A Wasserstein Minimax Framework for Mixed Linear Regression

Theo Diamandis*, Yonina Eldar*, Alireza Fallah*, Farzan Farnia*, Asu Ozdaglar*

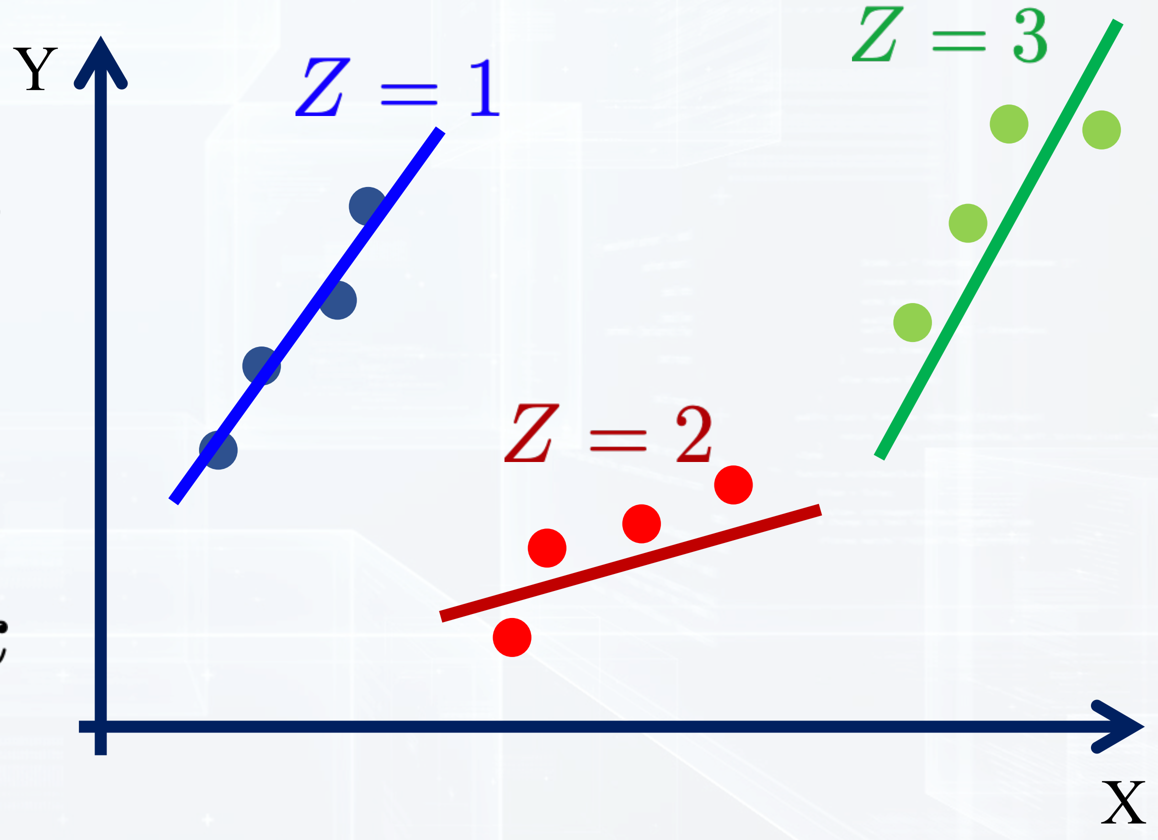
*: Massachusetts Institute of Technology, *: Weizmann Institute of Science

International Conference on Machine Learning 2021

Federated Learning and Mixture Models

Mixed Linear Regression Model

$$Y = \begin{cases} \boldsymbol{\beta}_1^\top \mathbf{X} + N, & \text{if } Z = 1 \\ \quad\vdots & \\ \boldsymbol{\beta}_k^\top \mathbf{X} + N, & \text{if } Z = k \end{cases}$$



Expectation-Maximization for Mixed Linear Regression

- The Maximum Likelihood (ML) approach for MLR is computationally hard.
- The EM algorithm is the standard method to solve ML for MLR models.
  - E-step: Estimate the latent variable from current parameters
  - M-step: Maximize the likelihood function based on the estimated latent variable.
- However, EM incurs great computational and communication costs in federated learning settings for the M-step at every iteration.

Wasserstein Mixed Linear Regression (WMLR)

- As in Maximum Likelihood, EM optimizes the KL-divergence:

$$\operatorname*{argmin}_{\boldsymbol{\beta}_{1:k}} D_{\mathrm{KL}}\left(P_{\mathrm{data}}, P_{\boldsymbol{\beta}_{1:k}}\right)$$

- We propose to minimize the Wasserstein distance in the WMLR approach

$$\operatorname*{argmin}_{\boldsymbol{\beta}_{1:k}} W_2\left(P_{\mathrm{data}}, P_{\boldsymbol{\beta}_{1:k}}\right)$$

WMLR as a Min-Max Optimization Problem

- We apply the Kantorovich duality to reduce WMLR to a minimax problem:

$$\min_{\boldsymbol{\beta}_{1:k}} \max_{\phi} \mathbb{E}_{P_{\text{data}}}\left[\phi(\mathbf{X}, Y)\right] - \mathbb{E}_{P_{\boldsymbol{\beta}_{1:k}}}\left[\phi^c(\mathbf{X}, Y)\right]$$

$$\phi^c(\mathbf{x}, y) := \max_{y'} \phi(\mathbf{x}, y) - \|y - y'\|^2$$

- Consider the minimax problem with unconstrained ϕ:
  - **Good news:** The population solution is the underlying MLR model. 🙂
  - **Bad news:** The computational and statistical costs are too heavy. 🙁

WMLR: Optimal Transport Theory for Optimization Design

- **Brenier's Theorem**: The optimal potential function's gradient in WMLR transports samples from the data distribution to learned model:

$$\left(\mathbf{X}_{\text{data}}, \frac{\partial}{\partial y}\phi^*(X_{\text{data}}, Y_{\text{data}})\right) \overset{\text{dist}}{=} \left(\mathbf{X}_{\text{model}}, Y_{\text{model}}\right)$$

MLR Models

WMLR: Optimal Transport Theory for Minimax Design

- Unimodal MLR: Linear transport map $\Longrightarrow$ Quadratic ϕ

Theorem: For a well-separable MLR with the component classification error p_e, the optimal ϕ can be approximated within $O(\sqrt[4]{p_e})$ error using the following softmax-based quadratic function:

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

WMLR Minimax Problem and Theoretical Guarantees

- Bounding the c-transform via a regularization term, we reduce WMLR to the following minimax problem with a nonconvex-concave structure:

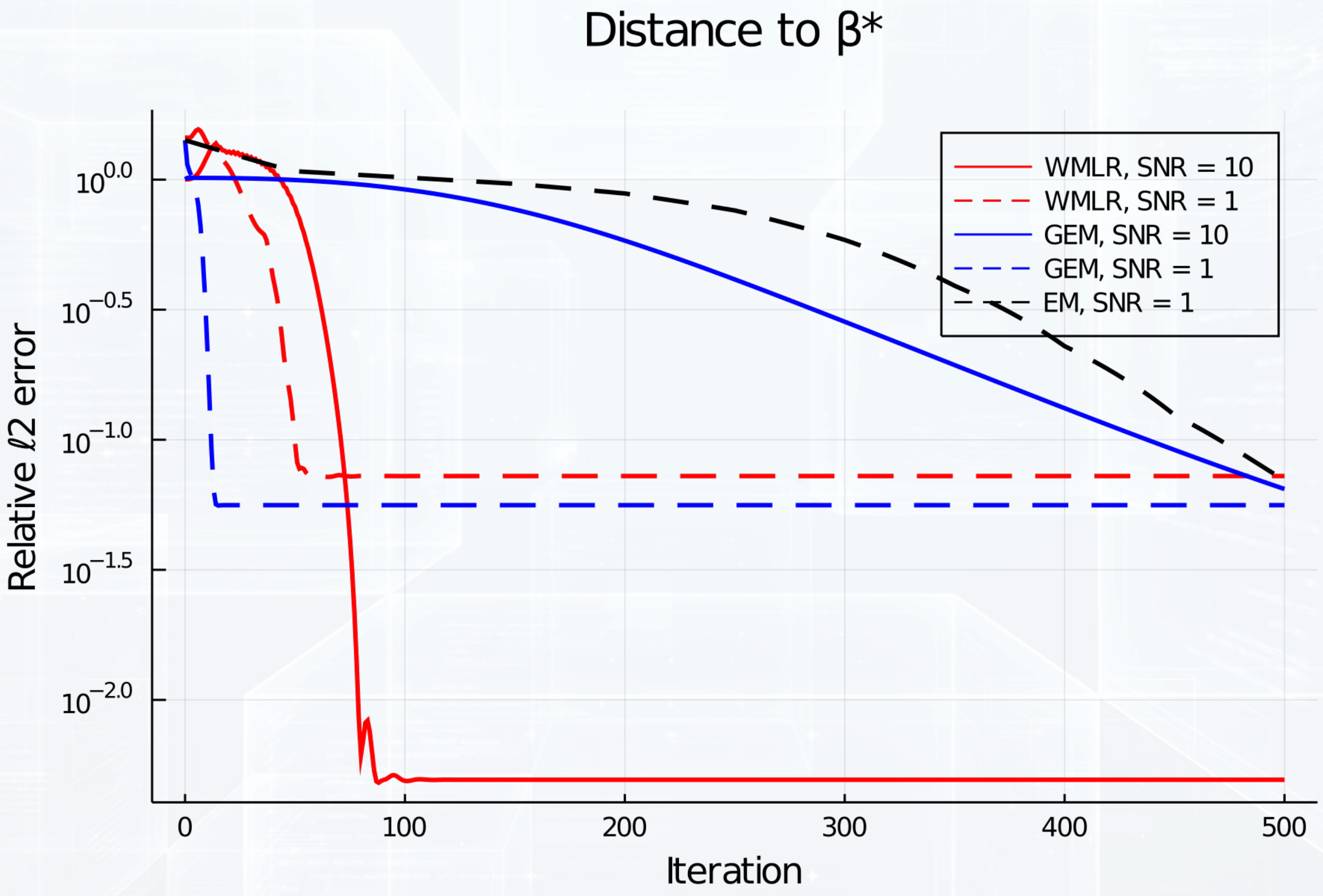
$$\min_{\beta_{1:k}} \max_{\gamma_{1:2k}} \mathbb{E}_{P_{\text{data}}}\left[\phi_{\gamma_{[2k]}}(\mathbf{X}, Y)\right] - \mathbb{E}_{P_{\beta_{1:k}}}\left[\phi_{\gamma_{[2k]}}(\mathbf{X}, Y)\right] - \lambda \left\| \gamma_{[2k]} \right\|_2^2$$

Theorem: (A) A gradient descent ascent (GDA) optimizer will solve the WMLR minimax problem to find a stationary minimax solution.

(B) For a mixture of two symmetric regression components, GDA can find the global minimax solution under the population distribution.

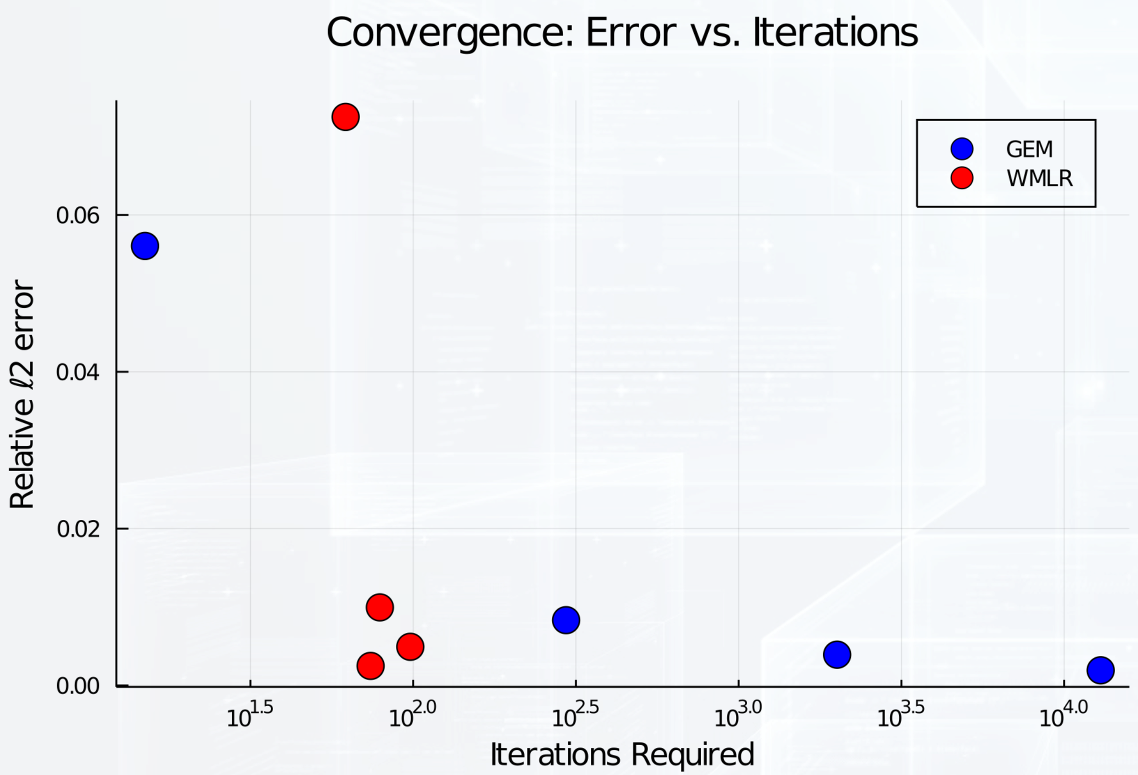
(C) GDA steps are capable of being decomposed to a distributed form.

Numerical Results: WMLR vs. EM baselines

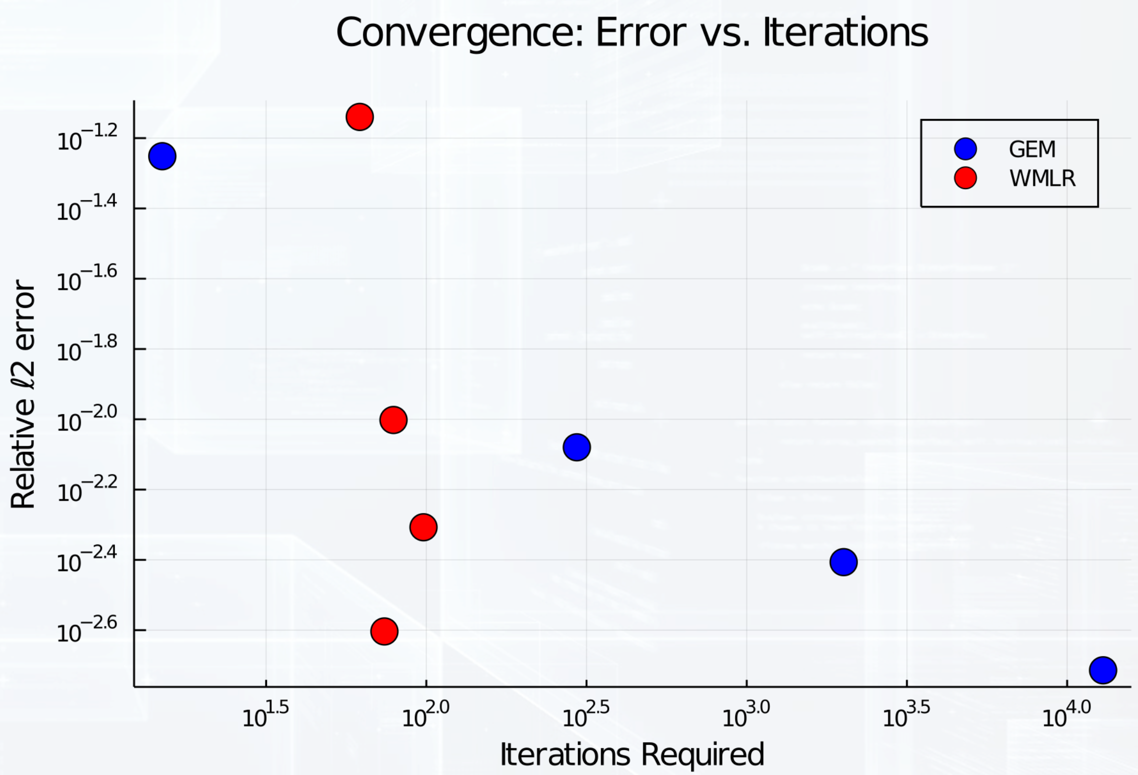


Federated Setting

Numerical Results: WMLR vs. EM baselines

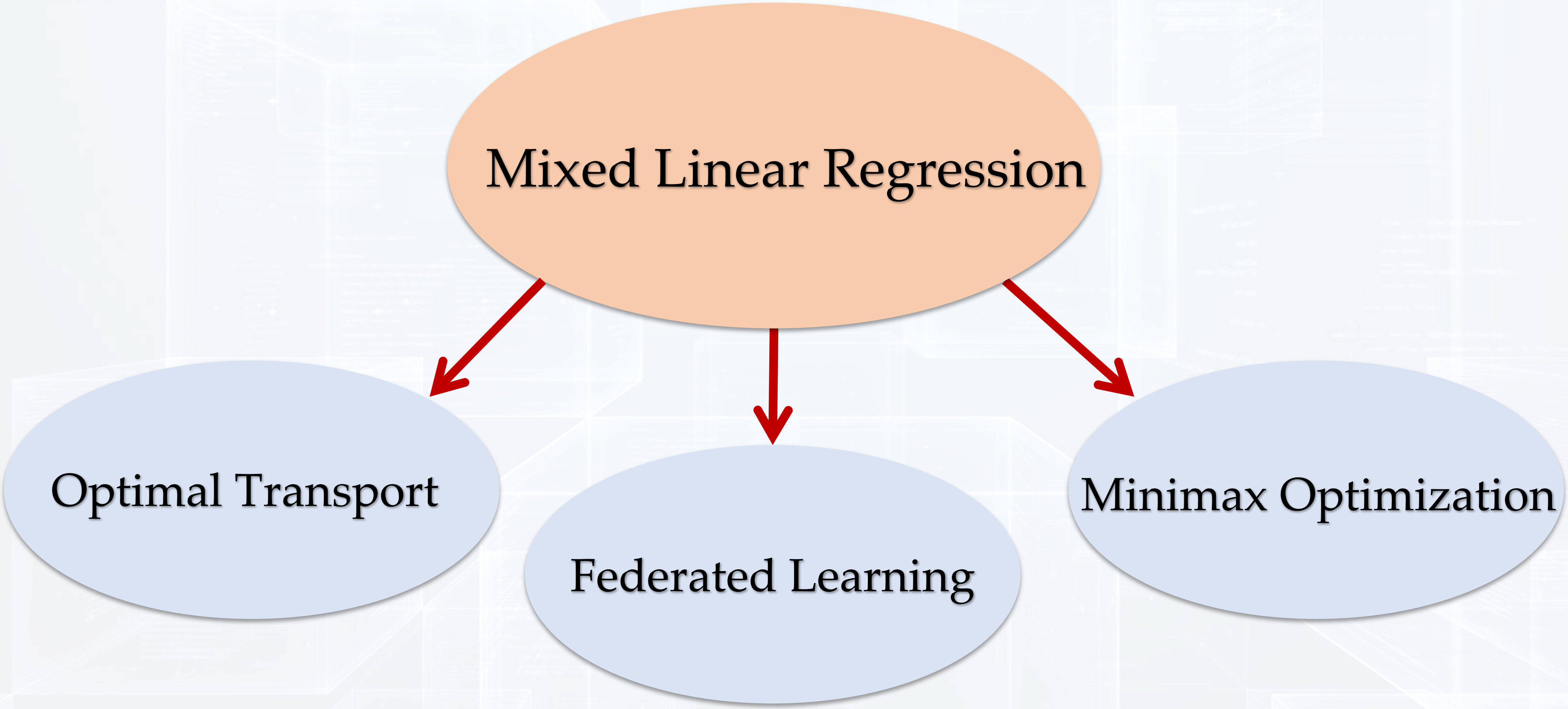


Centralized Setting



Federated Setting

Summary


Mixed Linear Regression
Optimal Transport
Federated Learning
Minimax Optimization