Synergies between Disentanglement and Sparsity: Generalization and Identifiability in Multi-Task Learning

Sébastien Lachapelle*, Tristan Deleu*, Divyat Mahajan,
Ioannis Mitliagkas, Yoshua Bengio, Simon Lacoste-Julien, Quentin Bertrand

Mila
Université de Montréal

Contributions

- We show how **disentangled representation + sparsity-regularized predictors** can **improve generalization** when the downstream task is "sparse"
- We introduce a **novel identifiability result**, showing how one can leverage **multiple sparse tasks** to learn a shared disentangled representation, by regularizing the task-specific predictors to be **maximally sparse across tasks**
- We propose a tractable **bilevel optimization problem** to learn this shared representation while regularizing task-specific predictors to be sparse
- We draw connections with the **meta-learning** algorithm MetaOptNet [3]

Disentanglement + Sparse Tasks = Generalization

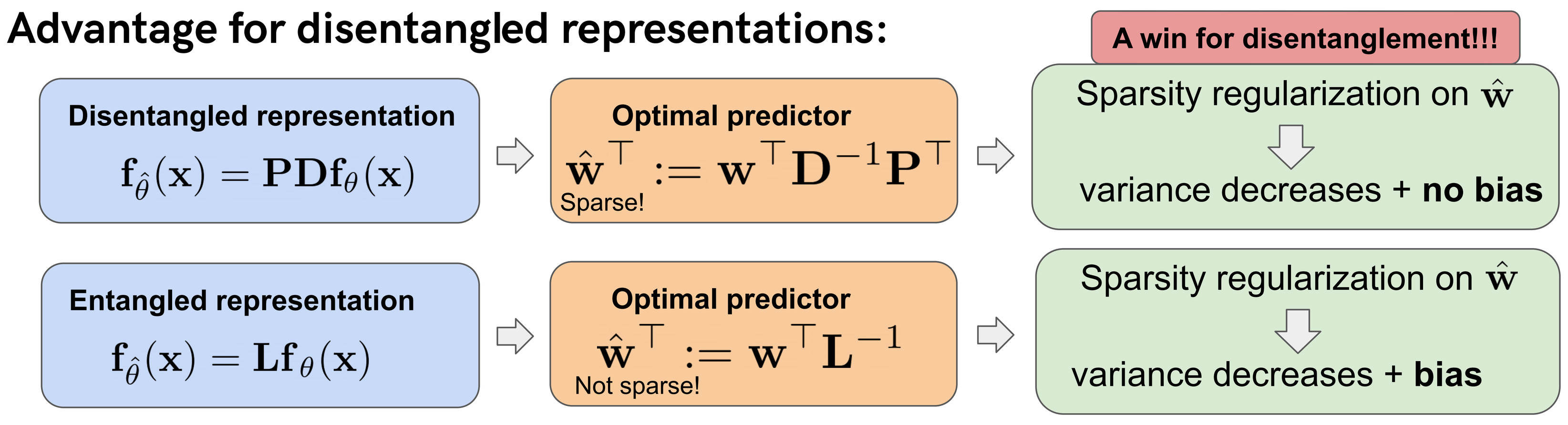
- **Sparse tasks:** Input-label pairs $(\boldsymbol{x}, y)$ are sampled from an unknown process:

$$\boldsymbol{x} \sim p(\boldsymbol{x}) \qquad y = \boldsymbol{w}^\top \boldsymbol{f}_\theta(\boldsymbol{x}) \quad \text{where } \boldsymbol{w} \text{ is sparse}$$

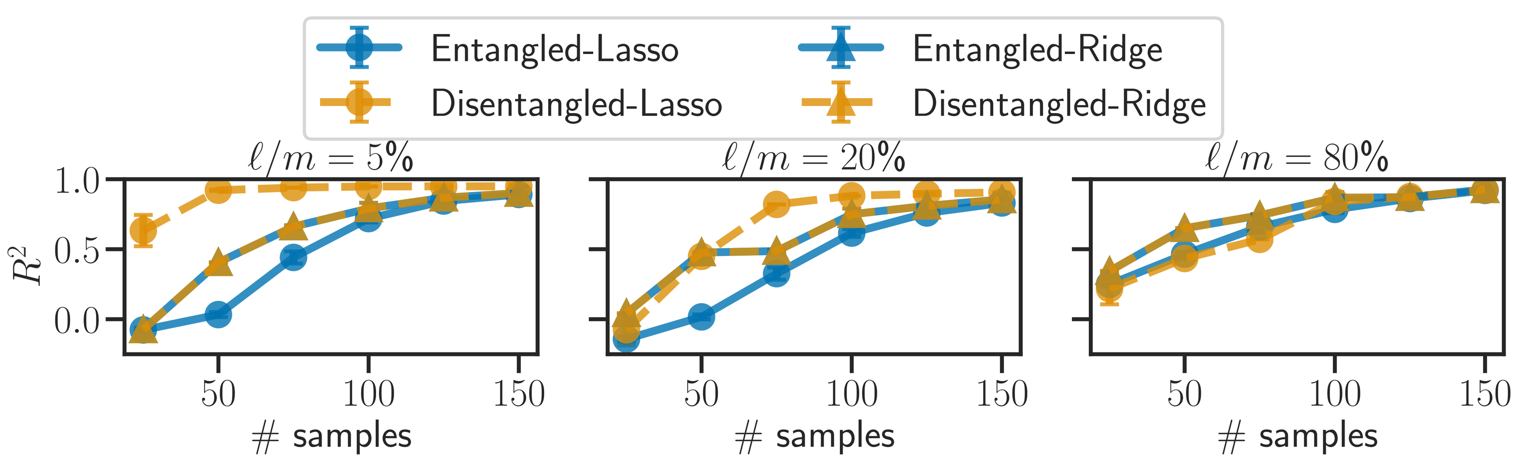
- **Assumption:** The learned representation is **linearly equivalent** to the ground-truth, i.e. there exists an invertible matrix $\boldsymbol{L}$ such that $\boldsymbol{f}_{\hat{\theta}}(\boldsymbol{x}) = \boldsymbol{L}\boldsymbol{f}_{\theta}(\boldsymbol{x})$ [4]
- **Optimal predictor for learned representation** is $\hat{\boldsymbol{w}}^\top := \boldsymbol{w}^\top \boldsymbol{L}^{-1}$ since

$$\hat{\boldsymbol{w}}^\top \boldsymbol{f}_{\hat{\theta}}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{L}^{-1}\boldsymbol{L}\boldsymbol{f}_{\theta}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{f}_{\theta}(\boldsymbol{x})$$

- **Definition:** A learned representation $\boldsymbol{f}_{\hat{\theta}}(\boldsymbol{x})$ is **disentangled** w.r.t. a ground-truth representation $\boldsymbol{f}_{\theta}(\boldsymbol{x})$ when $\boldsymbol{f}_{\hat{\theta}}(\boldsymbol{x}) = \boldsymbol{P}\boldsymbol{D}\boldsymbol{f}_{\theta}(\boldsymbol{x})$, where $\boldsymbol{P}$ is a permutation and $\boldsymbol{D}$ is an invertible diagonal matrix
- **Advantage for disentangled representations:**



- **Experiment with frozen representations:** (ℓ/m = ratio of useful features)



Disentanglement via Sparse Multi-Task Learning

Multi-Task Learning Setting:

- **Data generating process:** For each task t, $(\boldsymbol{x}, \boldsymbol{y})$ is distributed as

$$p(\boldsymbol{x}, y \mid \boldsymbol{W}^{(t)}) = p(\boldsymbol{x} \mid \boldsymbol{W}^{(t)})p(y; \eta = \boldsymbol{W}^{(t)}\boldsymbol{f}_{\theta}(\boldsymbol{x}))$$

where $p(y;\eta)$ is distribution parameterized by η. E.g. Gaussian with $\eta = (\mu, \sigma^2)$

- **Support of task** t: $S^{(t)} := \{j \in [m] \mid \boldsymbol{W}^{(t)}_{:,j} \neq 0\}$
- **Task generating process:**

$$\boldsymbol{W}^{(t)} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\boldsymbol{W}} = \textstyle\sum_S p(S)\mathbb{P}_{\boldsymbol{W}|S} \text{ where}$$

$$p(S) = \text{distribution over task support with support } \mathcal{S}$$

$$\mathbb{P}_{\boldsymbol{W}|S} = \text{conditional distribution of } \boldsymbol{W} \text{ given its support is } S$$

Theorem: Let $\hat{\theta}$ be a minimizer of

Task-specific estimator

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

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

Sparsity regularization $\|\boldsymbol{A}\|_{2,0} = \sum_{j=1}^{m} \mathbb{1}(\|\boldsymbol{A}_{:j}\|_2 \neq 0)$

then, under Assumptions 1 to 5, $\boldsymbol{f}_{\hat{\theta}}$ is **disentangled** w.r.t. $\boldsymbol{f}_{\theta}$

[1] K. Ahuja, D. Mahajan, V. Syrgkanis, and I. Mitliagkas. Towards efficient representation identification in supervised learning. In First Conference on Causal Learning and Reasoning, 2022.

[2] Q. Bertrand, Q. Klopfenstein, M. Massias, M. Blondel, S. Vaiter, A. Gramfort, and J. Salmon. Implicit differentiation for fast hyperparameter selection in non-smooth convex learning. JMLR, 2022.

[3] K. Lee, S.Maji, A. Ravichandran, and S. Soatto. Meta-learning with differentiable convex optimization. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10657--10665, 2019.

[4] G. Roeder, L. Metz, and D. P. Kingma. On linear identifiability of learned representations. In Proceedings of the 38th International Conference on Machine Learning, 2021.

Relaxation of the Bilevel Problem

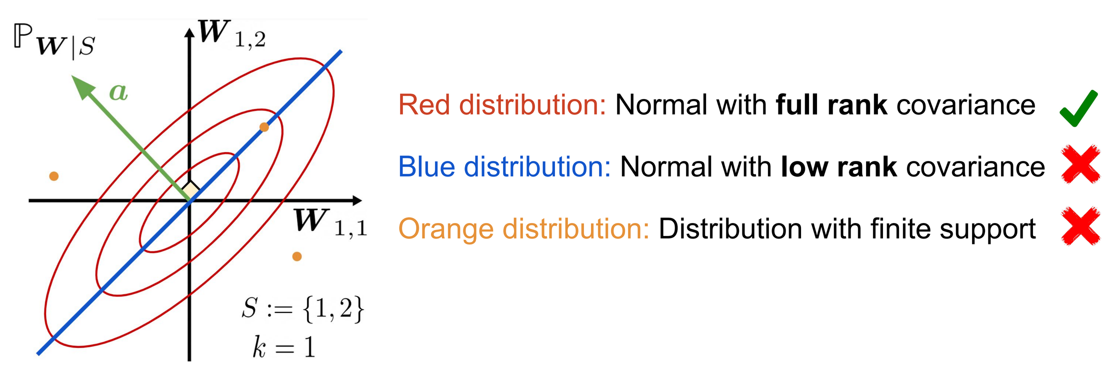
$$\min_{\hat{\theta}} \; -\frac{1}{Tn}\sum_{t=1}^{T}\sum_{(\boldsymbol{x},y)\in\mathcal{D}_t} \log p(y; \hat{\boldsymbol{W}}^{(t)}\boldsymbol{f}_{\hat{\theta}}(\boldsymbol{x}))$$

$$\text{s.t.} \; \hat{\boldsymbol{W}}^{(t)} \in \underset{\tilde{\boldsymbol{W}}}{\arg\min} -\frac{1}{n}\sum_{(\boldsymbol{x},y)\in\mathcal{D}_t} \log p(y; \tilde{\boldsymbol{W}}\boldsymbol{f}_{\hat{\theta}}(\boldsymbol{x})) + \lambda_t \underbrace{\|\tilde{\boldsymbol{W}}\|_{2,1}}_{\|\boldsymbol{A}\|_{2,1} = \sum_{j=1}^{m}\|\boldsymbol{A}_{:j}\|_2}$$

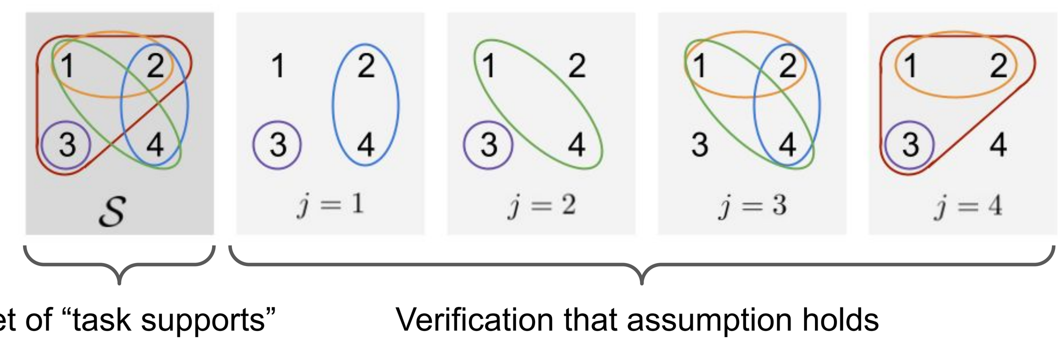
- We need to **"backpropagate through the solution of the inner problem"**
- We can compute the gradient of the (outer) objective w.r.t. $\hat{\theta}$ via backpropagation & **implicit differentiation**
- This can be done **even if the inner objective is non-smooth** [2]

Assumptions for Identifiability Result

- **Assumption 1** $\mathrm{KL}(p(y;\boldsymbol{\eta}) \,||\, p(y;\tilde{\boldsymbol{\eta}})) = 0 \implies \boldsymbol{\eta} = \tilde{\boldsymbol{\eta}}$, where KL denotes the Kullback-Leibler divergence
- **Assumption 2** There exists $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)} \in \mathcal{X}$ such that the matrix $\boldsymbol{F} := [\boldsymbol{f}_{\theta}(\boldsymbol{x}^{(1)}), \ldots, \boldsymbol{f}_{\theta}(\boldsymbol{x}^{(m)})]$ is invertible
- **Assumption 3** There exists $\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(m)} \in \mathcal{W}$ and indices $i_1, \ldots, i_m \in [k]$ such that the rows $\boldsymbol{W}^{(1)}_{i_1,:}, \ldots, \boldsymbol{W}^{(m)}_{i_m,:}$ are linearly independent
- **Assumption 4** For all $S \in \mathcal{S}$ and all $\boldsymbol{a} \in \mathbb{R}^{|S|}\setminus\{0\}$, $\mathbb{P}_{\boldsymbol{W}|S}[\boldsymbol{W}_{:S}\boldsymbol{a} = \boldsymbol{0}] = 0$

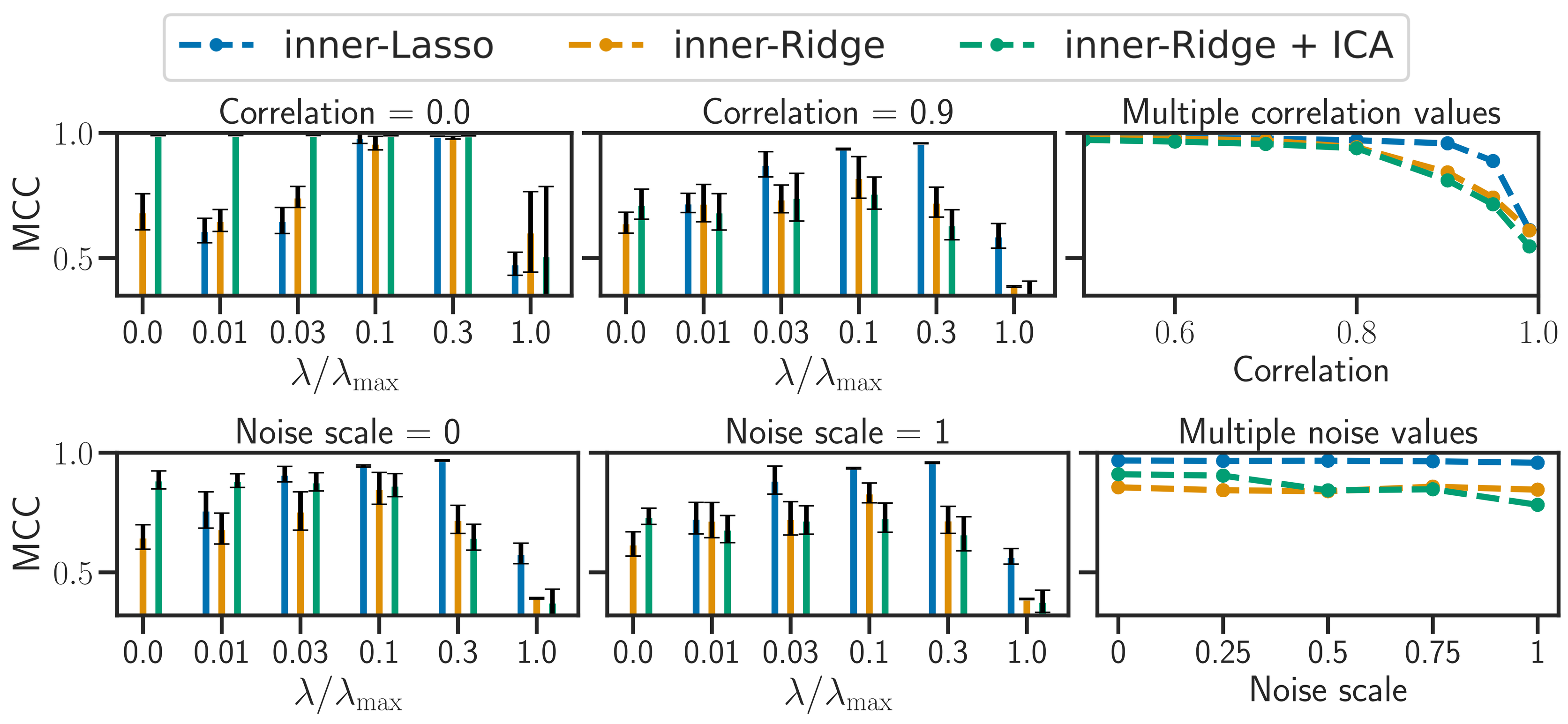


- **Assumption 5** For all $j \in [m]$, $\bigcup_{S\in\mathcal{S}|j\notin S} S = [m] \setminus \{j\}$



Semi-Synthetic Experiments on 3D Shapes

- We control the distribution over latents (various correlation & noise levels)
- Ground-truth labels are given by $y = \boldsymbol{w}^{(t)}\boldsymbol{f}_{\theta}(\boldsymbol{x}) + \epsilon$ where $\boldsymbol{w}^{(t)}$ are sampled from a **spike and slab** distribution to induce sparsity
- Inner-Ridge + ICA w/o regularization = [1] (assumes independent features)



- **Latent representation responses** to changing a single factor of variation (correlation 0.9 between latents, MCC=0.96):