SYNERGIES BETWEEN DISENTANGLEMENT AND SPARSITY: A MULTI-TASK LEARNING PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Although disentangled representations are often said to be beneficial for downstream tasks, current empirical and theoretical understanding is limited. In this work, we provide evidence that disentangled representations coupled with sparse base-predictors improve generalization. In the context of multi-task learning, we prove a new identifiability result that provides conditions under which maximally sparse base-predictors yield disentangled representations. Motivated by this theoretical result, we propose a practical approach to learn disentangled representations based on a sparsity-promoting bi-level optimization problem. Finally, we explore a meta-learning version of this algorithm based on group Lasso multiclass SVM base-predictors, for which we derive a tractable dual formulation. It obtains competitive results on standard few-shot classification benchmarks, while each task is using only a fraction of the learned representations.

1 INTRODUCTION

The recent literature on self-supervised learning has provided evidence that learning a representation on large corpuses of data can yield strong performances on a wide variety of downstream tasks (Devlin et al., 2018; Chen et al., 2020), especially in few-shot learning scenarios where the training data for these tasks is limited (Brown et al., 2020b; Dosovitskiy et al., 2021; Radford et al., 2021). Beyond transferring across multiple tasks, these learned representations also lead to improved robustness against distribution shifts (Wortsman et al., 2022) as well as stunning text-conditioned image generation (Ramesh et al., 2022). However, preliminary assessments of the latter has highlighted shortcomings related to compositionality (Marcus et al., 2022), suggesting new algorithmic innovations are needed to make further progress.

Another line of work has argued for the integration of ideas from causality to make progress towards more robust and transferable machine learning systems (Pearl, 2019; Schölkopf, 2019; Goyal & Bengio, 2021). *Causal representation learning* has emerged recently as a field aiming to define and learn representations suited for causal reasoning (Schölkopf et al., 2021). This set of ideas is strongly related to learning *disentangled representations* (Bengio et al., 2013). 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. Although a plethora of works have investigated theoretically under which conditions disentanglement is possible (Hyvarinen & Morioka, 2016; 2017; Hyvärinen et al., 2019; Khemakhem et al., 2020a; Locatello et al., 2020a; Klindt et al., 2021; Von Kügelgen et al., 2021; Gresele et al., 2021; Lachapelle et al., 2022; Lippe et al., 2022b; Ahuja et al., 2022c), fewer works have tackled *how a disentangled representation could be beneficial for downstream tasks*. Those who did mainly provide empirical rather than theoretical evidence for or against its usefulness (Locatello et al., 2019; van Steenkiste et al., 2019; Miladinović et al., 2019; Dittadi et al., 2021; Montero et al., 2021).

In this work, we explore synergies between disentanglement and sparse base-predictors in the context of multi-task learning. At the heart of our contributions is the assumption that 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. We will refer to such tasks as *sparse tasks*, and their corresponding sets of useful factors as their *supports*. This assumption was initially suggested by Bengio et al. (2013, Section 3.5): "the feature set being trained may be destined to be used in multiple tasks that may have distinct [and unknown] subsets of relevant features. Considerations such as these lead us to

the conclusion that the most robust approach to feature learning is to disentangle as many factors as possible, discarding as little information about the data as is practical". This strategy is very much in line with the current self-supervised learning trend (Radford et al., 2021), except for its focus on disentanglement.

Our main contributions are the following: (i) We formalize this "sparse task assumption" and argue theoretically and empirically how, in this context, disentangled representations coupled with sparsity-regularized base-predictors can obtain better generalization than their entangled counterparts (Section 2.1). (ii) We introduce a novel identifiability result (Theorem 1) which shows how one can leverage multiple sparse tasks to learn a shared disentangled representation by regularizing the task-specific predictors to be maximally sparse (Section 2.2.1). Crucially, Assumption 7 formalizes how diverse the task supports have to be in order to guarantee disentanglement. (iii) Motivated by this result, we propose a tractable bi-level optimization (Problem (5)) to learn the shared representation while regularizing the task-specific base-predictors to be sparse (Section 2.2.2). We validate our theory by showing our approach can indeed disentangle latent factors on tasks constructed from the 3D Shapes dataset (Burgess & Kim, 2018). (iv) Finally, we draw a connection between this bi-level optimization problem and some formulations from the meta-learning literature (Section 2.3). Inspired by our identifiability result, we enhance an existing method (Lee et al., 2019), where the base-learners are now group-sparse SVMs. We show that this new meta-learning algorithm achieves competitive performance on the *mini*ImageNet benchmark (Vinyals et al., 2016), while only using a fraction of the learned representation.

2 Synergies between Disentanglement and Sparsity

In this section, we formally introduce the notion of entangled and disentangled representations. First, we assume the existence of some ground-truth encoder function $\boldsymbol{f_\theta} : \mathbb{R}^d \to \mathbb{R}^m$ that maps observations $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$, *e.g.*, images, to its corresponding interpretable and usually lower dimensional representation $\boldsymbol{f_\theta}(\boldsymbol{x}) \in \mathbb{R}^m$, $m \leq d$. The exact form of this ground-truth encoder depends on the task at hand, but also on what the machine learning practitioner considers as interpretable. The learned encoder function is denoted by $\boldsymbol{f_{\hat\theta}} : \mathbb{R}^d \to \mathbb{R}^m$, and should not be conflated with the ground-truth representation $\boldsymbol{f_\theta}$. For example, $\boldsymbol{f_{\hat\theta}}$ can be parametrized by a neural network. Throughout, we are going to use the following definition of disentanglement.

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

Intuitively, a representation is disentangled when there is a one-to-one correspondence between its components and the components of the ground-truth representation, up to rescaling. Note that there exist less stringent notions of disentanglement which allow for component-wise nonlinear invertible transformations of the factors (Hyvarinen & Morioka, 2017; Hyvärinen et al., 2019).

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.

2.1 Disentanglement and sparse base-predictors for improved generalization

In this section, we compare the generalization performance of entangled and disentangled representations on sparse downstream tasks. We show that the maximum likelihood estimator (defined in Problem (1)) computed on *linearly equivalent* representations (entangled or disentangled) yield the same model (Proposition 1). However, disentangled representations have better generalization properties when combined with a sparse base-predictor (Proposition 2 and Figure 1).

First, the learned representation $\boldsymbol{f_{\hat\theta}}$ is assumed to be *linearly equivalent* to the ground-truth representation $\boldsymbol{f_\theta}$, i.e. there exists an invertible matrix $\boldsymbol{L}$ such that, for all $x \in \mathcal{X}$, $\boldsymbol{f_{\hat\theta}}(\boldsymbol{x}) = \boldsymbol{L}\boldsymbol{f_\theta}(\boldsymbol{x})$.

Note that despite being assumed linearly equivalent, the learned representation $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}$ might not be disentangled (Definition 1); in that case, we say the representation is *linearly entangled*. When we refer to a disentangled representation, we write $\boldsymbol{L} := \boldsymbol{D}\boldsymbol{P}$. Roeder et al. (2021) have shown that many common methods learn representations identifiable up to linear equivalence, such as deep neural networks for classification, contrastive learning (Oord et al., 2018; Radford et al., 2021) and autoregressive language models (Mikolov et al., 2010; Brown et al., 2020a).

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

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

where y denotes the label, $\mathcal{D} := \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^n$ is the dataset, $p(y;\boldsymbol{\eta})$ is a distribution over labels parameterized by $\boldsymbol{\eta} \in \mathbb{R}^k$, and $\hat{\boldsymbol{W}} \in \mathbb{R}^{k\times m}$ is the task-specific predictor[2]. The following result shows that the maximum likelihood estimator defined in Problem (1) is invariant to invertible linear transformations of the features. Note that it is an almost direct consequence of the invariance of MLE to reparametrization (Casella & Berger, 2001, Thm. 7.2.10). See Appendix A for a proof.

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.

In what follows, we assume the data is generated according to the following process.

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.*

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:

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

The following proposition will help us understand how this additional constraint interacts with representations that are disentangled or linearly entangled. See Appendix A for a proof.

Proposition 2 (Population MLE for Linearly Entangled Representations)**.** *Let $\hat{\boldsymbol{W}}_\infty^{(\hat{\boldsymbol{\theta}})}$ 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, $\hat{\boldsymbol{W}}_\infty^{(\hat{\boldsymbol{\theta}})} = \boldsymbol{W}\boldsymbol{L}^{-1}$.*

From Proposition 2, one can see that if the representation $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}$ is disentangled, then $\|\hat{\boldsymbol{W}}_\infty^{(\hat{\boldsymbol{\theta}})}\|_{2,0} = \|\boldsymbol{W}(\boldsymbol{D}\boldsymbol{P})^{-1}\|_{2,0} = \|\boldsymbol{W}\|_{2,0} = \ell$. 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.

Empirical validation. We now present a simple simulated experiment to validate the above claim. Figure 1 compares the generalization performance of the convex relaxation of Problem (2) (Lasso

[1] We assume the solution is unique.

[2] $p(y;\boldsymbol{\eta})$ could be a Gaussian density (regression) or a categorical distribution (classification).

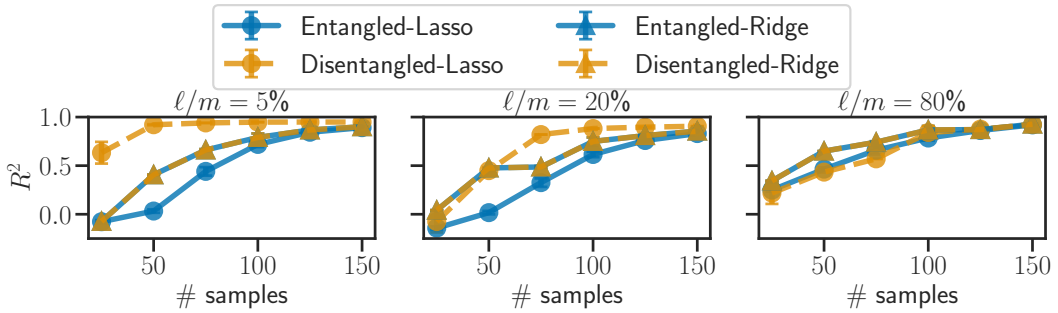

Figure 1: Test performance for the entangled and disentangled representation using Lasso and Ridge regression. All the results are averaged over 10 seeds, with standard error shown in error bars.

regression (Tibshirani, 1996)) and Ridge regression (Hoerl & Kennard, 1970) on both disentangled and linearly entangled representations. Lasso regression coupled with the disentangled representation obtains better generalization than the other alternatives when $\ell/m = 5\%$ and when the number of samples is very small. We can also see that, disentanglement, sparsity regularization and sufficient sparsity in the ground-truth data generating process are necessary to see a significant improvement, in line with our discussion. Lastly, the performance of all methods converge to the same value when the number of samples grows. See Appendix D.1 for more details and discussion on the results.

2.2 Disentanglement via sparse multitask learning

In Section 2.1, we argued that disentangled representations can improve generalization when combined with sparse base-predictors, but we did not provide an approach to learn them. We first provide a new identification result (Theorem 1, Section 2.2.1), which states that in the context of sparse multitask learning, sparse base-predictors yield disentangled representations. Then, in Section 2.2.2, we provide a practical way to learn disentangled representations motivated by our identifiability result.

Throughout this section, we assume the learner is given a set of T datasets $\{\mathcal{D}_1, \dots, \mathcal{D}_T\}$ where each dataset $\mathcal{D}_t := \{(\boldsymbol{x}^{(t,i)}, y^{(t,i)})\}_{i=1}^n$ consists of n couples of input $\boldsymbol{x} \in \mathbb{R}^d$ and label $y \in \mathcal{Y}$. The set of labels $\mathcal{Y}$ might contain either class indices or real values, depending on whether we are concerned with classification or regression tasks.

2.2.1 Identifiability analysis

We now present the main theoretical result of our work which shows how learning a shared representation across tasks while penalizing the task-specific base-predictor to be sparse can induce disentanglement. Our theory relies on the following ground-truth data generating process:

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. $\forall y;\ p(y; \boldsymbol{\eta}) = p(y; \tilde{\boldsymbol{\eta}}) \implies \boldsymbol{\eta} = \tilde{\boldsymbol{\eta}}$.*

This property holds, *e.g.*, when $p(y; \boldsymbol{\eta})$ is a Gaussian in the usual μ, σ^2 parameterization. Generally, it also holds for minimal parameterizations of exponential families (Wainwright & Jordan, 2008).

The following assumption requires the ground-truth representation $\boldsymbol{f}_\theta(\boldsymbol{x})$ 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 $\boldsymbol{F} := [\boldsymbol{f_\theta}(\boldsymbol{x}^{(1)}), \dots, \boldsymbol{f_\theta}(\boldsymbol{x}^{(m)})]$ is invertible.*

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

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

We note that, under Assumptions 2 to 5, the representation $\boldsymbol{f_\theta}$ is identifiable up to linear equivalence, as is shown by Theorem 2 in Appendix B. Similar results where shown by Roeder et al. (2021); Ahuja et al. (2022c). The next assumptions will guarantee disentanglement.

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, it will be useful to decompose $\mathbb{P}_{\boldsymbol{W}}$ as

$$\mathbb{P}_{\boldsymbol{W}} = \sum_{S \in \mathcal{P}([m])} p(S) \mathbb{P}_{\boldsymbol{W}|S} \,, \tag{3}$$

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.

The following assumption requires that $\mathbb{P}_{\boldsymbol{W}|S}$ does not concentrate some of its mass on certain proper subspaces of $\mathbb{R}^{k \times m}$.

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

We illustrate the above assumption in the simpler case where $k = 1$. The assumption holds for example when the distribution of $\boldsymbol{W}_{1,S} \mid S$ has a density w.r.t. the Lebesgue measure on $\mathbb{R}^{|S|}$, which is true for example when $\boldsymbol{W}_{1,S} \mid S \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ and the covariance matrix $\boldsymbol{\Sigma}$ is full rank. However, if $\boldsymbol{\Sigma}$ is not full rank, the probability distribution of $\boldsymbol{W}_{1,S} \mid S$ concentrates its mass on a proper linear subspace $V \subsetneq \mathbb{R}^{|S|}$, which violates Assumption 6. We provide more examples and details in Appendix B.2.

The following assumption requires that the support $\mathcal{S}$ of $p(S)$ is "rich enough". Note that this assumption bears a strong resemblance to the graphical criterion introduced by Lachapelle et al. (2022) in the context of *mechanism sparsity for disentanglement*, and was a source of inspiration.

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

Intuitively, Assumption 7 requires that, for every feature j, one can find a set of tasks such that their supports cover all features except j itself. We note that the sets $\mathcal{S}$ that contains all subsets of $[m]$ of size $1 \leq p \leq m-1$ satisfy this assumption. We provide an additional example is Appendix B.2. One could remove this assumption, but this would only guarantee *partial* disentanglement, as was explored by Lachapelle & Lacoste-Julien (2022) in the context of mechanism sparsity.

We are now ready to show the main theoretical result of this work, which provides a bi-level optimization problem for which the optimal representations are guaranteed to be disentangled. It assumes infinitely many tasks are observed, with task-specific ground-truth matrices $\boldsymbol{W}$ sampled from $\mathbb{P}_{\boldsymbol{W}}$. We denote by $\hat{\boldsymbol{W}}^{(\boldsymbol{W})}$ the task-specific estimator of $\boldsymbol{W}$. See Appendix B.1 for a proof. Note that we suggest a tractable relaxation in Section 2.2.2.

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

$$\begin{aligned} &\min_{\hat{\boldsymbol{\theta}}} \mathbb{E}_{\mathbb{P}_{\boldsymbol{W}}} \mathbb{E}_{p(\boldsymbol{x},y|\boldsymbol{W})} - \log p(y; \hat{\boldsymbol{W}}^{(\boldsymbol{W})} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})) \\ &\text{s.t.} \quad \forall\, \boldsymbol{W} \in \mathcal{W},\ \hat{\boldsymbol{W}}^{(\boldsymbol{W})} \in \underset{\substack{\tilde{\boldsymbol{W}} \text{ s.t.} \\ ||\tilde{\boldsymbol{W}}||_{2,0} \leq ||\boldsymbol{W}||_{2,0}}}{\arg\min} \mathbb{E}_{p(\boldsymbol{x},y|\boldsymbol{W})} - \log p(y; \tilde{\boldsymbol{W}} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})) \ . \end{aligned} \tag{4}$$

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

Intuitively, this optimization problem effectively selects a representation $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}$ that (i) allows a perfect fit of the data distribution, and (ii) allows the task-specific estimators $\hat{\boldsymbol{W}}^{(\boldsymbol{W})}$ to be as sparse as the ground-truth $\boldsymbol{W}$. Problem (46) in Appendix B presents a slight variation of Problem (4) which, instead of enforcing $\|\hat{\boldsymbol{W}}^{(\boldsymbol{W})}\|_{2,0} \leq \|\boldsymbol{W}\|_{2,0}$ for every task $\boldsymbol{W}$ individually, enforces the weaker constraint that $\mathbb{E}_{\mathbb{P}_{\boldsymbol{W}}}\|\hat{\boldsymbol{W}}^{(\boldsymbol{W})}\|_{2,0} \leq \mathbb{E}_{\mathbb{P}_{\boldsymbol{W}}}\|\boldsymbol{W}\|_{2,0}$ with the same guarantee that the learned representation $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}$ will be disentangled (Theorem 4).

2.2.2 Tractable bilevel optimization problems for sparse multitask learning

Problem (4) 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):

$$\begin{aligned} \min_{\hat{\boldsymbol{\theta}}} \quad & -\frac{1}{Tn}\sum_{t=1}^{T}\sum_{(\boldsymbol{x},y)\in\mathcal{D}_t} \log p(y; \hat{\boldsymbol{W}}^{(t)}\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})) \\ \text{s.t.} \quad & \forall\, t\in[T],\ \hat{\boldsymbol{W}}^{(t)} \in \arg\min_{\tilde{\boldsymbol{W}}} -\frac{1}{n}\sum_{(\boldsymbol{x},y)\in\mathcal{D}_t} \log p(y; \tilde{\boldsymbol{W}}\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})) + \lambda_t||\tilde{\boldsymbol{W}}||_{2,1} \ . \end{aligned} \tag{5}$$

Following Bengio (2000); Pedregosa (2016); Franceschi et al. (2018), one can compute the (hyper)gradient of the outer function using implicit differentiation, even if the inner optimization problem is non-smooth (Bertrand et al., 2020; Bolte et al., 2021; Malézieux et al., 2022; Bolte et al., 2022). Once the hypergradient is computed, one can optimize Problem (5) using usual first-order methods (Wright & Nocedal, 1999).

Note that the quantity $\hat{\boldsymbol{W}}^{(t)}\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$ is invariant to simultaneous rescaling of $\hat{\boldsymbol{W}}^{(t)}$ by a scalar and of $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$ by its inverse. Thus, without constraints on $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$, $\|\hat{\boldsymbol{W}}^{(t)}\|_{2,0}$ can be made arbitrarily small. This is a usual problem in sparse dictionary learning (Kreutz-Delgado et al., 2003; Mairal et al., 2008; 2009; 2011), where unit-norm constraints are usually imposed on the column of the dictionary. Here, since $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}$ is parametrized by a neural network, we suggest to apply batch or layer normalization (Ioffe & Szegedy, 2015; Ba et al., 2016) to control its norm. Since the number of relevant features might be task-dependent, Problem (5) has one regularization hyperparameter λ_t per task. To limit the number of hyperparameters in practice, we select $\lambda_t := \lambda$ for all $t \in [T]$.

2.3 A Meta-Learning formulation

Problem (5) is closely related to some bi-level optimization approaches previously existing in the meta-learning literature (Rajeswaran et al., 2019; Lorraine et al., 2020; Hospedales et al., 2021). For a large set of tasks $[T]$, we are given a training dataset $\mathcal{D}_t^{\text{train}}$, which usually contains a small number of samples. Unlike in the multi-task setting though (*i.e.,* unlike in Section 2.2), we are also given a separate test dataset $\mathcal{D}_t^{\text{test}}$ to evaluate how well the learned model generalizes to new test samples.

We focus here on metric-based meta-learning methods (Snell et al., 2017; Bertinetto et al., 2019; Lee et al., 2019), where a shared representation $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}$ is learned across all tasks, jointly with task-specific classifiers. These classifiers are usually optimization-based estimators, such as Ridge regressions (as in Bertinetto et al., 2019), or support-vector machines (as in Lee et al., 2019). Formally, this can be formulated as the following bi-level optimization problem:

$$\min_{\hat{\boldsymbol{\theta}}} \sum_{t=1}^{T} \mathcal{L}_{\text{out}}\big(\hat{\boldsymbol{W}}_t(\hat{\boldsymbol{\theta}}); \boldsymbol{F}_t^{\text{test}}(\hat{\boldsymbol{\theta}}), \boldsymbol{y}_t^{\text{test}}\big) \ \text{ s.t. } \hat{\boldsymbol{W}}_t(\hat{\boldsymbol{\theta}}) \in \arg\min_{\tilde{\boldsymbol{W}}} \mathcal{L}_{\text{in}}\big(\tilde{\boldsymbol{W}}; \boldsymbol{F}_t^{\text{train}}(\hat{\boldsymbol{\theta}}), \boldsymbol{y}_t^{\text{train}}\big), \tag{6}$$

where $\boldsymbol{F}_t^{\text{train}}$ are the representations of the inputs of $\mathcal{D}_t^{\text{train}}$ through $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}$, and similarly for $\boldsymbol{F}_t^{\text{test}}$. Inspired by Lee et al. (2019), where the base-classifiers were multiclass support-vector machines (SVMs) Crammer & Singer, 2001, we propose to use group Lasso penalized multiclass SVMs, in order to introduce sparsity in the base-learners:

$$\mathcal{L}_{\text{in}}(\boldsymbol{W}; \boldsymbol{F}, \boldsymbol{Y}) := \sum_{i=1}^{n} \max_{l\in[k]} \left(1 + (\boldsymbol{W}_{\boldsymbol{y}_i,:} - \boldsymbol{W}_{l,:})\cdot \boldsymbol{F}_{i,:} - \boldsymbol{Y}_{il}\right) + \lambda_1\|\boldsymbol{W}\|_{2,1} + \tfrac{\lambda_2}{2}\|\boldsymbol{W}\|^2 \ . \tag{7}$$

The primal objective of Problem (7) can be hard to minimize with modern solvers. Moreover in few-shot learning applications, the number of features m is usually much larger than the number of samples n (in Lee et al. 2019, $m = 1.6 \cdot 10^4$ and $n \leq 25$), hence we solve the dual of Problem (7).

Proposition 3. (Dual Group Lasso Soft-Margin Multiclass SVM.) *The dual of Problem (7) writes*

$$\min_{\boldsymbol{\Lambda} \in \mathbb{R}^{n \times k}} \frac{1}{\lambda_2} \sum_{j=1}^{m} \|\mathrm{BST}\left((\boldsymbol{Y} - \boldsymbol{\Lambda})^\top \boldsymbol{F}_{:,j}, \lambda_1\right)\|^2 + \langle \boldsymbol{Y}, \boldsymbol{\Lambda} \rangle + \sum_{i=1}^{n} \mathbb{1}_{\sum_{l=1}^{k} \boldsymbol{\Lambda}_{il} = 1} + \sum_{i=1}^{n} \sum_{l=1}^{k} \mathbb{1}_{\boldsymbol{\Lambda}_{il} \geq 0} \ , \tag{8}$$

with BST *the the block soft-thresholding operator:* $\mathrm{BST} : (x, \tau) \mapsto (1 - \tau / \|x\|)_+ \, x$. *In addition, the primal-dual link writes* $\boldsymbol{W}_{:,j} = \mathrm{BST}\left((\boldsymbol{Y} - \boldsymbol{\Lambda})^\top \boldsymbol{F}_{:,j}, \lambda_1\right) / \lambda_2$, *for all* $j \in [m]$.

Proof of Proposition 3 can be found in Appendix C.1. Problem (8) is an optimization problem composed of a smooth term and block separable non-smooth term. Hence Problem (8) can be solved efficiently using proximal block coordinate descent (Tseng, 2001; Richtárik & Takáč, 2014). As stated in Section 2.2, argmin differentiation of the solution of Problem (8) can be done using implicit differentiation (Blondel et al., 2022; Bertrand et al., 2022).

3 Related Work

Disentanglement. Since the work of Bengio et al. (2013), many methods have been proposed to learn disentangled representations based on various heuristics (Higgins et al., 2017; Chen et al., 2018; Kim & Mnih, 2018; Kumar et al., 2018; Bouchacourt et al., 2018). Following the work of Locatello et al. (2019), which highlighted the lack of identifiability in modern deep generative models, many works have proposed more or less weak forms of supervision motivated by identifiability analyses (Locatello et al., 2020a; Klindt et al., 2021; Von Kügelgen et al., 2021; Ahuja et al., 2022a;c; Zheng et al., 2022). A similar line of work have adopted the causal representation learning perspective (Lachapelle et al., 2022; Lachapelle & Lacoste-Julien, 2022; Lippe et al., 2022b;a; Ahuja et al., 2022b; Yao et al., 2022; Brehmer et al., 2022). The problem of identifiability was well known among the *independent component analysis* (ICA) community (Hyvärinen et al., 2001; Hyvärinen & Pajunen, 1999) which came up with solutions for general nonlinear mixing functions by leveraging auxiliary information (Hyvarinen & Morioka, 2016; 2017; Hyvärinen et al., 2019; Khemakhem et al., 2020a;b). Another approach is to consider restricted hypothesis classes of mixing functions (Taleb & Jutten, 1999; Gresele et al., 2021). Contrarily to most of the above works, we do not assume that the inputs $\boldsymbol{x}$ are generated by transforming a latent random variable $\boldsymbol{z}$ through a bijective decoder $\boldsymbol{g}$. Instead, we assume the existence of a not necessarily bijective ground-truth feature extractor $\boldsymbol{f_\theta}(\boldsymbol{x})$ from which the labels can be predicted using only a subset of its components in every tasks (Assumption 2). Many of these works make assumptions about the distribution of latent factors, e.g., (conditional) independence, exponential family or other parametric assumptions. In contrast, we make comparatively weaker assumptions on the support of the ground-truth features (Assumption 4), which are allowed to present dependencies (Section 4.1). Locatello et al. (2020b) proposed a semi-supervised learning approach to disentangle in cases where a few samples are labelled with the factors of variations themselves. This is different from our approach as the labels that we consider can be sampled from some $p(y; \boldsymbol{W} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}))$, which is more general. Ahuja et al. (2022c) consider a setting similar to ours, but they rely on the independence and non-gaussianity of the latent factors for disentanglement using linear ICA.

Multi-task, transfer & invariant learning. The statistical advantages of multi-task representation learning is well understood (Lounici et al., 2011a;b; Maurer et al., 2016). However, apart from Zhang et al. (2022), theoretical benefits of disentanglement for transfer learning are not clearly established. Some works have investigated this question empirically and obtained both positive (van Steenkiste et al., 2019; Miladinović et al., 2019; Dittadi et al., 2021) and negative results (Locatello et al., 2019; Montero et al., 2021). Invariant risk minimization (Arjovsky et al., 2020; Ahuja et al., 2020; Krueger et al., 2021; Lu et al., 2021) aims at learning a representation that elicits a task-invariant base-predictor. This differs from our approach which learns base-predictors that are task-specific .

Dictionary learning and sparse coding. We contrast our approach, which jointly learns a *dense representation* and sparse base-predictors (Problem (5)), with the line of work which consists in learning *sparse representations* (Chen et al., 1998; Gribonval & Lesage, 2006). For instance, sparse

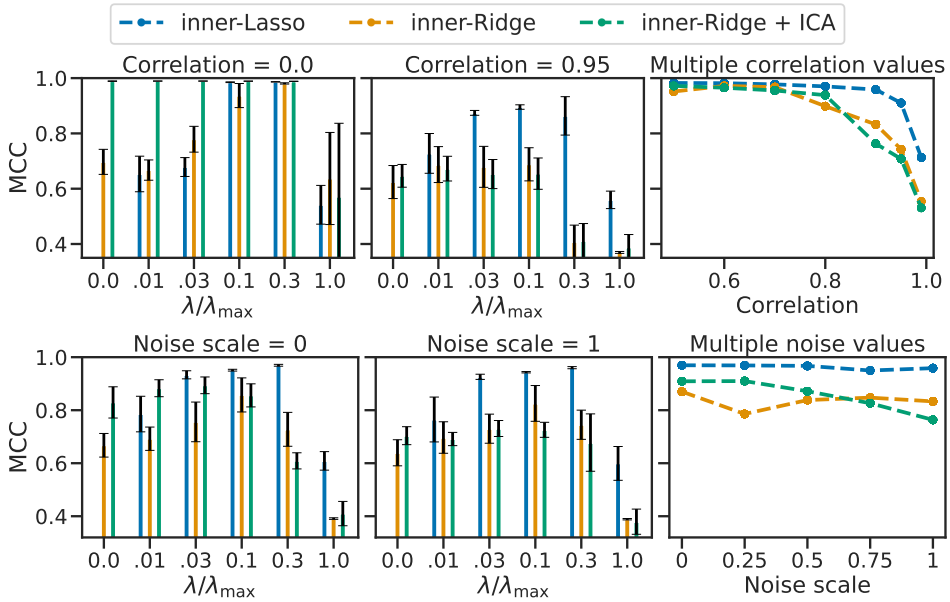

Figure 2: Disentanglement performance (MCC) for inner-Lasso, inner-Ridge and inner-Ridge combined with ICA as a function of the regularization parameter (left and middle). Varying level of correlation between latents (top) and of noise on the latents (bottom). The right columns shows performance of the best hyperparameter for different values of correlation and noise.

dictionary learning (Mairal et al., 2009; 2011; Maurer et al., 2013) is an unsupervised technique which aims at learning sparse representations that refer to atoms of a learned dictionary. Contrarily to our method which computes the representation of a single input $\boldsymbol{x}$ by evaluating a function approximator $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}$, in sparse dictionary learning, the representation of a single input is computed by minimizing a reconstruction loss. In the case of supervised dictionary learning (Mairal et al., 2008), an additional (potentially expressive) classifier is learned. This large literature has lead to a wide variety of estimators: for instance, Mairal et al. (2008, Eq. 4), which minimizes the sum of the classification error and the approximation error of the code, or Mairal et al. (2011); Malézieux et al. (2022), which introduce bi-level formulations which shares similarities with our formulations.

4 Experiments

4.1 Semi-real experiments on 3D Shapes

We now illustrate Theorem 1 by applying Problem (5) to tasks generated using the 3D Shapes dataset (Burgess & Kim, 2018).

Data generation. For all tasks t, the labelled dataset $\mathcal{D}_t = \{(\boldsymbol{x}^{(t,i)}), y^{(t,i)})\}_{i=1}^n$ is generated by first sampling the ground-truth latent variables $\boldsymbol{z}^{(t,i)} := \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}^{(t,i)})$ i.i.d. according to some distribution $p(\boldsymbol{z})$, while the corresponding input is obtained doing $\boldsymbol{x}^{(t,i)} := \boldsymbol{f}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{z}^{(t,i)})$ ($\boldsymbol{f}_{\boldsymbol{\theta}}$ is invertible in 3D Shapes). Then, a sparse weight vector $\boldsymbol{w}^{(t)}$ is sampled randomly to compute the labels of each example as $y^{(t,i)} := \boldsymbol{w}^{(t)} \cdot \boldsymbol{x}^{(t,i)} + \epsilon^{(t,i)}$, where $\epsilon^{(t,i)}$ is independent Gaussian noise. Figure 2 explores various choice of $p(\boldsymbol{z})$, i.e. by varying the level of correlation between the latent variables and by varying the level of noise on the ground-truth latents. See Appendix D.2 for more details about the data generating process.

Algorithms. In this setting where $p(y; \eta)$ is a Gaussian with fixed variance, the inner problem of Problem (5) amounts to Lasso regression, we thus refer to this approach as inner-Lasso. We also evaluate a simple variation of Problem (5) in which the L_1 norm is replaced by an L_2 norm, and refer to it as inner-Ridge. In addition we evaluate the representation obtained by performing linear

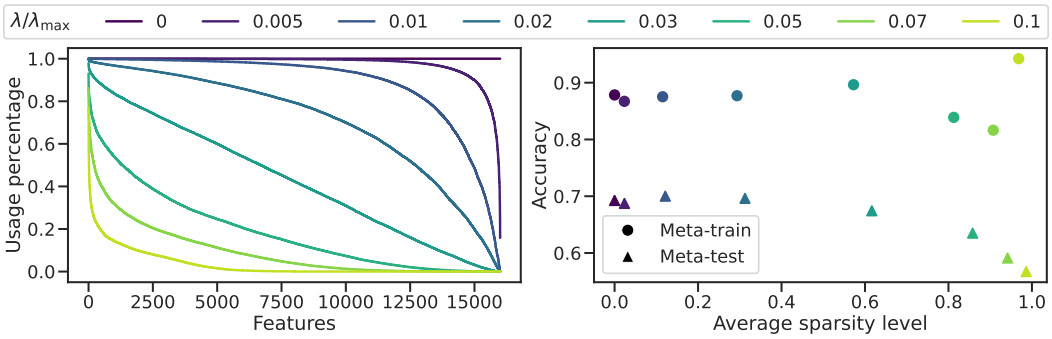

Figure 3: Effect of sparsity on the percentage of tasks using specific features, with our meta-learning objective, on *mini*ImageNet (left). The accuracy of the meta-learning algorithm and the average level of sparsity in the base-learners, as λ varies (right).

ICA (Comon, 1992) on the representation learned by inner-Ridge: the case $\lambda = 0$ corresponds to the approach of Ahuja et al. (2022c).

Discussion. Figure 2 reports disentanglement performance of the three methods, as measured by the *mean correlation coefficient*, or MCC (Hyvarinen & Morioka, 2016; Khemakhem et al., 2020a) (Appendix D.2). In all settings, inner-Lasso obtains high MCC for some values of λ, being on par or surpassing the baselines. As the theory suggests, it is robust to high levels of correlations between the latents, as opposed to inner-Ridge with ICA which is very much affected by strong correlations (since ICA assumes independence). We can also see how additional noise on the latent variables hurts inner-Ridge with ICA while leaving inner-Lasso unaffected. Figure 4 in Appendix D.2 shows that all methods find a representation which is linearly equivalent to the ground-truth representation, except for very large values of λ. Refer to Appendix D.2 for more details.

4.2 Few-shot learning experiments

Despite the lack of ground-truth latent factors in standard few-shot learning benchmarks, we also evaluate our meta-learning objective introduced in Section 2.3, using the dual formulation of the group Lasso penalized SVM as our base-learner, on the *mini*ImageNet dataset (Vinyals et al., 2016). Details about the experimental settings are provided in Appendix D.3.

Discussion. In Figure 3 (left), we report how frequently the learned features are used by the base-learner on meta-training tasks; the gradual decrease in usage suggests that the features are reused in different contexts, across different tasks. We also observe (Figure 3, right) that adding sparsity to the base learner may also improve performance on meta-training tasks, while only using a fraction of all the features available in the learned representation, supporting our observations in Section 2.1 on the effect of sparsity on generalization on natural images (see Appendix D.3 for further discussion about how this still tests generalization). We also observe that some level of sparsity improves the performance on novel meta-test tasks, albeit to a smaller extent; this setting is more akin to out-of-distribution generalization.

5 Conclusion

In this work, we investigated the synergies between sparsity, disentanglement and generalization. We showed that when the downstream task can be solved using only a fraction of the factors of variations, disentangled representations combined with sparse base-predictors can improve generalization (Section 2.1). Our novel identifiability result (Theorem 1) sheds light on how, in a multi-task setting, sparsity regularization on the task-specific predictors can induce disentanglement. This led to a practical bi-level optimization problem that was shown to yield disentangled representations on regression tasks based on the 3D Shapes dataset. Finally, we explored a meta-learning formulation extending this approach, and showed how sparse base-learners can help with generalization, while only using a small fraction of the features.

REFERENCES

K. Ahuja, K. Shanmugam, K. R. Varshney, and A. Dhurandhar. Invariant risk minimization games. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

K. Ahuja, J. Hartford, and Y. Bengio. Properties from mechanisms: an equivariance perspective on identifiable representation learning. In *International Conference on Learning Representations*, 2022a.

K. Ahuja, J. Hartford, and Y. Bengio. Weakly supervised representation learning with sparse perturbations, 2022b.

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

A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine learning*, 73 (3):243–272, 2008.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization, 2020.

J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Y. Bengio. Gradient-based optimization of hyperparameters. *Neural computation*, 12(8):1889–1900, 2000.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 2013.

L. Bertinetto, J. F. Henriques, P. HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. 2019.

Q. Bertrand, Q. Klopfenstein, M. Blondel, S. Vaiter, A. Gramfort, and J. Salmon. Implicit differentiation of lasso-type models for hyperparameter optimization. In *International Conference on Machine Learning*, pp. 810–821. PMLR, 2020.

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.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *The Annals of statistics*, 37(4):1705–1732, 2009.

M. Blondel, Q. Berthet, M. Cuturi, R. Frostig, S. Hoyer, F. Llinares-López, F. Pedregosa, and J.-P. Vert. Efficient and modular implicit differentiation. *NeurIPS*, 2022.

J. Bolte, T. Le, E., Pauwels, and T. Silveti-Falls. Nonsmooth implicit differentiation for machine-learning and optimization. *Advances in neural information processing systems*, 34:13537–13549, 2021.

J. Bolte, E. Pauwels, and S. Vaiter. Automatic differentiation of nonsmooth iterative algorithms. *NeurIPS*, 2022.

D. Bouchacourt, R. Tomioka, and S. Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

J. Brehmer, P. De Haan, P. Lippe, and T. Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, 2022.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020a.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020b.

Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

G. Casella and R. Berger. *Statistical Inference*. Duxbury Resource Center, 2001.

R. T. Q. Chen, X. Li, R. G., and D. Duvenaud. Isolating sources of disentanglement in vaes. In *Advances in Neural Information Processing Systems*, 2018.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 1998.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.

P. Comon. Independent component analysis. *Higher-Order Statistics*, 1992.

K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2(Dec):265–292, 2001.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

A. Dittadi, F. Träuble, F. Locatello, M. Wuthrich, V. Agrawal, O. Winther, S. Bauer, and B. Schölkopf. On the transfer of disentangled representations in realistic settings. In *International Conference on Learning Representations*, 2021.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021.

L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pp. 1568–1577. PMLR, 2018.

A. Goyal and Y. Bengio. Inductive biases for deep learning of higher-level cognition. *arXiv preprint arXiv:2011.15091*, 2021.

L. Gresele, J. Von Kügelgen, V. Stimper, B. Schölkopf, and M. Besserve. Independent mechanism analysis, a new concept? In *Advances in Neural Information Processing Systems*, 2021.

R. Gribonval and S. Lesage. A survey of sparse component analysis for blind source separation: principles, perspectives, and new challenges. In *ESANN'06 proceedings - 14th European Symposium on Artificial Neural Networks*, 2006.

I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021.

A. Hyvarinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In *Advances in Neural Information Processing Systems*, 2016.

A. Hyvarinen and H. Morioka. Nonlinear ICA of Temporally Dependent Stationary Sources. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 1999.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.

A. Hyvärinen, H. Sasaki, and R. E. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *AISTATS*. PMLR, 2019.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020a.

I. Khemakhem, R. Monti, D. Kingma, and A. Hyvarinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. In *Advances in Neural Information Processing Systems*, 2020b.

H. Kim and A. Mnih. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

D. A. Klindt, L. Schott, Y Sharma, I Ustyuzhaninov, W. Brendel, M. Bethge, and D. M. Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *9th International Conference on Learning Representations*, 2021.

K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural computation*, 15(2):349–396, 2003.

D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, R. Le Priol, D. Zhang, and A. Courville. Out-of-distribution generalization via risk extrapolation ({re}x), 2021.

A. Kumar, P. Sattigeri, and A. Balakrishnan. VARIATIONAL INFERENCE OF DISENTANGLED LATENT CONCEPTS FROM UNLABELED OBSERVATIONS. In *International Conference on Learning Representations*, 2018.

S. Lachapelle and S. Lacoste-Julien. Partial disentanglement via mechanism sparsity. In *UAI 2022 Workshop on Causal Representation Learning*, 2022.

S. Lachapelle, P. Rodriguez Lopez, Y. Sharma, K. E. Everett, R. Le Priol, A. Lacoste, and S. Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *First Conference on Causal Learning and Reasoning*, 2022.

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*, pp. 10657–10665, 2019.

P. Lippe, S. Magliacane, S. Löwe, Y. M Asano, T. Cohen, and E. Gavves. iCITRIS: Causal representation learning for instantaneous temporal effects. In *UAI 2022 Workshop on Causal Representation Learning*, 2022a.

P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and E. Gavves. CITRIS: Causal identifiability from temporal intervened sequences, 2022b.

F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

F. Locatello, B. Poole, G. Raetsch, B. Schölkopf, O. Bachem, and M. Tschannen. Weakly-supervised disentanglement without compromises. In *Proceedings of the 37th International Conference on Machine Learning*, 2020a.

F. Locatello, M. Tschannen, S. Bauer, G. Rätsch, B. Schölkopf, and O. Bachem. Disentangling factors of variations using few labels. In *International Conference on Learning Representations*, 2020b. URL https://openreview.net/forum?id=SygagpEKwB.

J. Lorraine, P. Vicol, and D. Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International Conference on Artificial Intelligence and Statistics*, pp. 1540–1552. PMLR, 2020.

K. Lounici, M. Pontil, A. B. Tsybakov, and S. Van De Geer. Taking advantage of sparsity in multi-task learning. *arXiv preprint arXiv:0903.1468*, 2009.

K. Lounici, M. Pontil, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of statistics*, 2011a.

K. Lounici, M. Pontil, S. Van De Geer, and A. B Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The annals of statistics*, 2011b.

Chaochao Lu, Yuhuai Wu, Jośe Miguel Hernández-Lobato, and Bernhard Schölkopf. Nonlinear invariant risk minimization: A causal approach, 2021.

J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. Bach. Supervised dictionary learning. *Advances in neural information processing systems*, 21, 2008.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pp. 689–696, 2009.

J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):791–804, 2011.

B. Malézieux, T. Moreau, and M. Kowalski. Dictionary and prior learning with unrolled algorithms for unsupervised inverse problems. *ICLR*, 2022.

G. Marcus, E. Davis, and S. Aaronson. A very preliminary analysis of dall-e 2. *arXiv preprint arXiv:2204.13807*, 2022.

A. Maurer, M. Pontil, and B. Romera-Paredes. Sparse coding for multitask and transfer learning. ICML'13, 2013.

A. Maurer, M. Pontil, and B. Romera-Paredes. The benefit of multitask representation learning. *J. Mach. Learn. Res.*, 2016.

T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur. Recurrent neural network based language model. ISCA, 2010.

Đ. Miladinović, M. W. Gondal, B. Schölkopf, J. M. Buhmann, and S. Bauer. Disentangled state space representations. *arXiv preprint arXiv:1906.03255*, 2019.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. MIT Press, 2018.

M. L. Montero, C. JH Ludwig, R. P. Costa, G. Malhotra, and J. Bowers. The role of disentanglement in generalisation. In *International Conference on Learning Representations*, 2021.

G. Obozinski, B. Taskar, and M. Jordan. Multi-task feature selection. *Statistics Department, UC Berkeley, Tech. Rep*, 2(2.2):2, 2006.

A. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *Advances in Neural Information Processing Systems*, 2018.

J. Pearl. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, 2019.

F. Pedregosa. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pp. 737–746. PMLR, 2016.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine. Meta-learning with implicit gradients. *Advances in Neural Information Processing Systems*, 32, 2019.

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1):1–38, 2014.

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.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE - Advances in Machine Learning and Deep Neural Networks*, 2021.

B. Schölkopf. Causality for machine learning, 2019.

J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30, 2017.

A. Taleb and C. Jutten. Source separation in post-nonlinear mixtures. *IEEE Transactions on Signal Processing*, 1999.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

S. van Steenkiste, F. Locatello, J. Schmidhuber, and O. Bachem. Are disentangled representations helpful for abstract visual reasoning? In *Advances in Neural Information Processing Systems*, 2019.

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

J. Von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 2008.

M. Wortsman, G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, R. G. Lopes, H. Hajishirzi, A. Farhadi, H. Namkoong, and L. Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7959–7971, June 2022.

S. Wright and J. Nocedal. Numerical optimization. *Springer Science*, 35(67-68):7, 1999.

W. Yao, Y. Sun, A. Ho, C. Sun, and K. Zhang. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations*, 2022.

H. Zhang, Y.-F. Zhang, W. Liu, A. Weller, B. Schölkopf, and E. Xing. Towards principled disentanglement for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Y. Zheng, I. Ng, and K. Zhang. On the identifiability of nonlinear ICA with unconditional priors. In *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*, 2022.

CONTENTS

1 Introduction **1**

2 Synergies between Disentanglement and Sparsity **2**

2.1 Disentanglement and sparse base-predictors for improved generalization 2

2.2 Disentanglement via sparse multitask learning 4

2.2.1 Identifiability analysis 4

2.2.2 Tractable bilevel optimization problems for sparse multitask learning . . . 6

2.3 A Meta-Learning formulation 6

3 Related Work **7**

4 Experiments **8**

4.1 Semi-real experiments on 3D Shapes 8

4.2 Few-shot learning experiments 9

5 Conclusion **9**

A Proofs of Section 2.1 **16**

B Identifiability Theory **17**

B.1 Proof of Theorem 1 19

B.2 Discussion of assumptions 20

C Optimization details **20**

C.1 Group Lasso SVM Dual 20

D Experimental details **22**

D.1 Disentangled representation coupled with sparsity regularization improves generalization 22

D.2 Semi-Real Experiments on 3D Shapes 23

D.2.1 Dataset generation 23

D.2.2 Metrics 24

D.2.3 Architecture, inner solver & hyperparameters 25

D.3 Meta-learning experiments 25

A PROOFS OF SECTION 2.1

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, $\forall \boldsymbol{x} \in \mathcal{X}$, $\hat{\boldsymbol{W}}_n^{(\hat{\boldsymbol{\theta}})}\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}) = \hat{\boldsymbol{W}}_n^{(\boldsymbol{\theta})}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})$.*

Proof. By definition of $\hat{\boldsymbol{W}}^{(\hat{\boldsymbol{\theta}})}$, we have that, for all $\hat{\boldsymbol{W}} \in \mathbb{R}^{k\times m}$,

$$\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y; \hat{\boldsymbol{W}}^{(\hat{\boldsymbol{\theta}})}\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})) \geq \sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y; \hat{\boldsymbol{W}}\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})) \tag{9}$$

$$\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y; \hat{\boldsymbol{W}}^{(\hat{\boldsymbol{\theta}})}\boldsymbol{L}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})) \geq \sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y; \hat{\boldsymbol{W}}\boldsymbol{L}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}))\,. \tag{10}$$

Because $\mathbb{R}^{k\times m}\boldsymbol{L} = \mathbb{R}^{k\times m}$, we have that, for all $\hat{\boldsymbol{W}} \in \mathbb{R}^{k\times m}$,

$$\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y; \hat{\boldsymbol{W}}^{(\hat{\boldsymbol{\theta}})}\boldsymbol{L}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})) \geq \sum_{(\boldsymbol{x},y)\in\mathcal{D}} \log p(y; \hat{\boldsymbol{W}}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}))\,, \tag{11}$$

which is to say that $\hat{\boldsymbol{W}}^{(\boldsymbol{\theta})} = \hat{\boldsymbol{W}}^{(\hat{\boldsymbol{\theta}})}\boldsymbol{L}$, or put differently, $\hat{\boldsymbol{W}}^{(\hat{\boldsymbol{\theta}})} = \hat{\boldsymbol{W}}^{(\boldsymbol{\theta})}\boldsymbol{L}^{-1}$. It implies

$$\hat{\boldsymbol{W}}^{(\hat{\boldsymbol{\theta}})}\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}) = \hat{\boldsymbol{W}}^{(\boldsymbol{\theta})}\boldsymbol{L}^{-1}\boldsymbol{L}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \hat{\boldsymbol{W}}^{(\boldsymbol{\theta})}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})\,, \tag{12}$$

which is what we wanted to show. □

Proposition 2 (Population MLE for Linearly Entangled Representations)**.** *Let $\hat{\boldsymbol{W}}_\infty^{(\hat{\boldsymbol{\theta}})}$ 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, $\hat{\boldsymbol{W}}_\infty^{(\hat{\boldsymbol{\theta}})} = \boldsymbol{W}\boldsymbol{L}^{-1}$.*

Proof. By definition of $\hat{\boldsymbol{W}}_\infty^{(\hat{\boldsymbol{\theta}})}$, we have that, for all $\tilde{\boldsymbol{W}} \in \mathbb{R}^{k\times m}$,

$$\mathbb{E}_{p(\boldsymbol{x},y)} \log p(y; \hat{\boldsymbol{W}}_\infty^{(\hat{\boldsymbol{\theta}})}\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})) \geq \mathbb{E}_{p(\boldsymbol{x},y)} \log p(y; \tilde{\boldsymbol{W}}\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})) \tag{13}$$

$$\mathbb{E}_{p(\boldsymbol{x},y)} \log p(y; \hat{\boldsymbol{W}}_\infty^{(\hat{\boldsymbol{\theta}})}\boldsymbol{L}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})) \geq \mathbb{E}_{p(\boldsymbol{x},y)} \log p(y; \tilde{\boldsymbol{W}}\boldsymbol{L}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}))\,. \tag{14}$$

In particular, the inequality holds for $\tilde{\boldsymbol{W}} := \boldsymbol{W}\boldsymbol{L}^{-1}$, which yields

$$\mathbb{E}_{p(\boldsymbol{x},y)} \log p(y; \hat{\boldsymbol{W}}_\infty^{(\hat{\boldsymbol{\theta}})}\boldsymbol{L}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})) \geq \mathbb{E}_{p(\boldsymbol{x},y)} \log p(y; \boldsymbol{W}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})) \tag{15}$$

$$0 \geq \mathbb{E}_{p(\boldsymbol{x},y)} \left[\log p(y; \boldsymbol{W}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})) - \log p(y; \hat{\boldsymbol{W}}_\infty^{(\hat{\boldsymbol{\theta}})}\boldsymbol{L}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}))\right] \tag{16}$$

$$0 \geq \mathbb{E}_{p(\boldsymbol{x})}\mathrm{KL}(p(y; \boldsymbol{W}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})) \mid\mid p(y; \hat{\boldsymbol{W}}_\infty^{(\hat{\boldsymbol{\theta}})}\boldsymbol{L}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})))\,. \tag{17}$$

Since the KL is always non-negative, we have that,

$$\mathbb{E}_{p(\boldsymbol{x})}\mathrm{KL}(p(y; \boldsymbol{W}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})) \mid\mid p(y; \hat{\boldsymbol{W}}_\infty^{(\hat{\boldsymbol{\theta}})}\boldsymbol{L}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}))) = 0\,, \tag{18}$$

which in turn implies

$$\mathbb{E}_{p(\boldsymbol{x},y)} \log p(y; \hat{\boldsymbol{W}}_\infty^{(\hat{\boldsymbol{\theta}})}\boldsymbol{L}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})) = \mathbb{E}_{p(\boldsymbol{x},y)} \log p(y; \boldsymbol{W}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})) \tag{19}$$

$$\mathbb{E}_{p(\boldsymbol{x},y)} \log p(y; \hat{\boldsymbol{W}}_\infty^{(\hat{\boldsymbol{\theta}})}\boldsymbol{L}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})) = \mathbb{E}_{p(\boldsymbol{x},y)} \log p(y; \boldsymbol{W}\boldsymbol{L}^{-1}\boldsymbol{L}\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})) \tag{20}$$

$$\mathbb{E}_{p(\boldsymbol{x},y)} \log p(y; \hat{\boldsymbol{W}}_\infty^{(\hat{\boldsymbol{\theta}})}\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})) = \mathbb{E}_{p(\boldsymbol{x},y)} \log p(y; \boldsymbol{W}\boldsymbol{L}^{-1}\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})) \tag{21}$$

$$\tag{22}$$

Since the solution to the population MLE from Proposition 2 is assumed to be unique, this equality holds if and only if $\hat{\boldsymbol{W}}_\infty^{(\hat{\boldsymbol{\theta}})} = \boldsymbol{W}\boldsymbol{L}^{-1}$. □

B IDENTIFIABILITY THEORY

The following lemma will be important for proving Theorem 3. The argument is taken from Lachapelle et al. (2022).

Lemma 1 (Sparsity pattern of an invertible matrix contains a permutation)**.** *Let $\boldsymbol{L} \in \mathbb{R}^{m\times m}$ be an invertible matrix. Then, there exists a permutation σ such that $\boldsymbol{L}_{i,\sigma(i)} \neq 0$ for all i.*

Proof. Since the matrix $\boldsymbol{L}$ is invertible, its determinant is non-zero, i.e.

$$\det(\boldsymbol{L}) := \sum_{\sigma\in\mathfrak{S}_m} \text{sign}(\sigma) \prod_{i=1}^{m} \boldsymbol{L}_{i,\sigma(i)} \neq 0\,, \tag{23}$$

where $\mathfrak{S}_m$ is the set of m-permutations. This equation implies that at least one term of the sum is non-zero, meaning there exists $\sigma \in \mathfrak{S}_m$ such that for all $i \in [m]$, $\boldsymbol{L}_{i,\sigma(i)} \neq 0$. □

For all $\boldsymbol{W} \in \mathcal{W}$, we are going to denote by $\hat{\boldsymbol{W}}^{(\boldsymbol{W})}$ some estimator of $\boldsymbol{W}$. The following result provides conditions under which if $\hat{\boldsymbol{W}}^{(\boldsymbol{W})}$ allows a perfect fit of the ground-truth distribution $p(y \mid \boldsymbol{x}, \boldsymbol{W})$, then the representation $\boldsymbol{f_\theta}$ and the parameter $\boldsymbol{W}$ are identified up to an invertible linear transformation. Many works have showed similar results in various context (Hyvarinen & Morioka, 2016; Khemakhem et al., 2020a; Roeder et al., 2021; Ahuja et al., 2022c). We reuse some of their proof techniques.

Theorem 2 (Linear identifiability)**.** *Let $\hat{\boldsymbol{W}}^{(\cdot)} : \mathcal{W} \to \mathbb{R}^{k\times m}$. Suppose Assumptions 2 to 5 hold and that, for all $\boldsymbol{W} \in \mathcal{W}$, $\boldsymbol{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$, the following holds*

$$p(y; \hat{\boldsymbol{W}}^{(\boldsymbol{W})} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})) = p(y; \boldsymbol{W}\boldsymbol{f_\theta}(\boldsymbol{x}))\ . \tag{24}$$

Then, there exists an invertible matrix $\boldsymbol{L} \in \mathbb{R}^{m\times m}$ such that, for all $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{f_\theta}(\boldsymbol{x}) = \boldsymbol{L}\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$ and such that, for all $\boldsymbol{W} \in \mathcal{W}$, $\hat{\boldsymbol{W}}^{(\boldsymbol{W})} = \boldsymbol{W}\boldsymbol{L}$

Proof. By Assumption 3, Equation (25) implies that $\boldsymbol{W}\boldsymbol{f_\theta}(\boldsymbol{x}) = \hat{\boldsymbol{W}}^{(\boldsymbol{W})}\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$. Assumption 5 ensures that we can construct an invertible matrix $\boldsymbol{U} := \begin{bmatrix} \boldsymbol{W}^{(1)}_{i_1,:} \\ \vdots \\ \boldsymbol{W}^{(d_z)}_{i_{d_z},:} \end{bmatrix}$. Construct analogously $\hat{\boldsymbol{U}} := \begin{bmatrix} \hat{\boldsymbol{W}}^{(\boldsymbol{W}^{(1)})}_{i_1,:} \\ \vdots \\ \hat{\boldsymbol{W}}^{(\boldsymbol{W}^{(d_z)})}_{i_{d_z},:} \end{bmatrix}$. This allows us to write $\boldsymbol{U}\boldsymbol{f_\theta}(\boldsymbol{x}) = \hat{\boldsymbol{U}}\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$. Left-multiplying by $\boldsymbol{U}^{-1}$ on both sides yields $\boldsymbol{f_\theta}(\boldsymbol{x}) = \boldsymbol{L}\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$, where $\boldsymbol{L} := \boldsymbol{U}^{-1}\hat{\boldsymbol{U}}$. Using the invertible matrix $\boldsymbol{F}$ from Assumption 4, we can thus write $\boldsymbol{F} = \boldsymbol{L}\hat{\boldsymbol{F}}$ where we defined $\hat{\boldsymbol{F}} := [\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}^{(1)}), \cdots, \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}^{(d_z)})]$. Since $\boldsymbol{F}$ is invertible, so are $\boldsymbol{L}$ and $\hat{\boldsymbol{F}}$.

By substituting $\boldsymbol{F} = \boldsymbol{L}\hat{\boldsymbol{F}}$ in $\boldsymbol{W}\boldsymbol{F} = \hat{\boldsymbol{W}}^{(\boldsymbol{W})}\hat{\boldsymbol{F}}$, we obtain $\boldsymbol{W}\boldsymbol{L}\hat{\boldsymbol{F}} = \hat{\boldsymbol{W}}^{(\boldsymbol{W})}\hat{\boldsymbol{F}}$. By right-multiplying both sides by $\hat{\boldsymbol{F}}^{-1}$, we obtain $\boldsymbol{W}\boldsymbol{L} = \hat{\boldsymbol{W}}^{(\boldsymbol{W})}$. □

The following theorem is where most of the theoretical contribution of this work lies. Note that Theorem 1, from the main text, is a straightforward application of this result.

Theorem 3. *(Disentanglement via task sparsity) Let $\hat{\boldsymbol{W}}^{(\cdot)} : \mathcal{W} \to \mathbb{R}^{k\times m}$. Suppose Assumptions 3 to 7 hold and that, for all $\boldsymbol{W} \in \mathcal{W}$, $\boldsymbol{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$, the following holds*

$$p(y; \hat{\boldsymbol{W}}^{(\boldsymbol{W})} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})) = p(y; \boldsymbol{W}\boldsymbol{f_\theta}(\boldsymbol{x}))\ . \tag{25}$$

Morevover, assume that $\mathbb{E}\|\hat{\boldsymbol{W}}^{(\boldsymbol{W})}\|_{2,0} \leq \mathbb{E}\|\boldsymbol{W}\|_{2,0}$, where both expectations are taken w.r.t. $\mathbb{P}_{\boldsymbol{W}}$ and $\|\boldsymbol{W}\|_{2,0} := \sum_{j=1}^{m} \mathbb{1}(\boldsymbol{W}_{:,j} \neq \boldsymbol{0})$ with $\mathbb{1}(\cdot)$ the indicator function. Then, $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}$ is disentangled w.r.t. $\boldsymbol{f_\theta}$ (Definition 1).

Proof. First of all, by Assumptions 3 to 5, we can apply Theorem 2 to conclude that $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{L}\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$ and $\boldsymbol{W}\boldsymbol{L} = \hat{\boldsymbol{W}}^{(\boldsymbol{W})}$ for some invertible matrix $\boldsymbol{L}$.

We can thus write $\mathbb{E}\|\boldsymbol{W}\boldsymbol{L}\|_{2,0} \leq \mathbb{E}\|\boldsymbol{W}\|_{2,0}$.

We can write

$$\mathbb{E}\|\boldsymbol{W}\|_{2,0} = \mathbb{E}_{p(S)}\mathbb{E}[\sum_{j=1}^{m} \mathbb{1}(\boldsymbol{W}_{:,j} \neq \boldsymbol{0}) \mid S] \tag{26}$$

$$= \mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbb{E}[\mathbb{1}(\boldsymbol{W}_{:,j} \neq \boldsymbol{0}) \mid S] \tag{27}$$

$$= \mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbb{P}_{\boldsymbol{W}|S}[\boldsymbol{W}_{:,j} \neq \boldsymbol{0}] \tag{28}$$

$$= \mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbb{1}(j \in S)\,, \tag{29}$$

where the last step follows from the definition of S.

We now perform similar steps for $\mathbb{E}\|\boldsymbol{W}\boldsymbol{L}\|_{2,0}$:

$$\mathbb{E}\|\boldsymbol{W}L\|_{2,0} = \mathbb{E}_{p(S)}\mathbb{E}[\sum_{j=1}^{m} \mathbb{1}(\boldsymbol{W}\boldsymbol{L}_{:,j} \neq \boldsymbol{0}) \mid S] \tag{30}$$

$$= \mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbb{E}[\mathbb{1}(\boldsymbol{W}\boldsymbol{L}_{:,j} \neq \boldsymbol{0}) \mid S] \tag{31}$$

$$= \mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbb{P}_{\boldsymbol{W}|S}[\boldsymbol{W}\boldsymbol{L}_{:,j} \neq \boldsymbol{0}] \tag{32}$$

$$= \mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbb{P}_{\boldsymbol{W}|S}[\boldsymbol{W}_{:,S}\boldsymbol{L}_{S,j} \neq \boldsymbol{0}]\,. \tag{33}$$

Notice that

$$\mathbb{P}_{\boldsymbol{W}|S}[\boldsymbol{W}_{:,S}\boldsymbol{L}_{S,j} \neq \boldsymbol{0}] = 1 - \mathbb{P}_{\boldsymbol{W}|S}[\boldsymbol{W}_{:,S}\boldsymbol{L}_{S,j} = \boldsymbol{0}] \tag{34}$$

Let N_j be the support of $\boldsymbol{L}_{:,j}$, i.e. $N_j := \{i \in [m] \mid \boldsymbol{L}_{i,j} \neq 0\}$. When $S \cap N_j = \emptyset$, $\boldsymbol{L}_{S,j} = \boldsymbol{0}$ and thus $\mathbb{P}_{\boldsymbol{W}|S}[\boldsymbol{W}_{:,S}\boldsymbol{L}_{S,j} = \boldsymbol{0}] = 1$. When $S \cap N_j \neq \emptyset$, $\boldsymbol{L}_{S,j} \neq \boldsymbol{0}$, by Assumption 6 we have that $\mathbb{P}_{\boldsymbol{W}|S}[\boldsymbol{W}_{:,S}\boldsymbol{L}_{S,j} = \boldsymbol{0}] = 0$. Thus

$$\mathbb{P}_{\boldsymbol{W}|S}[\boldsymbol{W}_{:,S}\boldsymbol{L}_{S,j} \neq \boldsymbol{0}] = 1 - \mathbb{1}(S \cap N_j = \emptyset) \tag{35}$$

$$= \mathbb{1}(S \cap N_j \neq \emptyset)\,, \tag{36}$$

which allows us to write

$$\mathbb{E}\|\boldsymbol{W}\boldsymbol{L}\|_{2,0} = \mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbb{1}(S \cap N_j \neq \emptyset)\,. \tag{37}$$

We thus have that

$$\mathbb{E}\|\boldsymbol{W}\boldsymbol{L}\|_{2,0} \leq \mathbb{E}\|\boldsymbol{W}\|_{2,0} \tag{38}$$

$$\mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbb{1}(S \cap N_j \neq \emptyset) \leq \mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbb{1}(j \in S)\,. \tag{39}$$

Since $\boldsymbol{L}$ is invertible, by Lemma 1, there exists a permutation $\sigma : [m] \to [m]$ such that, for all $j \in [m]$, $\boldsymbol{L}_{j,\sigma(j)} \neq 0$. In other words, for all $j \in [m]$, $j \in N_{\sigma(j)}$. Of course we can permute the

terms of the l.h.s. of eq. (39), which yields

$$\mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbb{1}(S \cap N_{\sigma(j)} \neq \emptyset) \leq \mathbb{E}_{p(S)} \sum_{j=1}^{m} \mathbb{1}(j \in S) \tag{40}$$

$$\mathbb{E}_{p(S)} \sum_{j=1}^{m} \big(\mathbb{1}(S \cap N_{\sigma(j)} \neq \emptyset) - \mathbb{1}(j \in S)\big) \leq 0\,. \tag{41}$$

We notice that each term $\mathbb{1}(S \cap N_{\sigma(j)} \neq \emptyset) - \mathbb{1}(j \in S) \geq 0$ since whenever $j \in S$, we also have that $j \in S \cap N_{\sigma(j)}$ (recall $j \in N_{\sigma(j)}$). Thus, the l.h.s. of eq. (41) is a sum of non-negative terms which is itself non-positive. This means that every term in the sum is zero:

$$\forall S \in \mathcal{S},\ \forall j \in [m],\ \mathbb{1}(S \cap N_{\sigma(j)} \neq \emptyset) = \mathbb{1}(j \in S)\,. \tag{42}$$

Importantly,

$$\forall j \in [m],\ \forall S \in \mathcal{S},\ j \notin S \implies S \cap N_{\sigma(j)} = \emptyset\,, \tag{43}$$

and since $S \cap N_{\sigma(j)} = \emptyset \iff N_{\sigma(j)} \subseteq S^c$ we have that

$$\forall j \in [m],\ \forall S \in \mathcal{S},\ j \notin S \implies N_{\sigma(j)} \subseteq S^c \tag{44}$$

$$\forall j \in [m],\ N_{\sigma(j)} \subseteq \bigcap_{S \in \mathcal{S} | j \notin S} S^c\,. \tag{45}$$

By Assumption 7, we have that $\bigcup_{S \in \mathcal{S} | j \notin S} S = [m] \setminus \{j\}$. By taking the complement on both sides and using De Morgan's law, we get $\bigcap_{S \in \mathcal{S} | j \notin S} S^c = \{j\}$, which implies that $N_{\sigma(j)} = \{j\}$ by Equation (45). Thus, $\boldsymbol{L} = \boldsymbol{D}\boldsymbol{P}$ where $\boldsymbol{D}$ is an invertible diagonal matrix and $\boldsymbol{P}$ is a permutation matrix. □

B.1 Proof of Theorem 1

Before presenting Theorem 1 from the main text, we first present a variation of it where we constrain $\mathbb{E}\|\hat{\boldsymbol{W}}^{(\boldsymbol{W})}\|_{2,0}$ to be smaller than $\mathbb{E}\|\boldsymbol{W}\|_{2,0}$. We note that this is weaker than imposing $\|\hat{\boldsymbol{W}}^{(\boldsymbol{W})}\|_{2,0} \leq \|\boldsymbol{W}\|_{2,0}$ for all $\boldsymbol{W} \in \mathcal{W}$, as is the case in Problem (4) of Theorem 1.

Theorem 4 (Sparse multitask learning for disentanglement)**.** *Let $\hat{\boldsymbol{\theta}}$ be a minimizer of*

$$\begin{aligned} &\min_{\hat{\boldsymbol{\theta}}} \mathbb{E}_{\mathbb{P}_{\boldsymbol{W}}} \mathbb{E}_{p(\boldsymbol{x},y|\boldsymbol{W})} - \log p(y; \hat{\boldsymbol{W}}^{(\boldsymbol{W})} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})) \\ &\text{s.t.} \quad \forall\, \boldsymbol{W} \in \mathcal{W},\ \hat{\boldsymbol{W}}^{(\boldsymbol{W})} \in \arg\min_{\tilde{\boldsymbol{W}}} \mathbb{E}_{p(\boldsymbol{x},y|\boldsymbol{W})} - \log p(y; \tilde{\boldsymbol{W}} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})) \\ &\qquad \mathbb{E}_{\mathbb{P}_{\boldsymbol{W}}} \|\hat{\boldsymbol{W}}^{(\boldsymbol{W})}\|_{2,0} \leq \mathbb{E}_{\mathbb{P}_{\boldsymbol{W}}} \|\boldsymbol{W}\|_{2,0}\ . \end{aligned} \tag{46}$$

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

Proof. First, notice that

$$0 \leq \mathbb{E}_{\mathbb{P}_{\boldsymbol{W}}} \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{W})} \mathrm{KL}(p(y; \boldsymbol{W} \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})) \,||\, p(y; \hat{\boldsymbol{W}}^{(\boldsymbol{W})} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}))) \tag{47}$$

$$\mathbb{E}_{\mathbb{P}_{\boldsymbol{W}}} \mathbb{E}_{p(\boldsymbol{x},y|\boldsymbol{W})} - \log p(y; \boldsymbol{W} \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})) \leq \mathbb{E}_{\mathbb{P}_{\boldsymbol{W}}} \mathbb{E}_{p(\boldsymbol{x},y|\boldsymbol{W})} - \log p(y; \hat{\boldsymbol{W}}^{(\boldsymbol{W})} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}))\,. \tag{48}$$

For a fixed value of $\boldsymbol{x}$ and $\boldsymbol{W}$, it is well known that $\mathrm{KL}(p(y; \boldsymbol{W} \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})) \,||\, p(y; \hat{\boldsymbol{W}}^{(\boldsymbol{W})} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}))) = 0$ if and only if, for all $y \in \mathcal{Y}$, $p(y; \boldsymbol{W} \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x})) = p(y; \hat{\boldsymbol{W}}^{(\boldsymbol{W})} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}))$. By Assumption 3, this is equivalent to $\boldsymbol{W} \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \hat{\boldsymbol{W}}^{(\boldsymbol{W})} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$. Thus, for the equality to hold in eq. (47), we need $\boldsymbol{W} \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \hat{\boldsymbol{W}}^{(\boldsymbol{W})} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$ everywhere. Of course, the global minimum can be achieved by respecting $\mathbb{E}_{\mathbb{P}_{\boldsymbol{W}}} \|\hat{\boldsymbol{W}}^{(\boldsymbol{W})}\|_{2,0} \leq \mathbb{E}_{\mathbb{P}_{\boldsymbol{W}}} \|\boldsymbol{W}\|_{2,0}$, simply by setting $\hat{\boldsymbol{\theta}} := \boldsymbol{\theta}$ and $\hat{\boldsymbol{W}}^{(\boldsymbol{W})} := \boldsymbol{W}$.

The above implies that if $\hat{\boldsymbol{\theta}}$ is some minimizer of Problem (46), we must have that $\boldsymbol{W} \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \hat{\boldsymbol{W}}^{(\boldsymbol{W})} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$ everywhere and $\mathbb{E}_{\mathbb{P}_{\boldsymbol{W}}} ||\hat{\boldsymbol{W}}^{(\boldsymbol{W})}||_0 \leq \mathbb{E}_{\mathbb{P}_{\boldsymbol{W}}} ||\boldsymbol{W}||_0$. Thus, Theorem 3 implies the desired conclusion. □

Based on Theorem 4, we can slightly adjust the argument to prove Theorem 1 from the main text.

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

$$\begin{aligned} &\min_{\hat{\boldsymbol{\theta}}} \mathbb{E}_{\mathbb{P}_{\boldsymbol{W}}} \mathbb{E}_{p(\boldsymbol{x},y|\boldsymbol{W})} - \log p(y; \hat{\boldsymbol{W}}^{(\boldsymbol{W})} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})) \\ &\text{s.t.} \quad \forall\, \boldsymbol{W} \in \mathcal{W},\ \hat{\boldsymbol{W}}^{(\boldsymbol{W})} \in \underset{\substack{\tilde{\boldsymbol{W}} \text{ s.t.} \\ ||\tilde{\boldsymbol{W}}||_{2,0} \leq ||\boldsymbol{W}||_{2,0}}}{\arg\min} \mathbb{E}_{p(\boldsymbol{x},y|\boldsymbol{W})} - \log p(y; \tilde{\boldsymbol{W}} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})) \ . \end{aligned} \tag{4}$$

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

Proof. The first part of the argument in the proof of Theorem 4 applies here as well, meaning: for the equality to hold in eq. (47), we need $\boldsymbol{W} \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \hat{\boldsymbol{W}}^{(\boldsymbol{W})} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$ everywhere. This global minimum can be achieved by respecting $\|\hat{\boldsymbol{W}}^{(\boldsymbol{W})}\|_{2,0} \leq \|\boldsymbol{W}\|_{2,0}$ for all $\boldsymbol{W} \in \mathcal{W}$ simply by setting $\hat{\boldsymbol{\theta}} := \boldsymbol{\theta}$ and $\hat{\boldsymbol{W}}^{(\boldsymbol{W})} := \boldsymbol{W}$.

This means that if $\hat{\boldsymbol{\theta}}$ is some minimizer of Problem (4), we must have that $\boldsymbol{W} \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \hat{\boldsymbol{W}}^{(\boldsymbol{W})} \boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$ holds everywhere and that, for all $\boldsymbol{W} \in \mathcal{W}$, $\|\hat{\boldsymbol{W}}^{(\boldsymbol{W})}\|_{2,0} \leq \|\boldsymbol{W}\|_{2,0}$. Of course, this means $\mathbb{E}_{\mathbb{P}_{\boldsymbol{W}}} ||\hat{\boldsymbol{W}}^{(\boldsymbol{W})}||_0 \leq \mathbb{E}_{\mathbb{P}_{\boldsymbol{W}}} ||\boldsymbol{W}||_0$, which allows us to apply Theorem 3 to obtain the desired conclusion. □

B.2 Discussion of assumptions

Assumption 6. We recall the example provided in Section 2.2.1: We illustrate Assumption 6 in the simpler case where $k = 1$. The assumption holds for example when the distribution of $\boldsymbol{W}_{1,S} \mid S$ has a density w.r.t. the Lebesgue measure on $\mathbb{R}^{|S|}$, which is true for example when $\boldsymbol{W}_{1,S} \mid S \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ and the covariance matrix $\boldsymbol{\Sigma}$ is full rank. However, if $\boldsymbol{\Sigma}$ is not full rank, the probability distribution of $\boldsymbol{W}_{1,S} \mid S$ concentrates its mass on a proper linear subspace $V \subsetneq \mathbb{R}^{|S|}$, and, thus, we can find a nonzero vector $\boldsymbol{a} \in V^{\perp}$. This means $\mathbb{P}_{\boldsymbol{W}|S}\{\boldsymbol{W} : \boldsymbol{W}_{1,S}\boldsymbol{a} = 0\} = 1$, which violates Assumption 6.

Another important counter-example is when $\mathbb{P}_{\boldsymbol{W}|S}$ concentrates some of its mass on a point $\boldsymbol{W}^{(0)}$, i.e. $\mathbb{P}_{\boldsymbol{W}|S}\{\boldsymbol{W}^{(0)}\} > 0$. In this case, one can simply pick $\boldsymbol{a} \in \text{span}(\boldsymbol{W}_{1,S}^{(0)})^{\perp}$ to see that $\mathbb{P}_{\boldsymbol{W}|S}\{\boldsymbol{W} : \boldsymbol{W}_{1,S}\boldsymbol{a} = 0\} \geq \mathbb{P}_{\boldsymbol{W}|S}\{\boldsymbol{W}^{(0)}\} > 0$.

Interestingly, there are distributions over $\boldsymbol{W}_{1,S} \mid S$ that do not have a density w.r.t. the Lebesgue measure, but still satisfy Assumption 6. As a simple example, consider the case where $\boldsymbol{W}_{1,S} \mid S$ puts uniform mass over a $(|S| - 1)$-dimensional sphere embedded in $\mathbb{R}^{|S|}$ and centered at zero. In that case, for all $\boldsymbol{a} \in \mathbb{R}^{|S|}$ not zero, the intersection of $\text{span}\{\boldsymbol{a}\}^{\perp}$, which is $(|S| - 1)$-dimensional, with the $(|S| - 1)$-dimensional sphere is $(|S| - 2)$-dimensional and thus has probability zero of occurring. One can certainly construct more exotic examples of measures satisfying Assumption 6 that concentrate mass on lower dimensional manifold.

Assumption 7. We provide an additional example. Consider the case $\mathcal{S} = \{\{1, 2, 3\}, \{2, 4\}, \{1, 2\}, \{3\}, \{1, 4\}\}$ with $m = 4$. The above shows that $\mathcal{S}$ satisfies Assumption 7.

$$\{2, 4\} \cup \{3\} = \{2, 3, 4\} \tag{49}$$
$$\{3\} \cup \{1, 4\} = \{1, 3, 4\} \tag{50}$$
$$\{2, 4\} \cup \{1, 4\} = \{1, 2, 4\} \tag{51}$$
$$\{1, 2, 3\} = \{1, 2, 3\}\,. \tag{52}$$

C Optimization details

C.1 Group Lasso SVM Dual

Notation. The Fenchel conjugate of a function $h : \mathbb{R}^d \to \mathbb{R}$ is written h^* and is defined for any $y \in \mathbb{R}^d$, by $h^*(y) = \sup_{x \in \mathbb{R}^d} \langle x, y \rangle - h(x)$.

Definition 2. (Primal Group Lasso Soft-Margin Multiclass SVM.) *The primal problem of the group Lasso soft-margin multiclass SVM is defined as*

$$\min_{\boldsymbol{W}\in\mathbb{R}^{K\times m}} \mathcal{L}_{\text{in}}(\boldsymbol{W};\boldsymbol{F},\boldsymbol{Y}) := \sum_{i=1}^{n} \max_{l\in[k]} (1+(\boldsymbol{W}_{\boldsymbol{y}_i,:}-\boldsymbol{W}_{l,:})\boldsymbol{F}_{i,:}-\boldsymbol{Y}_{il}) + \lambda_1\|\boldsymbol{W}\|_{2,1} + \tfrac{\lambda_2}{2}\|\boldsymbol{W}\|^2 \tag{53}$$

The primal objective 53 can be hard to minimize with modern solvers. Moreover in few-shot learning applications, the number of features m is usually much larger than the number of samples n (in Lee et al. 2019, $m = 1.6\cdot 10^4$ and $n \leq 25$), hence we solve the dual of Problem (53).

Proof of Proposition 3. Let $g : \boldsymbol{u} \mapsto \lambda_1\|\boldsymbol{u}\| + \frac{\lambda_2}{2}\|\boldsymbol{u}\|^2$. Proof of Proposition 3 is composed of the following lemmas.

Lemma 2. i) *The dual of Problem (53) is*

$$\begin{aligned} &\min_{\boldsymbol{\Lambda}\in\mathbb{R}^{n\times k}} \sum_{j=1}^{m} g^*((\boldsymbol{Y}-\boldsymbol{\Lambda})^\top \boldsymbol{F}_{:j}) + \langle \boldsymbol{Y},\boldsymbol{\Lambda}\rangle \\ &\text{s.t.} \quad \forall i\in[n],\ \sum_{l=1}^{k}\boldsymbol{\Lambda}_{il} = 1 \ , \quad \forall i\in[n], l\in[k],\ \boldsymbol{\Lambda}_{il}\geq 0 \ , \end{aligned} \tag{54}$$

where g^* is the Fenchel conjugate of the function g.

ii) *The Fenchel conjugate of the function g writes*

$$\forall \boldsymbol{v}\in\mathbb{R}^K,\ g^*(\boldsymbol{v}) = \frac{1}{\lambda_2}\|\text{BST}(\boldsymbol{v},\lambda_1)\|^2 \ . \tag{55}$$

Lemmas 4 *i)* and 4 *ii)* yields Proposition 3.

Proof of Lemma 4 i). The Lagrangian of Problem (53) writes:

$$\mathcal{L}(\boldsymbol{W},\boldsymbol{\xi},\boldsymbol{\Lambda}) = \sum_{j=1}^{m} g(\boldsymbol{W}_{:j}) + \sum_i \boldsymbol{\xi}_i + \sum_{i=1}^{n}\sum_{l=1}^{k}(1-\boldsymbol{\xi}_i - \boldsymbol{W}_{\boldsymbol{y}_i:}\cdot\boldsymbol{F}_{i:} + \boldsymbol{W}_{l:}\cdot\boldsymbol{F}_{i:} - \boldsymbol{Y}_{il})\boldsymbol{\Lambda}_{i,l} \ . \tag{56}$$

$\partial_{\boldsymbol{\xi}}\mathcal{L}(\boldsymbol{W},\boldsymbol{\xi},\boldsymbol{\Lambda}) = 0$ yields $\forall i\in[n], \sum_{l=1}^{k}\boldsymbol{\Lambda}_{i,l} = 1$. Then the Lagrangian rewrites

$$\begin{aligned} \min_{\boldsymbol{W}}\min_{\boldsymbol{\xi}} \mathcal{L}(\boldsymbol{W},\boldsymbol{\xi},\boldsymbol{\Lambda}) &= \min_{\boldsymbol{W},\boldsymbol{\xi}} \sum_{j=1}^{m} g(\boldsymbol{W}_{:j}) + \sum_{i=1}^{n}\boldsymbol{\xi}_i + \sum_{i=1}^{n}\sum_{l=1}^{k}(-\boldsymbol{\xi}_i - \boldsymbol{W}_{\boldsymbol{y}_i:}\cdot\boldsymbol{F}_{i:} + \boldsymbol{W}_{l:}\cdot\boldsymbol{F}_{i:} - \boldsymbol{Y}_{i,l})\boldsymbol{\Lambda}_{i,l} \\ &= \underbrace{\sum_{j=1}^{m}\min_{\boldsymbol{W}_{:j}} g(\boldsymbol{W}_{:j}) - \underbrace{\sum_{i=1}^{n}\sum_{l=1}^{k}(\boldsymbol{F}_{i:}\boldsymbol{Y}_{i,l} - \boldsymbol{F}_{i:}\boldsymbol{\Lambda}_{i,l})\boldsymbol{W}_{l:}}_{=\langle(\boldsymbol{Y}-\boldsymbol{\Lambda})^\top\boldsymbol{F}_{:j},\boldsymbol{W}_{:j}\rangle}}_{=-g^*((\boldsymbol{Y}-\boldsymbol{\Lambda})^\top\boldsymbol{F}_{:j})} - \sum_{i=1}^{n}\sum_{l=1}^{k}\boldsymbol{Y}_{i,l}\boldsymbol{\Lambda}_{i,l} \ . \end{aligned}$$

Then the dual problem writes:

$$\min_{\boldsymbol{\Lambda}\in\mathbb{R}^{n\times k}} \sum_{j=1}^{m} g^*\left((\boldsymbol{Y}-\boldsymbol{\Lambda})^\top\boldsymbol{F}_{:j}\right) + \langle\boldsymbol{Y},\boldsymbol{\Lambda}\rangle \tag{57}$$

$$\text{s. t.} \quad \forall i\in[n] \quad \sum_{l=1}^{k}\boldsymbol{\Lambda}_{il} = 1 \ , \forall i\in[n], l\in[k], \quad \boldsymbol{\Lambda}_{il}\geq 0 \ . \tag{58}$$

□

Proof of Lemma 4 ii). Let $h : \boldsymbol{u} \mapsto \|\boldsymbol{u}\|_2 + \frac{\kappa}{2}\|\boldsymbol{u}\|^2$. The proof of Lemma 4 *i)* is done using the following steps.

Lemma 3. i) $h^*(\boldsymbol{v}) = \frac{1}{2\kappa}\|\boldsymbol{v}\|_2^2 - \left(\frac{\kappa}{2}\|\cdot\|_2^2 \square \|\cdot\|_2\right)(\boldsymbol{v}/\kappa)$.

ii) $\left(\frac{\kappa}{2}\|\cdot\|_2^2 \square \|\cdot\|_2\right)(\boldsymbol{v}) = \frac{\kappa}{2}\|\boldsymbol{v}\|_2^2 - \frac{1}{2\kappa}\|\mathrm{BST}(\kappa\boldsymbol{v}, 1)\|^2$.

Proof of Lemma 4 i). With $\kappa = \lambda_2/\lambda_1$, the Fenchel transform of $h : \boldsymbol{w} \mapsto \|\boldsymbol{w}\|_2 + \kappa\|\boldsymbol{w}\|^2$.

$$
\begin{aligned}
h(\boldsymbol{u}) &= \|\boldsymbol{u}\|_2 + \tfrac{\kappa}{2}\|\boldsymbol{u}\|_2^2 \\
h^*(\boldsymbol{v}) &= \sup_{\boldsymbol{w}} \left(\boldsymbol{v}^\top \boldsymbol{w} - \|\boldsymbol{w}\|_2 - \tfrac{\kappa}{2}\|\boldsymbol{w}\|_2^2\right) \\
&= \tfrac{1}{2\kappa}\|\boldsymbol{v}\|_2^2 + \sup_{\boldsymbol{w}} \left(-\tfrac{\kappa}{2}\|\boldsymbol{w} - \boldsymbol{v}/\kappa\|_2^2 - \|\boldsymbol{w}\|_2\right) \\
&= \tfrac{1}{2\kappa}\|\boldsymbol{v}\|_2^2 - \inf_{\boldsymbol{w}} \left(\tfrac{\kappa}{2}\|\boldsymbol{w} - \boldsymbol{v}/\kappa\|_2^2 + \|\boldsymbol{w}\|_2\right) \\
&= \tfrac{1}{2\kappa}\|\boldsymbol{v}\|_2^2 - (\tfrac{\kappa}{2}\|\cdot\|_2^2 \square \|\cdot\|_2)(\boldsymbol{v}/\kappa) \ .
\end{aligned}
$$

□

Proof of Lemma 4 ii).

$$
\begin{aligned}
(\tfrac{\kappa}{2}\|\cdot\|_2^2 \square \|\cdot\|_2)(\boldsymbol{v}) &= (\tfrac{\kappa}{2}\|\cdot\|_2^2 \square \|\cdot\|_2)^{**}(\boldsymbol{v}) \\
&= (\tfrac{1}{2\kappa}\|\cdot\|_2^2 + \iota_{\mathcal{B}_2})^*(\boldsymbol{v}) \\
&= \sup_{\|\boldsymbol{w}\|_2 \leq 1} \left(\boldsymbol{v}^\top \boldsymbol{w} - \tfrac{1}{2\kappa}\|\boldsymbol{w}\|_2^2\right) \\
&= \tfrac{\kappa}{2}\|\boldsymbol{v}\|^2 + \sup_{\|\boldsymbol{w}\|_2 \leq 1} -\tfrac{1}{2\kappa}\|\kappa\boldsymbol{v} - \boldsymbol{w}\|_2^2 \\
&= \tfrac{\kappa}{2}\|\boldsymbol{v}\|^2 - \tfrac{1}{2\kappa}\|\mathrm{BST}(\kappa\boldsymbol{v}, 1)\|_2^2 \ .
\end{aligned}
$$

□

$$
\begin{aligned}
g^*(\boldsymbol{u}) &= \lambda_1 h^*(\boldsymbol{u}/\lambda_1) \\
&= \frac{\lambda_1}{2\kappa}\|\mathrm{BST}(\boldsymbol{u}/\lambda_1, 1)\|^2 \\
&= \frac{\lambda_1^2}{2\lambda_2}\|\mathrm{BST}(\boldsymbol{u}/\lambda_1, 1)\|^2 \\
&= \frac{1}{\lambda_2}\|\mathrm{BST}(\boldsymbol{u}, \lambda_1)\|^2 \ .
\end{aligned}
$$

□

□

D Experimental details

D.1 Disentangled representation coupled with sparsity regularization improves generalization

We consider the following data generating process: We sample the ground-truth features $\boldsymbol{f_\theta}(\boldsymbol{x})$ from a Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times m}$ and $\boldsymbol{\Sigma}_{i,j} = 0.9^{|i-j|}$. Moreover, the labels are given by $y = \boldsymbol{w} \cdot \boldsymbol{f_\theta}(\boldsymbol{x}) + \epsilon$ where $\boldsymbol{w} \in \mathbb{R}^m$, $\epsilon \sim \mathcal{N}(0, 0.04)$ and $m = 100$. The ground-truth weight vector $\boldsymbol{w}$ is sampled once from $\mathcal{N}(0, I_{m \times m})$ and mask some of its components to zero: we vary the fraction of meaningful features (ℓ/m) from very sparse ($\ell/m = 5\%$) to less sparse ($\ell/m = 80\%$)

settings. For each case, we study the sample complexity by varying the number of training samples from 25 to 150, but evaluating the generalization performance on a larger test dataset (1000 samples). To generate the entangled representations, we multiply the true latent variables $\boldsymbol{f_\theta}(\boldsymbol{x})$ by a randomly sampled orthogonal matrix $\boldsymbol{L}$, i.e., $\boldsymbol{f_{\hat{\theta}}}(\boldsymbol{x}) := \boldsymbol{L}\boldsymbol{f_\theta}(\boldsymbol{x})$. For the disentangled representation, we simply consider the true latents, i.e. $\boldsymbol{f_{\hat{\theta}}}(\boldsymbol{x}) := \boldsymbol{f_\theta}(\boldsymbol{x})$. Note that in principle we could have considered an invertible matrix $\boldsymbol{L}$ that is not orthogonal for the linearly entangled representation and a component-wise rescaling for the disentangled representation. The advantage of not doing so and opting for our approach is that the conditioning number of the covariance matrix of $\boldsymbol{f_{\hat{\theta}}}(\boldsymbol{x})$ is the same for both the entangled and the disentangled, hence offering a fairer comparison.

For both the case of entangled and disentangled representation, we solve the regression problem with Lasso and Ridge regression, where the associated hyperparameters (regularization strength) were inferred using 5-fold cross validation on the input training dataset. Using both lasso and ridge regression would help us to show the effect of encouraging sparsity.

In Figure 1 for the sparsest case ($\ell/m = 5\%$), we observe that that Disentangled-Lasso approach has the best performance when we have less training samples, while the Entangled-Lasso approach performs the worst. As we increase the number of training samples, the performance of Entangled-Lasso approaches that of Disentangled-Lasso, however, learning under the Disentangled-Lasso approach is sample efficient. Disentangled-Lasso obtains R^2 greater than 0.5 with only 25 training samples, while other approaches obtain R^2 close to zero. Also, Disentagled-Lasso converges to the optimal R^2 using only 50 training samples, while Entangled-Lasso does the same with 150 samples samples.

Note that the improvement due to disentanglement does not happen for the case of ridge regression as expected and there is no of a difference between the methods Disentangled-Ridge and Entangled-Ridge because the L2 norm is invariant to orthogonal transformation. Also, having sparsity in the underlying task is important. Disentangled-Lasso shows the max improvement for the case of $\ell/m = 5\%$, with the gains reducing as we decrease the sparsity in the underlying task ($l/m = 80\%$).

D.2 Semi-Real Experiments on 3D Shapes

D.2.1 Dataset generation

Details on 3D Shapes. The 3D Shapes dataset (Burgess & Kim, 2018) contains synthetic images of colored shapes resting in a simple 3D scene. These images vary across 6 factors: Floor hue (10 values linearly spaced in [0, 1]); Wall hue (10 values linearly spaced in [0, 1]); Object hue (10 values linearly spaced in [0, 1]); Scale (8 values linearly spaced in [0, 1]); Shape (4 values in [0, 1, 2, 3]); and Orientation (15 values linearly spaced in [-30, 30]). These are the factors we aim to disentangle. We standardize them to have mean 0 and variance 1. We denote by $\mathcal{Z} \subset \mathbb{R}^6$, the set of all possible latent factor combinations. In our framework, this corresponds to the support of the ground-truth features $\boldsymbol{f_\theta}(\boldsymbol{x})$. We note that the points in $\mathcal{Z}$ are arranged in a grid-like fashion in $\mathbb{R}^6$.

Task generation. For all tasks t, the labelled dataset $\mathcal{D}_t = \{(\boldsymbol{x}^{(t,i)}), y^{(t,i)})\}_{i=1}^n$ is generated by first sampling the ground-truth latent variables $\boldsymbol{z}^{(t,i)} := \boldsymbol{f_\theta}(\boldsymbol{x}^{(t,i)})$ i.i.d. according to some distribution $p(\boldsymbol{z})$ over $\mathcal{Z}$, while the corresponding input is obtained doing $\boldsymbol{x}^{(t,i)} := \boldsymbol{f_\theta}^{-1}(\boldsymbol{z}^{(t,i)})$ ($\boldsymbol{f_\theta}$ is invertible in 3D Shapes). Then, a sparse weight vector $\boldsymbol{w}^{(t)}$ is sampled randomly by doing $\boldsymbol{w}^{(t)} := \bar{\boldsymbol{w}}^{(t)} \odot \boldsymbol{s}^{(t)}$, were $\odot$ is the Hadamard (component-wise) product, $\bar{\boldsymbol{w}}^{(t)} \sim \mathcal{N}(\boldsymbol{0}, I)$ and $\boldsymbol{s} \in \{0,1\}^6$ is a binary vector with independent components sampled from a Bernoulli distribution with ($p = 0.5$). Then, the labels are computedfor each example as $y^{(t,i)} := \boldsymbol{w}^{(t)} \cdot \boldsymbol{x}^{(t,i)} + \epsilon^{(t,i)}$, where $\epsilon^{(t,i)}$ is independent Gaussian noise. In every tasks, the dataset has size $n = 50$. New tasks are generated continuously as we train. Figures 2 and 4 explores various choices of $p(\boldsymbol{z})$, i.e. by varying the level of correlation between the latent variables and by varying the level of noise on the ground-truth latents.

Noise on latents. To make the dataset slightly more realistic, we get rid of the artificial grid-like structure of the latents by adding noise to it. This procedure transforms $\mathcal{Z}$ into a new support $\mathcal{Z}_\alpha$, where α is the noise level. Formally, $\mathcal{Z}_\alpha := \bigcup_{\boldsymbol{z} \in \mathcal{Z}} \{\boldsymbol{z} + \boldsymbol{u_z}\}$ where the $\boldsymbol{u}_z$ are i.i.d samples from

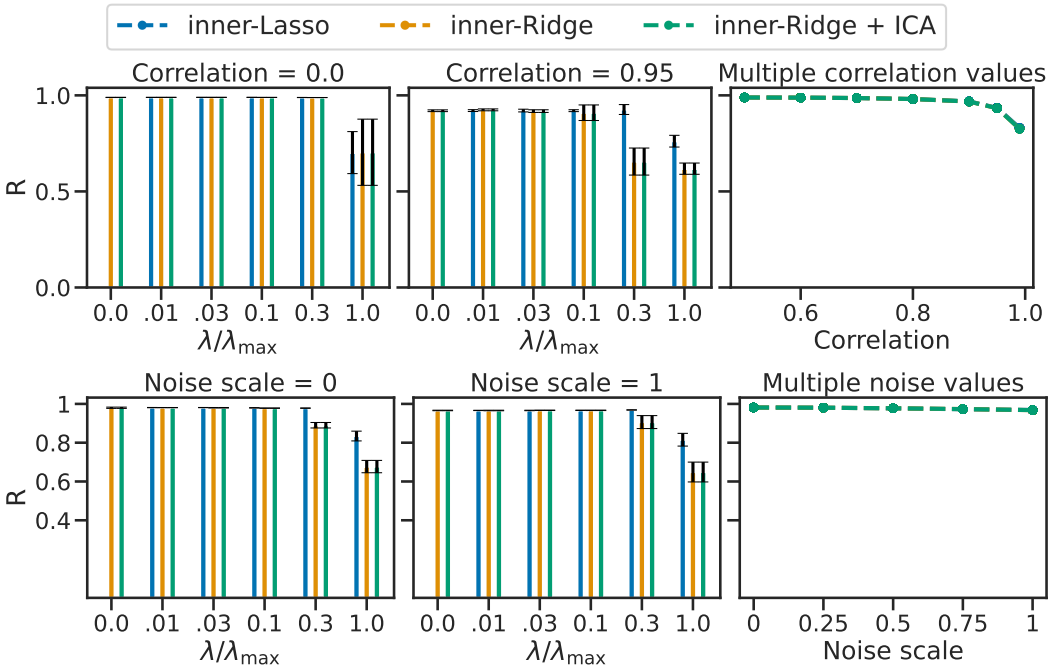

Figure 4: Prediction performance (R Score) for inner-Lasso, inner-Ridge and inner-Ridge combined with ICA as a function of the regularization parameter (left and middle). Varying level of correlation between latents (top) and noise on the latents (bottom). The right columns shows performance of the best hyperparameter for different values of correlation and noise levels.

the uniform over the hypercube

$$\left[-\alpha\frac{\Delta \boldsymbol{z}_1}{2}, \alpha\frac{\Delta \boldsymbol{z}_1}{2}\right] \times \left[-\alpha\frac{\Delta \boldsymbol{z}_2}{2}, \alpha\frac{\Delta \boldsymbol{z}_2}{2}\right] \times \ldots \times \left[-\alpha\frac{\Delta \boldsymbol{z}_6}{2}, \alpha\frac{\Delta \boldsymbol{z}_6}{2}\right] ,$$

where $\Delta \boldsymbol{z}_i$ denotes the gap between contiguous values of the factor $\boldsymbol{z}_i$. When $\alpha = 0$, no noise is added and the support $\mathcal{Z}$ is unchanged, i.e., $\mathcal{Z}_1 = \mathcal{Z}$. As long as $\alpha \in [0, 1]$, contiguous points in $\mathcal{Z}$ cannot be interchanged in $\mathcal{Z}_\alpha$. We also clarify that the ground-truth mapping $\boldsymbol{f}_{\boldsymbol{\theta}}$ is modified to $\boldsymbol{f}_{\boldsymbol{\theta},\alpha}$ consequently: for all $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{f}_{\boldsymbol{\theta},\alpha}(\boldsymbol{x}) := \boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}) + \boldsymbol{u}_{\boldsymbol{z}}$. We emphasize that the $\boldsymbol{u}_{\boldsymbol{z}}$ are sampled only once such that $\boldsymbol{f}_{\boldsymbol{\theta},\alpha}(\boldsymbol{x})$ is actually a deterministic mapping.

Varying correlations. To verify that our approach is robust to correlations in the latents, we construct $p(\boldsymbol{z})$ as follows: We consider a Gaussian density centered at $\mathbf{0}$ with covariance $\boldsymbol{\Sigma}_{i,j} := \rho + \mathbb{1}(i = j)\dot{(}1 - \rho)$. Then, we evaluate this density on the points of $\mathcal{Z}_\alpha$ and renormalize to have a well-defined probability distribution over $\mathcal{Z}_\alpha$. We denote by $p_{\alpha,\rho}(\boldsymbol{z})$ the distribution obtain by this construction.

In the top rows of Figures 2 and 4, the latents are sampled from $p_{\alpha=1,\rho}(\boldsymbol{z})$ and ρ varies between 0 and 0.99. In the bottom rows of Figures 2 and 4, the latents are sampled from $p_{\alpha,\rho=0.9}(\boldsymbol{z})$ and α varies from 0 to 1.

D.2.2 Metrics

We evaluate disentanglement via the *mean correlation coefficient* (Hyvarinen & Morioka, 2016; Khemakhem et al., 2020a) which is computed as follows: The Pearson correlation matrix C between the ground-truth features and learned ones is computed. Then, MCC $= \max_{\pi \in \text{permutations}} \frac{1}{m}\sum_{j=1}^{m} |C_{j,\pi(j)}|$. We also evaluate linear equivalence by performing linear regression to predict the ground-truth factors from the learned ones, and report the mean of the Pearson correlations between the ground-truth latents and the learned ones. This metric is known as the *coefficient of multiple correlation*, R, and turns out to be the square-root of the more widely

known *coefficient of determination*, R^2. The advantage of using R over R^2 is that we always have MCC $\leq R$.

D.2.3 ARCHITECTURE, INNER SOLVER & HYPERPARAMETERS

We use the four-layer convolutional neural network typically used in the disentanglement literature (Locatello et al., 2019). As mentioned in Section 2.2.2, the norm of the representation $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$ must be controlled to make sure the regularization remains effective. To do so, we apply batch normalization (Ioffe & Szegedy, 2015) at the very last layer of the neural network and do not learn its scale and shift parameters. Empirically, we do see the expected behavior that, without any normalization, the norm of $\boldsymbol{f}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$ explodes as we train, leading to instabilities and low sparsity.

In these experiments, the distribution $p(y; \boldsymbol{\eta})$ used for learning is a Gaussian with fixed variance. In that case, the inner problem of Section 2.2.2 reduces to Lasso regression. Computing the hypergradient w.r.t. $\boldsymbol{\theta}$ requires solving this inner problem. To do so, we use Proximal Coordinate Descent (Tseng, 2001; Richtárik & Takáč, 2014).

In Figures 2 and 4, we explore various levels of regularization λ. In our implementation of inner-Lasso, $\lambda_{\max} := \frac{1}{n}\|\boldsymbol{F}^\top \boldsymbol{y}\|_\infty$ where $\boldsymbol{F} \in \mathbb{R}^{n \times m}$ is the design matrix of the features of the samples of a task, while in the inner-Ridge implementation, $\lambda_{\max} := \frac{1}{n}\|\boldsymbol{F}\|^2$.

D.3 META-LEARNING EXPERIMENTS

Experimental settings. We evaluate the performance of our meta-learning algorithm based on a group-sparse SVM base-learner on the *mini*ImageNet (Vinyals et al., 2016) dataset. Following the standard nomenclature in few-shot classification (Hospedales et al., 2021) with k-shot N-way, where N is the number of classes in each classification task, and k is the number of samples per class in the training dataset $\mathcal{D}_t^{\text{train}}$, we consider 2 settings: 1-shot 5-way, and 5-shot 5-way. Note that the results presented in Figure 3 only show the performance on 5-shot classification. We use the same residual network architecture as in (Lee et al., 2019), with 12 layers and a representation of size $p = 1.6 \times 10^4$.

Even though we consider a similar base-learner as MetaOptNet (Lee et al., 2019) (namely, a SVM), our control experiment with $\lambda = 0$ cannot be directly compared to the performance of the model reported in that prior work. The reason is that in order to control for any other sources of "effective regularization" (e.g., data augmentation, label smoothing), we do not include the modifications made in MetaOptNet to improve performance. Moreover, we used a different solver (proximal block-coordinate descent, as opposed to a QP solver) to solve the inner problem Problem (8).

Generalization on meta-training tasks. In Section 2.3, we argued that evaluating the performance of the learned representations on meta-training tasks (i.e., tasks similar to the ones seen during meta-training) still shows the generalization capacity to new tasks. Indeed, those new tasks on which we evaluate performance were created using the same classes as the tasks used during meta-training, but using a combination of classes that may have not been seen in any tasks used for optimizing Problem (6). However, evaluation in meta-learning is typically done on *meta-test* tasks, i.e. tasks based on concepts that were never seen by any task during meta-training. This evaluation requires a stronger notion of generalization, closer to out-of-distribution generalization.

Base-learner	5-way 1-shot	5-way 5-shot
SVM ($\lambda = 0$)	$53.29 \pm 0.60\%$	$69.26 \pm 0.51\%$
Group-sparse SVM ($\lambda = 0.01$)	$54.22 \pm 0.61\%$	$70.01 \pm 0.51\%$
MetaOptNet (Lee et al., 2019)	$64.09 \pm 0.62\%$	$80.00 \pm 0.45\%$

Table 1: Performance of our meta-learning algorithm on the *mini*ImageNet benchmark. The performance is reported as the mean accuracy and 95% confidence interval on 1000 meta-test tasks. We also report the performance of MetOptNet (Lee et al., 2019) as reference, even though the performance is not directly comparable to our SVM baseline (see text for details).

Nonetheless, we observe in Table 1 that the performance of the meta-learning method improves as the base-learners are group-sparse.