Learning with Shared Representations: Statistical Rates & Optimal Algorithms

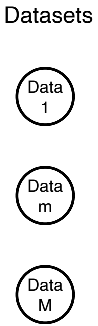
Jiaming Xu

The Fuqua School of Business
Duke University

Joint work with
Xiaochun Niu (Duke), Lili Su (Northeastern), Pengkun Yang (Tsinghua)

JSM, August 2025

Motivating Example: Transfer Learning



Motivating Example: Transfer Learning

Examples: LLMs

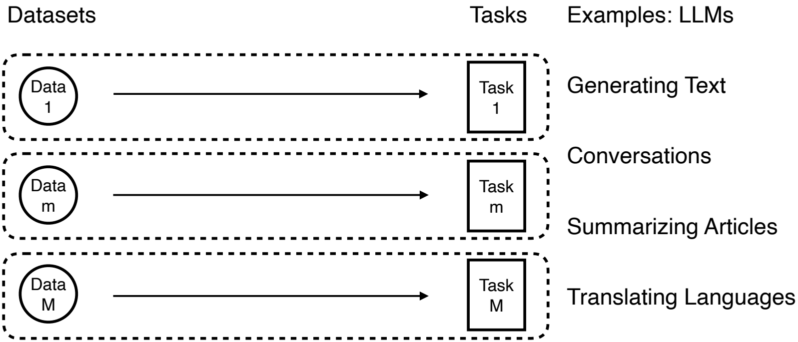
Generating Text

Conversations

Summarizing Articles

Translating Languages

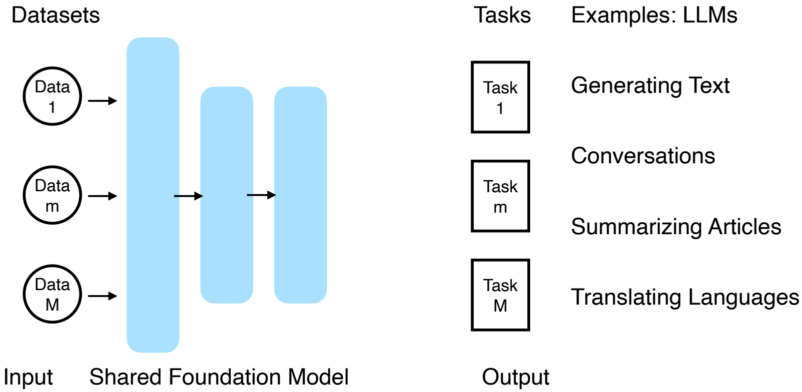
Motivating Example: Transfer Learning



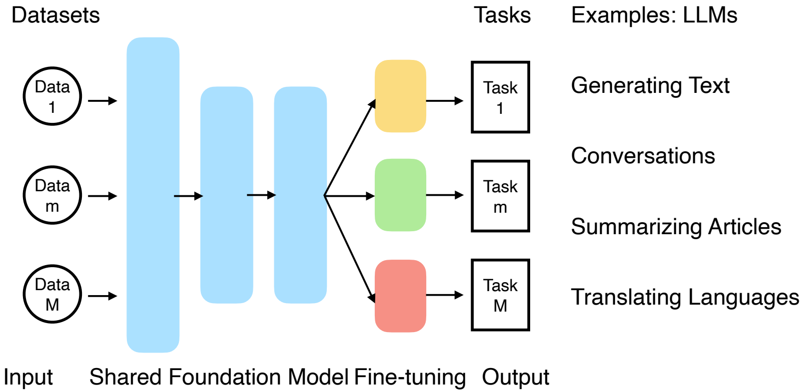
Separately train each task from scratch:

Inefficient; costly; limited task-specific data; less accurate.

Motivating Example: Transfer Learning



Motivating Example: Transfer Learning

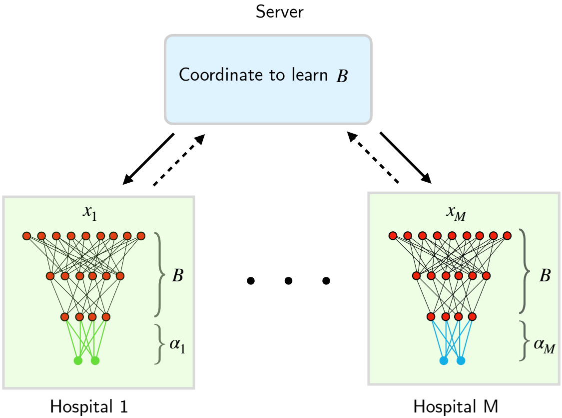


Improve model performance;

Reduce sample complexity;

Goes by many other names: meta-learning, multi-task learning, …

Example: Personalized Federated Learning



[1]

- Tolerate data heterogeneity: small hospitals benefit from large ones
- Achieve model personalization and protect privacy

[1] Figure: Collins et al. (2021)

Model: Learning with Shared Linear Representations

There are M clients, each with n_i data samples (x_{ij}, y_{ij}); $N = \sum_{i=1}^{M} n_i$.

$$y_{ij} = x_{ij}^{\intercal}\theta_i + \xi_{ij}, \qquad x_{ij} \in \mathbb{R}^d,\ y_{ij} \in \mathbb{R},\ j \in [n_i]\ i \in [M].$$

Here $\theta_i \in \mathbb{R}^d$ share a common low-dimensional representation B,

$$\theta_i = B\alpha_i, \qquad B \in \mathcal{O}^{d\times k},\ \alpha_i \in \mathbb{R}^k.$$

Model: Learning with Shared Linear Representations

There are M clients, each with n_i data samples (x_{ij}, y_{ij}); $N = \sum_{i=1}^{M} n_i$.

$$y_{ij} = x_{ij}^{\intercal}\theta_i + \xi_{ij}, \qquad x_{ij} \in \mathbb{R}^d,\ y_{ij} \in \mathbb{R},\ j \in [n_i]\ i \in [M].$$

Here $\theta_i \in \mathbb{R}^d$ share a common low-dimensional representation B,

$$\theta_i = B\alpha_i, \qquad B \in \mathcal{O}^{d\times k},\ \alpha_i \in \mathbb{R}^k.$$

Goal: Collaboratively learn B using datasets $\{(x_{ij}, y_{ij})_{j=1}^{n_i}\}_{i=1}^{M}$.

- Address high-dimensional challenge: $d \gg n_i$
- Tolerate data heterogeneity: different data distributions and sizes
- Estimated B can be further leveraged for (private) fine-tuning
- Can be extended to general non-linear models

Model: Shared Linear Representations

There are M clients, each with n_i data samples:

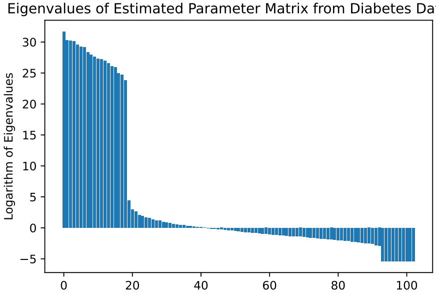
$$y_{ij} = x_{ij}^{\intercal} B \alpha_i + \xi_{ij}, \qquad B \in \mathcal{O}^{d \times k},\ \alpha_i \in \mathbb{R}^k.$$

Model: Shared Linear Representations

There are M clients, each with n_i data samples:

$$y_{ij} = x_{ij}^{\top} B \alpha_i + \xi_{ij}, \qquad B \in \mathcal{O}^{d \times k},\ \alpha_i \in \mathbb{R}^k.$$

Singular values of the estimated parameter matrix $[\widehat{\theta}_1, \cdots, \widehat{\theta}_M]$:

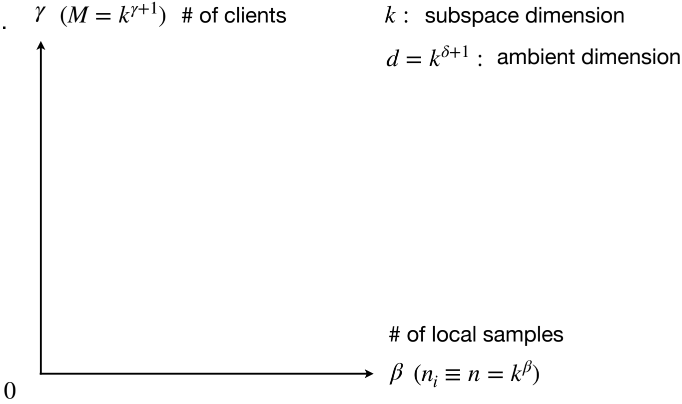


- $M = 102$;
- $d = 180$;
- $k \approx 20$.

Suboptimal Statistical Rates in Existing Works

To estimate B;

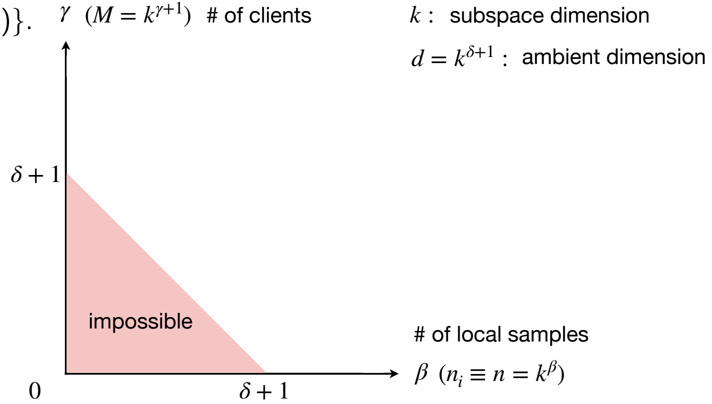
Data: $\{(x_{ij}, y_{ij})\}$.



Suboptimal Statistical Rates in Existing Works

To estimate B;

Data: $\{(x_{ij}, y_{ij})\}$.



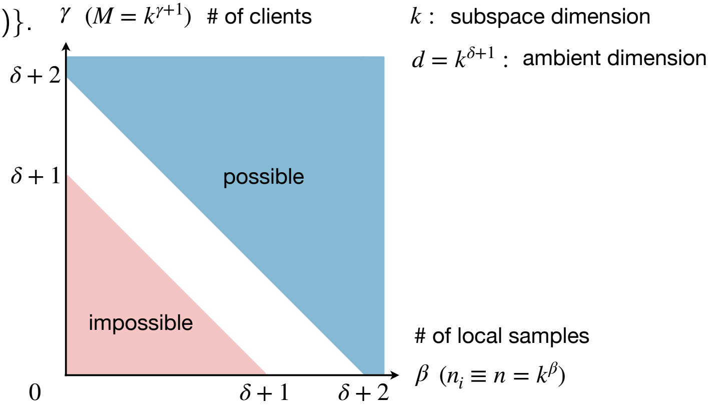
- Minimax lower bound: $\Omega(\sqrt{dk/(Mn)})$. Tripuraneni et al. (2021).

# of unknown parameters — # of total data samples

Suboptimal Statistical Rates in Existing Works

To estimate B;

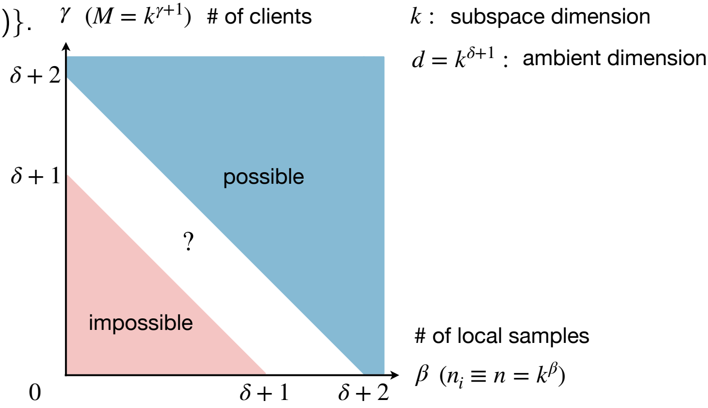
Data: $\{(x_{ij}, y_{ij})\}$.



- Minimax lower bound: $\Omega(\sqrt{dk/(Mn)})$. Tripuraneni et al. (2021).
- Best-known error upper bound: $O(\sqrt{dk^2/(Mn)})$. Tripuraneni et al. (2021); Du et al. (2021); Duchi et al. (2022).

Suboptimal Statistical Rates in Existing Works

To estimate B;
Data: $\{(x_{ij}, y_{ij})\}$.



Open Problem

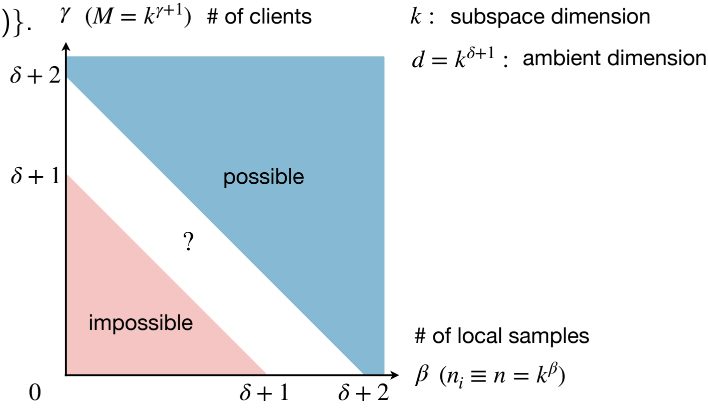
What is the optimal statistical rate to learn B?

Thekumparampil et al. (2021); Thaker et al. (2023); Tian et al. (2023).

Main Contributions

To estimate B;

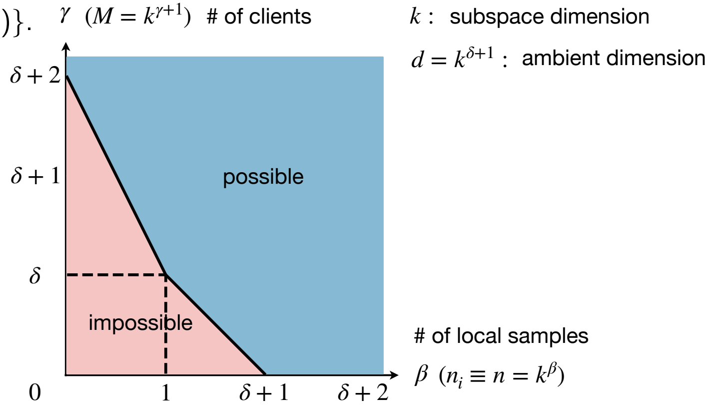
Data: $\{(x_{ij}, y_{ij})\}$.



Main Contributions

To estimate B;

Data: $\{(x_{ij}, y_{ij})\}$.

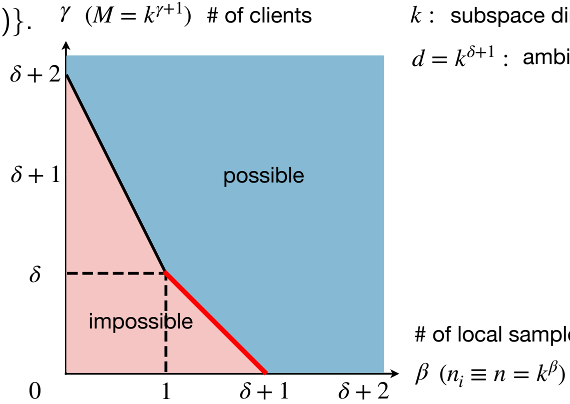


- The identified optimal rate is $\Theta(\sqrt{dk/(Mn)} + \sqrt{dk^2/(Mn^2)})$.

Main Contributions

To estimate B;

Data: $\{(x_{ij}, y_{ij})\}$.



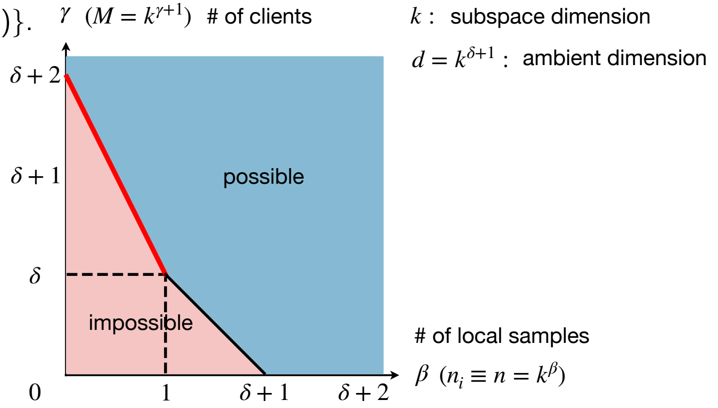
k : subspace dimension

$d = k^{\delta+1}$: ambient dimension

- The identified optimal rate is $\Theta(\sqrt{dk/(Mn)} + \sqrt{dk^2/(Mn^2)})$.
- Two distinct phases

Main Contributions

To estimate B;
Data: $\{(x_{ij}, y_{ij})\}$.

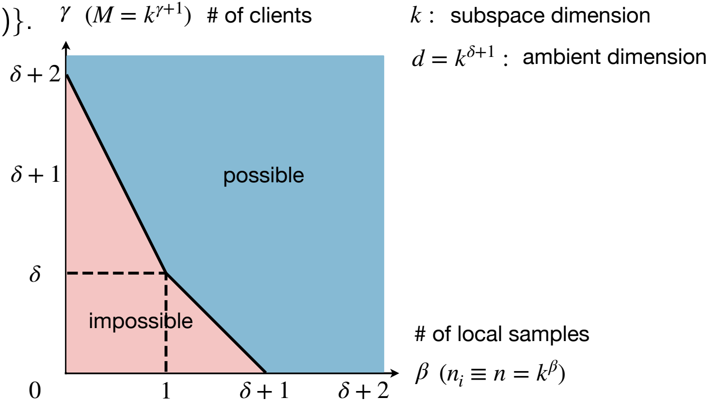


- The identified optimal rate is $\Theta(\sqrt{dk/(Mn)} + \sqrt{dk^2/(Mn^2)})$.
- Two distinct phases: statistical penalty when M is large or n is small.

Main Contributions

To estimate B;

Data: $\{(x_{ij}, y_{ij})\}$.



- The identified optimal rate is $\Theta(\sqrt{dk/(Mn)} + \sqrt{dk^2/(Mn^2)})$.
- Two distinct phases: statistical penalty when M is large or n is small.
- Apply to any small n_i; Some requires $n_i \gg d$ (Du et al., 2021; Duan and Wang, 2023; Tian et al., 2023).

Limitations of Existing Estimators

All the following methods lead to suboptimal error upper bounds.

Limitations of Existing Estimators

All the following methods lead to suboptimal error upper bounds.

- The method-of-moments estimator uses (Tripuraneni et al., 2021),

$$Z_{\mathrm{MoM}} = \sum_{i=1}^{M} \sum_{j=1}^{n_i} y_{ij}^2 x_{ij} x_{ij}^\top .$$

  - The analysis is limited to cases where $x_{ij} \overset{\text{i.i.d.}}{\sim} N(0, I_d)$.

Limitations of Existing Estimators

All the following methods lead to suboptimal error upper bounds.

- The method-of-moments estimator uses (Tripuraneni et al., 2021),

$$Z_{\mathrm{MoM}} = \sum_{i=1}^{M} \sum_{j=1}^{n_i} y_{ij}^2 x_{ij} x_{ij}^\top .$$

  - The analysis is limited to cases where $x_{ij} \overset{\text{i.i.d.}}{\sim} N(0, I_d)$.

- A subsequent estimator uses the matrix (Duchi et al., 2022),

$$Z'_{\mathrm{MoM}} = \sum_{i=1}^{M} \frac{1}{n_i - 1} \sum_{j_1 \neq j_2} y_{ij_1} y_{ij_2} x_{ij_1} x_{ij_2}^\top .$$

  - To handle cases where the noise ξ_{ij} may depend on x_{ij}.

Limitations of Existing Estimators

All the following methods lead to suboptimal error upper bounds.

- The method-of-moments estimator uses (Tripuraneni et al., 2021),

$$Z_{\mathrm{MoM}} = \sum_{i=1}^{M} \sum_{j=1}^{n_i} y_{ij}^2 x_{ij} x_{ij}^\top.$$

  - The analysis is limited to cases where $x_{ij} \overset{\text{i.i.d.}}{\sim} N(0, I_d)$.

- A subsequent estimator uses the matrix (Duchi et al., 2022),

$$Z'_{\mathrm{MoM}} = \sum_{i=1}^{M} \frac{1}{n_i - 1} \sum_{j_1 \neq j_2} y_{ij_1} y_{ij_2} x_{ij_1} x_{ij_2}^\top.$$

  - To handle cases where the noise ξ_{ij} may depend on x_{ij}.

- Alternating minimization methods is studied by Thekumparampil et al. (2021); Collins et al. (2021); Zhang et al. (2024),
  - Initialization via the method-of-moments estimator.

Warm-up Example: Mean Estimation Problems

Each client i observes n_i data sample vectors $u_{ij} \in \mathbb{R}^d$, where

$$u_{ij} = \theta_i + \xi_{ij} = B\alpha_i + \xi_{ij}.$$

Warm-up Example: Mean Estimation Problems

Each client i observes n_i data sample vectors $u_{ij} \in \mathbb{R}^d$, where

$$u_{ij} = \theta_i + \xi_{ij} = B\alpha_i + \xi_{ij}.$$

For the non-convex least squares minimization problem,

$$\min_{B \in \mathcal{O}^{d \times k}, \{\alpha_i\}} \sum_{i=1}^{M} \sum_{j=1}^{n_i} \|u_{ij} - B\alpha_i\|^2,$$

Warm-up Example: Mean Estimation Problems

Each client i observes n_i data sample vectors $u_{ij} \in \mathbb{R}^d$, where

$$u_{ij} = \theta_i + \xi_{ij} = B\alpha_i + \xi_{ij}.$$

For the non-convex least squares minimization problem,

$$\min_{B \in \mathcal{O}^{d \times k}, \{\alpha_i\}} \sum_{i=1}^{M} \sum_{j=1}^{n_i} \|u_{ij} - B\alpha_i\|^2,$$

The optimal solution is formed by the top-k eigenvectors of the matrix

$$\sum_{i=1}^{M} n_i \overline{u}_i \overline{u}_i^\intercal,$$

where $\overline{u}_i = (\sum_{j=1}^{n_i} u_{ij})/n_i$ is the local average at client i.

Our Optimal Estimator

The least squares problem for linear regression is,

$$\min_{B,\{\alpha_i\}} \sum_{i=1}^{M} \sum_{j=1}^{n_i} \left(y_{ij} - x_{ij}^{\intercal} B \alpha_i\right)^2.$$

Let $\widehat{z}_i = (\sum_{j=1}^{n_i} y_{ij} x_{ij})/n_i$. The top-$k$ eigenvectors of $\sum_{i=1}^{M} n_i \widehat{z}_i \widehat{z}_i^{\intercal}$ is an approximated optimal solution.

Our Optimal Estimator

The least squares problem for linear regression is,

$$\min_{B,\{\alpha_i\}} \sum_{i=1}^{M} \sum_{j=1}^{n_i} \left(y_{ij} - x_{ij}^\intercal B\alpha_i\right)^2.$$

Let $\widehat{z}_i = (\sum_{j=1}^{n_i} y_{ij}x_{ij})/n_i$. The top-$k$ eigenvectors of $\sum_{i=1}^{M} n_i \widehat{z}_i \widehat{z}_i^\intercal$ is an approximated optimal solution. However,

$$\mathbb{E}\Big[\sum_{i=1}^{M} n_i \widehat{z}_i \widehat{z}_i^\intercal\Big] = B\Big(\sum_{i=1}^{M}(n_i - 1)\alpha_i(\alpha_i)^\intercal\Big)(B)^\intercal + \sum_{i=1}^{M}\mathbb{E}[\xi_{ij}^2]I_d$$
$$+ \sum_{i=1}^{M}\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbb{E}[x_{ij}^\intercal\theta_i(\theta_i)^\intercal x_{ij}x_{ij}x_{ij}^\intercal].$$

Our Optimal Estimator

The least squares problem for linear regression is,

$$\min_{B,\{\alpha_i\}} \sum_{i=1}^{M} \sum_{j=1}^{n_i} \left(y_{ij} - x_{ij}^{\intercal} B \alpha_i\right)^2.$$

Let $\widehat{z}_i = (\sum_{j=1}^{n_i} y_{ij} x_{ij})/n_i$. The top-$k$ eigenvectors of $\sum_{i=1}^{M} n_i \widehat{z}_i \widehat{z}_i^{\intercal}$ is an approximated optimal solution. However,

$$\mathbb{E}\Big[\sum_{i=1}^{M} n_i \widehat{z}_i \widehat{z}_i^{\intercal}\Big] = B\Big(\sum_{i=1}^{M} (n_i - 1)\alpha_i(\alpha_i)^{\intercal}\Big)(B)^{\intercal} + \sum_{i=1}^{M} \mathbb{E}[\xi_{ij}^2] I_d$$

$$+ \sum_{i=1}^{M} \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbb{E}[x_{ij}^{\intercal} \theta_i (\theta_i)^{\intercal} x_{ij} x_{ij} x_{ij}^{\intercal}].$$

Unknown fourth-order moments

Our Optimal Estimator

Two independent replicas of local averages $\overline{z}_i = (2/n_i) \cdot \sum_{j=1}^{n_i/2} y_{ij} x_{ij}$ and $\widetilde{z}_i = (2/n_i) \cdot \sum_{j=n_i/2+1}^{n_i} y_{ij} x_{ij}$.

Our estimator $\widehat{B}$ is given by the top-k singular vectors of the matrix

$$Z = \sum_{i=1}^{M} n_i \overline{z}_i \widetilde{z}_i^{\mathsf{T}}.$$

Similar replica ideas appeared in related mixed linear regression problems (Kong et al., 2020; Su et al., 2024).

Our Optimal Estimator

Two independent replicas of local averages $\overline{z}_i = (2/n_i) \cdot \sum_{j=1}^{n_i/2} y_{ij} x_{ij}$ and $\widetilde{z}_i = (2/n_i) \cdot \sum_{j=n_i/2+1}^{n_i} y_{ij} x_{ij}$.

Our estimator $\widehat{B}$ is given by the top-k singular vectors of the matrix

$$Z = \sum_{i=1}^{M} n_i \overline{z}_i \widetilde{z}_i^{\mathsf{T}}.$$

Similar replica ideas appeared in related mixed linear regression problems (Kong et al., 2020; Su et al., 2024). In expectation,

$$\mathbb{E} Z = B \Big(\sum_{i=1}^{M} n_i \alpha_i \alpha_i^{\mathsf{T}} \Big) B^{\mathsf{T}}.$$

Our Optimal Estimator

Two independent replicas of local averages $\overline{z}_i = (2/n_i) \cdot \sum_{j=1}^{n_i/2} y_{ij} x_{ij}$ and $\widetilde{z}_i = (2/n_i) \cdot \sum_{j=n_i/2+1}^{n_i} y_{ij} x_{ij}$.

Our estimator $\widehat{B}$ is given by the top-k singular vectors of the matrix

$$Z = \sum_{i=1}^{M} n_i \overline{z}_i \widetilde{z}_i^{\mathsf{T}}.$$

Similar replica ideas appeared in related mixed linear regression problems (Kong et al., 2020; Su et al., 2024). In expectation,

$$\mathbb{E}Z = B \left(\sum_{i=1}^{M} n_i \alpha_i \alpha_i^{\mathsf{T}} \right) B^{\mathsf{T}}.$$

- Local averaging: reduces noises. Recall $Z_{\mathrm{MoM}} = \sum_{i=1}^{M} \sum_{j=1}^{n_i} y_{ij}^2 x_{ij} x_{ij}^{\mathsf{T}}$

Our Optimal Estimator

Two independent replicas of local averages $\overline{z}_i = (2/n_i) \cdot \sum_{j=1}^{n_i/2} y_{ij} x_{ij}$ and $\widetilde{z}_i = (2/n_i) \cdot \sum_{j=n_i/2+1}^{n_i} y_{ij} x_{ij}$.

Our estimator $\widehat{B}$ is given by the top-k singular vectors of the matrix

$$Z = \sum_{i=1}^{M} n_i \overline{z}_i \widetilde{z}_i^\intercal.$$

Similar replica ideas appeared in related mixed linear regression problems (Kong et al., 2020; Su et al., 2024). In expectation,

$$\mathbb{E} Z = B \Big(\sum_{i=1}^{M} n_i \alpha_i \alpha_i^\intercal \Big) B^\intercal.$$

- Local averaging: reduces noises. Recall $Z_{\mathrm{MoM}} = \sum_{i=1}^{M} \sum_{j=1}^{n_i} y_{ij}^2 x_{ij} x_{ij}^\intercal$
- Sending only local averages but not raw data: to preserve privacy.

Main Results: Key Factor for Learnability of B

Dataset at client i provides information about B along direction α_i

$$y_{ij} = x_{ij}^{\intercal} B \alpha_i + \xi_{ij}, \quad j \in [n_i]$$

Main Results: Key Factor for Learnability of B

Dataset at client i provides information about B along direction α_i

$$y_{ij} = x_{ij}^\intercal B \alpha_i + \xi_{ij}, \quad j \in [n_i]$$

The learnability of B is governed by the *client diversity matrix*:

$$\frac{1}{N} \sum_{i=1}^{M} n_i \alpha_i \alpha_i^\intercal \in \mathbb{R}^{k \times k}$$

with the largest eigenvalue λ_1 and the smallest eigenvalue λ_k.

Main Results: Key Factor for Learnability of B

Dataset at client i provides information about B along direction α_i

$$y_{ij} = x_{ij}^\intercal B \alpha_i + \xi_{ij}, \quad j \in [n_i]$$

The learnability of B is governed by the *client diversity matrix*:

$$\frac{1}{N} \sum_{i=1}^{M} n_i \alpha_i \alpha_i^\intercal \in \mathbb{R}^{k \times k}$$

with the largest eigenvalue λ_1 and the smallest eigenvalue λ_k.

B is well-represented if the condition number $\lambda_1 / \lambda_k = \Theta(1)$, which is satisfied when α_i's are spread out and local dataset sizes $\{n_i\}$ are not too unbalanced

Main Results

d: ambient dimension; M: # of clients; N: # of total data samples; λ_1 (λ_k): largest (smallest) eigenvalues of the client diversity matrix.

Theorem (Error Upper Bound)

For our spectral estimator with local averaging $\widehat{B}$, with high probability,

$$\|\sin\Theta(\widehat{B}, B)\| = \widetilde{O}\left(\left(\sqrt{\frac{d\lambda_1}{N\lambda_k^2}} + \sqrt{\frac{Md}{N^2\lambda_k^2}}\right) \wedge 1\right).$$

Main Results

d: ambient dimension; M: # of clients; N: # of total data samples; λ_1 (λ_k): largest (smallest) eigenvalues of the client diversity matrix.

Theorem (Error Upper Bound)

For our spectral estimator with local averaging $\widehat{B}$, with high probability,

$$\|\sin\Theta(\widehat{B}, B)\| = \widetilde{O}\left(\left(\sqrt{\frac{d\lambda_1}{N\lambda_k^2}} + \sqrt{\frac{Md}{N^2\lambda_k^2}}\right) \wedge 1\right).$$

Theorem (Minimax Lower Bound)

$$\inf_{\widehat{B}\in\mathcal{O}^{d\times k}} \sup_{B\in\mathcal{O}^{d\times k},\alpha,\vec{n}} \mathbb{E}[\|\sin\Theta(\widehat{B}, B)\|] = \Omega\left(\left(\sqrt{\frac{d}{N\lambda_k}} + \sqrt{\frac{Md}{N^2\lambda_k^2}}\right) \wedge 1\right).$$

Main Results

d: ambient dimension; M: # of clients; N: # of total data samples; λ_1 (λ_k): largest (smallest) eigenvalues of the client diversity matrix.

Theorem (Error Upper Bound)

For our spectral estimator with local averaging $\widehat{B}$, with high probability,

$$\|\sin\Theta(\widehat{B}, B)\| = \widetilde{O}\left(\left(\sqrt{\frac{d\lambda_1}{N\lambda_k^2}} + \sqrt{\frac{Md}{N^2\lambda_k^2}}\right) \wedge 1\right).$$

Theorem (Minimax Lower Bound)

$$\inf_{\widehat{B}\in\mathcal{O}^{d\times k}} \sup_{B\in\mathcal{O}^{d\times k},\alpha,\vec{n}} \mathbb{E}[\|\sin\Theta(\widehat{B}, B)\|] = \Omega\left(\left(\sqrt{\frac{d}{N\lambda_k}} + \sqrt{\frac{Md}{N^2\lambda_k^2}}\right) \wedge 1\right).$$

- If $\lambda_1 = \Theta(\lambda_k) = \Theta(1/k)$, we have $\widetilde{\Theta}(\sqrt{dk/N} + \sqrt{Mdk^2/N^2})$.

Minimax Lower Bound

	Our lower bound	Existing bound (Tripuraneni et al., 2021)
Rate	$\Omega\left(\sqrt{\frac{d}{N\lambda_k}} + \sqrt{\frac{Md}{N^2\lambda_k^2}}\right)$	$\Omega\left(\sqrt{\frac{1}{N\lambda_k}} + \sqrt{\frac{dk}{N}}\right)$
First term (deterministic α_i)	Packing set & Mutual information bound	Le Cam's two-point method
Second term	Gaussian-generated α_i	N/A

Applications: Fine Tuning for New Clients

At a new client $M+1$ with n_{M+1} data points and $\theta_{M+1} = B\alpha_{M+1}$.

Given a fixed $\widehat{B}$, learn α_{M+1} via regression on projected covariates:

$$\widehat{\alpha}_{M+1} = \underset{\alpha_{M+1}}{\operatorname{argmin}} \sum_{j=1}^{n_{M+1}} \left\| y_{M+1,j} - x_{M+1,j}^{\intercal} \widehat{B} \alpha_{M+1} \right\|^2.$$

Applications: Fine Tuning for New Clients

At a new client $M+1$ with n_{M+1} data points and $\theta_{M+1} = B\alpha_{M+1}$.

Given a fixed $\widehat{B}$, learn α_{M+1} via regression on projected covariates:

$$\widehat{\alpha}_{M+1} = \underset{\alpha_{M+1}}{\text{argmin}} \sum_{j=1}^{n_{M+1}} \left\| y_{M+1,j} - x_{M+1,j}^{\mathsf{T}} \widehat{B} \alpha_{M+1} \right\|^2.$$

Corollary (Fine-tuning)

For our estimator $\widehat{B}$ and $\widehat{\alpha}_{M+1}$, with high probability,

$$\left\| \widehat{B}\widehat{\alpha}_{M+1} - B\alpha_{M+1} \right\|^2 = \widetilde{O}\left(\frac{dk}{N} + \frac{Mdk^2}{N^2} + \frac{k}{n_{M+1}} \right).$$

Applications: Fine Tuning for New Clients

At a new client $M+1$ with n_{M+1} data points and $\theta_{M+1} = B\alpha_{M+1}$.

Given a fixed $\widehat{B}$, learn α_{M+1} via regression on projected covariates:

$$\widehat{\alpha}_{M+1} = \underset{\alpha_{M+1}}{\operatorname{argmin}} \sum_{j=1}^{n_{M+1}} \left\| y_{M+1,j} - x_{M+1,j}^{\mathsf{T}} \widehat{B} \alpha_{M+1} \right\|^2.$$

Corollary (Fine-tuning)

For our estimator $\widehat{B}$ and $\widehat{\alpha}_{M+1}$, with high probability,

$$\left\| \widehat{B}\widehat{\alpha}_{M+1} - B\alpha_{M+1} \right\|^2 = \widetilde{O}\left(\frac{dk}{N} + \frac{Mdk^2}{N^2} + \frac{k}{n_{M+1}} \right).$$

Can be also applied to private fine-tuning for new clients Thaker et al. (2023).

Numerical Experiments: Diabetes Dataset

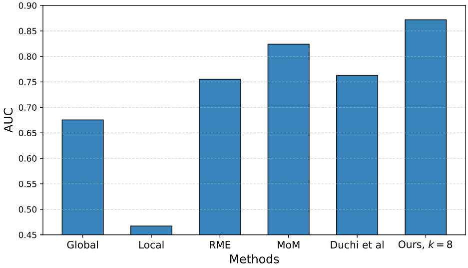


Figure: Hospital A.

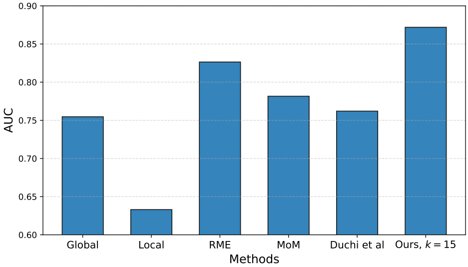
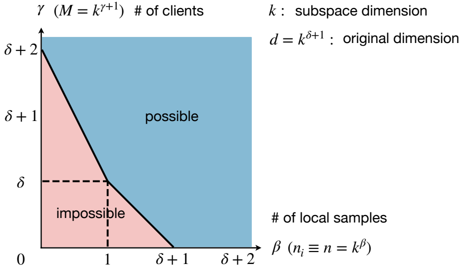


Figure: Hospital B.

Concluding Remarks

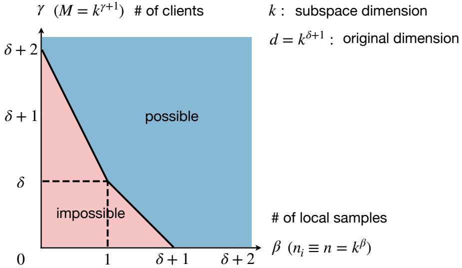
Main contributions:



- ✓ Designed a spectral estimator with local averaging.
- ✓ Extensions to general non-linear models: $\mathbb{E}[y_{ij} \mid x_{ij}] = f_i(B^\top x_{ij})$

Concluding Remarks

Main contributions:



✓ Designed a spectral estimator with local averaging.

✓ Extensions to general non-linear models: $\mathbb{E}[y_{ij} \mid x_{ij}] = f_i(B^\top x_{ij})$

Future directions:

- Non-identical representations B_i? Tian et al. (2023); Duan and Wang (2023)
- In-context learning

Niu, X., Su, L., Xu, J., & Yang, P. (2024). Learning with Shared Representations: Statistical Rates and Optimal Algorithms. arXiv:2409.04919.

Collins, L., Hassani, H., Mokhtari, A., and Shakkottai, S. (2021). Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pages 2089–2099. PMLR.

Duan, Y. and Wang, K. (2023). Adaptive and robust multi-task learning. *The Annals of Statistics*, 51(5):2015–2039.

Duchi, J. C., Feldman, V., Hu, L., and Talwar, K. (2022). Subspace recovery from heterogeneous data with non-isotropic noise. *Advances in Neural Information Processing Systems*, 35:5854–5866.

Kong, W., Somani, R., Song, Z., Kakade, S., and Oh, S. (2020). Meta-learning for mixed linear regression. In *International Conference on Machine Learning*, pages 5394–5404. PMLR.

Su, L., Xu, J., and Yang, P. (2024). Global convergence of federated learning for mixed regression. *IEEE Transactions on Information Theory*.

Thaker, P., Setlur, A., Wu, Z. S., and Smith, V. (2023). On the benefits of public representations for private transfer learning under distribution shift. *arXiv preprint arXiv:2312.15551*.

Thekumparampil, K. K., Jain, P., Netrapalli, P., and Oh, S. (2021). Statistically and computationally efficient linear meta-representation learning. *Advances in Neural Information Processing Systems*, 34:18487–18500.

Tian, Y., Gu, Y., and Feng, Y. (2023). Learning from similar linear representations: Adaptivity, minimaxity, and robustness. *arXiv preprint arXiv:2303.17765*.

Tripuraneni, N., Jin, C., and Jordan, M. (2021). Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pages 10434–10443. PMLR.

Zhang, T. T., Toso, L. F., Anderson, J., and Matni, N. (2024). Sample-efficient linear representation learning from non-iid non-isotropic data. In *The Twelfth International Conference on Learning Representations*.