "sharp" selection of clusters.

Gradient-based Algorithms for MLR The traditional EM algorithm fully solves a maximization at each step, resulting in the "sharp" behavior. Several alternative algorithms have been proposed that take a gradient descent approach. First-order EM, where only one gradient step in the maximization problem is taken, enjoys a local convergence guarantee (Balakrishnan et al., 2017). Zhong et al. (2016) show local convergence for a nonconvex objective function that solves the k-MLR problem. Chen et al. (2014) provide a convex formulation for the two component case, but it is unclear how this method generalizes to $k > 2$.

Federated Learning with Heterogeneous Data Several approaches have been proposed in the literature to deal with heterogeneity in FL, including correcting the local updates (Karimireddy et al., 2020) or using meta-learning techniques for achieving personalization (Fallah et al., 2020). In particular, clustering is one of these approaches where the idea is to group client population into clusters (Sattler et al., 2020; Ghosh et al., 2020; Mansour et al., 2020; Li et al., 2021). Most relevant to our work, Mansour et al. (2020) and Ghosh et al. (2020) propose alternating minimization algorithms, where at each step the agents find their cluster identity, compute the loss function gradient, and send them back to the server. Ghosh et al. (2020) further prove convergence guarantees for linear models and strongly convex loss functions under certain initialization assumptions. These frameworks include a much larger class of problems than MLR, but they do not enjoy the same global convergence and optimality guarantees that WMLR has for the MLR case.

Minimax Frameworks for Federated Learning Several related works explore the applications of minimax frameworks for improving the fairness and robustness of federated learning algorithms. Mohri et al. (2019) introduce Agnostic Federated Learning as a min-max framework that improves the fairness properties in federated learning tasks. Rei-

sizadeh et al. (2020) propose a minimax federated learning framework that is robust to affine distribution shifts. Similarly, Deng et al. (2021) develop a distributionally-robust federated learning algorithm using a minimax formulation. However, unlike our work the mentioned frameworks do not address the clustered federated learning problem.

Generative Adversarial Networks (GANs) Similar to our proposed framework, GANs (Goodfellow et al., 2014) reduce the distribution learning problem to a minimax optimization task. Optimal transport costs have been similarly used to formulate GAN problems (Arjovsky et al., 2017; Sanjabi et al., 2018; Farnia & Tse, 2018; Feizi et al., 2020). Also, Genevay et al. (2018) formulate a min-max problem for learning generative models using the optimal transport-based Sinkhorn loss functions. On the other hand, since standard GAN formulations perform suboptimally in learning multimodal distributions (Goodfellow, 2016), Farnia et al. (2020) propose a similar model-based minimax approach to successfully learn mixtures of Gaussians. Mena et al. (2020) introduce the optimal transport-based Sinkhorn EM framework for learning mixture models. However, while the mentioned minimax frameworks focus on unsupervised learning tasks, our proposed approach addresses the supervised MLR problem.

1.2. Notation

For two random variables Y and Y', $Y \overset{d}{=} Y'$ means that Y and Y' have the same distribution. For a finite set $\mathcal{A}$, Unif($\{\mathcal{A}\}$) stands for the uniform distribution over $\mathcal{A}$, and $I_A(u)$ is the indicator function of A, *i.e.,* $I_A(u) = 1$ if $u \in A$ and 0 otherwise. Given two distributions P and Q, defined over sets $\mathcal{Z}_P$ and $\mathcal{Z}_Q$, respectively, $\Pi(P, Q)$ denotes the set of joint distributions over $\mathcal{Z}_P \times \mathcal{Z}_Q$ such that its marginal over $\mathcal{Z}_P$ and $\mathcal{Z}_Q$ is equal to P and Q, respectively. The 2-Wasserstein cost between distributions P_Y and Q_Y on Y is defined as:

$$W_2(P_Y, Q_Y)^2 := \inf_{(Y,Y')\sim M\in\Pi(P_Y,Q_Y)} \mathbb{E}_M\big[\|Y - Y'\|_2^2\big], \tag{1}$$

where Y, Y' are constrained to be marginally distributed as P, Q, respectively. To extend this definition to the supervised learning setting, for joint distributions $P_{X,Y}$ and $Q_{X,Y}$ sharing the same marginal P_X we define:

$$W_2(P_{X,Y}, Q_{X,Y}) := \mathbb{E}_{P_X}\big[W_2(P_{Y|X=x}, Q_{Y|X=x})\big]. \tag{2}$$

2. Problem Formulation

We consider the mixed linear regression problem, where the output to each input vector is generated by one of k linear regression models. Specifically, we observe data points $\mathcal{S} := \{(x_i, y_i)\}_{i=1}^n$ where, for every i, $x_i \in \mathcal{X} \subset \mathbb{R}^d$, $y_i \in \mathbb{R}$, and

$$y_i = \sum_{j=1}^{k} \mathbb{1}\{z_i = j\}(\beta_j^*)^\top x_i + \epsilon_i, \quad i = 1, ..., n, \tag{3}$$

with latent variable $z_i \in \{1, 2, ..., k\}$. Each $\beta_j^* \in \mathbb{R}^d$ denotes the regression vector for one of the overall k components. We assume that the input data $\{x_i\}_{i=1}^n$ are norm-bounded random vectors with x_i drawn i.i.d. from p_x with $\sup_{x\in\mathcal{X}} \|x\| \le C$, that the noises $\{\epsilon_i\}_{i=1}^n$ are independent of the input data and drawn i.i.d. from the normal distribution $p_\epsilon := \mathcal{N}(0, \sigma^2)$ where σ^2 is known, and that each z_i is drawn from $\{1, 2, ..., k\}$ uniformly at random.

The MLR problem is to find the distribution $p^\star$ that best fits the data $\mathcal{S}$ (according to some metric). We know that $p^\star$ lies in the class of distributions $\mathcal{P}$, parameterized by $\beta_{[k]} := (\beta_j)_{j=1}^k$:

$$\mathcal{P} := \Big\{ p_{\beta_{[k]}}(X, Y) : X \sim p_x,\ Z \sim \text{Unif}(\{1, ..., k\}), \\ \mathbb{P}(Y \mid X = x, Z = j) \overset{d}{=} \mathcal{N}(\beta_j^\top x, \sigma^2) \Big\}. \tag{4}$$

The Expectation Maximization algorithm (EM) is commonly used to tackle this problem. The EM algorithm provides a widely-used heuristic for computing the maximum likelihood estimator (MLE) for the regressors $(\beta_j^*)_{j=1}^k$. However, implementing the EM algorithm in the federated learning setting can be challenging. We consider finding the $\beta_{[k]}$ which minimizes the distance between $p_{\beta_{[k]}}$ and $p^*_{\beta^*_{[k]}}$ with respect to a distribution distance measure, i.e.,

$$\operatorname*{argmin}_{\beta_{[k]}} D\left(p_{\beta^*_{[k]}}, p_{\beta_{[k]}}\right),$$

where $D(\cdot, \cdot)$ is a distribution distance metric to be chosen. In this work, we use the expected 2-Wasserstein cost as our metric, resulting in the problem

$$\operatorname*{argmin}_{\beta_{[k]}} W_c(p_{\beta^*_{[k]}}, p_{\beta_{[k]}}) \tag{5}$$

It is worth noting that, here, and similar to well-known EM analysis (Kwon et al., 2019), we assume σ is known to simplify the derivations. However, we could extend our framework to the case that σ is not known by parametrizing $\mathcal{P}$ by both $\beta_{[k]}$ and σ and minimzing over both of them in (5). Furthermore, as we will see in Section 4, one advantage of our proposed method works without the knowledge of σ for the symmetric case with $k = 2$.

In the next section, we use the properties of 2-Wasserstein distance to build a minimax framework for mixed linear regression and then show how it can be used in the federated learning setting.

3. A Wasserstein Minimax Approach to MLR

To formulate a minimax learning problem, we replace the Wasserstein cost in (5) with its dual representation according to the Kantorovich duality (Villani, 2008). This reformulation results in the following minimax optimization problem:

$$\underset{\beta_{[k]}}{\operatorname{argmin}} \max_{\psi} \mathbb{E}_{p_{\beta^*_{[k]}}}[\psi(x,y)] - \mathbb{E}_{p_{\beta_{[k]}}}[\psi^c(x,y)], \quad (6)$$

where the optimization variable $\psi : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ is an unconstrained function, and the c-transform $\psi^c(x,y)$ is defined as

$$\psi^c(x,y) = \sup_{y'} \psi(x,y') - \frac{1}{2}\|y-y'\|_2^2. \quad (7)$$

Note that the two distributions $p_{\beta_{[k]}}$ and $p_{\beta^*_{[k]}}$ have the same marginal p_x and only differ in the conditional distribution $p_{y|x}$. As a result, the optimal transport task requires to only move mass to match the conditional distribution. This observation results in the cost function used to define the c-transform operation in (7).

However, the above optimization problem for an unconstrained ψ is known to be statistically and computationally complex (Arora et al., 2017). In this section, our goal is to show that one can solve (6) over the following space of functions for ψ parameterized by $2k$ vectors $\gamma_{[2k]} \in \mathcal{F}$, with

$$\mathcal{F} = \Bigg\{ \psi_{\gamma_{[2k]}} : \psi_{\gamma_{[2k]}}(x,y) = \log\Bigg(\frac{\sum_{i=1}^k \exp\big(\frac{-1}{2\sigma^2}(y-\gamma_{2i-1}^\top x)^2\big)}{\sum_{i=1}^k \exp\big(\frac{-1}{2\sigma^2}(y-\gamma_{2i}^\top x)^2\big)} \Bigg) \Bigg\}.$$

This provides a tractable minimax optimization problem whose solution is provably close to that of (6). To find the above parameterized space for ψ, we apply Brenier's theorem connecting the optimal ψ to a transport map between two MLR models.

Lemma 1 (Brenier's Theorem, (Villani, 2008))**.** Assume $X \sim p_x$, and consider random variables Y and Y' such that $(X,Y) \sim p_{\beta^*_{[k]}}$ and $(X,Y') \sim p_{\beta_{[k]}}$ provide two MLR models according to $\beta^*_{[k]}$ and $\beta_{[k]}$, respectively. Then, the optimal ψ in (6) satisfies the following transportation property

$$(X, Y - \psi_y(X,Y)) \overset{d}{=} (X,Y'), \quad (8)$$

where $\psi_y(x,y) := \frac{\partial}{\partial y}\psi(x,y)$.

The above lemma shows that the optimal transport map's derivative will transport samples between the two domains. Therefore, we need to characterize the potential optimal transport maps and consider their integral for constraining ψ. To do this, we find an approximation of this optimal mapping in two steps: First, we use a randomized technique, adapted from (Farnia et al., 2020), to come up with a mapping Ψ that maps (X,Y,Z) to (X,Y') where Z is the regression index for (X,Y). Then, we obtain $\tilde{\Psi}$ by taking the expectation of Ψ with respect to Z to drop the dependence of Z. We bound the error of this approximation step in Theorem 1.

For the first step, consider the following randomized transportation map:

$$\Psi(X,Y,Z) := Y + \sum_{i=1}^k \mathbb{1}\{Z=i\}(\beta_i - \beta_i^*)^\top X, \quad (9)$$

where Z denotes the regression model index in the first mixture (X,Y). Note that the above randomized map will transport samples between the two MLR distributions, i.e.,

$$\big(X, Y + \sum_{i=1}^k \mathbb{1}\{Z=i\}(\beta_i - \beta_i^*)^\top X\big) \overset{d}{=} (X,Y').$$

However, the above mapping is a randomized function of x,y since Z remains random after observing the outcome for x,y. To obtain a deterministic map $\tilde{\Psi} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ from this randomized map, we consider its conditional expectation given (X,Y):

$$\begin{aligned}\tilde{\Psi}(x,y) &:= \mathbb{E}\big[\Psi(X,Y,Z)|X=x,Y=y\big] \\ &= y + \sum_{i=1}^k \mathbb{P}(Z=i|X=x,Y=y)(\beta_i - \beta_i^*)^\top x.\end{aligned} \quad (10)$$

In the above equation, by Bayes' rule we have

$$\begin{aligned}&\mathbb{P}(Z=i|X=x,Y=y) \\ &= \frac{\exp(\frac{-1}{2\sigma^2}(y-(\beta_i^*)^\top x)^2)}{\sum_{j=1}^k \exp(\frac{-1}{2\sigma^2}(y-(\beta_j^*)^\top x)^2)} \\ &= \frac{\exp(\frac{-1}{2\sigma^2}y(\beta_i^*)^\top x)\exp(\frac{-1}{2\sigma^2}((\beta_i^*)^\top x)^2)}{\sum_{j=1}^k \exp(\frac{-1}{2\sigma^2}y(\beta_j^*)^\top x)\exp(\frac{-1}{2\sigma^2}((\beta_j^*)^\top x)^2)}.\end{aligned}$$

Note that if $\tilde{\Psi}$ was the optimal transport with $\frac{\partial}{\partial y}\tilde{\psi}(x,y) = \tilde{\Psi}(x,y)$, then

$$W_2\big(p_{\beta_{[k]}}, p_{\beta^*_{[k]}}\big) = \mathbb{E}_{p_{\beta^*_{[k]}}}[\tilde{\psi}(x,y)] - \mathbb{E}_{p_{\beta_{[k]}}}[\tilde{\psi}^c(x,y)]$$

With $\tilde{\Psi}$ as an approximate solution, we next state the following result which bounds the duality gap of $\tilde{\Psi}$.

Theorem 1. Let $\tilde{\psi} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ be a convex function such that for every $x \in \mathcal{X}$, $\frac{\partial}{\partial y}\tilde{\psi}(x,Y)$ shares the same distribution with $\tilde{\Psi}(x,Y)$ in (10)[1]. Assume that for every $x \in \mathcal{X}$ and every β_i we have $|(\beta_i^*)^\top x| \le C'$. For an observation of input and output (X,Y) with regression index

[1]The existence of $\tilde{\psi}$ is guaranteed based on Brenier's theorem.

Z, we denote the optimal Bayes classifier of the cluster of (X, Y) as $Z^*(X, Y)$. Let $P_{\text{err}} := \mathbb{P}(Z \neq Z^*(X, Y))$ be the probability error of the Bayes classifier. Then, we have:

$$0 \leq W_2\big(p_{\beta_{[k]}}, p_{\beta^*_{[k]}}\big) - \mathbb{E}_{p_{\beta^*_{[k]}}}[\tilde{\psi}(x, y)] + \mathbb{E}_{p_{\beta_{[k]}}}[\tilde{\psi}^c(x, y)]$$
$$\leq 16(C'^2 + 2\sigma^2)\sqrt{P_{\text{err}}} + 2(C'^2 + \sigma^2)\sqrt[4]{P_{\text{err}}}.$$

Proof. See Appendix A. □

Finally, we estimate $\tilde{\psi}$ with a function from $\mathcal{F}$ that does not depend on the optimal $\beta^\star_{[k]}$.

Proposition 1. Assume that $\sum_{i=1}^k |\mathbb{P}_{(\beta_j)_{j=1}^k}(Z = i|X = x, Y = y) - \mathbb{P}_{(\beta^*_j)_{j=1}^k}(Z = i|X = x, Y = y)| \leq \delta$ and $\max_i |\beta_i^\top x|, \max_i |(\beta^*)_i^\top x| \leq C'$ for every $x \in \mathcal{X}$ and feasible β_i. Then, there exists $(\gamma_i)_{i=1}^{2k}$ such that the function

$$\psi_{\gamma_{[2k]}}(x, y) = \log\left(\frac{\sum_{i=1}^k \exp\big(\frac{-1}{2\sigma^2}(y - \gamma_{2i-1}^\top x)^2\big)}{\sum_{i=1}^k \exp\big(\frac{-1}{2\sigma^2}(y - \gamma_{2i}^\top x)^2\big)}\right), \tag{11}$$

approximates $\tilde{\psi}$, with error bounded by $C'\delta$.

Proof. See Appendix B. □

Combining (6) and Proposition 1, we formulate the following minimax problem which approximates (6):

$$\min_{\beta_{[k]}} \max_{\gamma_{[2k]}} \mathbb{E}_{p_{\beta^*_{[k]}}}[\psi_{\gamma_{[2k]}}(x, y)] - \mathbb{E}_{p_{\beta_{[k]}}}[\psi^c_{\gamma_{[2k]}}(x, y)]. \tag{12}$$

By Proposition 1 and Theorem 1, the approximation error is bounded when the clusters can be identified with high precision by the optimal Bayes classifier. This condition can be thought of as a separability condition.

3.1. Reducing c-transform to Norm Regularization

In order to simplify the c-transform operation, we introduce a regularization penalty term to substitute the c-transform term in (14). To do this, we bound the expected value of the c-transform $\psi^c(x, y)$ (7) by the expectation of $\psi(x, y)$ and a regularization term. This bound, given in the following proposition, allows us to formulate a strongly-concave maximization problem.

Proposition 2. Consider the discriminator function $\psi_{\gamma_{[2k]}}(x, y)$ in (17) and recall that $\|x\| \leq C$. Assume that $2kC^2 \max_i \|\gamma_i\|^2 \leq \eta < 1$. Then, for any set of vectors $\tilde{\gamma}_{[2k]} \in \mathbb{R}^{2k\times d}$, we have

$$\mathbb{E}\Big[\psi^c_{\gamma_{[2k]}}(x, y)\Big] \leq \mathbb{E}\big[\psi_{\gamma_{[2k]}}(x, y)\big] + \frac{kC^2\mathbb{E}\big[(1 + C|y|)^2\big]}{1 - \eta}$$
$$\times \left(\sum_{i=1}^k \|\gamma_i - \tilde{\gamma}_i\|^2 + \|\gamma_{i+k} - \tilde{\gamma}_i\|^2\right). \tag{13}$$

Algorithm 1 WMLR

Input: $(x_i, y_i)_{i\in[n]}$, $\beta^{(0)}_{[k]}, \gamma^{(0)}_{[2k]}$, step sizes $\alpha_{\min}, \alpha_{\max}$
for $t = 0$ **to** $T - 1$ **do**
  for $i = 1$ **to** k **do**
    $\beta_i^{(t+1)} = \beta_i^{(t)} - \alpha_{\min}\nabla_{\beta_i}\widehat{\mathcal{L}}(\beta^{(t)}_{[k]}, \gamma^{(t)}_{[2k]})$
    $\gamma_i^{(t+1)} = \gamma_i^{(t)} + \alpha_{\max}\nabla_{\gamma_i}\widehat{\mathcal{L}}(\beta^{(t)}_{[k]}, \gamma^{(t)}_{[2k]})$
    $\gamma_{i+k}^{(t+1)} = \gamma_{i+k}^{(t)} + \alpha_{\max}\nabla_{\gamma_{i+k}}\widehat{\mathcal{L}}(\beta^{(t)}_{[k]}, \gamma^{(t)}_{[2k]})$
  end for
end for

Proof. See Appendix C. □

3.2. WMLR Algorithm

It can be seen that (12) represents a nonconvex-nonconcave optimization problem. As shown in Proposition 2, we could bound the c-transform by adding a regularization, and, as a result, we obtain the following nonconvex strongly-concave minimax problem

$$\min_{\beta_{[k]}} \max_{\gamma_{[2k]}} \mathcal{L}(\beta_{[k]}, \gamma_{[2k]}) :=$$
$$\mathbb{E}_{p_{\beta^*_{[k]}}}[\psi_{\gamma_{[2k]}}(x, y)] - \mathbb{E}_{p_{\beta_{[k]}}}[\psi_{\gamma_{[2k]}}(x, y)]$$
$$- \lambda\left(\sum_{i=1}^k \|\gamma_i - \tilde{\gamma}_i\|^2 + \|\gamma_{i+k} - \tilde{\gamma}_i\|^2\right), \tag{14}$$

where $\tilde{\gamma}$ is a properly chosen reference vector.

Since we do not have access to $p_{\beta^*_{[k]}}$ in practice, we replace $\mathbb{E}_{p_{\beta^*_{[k]}}}[\psi_{\gamma_{[2k]}}(x, y)]$ above with $\mathbb{E}_{\hat{p}}[\psi_{\gamma_{[2k]}}(x, y)]$ where $\hat{p}$ is the empirical problem over the observed dataset $\mathcal{S} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. We denote the resulting function by $\widehat{\mathcal{L}}(\beta_{[k]}, \gamma_{[2k]})$.

WMLR, given in Algorithm 1, uses GDA to solve (14). Later, we show that solving (14) can recover the underlying $\beta_{[k]}$ that solves the original unregularized (12).

Federated Learning Since we use the gradient-based GDA algorithm to solve the minimax optimization problem, WMLR is particularly amenable to distributed computation. Here, we consider a federated learning setting with M agents, where each agent m has data samples $(x_{i,m}, y_{i,m})_{m\in[M],\, i\in[N]}$. This setting can model both the following scenarios: 1) each sample belongs to any of the k components with equal probability, as in the centralized case; or 2) all the samples for each individual agent are associated with the same cluster. The latter scenario arises when an unknown latent variable governs the regression model that best describes the relationship between y and x. We note that our proposed algorithm can apply to both these cases. Every agent only has access to its own data and

Algorithm 2 F-WMLR

Input: $(x_{i,m}, y_{i,m})_{m\in[M],\, i\in[N]}$, $\beta_{[k]}^{(0)}, \gamma_{[2k]}^{(0)}$, step sizes $\alpha_{\min}, \alpha_{\max}$.
for $t = 0$ **to** $T-1$ **do**
  Broadcast $\beta_{[k]}^{(0)}, \gamma_{[2k]}^{(0)}$ to all agents
  for each agent $m = 1$ **to** M **do**
    for $i = 1$ **to** k **do**
      $\beta_{i,m}^{(t+1)} = \beta_i^{(t)} - \alpha_{\min}\nabla_{\beta_i}\widehat{\mathcal{L}}_m(\beta_{[k]}^{(t)}, \gamma_{[2k]}^{(t)})$
      $\gamma_{i,m}^{(t+1)} = \gamma_i^{(t)} + \alpha_{\max}\nabla_{\gamma_i}\widehat{\mathcal{L}}_m(\beta_{[k]}^{(t)}, \gamma_{[2k]}^{(t)})$
      $\gamma_{i+k,m}^{(t+1)} = \gamma_{i+k}^{(t)} + \alpha_{\max}\nabla_{\gamma_{i+k}}\widehat{\mathcal{L}}_m(\beta_{[k]}^{(t)}, \gamma_{[2k]}^{(t)})$
    end for
    Send $\beta_{[k],m}^{(t+1)}, \gamma_{[2k],m}^{(t+1)}$ to server
  end for
  Collect $\beta_{[k],m}^{(t+1)}, \gamma_{[2k],m}^{(t+1)}$ from all agents $m \in [M]$
  for $i = 1$ **to** k **do**
    $\beta_i^{(t)} = \frac{1}{M}\sum_{m=1}^M \beta_{i,m}^{(t+1)}$
    $\gamma_i^{(t)} = \frac{1}{M}\sum_{m=1}^M \gamma_{i,m}^{(t+1)}$
    $\gamma_{i+k}^{(t)} = \frac{1}{M}\sum_{m=1}^M \gamma_{i+k,m}^{(t+1)}$
  end for
end for

therefore can only estimate its own minimax objective $\widehat{\mathcal{L}}_m$. Therefore, the total minimax objective in the network will be

$$\widehat{\mathcal{L}}(\beta_{[k]}^{(t)}, \gamma_{[2k]}^{(t)}) = \frac{1}{M}\sum_{m=1}^M \widehat{\mathcal{L}}_m(\beta_{[k]}^{(t)}, \gamma_{[2k]}^{(t)}), \tag{15}$$

where $\widehat{\mathcal{L}}_m$ computes $\mathbb{E}_{\hat{p}}$ using only the data on agent m. Our Federated WMLR (F-WMLR) algorithm adds a communication step after each GDA iteration, as described in Algorithm 2. This algorithm could be extended to include multiple GDA steps or partial agent participation at each round. We show that F-WMLR enjoys the same theoretical guarantees as WMLR in Section 4 below.

3.3. Generalization to Non-linear Models

The WMLR algorithm can also be used for the setting where the output is a mixture of linear regressions of a nonlinear transformation of the input vector that is common to all components. The corresponding ψ function will be

$$\psi_{\phi,\gamma_{[2k]}}(x,y) = \log\left(\frac{\sum_{i=1}^k \exp\left(\frac{-1}{2\sigma^2}(y - \gamma_{2i-1}^\top\phi(x))^2\right)}{\sum_{i=1}^k \exp\left(\frac{-1}{2\sigma^2}(y - \gamma_{2i}^\top\phi(x))^2\right)}\right). \tag{16}$$

Our theoretical results, discussed in the subsequent section, do not extend to the nonlinear case in general. However, WMLR will still convergence to a minimax stationary point when ψ has the form (16).

4. Convergence Guarantees for WMLR

In this section, we focus on the case $k = 2$, and further explore the minimax formulation (14). In particular, to simplify the derivations, we focus on the *symmetric* case, i.e., $\beta_2^* = -\beta_1^*$, which has been studied in the in EM literature as well (see (Kwon et al., 2019) and references therein).The non-symmetric case can be reduced to the symmetric case, by first estimating $\bar{\beta}$ as the mean of $\beta_{[2]}^*$, and then replacing each data point (x_i, y_i) by $(x_i, y_i - \bar{\beta}^\top x_i)$.

In the symmetric setting, we have that $\gamma_3 = -\gamma_1$ and $\gamma_4 = -\gamma_2$ in $\psi_{\gamma_{[4]}}$ and $\beta_2 = -\beta_1$ in $p_{\beta_{[2]}}$. We next observe that, in this case, $\psi_{\gamma_{[4]}}$ can be decomposed into the following two terms

$$\begin{aligned}
&\psi_{\gamma_{[4]}}(x,y)\\
&= \log\left(\frac{\exp\left(\frac{-1}{2\sigma^2}(y-\gamma_1^\top x)^2\right) + \exp\left(\frac{-1}{2\sigma^2}(y+\gamma_1^\top x)^2\right)}{\exp\left(\frac{-1}{2\sigma^2}(y-\gamma_2^\top x)^2\right) + \exp\left(\frac{-1}{2\sigma^2}(y+\gamma_2^\top x)^2\right)}\right)\\
&= x^\top A x + \log\left(\exp(\frac{y\gamma_1^\top x}{2\sigma^2}) + \exp(\frac{-y\gamma_1^\top x}{2\sigma^2})\right)\\
&\quad - \log\left(\exp(\frac{y\gamma_2^\top x}{2\sigma^2}) + \exp(\frac{-y\gamma_2^\top x}{2\sigma^2})\right),
\end{aligned}$$

where $A := \frac{\gamma_1\gamma_1^\top - \gamma_2\gamma_2^\top}{2\sigma^2}$.

Since the marginal distribution of $p_{\beta_{[k]}}$ over X is constant (and equal to p_x), we can ignore the quadratic term $x^\top A x$ as it will be canceled out in $\mathbb{E}_{p_{\beta_{[2]}^*}}[\psi_{\gamma_{[4]}}(x,y)] - \mathbb{E}_{p_{\beta_{[2]}}}[\psi^c_{\gamma_{[4]}}(x,y)]$. Furthermore, we can absorb $2\sigma^2$ into γ_1 and γ_2. Thus, we can replace $\psi_{\gamma_{[4]}}$ in (14) with $k = 2$ by

$$\begin{aligned}
\psi_{\gamma_1,\gamma_2}(x,y) &:= \log\left(\exp(y\gamma_1^\top x) + \exp(-y\gamma_1^\top x)\right)\\
&\quad - \log\left(\exp(y\gamma_2^\top x) + \exp(-y\gamma_2^\top x)\right).
\end{aligned} \tag{17}$$

As a result, and in this section, we work with $\psi_{\gamma_1,\gamma_2}(x,y)$ instead of $\psi_{\gamma_{[4]}}$ in (14). Also, we simplify $\mathcal{L}(\beta_{[2]}, \gamma_{[4]})$ and $\widehat{\mathcal{L}}(\beta_{[2]}, \gamma_{[4]})$ by $\mathcal{L}(\beta, \gamma_1, \gamma_2)$ and $\widehat{\mathcal{L}}(\beta, \gamma_1, \gamma_2)$, respectively.

Our goal is to solve the minimax problem (14) to a *minimax stationary point*, which we define below.

Definition 4.1. Consider a function $f(x,y)$, where $f(x,\cdot)$ is strongly concave for all x. The point $x^\star$ is an ϵ minimax stationary point of

$$\min_x \max_y f(x,y) \tag{18}$$

if $\|\nabla_x F(x^\star)\| \le \epsilon$, and $F(x) = \max_y f(x,y)$.

To discuss the convergence to stationary points in our setting, we define

$$\begin{aligned}
\mathcal{L}(\beta) &:= \max_{\gamma_{[2]}} \mathcal{L}(\beta, \gamma_1, \gamma_2)\\
\widehat{\mathcal{L}}(\beta) &:= \max_{\gamma_{[2]}} \widehat{\mathcal{L}}(\beta, \gamma_1, \gamma_2).
\end{aligned} \tag{19}$$

The outline of our theoretical results is as follows: We first show that the added regularization term forms a strongly concave inner maximization problem, and using that, in Theorem 2, we show WMLR finds the minimax stationary point solution of $\widehat{\mathcal{L}}$. Next, in Theorem 3, we show that under certain assumptions, this solution is optimal β^*. Finally, we provide bounds on the generalization error as well.

4.1. Local and Global Convergence of WMLR

In this subsection, we show that the GDA algorithm is guaranteed to converge to the optimal solution to (14).

Theorem 2. Consider the minimax problem (14). Assume that $C^2\mathbb{E}_{\hat{p}}[y^2] \leq \eta < \frac{\lambda}{2}$ and $C^2 < \frac{\lambda}{2}$. Then the WMLR algorithm (Algorithm 1) with step sizes $\alpha_{\max} = \frac{1}{L}$ and $\alpha_{\min} = \frac{1}{\kappa^2 L}$ for $L = \lambda + 4\eta(1+\eta/\lambda+\|\tilde{\gamma}\|)$ and $\kappa = \frac{L}{\lambda - 2\eta}$ will find an ϵ-approximate stationary point in the following number of iterations:

$$\mathcal{O}\left(\frac{\kappa^2 L\Delta + \kappa L^2(2\eta/\lambda)^2}{\epsilon^2}\right),$$

where $\Delta := \widehat{\mathcal{L}}(\beta^{(0)}) - \min_\beta \widehat{\mathcal{L}}(\beta)$.

Proof. See Appendix D. □

Remark 1. Consider the minimax problem (14) where x is replaced by non-linear $\phi(\cdot; w)$, a neural network parameterized by weights w. The weights w appear in the minimization problem; hence, the problem remains nonconvex strongly-concave and the guarantee in Theorem 2 also applies to the non-linear case, *i.e.,* WMLR still results in an approximate stationary point.

In Theorem 3 below, we show global convergence under correlated projections along β^* and $\tilde{\gamma}$.

Theorem 3. Consider two symmetric components for feature variables x. Suppose that the variables $\tilde{\gamma}^\top x$ and ${\beta^*}^\top x$ are correlated enough such that

$$\max\left\{\mathbb{P}\left(\tilde{\gamma}^\top x x^\top \beta^* \leq 0\right), \mathbb{P}\left(\tilde{\gamma}^\top x x^\top \beta^* \geq 0\right)\right\} = 1.$$

Then, any stationary minimax solution $\widehat{\beta}$ to the minimiax problem (14) which satisfies the above condition will further provide a global minimax solution to (14).

Proof. See Appendix E. □

The above theorem shows that if $\tilde{\gamma}$ and β^* are sufficiently aligned such that the random variables $\tilde{\gamma}^\top x$ and ${\beta^*}^\top x$ are correlated enough, then a stationary minimax point for the WMLR's minimax problem further leads to a global solution to the WMLR problem.

Note that the condition in the theorem statement automatically holds for a 1-dimensional scalar x. In general, the theorem condition suggests that we need to chose the reference vector $\tilde{\gamma}$ almost aligned to β^*. One way to do so is as follows: First, note that $\beta^*_{\text{norm}} := \beta^*/\|\beta^*\|$ is the top eigenvector of $\mathcal{M} := \mathbb{E}_x[(x^\top\beta^*)^2 x x^\top]$. Let us assume the top eigenvector of $\mathcal{M}$ is unique, i.e., $\beta^*_{\text{norm}}$is the only eigenvector corresponding to the maximum eigenvalue of $\mathcal{M}$. In that case, it can be shown that for sufficiently large n, β^*_{norm} is approximately the top eigenvector of $M_n := \frac{1}{n}\sum_{i=1}^n y_i^2 x_i x_i^\top$. To see this, we need to show the solution to $\text{argmax}_{v:\|v\|=1} v^\top M_n v$ is close to β^*_{norm}. To do so, first note that by classic concentration bounds we could show that, for sufficiently large n, $v^\top M_n v$ is close to $\mathbb{E}[\|v^\top(xy)\|^2]$. That said, maximizing $\mathbb{E}[\|v^\top(xy)\|^2]$ over v is equivalent to maximizing $\mathbb{E}[\|v^\top(xx^\top\beta^*)\|^2] = v^\top\mathbb{E}_x[(x^\top\beta^*)^2 x x^\top]v = v^\top\mathcal{M}v$ over v, and we assumed β^*_{norm} is the unique solution to the latter problem. We further evaluate this choice of refrence vector in the our numerical experiments.

Remark 2. (Federated Learning) F-WMLR (Algorithm 2) will produce the same sequence of iterates as the centralized WMLR algorithm by linearity of the gradient operator. Therefore, the above convergence results for WMLR will also apply to F-WMLR.

4.2. Generalization of WMLR

Here we establish generalization error bounds for the convergence of the value and gradient of the empirical objective to those of the underlying distribution.

Theorem 4. Recall the definition of $\mathcal{L}(\beta)$ and $\widehat{\mathcal{L}}(\beta)$ (19). Consider the minimax mixed regression setting with norm-bounded random vector X, $\|X\|_2 \leq C$ and noise vector $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Assume that $\max\{C, \sigma\} \leq 1$. Then, we have the following generalization bounds hold with probability at least $1-\delta$ for every $\|\beta\|_2 \leq \eta$:

$$|\mathcal{L}(\beta) - \widehat{\mathcal{L}}(\beta)| \leq O\Big(\sqrt{\frac{d\eta^4 \log(\eta/\lambda\delta)}{n}}\Big),$$

$$\|\nabla\mathcal{L}(\beta) - \nabla\widehat{\mathcal{L}}(\beta)\|_2 \leq O\Big(\sqrt{\frac{d\eta^4 \log(\eta/\lambda\delta)}{(1-\eta/\lambda)^2 n}}\Big).$$

5. Numerical Experiments

We consider $k = 2$ and focus on the symmetric case with $\beta^*_2 = -\beta^*_1$ for the numerical experiments. We implement[2] Algorithms 1 and 2 in Section 3 for both the centralized and federated learning settings. In both settings, we run experiments for a high SNR (10) and a low SNR (1) regime. We include two additional SNRs in the federated experiments. We set $d = 128$, draw x_i from $\mathcal{N}(0, I)$, set noise variance $\sigma^2 = 1$, and draw $\beta^\star$ uniformly at random from the spherical shell $\mathcal{S}_{\text{SNR}} = \{z \mid \|z\| = \text{SNR}\}$. We search over

[2] https://github.com/tjdiamandis/WMLR.

regularization parameter λ, and step sizes are $\alpha_{\max} = 1/2\lambda$ and $\alpha_{\min} = \alpha_{\max}/10$ (motivated by Theorem 2). Note that the algorithms operate without the knowledge of the noise variance or SNR.

Evaluation Metrics We evaluate methods in terms of the relative error $\frac{\|\hat{\beta}-\beta^*\|}{\|\beta^\star\|}$, where $\hat{\beta}$ is the last iterate of the applied method, and the negative log likelihood (NLL) for the symmetric 2-component MLR problem (4). Note that NLL can be computed without knowledge of the true underlying regressor $\beta^\star$ and noise variance σ^2.

Baselines We compare WMLR against EM and Gradient EM (GEM), which is similar to EM, but instead of solving the maximization problem at each iteration, it takes one gradient ascent step. The noise variance for EM and GEM is initialized as $\sigma^{2^{(0)}} = 1$. See Appendix G for additional details and discussion of the EM and GEM algorithms for two-component MLR. We do not compare these algorithms to GAN based methods in this work since GAN-based methods usually take thousands of iterations to converge, as shown by Farnia et al. (2020) for Gaussian mixture models.

5.1. Centralized Setting

For all experiments, the initial iterates $\beta^{(0)}$, $\gamma_1^{(0)}$ and $\gamma_2^{(0)}$ are all chosen i.i.d. from $\mathcal{N}(0, \frac{1}{d}I)$. Note that these initializations will have approximately unit norm (Vershynin, 2018). We use the eigenvector of $\mathbb{E}\left[y^2xx^T\right]$ associated with the largest eigenvalue as the reference vector $\tilde{\gamma}$; however, the algorithm is not sensitive to this parameter. WMLR simply needs a reference vector that has non-negligible correlation with $\beta^\star$ to avoid vanishing gradients (also see Theorem 3).

We compare the solution reached at iteration 100 of each algorithm in Table 1. We evaluate each algorithm over several hyperparameter choices, and we choose the run with the smallest final negative log likelihood. Both WMLR and EM converge quickly (under 100 iterations) while GEM often does not converge by that number of iterations, as seen in the higher SNR case. In the low SNR case, all three algorithms have similar performance. In the high SNR case, WMLR outperforms EM and GEM both in terms of negative log likelihood and the distance to the true parameter. However, one drawback of WMLR and GEM compared to EM is that WMLR and GEM require hyperparameter tuning. For additional discussion, implementation details, and the hyperparameter selection, see Appendix H.

5.2. Federated Setting

As described in Algorithm 2, we extend WMLR to F-WMLR by broadcasting the model to all agents from the central node at each iteration, having each agent take one gradient decent ascent step using his or her own data, and then averaging the resulting new iterates at the central node.

Table 1. Comparison of algorithms at iteration $T = 100$ ($\beta^{(T)}$) in terms of negative log likelihood (NLL) and relative ℓ_2 error, $\|\beta^{(T)} - \beta^\star\|/\|\beta^\star\|$.

Centralized Experiments, $n = 100,000$			
SNR	Method	NLL	Relative ℓ_2 error
10	EM	2.115	3.79×10^{-2}
	GEM	3.765	1.03
	WMLR	**2.059**	$\mathbf{5.31 \times 10^{-3}}$
1	EM	1.657	8.62×10^{-2}
	GEM	**1.656**	$\mathbf{5.20 \times 10^{-2}}$
	WMLR	**1.656**	7.78×10^{-2}
Centralized Experiments, $n = 10,000$			
10	EM	2.715	1.21×10^{-1}
	GEM	3.758	9.98×10^{-1}
	WMLR	**2.065**	$\mathbf{2.08 \times 10^{-2}}$
1	EM	1.671	2.95×10^{-1}
	GEM	**1.657**	$\mathbf{1.80 \times 10^{-1}}$
	WMLR	1.668	2.75×10^{-1}

Recall that the EM algorithm operates via two repeated steps: an expectation step and a *full* maximization step. However, in the federated setting, we cannot expect the average of the maximizers to be the maximizer of the average. Here, we implement EM in the following way: For each maximization, we perform several communication rounds to solve the maximization problem at each EM step via gradient ascent. We stop this inner maximization when the norm of the gradient is under the threshold $\nu = 0.01$ or after 50 iterations.

We simulate $M = 10,000$ agents with 10 data points each. We assume that each agent $m \in \{1, ..., M\}$ has all her samples drawn from only one of of the two regressors, *i.e.,* agent m's samples $(y_{m,i}, x_{m,i})_{i=1}^n$ satisfy

$$y_{m,i} = z_m(\beta^*)^\top x_{m,i} + \epsilon_{m,i}, \quad i = 1, ..., n, \tag{20}$$

where z_m is drawn from $\mathrm{Unif}(\{-1, 1\})$. Again, we draw $\beta^{(0)}, \gamma_1^{(0)}, \gamma_2^{(0)}$ from $\mathcal{N}(0, \frac{1}{d}I)$. The final solutions and the convergence behaviors of these algorithms are compared in Table 2 and Figure 1. EM does not converge in 10,000 iterations for the medium and high SNR cases. In our experiments, GEM and WMLR both converged to a comparable level of relative error (Table 2). Theoretically, both of these methods should converge to the optimal $\beta^\star$ in the population case, so the observed error is mainly due to their generalization performance. Although WMLR takes longer per iteration (about 3min for WMLR vs 1min for GEM in the federated case), WMLR is overall much faster due to the small number of iterations. WMLR consistently converges in 60 to 100 iterations regardless of SNR, whereas GEM is

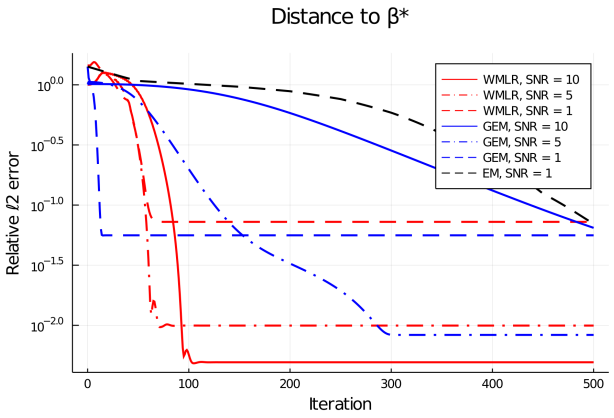

Figure 1. Convergence of $\hat{\beta}$ to $\beta^\star$ in the federated setting with 10,000 nodes with 10 samples each. EM is removed for tests which it did not converge to a reasonable value within 5,000 iterations.

fast in the low SNR cases but is up to 175x slower in the highest SNR case.

In addition, there is a significant communication cost in the federated setting. Therefore, WMLR's smaller iteration number is particularly important in this setting. While WMLR's implementation is more complex, WMLR enjoys higher robustness to the choice of hyperparameters than GEM. The same hyperparamaters work for all tested SNRs (Figure 2 and Table 3 in the appendix), and iteration count is comparable across all SNRs. Since communication rounds are very costly in the federated learning setting, these results suggest that WMLR may be better equipped than GEM or EM to handle distributed multimodal learning tasks. Additional details are provided in Appendix H.

Table 2. Comparison of algorithms at the final iterate in terms of relative ℓ_2 error, $\|\beta^{(T)} - \beta^\star\|/\|\beta^\star\|$. The iterations required for convergence is also compared. Note that EM did not convergence (d.n.c.) for SNR = 10 and SNR = 5 cases within 5,000 iterations.

Federated Experiments, Final Iterate			
SNR	Method	Iterations Req.	Relative ℓ_2 error
20	EM	d.n.c	d.n.c
	GEM	12,948	$\mathbf{1.93 \times 10^{-3}}$
	WMLR	**74**	2.49×10^{-3}
10	EM	d.n.c	d.n.c
	GEM	2,007	$\mathbf{3.92 \times 10^{-3}}$
	WMLR	**98**	4.93×10^{-3}
5	EM	d.n.c	d.n.c
	GEM	295	$\mathbf{8.32 \times 10^{-3}}$
	WMLR	**81**	9.95×10^{-3}
1	EM	544	5.60×10^{-2}
	GEM	**15**	$\mathbf{5.60 \times 10^{-2}}$
	WMLR	66	7.25×10^{-2}

6. Acknowledgment

This paper was partially funded by MIT-IBM Watson AI Lab and Defense Science and Technology Agency. This work was also supported by the QuantERA grant C'MON-QSENS!. Alireza Fallah acknowledges support from MathWorks Engineering Fellowship.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. Generalization and equilibrium in generative adversarial nets (gans). In *International Conference on Machine Learning*, pp. 224–232. PMLR, 2017.

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

Bilmes, J. A. et al. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126, 1998.

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

Chen, Y., Yi, X., and Caramanis, C. A convex formulation for mixed regression with two components: Minimax optimal rates. In *Conference on Learning Theory*, pp. 560–604, 2014.

De Veaux, R. D. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, 8(3):227–245, 1989.

Deng, Y., Kamani, M. M., and Mahdavi, M. Distributionally robust federated averaging. *arXiv preprint arXiv:2102.12660*, 2021.

Fallah, A., Mokhtari, A., and Ozdaglar, A. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3557–3568. Curran Associates, Inc., 2020.

Farnia, F. and Tse, D. A convex duality framework for gans. *arXiv preprint arXiv:1810.11740*, 2018.

Farnia, F., Wang, W., Das, S., and Jadbabaie, A. GAT-GMM: Generative adversarial training for gaussian mixture models. *arXiv preprint arXiv:2006.10293*, 2020.

Feizi, S., Farnia, F., Ginart, T., and Tse, D. Understanding gans in the lqg setting: Formulation, generalization and stability. *IEEE Journal on Selected Areas in Information Theory*, 1(1):304–311, 2020.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617. PMLR, 2018.

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

Goodfellow, I. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pp. 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

Jordan, M. I. and Jacobs, R. A. Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2): 181–214, 1994.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Kwon, J. and Caramanis, C. EM converges for a mixture of many linear regressions. In *International Conference on Artificial Intelligence and Statistics*, pp. 1727–1736, 2020.

Kwon, J., Qian, W., Caramanis, C., Chen, Y., and Davis, D. Global convergence of the EM algorithm for mixtures of two component linear regression. In *Conference on Learning Theory*, pp. 2055–2110, 2019.

Li, X., Jiang, M., Zhang, X., Kamp, M., and Dou, Q. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021.

Lin, T., Jin, C., and Jordan, M. On gradient descent ascent for nonconvex-concave minimax problems. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6083–6093. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/lin20a.html.

Mansour, Y., Mohri, M., Ro, J., and Suresh, A. T. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

Margossian, C. C. A review of automatic differentiation and its efficient implementation. *WIREs Data Mining and Knowledge Discovery*, 9(4):e1305, 2019. doi: https://doi.org/10.1002/widm.1305. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1305.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Singh, A. and Zhu, J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR. URL http://proceedings.mlr.press/v54/mcmahan17a.html.

Mena, G., Nejatbakhsh, A., Varol, E., and Niles-Weed, J. Sinkhorn em: an expectation-maximization algorithm based on entropic optimal transport. *arXiv preprint arXiv:2006.16548*, 2020.

Mohri, M., Sivek, G., and Suresh, A. T. Agnostic federated learning. In *International Conference on Machine Learning*, pp. 4615–4625. PMLR, 2019.

Reisizadeh, A., Farnia, F., Pedarsani, R., and Jadbabaie, A. Robust federated learning: The case of affine distribution shifts. *arXiv preprint arXiv:2006.08907*, 2020.

Sanjabi, M., Ba, J., Razaviyayn, M., and Lee, J. D. On the convergence and robustness of training gans with regularized optimal transport. *arXiv preprint arXiv:1802.08249*, 2018.

Sattler, F., Müller, K.-R., and Samek, W. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

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

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

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

Zhong, K., Jain, P., and Dhillon, I. S. Mixed linear regression with multiple components. In *Advances in neural information processing systems*, pp. 2190–2198, 2016.