# Robust Meta-learning for Mixed Linear Regression with Small Batches

Weihao Kong[1] Raghav Somani[1] Sham M. Kakade[1,2] Sewoong Oh[1]

[1]University of Washington & [2]Microsoft Research

NEURAL INFORMATION PROCESSING SYSTEMS

## Problem Setup

**Meta Learning**

Large number of heterogeneous users/sources.


Each provides a modest amount of data, insufficient to build a reliable model.

Question: *How to improve the estimate of each user's model by integrating the other users' data in the training process?*

**Mixed Linear Regression (MLR)**

- Meta-training dataset: $n$ tasks with $i$-th task with data $\{(\mathbf{x}_{i,j}, y_{i,j}) \in \mathbb{R}^d \times \mathbb{R}\}_{j=1}^{t_i}$.
- Each task comes from a **linear model**, parameterized by $(\beta_i, s_i) \in \mathbb{R}^d \times \mathbb{R}_+$:

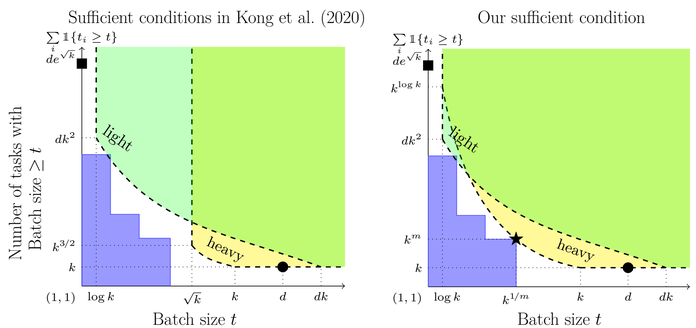
$$y_{i,j} = \langle \beta_i, \mathbf{x}_{i,j} \rangle + \epsilon_{i,j} \qquad \text{where} \quad \mathbf{x}_{i,j} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d),\ \epsilon_{i,j} \sim \mathcal{N}\big(0, s_i^2\big) \quad \forall\, j \in [t_i].$$

- **Mixture** of linear models: each linear model $\beta_i$ is drawn from a discrete set $\{(\mathbf{w}_\ell, s_\ell)\}_{\ell=1}^k$ from a multinomial distribution with $\mathbf{p} = [p_1, \dots, p_k]$, which are unknown.
- Goal: Train a model for a new arriving task from a small number of labelled examples.

## Comparisons to Existing Results

| Known results for MLR | # Samples $n$ |
|---|---|
| Chaganty and Liang (2013) | $d^6 \cdot \mathrm{poly}(k, 1/\sigma_k)$ |
| Yi et al. (2016) | $d \cdot \mathrm{poly}(k, 1/\sigma_k)$ |
| Zhong et al. (2016) | $d \cdot \exp(k \log(k \log d))$ |
| Li and Liang (2018) | $d \cdot \mathrm{poly}(k) + \exp(k^2 \log k)$ |
| Chen et al. (2020) | $d \cdot e^{\sqrt{k}}$ |

■ When $t = 1$, no algorithm can learn with $o\big(de^{\sqrt{k}}\big)$ samples in total.
● When $t \gtrsim d$, each task can be learned in isolation.




light tasks: $t_L \gtrsim \log k$ $\quad (\widetilde{\Omega}(dk^2)$ samples total)

heavy tasks: $t_H \gtrsim \sqrt{k}$ $\quad (\widetilde{\Omega}(k^2)$ samples total)

## Main Results

To account for heterogeneity, we consider Meta-training datasets consisting of $n_L$ light tasks each with at least $t_L$ samples, and $n_H$ heavy tasks each with $t_H$ samples.

**With no adversarial corruption**

For any $m \in \mathbb{N}$, there exists a polynomial time algorithm such that if

● $t_L \gtrsim 1$, ● $n_L t_L \gtrsim dk^2/\epsilon^2$, ● $t_H \gtrsim mk^{1/m}$, and ● $n_H t_H \gtrsim k^{\Theta(m)}$,

estimates $\{(\mathbf{w}_\ell, s_\ell, p_\ell)\}_{\ell=1}^k$ up to any desired accuracy of $\epsilon$ with high probability.

- When $m = 2$, we only use the second moment recovering the result of Kong et al. (2020).
- With larger $m$, using **$m$-th order moments** reduces $t_H$ at the cost of larger $t_H n_H$.
- Minimum batch size of $t_H \gtrsim \log k$ is achieved with $m = \log k$, requiring $n_H \gtrsim k^{\Theta(\log k)}$ (quasi-polynomial)
- We obtain a smooth trade-off between batch size $t_H$ and $n_H$, avoiding the prohibitive large requirement on $n_H$, i.e., $n_H = \exp\big(\Omega(\sqrt{k})\big)$, in the classical MLR setting ($t_H = 1$).

## Adversarial Setup

An adversary is allowed to inspect all the tasks in Meta-training dataset, remove examples associated with a subset of tasks, and replace the examples associated with those tasks with arbitrary points. Fraction of tasks corrupted is at most $\alpha$.

**When data is adversarially corrupted**

For any $\epsilon \in (0, 1/k^3)$, and $m \in \mathbb{N}$, there exists a polynomial time algorithm such that if

● $t_L \gtrsim 1$, ● $n_L t_L \gtrsim dk/\epsilon^2$, ● $t_H \gtrsim mk^{1/m}$, ● $n_H t_H \gtrsim k^{O(m)}$,

and $\alpha \lesssim \epsilon/k$, estimates $\{(\mathbf{w}_\ell, s_\ell, p_\ell)\}_{\ell=1}^k$ up to $\epsilon$-accuracy with high probability.

- This matches the fundamental limit of $\epsilon \geq k\alpha + dk/(n_L t_L + n_H t_H)$

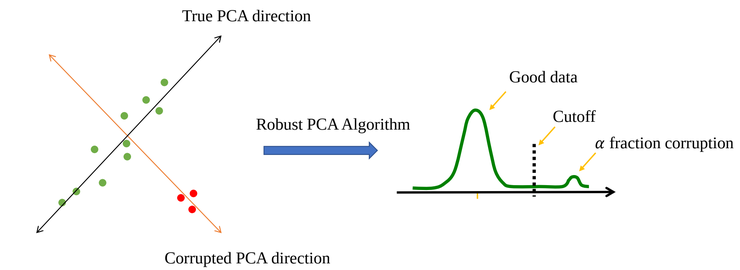
## Algorithms

1. **Robust Subspace Estimation**
   - Estimate an orthonormal basis $\mathbf{U} \in \mathbb{R}^{d \times k}$ of the rank-$k$ subspace span$\{\mathbf{w}_1, \dots, \mathbf{w}_k\}$.
   - Project the data in $d$ dimensions to $k$ ($\ll d$) dimensions using the estimated $\mathbf{U}$.
   - We propose an outlier-robust Principal Component Analysis algorithm to estimate $\mathbf{U}$.
2. **Robust Clustering**
   - Since each task has limited data points, we cluster tasks sharing the same regressor $\mathbf{w}_i$.
   - We leveraged a recently developed robust clustering algorithm Kothari et al. (2018) based on Sum-of-squares proofs to obtain a trade-off between the number of tasks $n_H$ and the batch size per task $t_H$.
3. **Robust Regression**
   - Once the tasks are clustered. We invoke standard robust linear regression algorithms (e.g., Diakonikolas et al. (2019)) to learn $\{\mathbf{w}_1, \dots, \mathbf{w}_k\}$.

## Robust Subspace Estimation

**Algorithm**: Alternate between

1. Computing the top-$k$ subspace.
2. Removing data points with abnormally large projection under the computed $k$ subspace.




**Result**: If $\alpha$-fraction of points sampled from $\mathcal{P}$ are corrupted, and $\mathcal{P}$ satisfies $\mathbb{E}_{\mathbf{x}\sim\mathcal{P}}[\mathbf{x}\mathbf{x}^\top] = \mathbf{\Sigma}$, and has a bounded support and bounded $4^{\text{th}}$ moment by $\nu^2$, then (Kong et al., 2020, Algorithm 2) returns a basis $\widehat{\mathbf{U}} \in \mathbb{R}^{d \times k}$ that with high probability satisfies

$$\mathrm{Tr}\Big[\widehat{\mathbf{U}}^\top \mathbf{\Sigma} \widehat{\mathbf{U}}\Big] \;\geq\; (1 - \mathcal{O}(\alpha)) \cdot \mathrm{Tr}[\mathcal{P}_k(\mathbf{\Sigma})] - \mathcal{O}\Big(\sqrt{\alpha} \cdot \nu\sqrt{k}\Big),$$

where $\mathcal{P}_k$ is the best rank-$k$ $\ell_2$-projection operator.

- We show that the term $\sqrt{\alpha} \cdot \nu\sqrt{k}$ cannot be improved as an information theoretic limit.

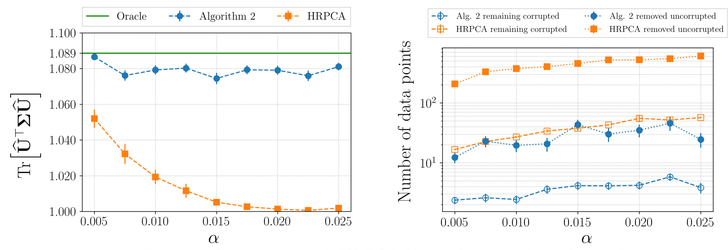



Figure 1: Comparison with HRPCA Xu et al. (2012) for a chosen $\mathcal{P}$.

## Robust Clustering

**Sum-of-Squares (SOS)**

- To efficiently exploit the higher order moments assumption for clustering, one need stronger condition than boundedness, i.e., the moments are SOS bounded, meaning there exist SOS proofs showing that the moments are bounded.
- We show that higher order moments of regression estimates are SOS bounded, allowing us to apply the result of Kothari et al. (2018) even in the presence of adversarial corruption.

## References

Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019.

Weihao Kong, Raghav Somani, Sham Kakade, and Sewoong Oh. Robust meta-learning for mixed linear regression with small batches. *arXiv preprint arXiv:2006.09702*, 2020.

Weihao Kong, Raghav Somani, Zhao Song, Sham Kakade, and Sewoong Oh. Meta-learning for mixed linear regression. *arXiv e-prints*, art. arXiv:2002.08936, February 2020.

Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046, 2018.

Huan Xu, Constantine Caramanis, and Shie Mannor. Outlier-robust pca: The high-dimensional case. *IEEE transactions on information theory*, 59(1):546–572, 2012.

$\lesssim$ and $\gtrsim$ are inequalities up to constants & all polylog factors.