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

Overview I

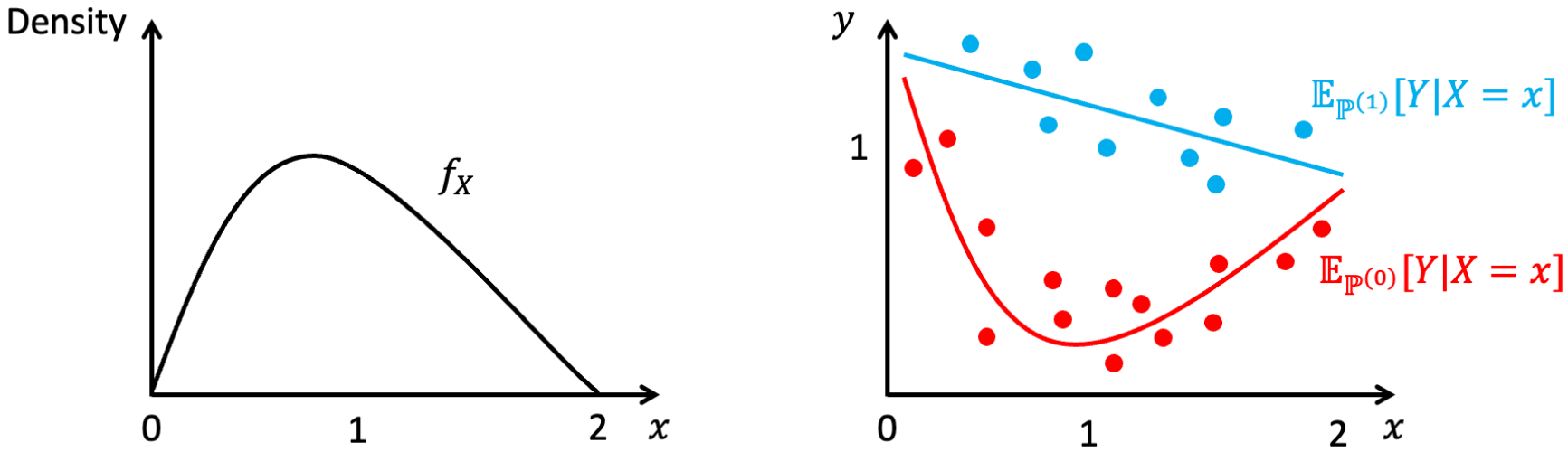
§4.1: Posterior drift and concept drift

Posterior drift and concept drift

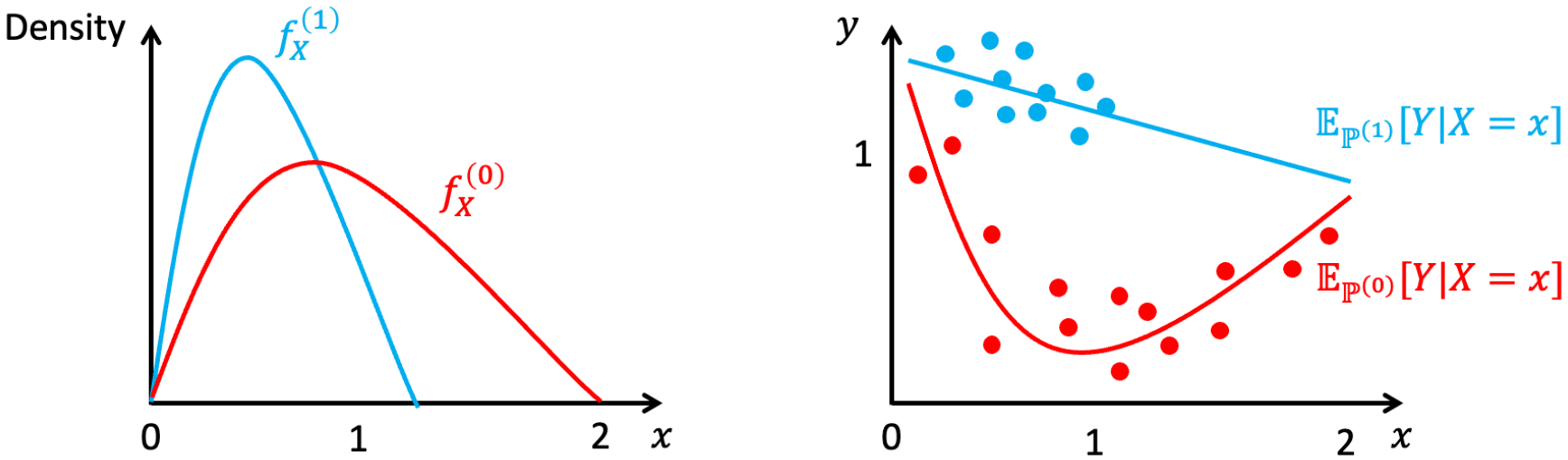
For simplicity, we consider the target and one single source in this section.

- 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

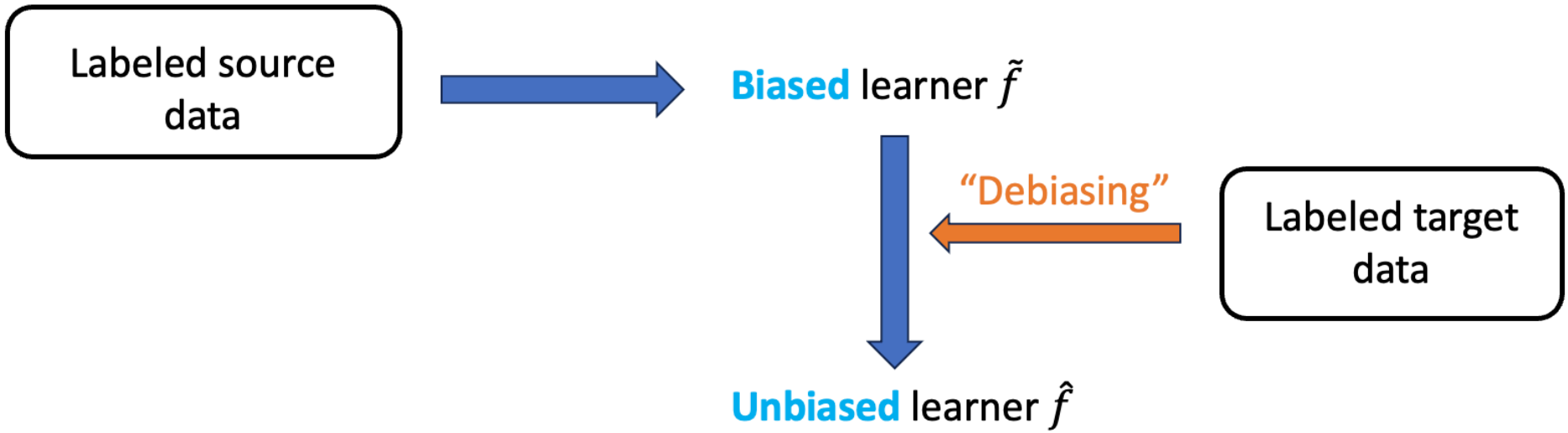
- §4.2.1 Motivation
- §4.2.2 Ridge penalty
- §4.2.3 An adaptive ℓ_2-penalty

§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)[1] 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 similarity 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 $R^{(0)}(f) := \mathbb{E}_{(X,Y)\sim\mu}[\ell(f(X), Y)]$ 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

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

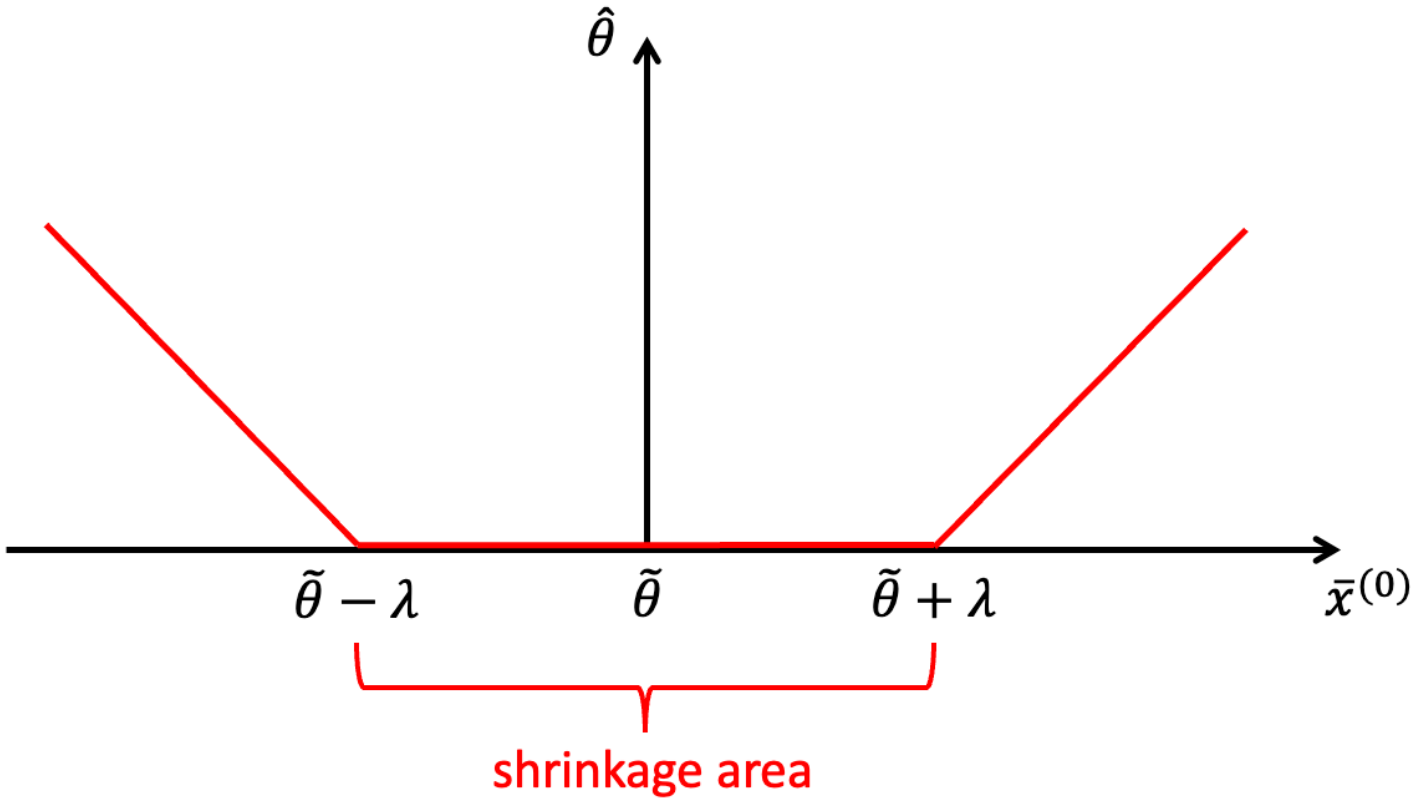
Intuition: Set $\lambda \approx 2|\bar{x}^{(0)} - \theta^*|$ then $|\hat{\theta} - \theta^*| \lesssim \min\{|\tilde{\theta} - \theta^*|, |\bar{x}^{(0)} - \theta^*|\}$.

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.

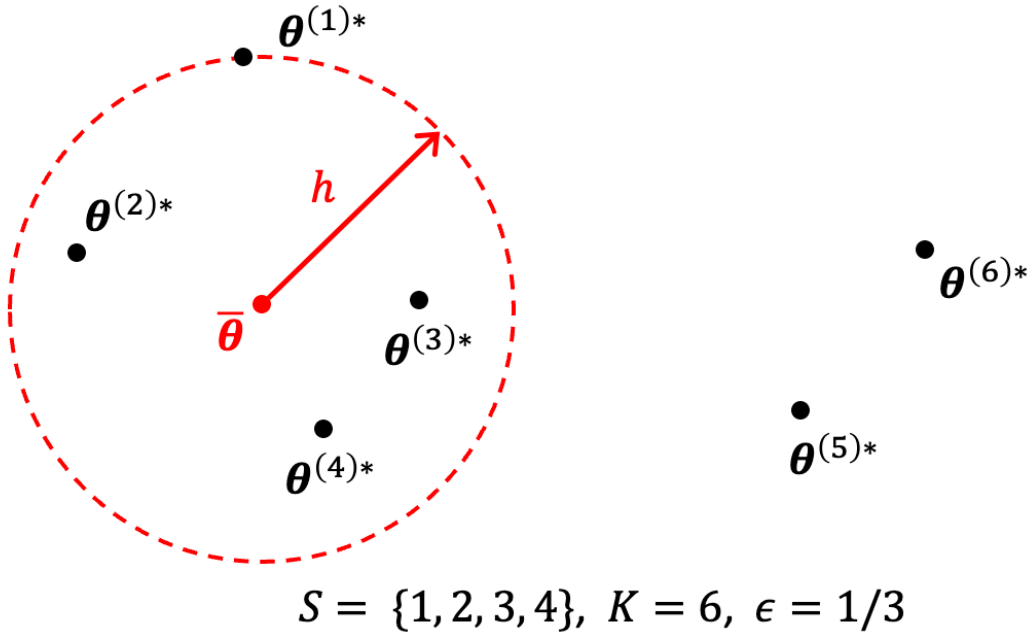
[4] Tian, Y., Weng, H., & Feng, Y. (2024). Towards the Theory of Unsupervised Federated Learning: Non-asymptotic Analysis of Federated EM Algorithms. arXiv preprint arXiv:2310.15330. (accepted by ICML 2024)

[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)

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

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

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

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^*?

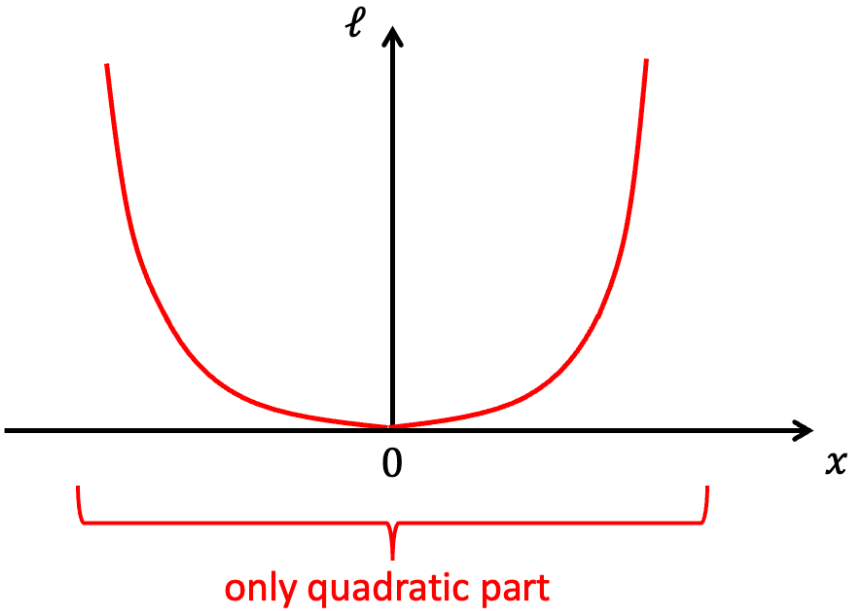
---

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

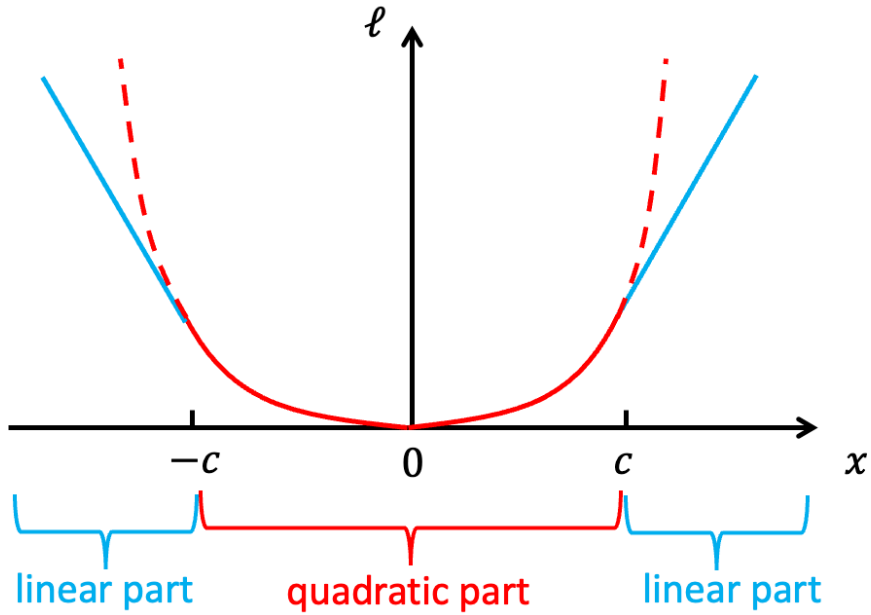
[2] Donoho, D., & Montanari, A. (2016). High dimensional robust m-estimation: Asymptotic variance via approximate message passing. Probability Theory and Related Fields, 166, 935-969.

ℓ_2-penalty in multi-task learning

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



Square loss



Huber loss

◦ **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:

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

- $(\dagger)$ can be seen as a variant of $(\star)$ by re-parameterization $\widehat{\boldsymbol{\theta}}^{(k)} = \widehat{\overline{\boldsymbol{\theta}}} + \widehat{\boldsymbol{\Delta}}^{(k)}$, 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: $\min_{\bar{\boldsymbol{\theta}}} \max_{k \in S} \|\boldsymbol{\theta}^{(k)*} - \bar{\boldsymbol{\theta}}\|_2 \le h$, 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

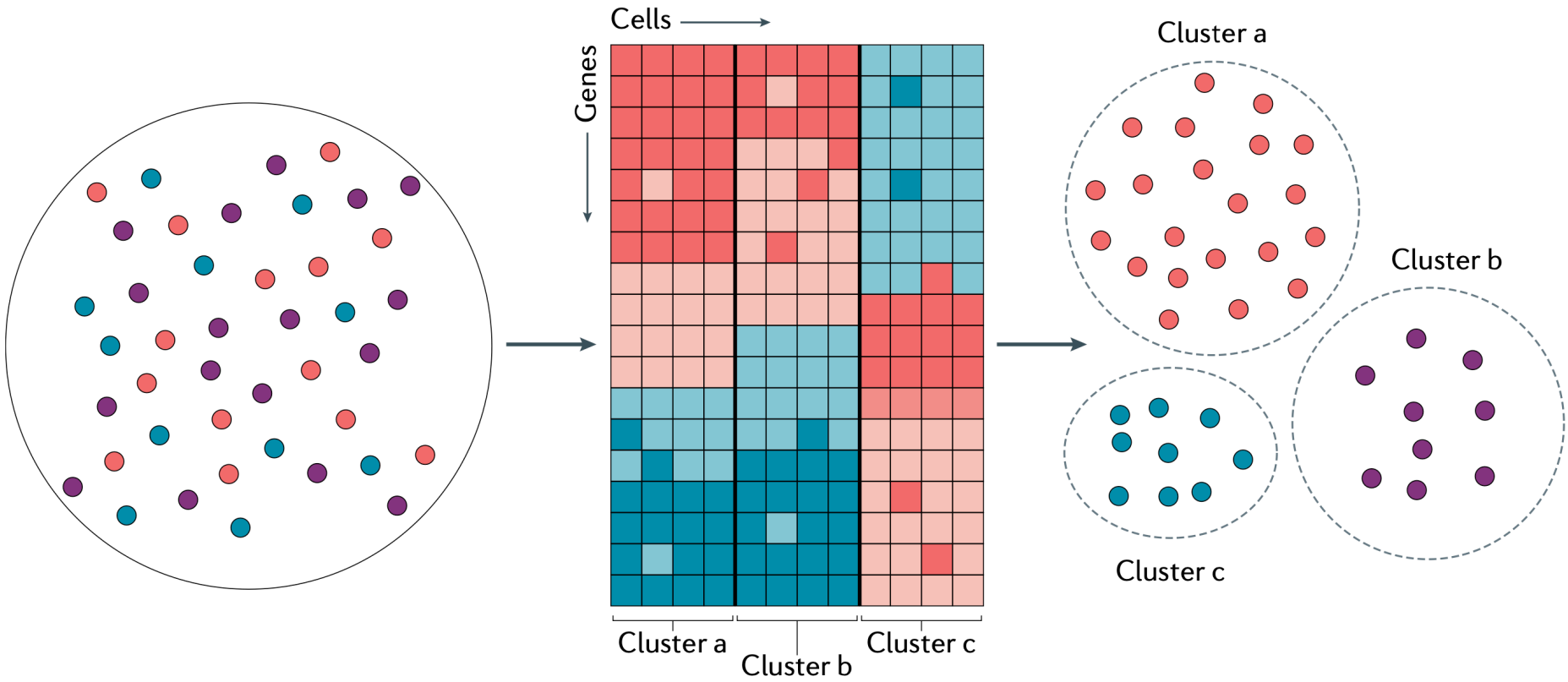
- §4.3.1 ℓ_1-penalty with GLMs
- §4.3.2 Go robust: better ways to aggregate data
- §4.3.3 Block penalty with multi-task learning

§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 (2023); 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, 2023)

◦ 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 (2023) does not.

ℓ_1-penalty with GLMs: theory

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

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 (2023))

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, 2023)

- 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 (2023); 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
$$\widehat{S} = \{1 \le k \le K : \widehat{R}^{(k)} - \widehat{R}^{(0)} \le \text{ threshold}\}$$
  - Run $\widehat{S}$-GLM-Trans
- The above is the likelihood-based version in Tian and Feng (2023). 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

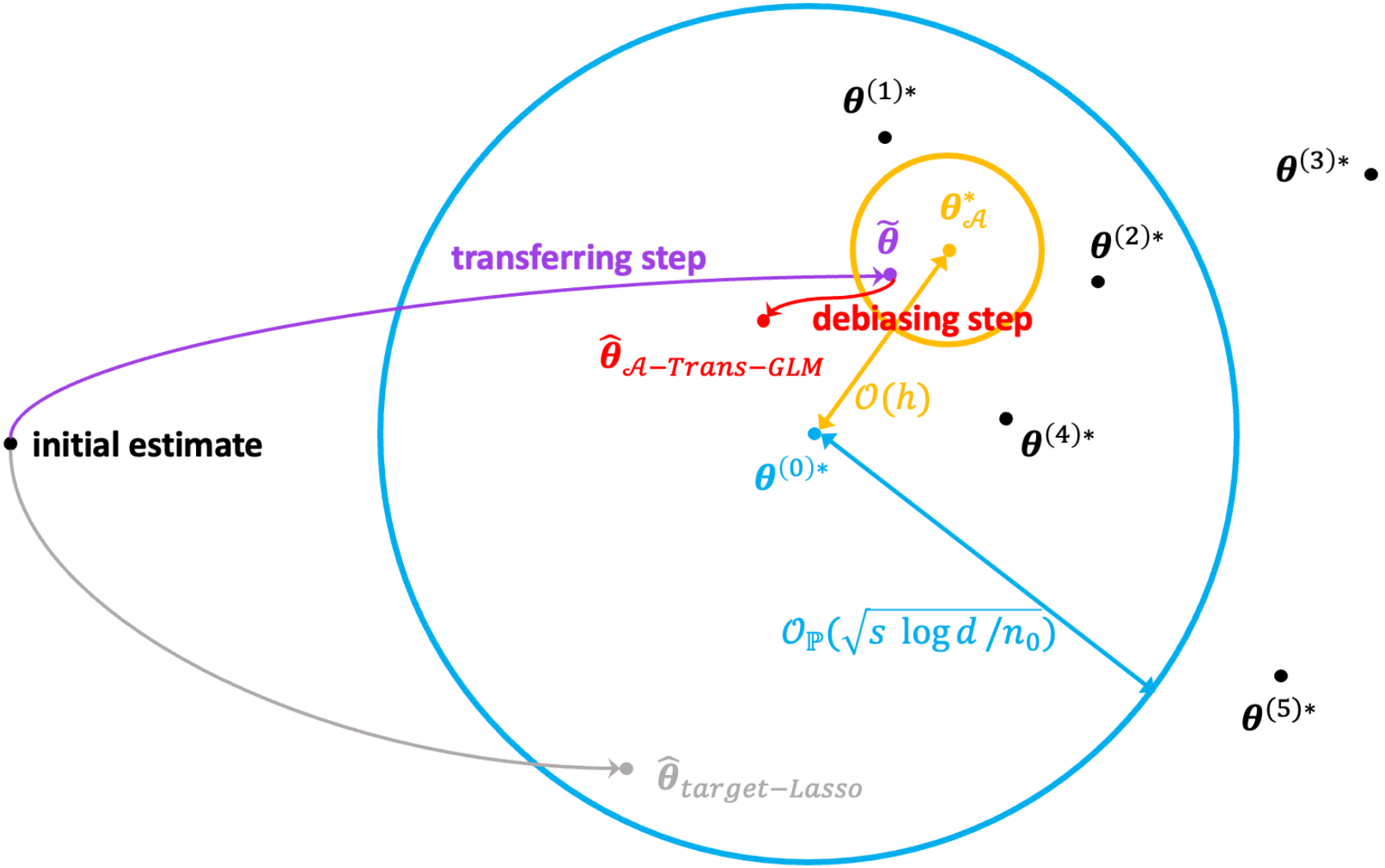
Given certain conditions, both aggregation and selection can guarantee that

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

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

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):

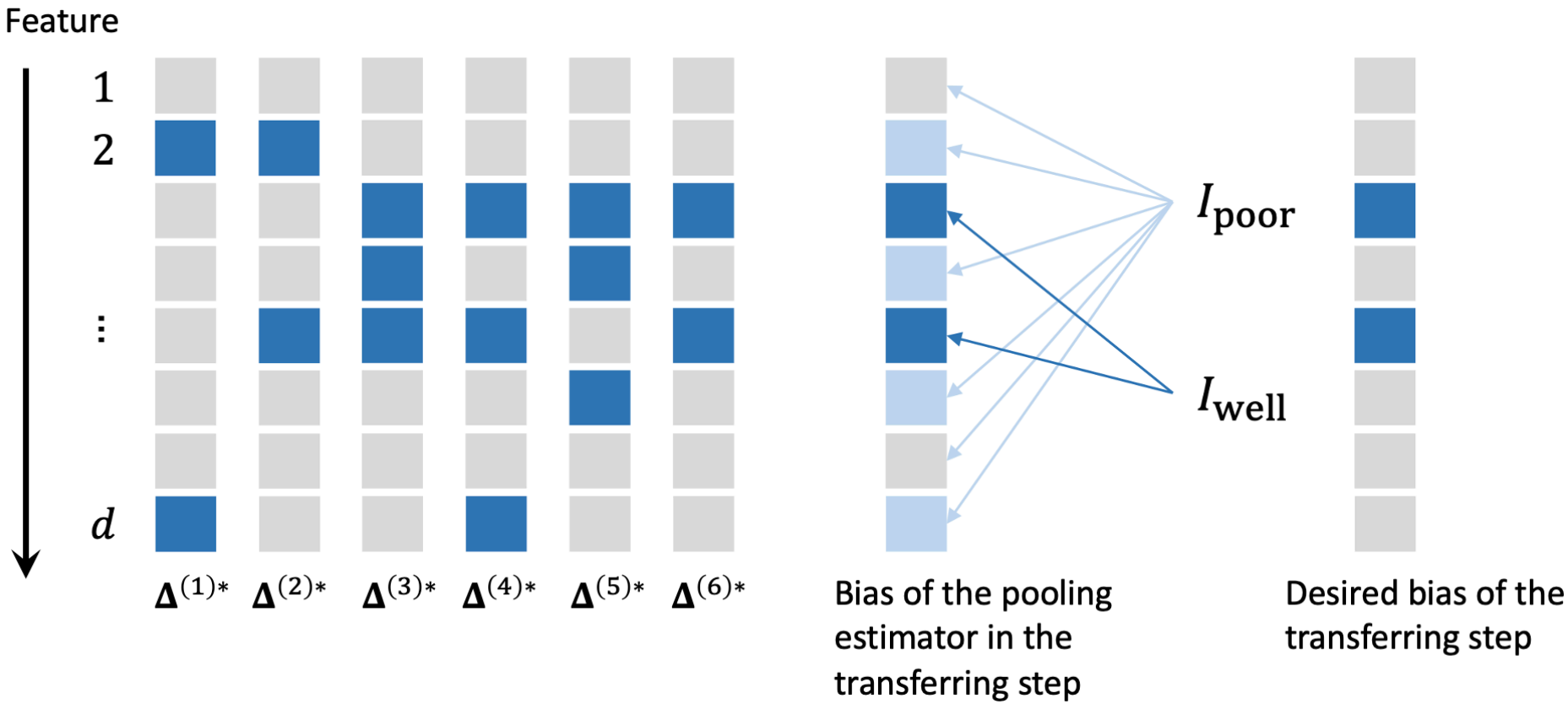
- $\{(\boldsymbol{x}_i^{(k)}, y_i^{(k)})\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} (X^{(k)}, Y^{(k)})$ for $k = 1 : K$, where $X^{(k)} \in \mathbb{R}^d$, $Y^{(k)} \in \mathbb{R}$, and

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

  $\mathbb{E}[X^{(k)}(X^{(k)})^\top] = \boldsymbol{\Sigma}^{(k)}$, $\epsilon^{(k)}$ 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)

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

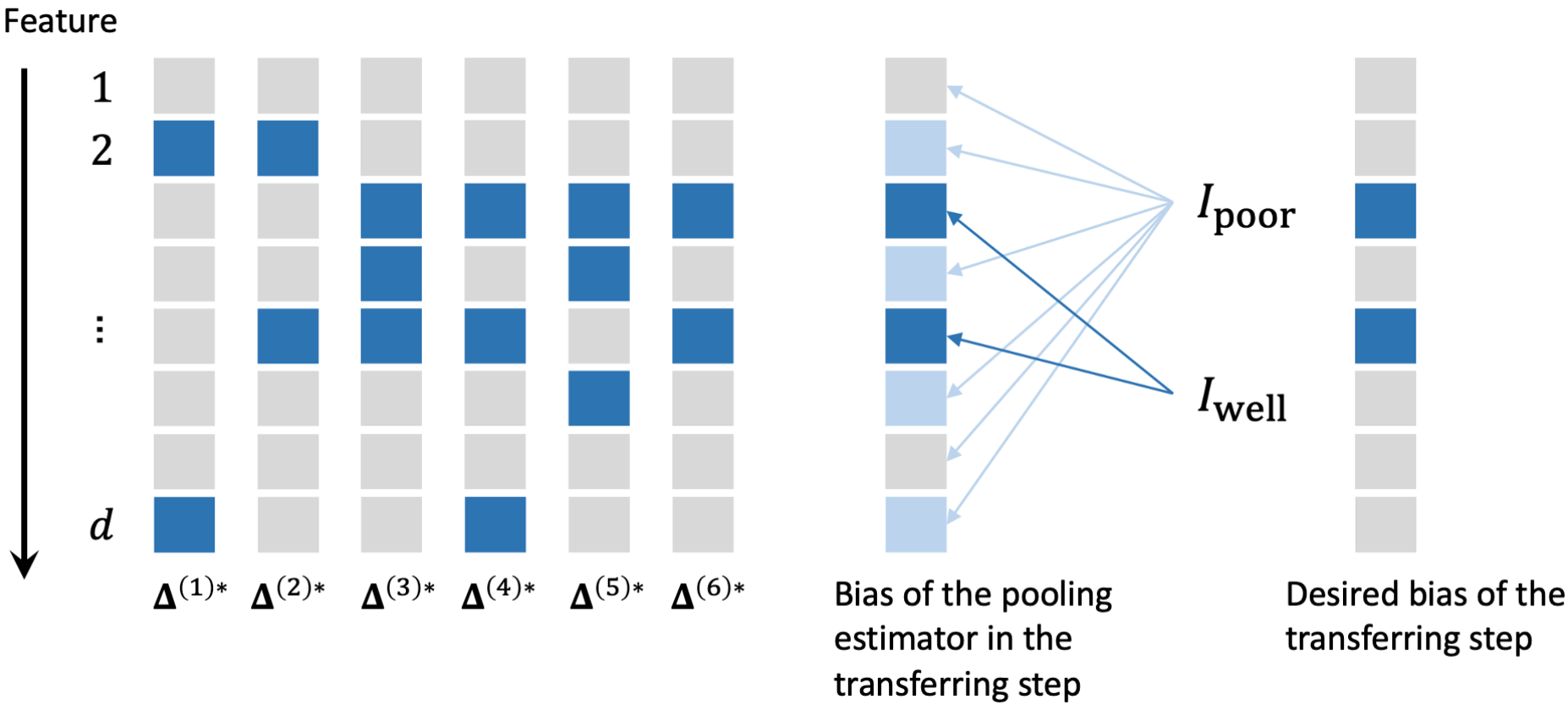
- Step 2: (Transferring)

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

- 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): $p = 1, q = \infty$

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

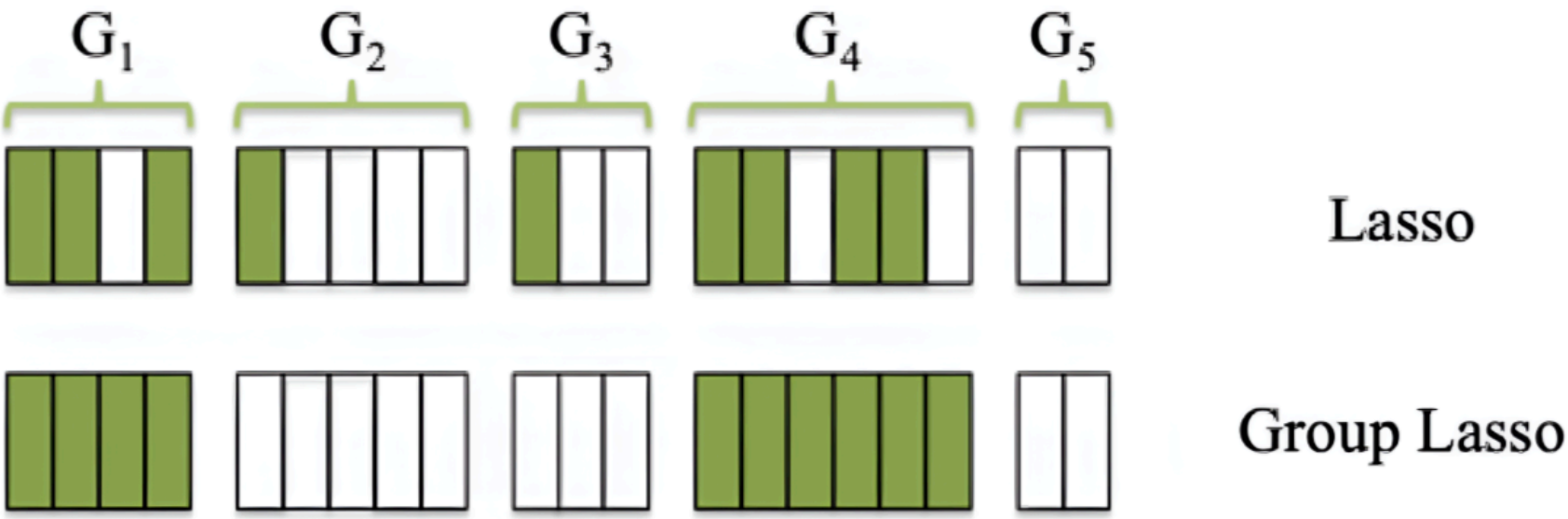
---

[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

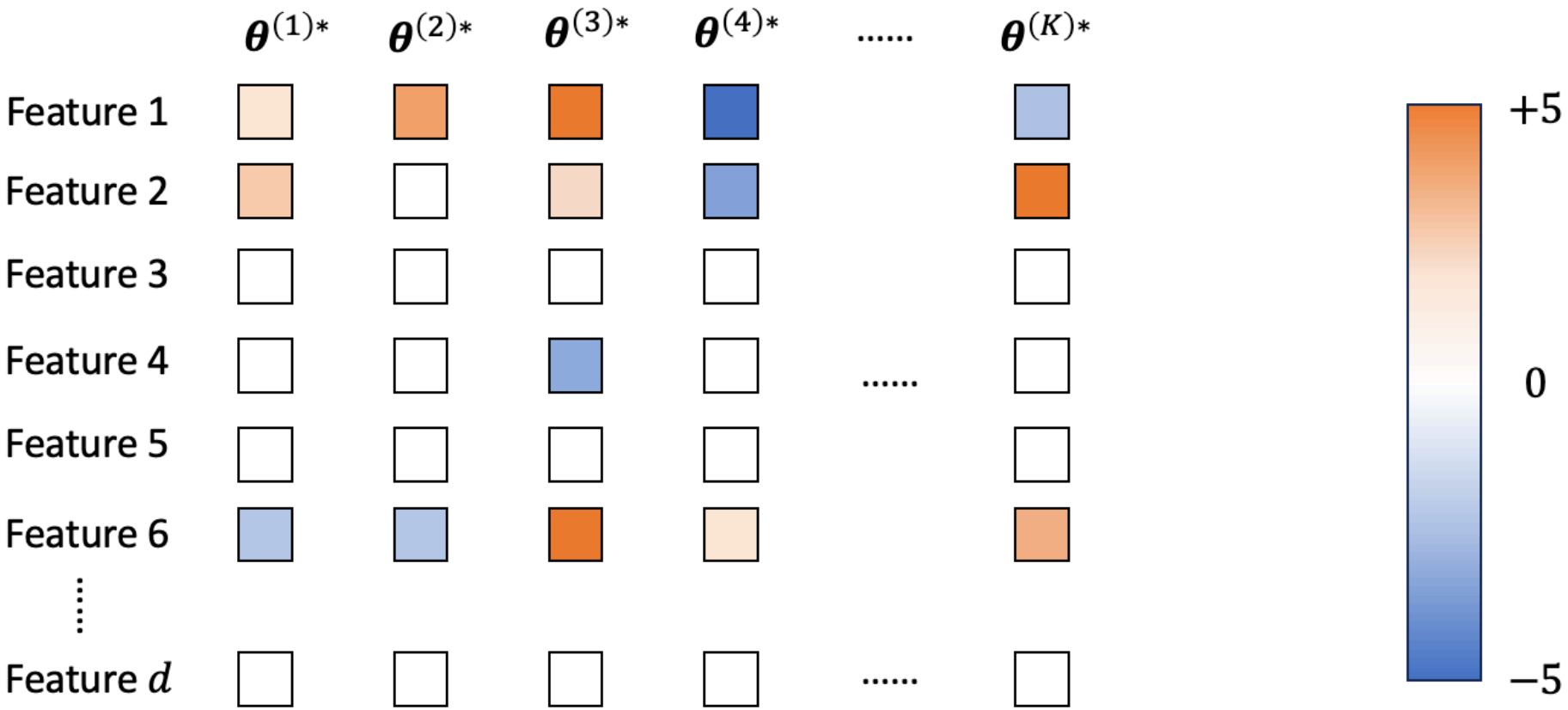
Let's consider the linear regression setting we considered before, but in an MTL framework (Lounici et al., 2011).

- 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* $n \gtrsim sK(K+\log d)$*:*

- $\mathrm{supp}(\widehat{\boldsymbol{\Theta}}) = \mathrm{supp}(\boldsymbol{\Theta}^*)$ *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 $K=2$, $|\mathrm{supp}(\boldsymbol{\theta}^{(1)*})| = |\mathrm{supp}(\boldsymbol{\theta}^{(2)*})| = s$, and the ``overlap proportion" $\alpha = |\mathrm{supp}(\boldsymbol{\theta}^{(1)*}) \cap \mathrm{supp}(\boldsymbol{\theta}^{(2)*})|/s$.

Theorem 4.3.4 (Negahban and Wainwright, 2011)

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

- *(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* $\mathrm{supp}(\widehat{\boldsymbol{\Theta}}) = \mathrm{supp}(\boldsymbol{\Theta}^*)$.

[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* $n \gtrsim s\log(dK) + sK(K+\log d)$*:*

- $\operatorname{supp}(\widehat{\boldsymbol{\Theta}}) = \operatorname{supp}(\boldsymbol{\Theta}^*)$ *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.

References I

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

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

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

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

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 II

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.

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

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.

References III

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.

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.

References IV

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*.

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.

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.

References V

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.

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.

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 VI

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.

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

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.

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

References VII

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)*.

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

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.

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

References VIII

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.

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.