Paper reading

# Synergies Between Disentanglement and Sparsity: a Multi-Task Learning Perspective

Zhiyuan Tang

20230206

Synergies Between Disentanglement and Sparsity: a Multi-Task Learning Perspective

# Disentangled representation

- Definition
  - Informally, a representation is considered disentangled when its components are in one-to-one correspondence with natural and interpretable factors of variations, such as object positions, colors or shape. (This paper)
  - Atomic, orthogonal, complete
- Related
  - Factorization, Dictionary learning, Sparse coding

# Disentangled representation

- Definition
  - Informally, a representation is considered disentangled when its components are in one-to-one correspondence with natural and interpretable factors of variations, such as object positions, colors or shape. (This paper)
  - Atomic, orthogonal, complete
- Related
  - Factorization, Dictionary learning, Sparse coding

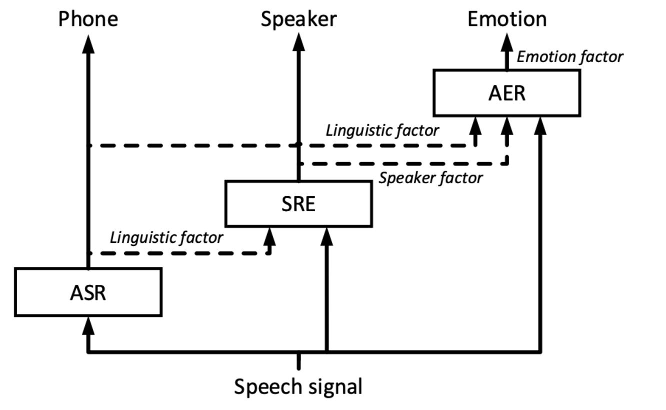


**Fig. 1**. The cascaded deep factorization approach applied to factorize emotional speech into three factors: linguistic, speaker and emotion.

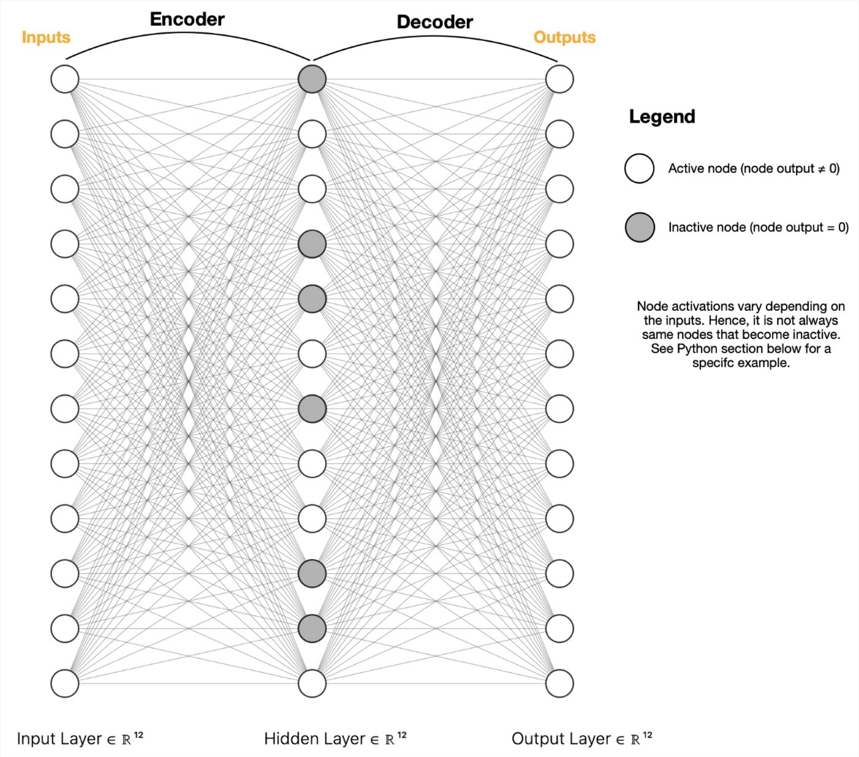
Deep factorization for speech signal

# Sparsity

- At the heart of the paper’s contributions is the assumption: only a small subset of all factors of variations are useful for each downstream task, and this subset might change from one task to another.
- Related
  - Sparse autoencoder

# Sparsity

- At the heart of the paper’s contributions is the assumption: only a small subset of all factors of variations are useful for each downstream task, and this subset might change from one task to another.
- Related
  - Sparse autoencoder

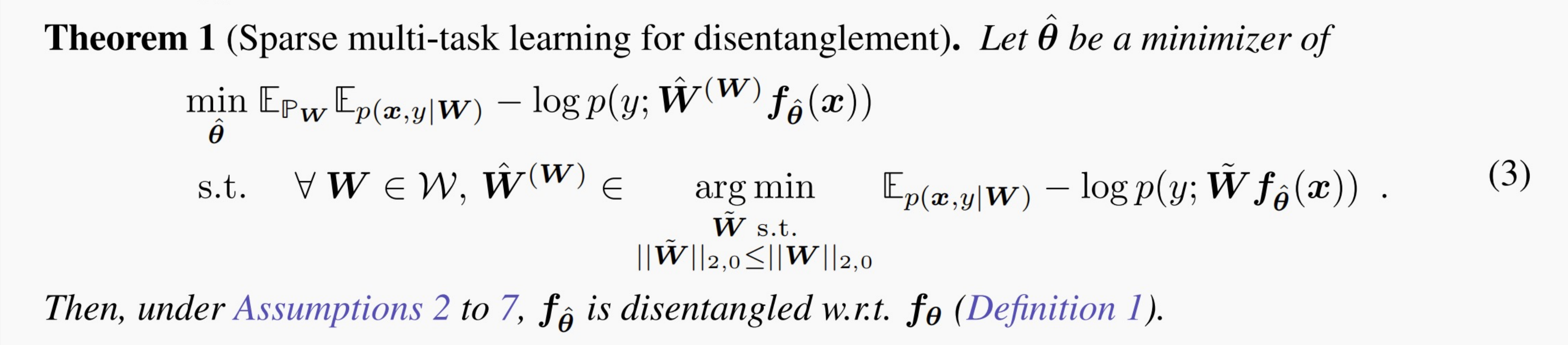

Encoder
Decoder
Inputs
Outputs
Legend
Active node (node output ≠ 0)
Inactive node (node output = 0)
Input Layer $\in \mathbb{R}^{12}$
Hidden Layer $\in \mathbb{R}^{12}$
Output Layer $\in \mathbb{R}^{12}$

Node activations vary depending on the inputs. Hence, it is not always same nodes that become inactive. See Python section below for a specifc example.

https://towardsdatascience.com/sparse-autoencoder-neural-networks-how-to-utilise-sparsity-for-robust-information-encoding-6aa9ff542bc9

# Take-away

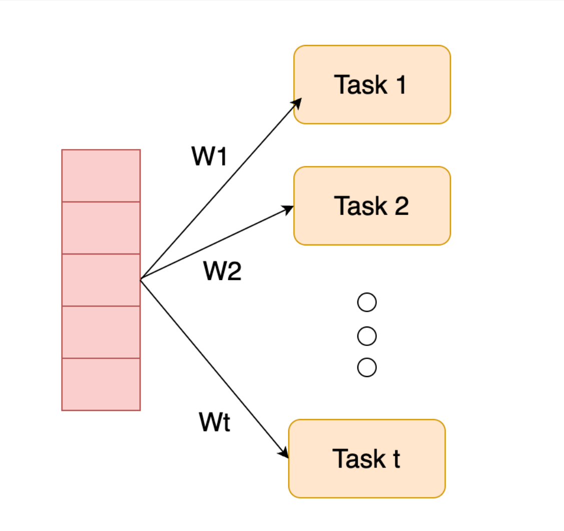
- Learning a shared representation across tasks while penalizing the task-specific base-predictor to be sparse can induce disentanglement.



**Theorem 1** (Sparse multi-task learning for disentanglement)**.** *Let $\hat{\boldsymbol{\theta}}$ be a minimizer of*

$$\begin{aligned} &\min_{\hat{\boldsymbol{\theta}}} \mathbb{E}_{\mathbb{P}_{\boldsymbol{W}}} \mathbb{E}_{p(\boldsymbol{x},y|\boldsymbol{W})} - \log p(y; \hat{\boldsymbol{W}}^{(\boldsymbol{W})} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})) \\ &\text{s.t.} \quad \forall\, \boldsymbol{W} \in \mathcal{W},\ \hat{\boldsymbol{W}}^{(\boldsymbol{W})} \in \underset{\substack{\tilde{\boldsymbol{W}} \text{ s.t.} \\ \|\tilde{\boldsymbol{W}}\|_{2,0} \leq \|\boldsymbol{W}\|_{2,0}}}{\arg\min} \mathbb{E}_{p(\boldsymbol{x},y|\boldsymbol{W})} - \log p(y; \tilde{\boldsymbol{W}} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}))\ . \end{aligned} \tag{3}$$

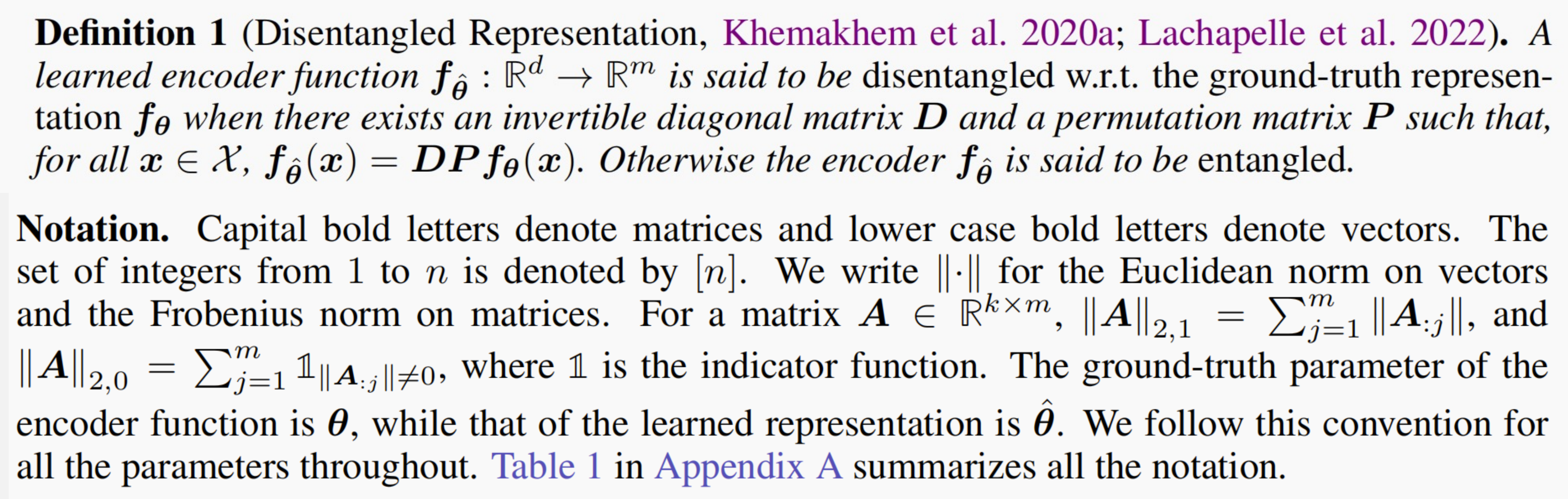
*Then, under Assumptions 2 to 7, $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}$ is disentangled w.r.t. $\boldsymbol{f}_{\boldsymbol{\theta}}$ (Definition 1).*



# Paper contents

- How disentangled representations coupled with sparsity-regularized base-predictors can obtain better generalization
- How one can leverage multiple sparse tasks to learn a shared disentangled representation by regularizing the task-specific predictors to be maximally sparse
- Learn disentangled representations based on a sparsity-promoting bi-level optimization problem
- Connection between this bi-level optimization problem and some formulations from the meta-learning literature

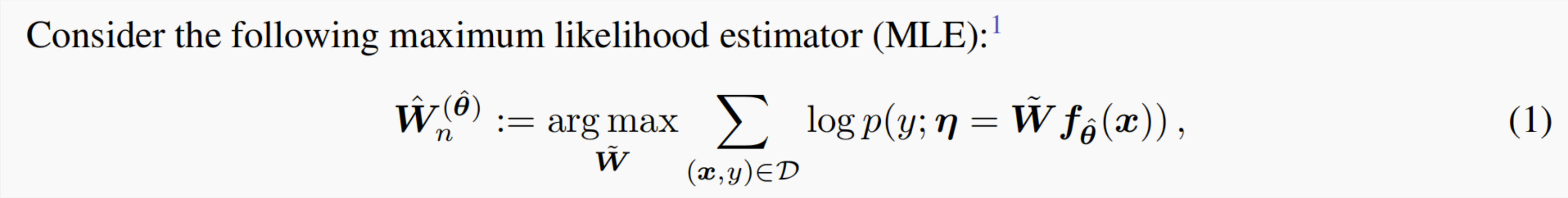
# Notation



**Definition 1** (Disentangled Representation, Khemakhem et al. 2020a; Lachapelle et al. 2022)**.** *A learned encoder function* $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}} : \mathbb{R}^d \to \mathbb{R}^m$ *is said to be* disentangled w.r.t. the ground-truth representation $\boldsymbol{f}_{\boldsymbol{\theta}}$ *when there exists an invertible diagonal matrix* $\boldsymbol{D}$ *and a permutation matrix* $\boldsymbol{P}$ *such that, for all* $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}) = \boldsymbol{D}\boldsymbol{P}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$. *Otherwise the encoder* $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}$ *is said to be* entangled.

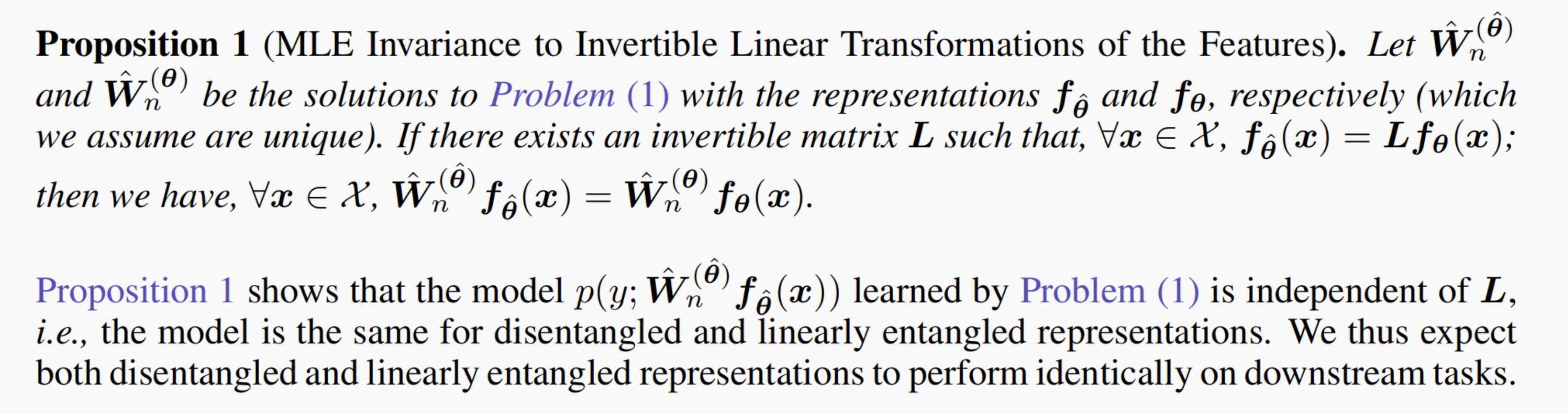
**Notation.** Capital bold letters denote matrices and lower case bold letters denote vectors. The set of integers from 1 to $n$ is denoted by $[n]$. We write $\|\cdot\|$ for the Euclidean norm on vectors and the Frobenius norm on matrices. For a matrix $\boldsymbol{A} \in \mathbb{R}^{k\times m}$, $\|\boldsymbol{A}\|_{2,1} = \sum_{j=1}^{m} \|\boldsymbol{A}_{:j}\|$, and $\|\boldsymbol{A}\|_{2,0} = \sum_{j=1}^{m} \mathbb{1}_{\|\boldsymbol{A}_{:j}\|\neq 0}$, where $\mathbb{1}$ is the indicator function. The ground-truth parameter of the encoder function is $\boldsymbol{\theta}$, while that of the learned representation is $\hat{\boldsymbol{\theta}}$. We follow this convention for all the parameters throughout. Table 1 in Appendix A summarizes all the notation.

# Disentanglement and sparse base-predictors for improved generalization



Consider the following maximum likelihood estimator (MLE):[1]

$$\hat{\boldsymbol{W}}_n^{(\hat{\boldsymbol{\theta}})} := \arg\max_{\tilde{\boldsymbol{W}}} \sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y; \boldsymbol{\eta} = \tilde{\boldsymbol{W}} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}))\,, \tag{1}$$

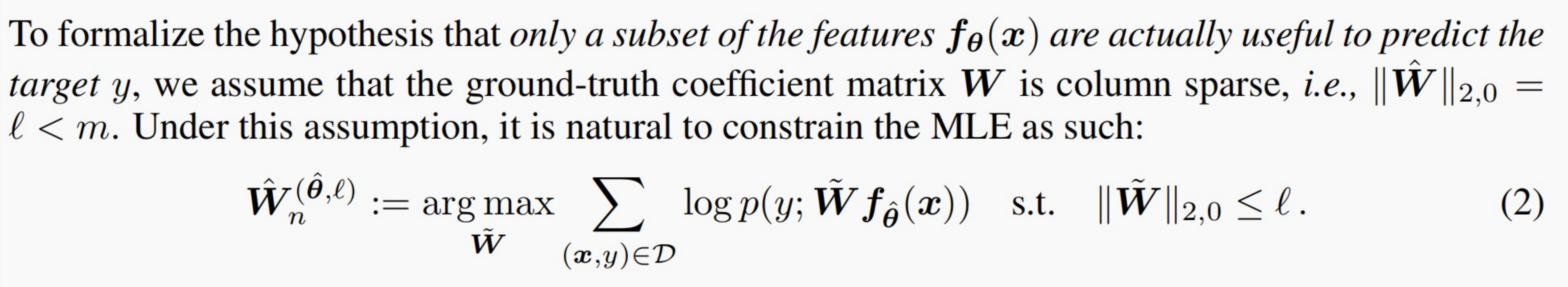


**Proposition 1** (MLE Invariance to Invertible Linear Transformations of the Features)**.** *Let* $\hat{\boldsymbol{W}}_n^{(\hat{\boldsymbol{\theta}})}$ *and* $\hat{\boldsymbol{W}}_n^{(\boldsymbol{\theta})}$ *be the solutions to Problem (1) with the representations* $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}$ *and* $\boldsymbol{f}_{\boldsymbol{\theta}}$*, respectively (which we assume are unique). If there exists an invertible matrix* $\boldsymbol{L}$ *such that,* $\forall \boldsymbol{x} \in \mathcal{X}$*,* $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}) = \boldsymbol{L}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$*; then we have,* $\forall \boldsymbol{x} \in \mathcal{X}$*,* $\hat{\boldsymbol{W}}_n^{(\hat{\boldsymbol{\theta}})} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}) = \hat{\boldsymbol{W}}_n^{(\boldsymbol{\theta})} \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$*.*

Proposition 1 shows that the model $p(y; \hat{\boldsymbol{W}}_n^{(\hat{\boldsymbol{\theta}})} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}))$ learned by Problem (1) is independent of $\boldsymbol{L}$, *i.e.*, the model is the same for disentangled and linearly entangled representations. We thus expect both disentangled and linearly entangled representations to perform identically on downstream tasks.

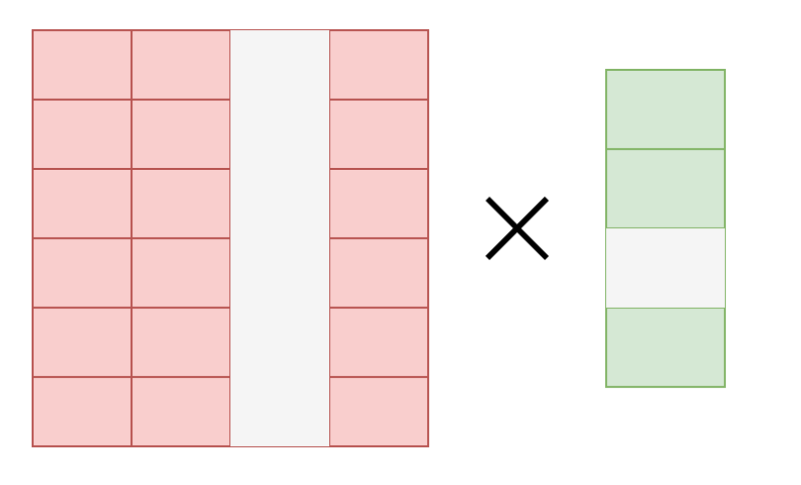
# Disentanglement and sparse base-predictors for improved generalization

- Sparsity constraint



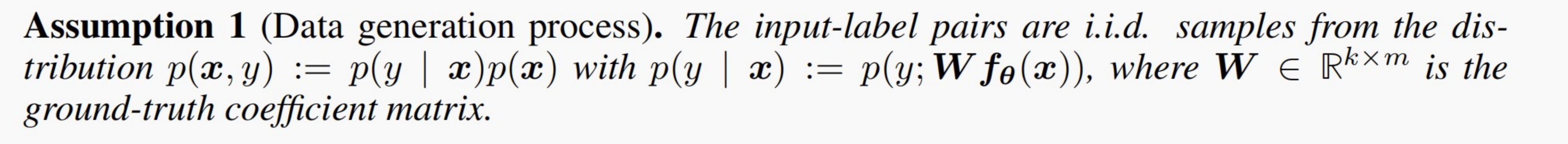
To formalize the hypothesis that *only a subset of the features* $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$ *are actually useful to predict the target* $y$, we assume that the ground-truth coefficient matrix $\boldsymbol{W}$ is column sparse, *i.e.*, $\|\hat{\boldsymbol{W}}\|_{2,0} = \ell < m$. Under this assumption, it is natural to constrain the MLE as such:

$$\hat{\boldsymbol{W}}_n^{(\hat{\boldsymbol{\theta}},\ell)} := \arg\max_{\tilde{\boldsymbol{W}}} \sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y; \tilde{\boldsymbol{W}} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})) \quad \text{s.t.} \quad \|\tilde{\boldsymbol{W}}\|_{2,0} \leq \ell\,. \tag{2}$$

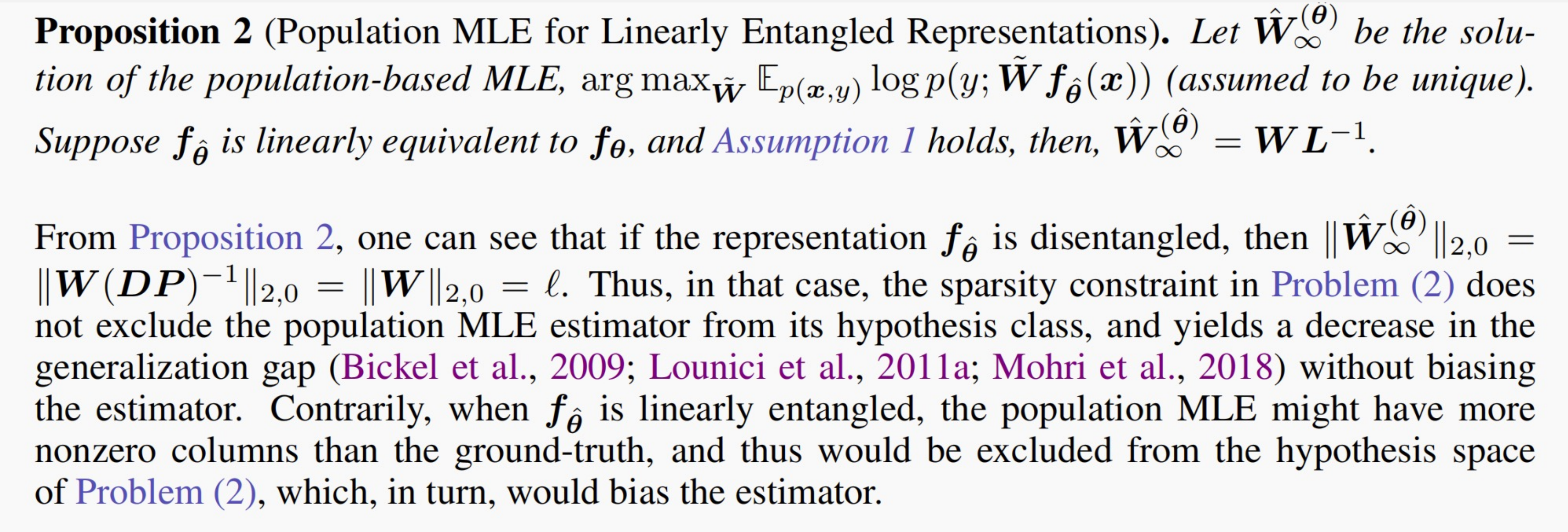


# Disentanglement and sparse base-predictors for improved generalization

- Theoretically



**Assumption 1** (Data generation process)**.** *The input-label pairs are i.i.d. samples from the distribution* $p(\boldsymbol{x}, y) := p(y \mid \boldsymbol{x})p(\boldsymbol{x})$ *with* $p(y \mid \boldsymbol{x}) := p(y; \boldsymbol{W}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}))$*, where* $\boldsymbol{W} \in \mathbb{R}^{k \times m}$ *is the ground-truth coefficient matrix.*

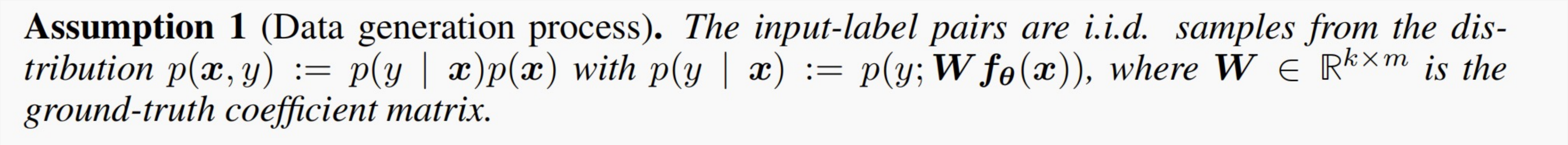


**Proposition 2** (Population MLE for Linearly Entangled Representations)**.** *Let* $\hat{\boldsymbol{W}}^{(\boldsymbol{\theta})}_{\infty}$ *be the solution of the population-based MLE,* $\arg\max_{\tilde{\boldsymbol{W}}} \mathbb{E}_{p(\boldsymbol{x},y)} \log p(y; \tilde{\boldsymbol{W}}\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}))$ *(assumed to be unique). Suppose* $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}$ *is linearly equivalent to* $\boldsymbol{f}_{\boldsymbol{\theta}}$*, and Assumption 1 holds, then,* $\hat{\boldsymbol{W}}^{(\hat{\boldsymbol{\theta}})}_{\infty} = \boldsymbol{W}\boldsymbol{L}^{-1}$.

From Proposition 2, one can see that if the representation $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}$ is disentangled, then $\|\hat{\boldsymbol{W}}^{(\hat{\boldsymbol{\theta}})}_{\infty}\|_{2,0} = \|\boldsymbol{W}(\boldsymbol{D}\boldsymbol{P})^{-1}\|_{2,0} = \|\boldsymbol{W}\|_{2,0} = \ell$. Thus, in that case, the sparsity constraint in Problem (2) does not exclude the population MLE estimator from its hypothesis class, and yields a decrease in the generalization gap (Bickel et al., 2009; Lounici et al., 2011a; Mohri et al., 2018) without biasing the estimator. Contrarily, when $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}$ is linearly entangled, the population MLE might have more nonzero columns than the ground-truth, and thus would be excluded from the hypothesis space of Problem (2), which, in turn, would bias the estimator.

# Disentanglement and sparse base-predictors for improved generalization

- Empirically



**Assumption 1** (Data generation process)**.** *The input-label pairs are i.i.d. samples from the distribution* $p(\boldsymbol{x}, y) := p(y \mid \boldsymbol{x})p(\boldsymbol{x})$ *with* $p(y \mid \boldsymbol{x}) := p(y; \boldsymbol{W} \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}))$*, where* $\boldsymbol{W} \in \mathbb{R}^{k \times m}$ *is the ground-truth coefficient matrix.*

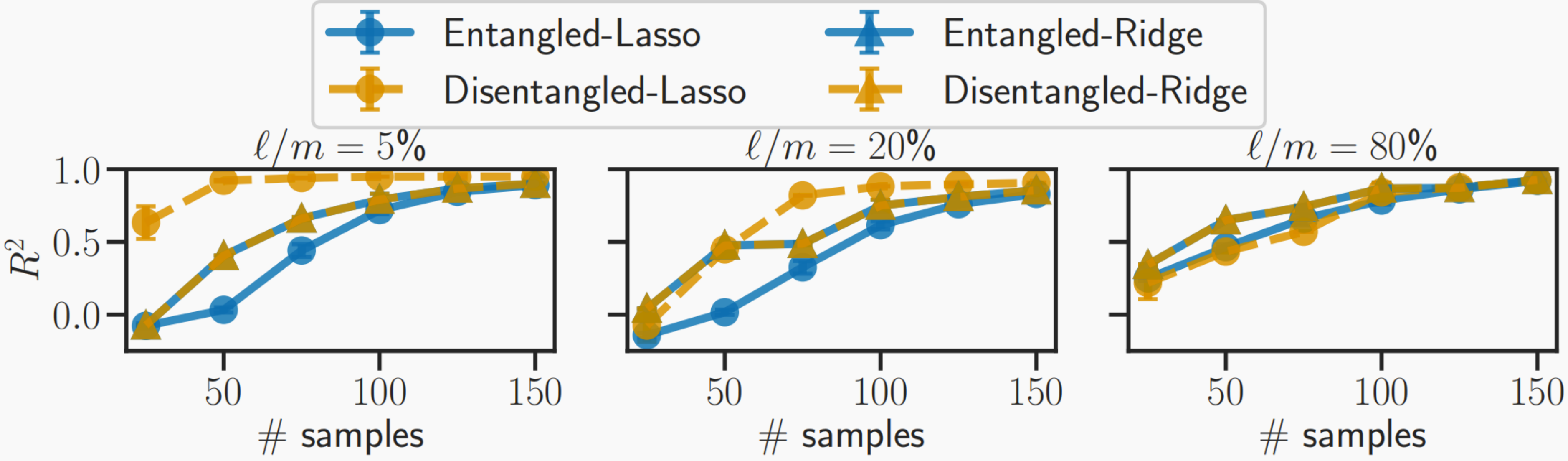
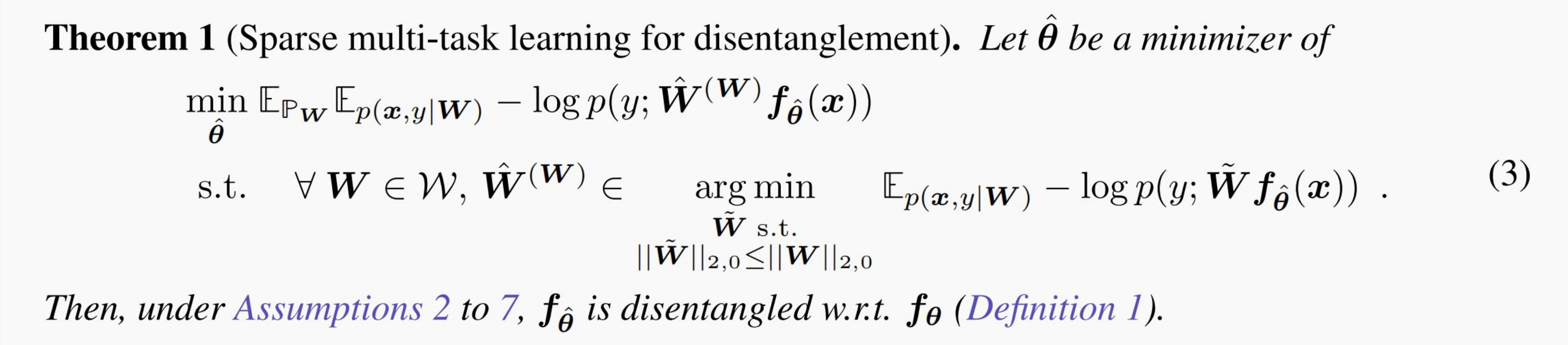


Figure 1: Test performance for the entangled and disentangled representation using Lasso and Ridge regression. All the results are averaged over 10 seeds, with standard error shown in error bars.

# Disentanglement via sparse multitask learning

- Learning a shared representation across tasks while penalizing the task-specific base-predictor to be sparse can induce disentanglement.

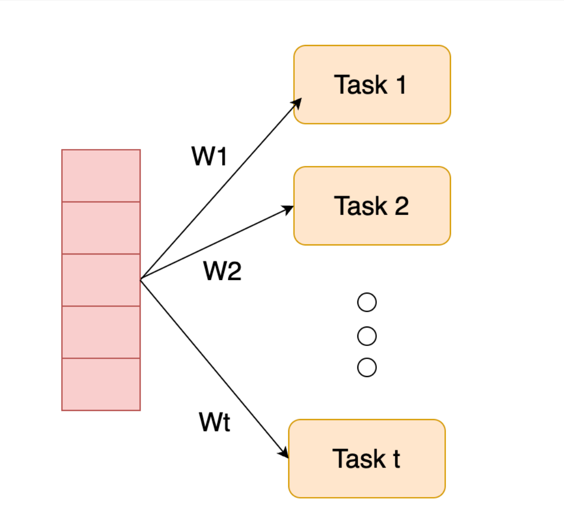


**Theorem 1** (Sparse multi-task learning for disentanglement)**.** *Let $\hat{\boldsymbol{\theta}}$ be a minimizer of*

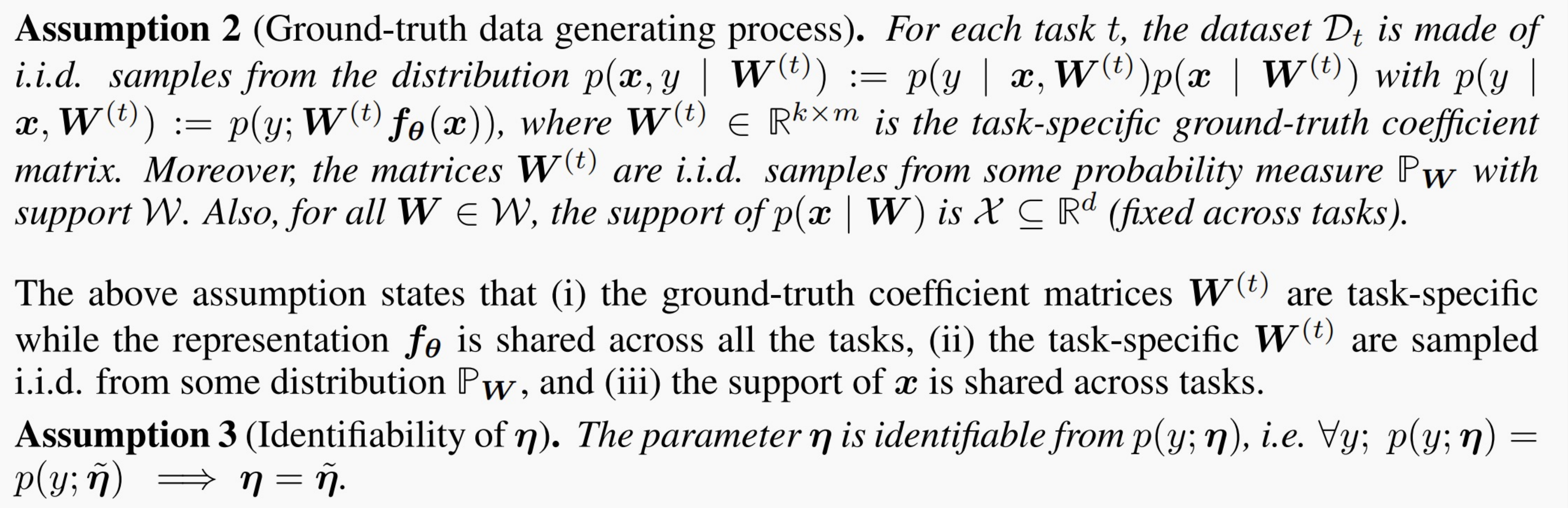
$$\min_{\hat{\boldsymbol{\theta}}} \mathbb{E}_{\mathbb{P}_{\boldsymbol{W}}} \mathbb{E}_{p(\boldsymbol{x},y|\boldsymbol{W})} -\log p(y; \hat{\boldsymbol{W}}^{(\boldsymbol{W})} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}))$$

$$\text{s.t.} \quad \forall\, \boldsymbol{W} \in \mathcal{W},\ \hat{\boldsymbol{W}}^{(\boldsymbol{W})} \in \underset{\substack{\tilde{\boldsymbol{W}} \text{ s.t.} \\ ||\tilde{\boldsymbol{W}}||_{2,0} \leq ||\boldsymbol{W}||_{2,0}}}{\arg\min} \mathbb{E}_{p(\boldsymbol{x},y|\boldsymbol{W})} -\log p(y; \tilde{\boldsymbol{W}} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}))\ . \qquad (3)$$

*Then, under Assumptions 2 to 7, $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}$ is disentangled w.r.t. $\boldsymbol{f}_{\boldsymbol{\theta}}$ (Definition 1).*



# Disentanglement via sparse multitask learning



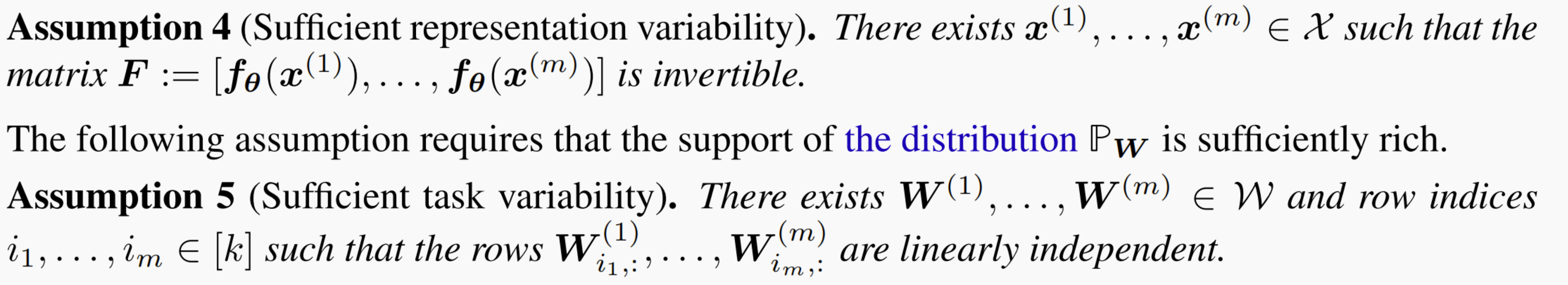
**Assumption 2** (Ground-truth data generating process)**.** *For each task $t$, the dataset $\mathcal{D}_t$ is made of i.i.d. samples from the distribution $p(\boldsymbol{x}, y \mid \boldsymbol{W}^{(t)}) := p(y \mid \boldsymbol{x}, \boldsymbol{W}^{(t)}) p(\boldsymbol{x} \mid \boldsymbol{W}^{(t)})$ with $p(y \mid \boldsymbol{x}, \boldsymbol{W}^{(t)}) := p(y; \boldsymbol{W}^{(t)} \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}))$, where $\boldsymbol{W}^{(t)} \in \mathbb{R}^{k \times m}$ is the task-specific ground-truth coefficient matrix. Moreover, the matrices $\boldsymbol{W}^{(t)}$ are i.i.d. samples from some probability measure $\mathbb{P}_{\boldsymbol{W}}$ with support $\mathcal{W}$. Also, for all $\boldsymbol{W} \in \mathcal{W}$, the support of $p(\boldsymbol{x} \mid \boldsymbol{W})$ is $\mathcal{X} \subseteq \mathbb{R}^d$ (fixed across tasks).*

The above assumption states that (i) the ground-truth coefficient matrices $\boldsymbol{W}^{(t)}$ are task-specific while the representation $\boldsymbol{f}_{\boldsymbol{\theta}}$ is shared across all the tasks, (ii) the task-specific $\boldsymbol{W}^{(t)}$ are sampled i.i.d. from some distribution $\mathbb{P}_{\boldsymbol{W}}$, and (iii) the support of $\boldsymbol{x}$ is shared across tasks.

**Assumption 3** (Identifiability of $\boldsymbol{\eta}$)**.** *The parameter $\boldsymbol{\eta}$ is identifiable from $p(y; \boldsymbol{\eta})$, i.e. $\forall y;\ p(y; \boldsymbol{\eta}) = p(y; \tilde{\boldsymbol{\eta}}) \implies \boldsymbol{\eta} = \tilde{\boldsymbol{\eta}}$.*

# Disentanglement via sparse multitask learning

- The following assumption requires the ground-truth representation fθ(x) to vary enough such that its image cannot be trapped inside a proper subspace.

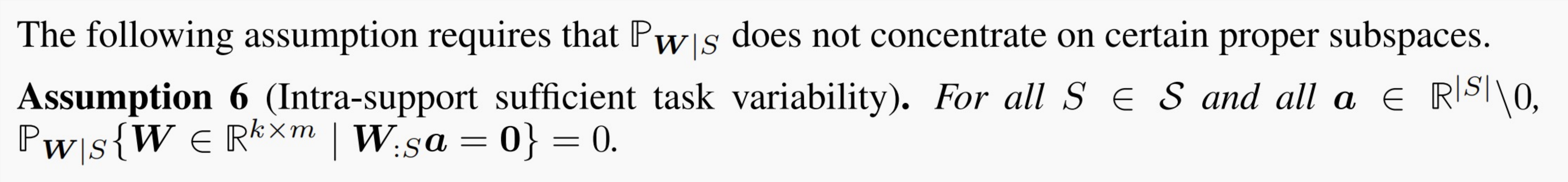


**Assumption 4** (Sufficient representation variability)**.** *There exists* $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(m)} \in \mathcal{X}$ *such that the matrix* $\boldsymbol{F} := [\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}^{(1)}), \dots, \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}^{(m)})]$ *is invertible.*

The following assumption requires that the support of the distribution $\mathbb{P}_{\boldsymbol{W}}$ is sufficiently rich.

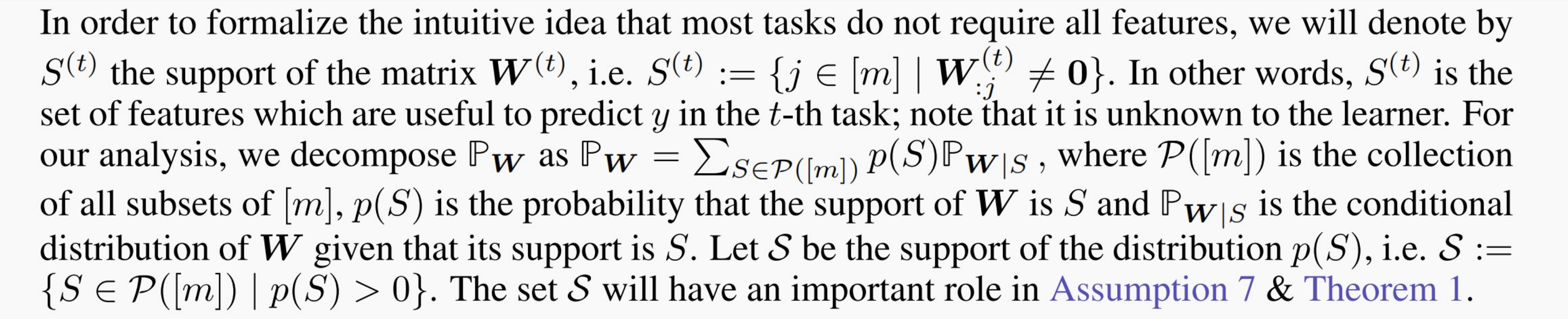
**Assumption 5** (Sufficient task variability)**.** *There exists* $\boldsymbol{W}^{(1)}, \dots, \boldsymbol{W}^{(m)} \in \mathcal{W}$ *and row indices* $i_1, \dots, i_m \in [k]$ *such that the rows* $\boldsymbol{W}^{(1)}_{i_1,:}, \dots, \boldsymbol{W}^{(m)}_{i_m,:}$ *are linearly independent.*

# Disentanglement via sparse multitask learning



The following assumption requires that $\mathbb{P}_{\boldsymbol{W}|S}$ does not concentrate on certain proper subspaces.

**Assumption 6** **(Intra-support sufficient task variability).** *For all $S \in \mathcal{S}$ and all $\boldsymbol{a} \in \mathbb{R}^{|S|} \backslash 0$, $\mathbb{P}_{\boldsymbol{W}|S}\{\boldsymbol{W} \in \mathbb{R}^{k \times m} \mid \boldsymbol{W}_{:S}\boldsymbol{a} = \boldsymbol{0}\} = 0$.*



In order to formalize the intuitive idea that most tasks do not require all features, we will denote by $S^{(t)}$ the support of the matrix $\boldsymbol{W}^{(t)}$, i.e. $S^{(t)} := \{j \in [m] \mid \boldsymbol{W}^{(t)}_{:j} \neq \boldsymbol{0}\}$. In other words, $S^{(t)}$ is the set of features which are useful to predict $y$ in the $t$-th task; note that it is unknown to the learner. For our analysis, we decompose $\mathbb{P}_{\boldsymbol{W}}$ as $\mathbb{P}_{\boldsymbol{W}} = \sum_{S \in \mathcal{P}([m])} p(S)\mathbb{P}_{\boldsymbol{W}|S}$ , where $\mathcal{P}([m])$ is the collection of all subsets of $[m]$, $p(S)$ is the probability that the support of $\boldsymbol{W}$ is $S$ and $\mathbb{P}_{\boldsymbol{W}|S}$ is the conditional distribution of $\boldsymbol{W}$ given that its support is $S$. Let $\mathcal{S}$ be the support of the distribution $p(S)$, i.e. $\mathcal{S} := \{S \in \mathcal{P}([m]) \mid p(S) > 0\}$. The set $\mathcal{S}$ will have an important role in Assumption 7 & Theorem 1.

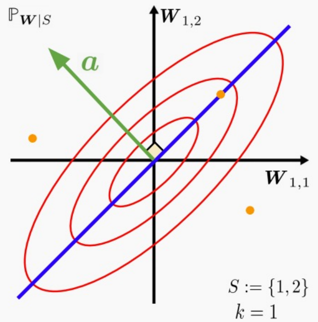
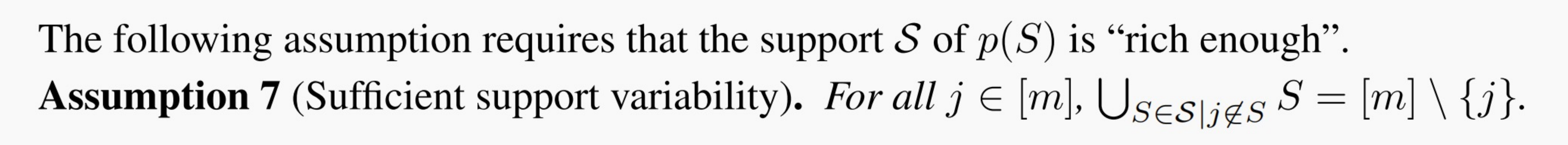


Figure 2: Illustration of Assumption 6 showing three examples of distribution $\mathbb{P}_{\boldsymbol{W}|S}$. The red distribution satisfies the assumption, but the blue and orange distributions do not. The red lines are level sets of a Gaussian distribution with full rank covariance. The blue line represents the support of a Gaussian distribution with a low rank covariance. The orange dots represents a distribution with finite support. The green vector $\boldsymbol{a}$ shows that the condition is violated for both the blue and the orange distribution, since, in both cases, $\boldsymbol{W}_{1,S}\boldsymbol{a} = 0$ (orthogonal) with probability greater than zero.

# Disentanglement via sparse multitask learning



The following assumption requires that the support $\mathcal{S}$ of $p(S)$ is “rich enough”.

**Assumption 7** (Sufficient support variability). *For all* $j \in [m]$, $\bigcup_{S \in \mathcal{S} \mid j \notin S} S = [m] \setminus \{j\}$.

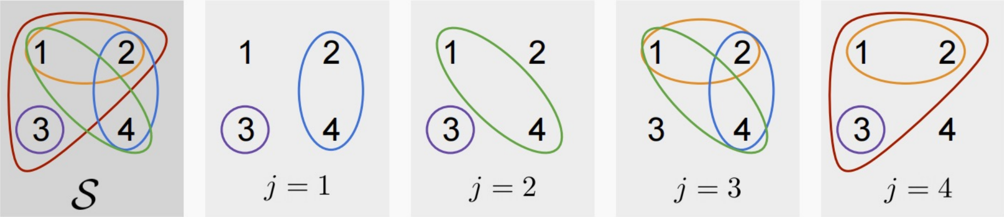
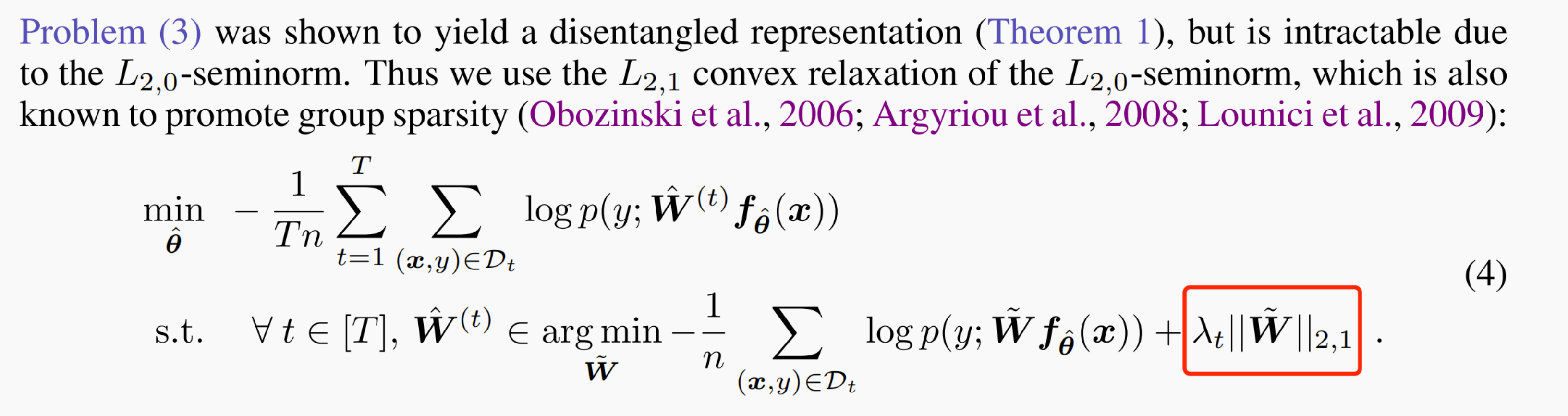


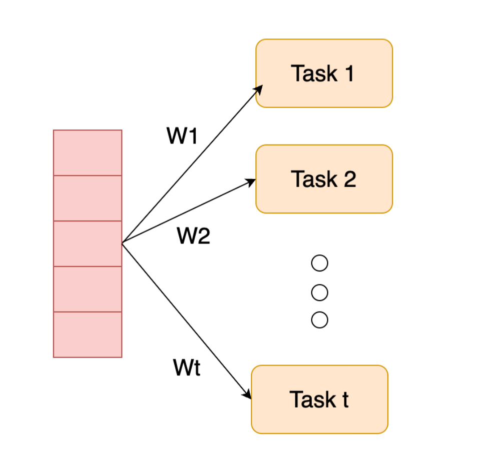
Figure 3: The leftmost figure represents $\mathcal{S}$, the support of some $p(S)$. The other figures form a verification that Assumption 7 holds for $\mathcal{S}$.

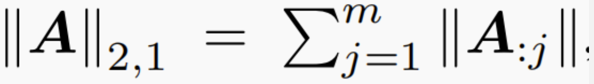
# Tractable bilevel optimization problems for sparse multitask learning



Problem (3) was shown to yield a disentangled representation (Theorem 1), but is intractable due to the $L_{2,0}$-seminorm. Thus we use the $L_{2,1}$ convex relaxation of the $L_{2,0}$-seminorm, which is also known to promote group sparsity (Obozinski et al., 2006; Argyriou et al., 2008; Lounici et al., 2009):

$$\begin{aligned} \min_{\hat{\boldsymbol{\theta}}} \quad & -\frac{1}{Tn}\sum_{t=1}^{T}\sum_{(\boldsymbol{x},y)\in\mathcal{D}_t} \log p(y; \hat{\boldsymbol{W}}^{(t)} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})) \\ \text{s.t.} \quad & \forall\, t \in [T],\ \hat{\boldsymbol{W}}^{(t)} \in \operatorname*{arg\,min}_{\tilde{\boldsymbol{W}}} -\frac{1}{n}\sum_{(\boldsymbol{x},y)\in\mathcal{D}_t} \log p(y; \tilde{\boldsymbol{W}} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})) + \boxed{\lambda_t \|\tilde{\boldsymbol{W}}\|_{2,1}}\ . \end{aligned} \tag{4}$$



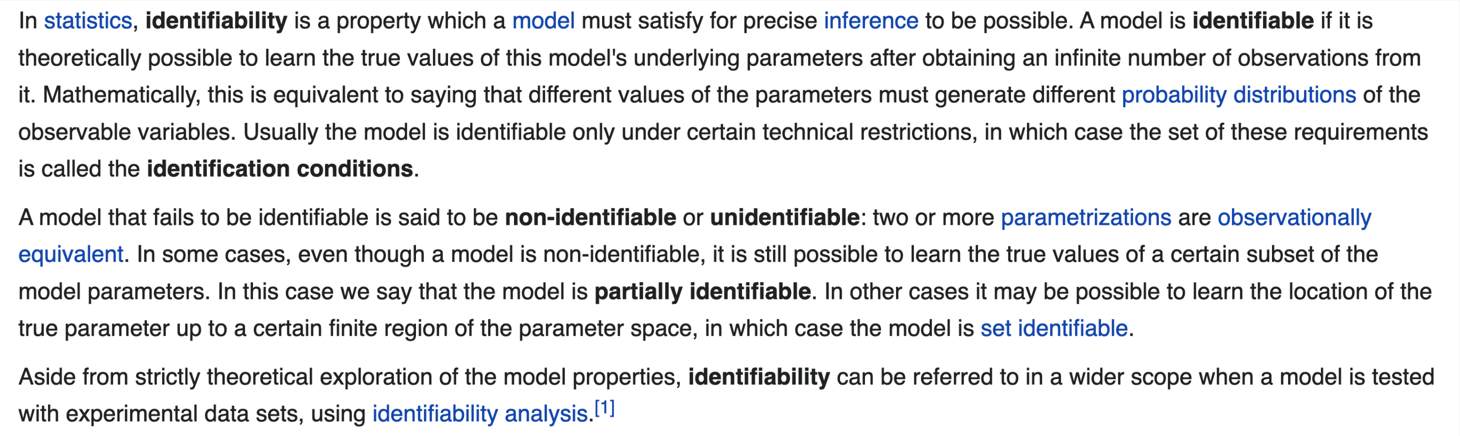


$$\|\boldsymbol{A}\|_{2,1} = \sum_{j=1}^{m} \|\boldsymbol{A}_{:j}\|$$

# Link with meta-learning

- Not done

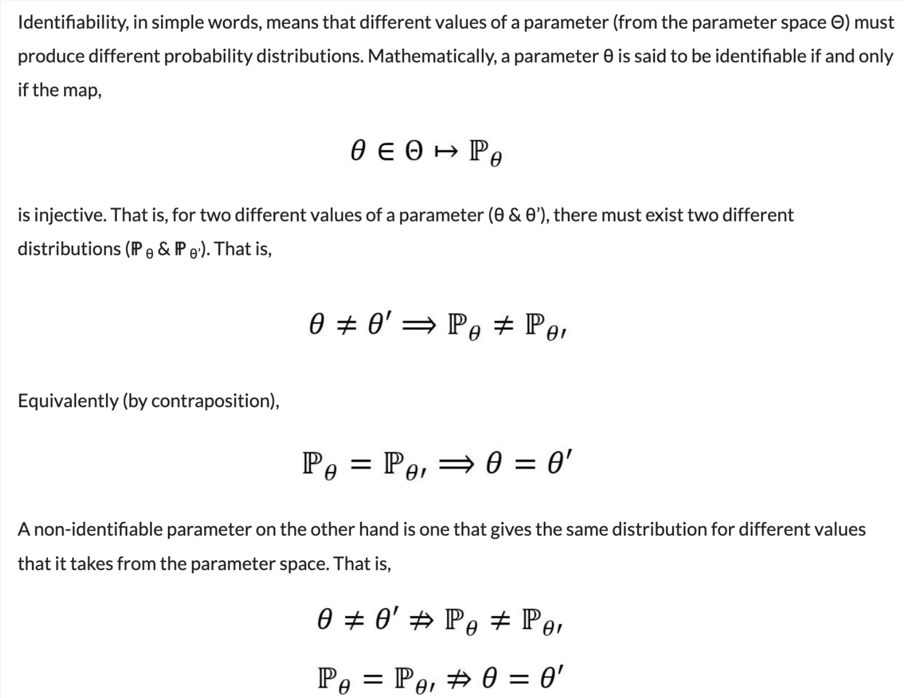
# Identifiability



In statistics, **identifiability** is a property which a model must satisfy for precise inference to be possible. A model is **identifiable** if it is theoretically possible to learn the true values of this model's underlying parameters after obtaining an infinite number of observations from it. Mathematically, this is equivalent to saying that different values of the parameters must generate different probability distributions of the observable variables. Usually the model is identifiable only under certain technical restrictions, in which case the set of these requirements is called the **identification conditions**.

A model that fails to be identifiable is said to be **non-identifiable** or **unidentifiable**: two or more parametrizations are observationally equivalent. In some cases, even though a model is non-identifiable, it is still possible to learn the true values of a certain subset of the model parameters. In this case we say that the model is **partially identifiable**. In other cases it may be possible to learn the location of the true parameter up to a certain finite region of the parameter space, in which case the model is set identifiable.

Aside from strictly theoretical exploration of the model properties, **identifiability** can be referred to in a wider scope when a model is tested with experimental data sets, using identifiability analysis.[1]



Identifiability, in simple words, means that different values of a parameter (from the parameter space Θ) must produce different probability distributions. Mathematically, a parameter θ is said to be identifiable if and only if the map,

$$\theta \in \Theta \mapsto \mathbb{P}_\theta$$

is injective. That is, for two different values of a parameter (θ & θ'), there must exist two different distributions ($\mathbb{P}_\theta$ & $\mathbb{P}_{\theta'}$). That is,

$$\theta \neq \theta' \implies \mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$$

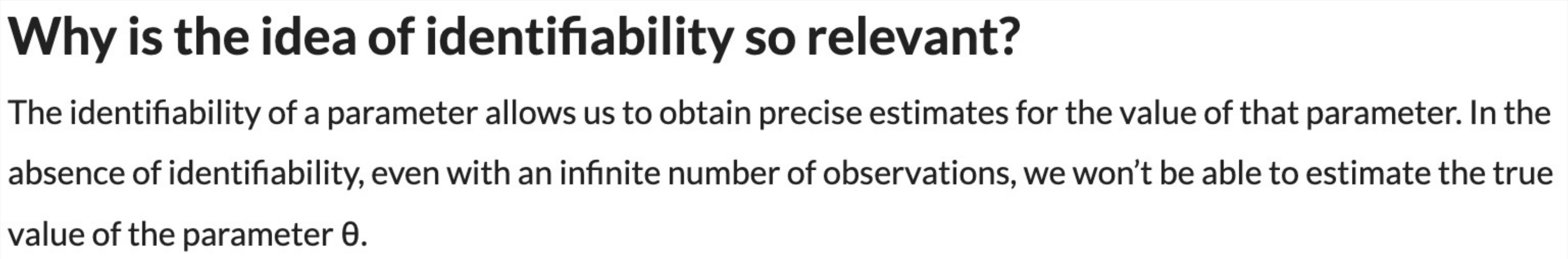
Equivalently (by contraposition),

$$\mathbb{P}_\theta = \mathbb{P}_{\theta'} \implies \theta = \theta'$$

A non-identifiable parameter on the other hand is one that gives the same distribution for different values that it takes from the parameter space. That is,

$$\theta \neq \theta' \not\Rightarrow \mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$$

$$\mathbb{P}_\theta = \mathbb{P}_{\theta'} \not\Rightarrow \theta = \theta'$$

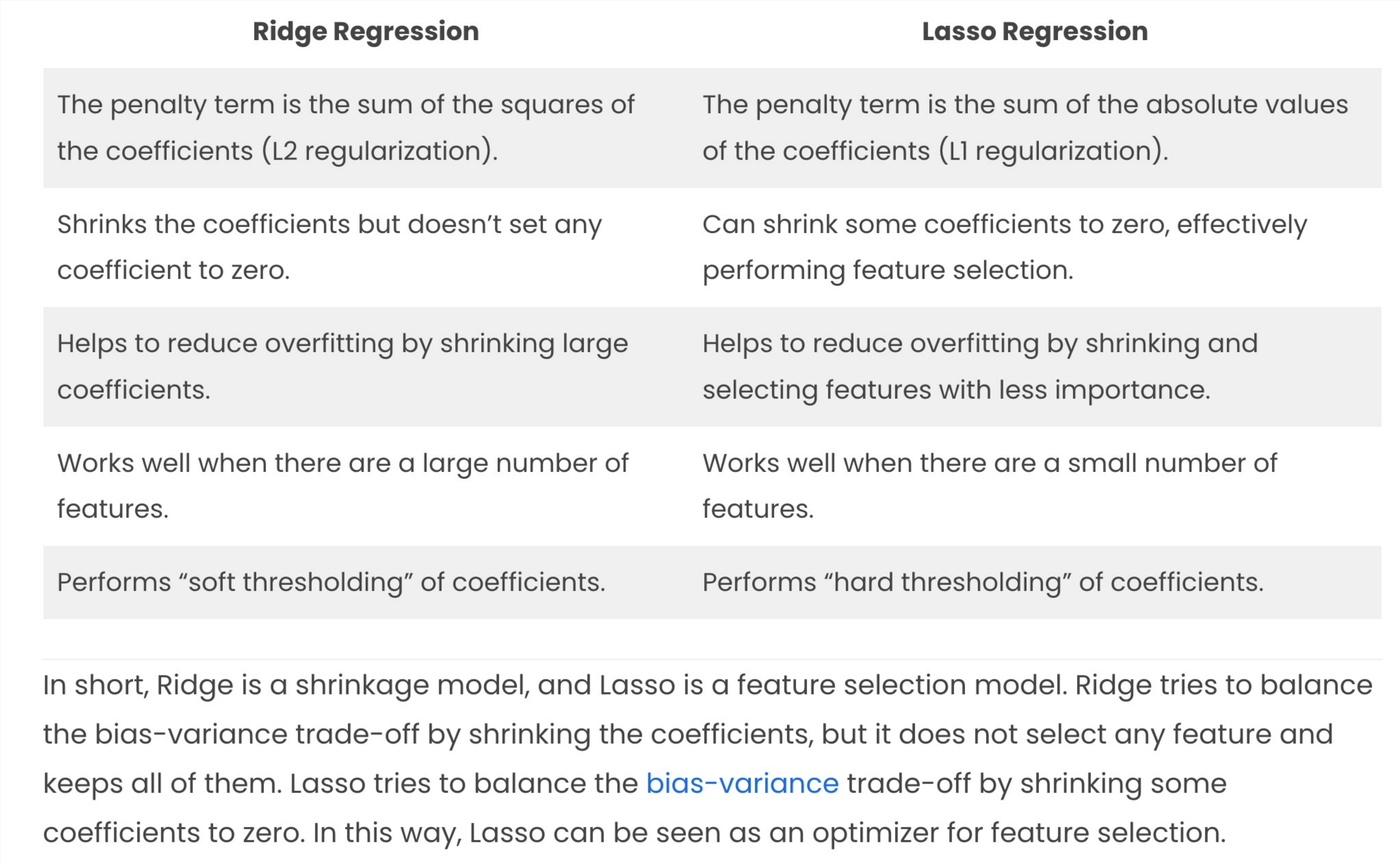


## Why is the idea of identifiability so relevant?

The identifiability of a parameter allows us to obtain precise estimates for the value of that parameter. In the absence of identifiability, even with an infinite number of observations, we won't be able to estimate the true value of the parameter θ.

https://en.wikipedia.org/wiki/Identifiability
https://www.analyticsvidhya.com/blog/2021/05/statistical-modelling-and-identifiability-of-parameters

# Lasso & Ridge Regression



| Ridge Regression | Lasso Regression |
|---|---|
| The penalty term is the sum of the squares of the coefficients (L2 regularization). | The penalty term is the sum of the absolute values of the coefficients (L1 regularization). |
| Shrinks the coefficients but doesn’t set any coefficient to zero. | Can shrink some coefficients to zero, effectively performing feature selection. |
| Helps to reduce overfitting by shrinking large coefficients. | Helps to reduce overfitting by shrinking and selecting features with less importance. |
| Works well when there are a large number of features. | Works well when there are a small number of features. |
| Performs “soft thresholding” of coefficients. | Performs “hard thresholding” of coefficients. |

In short, Ridge is a shrinkage model, and Lasso is a feature selection model. Ridge tries to balance the bias-variance trade-off by shrinking the coefficients, but it does not select any feature and keeps all of them. Lasso tries to balance the bias-variance trade-off by shrinking some coefficients to zero. In this way, Lasso can be seen as an optimizer for feature selection.

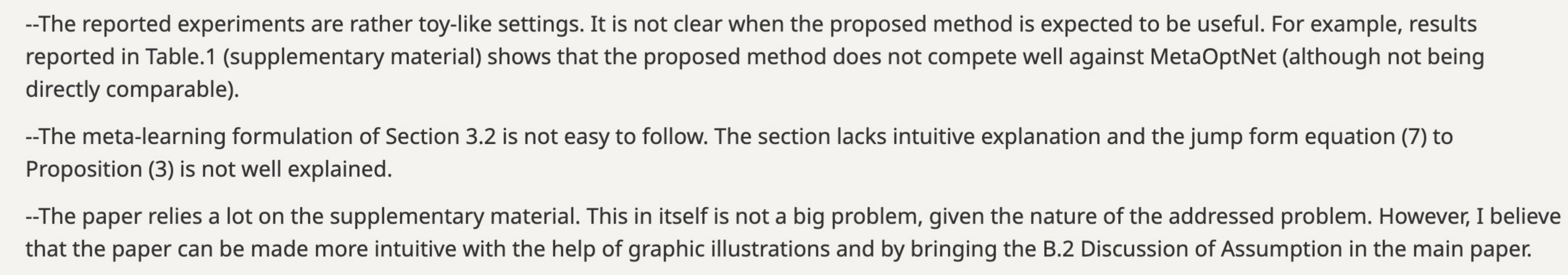
https://www.mygreatlearning.com/blog/understanding-of-lasso-regression

# Recall

- Unsupervised learning of disentangled representations is fundamentally impossible without inductive biases on both the models and the data (the lack of **identifiability**).
- Sparse base-predictors as inductive bias.

Challenging common assumptions in the unsupervised learning of disentangled representations

# Review



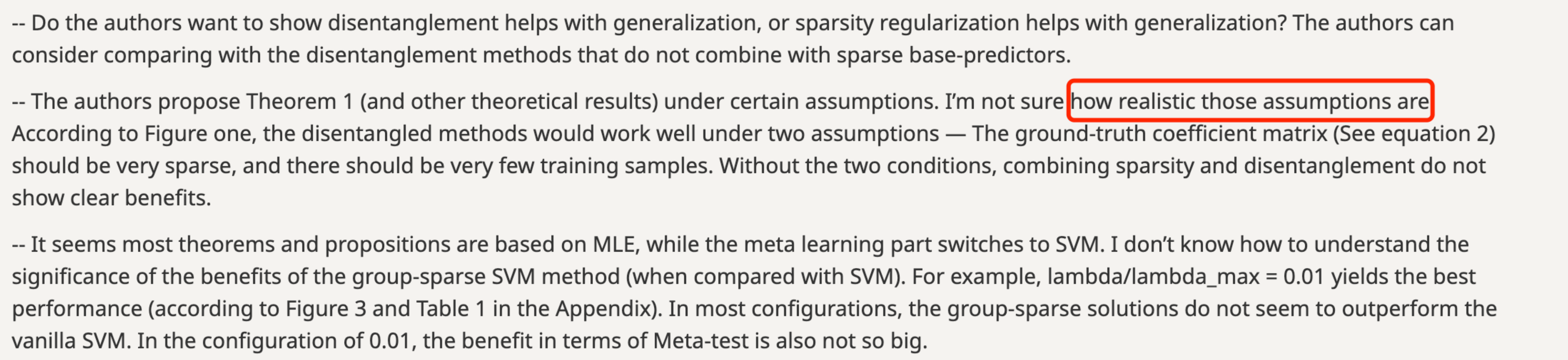
--The reported experiments are rather toy-like settings. It is not clear when the proposed method is expected to be useful. For example, results reported in Table.1 (supplementary material) shows that the proposed method does not compete well against MetaOptNet (although not being directly comparable).

--The meta-learning formulation of Section 3.2 is not easy to follow. The section lacks intuitive explanation and the jump form equation (7) to Proposition (3) is not well explained.

--The paper relies a lot on the supplementary material. This in itself is not a big problem, given the nature of the addressed problem. However, I believe that the paper can be made more intuitive with the help of graphic illustrations and by bringing the B.2 Discussion of Assumption in the main paper.

https://openreview.net/forum?id=7ZcyRF7Y3S

# Review



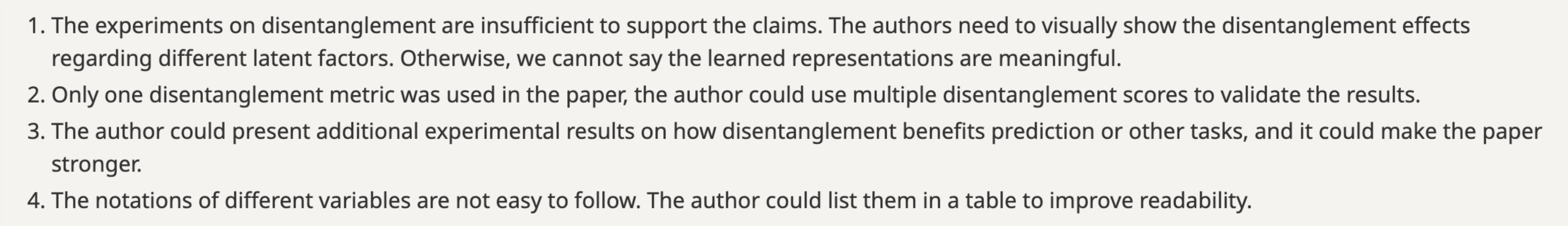
-- Do the authors want to show disentanglement helps with generalization, or sparsity regularization helps with generalization? The authors can consider comparing with the disentanglement methods that do not combine with sparse base-predictors.

-- The authors propose Theorem 1 (and other theoretical results) under certain assumptions. I'm not sure how realistic those assumptions are According to Figure one, the disentangled methods would work well under two assumptions — The ground-truth coefficient matrix (See equation 2) should be very sparse, and there should be very few training samples. Without the two conditions, combining sparsity and disentanglement do not show clear benefits.

-- It seems most theorems and propositions are based on MLE, while the meta learning part switches to SVM. I don't know how to understand the significance of the benefits of the group-sparse SVM method (when compared with SVM). For example, lambda/lambda_max = 0.01 yields the best performance (according to Figure 3 and Table 1 in the Appendix). In most configurations, the group-sparse solutions do not seem to outperform the vanilla SVM. In the configuration of 0.01, the benefit in terms of Meta-test is also not so big.

https://openreview.net/forum?id=7ZcyRF7Y3S

# Review



1. The experiments on disentanglement are insufficient to support the claims. The authors need to visually show the disentanglement effects regarding different latent factors. Otherwise, we cannot say the learned representations are meaningful.
2. Only one disentanglement metric was used in the paper, the author could use multiple disentanglement scores to validate the results.
3. The author could present additional experimental results on how disentanglement benefits prediction or other tasks, and it could make the paper stronger.
4. The notations of different variables are not easy to follow. The author could list them in a table to improve readability.

https://openreview.net/forum?id=7ZcyRF7Y3S

# Discussion

- Wahaha
- Tks