Lecture 4: Posterior Drift and Biased Regularization

Yang Feng[1], Ye Tian[2]

[1]Department of Biostatistics, School of Global Public Health, New York University
[2]Department of Statistics, Columbia University

May 20, 2024

The slides were initially used in a full-day short course "A (Selective) Introduction to the Statistical Foundations of Transfer Learning" at the 37th New England Statistics Symposium (NESS 2024), University of Connecticut. All the slides are publicly available at http://www.columbia.edu/~yt2661/STL.html and free to use in the educational context. The source should be mentioned.

Overview I

1. §4.1: Posterior drift and concept drift
2. §4.2: Biased regularization
   - §4.2.1: Motivation
   - §4.2.2: Ridge penalty
   - §4.2.3: An adaptive ℓ_2-penalty
3. §4.3: Extension to high-dimensional regressions
   - §4.3.1: ℓ_1-penalty with GLMs
   - §4.3.2: Go robust: better ways to aggregate data
   - §4.3.3: Block penalty with multi-task learning
4. §4.4: Other applications of biased regularization in supervised learning
   - §4.4.1: High-dimensional quantile regression
   - §4.4.2: Functional regression
   - §4.4.3: Causal inference
5. §4.5: Statistical inference
   - §4.5.1: Challenges in doing inference
   - §4.5.2: Inference in high-dimensional linear regression
   - §4.5.3: Inference of the prevailing model

Overview II

6. §4.6: Unsupervised multi-task learning

7. §4.7: Domain adaptation with representation learning
   - §4.7.1: Representation learning and fine tuning
   - §4.7.2: Learning the shared representation across tasks
   - §4.7.3: Go beyond the shared representation

8. References

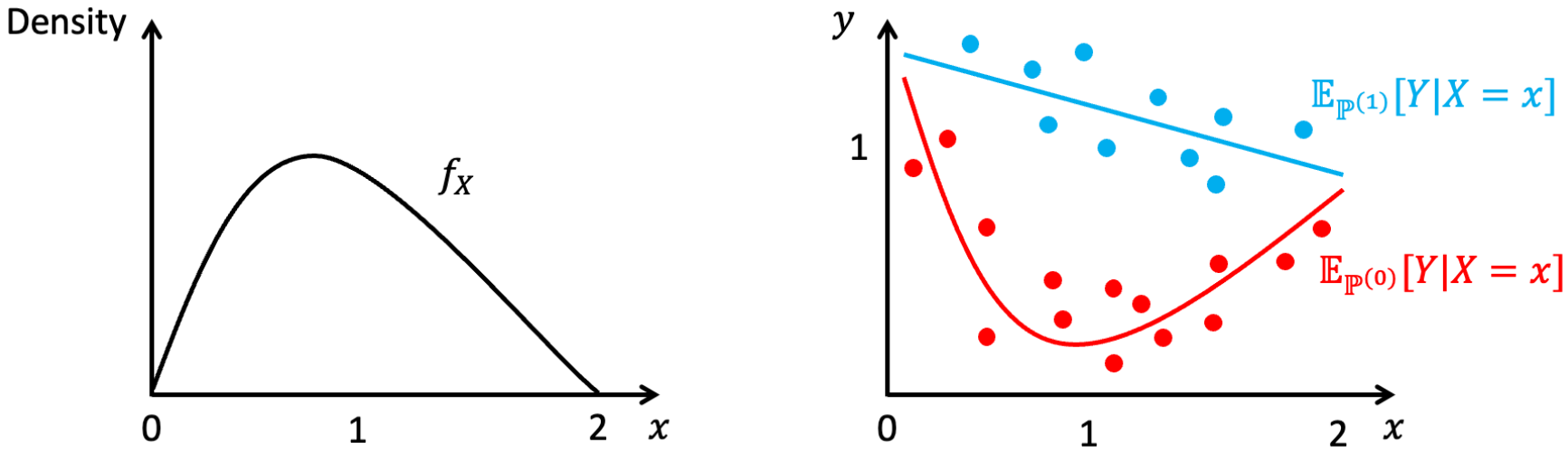
§4.1: Posterior drift and concept drift

Posterior drift and concept drift

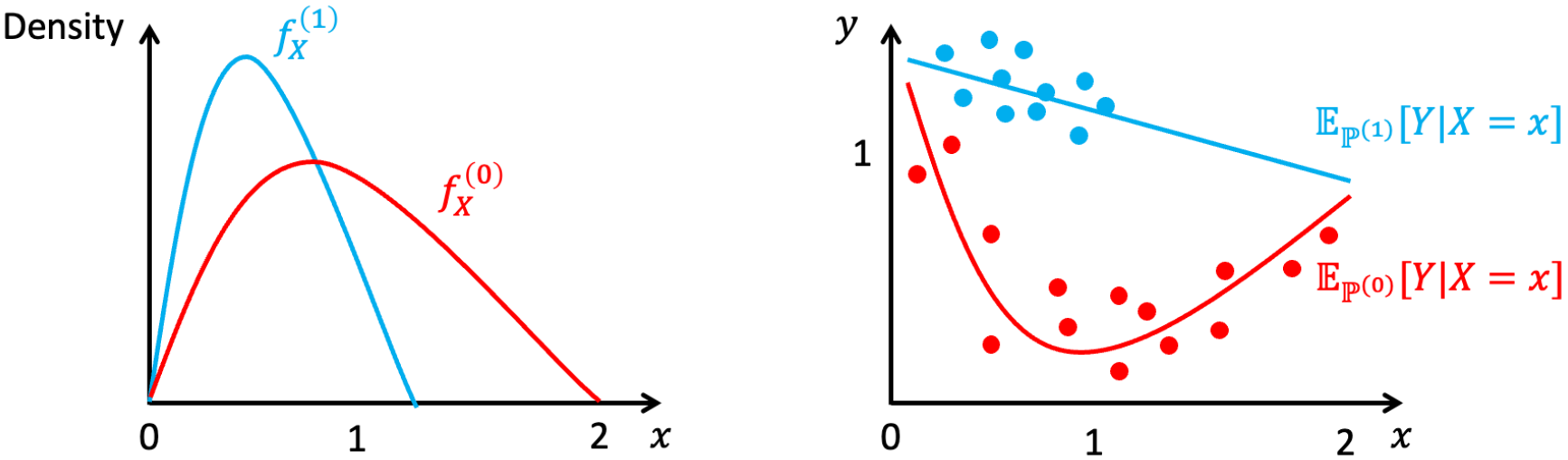
For simplicity, we consider the target and one single source in this section.

- Target distribution $X^{(0)} \sim \mathbb{P}_X^{(0)}$, $Y^{(0)}|X^{(0)} \sim \mathbb{P}_{Y|X}^{(0)}$
- Source distribution $X^{(1)} \sim \mathbb{P}_X^{(1)}$, $Y^{(1)}|X^{(1)} \sim \mathbb{P}_{Y|X}^{(1)}$
- **Posterior drift**: $\mathbb{P}_X^{(0)} = \mathbb{P}_X^{(1)}$, $\mathbb{P}_{Y|X}^{(0)} \neq \mathbb{P}_{Y|X}^{(1)}$
- **Concept drift**: $\mathbb{P}_X^{(0)} \neq \mathbb{P}_X^{(1)}$, $\mathbb{P}_{Y|X}^{(0)} \neq \mathbb{P}_{Y|X}^{(1)}$
- **Goal:** learn $\mathbb{E}_{\mathbb{P}^{(1)}}[Y|X = x]$ or make prediction on target domain

Posterior drift v.s. concept drift



"Posterior drift"



"Concept drift"

Posterior drift

Liu et al. (2023) pointed out that posterior drift can be more common in tabular data [1] compared to the covariate shift, due to the missing variables and hidden confounders.

They run different methods on 5 real tabular datasets with different source-target pairs. Out of 169 source-target pairs with significant performance degradation, 80% of them are primarily attributed to posterior drift. The evaluation is done via the diagnosis tool proposed by Cai et al. (2023).

#ID	Dataset	Type	#Samples	#Features	Outcome	#Domains	Selected Settings	Shift Patterns
1	`ACS Income`	Natural	1,599,229	9	Income≥50k	51	California → Puerto Rico	$Y\|X \gg X$
2	`ACS Mobility`	Natural	620,937	21	Residential Address	51	Mississippi → Hawaii	$Y\|X \gg X$
3	`Taxi`	Natural	1,506,769	7	Duration time≥30 min	4	New York City→ Botogá	$Y\|X \gg X$
4	`ACS Pub.Cov`	Natural	1,127,446	18	Public Ins. Coverage	51	Nebraska → Louisiana	$Y\|X > X$
5	`US Accident`	Natural	297,132	47	Severity of Accident	14	California→ Oregon	$Y\|X > X$
6	`ACS Pub.Cov`	Natural	859,632	18	Public Ins. Coverage	4	2010 (NY)→ 2017 (NY)	$Y\|X < X$
7	`ACS Income`	Synthetic	195,665	9	Income≥50k	2	Younger→ Older	$Y\|X \ll X$

[1] Tabular data refers to data organized in a table/data frame.

[1] Liu, J., Wang, T., Cui, P., & Namkoong, H. (2023). On the need for a language describing distribution shifts: Illustrations on tabular datasets. Advances in Neural Information Processing Systems, 36.

[2] Cai, T. T., Namkoong, H., & Yadlowsky, S. (2023). Diagnosing model performance under distribution shift. arXiv preprint arXiv:2303.02011.

Posterior drift

- We will mostly focus on posterior drift in well-specified parametric models
- As we mentioned in the Section 3.1, for well-specified parametric models with appropriate curvature, MLE on source domain can adapt to covariate shift **for free**
- Therefore, for many problems we will discuss in this section:

$$\text{taking care of } \textbf{posterior drift} \quad \Rightarrow \quad \text{taking care of } \textbf{concept drift}$$

Posterior drift: what we need

Recall that for **covariate shift**:

- Full (X, Y) data from both the source and the target?
  ✓
- Full (X, Y) data from the source, only X from the target?
  ✓ (in many cases)
- Full (X, Y) data from the source, no data from the target?
  ✗ (in general), possible with domain generalization

For **posterior drift**:

- Full (X, Y) data from both the source and the target?
  ✓
- Full (X, Y) data from the source, only X from the target?
  ✗ (in general)
- Full (X, Y) data from the source, no data from the target?
  ✗ (in general)

§4.2: Biased regularization

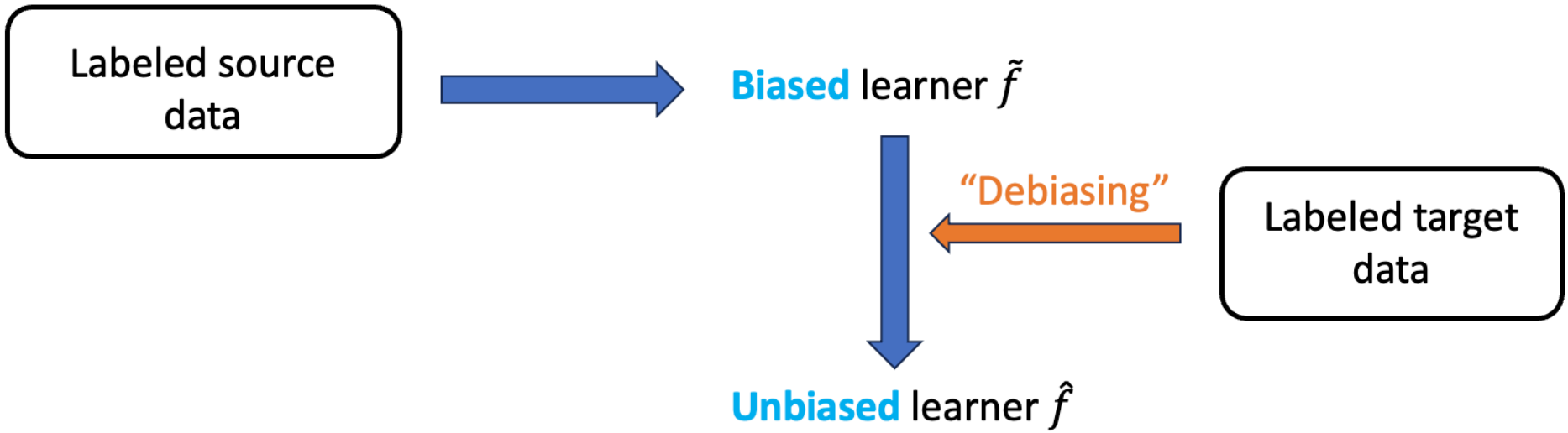
- §4.2.1 Motivation
- §4.2.2 Ridge penalty
- §4.2.3 An adaptive ℓ_2-penalty

§4.2: Biased regularization

- §4.2.1 Motivation
- §4.2.2 Ridge penalty
- §4.2.3 An adaptive ℓ_2-penalty

Biased regularization: motivation

- In Lecture 2, **with only source data available**, we fitted a learner on source data and directly applied it to target domain
- In Lecture 3, **with source data and unlabeled target data available**, we fitted the density ratio by unlabeled data, then fitted a learner on reweighted source data and applied it to target domain
- As we mentioned, when posterior drift exists, a learned fitted by source data can be **severely biased**. We need to remove this bias by using labeled target data.



Biased regularization: motivation

- This motivation leads to an idea called "**biased regularization**". This terminology first appeared in Schölkopf et al. (2001) in the context of penalized kernel methods.
- Schölkopf et al. (2001) studies a non-parametric learner $f \in \mathcal{H}$ by minimizing

$$\sum_{i=1}^{n} \ell(\boldsymbol{x}_i, y_i, f(x_i)) + g(\|f\|). \qquad (\star)$$

They commented in *"Remark 1 (Biased regularization)"* after their Theorem 2:

"When $g(\|f\|) = \frac{1}{2}\|f\|^2$, adding a term $-\langle f_0, f\rangle$ into $(\star)$ can be seen to correspond to an effective overall regularizer of the form $\frac{1}{2}\|f - f_0\|^2$. This, it is no longer the size of $\|f\|$ that is penalized, but the difference to f_0."

In their context, the regularizer $\|f - f_0\|$ is biased towards the pre-trained f_0. That's why it is called "**biased regularization**".

[1] Schölkopf, B., Herbrich, R., & Smola, A. J. (2001, July). A generalized representer theorem. In International conference on computational learning theory (pp. 416-426). Berlin, Heidelberg: Springer Berlin Heidelberg.

Biased regularization: motivation

Main idea of biased regularization:

- First fit a learner $\tilde{f}$ by using source data.
- Then debias $\tilde{f}$ to get $\hat{f}$ by ERM on target data with regularizer $\|f - \tilde{f}\|$
- Apply $\hat{f}$ on target domain

Remark: Do not get confused by "debias $\tilde{f}$ using biased regularization with penalty $\|f - \tilde{f}\|$". Every time we mention "bias", it refers to the bias relative to the **target** domain used as the baseline.

There are different regularizers we can use. The choice usually depends on the metric of similaity between different domains. We will discuss some of them in this section.

It is still an open question which penalty is more reliable to use in practice.

Biased regularization: application examples

Biased regularization has been used in many applications 15-20 years ago and achieved great success, even without comprehensive theoretical understandings.

- Orabona et al. (2009) solves the following modified SVM for hand prosthetics control:

$$\begin{aligned}\min_{\boldsymbol{a},b,\boldsymbol{u},\boldsymbol{v}} \ & \frac{1}{2}\|\boldsymbol{a}-\boldsymbol{a}'\|_2^2 + C(\mathbf{1}^\top \boldsymbol{u} + \mathbf{1}^\top \boldsymbol{v}) \\ \text{s.t. } & \boldsymbol{a}^\top \boldsymbol{x}_i^{(1)} + b \geq 1 - u_i, \quad i = 1:n_1 \\ & \boldsymbol{a}^\top \boldsymbol{x}_i^{(0)} + b \leq -1 + v_i, \quad i = 1:n_0 \\ & \boldsymbol{u} \geq \mathbf{0}, \boldsymbol{v} \geq \mathbf{0}, a \in \mathbb{R}^d, b \in \mathbb{R},\end{aligned}$$

  where they replaced $\|a\|_2$ by $\|\boldsymbol{a}-\boldsymbol{a}'\|_2$ with some pre-trained $\boldsymbol{a}'$.
- Yang et al. (2007) models $f - f'$ through an SVM with f' a pre-trained model, in cross-domained video concept detection.

---

[1] Orabona, F., Castellini, C., Caputo, B., Fiorilla, A. E., & Sandini, G. (2009, May). Model adaptation with least-squares SVM for adaptive hand prosthetics. In 2009 IEEE international conference on robotics and automation (pp. 2897-2903). IEEE.

[2] Yang, J., Yan, R., & Hauptmann, A. G. (2007, September). Cross-domain video concept detection using adaptive svms. In Proceedings of the 15th ACM international conference on Multimedia (pp. 188-197).

§4.2: Biased regularization

- §4.2.1 Motivation
- §4.2.2 Ridge penalty
- §4.2.3 An adaptive ℓ_2-penalty

Biased regularization: ridge penalty

In this section, we focus on a specific regularizer, **the ridge penalty**.
We adopt the following ERM setting from Kuzborskij and Orabona (2013, 2017).

- Target data $\mathcal{D}^{(0)} = \{\boldsymbol{x}_i^{(0)}, y_i^{(0)}\}_{i=1}^{n_0} \sim \mu^{\otimes n_0}$
- Source data $\Rightarrow$ a learner $\tilde{f}$, independent of $\mathcal{D}^{(0)}$
- **Goal:** Minimize $R^{(0)}(f) \coloneqq \mathbb{E}_{(X,Y)\sim\mu}[\ell(f(X), Y)]$ using $\mathcal{D}^{(0)}$ and $\tilde{f}$
- **Regularized ERM** (Kuzborskij and Orabona, 2017): consider

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + \tilde{f}(\boldsymbol{x}),$$

learn $\boldsymbol{w}$ through

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \left\{ \frac{1}{n_0} \sum_{i=1}^{n_0} \ell\big(\boldsymbol{w}^\top \boldsymbol{x}_i^{(0)} + \tilde{f}(\boldsymbol{x}_i^{(0)}), y_i^{(0)}\big) + \lambda \|\boldsymbol{w}\|_2^2 \right\},$$

and output $\hat{f}(\boldsymbol{x}) = \hat{\boldsymbol{w}}^\top \boldsymbol{x} + \tilde{f}(\boldsymbol{x})$.

---

[1] Kuzborskij, I., & Orabona, F. (2013, May). Stability and hypothesis transfer learning. In International Conference on Machine Learning (pp. 942-950). PMLR.

[2] Kuzborskij, I., & Orabona, F. (2017). Fast rates by transferring from auxiliary hypotheses. Machine Learning, 106, 171-195.

Biased regularization: ridge penalty

If $\tilde{f}(\boldsymbol{x}) = \tilde{\boldsymbol{w}}^\top \boldsymbol{x}$, then the regularized ERM

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \left\{ \frac{1}{n_0} \sum_{i=1}^{n_0} \ell\big(\boldsymbol{w}^\top \boldsymbol{x}_i^{(0)} + \tilde{f}(\boldsymbol{x}_i^{(0)}), y_i^{(0)}\big) + \lambda \|\boldsymbol{w}\|_2^2 \right\}, \qquad (1)$$

$$\hat{f}(\boldsymbol{x}) = \hat{\boldsymbol{w}}^\top \boldsymbol{x} + \tilde{f}(\boldsymbol{x}),$$

can be reparameterized as

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \left\{ \frac{1}{n_0} \sum_{i=1}^{n_0} \ell\big(\boldsymbol{w}^\top \boldsymbol{x}_i^{(0)}, y_i^{(0)}\big) + \lambda \|\boldsymbol{w} - \tilde{\boldsymbol{w}}\|_2^2 \right\}, \qquad (2)$$

$$\hat{f}(\boldsymbol{x}) = \hat{\boldsymbol{w}}^\top \boldsymbol{x}.$$

- (2) is in the form of biased regularization we defined before
- (1) is more flexible in general, because we can treat $\tilde{f}$ as a "black box", which means $\tilde{f}$ does not need to be linear
- $\hat{\boldsymbol{w}}$ in (2) can be viewed as the **proximal operator** of $\frac{1}{n_0} \sum_{i=1}^{n_0} \ell\big(\boldsymbol{w}^\top \boldsymbol{x}_i^{(0)}, y_i^{(0)}\big)$ (as a function of $\boldsymbol{w}$) at $\tilde{\boldsymbol{w}}$. The entire minimum is called **Moreau envelop** of $\frac{1}{n_0} \sum_{i=1}^{n_0} \ell\big(\boldsymbol{w}^\top \boldsymbol{x}_i^{(0)}, y_i^{(0)}\big)$ (as a function of $\boldsymbol{w}$).

Biased regularization: ridge penalty

Theorem 4.2.1 (Kuzborskij and Orabona, 2017)

Under certain conditions (second-order smooth and bounded ℓ, $\tilde{f}(\boldsymbol{x}) = \tilde{\boldsymbol{w}}^\top \boldsymbol{x}$ with $\|\tilde{\boldsymbol{w}}\|_2 \le C$ etc.), for $\lambda \asymp \tau^{-1} n_0^{-1/4} \le C$, we have

$$R^{(0)}(\hat{f}) \le \min_{\|\boldsymbol{w}\|_2 \le \tau} R^{(0)}(f_{\boldsymbol{w}}) + \mathcal{O}_{\mathbb{P}}\left(\frac{\tau}{n_0^{1/4}} + \sqrt{\frac{1}{n_0}} \right),$$

where $f_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + \tilde{f}(\boldsymbol{x})$, $\hat{f}(\boldsymbol{x}) = f_{\hat{\boldsymbol{w}}}(\boldsymbol{x}) = \hat{\boldsymbol{w}}^\top \boldsymbol{x} + \tilde{f}(\boldsymbol{x})$.

- The oracle inequality looks great: we can generalize to the target domain based on the pre-trained $\tilde{f}$ with an adjustment linear term $\hat{\boldsymbol{w}}^\top \boldsymbol{x}$
- But there are also some issues.

Biased regularization: ridge penalty

Main result from the last slide: $\lambda \asymp \tau^{-1} n_0^{-1/4} \le C$

$$R^{(0)}(\hat{f}) \le \min_{\|w\|_2 \le \tau} R^{(0)}(f_{\boldsymbol{w}}) + \mathcal{O}_{\mathbb{P}}\left(\frac{\tau}{n_0^{1/4}} + \sqrt{\frac{1}{n_0}} \right).$$

Consider $\tilde{f}(\boldsymbol{x}) = \tilde{\boldsymbol{w}}^\top \boldsymbol{x}$ and $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} R^{(0)}(f_{\boldsymbol{w}})$.

- Case 1: (transfer does help) $\|\boldsymbol{w}^*\|_2 = 0$.
  - ERM on target data $\Rightarrow R^{(0)}(\hat{f}) - \min_{\boldsymbol{w}} R^{(0)}(\boldsymbol{w}^\top \boldsymbol{x}) \lesssim_{\mathbb{P}} \sqrt{1/n_0}$
  - Biased regularization: let's set $\lambda \asymp C$, then $R^{(0)}(\hat{f}) - \min_{\boldsymbol{w}} R^{(0)}(\boldsymbol{w}^\top \boldsymbol{x}) \lesssim_{\mathbb{P}} \sqrt{1/n_0}$. No improvement.
- Case 2: (transfer doesn't help) $\|\boldsymbol{w}^*\|_2 = C'$ with some constant C'.
  - ERM on target data $\Rightarrow R^{(0)}(\hat{f}) - \min_{\boldsymbol{w}} R^{(0)}(\boldsymbol{w}^\top \boldsymbol{x}) \lesssim_{\mathbb{P}} \sqrt{1/n_0}$
  - Biased regularization: let's set $\lambda \asymp n_0^{-1/4}$, then $R^{(0)}(\hat{f}) - \min_{\boldsymbol{w}} R^{(0)}(\boldsymbol{w}^\top \boldsymbol{x}) \lesssim_{\mathbb{P}} n_0^{-1/4}$. Even worse.

Biased regularization: ridge penalty

Intuition:

- Ridge penalty is **not adaptive** to the similarity between source and target domains (which means we need different λ for different problems).
- In some problems, we are not improving the performance compared to ERM on target data (which should serve as our benchmark).

But the ridge penalty often leads to explicit formulas of the learner and is very popular in literature. Besides the previous two papers, the following papers also study ridge penalty in biased regularization (there are many more).

- Evgeniou and Pontil (2004): regularized SVM
- Chen et al. (2015): linear regression with one auxiliary source dataset
- Denevi et al. (2018): optimize over $\tilde{\boldsymbol{w}}$ in biased regularization with ridge penalty
- T Dinh et al. (2020): use of ridge penalty in federated learning

[1] Evgeniou, T., & Pontil, M. (2004, August). Regularized multi–task learning. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 109-117).

[2] Chen, A., & Shi, M. (2015). Data Enriched Linear Regression. Electronic Journal of Statistics, 9, 1078-1112.

[3] Denevi, G., Ciliberto, C., Stamos, D., & Pontil, M. (2018). Learning to learn around a common mean. Advances in neural information processing systems, 31.

[4] T Dinh, C., Tran, N., & Nguyen, J. (2020). Personalized federated learning with moreau envelopes. Advances in Neural Information Processing Systems, 33, 21394-21405.

Biased regularization: ridge penalty

We can further understand the non-adaptivity of ridge penalty through a simple example. Consider the following Gaussian mean estimation problem.

- Target data $\{x_i^{(0)}\}_{i=1}^{n_0} \overset{\text{i.i.d.}}{\sim} N(\theta^*, 1)$, we want to estimate θ^*
- Source data $\{x_i^{(1)}\}_{i=1}^{n_1} \Rightarrow$ an estimator $\tilde{\theta}$
- Biased regularization with ridge penalty:

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}} \left\{ \frac{1}{n_0} \sum_{i=1}^{n_0} |x_i^{(0)} - \theta|^2 + \lambda |\theta - \tilde{\theta}|^2 \right\}$$

It is easy to see that the objective function can be written as

$$|\bar{x}^{(0)} - \theta|^2 + \lambda |\theta - \tilde{\theta}|^2,$$

with sample mean $\bar{x}^{(0)} = n_0^{-1} \sum_{i=1}^{n_0} x_i^{(0)}$. Minimizing it leads to

$$\hat{\theta} = \frac{1}{1+\lambda} \bar{x}^{(0)} + \frac{\lambda}{1+\lambda} \tilde{\theta}.$$

Biased regularization: ridge penalty

Let's take a closer look at $\hat{\theta} = \frac{1}{1+\lambda}\bar{x}^{(0)} + \frac{\lambda}{1+\lambda}\tilde{\theta}$. The L_2^2-estimation error

$$\begin{aligned}\mathbb{E}|\hat{\theta} - \theta^*|^2 &= \left(\frac{1}{1+\lambda}\right)^2 \mathbb{E}|\bar{x}^{(0)} - \theta^*|^2 + \left(\frac{\lambda}{1+\lambda}\right)^2 \mathbb{E}|\tilde{\theta} - \theta^*|^2 \\ &= \left(\frac{1}{1+\lambda}\right)^2 \frac{1}{n_0} + \left(\frac{\lambda}{1+\lambda}\right)^2 \mathbb{E}|\tilde{\theta} - \theta^*|^2.\end{aligned}$$

- Optimize over $\lambda \geq 0$, we get $\lambda = \frac{1/n_0}{\mathbb{E}|\tilde{\theta}-\theta^*|^2}$, which leads to the risk

$$\mathbb{E}|\hat{\theta}-\theta^*|^2 = \frac{1/n_0 \cdot \mathbb{E}|\tilde{\theta} - \theta^*|^2}{1/n_0 + \mathbb{E}|\tilde{\theta} - \theta^*|^2} \asymp \min\left\{\frac{1}{n_0}, \mathbb{E}|\tilde{\theta}-\theta^*|^2\right\}. \to \text{minimax optimal}$$

  - If $\mathbb{E}|\tilde{\theta} - \theta^*|^2 \gtrsim 1/n_0$ (transfer doesn't help):
    we need a small λ to make $\hat{\theta}$ behave more like $\bar{x}^{(0)}$
  - If $\mathbb{E}|\tilde{\theta} - \theta^*|^2 \lesssim 1/n_0$ (transfer does help):
    we need a large λ to make $\hat{\theta}$ behave more like $\tilde{\theta}$

- No universal λ can achieve the optimal rate (similar to previous examples)

§4.2: Biased regularization

- §4.2.1 Motivation
- §4.2.2 Ridge penalty
- §4.2.3 An adaptive ℓ_2-penalty

Biased regularization: an adaptive ℓ_2-penalty

Question: Does there exist a penalty which is **adaptive** to the problem structure with a universal tuning parameter λ?

Let's consider the same Gaussian mean estimation problem, but a different ℓ_2-penalty:

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}} \left\{ \frac{1}{2n_0} \sum_{i=1}^{n_0} |x_i^{(0)} - \theta|^2 + \lambda |\theta - \tilde{\theta}| \right\}.$$

It can be seen that

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}} \left\{ \frac{1}{2} |\bar{x}^{(0)} - \theta|^2 + \lambda |\theta - \tilde{\theta}| \right\} = \begin{cases} \tilde{\theta}, & \text{if } |\tilde{\theta} - \bar{x}^{(0)}| \le \lambda; \\ \bar{x}^{(0)} - \lambda, & \text{if } \bar{x}^{(0)} > \tilde{\theta} + \lambda; \\ \bar{x}^{(0)} + \lambda, & \text{if } \bar{x}^{(0)} < \tilde{\theta} - \lambda. \end{cases}$$

And

$$\begin{aligned} |\hat{\theta} - \theta^*| \lesssim\ & |\tilde{\theta} - \theta^*| \cdot \mathbb{1}(|\tilde{\theta} - \theta^*| \le |\bar{x}^{(0)} - \theta^*| + \lambda) \\ & + (|\bar{x}^{(0)} - \theta^*| + \lambda) \mathbb{1}(|\tilde{\theta} - \theta^*| > \lambda - |\bar{x}^{(0)} - \theta^*|). \end{aligned}$$

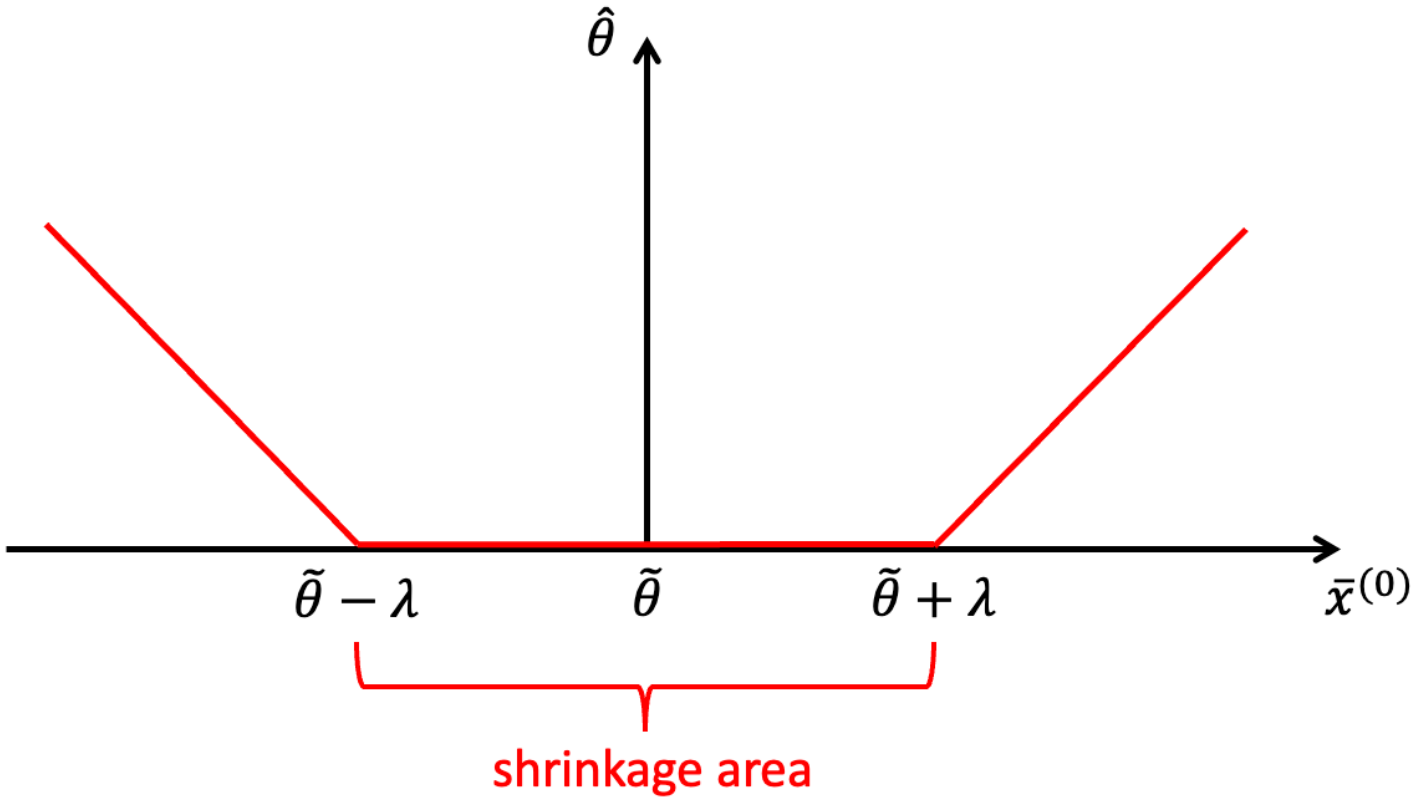
Intuition: Set $\lambda \approx 2|\bar{x}^{(0)} - \theta^*|$ then $|\hat{\theta} - \theta^*| \lesssim \min\{|\tilde{\theta} - \theta^*|, |\bar{x}^{(0)} - \theta^*|\}$.

Biased regularization: an adaptive ℓ_2-penalty

In fact, let $\sqrt{1/n_0} \gtrsim \lambda \geq C\sqrt{1/n_0}$ with a large C, it can be shown that

$$\mathbb{E}|\hat{\theta} - \theta^*| \lesssim \min\left\{\frac{1}{n_0}, \mathbb{E}|\tilde{\theta} - \theta^*|^2\right\}.$$

This is the adaptivity we want!



Biased regularization: an adaptive ℓ_2-penalty

- With ℓ_2-penalty and some $\lambda \asymp \sqrt{1/n_0}$, we have the adaptivity:

$$\mathbb{E}|\hat{\theta} - \theta^*| \lesssim \min\left\{\frac{1}{n_0}, \mathbb{E}|\tilde{\theta} - \theta^*|^2\right\}. \qquad (\star)$$

- Recall that $\tilde{\theta}$ is an estimator fitted on the source data $\{x_{i=1}^{(1)}\}^{n_1}$. If

$$\{x_{i=1}^{(1)}\}^{n_1} \overset{\text{i.i.d.}}{\sim} N(\theta', 1),$$

then a natural choice of $\tilde{\theta}$ would be the sample mean $\bar{x}^{(1)} = n_1^{-1}\sum_{i=1}^{n_1} x_i^{(1)}$, which satisfies $\mathbb{E}|\tilde{\theta} - \theta^*|^2 \lesssim \frac{1}{n_1} + |\theta' - \theta^*|^2$.

Plugging it into $(\star)$, we have

$$\mathbb{E}|\hat{\theta} - \theta^*|^2 \lesssim \min\left\{\underbrace{\frac{1}{n_0}}_{\text{target-only rate: only variance}}, \underbrace{\frac{1}{n_1}}_{\text{source variance}} + \underbrace{|\theta' - \theta^*|^2}_{\text{source bias}}\right\}.$$

"Bias-variance trade-off" in transfer learning

Biased regularization: literature about ℓ_2-penalty

- Li and Bilmes (2007) used ℓ_2-penalty in domain adaptation in classification. They motivate this penalty from a Beyesian perspective and show that it can be used to bound the cross-entropy between likelihood of two domains.
- The adaptivity of ℓ_2-penalty was first comprehensively studied in Duan and Wang (2022), in a multi-task learning context.
- A few of our follow-up works (Tian et al., 2022, 2024) have applied the ℓ_2-penalty on unsupervised problems [to be discussed later]
- Tian et al. (2023) extended this penalty to a representation learning setting and see similar adaptivity patterns [to be discussed later]

Now, let's generalize the 1-dimensional single-source problem to a multi-dimensional multi-source multi-task learning problem (Duan and Wang, 2022).

---

[1] Li, X., & Bilmes, J. (2007, March). A bayesian divergence prior for classiffier adaptation. In Artificial Intelligence and Statistics (pp. 275-282). PMLR.

[2] Duan, Y., & Wang, K. (2022). Adaptive and robust multi-task learning. arXiv preprint arXiv:2202.05250. (version 2)

[3] Tian, Y., Weng, H., & Feng, Y. (2022). Unsupervised multi-task and transfer learning on gaussian mixture models. arXiv preprint arXiv:2209.15224.

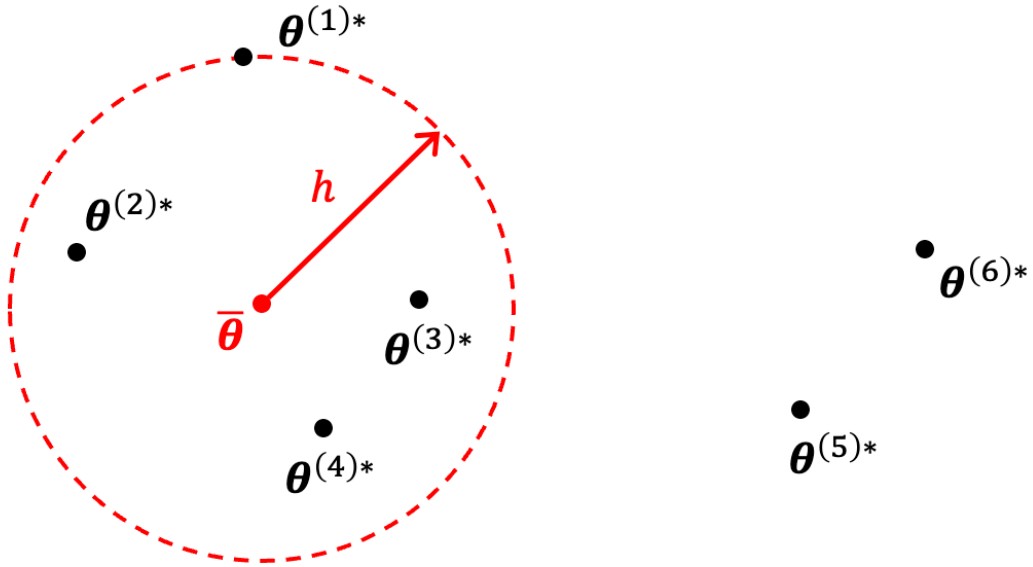
[4] Tian, Y., Weng, H., & Feng, Y. (2024). Towards the Theory of Unsupervised Federated Learning: Non-asymptotic Analysis of Federated EM Algorithms. arXiv preprint arXiv:2310.15330. (accepted by ICML 2024)

[5] Tian, Y., Gu, Y., & Feng, Y. (2023). Learning from similar linear representations: adaptivity, minimaxity, and robustness. arXiv preprint arXiv:2303.17765.

ℓ_2-penalty in multi-task learning

Consider the following multi-task Gaussian mean estimation problem.

- The k-th dataset $\mathcal{D}^{(k)} = \{\boldsymbol{x}_i^{(k)}\}_{i=1}^{n_k} \overset{\text{i.i.d.}}{\sim} N(\boldsymbol{\theta}^{(k)*}, \boldsymbol{I}_d)$, $k = 1 : K := [K]$
- $\boldsymbol{\theta}^{(k)*} \in \mathbb{R}^d$. For simplicity, assume $n_k \equiv n$ for all k
- Similarity between tasks: $\min_{\bar{\boldsymbol{\theta}}} \max_{k \in S} \|\boldsymbol{\theta}^{(k)*} - \bar{\boldsymbol{\theta}}\|_2 \leq h$, where h is unknown
- The set S is sometimes called the informative set, and $\epsilon = |S^c|/K$ is the contamination proportion or outlier proportion, where $S^c = [K] \backslash S$
- h characterizes the similarity level between tasks



$S = \{1, 2, 3, 4\}$, $K = 6$, $\epsilon = 1/3$

- **Goal:** Find a good estimator $\widehat{\boldsymbol{\theta}}^{(k)}$ for tasks in S to minimize

$$\max_{k \in S} \|\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{(k)*}\|_2. \quad \text{(worst-case performance)}$$

ℓ_2-penalty in multi-task learning

Biased regularization with ℓ_2-penalty: (Duan and Wang, 2022)

$$\{\widehat{\boldsymbol{\theta}}^{(k)}\}_{k=1}^K, \widehat{\bar{\boldsymbol{\theta}}} = \underset{\{\boldsymbol{\theta}^{(k)}\},\bar{\boldsymbol{\theta}}}{\arg\min} \left\{ \frac{1}{K} \sum_{k=1}^K \left(\frac{1}{n} \sum_{i=1}^n \|\boldsymbol{x}_i^{(k)} - \boldsymbol{\theta}^{(k)}\|_2^2 + \frac{\lambda}{\sqrt{n}} \|\boldsymbol{\theta}^{(k)} - \bar{\boldsymbol{\theta}}\|_2 \right) \right\}$$

- This is an extension of the previous single-source case
- λ controls the bias towards $\widehat{\bar{\boldsymbol{\theta}}}$
  - $\lambda \approx 0$: $\widehat{\boldsymbol{\theta}}^{(k)} \approx \bar{\boldsymbol{x}}^{(k)} = n_k^{-1} \sum_{k=1}^K \boldsymbol{x}_i^{(k)}$ $\quad \rightarrow$ good for large h
  - $\lambda \to \infty$: $\widehat{\boldsymbol{\theta}}^{(k)} \equiv \widehat{\bar{\boldsymbol{\theta}}}$ $\quad \rightarrow$ good for small h
  - As before, we will have a universal λ that is adaptive
- We can rewrite the optimization problem into

$$\{\widehat{\boldsymbol{\theta}}^{(k)}\}_{k=1}^K, \widehat{\bar{\boldsymbol{\theta}}} = \underset{\{\boldsymbol{\theta}^{(k)}\},\bar{\boldsymbol{\theta}}}{\arg\min} \left\{ \frac{1}{K} \sum_{k=1}^K \left(\|\bar{\boldsymbol{x}}^{(k)} - \boldsymbol{\theta}^{(k)}\|_2^2 + \frac{\lambda}{\sqrt{n}} \|\boldsymbol{\theta}^{(k)} - \bar{\boldsymbol{\theta}}\|_2 \right) \right\}$$

ℓ_2-penalty in multi-task learning

Theorem 4.2.1 (Duan and Wang, 2022)

With $\lambda \asymp \sqrt{p + \log K}$, *w.h.p.:*

$$\max_{k \in S} \|\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{(k)*}\|_2 \lesssim \underbrace{\sqrt{\frac{d}{nK}}}_{\text{oracle}} + \min\left\{ \underbrace{\sqrt{\frac{d + \log K}{n}}}_{\text{single-task rate}}, \underbrace{h}_{\text{heterogeneity}} \right\} + \underbrace{\epsilon\sqrt{\frac{d + \log K}{n}}}_{\text{outliers}},$$

$$\max_{k \in S^c} \|\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{(k)*}\|_2 \lesssim \sqrt{\frac{d + \log K}{n}}.$$

- For single-task methods, the minimax rate is $\sqrt{\frac{d+\log K}{n}}$
- The rate is faster than single-task rate, with:
  - Sufficient similarity: $h \ll \sqrt{\frac{d+\log K}{n}}$
  - Many tasks: $K \to \infty$
  - Small fraction of outlier tasks: $\epsilon \to 0$
- Therefore, we have achieved:
  - Adaptivity to task similarity h
  - Robustness against a small fraction of outliers

ℓ_2-penalty in multi-task learning

Question: We have explained the intuition of adaptivity before. But why do we have robustness against outliers? In fact, the same result holds even for **arbitrary** contamination on outlier tasks in S^c.

Answer: There are connections between **penalized over-parameterized models** and **robustified M-estimators**. (She and Owen, 2011; Donoho and Montanari, 2016)

Let's consider mean estimation in the single-task setting.

- $\{\boldsymbol{x}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} N(\theta^*, 1)$
- Huber contamination: An arbitrary contamination happens on $S^c \subseteq [n]$
- How can we consistently estimate θ^*?

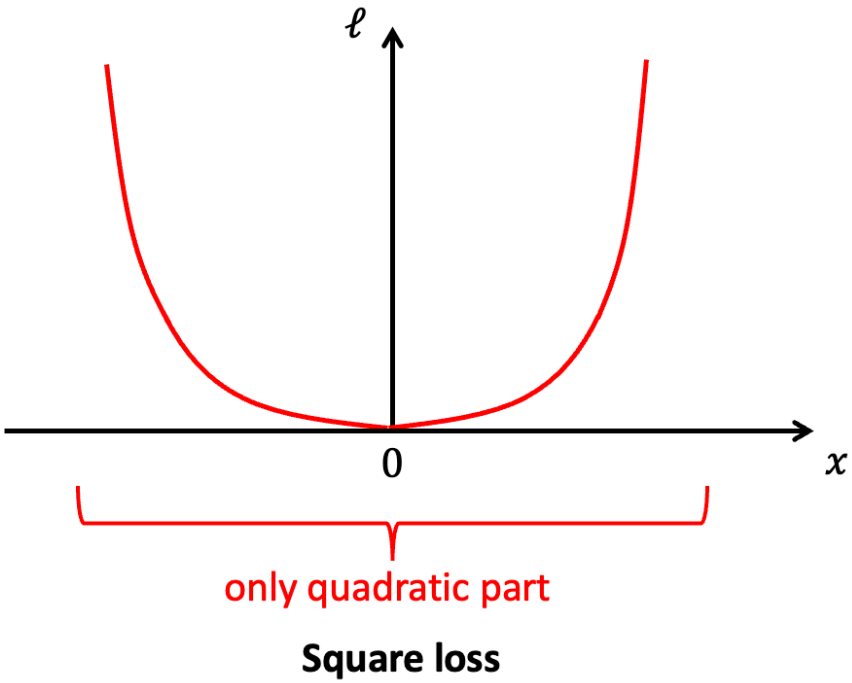
---

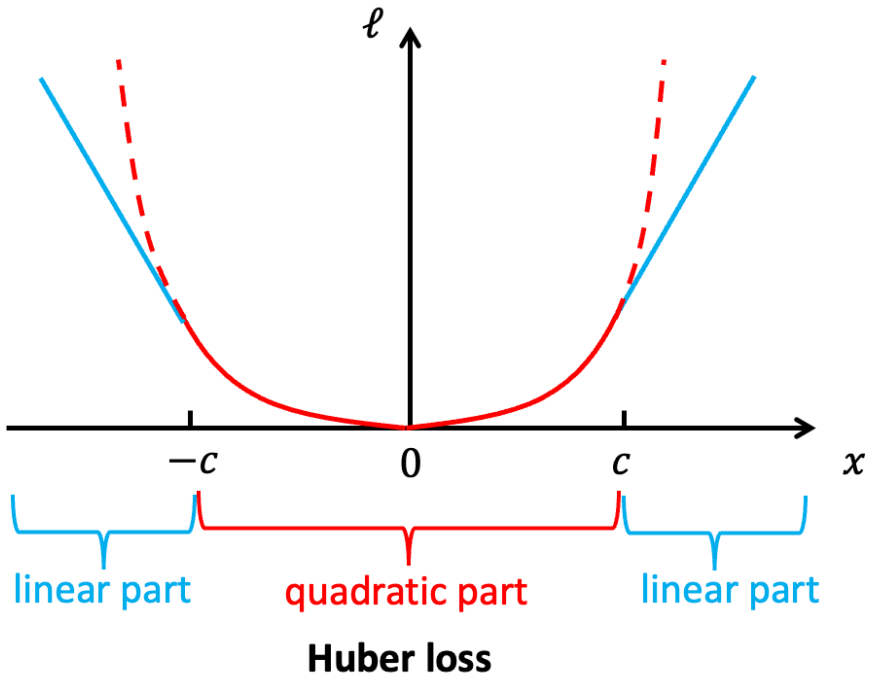
[1] She, Y., & Owen, A. B. (2011). Outlier detection using nonconvex penalized regression. Journal of the American Statistical Association, 106(494), 626-639.

[2] Donoho, D., & Montanari, A. (2016). High dimensional robust m-estimation: Asymptotic variance via approximate message passing. Probability Theory and Related Fields, 166, 935-969.

ℓ_2-penalty in multi-task learning

Why are we in trouble with square loss and sample mean?





◦ **Method 1:** M-estimation with Huber's loss (Huber, 1964)

$$\hat{\theta} = \arg\min_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho_c(x_i - \theta) \right\},$$

where $\rho_c(x) = \begin{cases} x^2/2, & \text{if } |x| \le c; \\ c|x| - c^2/2, & \text{if } |x| > c. \end{cases}$

[1] Huber, P. J. (1964). Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics, 73-101.

ℓ_2-penalty in multi-task learning

- **Method 1:** M-estimation with Huber's loss

$$\hat{\theta} = \arg\min_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho_c(x_i - \theta) \right\}. \quad (1)$$

- **Method 2:** Penalized over-parameterization (McCann and Welsch, 2007; Gannaz, 2007)

$$\hat{\theta}, \widehat{\boldsymbol{\Delta}} = \arg\min_{\theta, \boldsymbol{\Delta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} |x_i - \theta - \Delta_i|^2 + \lambda \|\boldsymbol{\Delta}\|_1 \right\}. \quad (2)$$

Theorem 4.2.2 (She and Owen, 2011)

(1) and (2) are equivalent and there is a one-to-one mapping between λ and c.

---

[1] McCann, L., & Welsch, R. E. (2007). Robust variable selection using least angle regression and elemental set sampling. Computational Statistics & Data Analysis, 52(1), 249-257.

[2] Gannaz, I. (2007). Robust estimation and wavelet thresholding in partially linear models. Statistics and Computing, 17, 293-310.

[3] She, Y., & Owen, A. B. (2011). Outlier detection using nonconvex penalized regression. Journal of the American Statistical Association, 106(494), 626-639.

ℓ_2-penalty in multi-task learning

Penalized over-parameterization (McCann and Welsch, 2007; Gannaz, 2007):

$$\hat{\theta}, \widehat{\boldsymbol{\Delta}} = \underset{\theta, \boldsymbol{\Delta}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(|x_i - \theta - \Delta_i|^2 + \lambda |\Delta_i| \right) \right\}. \quad (\star)$$

Recall the equivalent form of the ℓ_2-biased regularization:

$$\{\widehat{\boldsymbol{\theta}}^{(k)}\}_{k=1}^{K}, \widehat{\bar{\boldsymbol{\theta}}} = \underset{\{\boldsymbol{\theta}^{(k)}\}, \bar{\boldsymbol{\theta}}}{\arg\min} \left\{ \frac{1}{K} \sum_{k=1}^{K} \left(\|\bar{\boldsymbol{x}}^{(k)} - \boldsymbol{\theta}^{(k)}\|_2^2 + \frac{\lambda}{\sqrt{n}} \|\boldsymbol{\theta}^{(k)} - \bar{\boldsymbol{\theta}}\|_2 \right) \right\}. \quad (\dagger)$$

- $(\dagger)$ can be seen as a variant of $(\star)$ by re-parameterization $\widehat{\boldsymbol{\theta}}^{(k)} = \widehat{\bar{\boldsymbol{\theta}}} + \widehat{\boldsymbol{\Delta}}^{(k)}$, which illustrates the robustness against contamination.
- Note that the contamination in our setting is on the task level while the contamination in classical robust statistics is on the observation level.

ℓ_2-penalty in multi-task learning

Recall the upper bound of estimation error:

$$\max_{k\in S}\|\widehat{\boldsymbol{\theta}}^{(k)}-\boldsymbol{\theta}^{(k)*}\|_2 \lesssim_{\mathbb{P}} \underbrace{\sqrt{\frac{d}{nK}}}_{\text{oracle}} + \min\left\{\underbrace{\sqrt{\frac{d+\log K}{n}}}_{\text{single-task rate}}, \underbrace{h}_{\text{heterogeneity}}\right\} + \underbrace{\epsilon\sqrt{\frac{d+\log K}{n}}}_{\text{outliers}}$$

We also have a nearly matching **information-theoretic** lower bound.

Theorem 4.2.3 (Duan and Wang, 2022)

With prob. $\geq 1/10$,

$$\inf_{\{\widehat{\boldsymbol{\theta}}^{(k)}\}_{k=1}^K}\sup_{\substack{|S^c|/K\leq\epsilon\\ \{\boldsymbol{\theta}^{(k)*}\}_{k=1}^K}}\max_{k\in S}\|\widehat{\boldsymbol{\theta}}^{(k)}-\boldsymbol{\theta}^{(k)*}\|_2 \gtrsim \sqrt{\frac{d}{nK}}+\min\left\{\sqrt{\frac{d+\log K}{n}}, h\right\}+\frac{\epsilon}{\sqrt{n}}.$$

- The ϵ-related term doesn't match.
- In fact, in one of our ongoing works (Tian and Avella, 2024+), we showed that most biased regularization methods have the **algorithmic** lower bound $\epsilon\sqrt{\frac{d}{n}}$
- Time for new robust multi-task learning methods!

ℓ_2-penalty in multi-task learning

Finally, the method & theory of ℓ_2-biased regularization can be extended to an ERM setting.

- The k-th dataset $\mathcal{D}^{(k)} = \{\boldsymbol{x}_i^{(k)}, y_i^{(k)}\}_{i=1}^{n_k} \overset{\text{i.i.d.}}{\sim} (X^{(k)}, Y^{(k)})$, loss function $\ell(\boldsymbol{\theta}, (X, Y))$, and $\boldsymbol{\theta}^{(k)*} := \arg\min_{\boldsymbol{\theta}} \mathbb{E}\ell(\boldsymbol{\theta}, (X^{(k)}, Y^{(k)}))$, for $k = 1 : K$
- $\boldsymbol{\theta}^{(k)*} \in \mathbb{R}^d$. For simplicity, assume $n_k \equiv n$ for all k
- Similarity between tasks: $\min_{\bar{\boldsymbol{\theta}}} \max_{k \in S} \|\boldsymbol{\theta}^{(k)*} - \bar{\boldsymbol{\theta}}\|_2 \leq h$, where h is unknown
- The set S is sometimes called the informative set, and $\epsilon = |S^c|/K$ is the contamination proportion or outlier proportion, where $S^c = [K] \backslash S$

Theorem 4.2.4 (Duan and Wang, 2022)

Under certain assumptions, with $\lambda \asymp \sqrt{p + \log K}$, *w.h.p.:*

$$\max_{k \in S} \|\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{(k)*}\|_2 \lesssim \underbrace{\sqrt{\frac{d}{nK}}}_{\text{oracle}} + \min\left\{ \underbrace{\sqrt{\frac{d + \log K}{n}}}_{\text{single-task rate}}, \underbrace{h}_{\text{heterogeneity}} \right\} + \underbrace{\epsilon\sqrt{\frac{d + \log K}{n}}}_{\text{outliers}},$$

$$\max_{k \in S^c} \|\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{(k)*}\|_2 \lesssim \sqrt{\frac{d + \log K}{n}}.$$

Key components for adaptivity and robustness

There are a few important ingredients to help design a regularizer adaptive to the unknown task similarity and robust against outlier tasks.

- Some shrinkage regime: which leads to the oracle rate $1/\sqrt{nK}$
  - This usually requires singularity around 0 (Fan and Li, 2001)
  - The shrinkage radius should be $\approx$ the single-task error rate
  - This can be connected to Hodge's "super-efficiency" phenomenon (Van der Vaart, 2000)
- The regularized learner should be connected to some robustified M-estimator (She and Owen, 2011; Donoho and Montanari, 2016)
  - ℓ_2-penalty $\Leftrightarrow$ Huber's loss
  - SCAD-penalty $\Leftrightarrow$ Hampel's loss
  - ...

---

[1] Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association, 96(456), 1348-1360.

[2] Van der Vaart, A. W. (2000). Asymptotic statistics (Vol. 3). Cambridge university press.

[3] She, Y., & Owen, A. B. (2011). Outlier detection using nonconvex penalized regression. Journal of the American Statistical Association, 106(494), 626-639.

[4] Donoho, D., & Montanari, A. (2016). High dimensional robust m-estimation: Asymptotic variance via approximate message passing. Probability Theory and Related Fields, 166, 935-969.

§4.3: Extension to high-dimensional regressions

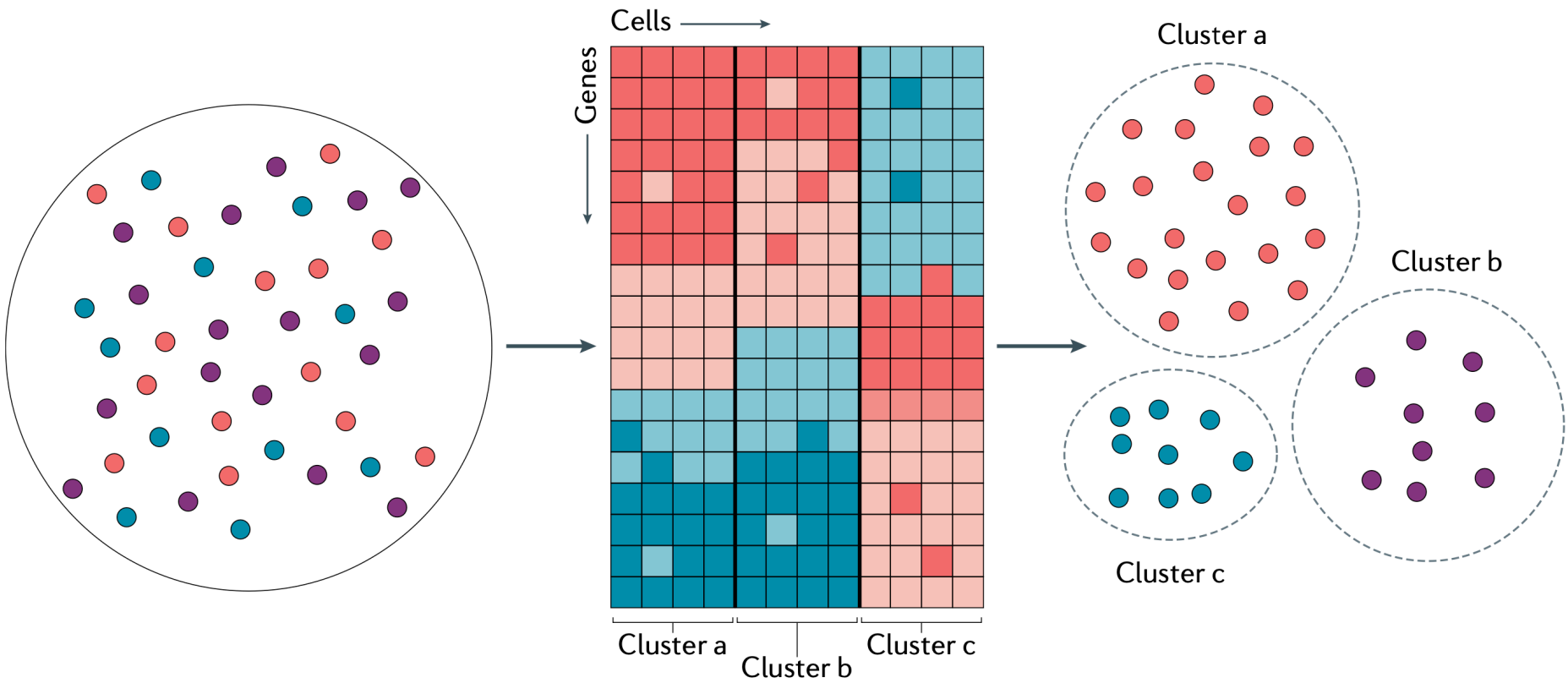
- §4.3.1 ℓ_1-penalty with GLMs
- §4.3.2 Go robust: better ways to aggregate data
- §4.3.3 Block penalty with multi-task learning

§4.3: Extension to high-dimensional regressions

- §4.3.1 ℓ_1-penalty with GLMs
- §4.3.2 Go robust: better ways to aggregate data
- §4.3.3 Block penalty with multi-task learning

ℓ_1-penalty with GLMs

Compared to low-dimensional problems, transfer learning in high-dimensional problems could be more helpful, because of the potentially limited target sample size and high dimensionality of the problem.



Picture is from: Wu, Y., & Zhang, K. (2020). Tools for the analysis of high-dimensional single-cell RNA sequencing data. Nature Reviews Nephrology, 16(7), 408-421.

ℓ_1-penalty with GLMs

Let us follow Li et al. (2021); Tian and Feng (2023b); Li et al. (2023), and consider the following high-dimensional regression setting.

- $\{(\boldsymbol{x}_i^{(k)}, y_i^{(k)})\}_{i=1}^{n_k} \overset{\text{i.i.d.}}{\sim} (X^{(k)}, Y^{(k)})$ for $k = 0 : K$, where $X^{(k)} \in \mathbb{R}^d$, $Y^{(k)} \in \mathbb{R}$, and

$$Y^{(k)} = (X^{(k)})^\top \boldsymbol{\theta}^{(k)*} + \epsilon^{(k)},$$

  $\mathbb{E}[X^{(k)}(X^{(k)})^\top] = \boldsymbol{\Sigma}^{(k)}$, $\epsilon^{(k)}$ is independent of $X^{(k)}$ and zero-mean sub-Gaussian (with a constant-level variance proxy).
- For simplicity, assume $n_k \equiv n$, $\boldsymbol{\Sigma}^{(k)} \equiv \boldsymbol{\Sigma}$
- **Sparsity:** $\|\boldsymbol{\theta}^{(0)*}\|_0 \leq s \ll d$
- **Relationship between tasks:** $\max_{k \in S} \|\boldsymbol{\theta}^{(k)*} - \boldsymbol{\theta}^{(0)*}\|_1 \leq h$ (unknown), where the **informative** set $S \subseteq [K]$ is also unknown
- **Goal:** estimate $\boldsymbol{\theta}^{(0)*}$

---

[1] Li, S., Cai, T. T., & Li, H. (2022). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. Journal of the Royal Statistical Society Series B: Statistical Methodology, 84(1), 149-173.

[2] Tian, Y., & Feng, Y. (2023). Transfer learning under high-dimensional generalized linear models. Journal of the American Statistical Association, 118(544), 2684-2697.

[3] Li, S., Zhang, L., Cai, T. T., & Li, H. (2023). Estimation and inference for high-dimensional generalized linear models with knowledge transfer. Journal of the American Statistical Association, 1-12.

ℓ_1-penalty with GLMs: S is known

Recall our bias regularization procedure. We need to:

(i) First aggregate the source datasets to obtain an estimator $\tilde{\boldsymbol{\theta}}$

(ii) Debias $\tilde{\boldsymbol{\theta}}$ using the target data by **penalization**

This motivates the following algorithm.

Two-step algorithm S-Trans-GLM: [2](Li et al., 2021; Tian and Feng, 2023b)

◦ Step 1: (Transferring) Obtain a global estimator from data aggregation:

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min} \left\{ \frac{1}{(|S|+1)} \sum_{k\in\{0\}\cup S} \frac{1}{n} \sum_{i=1}^{n} (y_i^{(k)} - \boldsymbol{\theta}^\top \boldsymbol{x}_i^{(k)})^2 + \lambda_1 \|\boldsymbol{\theta}\|_1 \right\}$$

◦ Step 2: (Debiasing) Debias $\tilde{\boldsymbol{\theta}}$ using the target data by **penalization**:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i^{(0)} - \boldsymbol{\theta}^\top \boldsymbol{x}_i^{(0)})^2 + \lambda_2 \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_1 \right\}$$

[2] Li et al. (2021) uses data splitting for two steps, while Tian and Feng (2023b) does not.

ℓ_1-penalty with GLMs: theory

Theorem 4.3.1 (Li et al. (2021); Tian and Feng (2023b))

With $\lambda_1 \asymp \sqrt{\frac{\log d}{(|S|+1)n}}$, $\lambda_2 \asymp \sqrt{\frac{\log d}{n}}$, *we have*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{(0)*}\|_2 \lesssim_{\mathbb{P}} \sqrt{\frac{s \log d}{(|S|+1)n}} + \left(\sqrt{\frac{\log d}{n}} h^{1/2}\right) \wedge h,$$

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{(0)*}\|_1 \lesssim_{\mathbb{P}} s\sqrt{\frac{\log d}{(|S|+1)n}} + h.$$

- The target-only minimax ℓ_2 and ℓ_1 estimation errors are $\sqrt{\frac{s \log d}{n}}$ and $s\sqrt{\frac{\log d}{n}}$, respectively.
- Transfer learning helps when the following conditions hold:
  - Sufficient similarity: $h \ll s\sqrt{\frac{\log d}{n}}$;
  - Many source datasets: $|S| \to \infty$.

ℓ_1-penalty with GLMs: theory

Theorem 4.3.2 (Li et al. (2021); Tian and Feng (2023b))

Suppose $d \gtrsim s^{1.01}$. *With prob. at least* $1/4$, *there exists a parameter setting* $\{\boldsymbol{\theta}^{(k)*}\}_{k\in\{0\}\cup S}$ *s.t.*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{(0)*}\|_2 \gtrsim \sqrt{\frac{s\log d}{(|S|+1)n}} + \left(\sqrt{\frac{\log d}{n}}h^{1/2}\right) \wedge h \wedge \sqrt{\frac{s\log d}{n}},$$

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{(0)*}\|_1 \gtrsim s\sqrt{\frac{\log d}{(|S|+1)n}} + h \wedge \left(s\sqrt{\frac{\log d}{n}}\right).$$

- The two-step algorithm is minimax optimal when $h \lesssim s\sqrt{\frac{s\log d}{n}}$

The method and theory can be extended to generalized linear models (GLMs).

ℓ_1-penalty with GLMs

GLMs: $Y^{(k)}|X^{(k)} = \boldsymbol{x} \sim \rho(y)\exp\{y\boldsymbol{x}^\top\boldsymbol{\theta}^{(k)*} - \psi(\boldsymbol{x}^\top\boldsymbol{\theta}^{(k)*})\}$ (density w.r.t. base measure σ), $\psi'(\boldsymbol{x}^\top\boldsymbol{\theta}^{(k)*}) = \mathbb{E}(Y^{(k)}|X^{(k)} = \boldsymbol{x})$ is the *inverse link function*

- Linear regression model: $\psi(x) = x^2/2$
- Logistic regression model: $\psi(x) = \log(1 + e^x)$
- Poisson regression model: $\psi(x) = e^x$

Two-step algorithm S-Trans-GLM: (Tian and Feng, 2023b)

- Step 1: (Transferring) Obtain a global estimator from data aggregation:

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min}\left\{\frac{1}{(|S|+1)}\sum_{k\in\{0\}\cup S}\frac{1}{n}\sum_{i=1}^{n}[-y_i^{(k)}\boldsymbol{\theta}^\top\boldsymbol{x}_i^{(k)} + \psi(\boldsymbol{\theta}^\top\boldsymbol{x}_i^{(k)})] + \lambda_1\|\boldsymbol{\theta}\|_1\right\}$$

- Step 2: (Debiasing) Debias $\tilde{\boldsymbol{\theta}}$ using the target data by **penalization**:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min}\left\{\frac{1}{n}\sum_{i=1}^{n}[-y_i^{(0)}\boldsymbol{\theta}^\top\boldsymbol{x}_i^{(0)} + \psi(\boldsymbol{\theta}^\top\boldsymbol{x}_i^{(0)})] + \lambda_2\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_1\right\}$$

ℓ_1-penalty with GLMs

- Two remaining problems:
  - We don't know S in practice
  - When h is large, the two-step algorithm may suffer from a bad estimation error
- Two solutions:
  - Aggregation: Li et al. (2021, 2022)
  - Selection: Tian and Feng (2023b); Li et al. (2024)

---

[1] Li, S., Cai, T. T., & Li, H. (2022). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. Journal of the Royal Statistical Society Series B: Statistical Methodology, 84(1), 149-173.

[2] Li, S., Zhang, L., Cai, T. T., & Li, H. (2023). Estimation and inference for high-dimensional generalized linear models with knowledge transfer. Journal of the American Statistical Association, 1-12.

[3] Tian, Y., & Feng, Y. (2023). Transfer learning under high-dimensional generalized linear models. Journal of the American Statistical Association, 118(544), 2684-2697.

[4] Li, M., Tian, Y., Feng, Y., & Yu, Y. (2024). Federated Transfer Learning with Differential Privacy. arXiv preprint arXiv:2403.11343.

When S is unknown: aggregation

- Aggregation technique originates from the statistical aggregation literature (Rigollet and Tsybakov, 2012; Dai et al., 2012)
- The main idea: construct estimators based on different candidates S, then combine them by a weighted average
  - Construct an estimator $\hat{R}^{(k)}$ to estimate the "sparsity index"

$$R^{(k)} := \|\boldsymbol{\Sigma}(\boldsymbol{\theta}^{(k)*} - \boldsymbol{\theta}^{(0)*})\|_2.$$

  - Sort the K sources by $\hat{R}^{(k)}$ values with increasing order
  - Construct candidate sets ($\widehat{G}_0 := \emptyset$)

$$\widehat{G}_k = \{1 \le k \le K : \hat{R}^{(k)} \text{ is among the first } k \text{ smallest ones}\}.$$

  - Output $\widehat{\boldsymbol{\theta}} = \sum_{k=0}^{K} \hat{w}_k \cdot [(\{0\} \cup \widehat{G}_k)\text{-Trans-GLM}]$, with $\{\hat{w}_k\}_{k=0}^{K}$ solved from a variant of Lasso
- See Li et al. (2021) for details.

[1] Rigollet, P., & Tsybakov, A. B. (2012). Sparse Estimation by Exponential Weighting. Statistical Science, 27(4), 558-575.

[2] Dai, D., Rigollet, P., & Zhang, T. (2012). Deviation optimal learning using greedy q-aggregation. Annals of Statistics, 40(3), 1878-1905.

[3] Li, S., Cai, T. T., & Li, H. (2022). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. Journal of the Royal Statistical Society Series B: Statistical Methodology, 84(1), 149-173.

When S is unknown: selection

- Selection technique originates from the diagnosis in outlier detection (Cook, 2000; Belsley et al., 2005; Kwon and Zou, 2022).
- Main idea of informative source selection: evaluate source data quality by *the likelihood on target data* or *the distance between target-only estimator and source-only estimator*
  - Split the target data into two folds $\mathcal{D}_1^{(0)}$ and $\mathcal{D}_2^{(0)}$
  - Fit local estimators on target dataset $\mathcal{D}_1^{(0)}$ and each source dataset $\Rightarrow \widehat{\boldsymbol{\theta}}^{(k)}$
  - Calculate the likelihood of $\mathcal{D}_2^{(0)}$ based on each $\widehat{\boldsymbol{\theta}}^{(k)} \Rightarrow \widehat{R}^{(k)}$
  - Threshold and select the informative source by

$$\widehat{S} = \{1 \leq k \leq K : \widehat{R}^{(k)} - \widehat{R}^{(0)} \leq \text{ threshold}\}$$

  - Run $\widehat{S}$-GLM-Trans
- The above is the likelihood-based version in Tian and Feng (2023b). A cleaner distance-based version can be found in Li et al. (2024).

---

[1] Cook, R. D. (2000). Detection of influential observation in linear regression. Technometrics, 42(1), 65-68.

[2] Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). Regression diagnostics: Identifying influential data and sources of collinearity. John Wiley & Sons.

[3] Kwon, Y., & Zou, J. (2022, May). Beta Shapley: a Unified and Noise-reduced Data Valuation Framework for Machine Learning. In International Conference on Artificial Intelligence and Statistics (pp. 8780-8802). PMLR.

When S is unknown

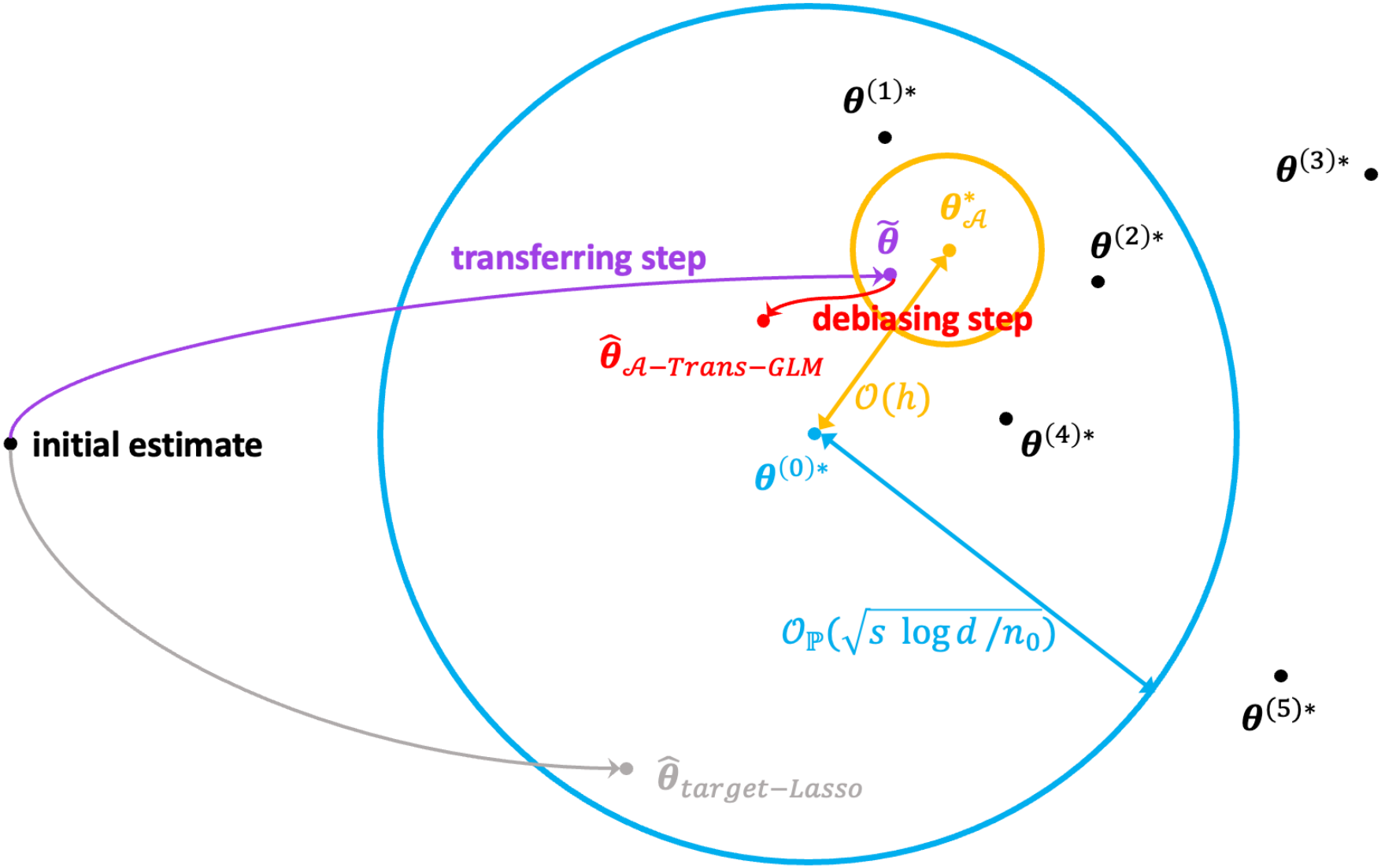
Given certain conditions, both aggregation and selection can guarantee that

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{(0)*}\|_2 \lesssim_{\mathbb{P}} \sqrt{\frac{s \log d}{(|S|+1)n}} + \left(\sqrt{\frac{\log d}{n}} h^{1/2}\right) \wedge h \wedge \sqrt{\frac{s \log d}{n}},$$

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{(0)*}\|_1 \lesssim_{\mathbb{P}} s\sqrt{\frac{\log d}{(|S|+1)n}} + h \wedge \left(s\sqrt{\frac{\log d}{n}}\right).$$

This finally matches with the lower bound and makes our two-step algorithm free of negative transfer.

ℓ_1-penalty with GLMs: overall review



Picture is adapted from: Tian, Y., & Feng, Y. (2023). Transfer learning under high-dimensional generalized linear models. Journal of the American Statistical Association, 118(544), 2684-2697.

ℓ_1-penalty with GLMs: history and other literature

Early explorations of ℓ_1-based regularization date back to ~ 10 years ago.

- Gross and Tibshirani (2016) studies the stratified linear model $y_i = \boldsymbol{x}_i^\top \boldsymbol{\theta}^*_{g_i} + \epsilon_i$, $g_i \in 1:K$, $i = 1:n$, with *data shared Lasso*:

$$\tilde{\boldsymbol{\theta}}, \{\widehat{\boldsymbol{\Delta}}_k\}_{k=1}^K = \underset{\boldsymbol{\theta}, \{\boldsymbol{\Delta}_k\}_{k=1}^K}{\arg\min} \left\{ \frac{1}{2} \sum_{i=1}^n [y_i - \boldsymbol{x}_i^\top (\boldsymbol{\theta} + \boldsymbol{\Delta}_{g_i})]^2 + \lambda \|\boldsymbol{\theta}\|_1 + \sum_{k=1}^K \lambda_k \|\boldsymbol{\Delta}_k\|_1 \right\}$$

- Ollier and Viallon (2017) studies the same model with some theory on variable selection consistency under strong conditions (e.g. irrepresentative condition)

[1] Gross, S. M., & Tibshirani, R. (2016). Data Shared Lasso: A novel tool to discover uplift. Computational statistics & data analysis, 101, 226-235.

[2] Ollier, E., & Viallon, V. (2017). Regression modelling on stratified data with the lasso. Biometrika, 104(1), 83-96.

ℓ_1-penalty with GLMs: history and other literature

- Bastani (2021) studies a single-source transfer learning problem on linear model

$$y_i^{(k)} = (\boldsymbol{x}_i^{(k)})^\top \boldsymbol{\theta}^{(k)*} + \epsilon_i^{(k)}, \ k = 0:1, \ i = 1:n_k,$$

where $n_0 \ll n_1$, and proposes a similar two-step approach.

  - ▷ Step 1: $\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min} \frac{1}{n_1} \sum_{i=1}^{n_1} [y_i^{(1)} - (\boldsymbol{x}_i^{(1)})^\top \boldsymbol{\theta}]^2$
  - ▷ Step 2: $\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min} \frac{1}{n_0} \sum_{i=1}^{n_0} [y_i^{(0)} - (\boldsymbol{x}_i^{(0)})^\top \boldsymbol{\theta}]^2 + \lambda \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_1$

They assume $\|\boldsymbol{\theta}^{(1)*} - \boldsymbol{\theta}^{(0)*}\|_0 \le s \ll d$ and $\boldsymbol{\theta}^{(0)*}$ can be dense. They show that

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{(0)*}\|_1 \lesssim_{\mathbb{P}} \frac{s\log(dn_0)}{\sqrt{n_0}} + \frac{sd\log(dn_1)}{\sqrt{n_1}} \ll \underbrace{\frac{d}{\sqrt{n_0}}}_{\text{target-only OLS}},$$

when $n_1 \gg n_0 s^2 \log^2(dn_1)$, $d \gg s\log(dn_0)$.

[1] Bastani, H. (2021). Predicting with proxies: Transfer learning in high dimension. Management Science, 67(5), 2964-2984.

§4.3: Extension to high-dimensional regressions

- §4.3.1 ℓ_1-penalty with GLMs
- §4.3.2 Go robust: better ways to aggregate data
- §4.3.3 Block penalty with multi-task learning

Issues of the previous two-step approach

- Recall our previous linear regression setting: $\{(\boldsymbol{x}_i^{(k)}, y_i^{(k)})\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} (X^{(k)}, Y^{(k)})$ for $k = 0 : K$, where $X^{(k)} \in \mathbb{R}^d$, $Y^{(k)} \in \mathbb{R}$, and

$$Y^{(k)} = (X^{(k)})^\top \boldsymbol{\theta}^{(k)*} + \epsilon^{(k)},$$

  $\mathbb{E}[X^{(k)}(X^{(k)})^\top] = \boldsymbol{\Sigma}^{(k)}$, $\epsilon^{(k)}$ is independent of $X^{(k)}$ and zero-mean sub-Gaussian.
- Recall the first transferring step of our algorithm (consider the case $S = [K]$):

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\arg\min} \left\{ \frac{1}{(K+1)} \sum_{k=0}^{K} \frac{1}{n} \sum_{i=1}^{n} (y_i^{(k)} - \boldsymbol{\theta}^\top \boldsymbol{x}_i^{(k)})^2 + \lambda_1 \|\boldsymbol{\theta}\|_1 \right\}$$

- In the second debiasing step, we penalize the bias $\|\boldsymbol{\theta}^{(0)} - \tilde{\boldsymbol{\theta}}\|_1$ to learn $\boldsymbol{\Delta}^{(k)*} = \boldsymbol{\theta}^{(0)*} - \tilde{\boldsymbol{\theta}}^*$, where $\tilde{\boldsymbol{\theta}} \xrightarrow{\mathbb{P}} \tilde{\boldsymbol{\theta}}^*$ when $n \to \infty$.
- **An underlying assumption:** $\boldsymbol{\Delta}^{(k)*}$ is "sparse" in some sense so that the debiasing step can succeed

Issues of the previous two-step approach

If we ignore the regularizer, then our transferring step is equivalent to **data pooling**:

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min} \left\{ \frac{1}{(K+1)} \sum_{k=0}^{K} \frac{1}{n} \sum_{i=1}^{n} (y_i^{(k)} - \boldsymbol{\theta}^\top \boldsymbol{x}_i^{(k)})^2 \right\}.$$

From the population-level, we are estimating

$$\tilde{\boldsymbol{\theta}}^* = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min} \left\{ \frac{1}{(K+1)} \sum_{k=0}^{K} \mathbb{E}(Y^{(k)} - \boldsymbol{\theta}^\top X^{(k)})^2 \right\} = \left(\sum_{k=0}^{K} \boldsymbol{\Sigma}^{(k)} \right)^{-1} \sum_{k=0}^{K} \boldsymbol{\Sigma}^{(k)} \boldsymbol{\theta}^{(k)*}.$$

Therefore the actual bias is

$$\boldsymbol{\theta}^{(0)*} - \tilde{\boldsymbol{\theta}}^* = \left(\sum_{k=0}^{K} \boldsymbol{\Sigma}^{(k)} \right)^{-1} \sum_{k=1}^{K} \boldsymbol{\Sigma}^{(k)} (\boldsymbol{\theta}^{(k)*} - \boldsymbol{\theta}^{(0)*}).$$

- **Problem:** In general, there is no guarantee that this bias would be "sparse" in any sense!
- Previous we don't have this issue because we assume $\boldsymbol{\Sigma}^{(k)} \equiv \boldsymbol{\Sigma}$ for all k

Issues of the previous two-step approach

- Bias of the transferring step:

$$\boldsymbol{\theta}^{(0)*} - \tilde{\boldsymbol{\theta}}^* = \left(\sum_{k=0}^{K} \boldsymbol{\Sigma}^{(k)}\right)^{-1} \sum_{k=1}^{K} \boldsymbol{\Sigma}^{(k)} (\boldsymbol{\theta}^{(k)*} - \boldsymbol{\theta}^{(0)*}).$$

- **Problem:** In general, there is no guarantee that this bias would be "sparse" in any sense!
- Previous we don't have this issue because we assume $\boldsymbol{\Sigma}^{(k)} \equiv \boldsymbol{\Sigma}$ for all k. Then under the assumption $\max_{k\in[K]} \|\boldsymbol{\theta}^{(k)*} - \boldsymbol{\theta}^{(0)*}\|_1 \leq h$, we have $\|\boldsymbol{\theta}^{(0)*} - \tilde{\boldsymbol{\theta}}^*\|_1 \leq h$
- If our similarity assumption is of ℓ_0-pseudo norm, in the sense that $\max_{k\in[K]} \|\boldsymbol{\theta}^{(k)*} - \boldsymbol{\theta}^{(0)*}\|_0 \leq h$, then:
  - In general, $\boldsymbol{\theta}^{(0)*} - \tilde{\boldsymbol{\theta}}^*$ could be dense in the sense that $\|\boldsymbol{\theta}^{(0)*} - \tilde{\boldsymbol{\theta}}^*\|_0 \asymp d$
  - Even when $\boldsymbol{\Sigma}^{(k)} \equiv \boldsymbol{\Sigma}$ for all k, $\|\boldsymbol{\theta}^{(0)*} - \tilde{\boldsymbol{\theta}}^*\|_0$ could still be as large as Kh!

Let's formulate the problem into a MTL framework and see how we can solve it.

Multi-task linear regression

Multi-task linear regression (Xu and Bastani, 2021):

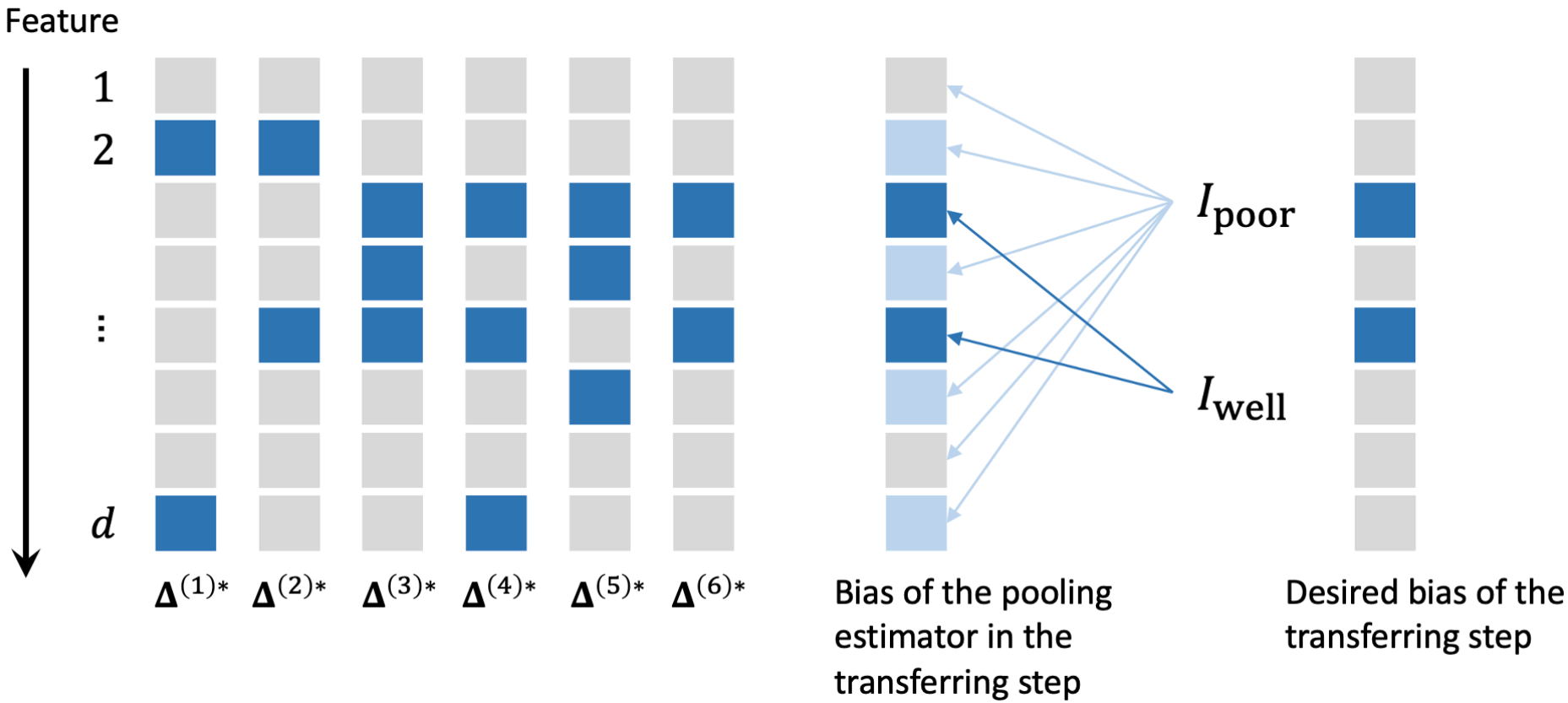
- $\{(\boldsymbol{x}_i^{(k)}, y_i^{(k)})\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} (X^{(k)}, Y^{(k)})$ for $k = 1 : K$, where $X^{(k)} \in \mathbb{R}^d$, $Y^{(k)} \in \mathbb{R}$, and

$$Y^{(k)} = (X^{(k)})^\top \boldsymbol{\theta}^{(k)*} + \epsilon^{(k)},$$

  $\mathbb{E}[X^{(k)}(X^{(k)})^\top] = \boldsymbol{\Sigma}^{(k)}$, $\epsilon^{(k)}$ is independent of $X^{(k)}$ and zero-mean sub-Gaussian.
- **Decomposition:** $\boldsymbol{\theta}^{(k)*} = \boldsymbol{\theta}^* + \boldsymbol{\Delta}^{(k)*}$
- ℓ_0**-similarity:** $\max_{k\in[K]} \|\boldsymbol{\Delta}^{(k)*}\|_0 \leq s$, $\boldsymbol{\theta}^*$ can be dense
- **Goal:** Learn all $\boldsymbol{\theta}^{(k)*}$'s simultaneously and borrow information to perform better than *single-task estimators*

[1] Xu, K., & Bastani, H. (2021). Learning across bandits in high dimension via robust statistics. arXiv preprint arXiv:2112.14233, 52(7).

Multi-task linear regression



- $\boldsymbol{\theta}^{(k)*} = \boldsymbol{\theta}^* + \boldsymbol{\Delta}^{(k)*}$
- If we can void out the "poorly aligned" features, then we only need to debias $|I_{\text{well}} \cup \text{supp}(\boldsymbol{\Delta}^{(k)})| \lesssim s$ coordinates!

Picture adapted from: Xu, K., & Bastani, H. (2021). Learning across bandits in high dimension via robust statistics. arXiv preprint arXiv:2112.14233, 52(7).

Multi-task linear regression

Xu and Bastani (2021) proposes to use coordinate-wise trimmed mean as the global estimation in the transferring step

Two-step algorithm with trimmied mean: (Xu and Bastani, 2021)

- Step 1: (Single-task OLS)

$$\tilde{\boldsymbol{\theta}}^{(k)} = \text{OLS on data } \{\boldsymbol{x}_i^{(k)}, y_i^{(k)}\}_{i=1}^n \text{ from the } k\text{-th task}$$

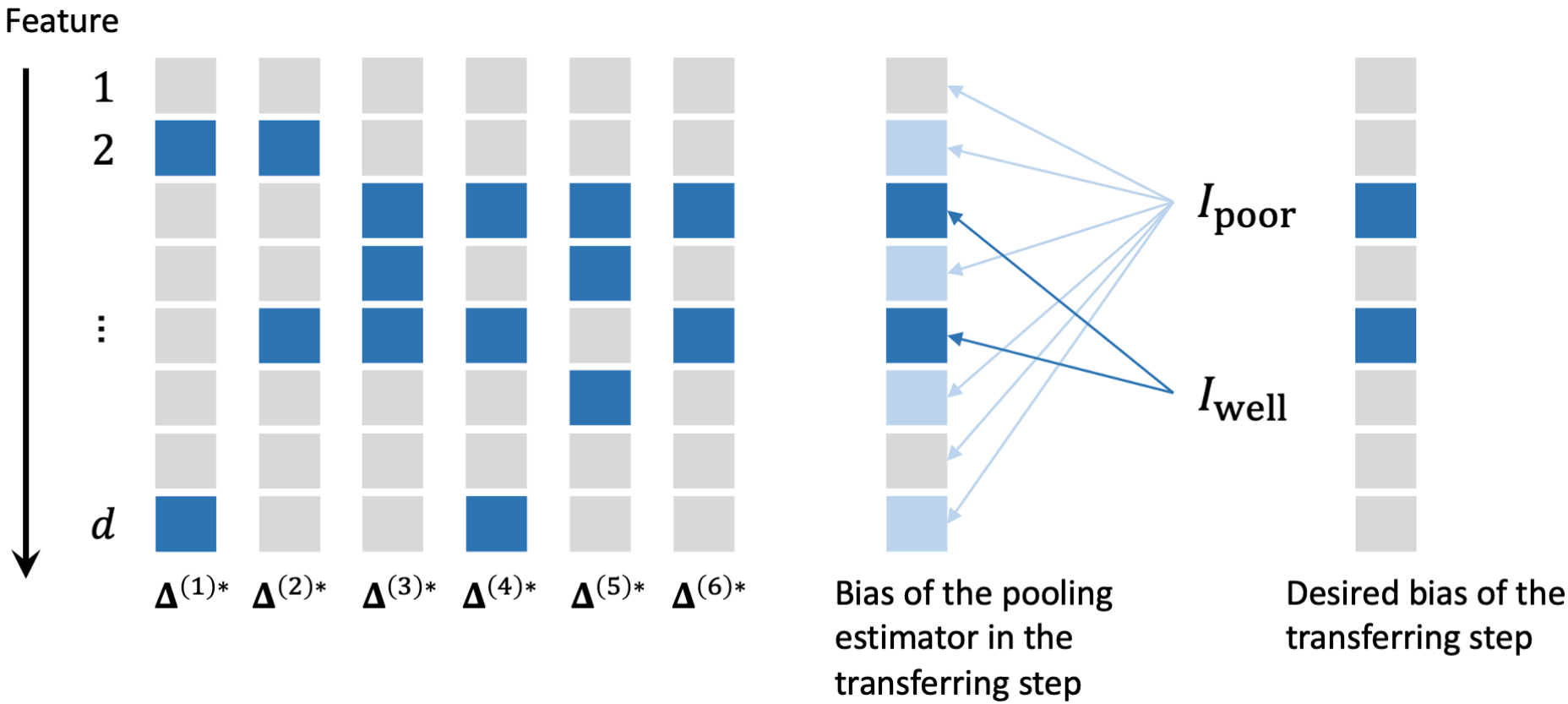
- Step 2: (Transferring)

$$\tilde{\boldsymbol{\theta}} = \text{coordinate-wise trimmed mean of } \{\tilde{\boldsymbol{\theta}}^{(k)}\}_{k=1}^K \text{ with trimming proportion } w$$

- Step 3: (Debiasing) Debias $\tilde{\boldsymbol{\theta}}$ for each task using by **penalization**:

$$\widehat{\boldsymbol{\theta}}^{(k)} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min} \left\{ \frac{1}{n}\sum_{i=1}^n (y_i^{(k)} - \boldsymbol{\theta}^\top \boldsymbol{x}_i^{(k)})^2 + \lambda \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_1 \right\}, \quad k = 1:K.$$

Multi-task linear regression: intuition revisited



- Trimmed mean can "zero out" the "poorly aligned" features in the transferring step, then we only need to debias $|I_{\text{well}} \cup \text{supp}(\boldsymbol{\Delta}^{(k)})| \lesssim s$ coordinates!

Picture adapted from: Xu, K., & Bastani, H. (2021). Learning across bandits in high dimension via robust statistics. arXiv preprint arXiv:2112.14233, 52(7).

Multi-task linear regression with trimmed mean: theory

Theorem 4.3.1 (Xu and Bastani, 2021)

Under certain conditions, let $w \asymp \sqrt{s/d}$ *and* $\lambda \asymp \sqrt{\frac{\log d}{n}}$*, then up to a logarithmic factor:*

$$\max_{k\in[K]} \|\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{(k)*}\|_1 \lesssim_{\mathbb{P}} \sqrt{\frac{sd}{n}} + d\sqrt{\frac{1}{nK}} \ll \underbrace{d\sqrt{\frac{1}{n}}}_{\text{single-task error}} . \quad (\star)$$

- Compared to the rate $s\sqrt{\frac{1}{n_0}} + sd\sqrt{\frac{1}{n_1}}$ obtained by Bastani (2021) for the case $K = 2$ in a transfer learning context, the second term in $(\star)$ is better while the first term is worse.
- The minimax rate is proved to be $s\sqrt{\frac{1}{n}} + d\sqrt{\frac{1}{nK}}$ and achieved by a coordinate-wise median transferring step (Huang et al., 2023)
- The original paper (Xu and Bastani, 2021) applies the method to a multi-armed contextual bandit problem.

[1] Xu, K., & Bastani, H. (2021). Learning across bandits in high dimension via robust statistics. arXiv preprint arXiv:2112.14233, 52(7).

[2] Bastani, H. (2021). Predicting with proxies: Transfer learning in high dimension. Management Science, 67(5), 2964-2984.

[3] Huang, X., Xu, K., Lee, D., Hassani, H., Bastani, H., & Dobriban, E. (2023). Optimal Heterogeneous Collaborative Linear Regression and Contextual Bandits. arXiv preprint arXiv:2306.06291.

Multi-task linear regression: other aggregation methods

- Maity et al. (2022) discusses two other options as the transferring step. They propose to use single-task debiased Lasso estimators as $\{\tilde{\boldsymbol{\theta}}^{(k)}\}_{k=1}^K$
  - ▷ A re-descending loss: $\tilde{\theta}_j = \arg\min_{\theta\in\mathbb{R}} \sum_{k=1}^K \Psi_{\eta_j}(\theta_j^{(k)} - \theta)$ for $j = 1:d$, where $\Psi_\eta(x) = x^2 \wedge \eta^2$.
  - ▷ Quadratic + ℓ_1 loss: $\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \{ \sum_{k=1}^K \frac{1}{1+\lambda}(\lambda\|\tilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}\|_1 + \frac{1}{2}\|\tilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}\|_2)\}$
- They use coordinate-wise hard/soft-thresholding in the debiasing step to obtain $\widehat{\boldsymbol{\theta}}^{(k)}$ instead of penalization (approx. equiv. to hard-thresholding/ℓ_1 penalty).
- With certain conditions [3], they show the following ℓ_∞ error bound.

Theorem 4.3.2 (Maity et al., 2022)

Up to logarithmic factors, we have

$$\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \lesssim_{\mathbb{P}} \sqrt{\frac{\log d}{nK}}, \quad \max_{k=1:K} \|\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{(k)*}\|_\infty \lesssim_{\mathbb{P}} \sqrt{\frac{\log d}{n}}.$$

[3] We need assumptions to make the global parameter $\boldsymbol{\theta}^*$ identifiable. Recall that $\boldsymbol{\theta}^{(k)*} = \boldsymbol{\theta}^* + \boldsymbol{\Delta}^{(k)*}$.

[1] Maity, S., Sun, Y., & Banerjee, M. (2022). Meta-analysis of heterogeneous data: integrative sparse regression in high-dimensions. Journal of Machine Learning Research, 23(198), 1-50.

§4.3: Extension to high-dimensional regressions

- §4.3.1 ℓ_1-penalty with GLMs
- §4.3.2 Go robust: better ways to aggregate data
- §4.3.3 Block penalty with multi-task learning

Block regularization

Definition 4.3.1

We define the $L_{p,q}$ or ℓ_p/ℓ_q block norm ($1 \leq p, q \leq \infty$) of a matrix $\boldsymbol{B} \in \mathbb{R}^{m \times n}$ as

$$\|\boldsymbol{B}\|_{p,q} = \left(\sum_{i=1}^{m} \|\boldsymbol{b}_i\|_q^p \right)^{1/p} = \left[\sum_{i=1}^{m} \left(\sum_{j=1}^{n} |b_{ij}|^q \right)^{p/q} \right]^{1/p},$$

where $\boldsymbol{b}_i$ is the i-th row of $\boldsymbol{B}$.

Some examples:

- Group Lasso penalty (Yuan and Lin, 2006): $p = 1, q = 2$

$$\|\boldsymbol{B}\|_{1,2} = \sum_{i=1}^{m} \|\boldsymbol{b}_i\|_2.$$

- ℓ_1/ℓ_∞-penalty (Negahban and Wainwright, 2011): $p = 1, q = \infty$

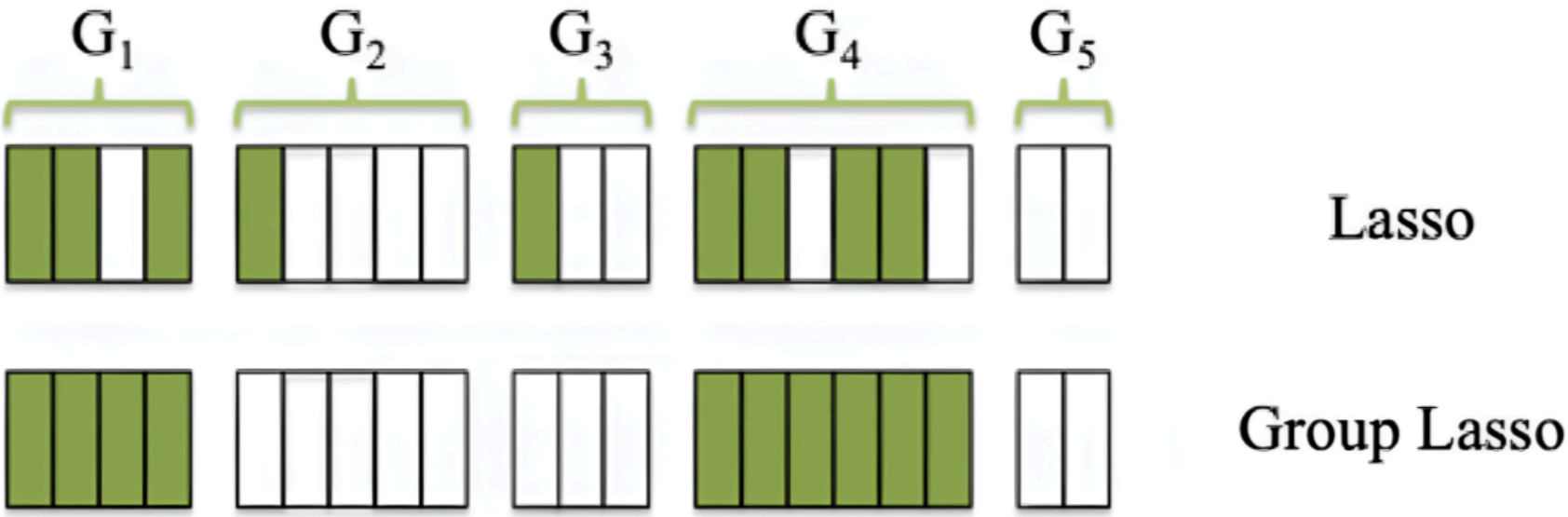
$$\|\boldsymbol{B}\|_{1,\infty} = \sum_{i=1}^{m} \max_{j=1:n} |b_{ij}|.$$

[1] Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society Series B: Statistical Methodology, 68(1), 49-67.

[2] Negahban, S. N., & Wainwright, M. J. (2011). Simultaneous Support Recovery in High Dimensions: Benefits and Perils of Block ℓ_1/ℓ_∞-Regularization. IEEE Transactions on Information Theory, 57(6), 3841-3863.

Block regularization

Block regularization is used in high-dimensional statistics when there is some "block/group" structure.



We can utilize this block penalty for bias regularization in multi-task learning.

Image source: Bai, Y., Calhoun, V. D., & Wang, Y. P. (2020, February). Integration of multi-task fmri for cognitive study by structure-enforced collaborative regression. In Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging (Vol. 11317, pp. 515-520). SPIE.

Group Lasso with multi-task learning

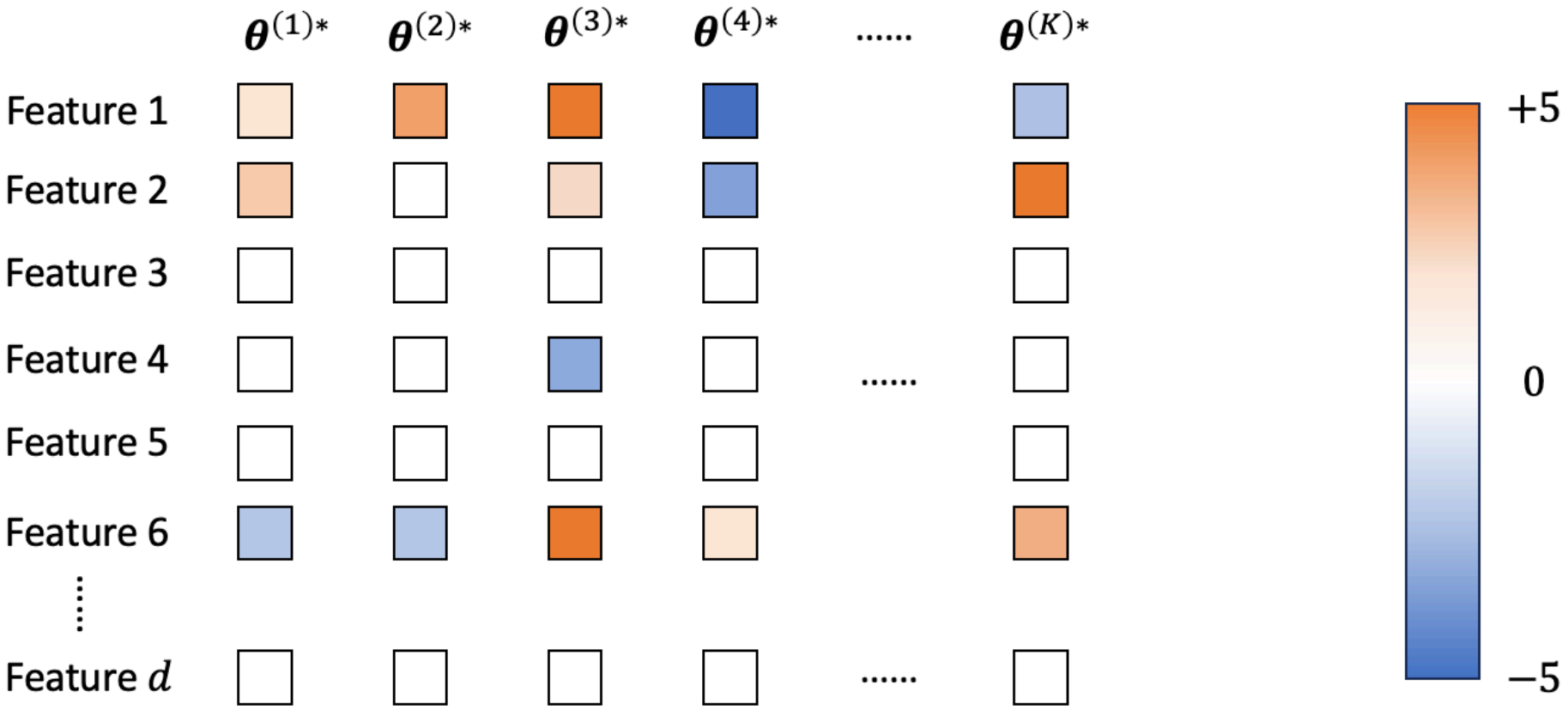
Let's consider the linear regression setting we considered before, but in an MTL framework (Lounici et al., 2011).

- We observe dataset $\mathcal{D}^{(k)} = \{\boldsymbol{x}_i^{(k)}, y_i^{(k)}\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} (X^{(k)}, Y^{(k)})$, where $X^{(k)} \in \mathbb{R}^d$, $Y^{(k)} \in \mathbb{R}$, and
$$Y^{(k)} = (X^{(k)})^\top \boldsymbol{\theta}^{(k)*} + \epsilon^{(k)}.$$
- Denote the coefficient matrix $\boldsymbol{\Theta}^* = (\boldsymbol{\theta}^{(1)*}, \ldots, \boldsymbol{\theta}^{(K)*}) \in \mathbb{R}^{d \times K}$, where $\boldsymbol{\theta}^{(k)*}$ is the k-th column of $\boldsymbol{\Theta}^*$. Denote the j-th row of $\boldsymbol{\Theta}^*$ as $\boldsymbol{\theta}_j^*$.
- **Sparsity:** $S := \{j \in [d] : \boldsymbol{\theta}_j^* \neq \mathbf{0}_K\}$, $|S| \leq s$.
- **Intuition:** The support $\text{supp}(\boldsymbol{\theta}^{(k)*})$ overlaps a lot across tasks, but the values of the same coordinate can differ.
- **What we expect:**
  - The simultaneous sparsity could help, because it might be easier for variable selection
  - But the estimation error for each task may not improve a lot due to heterogeneity

[1] Tsybakov, A. B., Lounici, K., Pontil, M., & van de Geer, S. (2011). Oracle inequalities and optimal inference under group sparsity. Annals of Statistics, 39(4), 2164-2204.

Group Lasso with multi-task learning

Sparsity: $S := \{j \in [d] : \boldsymbol{\theta}_j^* \neq \mathbf{0}_K\},\ |S| \leq s.$



$S = \{1, 2, 4, 6\}, \qquad s = 4$

Group Lasso with multi-task learning

Lounici et al. (2011) proposes to use group Lasso regularization

$$\widehat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}\in\mathbb{R}^{d\times K}}{\arg\min} \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n}[(\boldsymbol{x}_i^{(k)})^\top\boldsymbol{\theta}^{(k)} - y_i^{(k)}]^2 + \lambda\|\boldsymbol{\Theta}\|_{1,2},$$

where $\boldsymbol{\theta}^{(k)}$ and $\boldsymbol{\theta}_j$ represent the k-th column and j-th row of $\boldsymbol{\Theta}$, and

$$\|\boldsymbol{\Theta}\|_{1,2} = \sum_{j=1}^{d}\|\boldsymbol{\theta}_j\|_2.$$

Under certain conditions (regular eigenvalues of covariance matrices etc.), we have the following result.

Theorem 4.3.2 (Lounici et al., 2011)

Let $\lambda \asymp \frac{1}{\sqrt{nK}}\sqrt{1+\frac{\log d}{K}}$, *then*

$$\frac{1}{T}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathsf{F}}^2 \lesssim_{\mathbb{P}} \frac{s}{n}\left(1+\frac{\log d}{K}\right)$$

[1] Tsybakov, A. B., Lounici, K., Pontil, M., & van de Geer, S. (2011). Oracle inequalities and optimal inference under group sparsity. Annals of Statistics, 39(4), 2164-2204.

Group Lasso with multi-task learning

Let's compare the result with single-task Lasso.

- Single-task Lasso: $\frac{1}{T}\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_{\mathsf{F}}^2 \lesssim_{\mathbb{P}} \frac{s \log d}{n}$
- Group Lasso regularization: $\frac{1}{T}\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_{\mathsf{F}}^2 \lesssim_{\mathbb{P}} \frac{s}{n}\left(1 + \frac{\log d}{K}\right)$

Our previous intuition is correct:

- Regularization helps, but we do not achieve big improvement.
- When $K \gtrsim \log d$, we completely get rid of the full dimension d by using group Lasso regularization.

ℓ_1/ℓ_∞-penalty with multi-task learning

Besides Group Lasso penalty, Negahban and Wainwright (2011) explores the following ℓ_1/ℓ_∞-regularization in the same problem.

$$\widehat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}\in\mathbb{R}^{d\times K}}{\arg\min} \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n}[(\boldsymbol{x}_i^{(k)})^\top\boldsymbol{\theta}^{(k)} - y_i^{(k)}]^2 + \lambda\|\boldsymbol{\Theta}\|_{1,\infty},$$

where $\boldsymbol{\theta}^{(k)}$ and $\boldsymbol{\theta}_j$ represent the k-th column and j-th row of $\boldsymbol{\Theta}$, and

$$\|\boldsymbol{\Theta}\|_{1,\infty} = \sum_{j=1}^{d}\max_{k=1:K}|\theta_{jk}|.$$

Under certain conditions (irrepresentative condition, minimum signal strength etc. for variable selection consistency), we have the following results.

[1] Negahban, S. N., & Wainwright, M. J. (2011). Simultaneous Support Recovery in High Dimensions: Benefits and Perils of Block ℓ_1/ℓ_∞-Regularization. IEEE Transactions on Information Theory, 57(6), 3841-3863.

ℓ_1/ℓ_∞-penalty with multi-task learning

Theorem 4.3.3 (Negahban and Wainwright, 2011)

Let $\lambda \asymp \sqrt{\frac{K^2+K\log d}{n}}$, *then under Gaussian design, when* $n \gtrsim sK(K+\log d)$*:*

- $\text{supp}(\widehat{\boldsymbol{\Theta}}) = \text{supp}(\boldsymbol{\Theta}^*)$ *w.h.p.*
- $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\max} \lesssim_{\mathbb{P}} \sqrt{\frac{K^2+K\log d}{n}}$

The following phase transition result highlights the benefit of regularization in MTL setting. Consider the case $K = 2$, $|\text{supp}(\boldsymbol{\theta}^{(1)*})| = |\text{supp}(\boldsymbol{\theta}^{(2)*})| = s$, and the "overlap proportion" $\alpha = |\text{supp}(\boldsymbol{\theta}^{(1)*}) \cap \text{supp}(\boldsymbol{\theta}^{(2)*})|/s$.

Theorem 4.3.4 (Negahban and Wainwright, 2011)

We have the following phase transition when $\max_{j \in \text{supp}(\boldsymbol{\theta}^{(1)*}) \cap \text{supp}(\boldsymbol{\theta}^{(2)*})} |\theta_j^{(1)*} - \theta_j^{(2)*}|$ $\ll \lambda$*:*

- *(Success) When* $\frac{n}{s\log(d-(2-\alpha)s)} > 4 - 3\alpha$, *results in Theorem 4.3.3 hold.*
- *(Failure) When* $\frac{n}{s\log(d-(2-\alpha)s)} < 4 - 3\alpha$, *no* λ *can make* $\text{supp}(\widehat{\boldsymbol{\Theta}}) = \text{supp}(\boldsymbol{\Theta}^*)$.

[1] Negahban, S. N., & Wainwright, M. J. (2011). Simultaneous Support Recovery in High Dimensions: Benefits and Perils of Block ℓ_1/ℓ_∞-Regularization. IEEE Transactions on Information Theory, 57(6), 3841-3863.

ℓ_1/ℓ_∞-penalty with multi-task learning

- The phase transition of ℓ_1/ℓ_∞ regularization happens at

$$\frac{n}{s\log(d-(2-\alpha)s)} = 4 - 3\alpha.$$

- Lasso has similar phase transition phenomenon (Wainwright, 2009) which happens at

$$\frac{n}{s\log(d-s)} = 2.$$

When $d \gg s$, the LHS are almost the same. Then we can make the following conclusion:

- When $\alpha < 2/3$ (less sharing), Lasso performs better in the sense that its transition is at a smaller sample size
- When $\alpha \in (2/3, 1]$ (more sharing), ℓ_1/ℓ_∞ regularization performs better in the sense that its transition is at a smaller sample size

[1] Wainwright, M. J. (2009). Sharp thresholds for High-Dimensional and noisy sparsity recovery using ℓ_1-Constrained Quadratic Programming (Lasso). IEEE transactions on information theory, 55(5), 2183-2202.

ℓ_1/ℓ_∞-penalty with multi-task learning

To further fix the inferiority of ℓ_1/ℓ_∞ regularization in Negahban and Wainwright (2011) compared to Lasso when there are not a lot of support overlaps across tasks, Jalali et al. (2010, 2013) propose the following variant of ℓ_1/ℓ_∞ regularization in the same MTL setting.

$$\widehat{\boldsymbol{S}}, \widehat{\boldsymbol{B}} = \underset{\boldsymbol{S}, \boldsymbol{B} \in \mathbb{R}^{d \times K}}{\arg\min} \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{n} [(\boldsymbol{x}_i^{(k)})^\top \boldsymbol{\theta}^{(k)} - y_i^{(k)}]^2 + \lambda_{\boldsymbol{S}} \|\boldsymbol{S}\|_{1,1} + \lambda_{\boldsymbol{B}} \|\boldsymbol{B}\|_{1,\infty},$$

$$\widehat{\boldsymbol{\Theta}} = \widehat{\boldsymbol{S}} + \widehat{\boldsymbol{B}}.$$

This can be seen as a combination of Lasso and the ℓ_1/ℓ_∞ regularization in Negahban and Wainwright (2011).

- Push $\lambda_{\boldsymbol{B}} \to \infty$: we will force $\boldsymbol{B} = \boldsymbol{0}$ and obtain K Lassos.
- Push $\lambda_{\boldsymbol{S}} \to \infty$: we will force $\boldsymbol{S} = \boldsymbol{0}$ and recover the ℓ_1/ℓ_∞ regularization in Negahban and Wainwright (2011).

ℓ_1/ℓ_∞-penalty with multi-task learning

Under similar conditions as before, we have the following results.

Theorem 4.3.5 (Jalali et al., 2010, 2013)

Let $\lambda_{\boldsymbol{S}} \asymp \sqrt{\frac{\log d}{n}}$ *and* $\lambda_{\boldsymbol{B}} \asymp \sqrt{\frac{r(r+\log d)}{n}}$, *then under Gaussian design, when* $n \gtrsim s\log(dK) + sK(K+\log d)$*:*

- $\mathrm{supp}(\widehat{\boldsymbol{\Theta}}) = \mathrm{supp}(\boldsymbol{\Theta}^*)$ *w.h.p.*
- $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\max} \lesssim_{\mathbb{P}} \sqrt{\frac{\log(dK)}{n}}$

The $\max$-estimation error rate is better than the rate $\sqrt{\frac{K^2+K\log d}{n}}$ by ℓ_1/ℓ_∞ regularization in Negahban and Wainwright (2011).

Next, let's look at the phase transition.

---

[1] Jalali, A., Sanghavi, S., Ruan, C., & Ravikumar, P. (2010). A dirty model for multi-task learning. Advances in neural information processing systems, 23.

[2] Jalali, A., Ravikumar, P., & Sanghavi, S. (2013). A dirty model for multiple sparse regression. IEEE Transactions on Information Theory, 59(12), 7947-7968.

[3] Negahban, S. N., & Wainwright, M. J. (2011). Simultaneous Support Recovery in High Dimensions: Benefits and Perils of Block ℓ_1/ℓ_∞-Regularization. IEEE Transactions on Information Theory, 57(6), 3841-3863.

ℓ_1/ℓ_∞-penalty with multi-task learning

Consider the case $K = 2$, $|\text{supp}(\boldsymbol{\theta}^{(1)*})| = |\text{supp}(\boldsymbol{\theta}^{(2)*})| = s$, and the "overlap proportion" $\alpha = |\text{supp}(\boldsymbol{\theta}^{(1)*}) \cap \text{supp}(\boldsymbol{\theta}^{(2)*})|/s$.

Theorem 4.3.6 (Jalali et al., 2010, 2013)

We have the following phase transition when $\max_{j \in \text{supp}(\boldsymbol{\theta}^{(1)*}) \cap \text{supp}(\boldsymbol{\theta}^{(2)*})} |\theta_j^{(1)*} - \theta_j^{(2)*}| \ll \lambda_{\boldsymbol{S}}$*:*

- *(Success) When* $\frac{n}{s \log(d-(2-\alpha)s)} > 2 - \alpha$*, results in Theorem 4.3.5 hold.*
- *(Failure) When* $\frac{n}{s \log(d-(2-\alpha)s)} < 2 - \alpha$*, no* λ *can make* $\text{supp}(\widehat{\boldsymbol{\Theta}}) = \text{supp}(\boldsymbol{\Theta}^*)$.

The transition point $2 - \alpha$ is better than $4 - 3\alpha$ by ℓ_1/ℓ_∞ regularization in Negahban and Wainwright (2011).

Let us make a more comprehensive summary.

ℓ_1/ℓ_∞-penalty with multi-task learning

- The phase transition of ℓ_1/ℓ_∞ regularization in Negahban and Wainwright (2011) happens at

$$\frac{n}{s\log(d-(2-\alpha)s)} = 4 - 3\alpha.$$

- Lasso has similar phase transition phenomenon (Wainwright, 2009) which happens at

$$\frac{n}{s\log(d-s)} = 2.$$

- The phase transition of ℓ_1/ℓ_∞ regularization variant in Jalali et al. (2010, 2013) happens at

$$\frac{n}{s\log(d-(2-\alpha)s)} = 2 - \alpha.$$

[1] Jalali, A., Sanghavi, S., Ruan, C., & Ravikumar, P. (2010). A dirty model for multi-task learning. Advances in neural information processing systems, 23.

[2] Jalali, A., Ravikumar, P., & Sanghavi, S. (2013). A dirty model for multiple sparse regression. IEEE Transactions on Information Theory, 59(12), 7947-7968.

[3] Negahban, S. N., & Wainwright, M. J. (2011). Simultaneous Support Recovery in High Dimensions: Benefits and Perils of Block ℓ_1/ℓ_∞-Regularization. IEEE Transactions on Information Theory, 57(6), 3841-3863.

ℓ_1/ℓ_∞-penalty with multi-task learning

When $d \gg s$, the LHS are almost the same. Then we can conclude as follows:

- When $\alpha \in (0, 1)$, the ℓ_1/ℓ_∞ regularization variant in Jalali et al. (2010, 2013) performs the best
- When $\alpha = 0$ (zero sharing), the ℓ_1/ℓ_∞ regularization variant in Jalali et al. (2010, 2013) performs similarly as Lasso
- When $\alpha = 1$ (full sharing), the ℓ_1/ℓ_∞ regularization variant in Jalali et al. (2010, 2013) performs similarly as ℓ_1/ℓ_∞ regularization in Negahban and Wainwright (2011)

---

[1] Jalali, A., Sanghavi, S., Ruan, C., & Ravikumar, P. (2010). A dirty model for multi-task learning. Advances in neural information processing systems, 23.

[2] Jalali, A., Ravikumar, P., & Sanghavi, S. (2013). A dirty model for multiple sparse regression. IEEE Transactions on Information Theory, 59(12), 7947-7968.

[3] Negahban, S. N., & Wainwright, M. J. (2011). Simultaneous Support Recovery in High Dimensions: Benefits and Perils of Block ℓ_1/ℓ_∞-Regularization. IEEE Transactions on Information Theory, 57(6), 3841-3863.

§4.4: Other applications of biased regularization in supervised learning

- §4.4.1 High-dimensional quantile regression
- §4.4.2 Functional regression
- §4.4.3 Causal inference

§4.4: Other applications of biased regularization in supervised learning

- §4.4.1 High-dimensional quantile regression
- §4.4.2 Functional regression
- §4.4.3 Causal inference

High-dimensional quantile regression

- **Quantile regression** (assuming $\epsilon^{(k)}$ has Lebesgue density, for simplicity):

$$Y^{(k)} = (X^{(k)})^\top \boldsymbol{\theta}^{(k)*} + \epsilon^{(k)}, \quad \mathbb{P}(\epsilon^{(k)} \leq 0) = \tau, \quad \epsilon^{(k)} \perp\!\!\!\perp X^{(k)},$$

  which implies that $(X^{(k)})^\top \boldsymbol{\theta}^{(k)*}$ is the τ-quantile of the conditional distribution of $Y^{(k)}$ given $X^{(k)}$.

- We observe data $\{\boldsymbol{x}_i^{(k)}, y_i^{(k)}\}_{i=1}^{n_k} \overset{\text{i.i.d.}}{\sim} (X^{(k)}, Y^{(k)})$, $k = 0 : K$.
- **Sparsity:** $\|\boldsymbol{\theta}^{(0)*}\|_0 \leq s$
- **Similarity:** $\max_{k \in S} \|\boldsymbol{\theta}^{(k)*} - \boldsymbol{\theta}^{(0)*}\|_1 \leq h$
- **Goal:** with a fixed τ, estimate $\boldsymbol{\theta}^{(0)*}$
- **Main method:** the previous two-step method (transferring - debiasing) based on the quantile loss function

$$\ell_\tau(x) = x[\tau - \mathbb{1}(x \leq 0)].$$

High-dimensional quantile regression

There are at least five papers studying this problem with slightly different approaches and focuses.

- Huang et al. (2022); Li and Song (2024) directly uses the quantile loss functions in the two-step method. The original quantile loss $\ell_\tau(x) = x[\tau - \mathbb{1}(x \leq 0)]$ is non-differentiable at 0, which leads to a slower rate and computational challenges.
- Jin et al. (2024) creates pseudo labels then adopts squaure loss in the transferring step. But creating pseudo labels requires a good local estimator on each source, which necessitates an extra ℓ_0-sparsity assumption on $\{\boldsymbol{\theta}^{(k)*}\}_{k=1}^K$.
- Zhang and Zhu (2022); Qiao et al. (2023) apply the kernel convolution smoothing to smooth the loss first, then use the smoothes loss in the two-step method.

[1] Huang, J., Wang, M., & Wu, Y. (2022). Estimation and inference for transfer learning with high-dimensional quantile regression. arXiv preprint arXiv:2211.14578.

[2] Li, J., & Song, Y. (2024). Transfer learning with high dimensional composite quantile regression. Journal of Statistical Computation and Simulation, 1-18.

[3] Jin, J., Yan, J., Aseltine, R. H., & Chen, K. (2024). Transfer Learning with Large-Scale Quantile Regression. Technometrics, (just-accepted), 1-30.

[4] Zhang, Y., & Zhu, Z. (2022). Transfer learning for high-dimensional quantile regression via convolution smoothing. arXiv preprint arXiv:2212.00428.

[5] Qiao, S., He, Y., & Zhou, W. (2023). Transfer Learning for High-dimensional Quantile Regression with Statistical Guarantee. Transactions on Machine Learning Research.

§4.4: Other applications of biased regularization in supervised learning

- §4.4.1 High-dimensional quantile regression
- §4.4.2 Functional regression
- §4.4.3 Causal inference

Functional regression

- **Functional linear regression**: (Lin and Reimherr, 2024a)

$$Y^{(k)} = \langle X^{(k)}, \theta^{(k)*} \rangle_{L^2} + \epsilon^{(k)},$$

  where $X^{(k)}(\cdot)$ and $\theta^{(k)*}(\cdot)$ are square integrable real functions over a compact domain in $\mathbb{R}$.
- We observe data $\{x_i^{(k)}, y_i^{(k)}\}_{i=1}^{n_k} \overset{\text{i.i.d.}}{\sim} (X^{(k)}, Y^{(k)})$, $k = 0 : K$.
- Assume $\theta^{(k)*} \in$ some RKHS $\mathcal{H}$ induced by kernel $\mathcal{K}$
- **Similarity:** $\max_{k \in S} \|\theta^{(k)*} - \theta^{(0)*}\|_{\mathcal{H}} \leq h$
- **Goal:** estimate $\theta^{(0)*}$ and bound the excess risk under square loss

[1] Lin, H., & Reimherr, M. (2024). On Hypothesis Transfer Learning of Functional Linear Models. arXiv preprint arXiv:2206.04277. (accepted by ICML 2024)

Functional regression

- **Main method:** (Lin and Reimherr, 2024a) the previous two-step method (transferring - debiasing) with square loss and ridge regularizer
  - Step 1: (transferring)

$$\tilde{\theta} = \arg\min_{\theta \in \mathcal{H}} \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} [y_i^{(k)} - \langle x_i^{(k)}, \theta \rangle_{L^2}]^2 + \lambda_1 \|\theta\|_{\mathcal{H}}^2$$

  - Step 2: (debiasing)

$$\hat{\theta}^{(0)} = \arg\min_{\theta \in \mathcal{H}} \frac{1}{n_0} \sum_{i=1}^{n_0} [y_i^{(0)} - \langle x_i^{(0)}, \theta \rangle_{L^2}]^2 + \lambda_2 \|\theta - \tilde{\theta}\|_{\mathcal{H}}^2$$

- The method is minimax optimal in terms of excess risk.
- The method and theory can be extended to functional GLMs.

Adaptitivty to function smoothness

There some issues of previous two-step method and the theory:

- We need $\theta^{(k)*} \in \mathcal{H}$ with $\mathcal{H}$ known
- We use the same $\|\cdot\|_{\mathcal{H}}^2$ in two steps. However, the bias function $\theta^{(k)*} - \theta^{(0)*}$ might be simpler and smoother than $\theta^{(k)*}$ and $\theta^{(0)*}$

Questions: Is it possible to use the same kernel in two steps (as we did before) to be adaptive to potentially different smoothness of the bias $\theta^{(k)*} - \theta^{(0)*}$ and the target function $\theta^{(0)*}$?

In a non-parametric regression context (different from the functional regression we discussed before), Lin and Reimherr (2024b) proposes to use Gaussian kernel in both steps, which is shown to be adaptive to the smoothness of the bias $\theta^{(k)*} - \theta^{(0)*}$ and the target function $\theta^{(0)*}$, by setting the penalty parameter λ_1 and λ_2 in a proper way.

---

[1] Lin, H., & Reimherr, M. (2024). Smoothness Adaptive Hypothesis Transfer Learning. arXiv preprint arXiv:2402.14966. (accepted by ICML 2024)

§4.4: Other applications of biased regularization in supervised learning

- §4.4.1 High-dimensional quantile regression
- §4.4.2 Functional regression
- §4.4.3 Causal inference

Causal inference

Consider an observational study and potential outcome framework.

- $(X^{(k)}, Z^{(k)}, Y^{(k)}(0), Y^{(k)}(1))$, where $X^{(k)} \in \mathbb{R}^d$ are the features, $Z^{(k)} \in \{0, 1\}$ is the treatment, and $Y^{(k)}(1)$ and $Y^{(k)}(0)$ are the outcomes, $K = 0, 1$.
- Observe $(\boldsymbol{x}_i^{(k)}, z_i^{(k)}, y_i^{(k)}) \overset{\text{i.i.d.}}{\sim} (X^{(k)}, Z^{(k)}, Y^{(k)}(Z^{(k)}))$, where $X^{(k)} \in \mathbb{R}^d$ are the features, $Z^{(k)} \in \{0, 1\}$ is the treatment, and $Y^{(k)}$ is the outcome.
- **Goal:** estimate the average treatment effect (ATE) of the target study:

$$\tau = \mathbb{E}[Y^{(0)}(1)] - \mathbb{E}[Y^{(0)}(0)].$$

- Under the ignorability assumption, i.e.

$$(Y^{(0)}(1), Y^{(0)}(1)) \perp\!\!\!\perp Z^{(0)} | X^{(0)},$$

we have

$$\mathbb{E}[Y^{(0)}(1)] = \mathbb{E}\left[\frac{Z^{(0)} Y^{(0)}}{e^{(0)}(X^{(0)})}\right], \quad \mathbb{E}[Y^{(1)}(1)] = \mathbb{E}\left[\frac{(1 - Z^{(0)}) Y^{(0)}}{1 - e^{(0)}(X^{(0)})}\right],$$

where $e^{(0)}(\boldsymbol{x}) = \mathbb{P}(Z^{(0)} = 1 | X^{(0)} = \boldsymbol{x})$ is the propensity score.

Causal inference

This motivates the inverse propensity score weighting (IPW) estimator

$$\hat{\tau} = \frac{1}{n_0} \sum_{i=1} \frac{z_i^{(0)} y_i^{(0)}}{\hat{e}^{(0)}(\boldsymbol{x}_i^{(0)})} - \frac{1}{n_0} \sum_{i=1} \frac{(1 - z_i^{(0)}) y_i^{(0)}}{1 - \hat{e}^{(0)}(\boldsymbol{x}_i^{(0)})},$$

where $\hat{e}^{(0)}(\cdot)$ is the estimated propensity score.

- Wei et al. (2023) assumes the propensity score model $e^{(k)}(\boldsymbol{x}) \sim$ a GLM with parameter $\boldsymbol{\theta}^{(k)*} \in \mathbb{R}^d$
- Then with an ℓ_0-sparsity assumption on $\max_{k=1:K} \|\boldsymbol{\theta}^{(k)*} - \boldsymbol{\theta}^{(0)*}\|_0$, they propose the following learning procedure:
  - First, apply the two-step method in Bastani (2021), with an OLS on source data to fit $\tilde{\boldsymbol{\theta}}$, then debias $\tilde{\boldsymbol{\theta}}$ with ℓ_1-penalty to get $\widehat{\boldsymbol{\theta}}^{(0)}$
  - Next, plugging $\hat{e}^{(0)}(\cdot) =$ GLM with $\widehat{\boldsymbol{\theta}}^{(0)}$ into the IPW estimator on target domain to get $\hat{\tau}$
- They show an upper bound on $|\hat{\tau} - \tau|$

---

[1] Wei, S., Moore, R., Zhang, H., Xie, Y., & Kamaleswaran, R. (2023, July). Transfer Causal Learning: Causal Effect Estimation with Knowledge Transfer. In ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH).

[2] Bastani, H. (2021). Predicting with proxies: Transfer learning in high dimension. Management Science, 67(5), 2964-2984.

§4.5: Statistical inference

- §4.5.1 Challenges in doing inference
- §4.5.2 Inference in high-dimensional linear regression
- §4.5.3 Inference of the prevailing model

§4.5: Statistical inference

- §4.5.1 Challenges in doing inference
- §4.5.2 Inference in high-dimensional linear regression
- §4.5.3 Inference of the prevailing model

Challenges in doing inference

- We have shown that it is possible to construct an estimation/prediction procedure that is **adaptive** to the similarity between target and sources (or tasks in the context of MTL)
- But statistical inference is much harder than estimation/prediction.

Lemma 4.5.1 (A good CI implies a good point estimator)

Consider observations $\{x_i\}_{i=0}^n \overset{\text{i.i.d.}}{\sim} N(\theta^*, 1)$. *If we can construct a* 95% *(can be asymptotic) confidence interval (CI) of length* $2r$ *that covers* θ^**, then any point estimator inside the CI should has estimation error* $\leq r$ *with prob.* $\geq 95\%$.

Let's look at the following simple example.

Challenges in doing inference

Example: (Tian and Feng, 2023a) $\{x_i^{(k)}\}_{i=0}^{n_k} \overset{\text{i.i.d.}}{\sim} N(\theta^{(k)*}, 1)$, $|\theta^{(1)*} - \theta^{(0)*}| \leq h$. Construct a $(1-\alpha)$-confidence interval (CI) for $\theta^{(0)*}$ with minimum length.

- Define $\bar{x}^{(k)} = \frac{1}{n_k}\sum_{i=1}^{n_k} x_i^{(k)}$, and $\bar{x} = \frac{n_0}{n_0+n_1}\bar{x}^{(0)} + \frac{n_1}{n_0+n_1}\bar{x}^{(1)}$.
- Consider the following cases:
  - If $h \ll \sqrt{\frac{1}{n_0+n_1}}$: $\bar{x} \pm \frac{z_{\alpha/2}}{\sqrt{n_0+n_1}}$ should be optimal [4]
  - If $h \gtrsim \sqrt{\frac{1}{n_0}}$: $\bar{x}^{(0)} \pm \frac{z_{\alpha/2}}{\sqrt{n_0}}$ should be optimal
  - If $\sqrt{\frac{1}{n_0+n_1}} \lesssim h \ll \sqrt{\frac{1}{n_0}}$: **?**
- Another question is how to combine these CIs to make the inference procedure adaptive to unknown h?

If we don't care about optimality, then we can always construct a CI based on

(i) the good point estimator learned by bias regularization;

(ii) the other parts (e.g. standard deviation) learned on target data

[4] $z_{\alpha/2}$ is the upper $\alpha/2$-quantile of $N(0,1)$

[1] Tian, Y., & Feng, Y. (2023). Comments on: Statistical inference and large-scale multiple testing for high-dimensional regression models. Test, 32(4), 1172-1176.

§4.5: Statistical inference

- §4.5.1 Challenges in doing inference
- §4.5.2 Inference in high-dimensional linear regression
- §4.5.3 Inference of the prevailing model

Inference in high-dimensional linear regression

Recall the following high-dimensional regression setting in Li et al. (2021).

- $\{(\boldsymbol{x}_i^{(k)}, y_i^{(k)})\}_{i=1}^{n_k} \overset{\text{i.i.d.}}{\sim} (X^{(k)}, Y^{(k)})$ for $k = 0 : K$, where $X^{(k)} \in \mathbb{R}^d$, $Y^{(k)} \in \mathbb{R}$, and

$$Y^{(k)} = (X^{(k)})^\top \boldsymbol{\theta}^{(k)*} + \epsilon^{(k)},$$

  $\mathbb{E}[X^{(k)}(X^{(k)})^\top] = \boldsymbol{\Sigma}^{(k)}$, $\epsilon^{(k)}$ is independent of $X^{(k)}$ and zero-mean sub-Gaussian (with a constant-level variance proxy).
- For simplicity, assume $n_k \equiv n$
- **Sparsity:** $\|\boldsymbol{\theta}^{(0)*}\|_0 \leq s \ll d$
- **Relationship between tasks:** $\max_{k \in S} \|\boldsymbol{\theta}^{(k)*} - \boldsymbol{\theta}^{(0)*}\|_1 \leq h$ (unknown), where the **informative** set $S \subseteq [K]$ is also unknown
- **Goal:** construct a $(1-\alpha)$-CI for $\theta_j^{(0)*}$

[1] Li, S., Cai, T. T., & Li, H. (2022). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. Journal of the Royal Statistical Society Series B: Statistical Methodology, 84(1), 149-173.

Inference in high-dimensional linear regression

If we only have target data and we apply Lasso

$$\widehat{\boldsymbol{\theta}}^{(0)} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2n_0}\|\boldsymbol{Y}^{(0)} - \boldsymbol{X}^{(0)}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_1$$

to fit $\boldsymbol{\theta}^{(0)*}$, then **debiased Lasso** estimator (Van de Geer et al., 2014; Zhang and Zhang, 2014)

$$\hat{\boldsymbol{b}} = \widehat{\boldsymbol{\theta}}^{(0)} + \frac{1}{\sqrt{n_0}}\widehat{\boldsymbol{\Theta}}^{(0)}(\boldsymbol{X}^{(0)})^\top(\boldsymbol{Y}^{(0)} - \boldsymbol{X}^{(0)}\widehat{\boldsymbol{\theta}}^{(0)})$$

can be used to derive an (asymptotic) $(1-\alpha)$-CI

$$\hat{b}_j \pm \frac{(\widehat{\boldsymbol{\Theta}}_j^{(0)})^\top\widehat{\boldsymbol{\Sigma}}^{(0)}\widehat{\boldsymbol{\Theta}}_j^{(0)}}{\sqrt{n_0}} z_{\alpha/2},$$

where $\widehat{\boldsymbol{\Theta}}^{(0)}$ is an estimator of $\boldsymbol{\Sigma}^{-1}$ and $\widehat{\boldsymbol{\Theta}}_j$ is the j-th column of $\widehat{\boldsymbol{\Theta}}^{(0)}$, obtained by regressing the j-th column of $\boldsymbol{X}^{(0)}$ with other $(d-1)$ columns by Lasso.

[1] Van de Geer, S., Bühlmann, P., Ritov, Y. A., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models.

[2] Zhang, C. H., & Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society Series B: Statistical Methodology, 76(1), 217-242.

Inference in high-dimensional linear regression

- The target-only $(1-\alpha)$-CI:

$$\hat{b}_j \pm \frac{(\widehat{\boldsymbol{\Theta}}_j^{(0)})^\top \widehat{\boldsymbol{\Sigma}}^{(0)} \widehat{\boldsymbol{\Theta}}_j^{(0)}}{\sqrt{n_0}},$$

  with $\hat{\boldsymbol{b}} = \widehat{\boldsymbol{\theta}}^{(0)} + \frac{1}{\sqrt{n_0}} \widehat{\boldsymbol{\Theta}}^{(0)} (\boldsymbol{X}^{(0)})^\top (\boldsymbol{Y}^{(0)} - \boldsymbol{X}^{(0)} \widehat{\boldsymbol{\theta}}^{(0)})$.

- We can modify the CI with

$$\hat{b}_j \pm \frac{\widehat{\boldsymbol{\Theta}}_j^\top \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\Theta}}_j}{\sqrt{n_0}}$$

  - We use $\hat{\boldsymbol{b}} = \widehat{\boldsymbol{\theta}} + \frac{1}{\sqrt{n_0}} \widehat{\boldsymbol{\Theta}} (\boldsymbol{X}^{(0)})^\top (\boldsymbol{Y}^{(0)} - \boldsymbol{X}^{(0)} \widehat{\boldsymbol{\theta}})$, where $\widehat{\boldsymbol{\theta}}$ is obtained by our previous two-step ℓ_1-biased regularization
  - We run our two-step method to obtain each row $\widehat{\boldsymbol{\Theta}}_j$, which should perform better than $\widehat{\boldsymbol{\Theta}}_j^{(0)}$ fitted by target data only.
  - $\widehat{\boldsymbol{\Sigma}}$ is the sample-size weighted average of all empirical covariances across tasks.

Inference in high-dimensional linear regression

An (asymptotic) $(1-\alpha)$-CI: (Tian and Feng, 2023b)

$$\hat{b}_j \pm \frac{\widehat{\boldsymbol{\Theta}}_j^\top \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\Theta}}_j}{\sqrt{n_0}}$$

where $\hat{b}_j \pm \frac{\widehat{\boldsymbol{\Theta}}_j^\top \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\Theta}}_j}{\sqrt{n_0}}$.

- The order of CI width is still $\asymp \sqrt{\frac{1}{n_0}}$, which doesn't improve
- We may improve on the constant level due to the use of new $\widehat{\boldsymbol{\Theta}}$ and $\widehat{\boldsymbol{\theta}}$
- **Challenges:**
  - Lasso itself is biased.
  - Introducing source data in the debiased Lasso will bring the bias of source data into the game
  - The composition of two different biases makes the inference difficult
  - To avoid the second bias, we used the target data to debias the two-step estimator $\widehat{\boldsymbol{\theta}}$

[1] Tian, Y., & Feng, Y. (2023). Transfer learning under high-dimensional generalized linear models. Journal of the American Statistical Association, 118(544), 2684-2697.

Inference in high-dimensional linear regression

- Li et al. (2023) proposes a similar debiased Lasso procedure, but they use only the target data to construct $\widehat{\mathbf{\Theta}}$.
- When $\mathbf{\Sigma}^{(k)} = \mathbb{E}[X^{(k)}(X^{(k)})^\top]$ are similar across different sources (e.g., only posterior drift happens), our procedure (Tian and Feng, 2023b) is better because we utilize this similarity
- However, the procedure in Li et al. (2023) is more safe and conservative, and it is valid even when $\mathbf{\Sigma}^{(k)} = \mathbb{E}[X^{(k)}(X^{(k)})^\top]$ are very different across different sources

[1] Li, S., Zhang, L., Cai, T. T., & Li, H. (2023). Estimation and inference for high-dimensional generalized linear models with knowledge transfer. Journal of the American Statistical Association, 1-12.

[2] Tian, Y., & Feng, Y. (2023). Transfer learning under high-dimensional generalized linear models. Journal of the American Statistical Association, 118(544), 2684-2697.

§4.5: Statistical inference

- §4.5.1 Challenges in doing inference
- §4.5.2 Inference in high-dimensional linear regression
- §4.5.3 Inference of the prevailing model

Inference of the prevailing model

Before we conclude this section, let's take a look at another related inference problem studied in Guo et al. (2023).

- $\{(\boldsymbol{z}_i^{(k)})\}_{i=1}^{n_k} \overset{\text{i.i.d.}}{\sim} Z^{(k)} \sim \mathbb{P}_{\boldsymbol{\theta}^{(k)*}}$ for $k = 1 : K$
- **Prevailing model**: $\exists$ an unknown $S \subseteq [K]$ with $|S| > K/2$, s.t.

$$\boldsymbol{\theta}^{(k)*} = \boldsymbol{\theta}^*, \quad \forall k \in S.$$

- **Goal:** construct a $(1-\alpha)$-CI for $g(\boldsymbol{\theta}^*)$ with some known $g : \mathbb{R}^d \to \mathbb{R}$ [5]
- **Examples of g:**
  - Single coefficient: $g(\boldsymbol{\theta}^*) = \theta_j^*$
  - Linear transform: $g(\boldsymbol{\theta}^*) = \boldsymbol{x}^\top \boldsymbol{\theta}^*$
  - Quadratic form: $g(\boldsymbol{\theta}^*) = \|\boldsymbol{\theta}^*\|_2^2$
- For simplicity, let's consider $g(\boldsymbol{\theta}^*) = \theta_j^*$.

[5] There are other forms of g considered in Guo et al. (2023), please refer to their paper for details.

[1] Guo, Z., Li, X., Han, L., & Cai, T. (2023). Robust inference for federated meta-learning. arXiv preprint arXiv:2301.00718.

Inference of the prevailing model

- Suppose we have a CLT result for each of $\{\widehat{\theta}_j^{(k)}, \hat{\sigma}_k\}_{k=1}^K$

$$\frac{1}{\sigma_k}(\widehat{\theta}_j^{(k)} - \theta_j^{(k)*}) \xrightarrow{d} N(0,1), \quad \hat{\sigma}_k \xrightarrow{d} \sigma_k,$$

  where $\{\widehat{\theta}_j^{(k)}, \hat{\sigma}_k\}_{k=1}^K$ are independent (i.e. constructed by local data). Then we have a $(1-\alpha)$-CI constructed from each dataset:

$$\widehat{\theta}_j^{(k)} \pm \hat{\sigma}_k z_{\alpha/2}.$$

- If we know S beforehand, then we can use the following $(1-\alpha)$-CI:

$$\widehat{\theta}_j \pm z_{\alpha/2} \frac{1}{\sqrt{\sum_{k\in S} 1/\hat{\sigma}_k^2}},$$

  where $\widehat{\theta}_j = \frac{\sum_{k\in S} \widehat{\theta}_j^{(k)}/\hat{\sigma}_k^2}{\sum_{k\in S} 1/\hat{\sigma}_k^2}$.

Inference of the prevailing model

- We can use the following CLT result to estimate S:

$$\frac{1}{\sqrt{\sigma_k^2+\sigma_l^2}}[\underbrace{(\widehat{\theta}_j^{(k)}-\widehat{\theta}_j^{(l)})}_{:=\widehat{L}_{l,k}}-(\theta_j^{(k)*}-\theta_j^{(l)*})] \xrightarrow{d} N(0,1), \quad \hat{\sigma}_k^2+\hat{\sigma}_l^2 \xrightarrow{d} \sigma_k^2+\sigma_l^2.$$

- Consider

$$\widehat{S}_{l,k} := \frac{|\widehat{\theta}_j^{(k)}-\widehat{\theta}_j^{(l)}|}{\sqrt{\hat{\sigma}_k^2+\hat{\sigma}_l^2}}, \quad k \neq l \in [K],$$

$\widehat{H}_{l,k} = \mathbb{1}(\widehat{S}_{l,k} \leq z_{\alpha/[2K(K-1)]})$, and $\widehat{\boldsymbol{H}} = (\widehat{H}_{l,k})_{l,k=1}^K$. We can estimate S by

$$\widehat{S} = \{l \in [K] : \|\widehat{\boldsymbol{H}}_{l,:}\|_0 > K/2\}.$$

Inference of the prevailing model

- Can we directly use

$$\widehat{\theta}_j \pm z_{\alpha/2} \frac{1}{\sqrt{\sum_{k\in\widehat{S}} 1/\hat{\sigma}_k^2}}$$

  as the $(1-\alpha)$-CI with $\widehat{\theta}_j = \frac{\sum_{k\in\widehat{S}} \widehat{\theta}_j^{(k)}/\hat{\sigma}_k^2}{\sum_{k\in\widehat{S}} 1/\hat{\sigma}_k^2}$?

  NO! We haven't considered the effect of $\widehat{S}$ in the confidence level. This is related to the post-selection inference in high-dimensional statistics (Taylor and Tibshirani, 2015).

- Guo et al. (2023) proposes the re-sampling procedure:
  - Step 1: Sample $\widehat{L}_{l,k}^{[m]} \sim N(\widehat{L}_{l,k}, \hat{\sigma}_k^2)$, and construct $\widehat{S}^{[m]}$
  - Step 2: Define $\text{CI}^{[m]} = \widehat{\theta}_j \pm z_{\alpha/2} \frac{1}{\sqrt{\sum_{k\in\widehat{S}^{[m]}} 1/\hat{\sigma}_k^2}}$ with $\widehat{\theta}_j = \frac{\sum_{k\in\widehat{S}^{[m]}} \widehat{\theta}_j^{(k)}/\hat{\sigma}_k^2}{\sum_{k\in\widehat{S}^{[m]}} 1/\hat{\sigma}_k^2}$.

    Output $\text{CI} = \cup_{m\in[M]} \text{CI}^{[m]}$.

---

[1] Taylor, J., & Tibshirani, R. J. (2015). Statistical learning and selective inference. Proceedings of the National Academy of Sciences, 112(25), 7629-7634.

[2] Guo, Z., Li, X., Han, L., & Cai, T. (2023). Robust inference for federated meta-learning. arXiv preprint arXiv:2301.00718.

Inference of the prevailing model

Theorem 4.5.1 (Guo et al., 2023)

Under certain conditions:

- *As $n, M \to \infty$, the CI we constructed has confidence level $1 - \alpha$.*
- *As $n, M \to \infty$, the CI we constructed is the same as the oracle CI based on known S.*

- In the Gaussian mean estimation problem we discussed before

$$\text{CI width} \asymp \frac{1}{\sqrt{\sum_{k \in S} 1/\sigma_k^2}} \asymp \frac{1}{\sqrt{\sum_{k \in S} n_k}}$$

- This optimal rate is possible because of the "prevailing model" assumption, which differs from our previous example where the majority model doesn't exist.
- A generic inference framework with heterogeneous datasets is still an open problem.

[1] Guo, Z., Li, X., Han, L., & Cai, T. (2023). Robust inference for federated meta-learning. arXiv preprint arXiv:2301.00718.

§4.6: Unsupervised multi-task learning

Unsupervised MTL

Existing works of domain adaptation:

- Mainly focused on the goal of supervised learning or semi-supervised learning
- Some studied pure unsupervised learning setting but are lack of theoretical understanding (Gu et al., 2011; Zhang and Zhang, 2011; Yang et al., 2014)
- Recently, there are a few works proposing a so-called federated Expectation-Maximization (EM) algorithms (Marfoq et al., 2021; Dieuleveut et al., 2021; Wu et al., 2023) based on **pooling**, and these federated EM algorithms have achieved great success empirically
- We extended the pooling-based EM to biased regularized EM with ℓ_2-penalty (Duan and Wang, 2022), and first explore the non-asymptotic theory of federated EM algorithms in Gaussian mixture models (Tian et al., 2022). Then we extended the result to general mixture models (Tian et al., 2024).

[1] Marfoq, O., Neglia, G., Bellet, A., Kameni, L., & Vidal, R. (2021). Federated multi-task learning under a mixture of distributions. Advances in Neural Information Processing Systems, 34, 15434-15447.

[2] Dieuleveut, A., Fort, G., Moulines, E., & Robin, G. (2021). Federated-EM with heterogeneity mitigation and variance reduction. Advances in Neural Information Processing Systems, 34, 29553-29566.

[3] Wu, Y., Zhang, S., Yu, W., Liu, Y., Gu, Q., Zhou, D., ... & Cheng, W. (2023, July). Personalized federated learning under mixture of distributions. In International Conference on Machine Learning (pp. 37860-37879). PMLR.

[4] Tian, Y., Weng, H., & Feng, Y. (2022). Unsupervised multi-task and transfer learning on gaussian mixture models. arXiv preprint arXiv:2209.15224.

[5] Tian, Y., Weng, H., & Feng, Y. (2024). Towards the Theory of Unsupervised Federated Learning: Non-asymptotic Analysis of Federated EM Algorithms. arXiv preprint arXiv:2310.15330. (accepted by ICML 2024)

Binary GMMs

We will focus on the setting in Tian et al. (2022, 2024) and understand how biased regularization can be used in unsupervised multi-task/federated learning.

Let's start from the binary GMMs.

- The observations from the k-th task ($k \in S$) is from a binary GMM (in $\mathbb{R}^d$):

$$\boldsymbol{z}_i^{(k)} \overset{\text{i.i.d.}}{\sim} (1 - w^{(k)*}) \cdot N(\boldsymbol{\mu}_1^{(k)*}, \boldsymbol{\Sigma}^{(k)*}) + w^{(k)*} \cdot N(\boldsymbol{\mu}_2^{(k)*}, \boldsymbol{\Sigma}^{(k)*}), \quad i = 1 : n.$$

  Equivalent to:

$$y_i^{(k)} = \begin{cases} 1, & \text{w.p. } 1 - w^{(k)*}, \\ 2, & \text{w.p. } w^{(k)*}, \end{cases} \qquad \boldsymbol{z}_i^{(k)} | y_i^{(k)} = r \sim N(\boldsymbol{\mu}_r^{(k)*}, \boldsymbol{\Sigma}^{(k)*}).$$

- Datasets outside S: $\{\boldsymbol{z}_1^{(k)}, \ldots, \boldsymbol{z}_n^{(k)}\}_{k \in S^c} \sim \mathbb{Q}$ (arbitrary contamination)
- **Goal:** Improve mis-clustering error on tasks in S

---

[1] Tian, Y., Weng, H., & Feng, Y. (2022). Unsupervised multi-task and transfer learning on gaussian mixture models. arXiv preprint arXiv:2209.15224.

[2] Tian, Y., Weng, H., & Feng, Y. (2024). Towards the Theory of Unsupervised Federated Learning: Non-asymptotic Analysis of Federated EM Algorithms. arXiv preprint arXiv:2310.15330. (accepted by ICML 2024)

Binary GMMs

Similarity between GMMs:

- A key quantity "discriminative coefficient":

$$\boldsymbol{\theta}^{(k)*} := (\boldsymbol{\Sigma}^{(k)*})^{-1}(\boldsymbol{\mu}_2^{(k)*} - \boldsymbol{\mu}_1^{(k)*})$$

- Assume

$$\min_{\bar{\boldsymbol{\theta}} \in \mathbb{R}^d} \max_{k \in S} \|\boldsymbol{\theta}^{(k)*} - \bar{\boldsymbol{\theta}}\|_2 \leq h$$

- Why similarity between $\boldsymbol{\theta}^{(k)*}$? Because it is the key in clustering!

  Recall the Bayes clustering method:

$$\mathcal{C}^{(k)*}(\boldsymbol{z}) = \begin{cases} 1, & \text{if } \left(\boldsymbol{z} - \frac{\boldsymbol{\mu}_1^{(k)*} + \boldsymbol{\mu}_2^{(k)*}}{2}\right)^\top \boldsymbol{\theta}^{(k)*} \leq \log\left(\frac{1-w^{(k)*}}{w^{(k)*}}\right); \\ 2, & \text{otherwise.} \end{cases}$$

  - $h = 0$: $\boldsymbol{\theta}^{(k)*}$'s for $k \in S$ are the same
  - $h = \infty$: No relationship between GMMs
  - $0 < h < \infty$: "Some similarities" between GMMs in S

Expectation-Maximization (EM) algorithm

Consider a finite-component mixture model $\{\boldsymbol{z}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \sum_y p(\boldsymbol{z}, y; \theta)$.

- Full log-likelihood is NOT decomposable: $\sum_{i=1}^n \log\left[\sum_y p(\boldsymbol{z}_i, y; \theta)\right]$
- A surrogate: For any θ':
  $\sum_{i=1}^n \log\left[\sum_y p(\boldsymbol{z}_i, y|\theta)\right] \geq \sum_{i=1}^n \sum_y p(y|\boldsymbol{z}_i; \theta') \cdot \log[p(y, \boldsymbol{z}_i; \theta)]$
- Iterations of E-step and M-step:
  - E-step: $G^{[t+1]}(\theta; \theta^{[t]}) = \sum_{i=1}^n \sum_y p(y|\boldsymbol{z}_i; \theta^{[t]}) \cdot \log[p(y, \boldsymbol{z}_i; \theta)]$
  - M-step: $\theta^{[t+1]} \leftarrow \arg\max_\theta G^{[t+1]}(\theta; \theta^{[t]})$

EM algorithm for GMM

Consider: GMM $(1-w)\cdot N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + w \cdot N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$

$\theta = (w, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ and $\theta^{[t+1]}$ has an explicit expression in M-step:

- $\hat{w}^{[t+1]} = \frac{1}{n}\sum_{i=1}^n \mathbb{P}(y=2|\boldsymbol{z}_i; \widehat{\boldsymbol{\theta}}^{[t]}) := \frac{1}{n}\sum_{i=1}^n \gamma_{\widehat{\boldsymbol{\theta}}^{[t]}}(\boldsymbol{z}_i)$
- $\widehat{\boldsymbol{\mu}}_1^{[t+1]} = \frac{\sum_{i=1}^n [1-\gamma_{\widehat{\boldsymbol{\theta}}^{[t]}}(\boldsymbol{z}_i)]\boldsymbol{z}_i}{\sum_{i=1}^n [1-\gamma_{\widehat{\boldsymbol{\theta}}^{[t]}}(\boldsymbol{z}_i)]}, \quad \widehat{\boldsymbol{\mu}}_2^{[t+1]} = \frac{\sum_{i=1}^n \gamma_{\widehat{\boldsymbol{\theta}}^{[t]}}(\boldsymbol{z}_i)\boldsymbol{z}_i}{\sum_{i=1}^n \gamma_{\widehat{\boldsymbol{\theta}}^{[t]}}(\boldsymbol{z}_i)}$
- $\widehat{\boldsymbol{\Sigma}}^{[t+1]} = \frac{1}{n}\sum_{i=1}^n \Big\{ [1-\gamma_{\widehat{\boldsymbol{\theta}}^{[t]}}(\boldsymbol{z}_i)] \cdot (\boldsymbol{z}_i - \widehat{\boldsymbol{\mu}}_1^{[t+1]})(\boldsymbol{z}_i - \widehat{\boldsymbol{\theta}}_1^{[t+1]})^\top + \gamma_{\widehat{\boldsymbol{\theta}}^{[t]}}(\boldsymbol{z}_i) \cdot (\boldsymbol{z}_i - \widehat{\boldsymbol{\mu}}_2^{[t+1]})(\boldsymbol{z}_i^{(k)} - \widehat{\boldsymbol{\mu}}_2^{[t+1]})^\top \Big\}$
- $\widehat{\boldsymbol{\theta}}^{[t+1]} = (\widehat{\boldsymbol{\Sigma}}^{[t+1]})^{-1}(\widehat{\boldsymbol{\mu}}_2^{[t+1]} - \widehat{\boldsymbol{\mu}}_1^{[t+1]})$

Finally, we plug the estimates in the formula of Bayes clustering method.

Multi-task EM algorithm

- Initializations $\{\hat{w}^{(k)[0]}\}_{k=1}^K$, $\{\widehat{\boldsymbol{\mu}}_1^{(k)[0]}\}_{k=1}^K$, $\{\widehat{\boldsymbol{\mu}}_2^{(k)[0]}\}_{k=1}^K$, $\{\widehat{\boldsymbol{\theta}}^{(k)[0]}\}_{k=1}^K$
- For $t = 0 : (T-1)$: $\lambda^{[t+1]} = \kappa\lambda^{[t]} + C\sqrt{p + \log K}$
  - ▷ Local update: Update $\hat{w}^{(k)[t+1]}, \widehat{\boldsymbol{\mu}}_1^{(k)[t+1]}, \widehat{\boldsymbol{\mu}}_2^{(k)[t+1]}$ and $\widehat{\boldsymbol{\Sigma}}^{(k)[t+1]}$ as in single-task EM, but with posterior calculated by $\{\widehat{\boldsymbol{\theta}}^{(k)[t]}\}_{k=1}^K$
  - ▷ Central update:

$$\{\widehat{\boldsymbol{\theta}}^{(k)[t+1]}\}_{k=1}^K, \widehat{\bar{\boldsymbol{\theta}}}^{[t+1]} = \underset{\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(K)},\bar{\boldsymbol{\theta}}}{\arg\min} \left\{ \sum_k \frac{\lambda^{[t+1]}}{\sqrt{n}} \|\boldsymbol{\theta}^{(k)} - \bar{\boldsymbol{\theta}}\|_2 + \sum_k \left[\frac{1}{2} (\boldsymbol{\theta}^{(k)})^\top \widehat{\boldsymbol{\Sigma}}^{(k)[t+1]} \boldsymbol{\theta}^{(k)} - (\boldsymbol{\theta}^{(k)})^\top (\widehat{\boldsymbol{\mu}}_2^{(k)[t+1]} - \widehat{\boldsymbol{\mu}}_1^{(k)[t+1]}) \right] \right\}$$

- Output $\{\hat{w}^{(k)[T]}\}_{k=1}^K$, $\{\widehat{\boldsymbol{\mu}}_1^{(k)[T]}\}_{k=1}^K$, $\{\widehat{\boldsymbol{\mu}}_2^{(k)[T]}\}_{k=1}^K$, $\{\widehat{\boldsymbol{\theta}}^{(k)[T]}\}_{k=1}^K$, $\{\widehat{\boldsymbol{\Sigma}}^{(k)[T]}\}_{k=1}^K$
- Clustering: Construct Bayes clustering methods $\{\widehat{\mathcal{C}}^{(k)}\}_{k=1}^K$ by the estimates

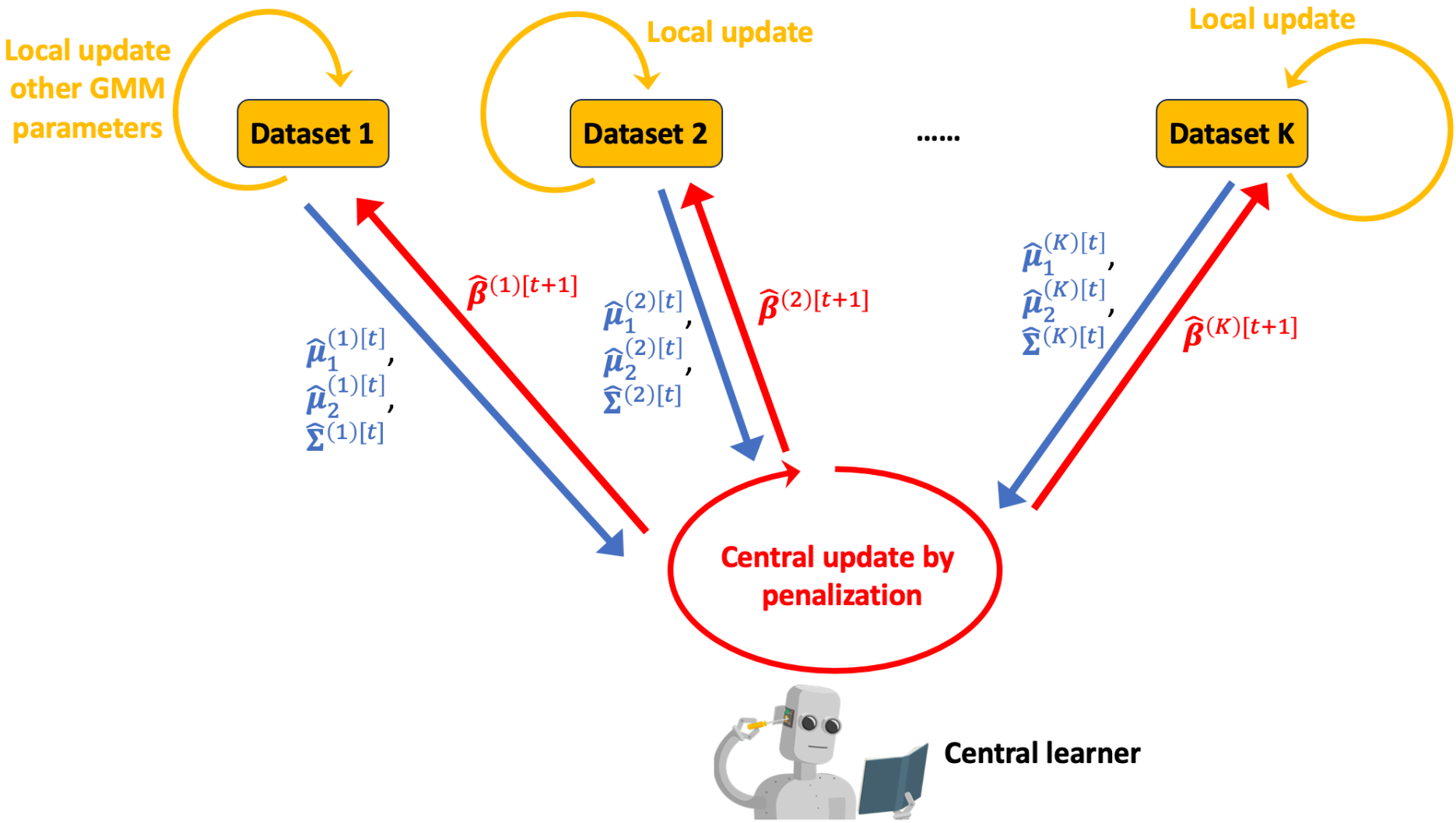
Remark: By setting $\lambda^{[t]} = +\infty$, the federated EM in Marfoq et al. (2021); Dieuleveut et al. (2021); Wu et al. (2023) can be seen as a variant of our method.

[1] Marfoq, O., Neglia, G., Bellet, A., Kameni, L., & Vidal, R. (2021). Federated multi-task learning under a mixture of distributions. Advances in Neural Information Processing Systems, 34, 15434-15447.

[2] Dieuleveut, A., Fort, G., Moulines, E., & Robin, G. (2021). Federated-EM with heterogeneity mitigation and variance reduction. Advances in Neural Information Processing Systems, 34, 29553-29566.

[3] Wu, Y., Zhang, S., Yu, W., Liu, Y., Gu, Q., Zhou, D., ... & Cheng, W. (2023, July). Personalized federated learning under mixture of distributions. In International Conference on Machine Learning (pp. 37860-37879). PMLR.

Multi-task EM algorithm: a federated learning fashion



Excess mis-clustering error: upper bound

Assumptions:

- Initializations of k-th task ($k \in S$) are close enough to the truth
- Signal-noise-ratio of each task in S is not very small
- The outlier fraction $\epsilon = |S^c|/K$ is small

The mis-clustering error of any clustering method $\mathcal{C}$ on the k-th GMM is

$$R^{(k)}(\mathcal{C}) = \min_{\pi:\{1,2\}\to\{1,2\}} \mathbb{P}_{(Z^{\text{new}},Y^{\text{new}})\sim k\text{-th GMM}}\big(\mathcal{C}(Z^{\text{new}}) \neq \pi(Y^{\text{new}})\big).$$

Denote $\mathcal{C}^{(k)*}$ as Bayes clustering method of the k-th GMM.

Theorem 4.6.1 (Tian et al., 2022)

The excess risk of $\{\widehat{\mathcal{C}}^{(k)}\}_{k=1}^K$ learned by multi-task EM satisfies:

$$\max_{k\in S}[R^{(k)}(\widehat{\mathcal{C}}^{(k)}) - R^{(k)}(\mathcal{C}^{(k)*})] \lesssim_{\mathbb{P}} T^4(\kappa')^{2T} + \frac{d}{nK} + h^2 \wedge \left(\frac{d+\log K}{n}\right) + \frac{1}{n} + \epsilon^2\left(\frac{d+\log K}{n}\right)$$

where $\kappa' \in (0,1)$.

[1] Tian, Y., Weng, H., & Feng, Y. (2022). Unsupervised multi-task and transfer learning on gaussian mixture models. arXiv preprint arXiv:2209.15224.

Excess mis-clustering error: upper bound

Consequence of this upper bound: (let $T \gtrsim \log n$)

$$\max_{k \in S} [R^{(k)}(\widehat{\mathcal{C}}^{(k)}) - R^{(k)}(\mathcal{C}^{(k)*})] \lesssim_{\mathbb{P}} \frac{d}{nK} + h^2 \wedge \left(\frac{d + \log K}{n} \right) + \frac{1}{n} + \epsilon^2 \left(\frac{d + \log K}{n} \right)$$

- The excess risk of any single-task algorithm $\gtrsim_{\mathbb{P}} \frac{d+\log K}{n}$ (Cai et al., 2019)
- When
  - $h \ll \sqrt{\frac{d+\log K}{n}}$ (similar datasets);
  - $d \gg 1$ (diverging dimension);
  - $\epsilon \ll (d + \log K)^{-1/2}$ (not many contaminated datasets),

  the multi-task EM performs better than all single-task algorithms.
- Multi-task EM always achieves the single-task rate $\frac{d+\log K}{n}$
- We achieved **adaptivity** and **robustness**:
  - Adaptive to similarity level h
  - Robust to a small fraction of dataset contaminations

[1] Cai, T. T., Ma, J., & Zhang, L. (2019). CHIME: Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality. Annals of Statistics, 47(3), 1234-1267.

Excess mis-clustering error: lower bound

Theorem 4.6.2 (Tian et al., 2022)

For any clustering methods $\{\widehat{\mathcal{C}}^{(k)}\}_{k=1}^{K}$, $\exists$ *a parameter setting and a contamination set* $S^c \subseteq 1:K$ *s.t. with prob.* $\geq 1/10$*:*

$$\max_{k\in S}[R^{(k)}(\widehat{\mathcal{C}}^{(k)}) - R^{(k)}(\mathcal{C}^{(k)*})] \gtrsim \frac{d}{nK} + \min\left\{h^2, \frac{d+\log K}{n}\right\} + \frac{1}{n} + \epsilon^2\left(\frac{1}{n}\right)$$

Remarks:

- Recall the upper bound $\frac{d}{nK} + \min\left\{h^2, \frac{d+\log K}{n}\right\} + \frac{1}{n} + \epsilon^2\left(\frac{d+\log K}{n}\right)$, only the last term is sub-optimal.
- As we mentioned before when discussing the ℓ_2-penalty, this mismatch is shown to exist for biased regularization with many penalties in one of our ongoing works (Tian and Avella, 2024+).

Federated EM: estension to general mixture models

- Marfoq et al. (2021); Dieuleveut et al. (2021); Wu et al. (2023) show that federated EM methods performed well in more than just GMMs, but also mixture of regressions (MoRs).
- Our recent paper (Tian et al., 2024) extends our method and theory from GMMs (Tian et al., 2022) to general mixture models, which include GMMs and MoRs as special cases.
- Our result shows how biased regularization can help in unsupervised multi-task/federated learning, and our theory illustrates the success of the empirical findings of existing federated EM methods.

---

[1] Marfoq, O., Neglia, G., Bellet, A., Kameni, L., & Vidal, R. (2021). Federated multi-task learning under a mixture of distributions. Advances in Neural Information Processing Systems, 34, 15434-15447.

[2] Dieuleveut, A., Fort, G., Moulines, E., & Robin, G. (2021). Federated-EM with heterogeneity mitigation and variance reduction. Advances in Neural Information Processing Systems, 34, 29553-29566.

[3] Wu, Y., Zhang, S., Yu, W., Liu, Y., Gu, Q., Zhou, D., ... & Cheng, W. (2023, July). Personalized federated learning under mixture of distributions. In International Conference on Machine Learning (pp. 37860-37879). PMLR.

[4] Tian, Y., Weng, H., & Feng, Y. (2024). Towards the Theory of Unsupervised Federated Learning: Non-asymptotic Analysis of Federated EM Algorithms. arXiv preprint arXiv:2310.15330. (accepted by ICML 2024)

[5] Tian, Y., Weng, H., & Feng, Y. (2022). Unsupervised multi-task and transfer learning on gaussian mixture models. arXiv preprint arXiv:2209.15224.

§4.7: Domain adaptation with representation learning

- §4.7.1 Representation learning and fine tuning
- §4.7.2 Learning the shared representation across tasks
- §4.7.3 Go beyond the shared representation

§4.7: Domain adaptation with representation learning

- §4.7.1 Representation learning and fine tuning
- §4.7.2 Learning the shared representation across tasks
- §4.7.3 Go beyond the shared representation

A quote from Vapnik (1996)

The following quote is from Chapter 5.13.1 of Vapnik (1996):

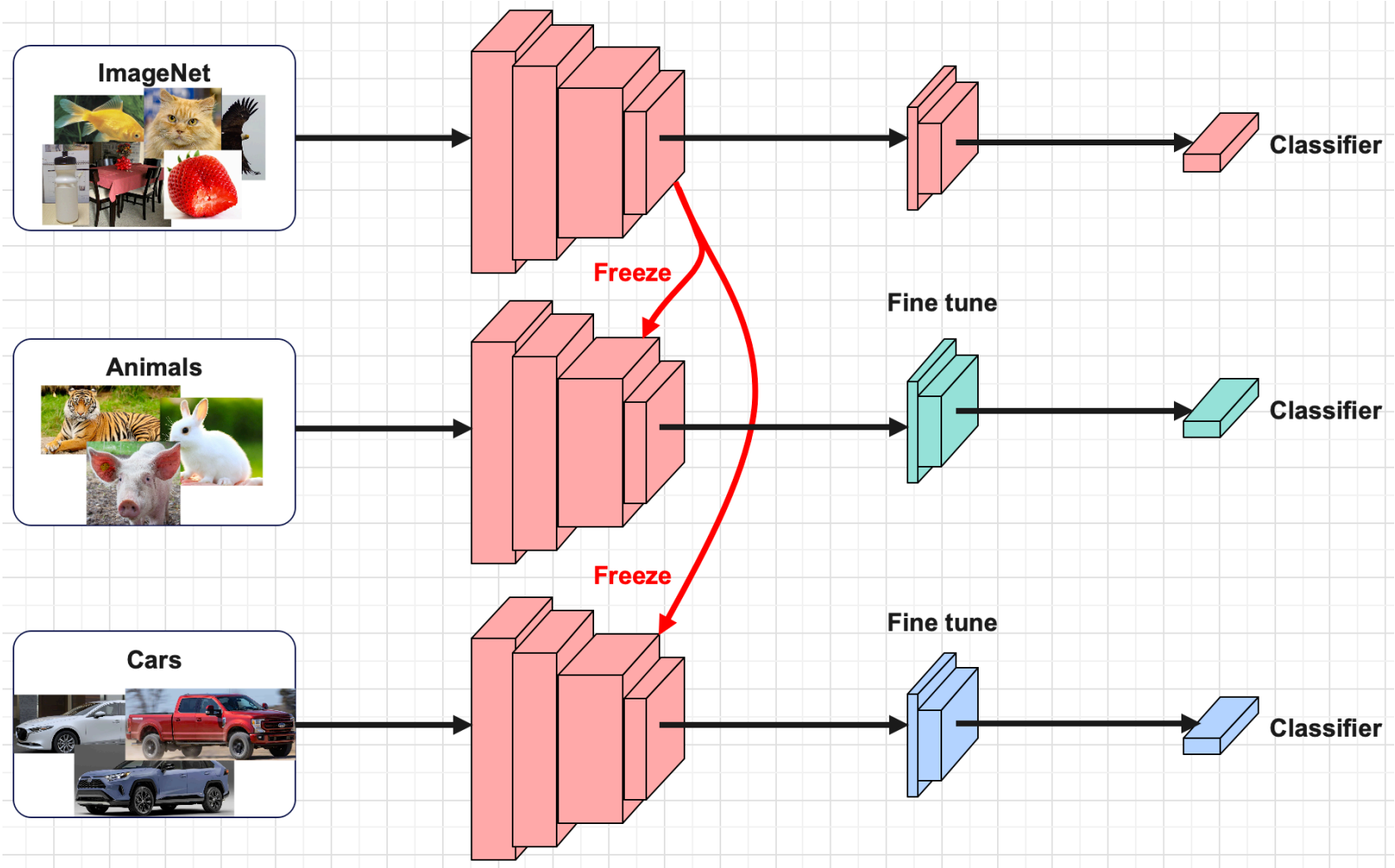
> The classical approach to estimating multidimensional functional dependencies is based on the following belief:
>
> *Real-life problems are such that there exists a small number of "strong features," simple functions of which (say linear combinations) approximate well the unknown function. Therefore, it is necessary to carefully choose a low-dimensional feature space and then to use regular statistical techniques to construct an approximation.*

We have used sparsity (of the offset) and divergence (between distributions) to describe the similarity between different tasks.

Question: Can we describe the similarity by assuming all tasks share some **low-dimensional structure**?

[1] Vapnik, V. (1996). The nature of statistical learning theory. Springer science & business media.

An example: image classification

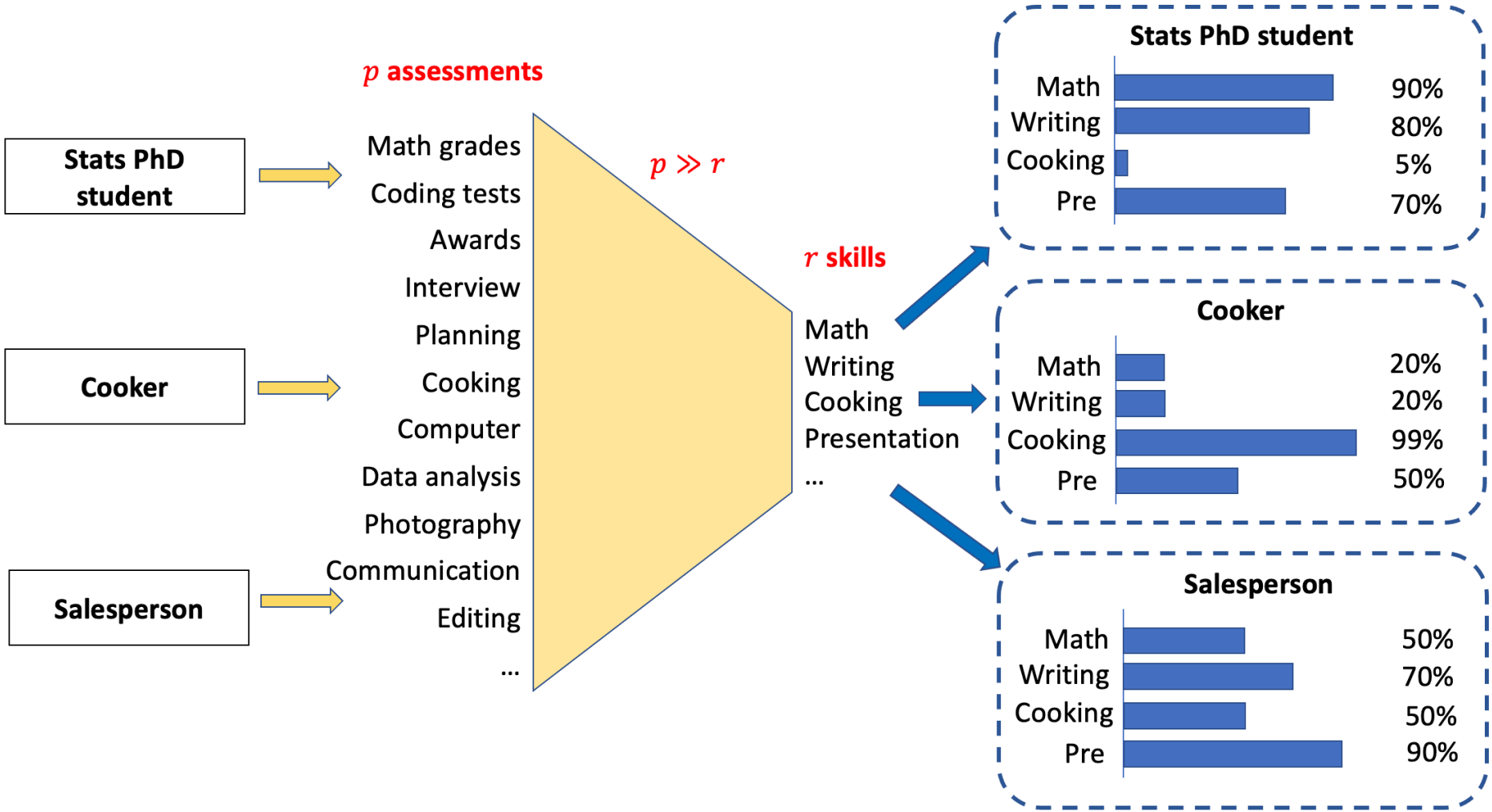


See applications of this idea in, e.g., Donahue et al. (2014); Vinyals et al. (2016).

[1] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014, January). Decaf: A deep convolutional activation feature for generic visual recognition. In International conference on machine learning. PMLR.

[2] Vinyals, O., Blundell, C., Lillicrap, T., & Wierstra, D. (2016). Matching networks for one shot learning. Advances in neural information processing systems, 29.

Another example: the skill assessment



Representation learning and latent factorization are popular in education assessment applications. See, e.g., Chen et al. (2023); Scarlatos et al. (2022).

[1] Chen, Y., Li, X., Liu, J., & Ying, Z. (2023). Item response theory: a statistical framework for educational and psychological measurement. Statistical Science.

[2] Scarlatos, A., Brinton, C., & Lan, A. (2022). Process-BERT: A framework for representation learning on educational process data. arXiv preprint arXiv:2204.13607.

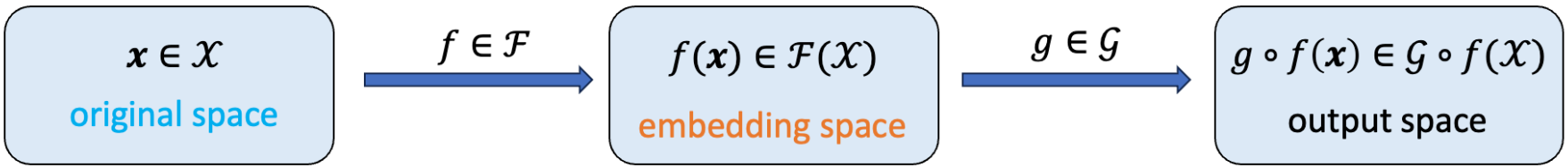
§4.7: Domain adaptation with representation learning

- §4.7.1 Representation learning and fine tuning
- §4.7.2 Learning the shared representation across tasks
- §4.7.3 Go beyond the shared representation

Understanding shared representation: history

Baxter (2000) originally formulate this problem into a "bias learning" model.

- Consider an ERM setting with $\{\boldsymbol{x}_i^{(k)}, y_i^{(k)}\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{P}^{(k)}$, for $k = 1 : K$, and $\{\mathbb{P}^{(k)}\}_{k=1}^K \overset{\text{i.i.d.}}{\sim} \mu$.
- Their goal is to find a **hypothesis space family** $\mathbb{H}$ s.t. one hypothesis class $\mathcal{H} \in \mathbb{H}$ incurs small excess error over $\mathbb{P} \sim \mu$
- Then they specialize $\mathbb{H} = \{\mathcal{G} \circ f : f \in \mathcal{F}\}$, where $\mathcal{G} \circ f := \{g \circ f : g \in \mathcal{G}\}$.
  - f serves as a "representation", which they call "a set of strong features" [6]
  - g is the learner in the embedding space
  - They want to look for an f that works for all tasks



[6] The name is credited to the quote from Vapnik (1996).

[1] Baxter, J. (2000). A model of inductive bias learning. Journal of artificial intelligence research, 12, 149-198.

[2] Vapnik, V. (1996). The nature of statistical learning theory. Springer science & business media.

Understanding shared representation: history

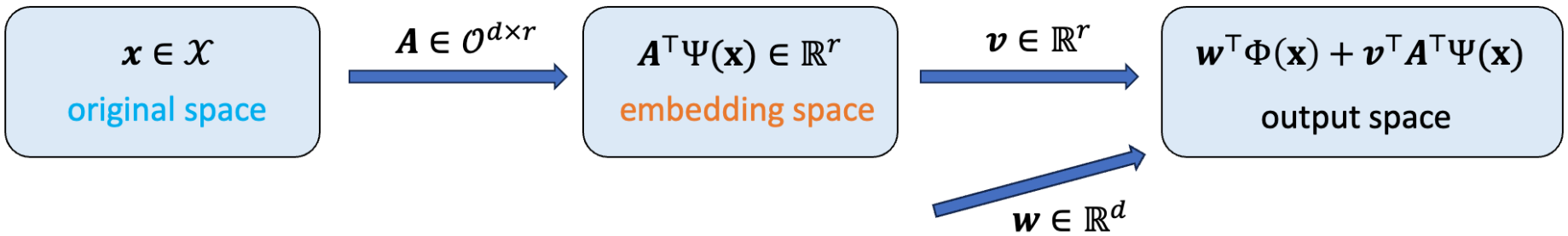
Ando and Zhang (2005) considers a linear predictor

$$h(\boldsymbol{x}) = \boldsymbol{w}^\top \Phi(\boldsymbol{x}) + \boldsymbol{v}^\top \boldsymbol{A}^\top \Psi(\boldsymbol{x}),$$

with $\boldsymbol{w} \in \mathbb{R}^d$, $\boldsymbol{v} \in \mathbb{R}^r$, and $\boldsymbol{A} \in \mathbb{R}^{d \times r}$, $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}_r$. Φ and Ψ are given functions.

- For each task, $\boldsymbol{A}$ serves as a "representation" and is **shared** across tasks
- $\boldsymbol{v}$ can be seen as the coordinate in the embedding space, which can differ across tasks
- $\boldsymbol{w}^\top \Phi(\boldsymbol{x})$ allows some deviation from the shared representation, where $\boldsymbol{w}$ can differ across tasks
- Overall, they propose to consider the following learner for task k:

$$h^{(k)}(\boldsymbol{x}) = (\boldsymbol{w}^{(k)})^\top \Phi(\boldsymbol{x}) + (\boldsymbol{v}^{(k)})^\top \boldsymbol{A}^\top \Psi(\boldsymbol{x}).$$



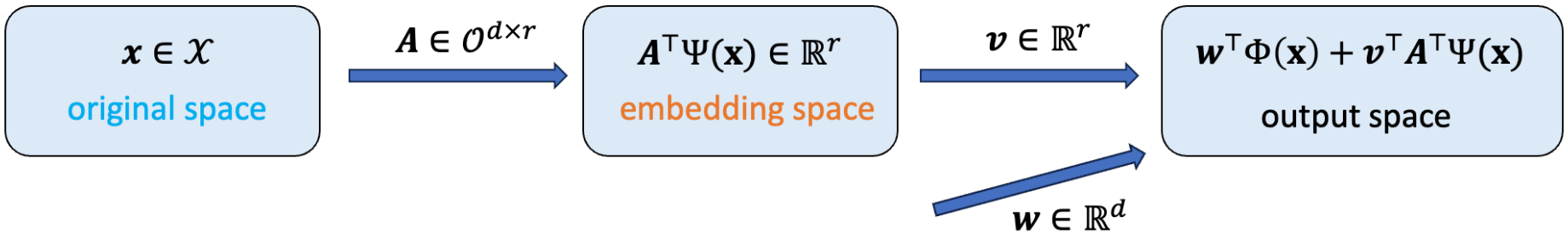
[1] Ando, R. K., & Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research, 6, 1817.

Understanding shared representation: history

The learner for task k:

$$h^{(k)}(\boldsymbol{x}) = (\boldsymbol{w}^{(k)})^\top \Phi(\boldsymbol{x}) + (\boldsymbol{v}^{(k)})^\top \boldsymbol{A}^\top \Psi(\boldsymbol{x}).$$

Ando and Zhang (2005) proposes an algorithm jointly estimating $\boldsymbol{w}^{(k)}$, $\boldsymbol{v}^{(k)}$, and $\boldsymbol{A}$, in an alternating optimization framework.



[1] Ando, R. K., & Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research, 6, 1817.

Understanding shared representation: history

Maurer et al. (2016) considers a similar setting as in Baxter (2000) and refines their analysis.

- Data $\{\boldsymbol{x}_i^{(k)}, y_i^{(k)}\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{P}^{(k)}$, for $k = 1 : K$, and $\{\mathbb{P}^{(k)}\}_{k=1}^K \overset{\text{i.i.d.}}{\sim} \mu$.
- Multi-task learning: They consider an ERM procedure to learn a representation $f \in \mathcal{F}$ shared by all tasks and learners $\{g^{(k)}\}_{k=1}^K$ in the embedding space $g(\mathcal{X})$ for each task. The final learner for task k is $g^{(k)} \circ f$. The average excess risk can be bounded by

$$\underbrace{\text{cost of learning } f \in \mathcal{F}}_{\text{depends on } nK} + \underbrace{\text{cost of learning } g \text{ given } f}_{\text{depends on } n\text{, but much smaller than single-task rate}}.$$

  Note that the second term depends on the complexity of $\mathcal{G} \circ f$, which is **much smaller** than the complexity of single-task function class $\mathcal{G} \circ \mathcal{F}$.

  Therefore, when $K \gg 1$, this rate is better than single-task rate.

[1] Maurer, A., Pontil, M., & Romera-Paredes, B. (2016). The benefit of multitask representation learning. Journal of Machine Learning Research, 17(81), 1-32.

Understanding shared representation: history

- Transfer learning: Suppose they have learned a representation $\hat{f}$. For a **new tasks** generated from $\mathbb{P} \sim \mu$, they first project the observations from $\mathcal{X}$ to $\hat{f}(\mathcal{X})$, then learn g via ERM. The **expected** target excess risk (under μ) can be bounded by

$$\underbrace{\text{cost of learning } f \in \mathcal{F}}_{\text{depends on } nK} + \underbrace{\text{cost of learning } g \text{ given } f}_{\text{depends on } n\text{, but much smaller than single-task rate}} + \underbrace{\sqrt{1/K}}_{\text{cost of generalization}}$$

  - **Question:** Is $\sqrt{1/K}$ improvable?
  - **Answer:** Unfortunately, no, under the **expected** target excess risk (under μ), as in Maurer et al. (2016).
  - Imagine a case that $\mu =$ a finite mixture of point masses. Then nK samples are equivalently K samples, which incurs a $\sqrt{1/K}$ error.

However, in practice, even borrowing the representation from one giant model (e.g. the deep NN learned on ImageNet) can help generalization. It is better to focus on the target excess risk with data from **a specific** $\mathbb{P} \sim \mu$.

---

[1] Maurer, A., Pontil, M., & Romera-Paredes, B. (2016). The benefit of multitask representation learning. Journal of Machine Learning Research, 17(81), 1-32.

Shared linear representation

From now on, we will focus on a very simple linear regression set-up with shared linear representation across tasks, to see how representation learning helps and how we break the $\sqrt{1/K}$ barrier.

- **Multi-task linear regression:** Observe $\{\boldsymbol{x}_i^{(k)}, y_i^{(k)}\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} (X^{(k)}, Y^{(k)})$, $k = 1 : K$, with

$$Y^{(k)} = (X^{(k)})^\top \boldsymbol{\beta}^{(k)*} + \epsilon^{(k)},$$

  where $\epsilon^{(k)} \perp\!\!\!\perp X^{(k)}$ are sub-Gaussian with constant variance proxy, $X^{(k)} \in \mathbb{R}^d$
- **Shared linear representation:**

$$\boldsymbol{\beta}^{(k)*} = \boldsymbol{A}^* \boldsymbol{\theta}^{(k)*},$$

  where $\boldsymbol{A}^* \in \mathcal{O}^{d\times r}$ [7], $\boldsymbol{\theta}^{(k)*} \in \mathbb{R}^r$.
- **Goal:** estimate $\{\boldsymbol{\beta}^{(k)*}\}_{k=1}^K$ and $\boldsymbol{A}^*$, generalize to new tasks
- Note that $\boldsymbol{A}^*$ and $\{\boldsymbol{\theta}^{(k)*}\}_{k=1}^K$ are **NOT** identifiable, but the column space of $\boldsymbol{A}^*$ and $\{\boldsymbol{\beta}^{(k)*}\}_{k=1}^K$ are identifiable, which are enough for our goal.

[7] $\mathcal{O}^{d\times r} := \{\boldsymbol{A} \in \mathbb{R}^{d\times r} : \boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}_r\}$.

Shared linear representation

- **Generalization to a new task (transfer learning):** In additional to the observations from K tasks, we have samples from the new task $\{\boldsymbol{x}_i^{(0)}, y_i^{(0)}\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} (X^{(0)}, Y^{(0)})$, with

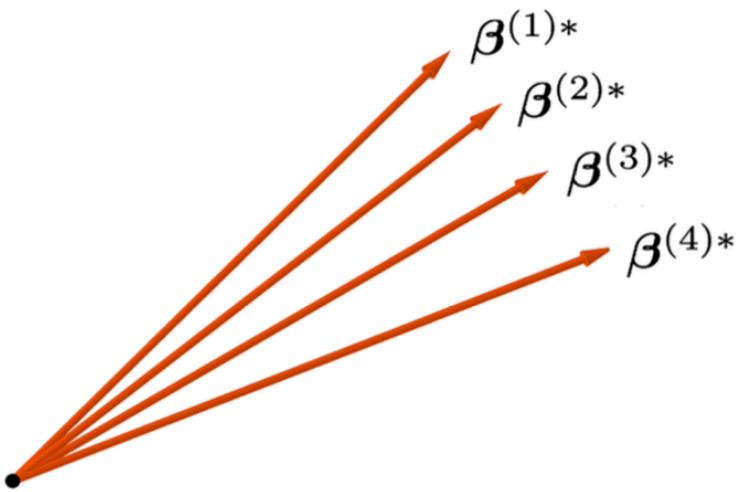
$$Y^{(0)} = (X^{(0)})^\top \boldsymbol{\beta}^{(0)*} + \epsilon^{(0)},$$

where $\epsilon^{(0)} \perp\!\!\!\perp X^{(0)}$ are sub-Gaussian with constant variance proxy, $X^{(0)} \in \mathbb{R}^d$. Assume

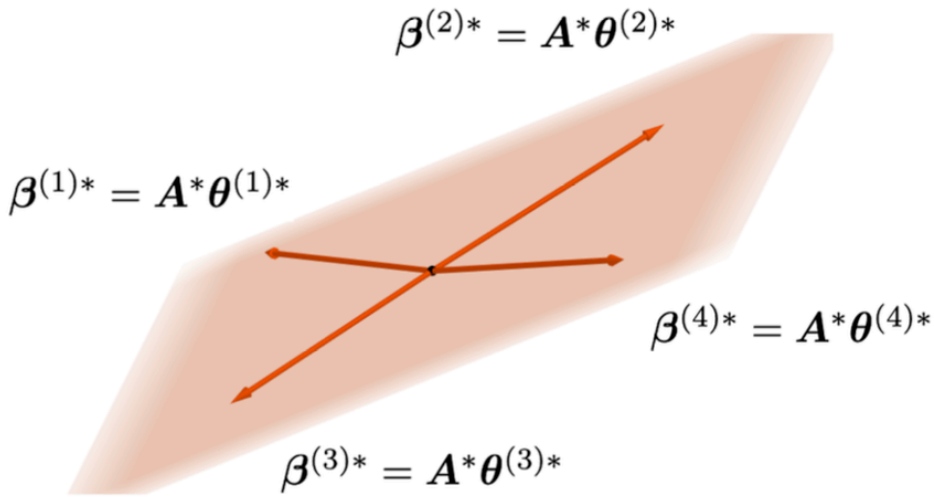
$$\boldsymbol{\beta}^{(0)*} = \boldsymbol{A}^* \boldsymbol{\theta}^{(0)*}.$$

- As we have shown before, under the generative assumption (i.e. $\exists$ a hyper distribution generating target tasks), the expected target excess risk suffers from a $\sqrt{1/K}$ term. Later we will see this is not the case in our current set-up.

Shared linear representation



Distance-based similarity



Representation-based similarity

Picture source: Tian, Y., Gu, Y., & Feng, Y. (2023). Learning from similar linear representations: adaptivity, minimaxity, and robustness. arXiv preprint arXiv:2303.17765.

The 1st method: ERM

Recall the data we have:

- **MTL setting:** $\{\boldsymbol{x}_i^{(k)}, y_i^{(k)}\}_{i=1}^n$, $k = 1 : K$
- **Generalization to new tasks (TL):** Also $\{\boldsymbol{x}_i^{(0)}, y_i^{(0)}\}_{i=1}^n$ from a new task

ERM: (Du et al., 2021)

$$\widehat{\boldsymbol{A}}, \{\widehat{\boldsymbol{\theta}}^{(k)}\}_{k=1}^K \leftarrow \underset{\boldsymbol{A}\in\mathcal{O}^{d\times r}, \{\boldsymbol{\theta}^{(k)}\}_{k=1}^K \subseteq \mathbb{R}^r}{\arg\min} \left\{ \frac{1}{2nK} \sum_{k=1}^K \sum_{i=1}^n |y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top (\boldsymbol{A}\boldsymbol{\theta}^{(k)})|^2 \right\}$$

- Tripuraneni et al. (2021) proposed a more practical variant of this ERM by replacing the $\boldsymbol{A} \in \mathcal{O}^{d\times r}$ restriction with an F-norm penalty.

$$\widehat{\boldsymbol{A}}, \{\widehat{\boldsymbol{\theta}}^{(k)}\}_{k=1}^K \leftarrow \underset{\boldsymbol{A}\in\mathcal{R}^{d\times r}, \{\boldsymbol{\theta}^{(k)}\}_{k=1}^K \subseteq \mathbb{R}^r}{\arg\min} \left\{ \frac{1}{2nK} \sum_{k=1}^K \sum_{i=1}^n |y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top (\boldsymbol{A}\boldsymbol{\theta}^{(k)})|^2 + \frac{1}{2}\|\boldsymbol{A}^\top\boldsymbol{A} - \boldsymbol{\Theta}\boldsymbol{\Theta}^\top\|_{\mathsf{F}}^2 \right\}, \text{ with } \boldsymbol{\Theta} = \{\boldsymbol{\theta}^{(k)}\}_{k=1}^K \in \mathbb{R}^{r\times K}.$$

For simplicity, we will mainly focus on the form in Du et al. (2021).

[1] Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., & Lei, Q. (2020, October). Few-Shot Learning via Learning the Representation, Provably. In International Conference on Learning Representations.

[2] Tripuraneni, N., Jin, C., & Jordan, M. (2021, July). Provable meta-learning of linear representations. In International Conference on Machine Learning (pp. 10434-10443). PMLR.

The 1st method: ERM

After learning $\widehat{\boldsymbol{A}}, \{\widehat{\boldsymbol{\theta}}^{(k)}\}_{k=1}^{K}$, we will set the estimator of regression coefficient as $\widehat{\boldsymbol{\beta}}^{(k)} = \widehat{\boldsymbol{A}}\widehat{\boldsymbol{\theta}}^{(k)}$, $k = 1 : K$.

When generalizing to new tasks, we first learn a representation $\widehat{\boldsymbol{A}}$, then it remains to learn $\boldsymbol{\theta}^{(k)*}$ via target-only ERM:

$$\widehat{\boldsymbol{\theta}}^{(0)} \leftarrow \underset{\boldsymbol{\theta} \in \mathbb{R}^r}{\arg\min} \left\{ \frac{1}{2n} \sum_{i=1}^{n} |y_i^{(0)} - (\boldsymbol{x}_i^{(0)})^\top (\widehat{\boldsymbol{A}}\boldsymbol{\theta})|^2 \right\}$$

Then we set $\widehat{\boldsymbol{\beta}}^{(0)} = \widehat{\boldsymbol{A}}\widehat{\boldsymbol{\theta}}^{(0)}$ as the target coefficient estimator.

The 1st method: ERM

Before we introduce the theory of ERM method, let's review the defition of **principal angles** between two subspaces.

Consider $\boldsymbol{A} \in \mathcal{O}^{d\times r}$ and its orthogonal $\boldsymbol{A}_\perp \in \mathcal{O}^{d\times(d-r)}$, which satisfy $\boldsymbol{A}^\top \boldsymbol{A}_\perp = \boldsymbol{0}_{r\times(d-r)}$. Then $\mathrm{Col}((\boldsymbol{A}, \boldsymbol{A}_\perp)) = \mathbb{R}^d$.

Definition 4.7.1 (Chen et al., 2021)

The i-th **principal angle** between $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{O}^{d\times r}$ is $\arccos \sigma_i(\boldsymbol{A}^\top \boldsymbol{B}) \in [0, \pi/2]$. We denote the sine of 1st principal angle as $\sin\theta(\boldsymbol{A}, \boldsymbol{B})$.

Theorem 4.7.2 (Chen et al., 2021)

The following distances between $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{O}^{d\times r}$ are "equivalent" in the sense that they differ from each other up to a constant factor:

- $\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \|\boldsymbol{A}\boldsymbol{R} - \boldsymbol{B}\|_2$;
- $\|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{B}\boldsymbol{B}^\top\|_2$;
- $\sin\theta(\boldsymbol{A}, \boldsymbol{B})$.

[1] Chen, Y., Chi, Y., Fan, J., & Ma, C. (2021). Spectral methods for data science: A statistical perspective. Foundations and Trends® in Machine Learning, 14(5), 566-806.

The 1st method: ERM

Additional assumptions:

- (Bounded coefficients) $\boldsymbol{\theta}^{(k)*} \leq C$ for all k
- (Task diversity) $\lambda_r(\frac{1}{K}\boldsymbol{\Theta}^*(\boldsymbol{\Theta}^*)^\top) \gtrsim \frac{1}{r}$, where $\boldsymbol{\Theta}^* = \{\boldsymbol{\theta}^{(k)*}\}_{k=1}^K \in \mathbb{R}^{r\times K}$ [8]

Theorem 4.7.3 (Du et al., 2021; Tripuraneni et al., 2021)

- $\sin\theta(\widehat{\boldsymbol{A}}, \boldsymbol{A}^*) \lesssim_{\mathbb{P}} r\sqrt{\frac{d}{nK}} + r\sqrt{\frac{1}{n}}.$
- $\max_{k=1:K} \|\widehat{\boldsymbol{A}}\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\beta}^{(k)*}\|_2 \lesssim_{\mathbb{P}} r\sqrt{\frac{d}{nK}} + \sqrt{r}\sqrt{\frac{r+\log K}{n}}$
- $\|\widehat{\boldsymbol{A}}\widehat{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\beta}^{(0)*}\|_2 \lesssim_{\mathbb{P}} r\sqrt{\frac{d}{nK}} + r\sqrt{\frac{1}{n}}$

[8] λ_r means the r-th largest eigenvalue.

[1] Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., & Lei, Q. (2020, October). Few-Shot Learning via Learning the Representation, Provably. In International Conference on Learning Representations.

[2] Tripuraneni, N., Jin, C., & Jordan, M. (2021, July). Provable meta-learning of linear representations. In International Conference on Machine Learning (pp. 10434-10443). PMLR.

The 1st method: ERM

Let us discuss the second and third bounds.

Main results from last slide:

- $\max_{k=1:K} \|\widehat{\boldsymbol{A}}\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\beta}^{(k)*}\|_2 \lesssim_{\mathbb{P}} \underbrace{r\sqrt{\frac{d}{nK}}}_{\text{cost of learning representation } \boldsymbol{A}^*} + \underbrace{\sqrt{r}\sqrt{\frac{r+\log K}{n}}}_{\text{cost of learning } \{\boldsymbol{\theta}^{(k)*}\}_{k=1}^K}$
- $\|\widehat{\boldsymbol{A}}\widehat{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\beta}^{(0)*}\|_2 \lesssim_{\mathbb{P}} r\sqrt{\frac{d}{nK}} + r\sqrt{\frac{1}{n}}$

- The single-task linear regression leads to $\max_{k=1:K} \|\widehat{\boldsymbol{A}}\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\beta}^{(k)*}\|_2 \lesssim_{\mathbb{P}} \sqrt{\frac{d+\log K}{n}}$.
  Hence representation MTL helps when
  - Many tasks: $K \gg r^2$;
  - Low intrinsic dimension: $d \gg r\log K + r^2$;
- The TL bound does not involve $\sqrt{1/K}$ term, which appears in the expected target excess risk when the target task is randomly generated (Maurer et al., 2016).

[1] Maurer, A., Pontil, M., & Romera-Paredes, B. (2016). The benefit of multitask representation learning. Journal of Machine Learning Research, 17(81), 1-32.

Local minimizers?

Recall the two forms of ERM:

- **Form 1:** (Du et al., 2021)

$$\widehat{\boldsymbol{A}}, \{\widehat{\boldsymbol{\theta}}^{(k)}\}_{k=1}^K \leftarrow \underset{\boldsymbol{A}\in\mathcal{O}^{d\times r}, \{\boldsymbol{\theta}^{(k)}\}_{k=1}^K \subseteq \mathbb{R}^r}{\arg\min} \left\{ \frac{1}{2nK} \sum_{k=1}^K \sum_{i=1}^n |y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top (\boldsymbol{A}\boldsymbol{\theta}^{(k)})|^2 \right\}$$

- **Form 2:** (Tripuraneni et al., 2021)

$$\widehat{\boldsymbol{A}}, \{\widehat{\boldsymbol{\theta}}^{(k)}\}_{k=1}^K \leftarrow \underset{\boldsymbol{A}\in\mathcal{R}^{d\times r}, \{\boldsymbol{\theta}^{(k)}\}_{k=1}^K \subseteq \mathbb{R}^r}{\arg\min} \left\{ \frac{1}{2nK} \sum_{k=1}^K \sum_{i=1}^n |y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top (\boldsymbol{A}\boldsymbol{\theta}^{(k)})|^2 + \frac{1}{2} \|\boldsymbol{A}^\top \boldsymbol{A} - \boldsymbol{\Theta}\boldsymbol{\Theta}^\top\|_{\mathsf{F}}^2 \right\}, \text{ with } \boldsymbol{\Theta} = \{\boldsymbol{\theta}^{(k)}\}_{k=1}^K \in \mathbb{R}^{r\times K}.$$

Both forms are non-convex. But:

Theorem 4.7.4 (Tripuraneni et al., 2021; Tian et al., 2023)

The MTL bounds we presented before hold for any ***local minimizers***. [9]

[9] Rigorously, Form 2 doesn't have local minimizers in a certain regime. Form 1 doesn't have local minimizers in the entire regime. Tripuraneni et al. (2021) and Tian et al. (2023) prove the cases of Forms 2 and 1, respectively.

[1] Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., & Lei, Q. (2020, October). Few-Shot Learning via Learning the Representation, Provably. In International Conference on Learning Representations.

[2] Tripuraneni, N., Jin, C., & Jordan, M. (2021, July). Provable meta-learning of linear representations. In International Conference on Machine Learning (pp. 10434-10443). PMLR.

[3] Tian, Y., Gu, Y., & Feng, Y. (2023). Learning from similar linear representations: adaptivity, minimaxity, and robustness. arXiv preprint arXiv:2303.17765.

The 2nd method: method of moments

Tripuraneni et al. (2021) proposes the method-of-moments estimator for $\boldsymbol{A}^*$.

Method of moments: (Tripuraneni et al., 2021)

(i) $\widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{U}}^\top \leftarrow$ PCA of $\frac{1}{nK}\sum_{k=1}^K\sum_{i=1}^n (y_i^{(k)})^2 \boldsymbol{x}_i^{(k)}(\boldsymbol{x}_i^{(k)})^\top$

(ii) $\widehat{\boldsymbol{A}} \leftarrow$ top-r columns of $\widehat{\boldsymbol{U}}$

(iii) $\widehat{\boldsymbol{\theta}}^{(k)} \leftarrow \underset{\boldsymbol{\theta}\in\mathbb{R}^r}{\arg\min}\left\{\frac{1}{2n}\sum_{i=1}^n |y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top(\widehat{\boldsymbol{A}}\boldsymbol{\theta})|^2\right\}$, for $k = 1:K$

- **Caveat**: Only works when $\boldsymbol{x}_i^{(k)} \overset{\text{i.i.d.}}{\sim} X^{(k)} \sim N(\boldsymbol{0}, \sigma_k^2 \boldsymbol{I}_d)$ [10]
- The intuition comes from the second-order Stein's method (Janzamin et al., 2014), which implies

$$\mathbb{E}\big[\frac{1}{nK}\sum_{k=1}^K\sum_{i=1}^n (y_i^{(k)})^2 \boldsymbol{x}_i^{(k)}(\boldsymbol{x}_i^{(k)})^\top\big] = \frac{1}{K}\boldsymbol{A}^*\boldsymbol{\Theta}^*(\boldsymbol{\Theta}^*)^\top(\boldsymbol{A}^*)^\top.$$

  The top-r eigenspace will recover $\text{Col}(\boldsymbol{A}^*)$.

[10] Tripuraneni et al. (2021) assumes standard Gaussian, but it can be extended to the isotropic case.

[1] Tripuraneni, N., Jin, C., & Jordan, M. (2021, July). Provable meta-learning of linear representations. In International Conference on Machine Learning (pp. 10434-10443). PMLR.

[2] Janzamin, M., Sedghi, H., & Anandkumar, A. (2014). Score function features for discriminative learning: Matrix and tensor framework. arXiv preprint arXiv:1412.2863.

The 2nd method: method of moments

Theorem 4.7.5 (Tripuraneni et al., 2021)

Under similar assumptions as before (together with the standard Gaussian $X^{(k)}$):

- $\sin\theta(\widehat{\boldsymbol{A}}, \boldsymbol{A}^*) \lesssim_{\mathbb{P}} r\sqrt{\frac{d}{nK}}$.
- $\max_{k=1:K} \|\widehat{\boldsymbol{A}}\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\beta}^{(k)*}\|_2 \lesssim_{\mathbb{P}} r\sqrt{\frac{d}{nK}} + \sqrt{\frac{r+\log K}{n}}$
- $\|\widehat{\boldsymbol{A}}\widehat{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\beta}^{(0)*}\|_2 \lesssim_{\mathbb{P}} r\sqrt{\frac{d}{nK}} + \sqrt{\frac{r}{n}}$

- The bounds have better dependence on r compared to the previous results for ERM method.

[1] Tripuraneni, N., Jin, C., & Jordan, M. (2021, July). Provable meta-learning of linear representations. In International Conference on Machine Learning (pp. 10434-10443). PMLR.

The 3rd method: concatenation method

- The method of moments estimates $\boldsymbol{B}^*(\boldsymbol{B}^*)^\top$ through the moment estimator then conducts PCA, where $\boldsymbol{B}^* = \{\boldsymbol{\beta}^{(k)*}\}_{k=1}^K \in \mathbb{R}^{d\times K}$
- Can we estimate each column of $\boldsymbol{B}^*$ by single-task OLS, then concatenate them to a matrix?

Concatenation method: (Tian et al., 2023)

(i) $\widehat{\boldsymbol{\beta}}^{(k)} \leftarrow \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\arg\min} \left\{ \frac{1}{2n}\sum_{i=1}^n |y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top \boldsymbol{\beta}|^2 \right\}$, for $k = 1:K$, then contatenate into $\widehat{\boldsymbol{B}} = \{\widehat{\boldsymbol{\beta}}^{(k)}\}_{k=1}^K \in \mathbb{R}^{d\times K}$

(ii) $\widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{U}}^\top \leftarrow$ PCA of $\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top$

(iii) $\widehat{\boldsymbol{A}} \leftarrow$ top-r columns of $\widehat{\boldsymbol{U}}$

(iv) $\widehat{\boldsymbol{\theta}}^{(k)} \leftarrow \underset{\boldsymbol{\theta}\in\mathbb{R}^r}{\arg\min} \left\{ \frac{1}{2n}\sum_{i=1}^n |y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top (\widehat{\boldsymbol{A}}\boldsymbol{\theta})|^2 \right\}$, for $k = 1:K$

[1] Tian, Y., Gu, Y., & Feng, Y. (2023). Learning from similar linear representations: adaptivity, minimaxity, and robustness. arXiv preprint arXiv:2303.17765.

The 3rd method: concatenation method

Theorem 4.7.6 (Tian et al., 2023)

Under similar assumptions for ERM method:

- $\sin\theta(\widehat{\boldsymbol{A}}, \boldsymbol{A}^*) \lesssim_{\mathbb{P}} \sqrt{\frac{dr}{nK}} + \sqrt{\frac{r}{n}}$.
- $\max_{k=1:K} \|\widehat{\boldsymbol{A}}\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\beta}^{(k)*}\|_2 \lesssim_{\mathbb{P}} \sqrt{\frac{dr}{nK}} + \sqrt{\frac{r+\log K}{n}}$
- $\|\widehat{\boldsymbol{A}}\widehat{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\beta}^{(0)*}\|_2 \lesssim_{\mathbb{P}} \sqrt{\frac{dr}{nK}} + \sqrt{\frac{r}{n}}$

- The second and third bounds have better dependence on r compared to the previous results for ERM method and method of moments.
- They turn out to be optimal.

Theorem 4.7.7 (Tian et al., 2023)

The bounds on $\max_{k=1:K} \|\widehat{\boldsymbol{A}}\widehat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\beta}^{(k)*}\|_2$ *and* $\|\widehat{\boldsymbol{A}}\widehat{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\beta}^{(0)*}\|_2$ *are* ***minimax optimal****.*

[1] Tian, Y., Gu, Y., & Feng, Y. (2023). Learning from similar linear representations: adaptivity, minimaxity, and robustness. arXiv preprint arXiv:2303.17765.

Discussion of three methods

- Comparison of three methods:
  - The ERM method is the most intuitive one;
  - The method of moments has better rates then ERM, and the best rate for estimating $\boldsymbol{A}^*$, but requires isotropic Gaussian predictors;
  - The concatenation method achieves minimax rates for MTL and TL.
- The method of moments and concatenation method are motivated by the spectrum decomposition of $\boldsymbol{B}^*(\boldsymbol{B}^*)^\top$. This is closely connected to the subspace recovery in single-index models (Brillinger, 2012) and multi-index models (Samarov, 1993; Yang et al., 2017).

  See Yuan et al. (2023) for an efficient algorithm using this idea to recover the subspace in multi-index model and more related discussions therein.

---

[1] Brillinger, D. R. (2012). A generalized linear model with "Gaussian" regressor variables. Selected Works of David Brillinger, 589-606.

[2] Samarov, A. M. (1993). Exploring regression structure using nonparametric functional estimation. Journal of the American Statistical Association, 88(423), 836-847.

[3] Yang, Z., Balasubramanian, K., Wang, Z., & Liu, H. (2017). Learning non-gaussian multi-index model via second-order stein's method. Advances in Neural Information Processing Systems, 30, 6097-6106.

[4] Yuan, G., Xu, M., Kpotufe, S., & Hsu, D. (2023). Efficient estimation of the central mean subspace via smoothed gradient outer products. arXiv preprint arXiv:2312.15469.

Discussion of three methods

- **Unknown intrinsic dimension r:**
  - Almost all existing works assume r is known
  - Tian et al. (2023) develops a method to estimate r, with nearly no additional assumptions. The idea is based on thresholding the singular value of $\widehat{\boldsymbol{B}}/\sqrt{K}$, where $\widehat{\boldsymbol{B}} \in \mathbb{R}^{d\times K}$ is the concatenated coefficient matrix

    Intuition: $\sigma_r(\boldsymbol{B}^*/\sqrt{K}) \gtrsim 1/\sqrt{r}$, $\sigma_{r+1}(\boldsymbol{B}^*/\sqrt{K}) = 0$, [11]
    where $\boldsymbol{B}^* = \{\boldsymbol{\beta}^{(k)*}\}_{k=1}^K \in \mathbb{R}^{d\times K}$, when $h = \epsilon = 0$.

[11] $\sigma_r(\boldsymbol{B}^*/\sqrt{K})$ refers to the r-th largest singular value of $\boldsymbol{B}^*/\sqrt{K}$.

[1] Tian, Y., Gu, Y., & Feng, Y. (2023). Learning from similar linear representations: adaptivity, minimaxity, and robustness. arXiv preprint arXiv:2303.17765.

Discussion of three methods

- Thekumparampil et al. (2021) proposes an alternating gradient descent algorithm to solve the ERM, which can achieve the same accuracy with polynomial time complexity.
- **Non-linear models and non-linear representations:**
  - Du et al. (2021) also considers linear model with non-linear representations.
  - Tripuraneni et al. (2020) considers non-linear models and non-linear representations. They use logistic regression and deep neural nets with tanh activation function as running examples.
  - Tian et al. (2023) considers also considers GLMs and non-linear regression models with linear representations.

---

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

[2] Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., & Lei, Q. (2020, October). Few-Shot Learning via Learning the Representation, Provably. In International Conference on Learning Representations.

[3] Tripuraneni, N., Jordan, M., & Jin, C. (2020). On the theory of transfer learning: The importance of task diversity. Advances in neural information processing systems, 33, 7852-7862.

[4] Tian, Y., Gu, Y., & Feng, Y. (2023). Learning from similar linear representations: adaptivity, minimaxity, and robustness. arXiv preprint arXiv:2303.17765.

§4.7: Domain adaptation with representation learning

- §4.7.1 Representation learning and fine tuning
- §4.7.2 Learning the shared representation across tasks
- §4.7.3 Go beyond the shared representation

Go beyond the shared representation

- We have assumed that all tasks share the same representation in the last section. However, freezing the representation may not always be a good idea.

  Question: Can we allow **similar but not exactly the same** representations across tasks?

- We have seen the robustness against adversarial contaminations (or outlier tasks) for ℓ_2-based regularization (Duan and Wang, 2022). Until now, we haven't explored such contaminated setting under the representation MTL yet.

  Question: Can we develop a method that is **robust to a small fraction of outlier tasks**?

Go beyond the shared representation

There have been a few attempts on both questions. Let's still focus on the multi-task linear regression setting $Y^{(k)} = (X^{(k)})^\top \boldsymbol{\beta}^{(k)*} + \epsilon^{(k)}$, where $\boldsymbol{\beta}^{(k)*} \in \mathbb{R}^d$, $\epsilon^{(k)} \perp\!\!\!\perp X^{(k)}$ are sub-Gaussian.

- Chua et al. (2021) considers

$$\boldsymbol{\beta}^{(k)*} = (\boldsymbol{A}^* + \boldsymbol{\Delta}^{(k)*})\boldsymbol{\theta}^{(k)*},$$

  where $\boldsymbol{A}^*, \boldsymbol{\Delta}^{(k)*} \in \mathbb{R}^{d\times r}$.

- Duan and Wang (2022) considers

$$\max_{k=1:K} \|\boldsymbol{\beta}^{(k)*} - \boldsymbol{A}^*\boldsymbol{\theta}^{(k)*}\|_2 \leq h,$$

  where $h >\geq 0$ is an unknown parameter.

- Tian et al. (2023) considers $\boldsymbol{\beta}^{(k)*} = \boldsymbol{A}^{(k)*}\boldsymbol{\theta}^{(k)*}$ for $k \in S$ and

$$\max_{k\in S} \|\boldsymbol{A}^{(k)*}(\boldsymbol{A}^{(k)*})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \leq h,$$

  where $S \subseteq [K]$ is an unknown subset, and $h \in [0, 1]$ is an unknown parameter. The tasks outside S can follow **arbitrary** distribution.

[1] Duan, Y., & Wang, K. (2022). Adaptive and robust multi-task learning. arXiv preprint arXiv:2202.05250. (version 2)

[2] Chua, K., Lei, Q., & Lee, J. D. (2021). How fine-tuning allows for effective meta-learning. Advances in Neural Information Processing Systems, 34, 8871-8884.

[3] Tian, Y., Gu, Y., & Feng, Y. (2023). Learning from similar linear representations: adaptivity, minimaxity, and robustness. arXiv preprint arXiv:2303.17765.

Go beyond the shared representation

We will discuss the relationship between these three settings later. For now, let's focus on the setting of Tian et al. (2023).

- **Multi-task linear regression setting:** $\{\boldsymbol{x}_i^{(k)}, y_i^{(k)}\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} (X^{(k)}, Y^{(k)})$, and

$$Y^{(k)} = (X^{(k)})^\top \boldsymbol{\beta}^{(k)*} + \epsilon^{(k)},$$

  where $\boldsymbol{\beta}^{(k)*} \in \mathbb{R}^d$, $\epsilon^{(k)} \perp\!\!\!\perp X^{(k)}$ are sub-Gaussian, for $k \in S \subseteq [K]$.

- **Similar representations in S:**

$$\max_{k \in S} \|\boldsymbol{A}^{(k)*}(\boldsymbol{A}^{(k)*})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \leq h, \quad [12]$$

  where $h \in [0, 1]$ is an unknown parameter.

- **Outlier tasks/Contaminations:** The tasks in $S^c := [K] \backslash S$ can follow arbitrary distribution, in the sense that

$$\{\boldsymbol{x}_i^{(k)}, y_i^{(k)}\}_{i=1:n, k \in S^c} \sim \mathbb{Q} \text{ (arbitrary distribution)}.$$

  Denote $\epsilon = |S^c|/K$ as the **contamination proportion**.

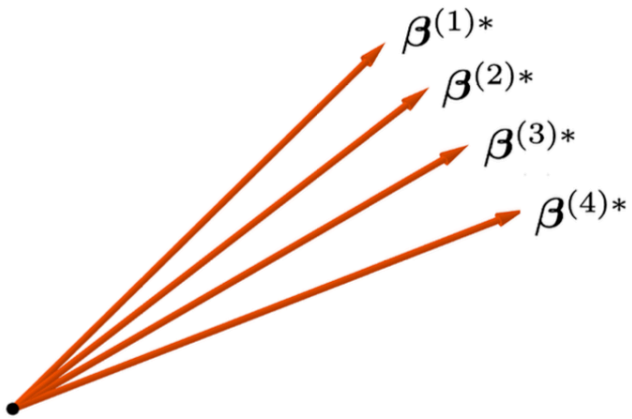
[12] Recall that $\|\boldsymbol{A}^{(k)*}(\boldsymbol{A}^{(k)*})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \asymp \sin\theta(\boldsymbol{A}^{(k)*}, \overline{\boldsymbol{A}})$

Connections to other works

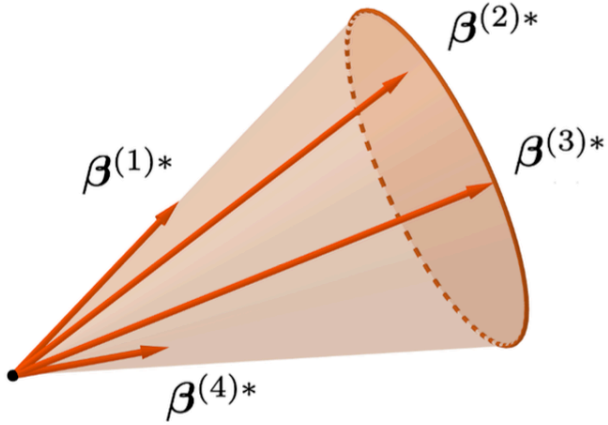
- Gu et al. (2022) considers a special case of $r = 1$ under a transfer learning setting, where $\sin\theta(\boldsymbol{\beta}^{(0)*}, \boldsymbol{\beta}^{(1)*})$ is small. They call it "angle-based similarity".
- The settings of Chua et al. (2021) and Duan and Wang (2022) are equivalent. And it is more general than our current setting (with no contamination). But our upcoming revision pushes our setting to a broader situation.
- We will focus on the current simple setting for convenience. We will see that our estimation error is better than the results in Chua et al. (2021) and Duan and Wang (2022).

[1] Gu, T., Han, Y., & Duan, R. (2022). Robust angle-based transfer learning in high dimensions. arXiv preprint arXiv:2210.12759.

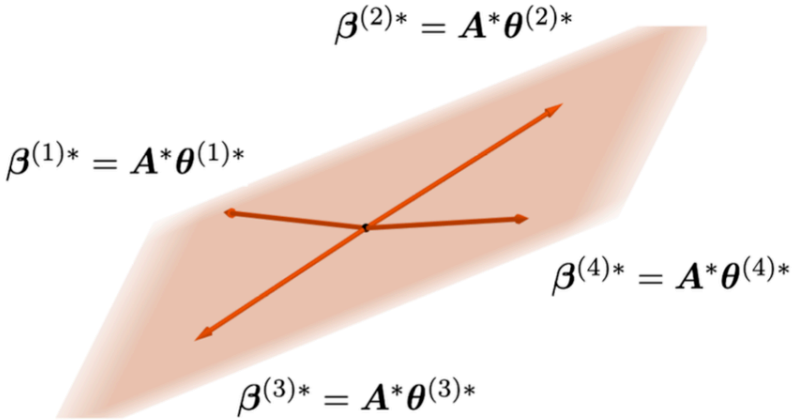
Different similarity metrics



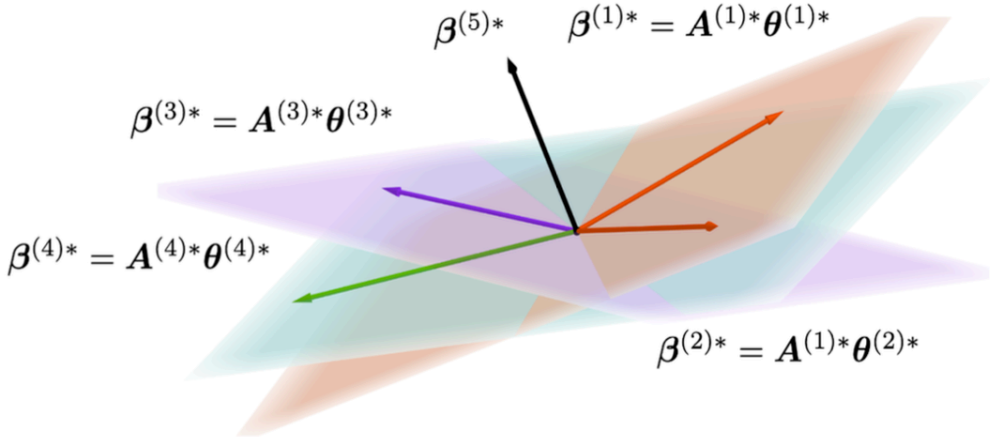
Distance-based similarity



Angle-based similarity



The same representation across tasks



Similar representation with outliers

Picture source: Tian, Y., Gu, Y., & Feng, Y. (2023). Learning from similar linear representations: adaptivity, minimaxity, and robustness. arXiv preprint arXiv:2303.17765.

Penalized ERM

Penalized ERM: (Tian et al., 2023)

(i) $\{\widehat{\boldsymbol{A}}^{(k)}\}_{k=1}^K, \widehat{\overline{\boldsymbol{A}}}, \{\widehat{\boldsymbol{\theta}}^{(k)}\}_{k=1}^K \leftarrow \underset{\substack{\boldsymbol{A}^{(k)}, \overline{\boldsymbol{A}} \in \mathcal{O}^{d\times r} \\ \boldsymbol{\theta}^{(k)} \in \mathbb{R}^r}}{\arg\min} \left\{ \frac{1}{2nK} \sum_{k=1}^K \sum_{i=1}^n |y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top (\boldsymbol{A}^{(k)} \boldsymbol{\theta}^{(k)})|^2 + \frac{\lambda_1}{\sqrt{nK}} \sum_{k=1}^K \|\boldsymbol{A}^{(k)} (\boldsymbol{A}^{(k)})^\top - \overline{\boldsymbol{A}} (\overline{\boldsymbol{A}})^\top\|_2 \right\}$

(ii) $\widehat{\boldsymbol{\beta}}^{(k)} \leftarrow \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\arg\min} \left\{ \frac{1}{2n} \sum_{i=1}^n |y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top \boldsymbol{\beta}^{(k)}|^2 + \frac{\lambda_2}{\sqrt{n}} \|\boldsymbol{\beta} - \widehat{\boldsymbol{A}}^{(k)} \widehat{\boldsymbol{\theta}}^{(k)}\|_2 \right\}$, for $k = 1 : K$

- The penalization in step (i) is motivated by the similarity assumption
- The penalization in step (ii) is an ℓ_2-penalty.

Penalized ERM

Theorem 4.7.1 (Tian et al., 2023)

Under similar conditions as we assumed for ERM method when representation is the same across tasks:

$$\max_{k\in S}\|\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}^{(k)*}\|_2 \lesssim_{\mathbb{P}} \left\{ r\sqrt{\frac{d}{nK}}+\sqrt{r}\sqrt{\frac{r+\log K}{n}}+\sqrt{r}h+r^{7/4}\epsilon\sqrt{\frac{d+\log K}{n}} \right\} \wedge \sqrt{\frac{d+\log K}{n}}.$$

- **Error decomposition:**
  - $r\sqrt{\frac{d}{nK}}$: cost of learning the "central" representation $\overline{\boldsymbol{A}}$
  - $\sqrt{r}\sqrt{\frac{r+\log K}{n}}$: cost of learning the low-dimensional coefficients $\boldsymbol{\theta}^{(k)*}$'s
  - $\sqrt{r}h$: heterogeneity between representations $\boldsymbol{A}^{(k)*}$'s
  - $r^{7/4}\epsilon\sqrt{\frac{d+\log K}{n}}$: cost of contamination
  - $\sqrt{\frac{d+\log K}{n}}$: single-task rate

Penalized ERM

Main result from last slide:

$$\max_{k\in S}\|\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}^{(k)*}\|_2 \lesssim_{\mathbb{P}} \left\{ r\sqrt{\frac{d}{nK}}+\sqrt{r}\sqrt{\frac{r+\log K}{n}}+\sqrt{r}h+r^{7/4}\epsilon\sqrt{\frac{d+\log K}{n}} \right\} \wedge \sqrt{\frac{d+\log K}{n}}.$$

- The rate is faster than the single-task rate, when
  - Many tasks: $K \gg r^2$;
  - Low intrinsic dimension: $d \gg r^2$;
  - Sufficient similar representations: $\sqrt{r}h \ll \sqrt{\frac{d+\log K}{n}}$;
  - Not too many outlier tasks: $r^{7/4}\epsilon \ll 1$
- The rate is never worse than the single-task rate.
- The minimax rate is $\left\{ \sqrt{\frac{rd}{nK}}+\sqrt{\frac{r+\log K}{n}}+h+r\epsilon\sqrt{\frac{1}{n}} \right\} \wedge \sqrt{\frac{d+\log K}{n}}$
  - The dependence on r is not optimal
  - The full dimension d in the contamination-related term may be not unavoidable for biased regularization. See our discussions in Section §4.2.3.

Penalized concatenation method

Previously, we see that concatenation method achieves a better estimation error than ERM. Can we use the concatenation method here?

Penalized concatenation method: (Tian et al., 2023)

(i) $\widehat{\boldsymbol{\beta}}^{(k)} \leftarrow \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\arg\min} \left\{ \frac{1}{2n} \sum_{i=1}^n |y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top \boldsymbol{\beta}|^2 \right\}$, for $k = 1 : K$, then contatenate into $\widehat{\boldsymbol{B}} = \{\widehat{\boldsymbol{\beta}}^{(k)}\}_{k=1}^K \in \mathbb{R}^{d\times K}$

(ii) $\widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{U}}^\top \leftarrow$ PCA of $\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top$

(iii) $\widehat{\boldsymbol{A}} \leftarrow$ top-r columns of $\widehat{\boldsymbol{U}}$

(iv) $\widehat{\boldsymbol{\theta}}^{(k)} \leftarrow \underset{\boldsymbol{\theta}\in\mathbb{R}^r}{\arg\min} \left\{ \frac{1}{2n} \sum_{i=1}^n |y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top (\widehat{\boldsymbol{A}}\boldsymbol{\theta})|^2 \right\}$, for $k = 1 : K$

(v) $\widehat{\boldsymbol{\beta}}^{(k)} \leftarrow \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\arg\min} \left\{ \frac{1}{2n} \sum_{i=1}^n |y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^\top \boldsymbol{\beta}^{(k)}|^2 + \frac{\lambda_2}{\sqrt{n}} \|\boldsymbol{\beta} - \widehat{\boldsymbol{A}}^{(k)}\widehat{\boldsymbol{\theta}}^{(k)}\|_2 \right\}$, $k = 1 : K$

- (i)-(iv) are the same as the concatenation method we saw before.

Penalized concatenation method

We have the following result for the penalized concatenation method, without contaminations ($\epsilon = 0$).

Theorem 4.7.2 (Tian et al., 2023)

Under similar conditions as we assumed for ERM method when representation is the same across tasks, when $\epsilon = 0$, $\lambda_1 \asymp r^{3/4}\sqrt{d + \log K}$, $\lambda_2 \asymp \sqrt{d + \log K}$*:*

$$\max_{k \in S} \|\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)*}\|_2 \lesssim_{\mathbb{P}} \left\{ \sqrt{\frac{rd}{nK}} + \sqrt{\frac{r + \log K}{n}} + \sqrt{r}h \right\} \wedge \sqrt{\frac{d + \log K}{n}}.$$

- When $\epsilon = 0$, the minimax rate is $\left\{ \sqrt{\frac{rd}{nK}} + \sqrt{\frac{r+\log K}{n}} + h \right\} \wedge \sqrt{\frac{d+\log K}{n}}$
- When $X^{(k)}$'s have the same covariance matrices, $\sqrt{r}h$ can be replaced by h multiplied by a conditional number, which makes the penalized concatenation method nearly optimal
- For more details, please wait for the upcoming version of our paper.

[1] Tian, Y., Gu, Y., & Feng, Y. (2023). Learning from similar linear representations: adaptivity, minimaxity, and robustness. arXiv preprint arXiv:2303.17765.

References I

Ando, R. K. and Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817.

Bastani, H. (2021). Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67(5):2964–2984.

Baxter, J. (2000). A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198.

Belsley, D. A., Kuh, E., and Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons.

Brillinger, D. R. (2012). A generalized linear model with "gaussian" regressor variables. *Selected Works of David Brillinger*, pages 589–606.

Cai, T. T., Ma, J., and Zhang, L. (2019). Chime: Clustering of high-dimensional gaussian mixtures with em algorithm and its optimality. *The Annals of Statistics*, 47(3):1234–1267.

Cai, T. T., Namkoong, H., and Yadlowsky, S. (2023). Diagnosing model performance under distribution shift. *arXiv preprint arXiv:2303.02011*.

References II

Chen, A., Owen, A. B., Shi, M., et al. (2015). Data enriched linear regression. *Electronic journal of statistics*, 9(1):1078–1112.

Chen, Y., Chi, Y., Fan, J., Ma, C., et al. (2021). Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806.

Chen, Y., Li, X., Liu, J., and Ying, Z. (2023). Item response theory: a statistical framework for educational and psychological measurement. *Statistical Science*.

Chua, K., Lei, Q., and Lee, J. D. (2021). How fine-tuning allows for effective meta-learning. *Advances in Neural Information Processing Systems*, 34:8871–8884.

Cook, R. D. (2000). Detection of influential observation in linear regression. *Technometrics*, 42(1):65–68.

Dai, D., Rigollet, P., and Zhang, T. (2012). Deviation optimal learning using greedy q-aggregation. *Annals of Statistics*, 40(3):1878–1905.

Denevi, G., Ciliberto, C., Stamos, D., and Pontil, M. (2018). Learning to learn around a common mean. *Advances in neural information processing systems*, 31.

References III

Dieuleveut, A., Fort, G., Moulines, E., and Robin, G. (2021). Federated-em with heterogeneity mitigation and variance reduction. *Advances in Neural Information Processing Systems*, 34:29553–29566.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655. PMLR.

Donoho, D. and Montanari, A. (2016). High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166:935–969.

Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., and Lei, Q. (2021). Few-shot learning via learning the representation, provably. In *9th International Conference on Learning Representations, ICLR 2021*.

Duan, Y. and Wang, K. (2022). Adaptive and robust multi-task learning. *arXiv preprint arXiv:2202.05250*.

References IV

Evgeniou, T. and Pontil, M. (2004). Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Gannaz, I. (2007). Robust estimation and wavelet thresholding in partially linear models. *Statistics and Computing*, 17:293–310.

Gross, S. M. and Tibshirani, R. (2016). Data shared lasso: A novel tool to discover uplift. *Computational statistics & data analysis*, 101:226–235.

Gu, Q., Li, Z., and Han, J. (2011). Learning a kernel for multi-task clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 368–373.

Gu, T., Han, Y., and Duan, R. (2022). Robust angle-based transfer learning in high dimensions. *arXiv preprint arXiv:2210.12759*.

Guo, Z., Li, X., Han, L., and Cai, T. (2023). Robust inference for federated meta-learning. *arXiv preprint arXiv:2301.00718*.

References V

Huang, J., Wang, M., and Wu, Y. (2022). Estimation and inference for transfer learning with high-dimensional quantile regression. *arXiv preprint arXiv:2211.14578*.

Huang, X., Xu, K., Lee, D., Hassani, H., Bastani, H., and Dobriban, E. (2023). Optimal heterogeneous collaborative linear regression and contextual bandits. *arXiv preprint arXiv:2306.06291*.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 73–101.

Jalali, A., Ravikumar, P., and Sanghavi, S. (2013). A dirty model for multiple sparse regression. *IEEE Transactions on Information Theory*, 59(12):7947–7968.

Jalali, A., Sanghavi, S., Ruan, C., and Ravikumar, P. (2010). A dirty model for multi-task learning. *Advances in neural information processing systems*, 23.

Janzamin, M., Sedghi, H., and Anandkumar, A. (2014). Score function features for discriminative learning: Matrix and tensor framework. *arXiv preprint arXiv:1412.2863*.

Jin, J., Yan, J., Aseltine, R. H., and Chen, K. (2024). Transfer learning with large-scale quantile regression. *Technometrics*, (just-accepted):1–30.

References VI

Kuzborskij, I. and Orabona, F. (2013). Stability and hypothesis transfer learning. In *International Conference on Machine Learning*, pages 942–950. PMLR.

Kuzborskij, I. and Orabona, F. (2017). Fast rates by transferring from auxiliary hypotheses. *Machine Learning*, 106:171–195.

Kwon, Y. and Zou, J. (2022). Beta shapley: a unified and noise-reduced data valuation framework for machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 8780–8802. PMLR.

Li, J. and Song, Y. (2024). Transfer learning with high dimensional composite quantile regression. *Journal of Statistical Computation and Simulation*, pages 1–18.

Li, M., Tian, Y., Feng, Y., and Yu, Y. (2024). Federated transfer learning with differential privacy. *arXiv preprint arXiv:2403.11343*.

Li, S., Cai, T. T., and Li, H. (2021). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 1–25.

Li, S., Cai, T. T., and Li, H. (2022). Estimation and inference with proxy data and its genetic applications. *arXiv preprint arXiv:2201.03727*.

References VII

Li, S., Zhang, L., Cai, T. T., and Li, H. (2023). Estimation and inference for high-dimensional generalized linear models with knowledge transfer. *Journal of the American Statistical Association*, pages 1–12.

Li, X. and Bilmes, J. (2007). A bayesian divergence prior for classiffier adaptation. In *Artificial Intelligence and Statistics*, pages 275–282. PMLR.

Lin, H. and Reimherr, M. (2024a). On hypothesis transfer learning of functional linear models. *arXiv preprint arXiv:2206.04277 (accepted by ICML 2024)*.

Lin, H. and Reimherr, M. (2024b). Smoothness adaptive hypothesis transfer learning. *arXiv preprint arXiv:2402.14966. (accepted by ICML 2024)*.

Liu, J., Wang, T., Cui, P., and Namkoong, H. (2023). On the need for a language describing distribution shifts: Illustrations on tabular datasets. *Advances in Neural Information Processing Systems*, 36.

Lounici, K., Pontil, M., van de Geer, S., and Tsybakov, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, pages 2164–2204.

References VIII

Maity, S., Sun, Y., and Banerjee, M. (2022). Meta-analysis of heterogeneous data: integrative sparse regression in high-dimensions. *The Journal of Machine Learning Research*, 23(1):8975–9024.

Marfoq, O., Neglia, G., Bellet, A., Kameni, L., and Vidal, R. (2021). Federated multi-task learning under a mixture of distributions. *Advances in Neural Information Processing Systems*, 34:15434–15447.

Maurer, A., Pontil, M., and Romera-Paredes, B. (2016). The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32.

McCann, L. and Welsch, R. E. (2007). Robust variable selection using least angle regression and elemental set sampling. *Computational Statistics & Data Analysis*, 52(1):249–257.

Negahban, S. N. and Wainwright, M. J. (2011). Simultaneous support recovery in high dimensions: Benefits and perils of block ℓ_1/ℓ_∞-regularization. *IEEE Transactions on Information Theory*, 57(6):3841–3863.

Ollier, E. and Viallon, V. (2017). Regression modelling on stratified data with the lasso. *Biometrika*, 104(1):83–96.

References IX

Orabona, F., Castellini, C., Caputo, B., Fiorilla, A. E., and Sandini, G. (2009). Model adaptation with least-squares svm for adaptive hand prosthetics. In *2009 IEEE international conference on robotics and automation*, pages 2897–2903. IEEE.

Qiao, S., He, Y., and Zhou, W. (2023). Transfer learning for high-dimensional quantile regression with statistical guarantee. *Transactions on Machine Learning Research*.

Rigollet, P. and Tsybakov, A. B. (2012). Sparse estimation by exponential weighting. *Statistical Science*, 27(4):558–575.

Samarov, A. M. (1993). Exploring regression structure using nonparametric functional estimation. *Journal of the American Statistical Association*, 88(423):836–847.

Scarlatos, A., Brinton, C., and Lan, A. (2022). Process-bert: A framework for representation learning on educational process data. *arXiv preprint arXiv:2204.13607*.

References X

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer.

She, Y. and Owen, A. B. (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106(494):626–639.

T Dinh, C., Tran, N., and Nguyen, J. (2020). Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405.

Taylor, J. and Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634.

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. and Feng, Y. (2023a). Comments on: Statistical inference and large-scale multiple testing for high-dimensional regression models. *Test*, 32(4):1172–1176.

References XI

Tian, Y. and Feng, Y. (2023b). Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 118(544):2684–2697.

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

Tian, Y., Weng, H., and Feng, Y. (2022). Unsupervised multi-task and transfer learning on gaussian mixture models. *arXiv preprint arXiv:2209.15224*.

Tian, Y., Weng, H., and Feng, Y. (2024). Towards the theory of unsupervised federated learning: Non-asymptotic analysis of federated em algorithms. *arXiv preprint arXiv:2310.15330 (accepted by ICML 2024)*.

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

Tripuraneni, N., Jordan, M., and Jin, C. (2020). On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33:7852–7862.

References XII

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Vapnik, V. (1996). *The nature of statistical learning theory*. Springer science & business media.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202.

Wei, S., Moore, R., Zhang, H., Xie, Y., and Kamaleswaran, R. (2023). Transfer causal learning: Causal effect estimation with knowledge transfer. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*.

References XIII

Wu, Y., Zhang, S., Yu, W., Liu, Y., Gu, Q., Zhou, D., Chen, H., and Cheng, W. (2023). Personalized federated learning under mixture of distributions. In *International Conference on Machine Learning*, pages 37860–37879. PMLR.

Xu, K. and Bastani, H. (2021). Learning across bandits in high dimension via robust statistics. *arXiv preprint arXiv:2112.14233*.

Yang, J., Yan, R., and Hauptmann, A. G. (2007). Cross-domain video concept detection using adaptive svms. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 188–197.

Yang, Y., Ma, Z., Yang, Y., Nie, F., and Shen, H. T. (2014). Multitask spectral clustering by exploring intertask correlation. *IEEE transactions on cybernetics*, 45(5):1083–1094.

Yang, Z., Balasubramanian, K., Wang, Z., and Liu, H. (2017). Learning non-gaussian multi-index model via second-order stein's method. *Advances in Neural Information Processing Systems*, 30:6097–6106.

Yuan, G., Xu, M., Kpotufe, S., and Hsu, D. (2023). Efficient estimation of the central mean subspace via smoothed gradient outer products. *arXiv preprint arXiv:2312.15469*.

References XIV

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 217–242.

Zhang, J. and Zhang, C. (2011). Multitask bregman clustering. *Neurocomputing*, 74(10):1720–1734.

Zhang, Y. and Zhu, Z. (2022). Transfer learning for high-dimensional quantile regression via convolution smoothing. *arXiv preprint arXiv:2212.00428*.

Special thanks to

Haotian Lin (Penn State University, Statistics)
Binghe Zhu (Columbia University, Statistics)
Gan Yuan (Columbia University, Statistics)
Haiyan Zheng (University of Bath, Math)

for helpful discussions and providing references