Learning Visual-Semantic Subspace Representations

Gabriel Moreira
Instituto Superior Técnico
Carnegie Mellon University

Manuel Marques
Institute for Systems and Robotics
Instituto Superior Técnico

João Paulo Costeira
Institute for Systems and Robotics
Instituto Superior Técnico

Alexander Hauptmann
Carnegie Mellon University

Abstract

Learning image representations that capture rich semantic relationships remains a significant challenge. Existing approaches are either contrastive, lacking robust theoretical guarantees, or struggle to effectively represent the partial orders inherent to structured visual-semantic data. In this paper, we introduce a nuclear norm-based loss function, grounded in the same information theoretic principles that have proved effective in self-supervised learning. We present a theoretical characterization of this loss, demonstrating that, in addition to promoting class orthogonality, it encodes the spectral geometry of the data within a subspace lattice. This geometric representation allows us to associate logical propositions with subspaces, ensuring that our learned representations adhere to a predefined symbolic structure.

1 INTRODUCTION

In this paper, we present a methodology for learning image representations that adhere to a predefined symbolic structure. Consider a representation space $\mathcal{H}$ and an image encoded therein. Answering any Boolean proposition conditioned on it amounts to specifying a binary value and corresponds thus, to a subset of an observation space $\mathcal{Y}$. Intuitively, we expect the representation space and the observation space to be correlated – if proposition p entails q, the respective observations of p and q in $\mathcal{Y}$ preserve the set-theoretic inclusion, and so should their corresponding representations in $\mathcal{H}$. Notably, works on semantic linguistics have even conceptualized representations as regions, with the set-theoretic inclusion corresponding to semantic entailment [Geffet and Dagan, 2005, Gardenfors, 2004].

Despite their ubiquity, Euclidean embeddings with the usual inner product do not represent partial orders and symbolic operations in the most natural manner. While the commonly used cosine similarity encodes a notion of *closeness*, useful for simple retrieval tasks, its symmetry imposes a latent space geometry that is *incoherent* with that of the semantics [Palel et al., 2022]. Indeed, state-of-the-art vision-language models trained via contrastive objectives exhibit poor performance at tasks requiring compositionality [Yuksekgonul et al., 2023] and logical reasoning, such as understanding negations [Singh et al., 2024, Quantmeyer et al., 2024]. This disconnection limits our ability to harness structural constraints, compromising reasoning capabilities and hindering interpretability [Steck et al., 2024].

Previous efforts to incorporate symbolic structure into representation spaces, particularly focusing on transitive relations and the asymmetries present in ontologies and knowledge graphs, have proposed *ad-hoc* contrastive losses or geometry-inspired approaches, such as measure-based embeddings [Ren and Leskovec, 2020, Vilnis et al., 2018, Li et al., 2018], quasimetrics [Mémoli et al., 2018] and hyperbolic representations [Moreira et al., 2024, Chami et al., 2020, Chamberlain et al., 2017, Nickel and Kiela, 2017, Atigh et al., 2021, Ganea et al., 2018]. Although these approaches are effective, many of them lack robust theoretical guarantees.

In contrast, we demonstrate that a rich and consistent geometrical structure arises naturally from minimizing a nuclear norm-based loss, grounded

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

in the same information theoretic principles that underlie recent advances in self-supervised learning [Bardes et al., 2021, Lecomte et al., 2023, Yu et al., 2020, Yerxa et al., 2023]. Our work contributes to the growing interest in utilizing rank-based non-contrastive losses for representation learning. While prior results focused on enforcing class orthogonality, we identify the minimizer of the proposed loss as a spectral embedding of the data, in the form of a Boolean subspace lattice [Mittelstaedt, 1972, Birkhoff and Von Neumann, 1975].

By encoding semantic partial orders, our representations enable a probabilistic formulation of propositional queries as subspaces [van Rijsbergen, 2004]. These, in turn, facilitate the representation of logical operations such as conjunction, disjunction and negation in terms of subspace calculus. Such a capability is crucial for complex symbolic reasoning and for faithfully representing relations that extend beyond mere similarity, such as hypernymy, inclusion, and entailment, which form the cornerstone of semantic understanding.

In summary, our contributions are three-fold:

1. A new loss function that acts as a surrogate for the mutual information between the semantic, or symbolic, distribution and the embedding distribution (Section 3.1).
2. We prove that the minimization of the proposed loss guarantees that the visual representations adhere to the spectral geometry of the underlying semantics (Section 3.2).
3. We show that the learned representations form a subspace Boolean lattice, where propositions are encoded as projection operators (Section 4.1).

Our general learning framework yields representations suitable for single-label and multi-label classification, as well as for retrieval from complex propositional queries. While we present our approach in the context of visual-semantic data, any modalities may be considered. Code is available at `https://github.com/gabmoreira/subspaces`.

2 RELATED WORK

Structured Representations A large number of works have addressed the problem of endowing representation spaces with an interpretable structure, capable of encoding the partial orders inherent to knowledge graphs and image-caption hierarchies [Vendrov et al., 2015]. A notable subset of these include measure-based representations, such as Gaussian embeddings [Vilnis and McCallum, 2015], Gaussian mixtures [Choudhary et al., 2021], box embeddings [Vilnis et al., 2018, Li et al., 2018, Palel et al., 2022] and Beta embeddings [Ren and Leskovec, 2020]. An alternative direction has been to look at the problem from a geometrical standpoint, namely via negatively curved manifolds. In fact, hyperbolic embeddings have seen widespread adoption in the literature, owing to their ability to naturally represent tree-like structures [Moreira et al., 2024, Nickel and Kiela, 2018, Chamberlain et al., 2017, Nickel and Kiela, 2017, Chami et al., 2020, Atigh et al., 2021, Ganea et al., 2018].

Information Theoretic Representations Our work bridges the gap between the aforementioned research on structured representations and works on representation learning based on information theoretic principles, such as MCR^2 [Yu et al., 2020] and MMCR [Yerxa et al., 2023, Lecomte et al., 2023]. The latter, in particular, while finding applications in self-supervised learning, uses a nuclear norm-based loss similar to the regularization proposed in OLÉ [Lezama et al., 2018], as a means to improve image classification. Both OLÉ and MMCR promote intra-class low-rank and inter-class high-rank via the nuclear norm and either enforce unit-norm embeddings or impose a lower bound on the intra-class nuclear norm to avoid representation collapse. In contrast, our method focuses on the low-rank assumption of the joint visual-semantic distribution and we provide a theoretical characterization of the solution for the general multi-label setting, instead of considering only the disjoint case.

3 LEARNING SUBSPACES VIA THE NUCLEAR NORM

3.1 A Joint Low-Rank Formulation

In this section, we informally derive the proposed loss, by establishing a connection between multi-label classification via low-rank matrix completion [Cabral et al., 2011, Goldberg et al., 2010] and recent works that employ the nuclear norm as a loss [Yerxa et al., 2023, Lecomte et al., 2023], or a regularization thereof [Lezama et al., 2018]. Consider an image dataset with c labels and n images. Both the images and the labels shall be represented in some d-dimensional vector space. We define the label matrix as $\mathbf{Y} = [\mathbf{y}_1 \dots \mathbf{y}_n] \in \{0,1\}^{c \times n}$, and assume that $\mathbf{Y}$ contains all the structure that we want to impose on the representations $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_n] \in \mathbb{R}^{d \times n}$. If the latter are amenable to a linear probe, the joint distribution

$$\mathbf{Z} := \begin{bmatrix} \mathbf{Y} \\ \mathbf{X} \end{bmatrix} \in \mathbb{R}^{(c+d) \times n} \tag{1}$$

is low-rank. Inferring missing labels, or class prototypes, can thus be formulated as finding a complete low-rank matrix that is close to $\mathbf{Z}$ (for some metric) in all the observed entries [Cabral et al., 2011, Goldberg et al., 2010].

If we assume instead that we have access to $\mathbf{Y}$ and that we wish to learn $\mathbf{X}$, the problem becomes one of learning representations. In order to avoid it being ill-posed *i.e.*, representation collapse, we need to ensure that $\mathbf{X}$ is non-trivial. This can be achieved by forcing the representations to be full-rank. Using the nuclear norm $\|\cdot\|_*$ as a rank-surrogate, the proposed loss function reads as

$$l(\mathbf{X}) := \|\mathbf{Z}\|_* - \alpha\|\mathbf{X}\|_* + \beta\|\mathbf{X}\|_2^2, \tag{2}$$

for some $\beta \in (0,1)$ and $\alpha \in (0,1)$, and where $\|\cdot\|_2$ is the spectral norm. The low-rank we seek for $\mathbf{Z}$ encodes the fact that $\mathbf{X}$ should require no more information than that specified in $\mathbf{Y}$ *i.e.*, $\mathbf{Z}$ should be be redundant. Conversely, the high rank we require for $\mathbf{X}$ forces it to be informative, unless $\mathbf{Y}$ is given. This objective is *tantamount* to requiring the joint distribution of $\mathbf{X}, \mathbf{Y}$ to have low entropy and $\mathbf{X}$ to have high entropy, which corresponds to minimizing $H(\mathbf{X},\mathbf{Y}) - H(\mathbf{X}) - H(\mathbf{Y})$ *i.e.*, the negative mutual information between $\mathbf{Y}$ and the representations $\mathbf{X}$. As we state formally in Section 3.2, minimizing this proxy of the negative mutual information yields a spectral embedding of $\mathbf{Y}$.

Minimizing the Loss One of our key results rests on the minimization of the loss $l(\mathbf{X})$, formally stated by Theorem 3.3, in Section 3.2. We show that the representations that minimize (2) are proportional to the eigenvectors of the Gram matrix of the labels $\mathbf{Y}^\top\mathbf{Y}$, up to an orthogonal transformation.

The ℓ_2-penalty ensures that the top singular values of the minimizer $\mathbf{X}^*$ are equal. The hyperparameter $\alpha < 1$ guarantees that the rank of the minimizers is exactly that of $\mathbf{Y}$. Table 1 provides a summary of different nuclear norm-based losses and their minimizers, highlighting the contribution of each term.

3.2 Theoretical Guarantees and Derivations

Our objective for this section is formally derive the minimizer of (2), with the required auxiliary results. We will denote by $O(d) = \{\mathbf{U} \in \mathbb{R}^{d\times d} : \mathbf{U}^\top\mathbf{U} = \mathbf{U}\mathbf{U}^\top = \mathbf{I}_d\}$ the orthogonal group. The Stiefel manifold of orthonormal d-frames in $\mathbb{R}^n$ reads as $\mathrm{St}_d(\mathbb{R}^n) \subset \mathbb{R}^{n\times d}$ and the nullspace of a matrix as $\mathcal{N}$. The nuclear and spectral norms are denoted as $\|\mathbf{X}\|_*$ and $\|\mathbf{X}\|_2$, respectively. Recalling that the nuclear norm is the sum of the singular values $\|\mathbf{X}\|_* = \mathrm{Tr}((\mathbf{X}^\top\mathbf{X})^{1/2})$, we have that it is invariant to orthogonal transformations (proofs are deferred to Appendix B).

Lemma 3.1 (Symmetry)**.** *Let $\mathbf{Y} \in \mathbb{R}^{c\times n}$ and $\mathbf{X} \in \mathbb{R}^{d\times n}$. For any $\mathbf{U}_1 \in O(c), \mathbf{U}_2 \in O(d)$ and $\mathbf{V} \in O(n)$*

$$\left\|\begin{bmatrix}\mathbf{U}_1\mathbf{Y}\\ \mathbf{U}_2\mathbf{X}\mathbf{V}\end{bmatrix}\right\|_* = \left\|\begin{bmatrix}\mathbf{Y}\mathbf{V}^\top\\ \mathbf{X}\end{bmatrix}\right\|_* \tag{3}$$

If we fix the singular values of $\mathbf{X}$, the invariance from Lemma 3.1 turns the minimization of $\|\cdot\|_*$ over $\mathbb{R}^{d\times n}$ into a minimization over $\mathrm{St}_d(\mathbb{R}^n)$. The solution is given by the following Theorem.

Theorem 3.2. *Let $\mathbf{Y} \in \mathbb{R}^{c\times n}$ be a rank-c matrix with SVD $\mathbf{U}_Y\mathbf{\Sigma}_Y\mathbf{V}_Y^\top$, where $\mathbf{U}_Y \in O(c)$, $\mathbf{\Sigma}_Y \in \mathbb{R}^{c\times c}$ is diagonal, with singular values $\mu_1 \geq \cdots \geq \mu_c$, and $\mathbf{V}_Y \in \mathrm{St}_c(\mathbb{R}^n)$. For a rank-$d$ matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$ with SVD $\mathbf{U}_X\mathbf{\Sigma}_X\mathbf{V}_X^\top$ and singular values $\sigma_1 \geq \cdots \geq \sigma_d$, let*

$$\mathbf{V}^* = \operatorname*{arg\,min}_{\mathbf{V}_X \in \mathrm{St}_d(\mathbb{R}^n)} \left\|\begin{bmatrix}\mathbf{Y}\\ \mathbf{U}_X\mathbf{\Sigma}_X\mathbf{V}_X^\top\end{bmatrix}\right\|_*. \tag{4}$$

It follows that,

- *if $c = d$, then $\mathbf{V}^* = \mathbf{V}_Y$ and the min is $\sum_{i=1}^c \sqrt{\mu_i^2 + \sigma_i^2}$.*
- *if $d < c$, then $\mathbf{V}^* = \mathbf{V}_Y \begin{bmatrix}\mathbf{I}_d & \mathbf{0}_{d\times(c-d)}\end{bmatrix}^\top$ and the min is $\sum_{i=1}^d \sqrt{\mu_i^2 + \sigma_i^2} + \sum_{i=d+1}^c \mu_i$.*
- *if $d > c$, then $\mathbf{V}^* = \begin{bmatrix}\mathbf{V}_Y & \mathbf{V}\end{bmatrix}$, with $\mathbf{V}^\top\mathbf{V}_Y = \mathbf{0}$, and the min is $\sum_{i=1}^c \sqrt{\mu_i^2 + \sigma_i^2} + \sum_{i=c+1}^d \sigma_i$.*

From Lemma 3.1 and Theorem 3.2 we can derive our main result, which states that, for the appropriate choice of α and β, the minimizer of (2) is a spectral embedding of $\mathbf{Y}^\top\mathbf{Y}$ with the rank of $\mathbf{Y}$.

Theorem 3.3. *Let $\mathbf{Y} \in \mathbb{R}^{c\times n}$ be a rank-c matrix with SVD $\mathbf{U}_Y\mathbf{\Sigma}_Y\mathbf{V}_Y^\top$, where $\mathbf{U}_Y \in O(c)$, $\mathbf{\Sigma}_Y \in \mathbb{R}^{c\times c}$ is diagonal, with singular values $\mu_1 \geq \cdots \geq \mu_c$, and $\mathbf{V}_Y \in \mathrm{St}_c(\mathbb{R}^n)$. For $\beta \in (0,1)$ and $\sqrt{\max\left\{0, 1 - \frac{4\beta^2\mu_c^2}{c^2}\right\}} \leq \alpha < 1$, define the set*

$$\mathcal{X} := \operatorname*{arg\,min}_{\mathbf{X}\in\mathbb{R}^{d\times n}} \left\{\left\|\begin{bmatrix}\mathbf{Y}\\ \mathbf{X}\end{bmatrix}\right\|_* - \alpha\|\mathbf{X}\|_* + \beta\|\mathbf{X}\|_2^2\right\}, \tag{5}$$

for $d \geq c$. Then,

$$\mathcal{X} = \left\{\mathbf{U}(t^*\mathbf{I}_c)\mathbf{V}_Y^\top :\ \mathbf{U} \in \mathrm{St}_c(\mathbb{R}^d)\right\}, \tag{6}$$

where $t^* > 0$ is the solution to $\sum_{i=1}^c \frac{t}{\sqrt{\mu_i^2 + t^2}} = \alpha c - 2\beta t$.

As formalized in the following lemma, the ℓ_2-penalty guarantees that the non-trivial singular values of the minimizers of (2) are equal. Without this penalty, the minimizer set for $d = c$ is the orbit $\{\mathbf{R}\mathbf{Y}\alpha/\sqrt{1-\alpha^2} : \mathbf{R} \in O(d)\}$, akin to MMCR.

Table 1: Losses and minimizers for $\mathbf{Y} = \mathbf{U}_Y \mathbf{\Sigma}_Y \mathbf{V}_Y^\top \in \mathbb{R}^{c\times n}$ rank-c and $\mathbf{X} \in \mathbb{R}^{d\times n}$, with $d \geq c$.

Loss $l(\mathbf{X})$	Dim	Argmin
$\|\mathbf{Z}\|_* - \alpha\|\mathbf{X}\|_*$	$d = c$	$\frac{\alpha}{\sqrt{1-\alpha^2}}\mathbf{U}\mathbf{\Sigma}_Y\mathbf{V}_Y^\top \;:\; \mathbf{U} \in O(d)$
	$d > c$	$\mathbf{U}\begin{bmatrix}\frac{\alpha}{\sqrt{1-\alpha^2}}\mathbf{\Sigma}_Y & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}_Y^\top\\ \mathbf{V}^\top\end{bmatrix} \;:\; \mathbf{U} \in O(d), \mathbf{V} \in \mathcal{N}(\mathbf{V}_Y)$
$\|\mathbf{Z}\|_* - \|\mathbf{X}\|_* + \beta\|\mathbf{X}\|_2$	$d = c$	$\mathbf{U}(t^*\mathbf{I}_d)\mathbf{V}_Y^\top :\; \mathbf{U} \in O(d)$
	$d > c$	$\mathbf{U}\begin{bmatrix}t^*\mathbf{I}_c & \mathbf{0}\\ \mathbf{0} & \mathbf{\Sigma}\end{bmatrix}\begin{bmatrix}\mathbf{V}_Y^\top\\ \mathbf{V}^\top\end{bmatrix} \;:\; \mathbf{U} \in O(d), t^* \succeq \mathbf{\Sigma}, \mathbf{V} \in \mathcal{N}(\mathbf{V}_Y)$
$\|\mathbf{Z}\|_* - \alpha\|\mathbf{X}\|_* + \beta\|\mathbf{X}\|_2^2$	$d = c$	$\mathbf{U}(t^*\mathbf{I}_c)\mathbf{V}_Y^\top \;:\; \mathbf{U} \in O(d)$
	$d > c$	$\mathbf{U}(t^*\mathbf{I}_c)\mathbf{V}_Y^\top :\; \mathbf{U} \in \mathrm{St}_c(\mathbb{R}^d)$

Lemma 3.4 (No ℓ_2-penalty)**.** *Let $\mathbf{Y} \in \mathbb{R}^{c\times n}$ be a rank-c matrix with SVD given by $\mathbf{U}_Y\mathbf{\Sigma}_Y\mathbf{V}_Y^\top$, where $\mathbf{U}_Y \in O(c)$, $\mathbf{\Sigma}_Y \in \mathbb{R}^{c\times c}$ is diagonal, with singular values $\mu_1 \geq \cdots \geq \mu_c$, and $\mathbf{V}_Y \in \mathrm{St}_c(\mathbb{R}^n)$. For $\alpha \in (0,1)$, define the set*

$$\mathcal{X} := \underset{\mathbf{X}\in\mathbb{R}^{d\times n}}{\arg\min}\left\{\left\|\begin{bmatrix}\mathbf{Y}\\ \mathbf{X}\end{bmatrix}\right\|_* - \alpha\|\mathbf{X}\|_*\right\}, \tag{7}$$

for $d \geq c$. Then,

$$\mathcal{X} = \left\{\mathbf{U}\begin{bmatrix}\frac{\alpha}{\sqrt{1-\alpha^2}}\mathbf{\Sigma}_Y & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}_Y^\top\\ \mathbf{V}^\top\end{bmatrix} : \mathbf{U} \in O(d), \mathbf{V} \in \mathcal{N}(\mathbf{V}_Y)\right\}. \tag{8}$$

Lemma 3.5. *Let $\mathbf{Y} \in \mathbb{R}^{c\times n}$ be a rank-c matrix with SVD $\mathbf{U}_Y\mathbf{\Sigma}_Y\mathbf{V}_Y^\top$, where $\mathbf{U}_Y \in O(c)$, $\mathbf{\Sigma}_Y \in \mathbb{R}^{c\times c}$ is diagonal, with singular values $\mu_1 \geq \cdots \geq \mu_c$, and $\mathbf{V}_Y \in \mathrm{St}_c(\mathbb{R}^n)$. For $\beta \in (0,1)$, define the set*

$$\mathcal{X} := \underset{\mathbf{X}\in\mathbb{R}^{d\times n}}{\arg\min}\left\{\left\|\begin{bmatrix}\mathbf{Y}\\ \mathbf{X}\end{bmatrix}\right\|_* - \|\mathbf{X}\|_* + \beta\|\mathbf{X}\|_2\right\}, \tag{9}$$

for $d \geq c$. Then,

$$\mathcal{X} = \left\{\mathbf{U}\begin{bmatrix}t^*\mathbf{I}_c & \mathbf{0}\\ \mathbf{0} & \mathbf{\Sigma}\end{bmatrix}\begin{bmatrix}\mathbf{V}_Y^\top\\ \mathbf{V}^\top\end{bmatrix} : \mathbf{U} \in O(d),\ t^*\mathbf{I} \succeq \mathbf{\Sigma},\ \mathbf{V} \in \mathcal{N}(\mathbf{V}_Y)\right\}, \tag{10}$$

where $t^* > 0$ is the solution to $\sum_{i=1}^c \frac{t}{\sqrt{\mu_i^2+t^2}} = c - \beta$.

Corollary 3.6 (Orthogonal disjoint classes)**.** *Let $\mathbf{Y}$ contain n samples from c disjoint classes and*

$$\mathbf{X} \in \underset{\mathbf{X}\in\mathbb{R}^{d\times n}}{\arg\min}\left\{\left\|\begin{bmatrix}\mathbf{Y}\\ \mathbf{X}\end{bmatrix}\right\|_* - \alpha\|\mathbf{X}\|_* + \beta\|\mathbf{X}\|_2^2\right\}. \tag{11}$$

Then, $\mathbf{y}_i \neq \mathbf{y}_j \implies \langle \mathbf{x}_i, \mathbf{x}_j\rangle = 0$.

Comparison with OLÉ, MMCR and MCR^2 MMCR maximizes the nuclear norm of the matrix of ℓ_2-normalized class centroids. The optimal representations for this loss have the same right-singular vectors as the label matrix $\mathbf{Y}$, with the singular values of the former proportional to those of the latter. MMCR behaves thus similarly to the loss in the first row of Table 1. While it represents disjoint classes as orthogonal subspaces, the solution for correlated labels is unclear. The orthogonalization proposed in OLÉ acts as a regularizer for cross-entropy training, whereas the loss we propose (2) is standalone. In MCR^2 [Yu et al., 2020], disjoint classes are encoded as orthogonal subspaces, with their sum spanning the entire representation space. In our case, the corresponding subspaces have the smallest dimension necessary to represent $\mathbf{Y}$ *i.e.*, the rank of our representations is that of the semantics, regardless of the ambient space. This key difference can be attributed to the use of the nuclear norm which, being the convex envelope of the rank, naturally leads to low-rank solutions.

Comparison with Covariance Regularization From Lemma 3.1, the orbit $\{\mathbf{RX} : \mathbf{R} \in O(d)\}$ contains equivalent embedding matrices. Denote the SVD of $\mathbf{X}$ by $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. Since $\mathbf{U} \in O(d)$, we can define an equivalent embedding matrix by writing $\mathbf{X}$ in a new basis $\mathbf{U}$ such that its covariance is diagonal (assuming centered embeddings) *i.e.*, $\mathbf{X}' := \mathbf{U}^\top\mathbf{X}$ and $\mathbf{X}'\mathbf{X}'^\top = \mathbf{U}^\top\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top\mathbf{U} = \mathbf{\Sigma}^2$. Thus, instead of searching for $\mathbf{X}$ that minimizes $\|\mathbf{Z}\|_*$ and maximizes $\|\mathbf{X}\|_*$, we can limit the search space to the set of matrices which admit the canonical basis of $\mathbb{R}^c$ as left singular vectors *i.e.*, with an SVD $\mathbf{X}' = \mathbf{\Sigma}\mathbf{V}^\top$. For such matrices, the nuclear norm is given by the trace of $(\mathbf{X}\mathbf{X}^\top)^{\frac{1}{2}}$ and the problem of maximizing $\|\mathbf{X}\|_*$ becomes

$$\max_{\mathbf{X}\in\mathbb{R}^{c\times n}} \sum_{i=1}^c \sqrt{(\mathbf{X}\mathbf{X}^\top)_{ii}} \quad \text{s.t. } (\mathbf{X}\mathbf{X}^\top)_{ij} = 0,\ i \neq j. \tag{12}$$

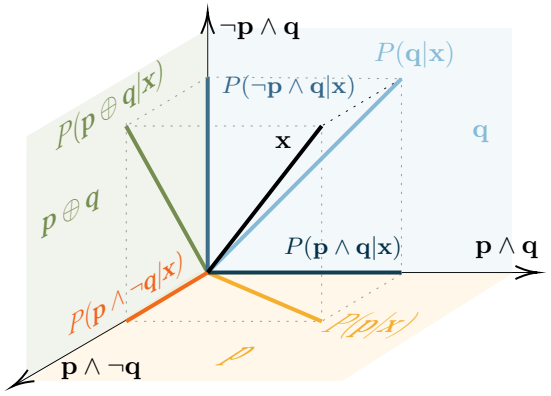


Figure 1: Subspace Boolean lattice. Each axis encodes a minterm of 2 literals: $\mathbf{p} \wedge \mathbf{q}$, $\neg\mathbf{p} \wedge \mathbf{q}$ and $\mathbf{p} \wedge \neg\mathbf{q}$. The propositions $\mathbf{p}$ and $\mathbf{q}$ are represented by two orthogonal 2-d subspaces. The squared norm of the projection of $\mathbf{x}$, with $\|\mathbf{x}\| = 1$, over each subspace yields the posterior probability of the corresponding proposition.

The terms of the sum are the standard deviations of each coordinate *i.e.*, $\sqrt{(\mathbf{X}\mathbf{X}^\top)_{ii}} = \sqrt{\text{Cov}[\mathbf{X}_{\cdot,i}]}$. This is simply the covariance regularization proposed in VICReg [Bardes et al., 2021].

4 SUBSPACES AS PROPOSITIONS

Given a minimizer of (2) we can derive probabilistic answers to propositions conditioned on the representations. This geometry-induced probabilistic formulation encodes the semantic partial orders, allowing us to perform propositional calculus on the representations, as illustrated in Fig. 1.

4.1 Boolean Subspace Lattice

Given a full-row rank $\mathbf{Y} \in \{0,1\}^{c\times n}$, the minimizer of (2) embeds an inclusion Boolean sublattice of the power set $2^{[c]}$, as a subspace lattice of $\mathbb{R}^d$. We start by showing that, if we consider the columns of $\mathbf{Y} = [\mathbf{y}_1 \dots \mathbf{y}_n]$, denoted $\mathbf{y}_i \in \{0,1\}^c$, as minterms of c literals (logical propositions), then for any $\mathbf{y}_i \neq \mathbf{y}_j$, the corresponding embeddings are orthogonal *i.e.*, $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = 0$. This allows us to associate more general propositions $\mathbf{q}$ to subspaces $\mathcal{S}_q \subseteq \mathbb{R}^c$. Logical implication $\mathbf{p} \implies \mathbf{q}$ or equivalently, the semantic partial order $\mathbf{q} \geq \mathbf{p}$, corresponds in the representation space to the subspace inclusion $\mathcal{S}_p \subseteq \mathcal{S}_q$. This is a more general version of the orthogonality of disjoint classes of MMCR and MCR^2.

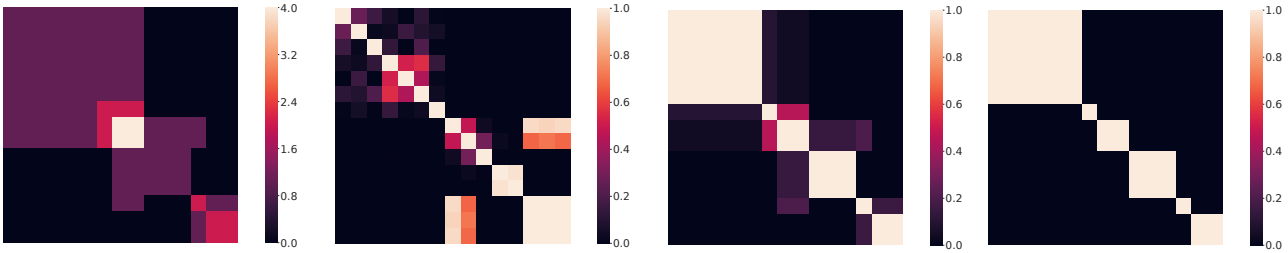
Lemma 4.1 (Minterm orthogonality)**.** *Let $\mathbf{Y} \in \{0,1\}^{c\times n}$ be a rank-c matrix with SVD $\mathbf{U}_Y\mathbf{\Sigma}_Y\mathbf{V}_Y^\top$, where $\mathbf{U}_Y \in O(c)$, $\mathbf{\Sigma}_Y \in \mathbb{R}^{c\times c}$ and $\mathbf{V}_Y \in \text{St}_c(\mathbb{R}^n) \subset \mathbb{R}^{n\times c}$, with rows $\{\mathbf{v}_i^\top\}_{i\in[n]}$. Let $\mathcal{I}$ be the largest index set such that for $i,j \in \mathcal{I}$ $\mathbf{y}_i \neq \mathbf{y}_j$. If* rank $\mathbf{Y} = |\mathcal{I}|$ *then $\{\mathbf{v}_i\}_{i\in\mathcal{I}}$ is an orthogonal basis for $\mathbb{R}^c$.*

Recall from Theorem 3.3, that the minimizer of (2) is precisely $\mathbf{V}_Y^\top$, up to a global scale and an orthogonal transformation. Combining this with Lemma 4.1, we have that the normalized embeddings corresponding to the unique minterms $\{\mathbf{x}_i/\|\mathbf{x}_i\|_2\}_{i\in\mathcal{I}}$ form an orthonormal basis for the representation space $\mathbb{R}^c$. We will henceforth write this basis as $\{\mathbf{e}_i\}_{i\in\mathcal{I}}$. Given a unit ℓ_2-norm representation $\mathbf{x}$, we have $\sum_{i\in\mathcal{I}}\langle \mathbf{x}, \mathbf{e}_i\rangle^2 = 1$ and we can define the probability of the i-th minterm $\mathbf{y}_i$, given the image, as $P(\mathbf{y}_i|\mathbf{x}) := \langle \mathbf{x}, \mathbf{e}_i\rangle^2$ (by abuse of notation we will use the same symbols for a proposition and the corresponding random variable). Since $0 \leq \langle \mathbf{x}, \mathbf{e}_i\rangle^2 \leq 1$, the representations induce a categorical distribution over the dictionary $\{\mathbf{y}_i\}_{i\in\mathcal{I}}$. More general propositions $\mathbf{q}$ over the c labels can be written as a disjunction of the conjunctions $\{\mathbf{y}_i\}_{i\in\mathcal{I}}$ *i.e.*, in the disjunctive normal form (DNF).

Lemma 4.2 (Propositions as projections)**.** *Let $\mathbf{Y} \in \{0,1\}^{c\times n}$ verify the conditions of Lemma 4.1. Given $\mathbf{x} \in \mathbb{R}^d$, with $\|\mathbf{x}\|_2 = 1$, define the posterior probability of the Bernoulli variable associated with the minterm $\mathbf{y}_i$ as $P(\mathbf{y}_i|\mathbf{x}) := \langle \mathbf{x}, \mathbf{e}_i\rangle^2$. Then, $\forall \mathbf{q}$ such that $\mathbf{q} = \bigvee_{i\in\mathcal{J}} \mathbf{y}_i$, for some $\mathcal{J} \subseteq \mathcal{I}$, there is a unique projection operator $\mathbf{P}_q$ such that $P(\mathbf{q}|\mathbf{x}) = \langle \mathbf{x}, \mathbf{P}_q\mathbf{x}\rangle$.*

If given $\mathbf{x}$, proposition $\mathbf{q}$ holds, $\mathbf{x}$ lies in the corresponding subspace *i.e.*, $\mathbf{P}_q\mathbf{x} = \mathbf{x}$ and $P(\mathbf{q}|\mathbf{x}) = 1$. Conversely, if $\mathbf{q}$ is false then $\mathbf{P}_q\mathbf{x} = 0$ and $P(\mathbf{q}|\mathbf{x}) = 0$. The angle between the proposition subspace and the representation determines the probability $P(\mathbf{q}|\mathbf{x}) = \langle \mathbf{x}, \mathbf{P}_q\mathbf{x}\rangle$. Therefore, this formulation allows us to update representations via projections. If new information implies that $\mathbf{q}$ holds for $\mathbf{x}$, we consider the new embedding $\mathbf{P}_q\mathbf{x}$. The same reasoning applies to disjunction, conjunction or negation of propositions, with $\mathbf{P}_{p\wedge q} = \mathbf{P}_p\mathbf{P}_q$, $\mathbf{P}_{p\vee q} = \mathbf{P}_p + \mathbf{P}_q - \mathbf{P}_p\mathbf{P}_q$ and $\mathbf{P}_{\neg q} = \mathbf{I} - \mathbf{P}_q$.

The set of subspaces of a vector space forms a lattice under subspace inclusion [Von Neumann, 2018]. Given subspaces $\mathcal{S}_p$ and $\mathcal{S}_q$, $\mathcal{S}_p \wedge \mathcal{S}_q$ is the set-theoretic intersection *i.e.*, the greatest lower bound or *meet*. The least upper bound or *join*, $\mathcal{S}_p \vee \mathcal{S}_q$ is given by the closure of the sum $\mathcal{S}_p + \mathcal{S}_q$. From the one-to-one correspondence between projections and closed subspaces, the lattice structure has an algebraic characterization as $\mathcal{S}_p \leq \mathcal{S}_q \Leftrightarrow \mathbf{P}_p \leq \mathbf{P}_q \Leftrightarrow \mathbf{P}_p = \mathbf{P}_p\mathbf{P}_q$. If the projections commute then $\mathbf{P}_p \wedge \mathbf{P}_q = \mathbf{P}_p\mathbf{P}_q$ and $\mathbf{P}_p \vee \mathbf{P}_q = \mathbf{P}_p + \mathbf{P}_q - \mathbf{P}_p\mathbf{P}_q$. The *largest* projection is the identity $\mathbf{I}$, and the smallest one is the zero operator $\mathbf{0}$. We have that $\mathbf{P}_p \vee \mathbf{P}_p^\perp = \mathbf{I}$, $\mathbf{P}_p \wedge \mathbf{P}_p^\perp = \mathbf{0}$ and $\mathbf{P}_p = \mathbf{P}_p^{\perp\perp}$, where $\mathbf{P}_p^\perp = \mathbf{I} - \mathbf{P}_p$ is called the orthocomplement of $\mathbf{P}_p$. Every subspace lattice is orthomodular [Holland, 1975], which is a weaker version of the distributive property of Boolean algebras. In our case, the subspace lattice is Boolean since all the projections share the eigenbasis $\{\mathbf{e}_i\}_{i\in\mathcal{I}}$ and thus commute.



(a) $\mathbf{Y}^\top\mathbf{Y}$ (b) OLÉ (c) MMCR (d) Ours

Figure 2: Gram matrix of $\mathbf{Y} \in \{0,1\}^{6\times 15}$ and of the representations optimized with OLÉ, MMCR and our loss.

5 EXPERIMENTS

Given the fundamental differences in how representations are learned and inference conducted, it is important to verify that the unique properties of our methodology do not come at the expense of performance in canonical benchmarks. Thus, we first evaluate our approach in standard classification and then present results for retrieval from propositional queries. In both cases, we only train encoder models, not employing linear probes. At inference time, proposition probabilities are derived from the geometry of the representations (Lemma 4.2). Our results show that the proposed loss can surpass traditional cross-entropy training in standard classification. In multi-label settings, our method captures the true semantic geometry, unlike MMCR or OLÉ, which allows for image retrieval from complex propositional queries.

5.1 Implementation Details

We implemented the loss (2) in PyTorch, using the subdifferential of the nuclear norm [Recht et al., 2010] $\mathbf{U}_X\mathbf{V}_X^\top \in \partial_{\mathbf{X}}\|\mathbf{X}\|_*$ as the descent direction in SGD. While the nuclear norm is non-smooth, we observed empirical convergence in the all the experiments conducted, akin to prior works. All experiments were performed on a Tesla T4 GPU with 16GB of memory. During training, the batched output of the image encoder is plugged directly in (2), without any normalization or centering. The subspace representation of each proposition was computed from the training data, via the SVD of the embedding matrix of images for which the proposition holds. If $\mathbf{X}$ is the matrix of all the representations from the training set that verify $\mathbf{q}$, then $\mathbf{q}$ is represented by the subspace spanned by the singular vectors of $\mathbf{X}$ corresponding to the largest singular values. These are the only non-null singular values if (2) attains its minimum.

In order for the label matrix $\mathbf{Y}$ to have full row-rank equal to the number of minterms, as required by Theorem 3.3, we construct a minterm batch sampler so that every batch has as many minterms as labels, with each minterm having the same number of samples in the batch.

5.2 Synthetic Experiments

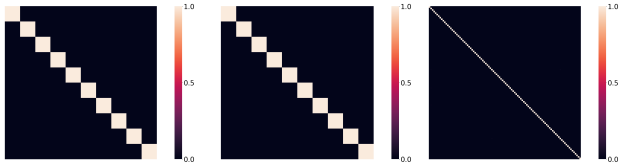
In order to shed light on the differences between our approach and the nuclear norm-based losses from OLÉ and MMCR, we optimized representations $\mathbf{X}$ for synthetic binary label matrices $\mathbf{Y}$. Since all methods, ours included, provide theoretical guarantees for the orthogonality of disjoint classes, we focus here on the multi-label case. In Fig. 2 we plot the Gram matrix of a label matrix $\mathbf{Y}^\top\mathbf{Y} \in \{0,1\}^{n\times n}$, and the Gram matrices of the ℓ_2-normalized solutions $\mathbf{X}^\top\mathbf{X} \in \mathbb{R}^{n\times n}$. OLÉ, MMCR were implemented as described in the respective papers, with MMCR requiring embedding normalization and OLÉ using a lower bound on the nuclear norm to avoid collapse during training. The hyperparameters of our loss (2) were set to $\alpha = 0.99$ and $\beta = 0.7$, with optimization performed via gradient descent. In Fig. 2d, we observe that our representations verify $\forall \mathbf{y}_i \neq \mathbf{y}_j \implies \langle \mathbf{x}_i, \mathbf{x}_j \rangle = 0$ *i.e.*, different minterms correspond to orthogonal directions, as consequence of Theorem 3.3 and Lemma 4.1. Thus, our approach generalizes the disjoint classes setting, unlike the others methods considered, which converge to representations with no clear interpretation. In Appendix C.2, we present additional experiments, showing that in standard classification, all three methods represent disjoint classes as orthogonal subspaces, as predicted theoretically. We also plot the convergence of the singular values of the our representations during training, which are in accordance with Theorem 3.3.

5.3 Standard Classification

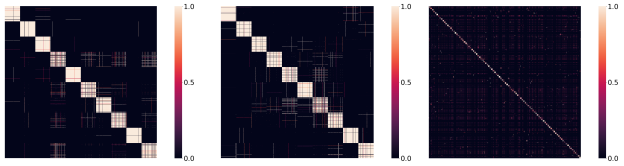
We report the classification performance of our method on MNIST [Deng, 2012], FashionMNIST [Xiao et al., 2017], CIFAR-10 and CIFAR-100 [Krizhevsky et al., 2009], comparing it with the standard approach of using linear classifiers and optimizing a cross-entropy loss.

Table 2: Classification accuracy

Dataset	Backbone	Ours	CrossEnt.
MNIST	ConvNet	0.992	0.992
	ResNet-18	**0.996**	0.995
FashionMNIST	ConvNet	**0.932**	0.930
	ResNet-18	0.935	0.935
CIFAR-10	ResNet-18	**0.934**	0.929
CIFAR-100	ResNet-18	**0.728**	0.705



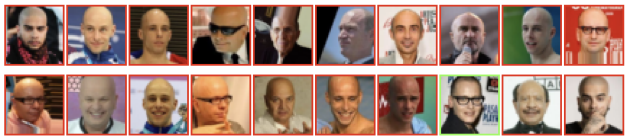
(a) FashionMNIST (b) CIFAR-10 (c) CIFAR-100



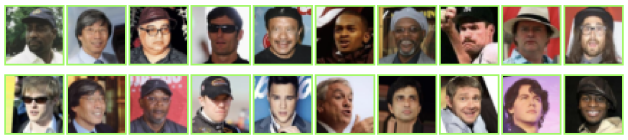
(d) FashionMNIST (e) CIFAR-10 (f) CIFAR-100

Figure 3: **Top**: Inner products between the principal direction of each class. **Bottom**: Inner products between the unit ℓ_2-norm test set embeddings.

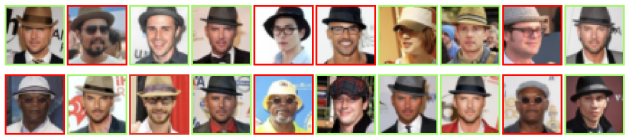
We conducted experiments with two backbones, a standard ConvNet with 2 convolutional layers and a ResNet-18 [He et al., 2016], both with leaky ReLU activations and a fully connected output layer with dimension equal to the number of classes. Minimization of (2) was carried out via SGD, with a batch size of 512 and no weight decay. For CIFAR-10 and CIFAR-100, we applied the standard data augmentations, including random cropping with a 4-pixel padding resized to 32×32 and random horizontal flipping. Full training details can be consulted in Appendix C.3. We report the results in Table 2, which show that we perform on par with, or better than, the *de facto* classification approach of using linear probes and optimizing a cross-entropy loss. To provide further insight into the geometry of the representation space, we plot in Fig. 3 the squared inner products between the 1-dimensional subspaces of each class, and the squared inner products between the ℓ_2-normalized embeddings of the test set, for three of the datasets. Notice that, in each dataset, the set of 1-d subspaces representing the different classes forms an orthonormal basis for the corresponding representation space, as predicted theoretically.



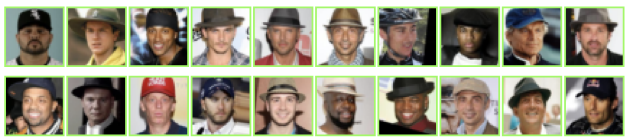
(a) **CLIP**: *A male person that is not bald*



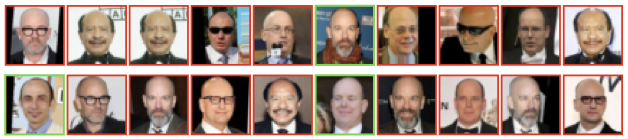
(b) **Ours**: `Male` $\wedge\neg$ `Bald`



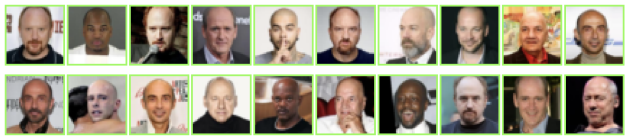
(c) **CLIP**: *A male person without eyeglasses and wearing a hat*



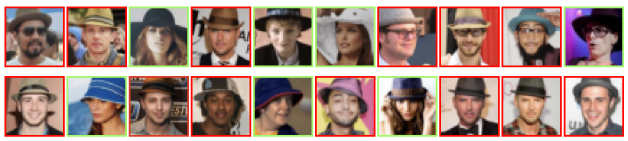
(d) **Ours**: `Male` $\wedge$ `Wearing Hat` $\wedge\neg$ `Eyeglasses`



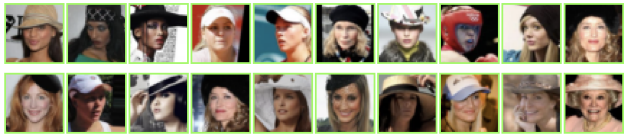
(e) **CLIP**: *A bald male person without eyeglasses and not wearing a necktie*



(f) **Ours**: `Male` $\wedge$ `Bald` $\wedge\neg$ `Wearing Necktie` $\wedge\neg$ `Eyeglasses`



(g) **CLIP**: *A person that is not male wearing a hat*



(h) **Ours**: `Wearing Hat` $\wedge\neg$ `Male`

Figure 4: Ours vs CLIP's top-20 retrieved images from the test set of Celeb-A using propositional and the corresponding natural language queries.

5.4 Retrieval

We showcase the applicability of our approach to the multi-label setting by using a subset of CelebA [Liu et al., 2015] corresponding to five of the most consistent labels: `Men`, `Bald`, `Eyeglasses`, `Wearing Hat`, and `Wearing Necktie`. This subset comprises 72,900 training images, 8,952 validation images, and 8,303 test images, with 25 minterms. We used a ResNet-18 [He et al., 2016] backbone with output dimension 25, optimized via SGD, with momentum set to 0.98 and an initial learning rate of 10^{-4}, decayed by 0.5 after 100 epochs. We used $\alpha = 0.999$ and $\beta = 0.01$.

To evaluate the retrieval performance of our approach, we constructed propositional queries using subsets of the 5 labels as literals, along with the corresponding natural language counterparts. For example, the proposition with 3 literals $\texttt{Bald} \wedge \texttt{Male} \wedge \neg\texttt{Eyeglasses}$ translates to the query '*a bald male person without eyeglasses*'. More examples are available in Appendix C.4. In total, there are 242 queries, of which 171 have at least 10 examples in the test set. With these logical queries and their natural language counterparts, we compared our retrieval results against zero-shot (ZS) CLIP [Radford et al., 2021], using the ViT-B/32 visual encoder. Motivated by [Yuksekgonul et al., 2023], we also present results for Bag-of-Words CLIP (BoW), where we prompt CLIP with an enumeration of the labels, without logical connectives. In the aforementioned example, the BoW query is simply '*bald male eyeglasses*'. We report the mean average precision (mAP) and Precision@10 (Pr@10) in Table 3. Given prior observations that vision-language models struggle with negation, we present results separately for positive queries (no negated literals) and queries containing at least one negated literal.

Table 3: Celeb-A retrieval results

Model	Positive		w/ negations	
	Pr@10	mAP	Pr@10	mAP
Ours	**0.93**	**0.75**	**0.88**	**0.79**
ZS-CLIP	0.67	0.46	0.21	0.30
ZS-CLIP (BoW)	0.66	0.48	0.11	0.24

Our method consistently outperforms CLIP, with the performance gap widening for queries involving negation. As expected, CLIP performs similarly for natural language and BoW in the positive case. For queries with negated literals, removing logical connectives does not have a drastic impact on CLIP's performance, suggesting that CLIP represents these queries in a manner that is closer to a bag-of-words, rather than a logically meaningful embedding. We present retrieval examples for multiple queries in Fig. 4, showing the top-20 retrieved images by our model and by CLIP. Green and red image borders indicate true and false positives, respectively.

6 CONCLUSION

In this work, we introduced a novel approach for learning image representations that adhere to the semantic geometry of the data. Motivated by recent advances in self-supervised learning, our method is based on a new nuclear norm-based loss function, which we show to yield the spectral embeddings of the data. Our approach, being theoretically grounded, diverges both from contrastive learning approaches and methods that simply emphasize class orthogonality. Instead, we show that our representations form a subspace Boolean lattice, which facilitates the definition of probabilistic propositional queries through projection operators. We believe that our work can thus reveal new possibilities for handling complex retrieval tasks and multi-label classification. Finally, while we presented our approach targeting visual-semantic representations, the proposed loss is modality agnostic.

Limitations and Future Work While we provide theoretical guarantees for our approach, some limitations are worthy of mention. Firstly, the use of the nuclear norm leads to the non-smoothness of the loss function. This can make optimization not as straightforward as that of other losses, namely the cross-entropy or the log-determinant. Nevertheless, all the experiments conducted point to good convergence properties when employing the sub-differential as the descent direction. Secondly, and akin to similar works in representation learning, we did not investigate the convergence properties of the batched version of the proposed nuclear norm-based loss, relying uniquely on empirical validation. Finally, we consider our main contribution to be theoretical and additional experiments would showcase the full potential of the proposed methodology in multi-label classification and retrieval.

Acknowledgements

This work was supported by LARSyS funding (DOI: `10.54499/LA/P/0083/2020`, `10.54499/UIDP/50009/2020`, and `10.54499/UIDB/50009/2020`), through Fundação para a Ciência e a Tecnologia. G Moreira was also supported via grant `SFRH/BD/151466/2021` through the Carnegie Mellon Portugal program. M Marques and J Costeira were also supported by the PT Smart Retail project (PRR - `02/C05-i11/2024.C645440011-00000062`), through IAPMEI - Agência para a Competitividade e Inovação.

References

[Atigh et al., 2021] Atigh, M. G., Keller-Ressel, M., and Mettes, P. (2021). Hyperbolic Busemann Learning with Ideal Prototypes. In *Advances in Neural Information Processing Systems*, pages 103–115.

[Bardes et al., 2021] Bardes, A., Ponce, J., and LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*.

[Birkhoff and Von Neumann, 1975] Birkhoff, G. and Von Neumann, J. (1975). *The logic of quantum mechanics*. Springer.

[Cabral et al., 2011] Cabral, R., Torre, F., Costeira, J. P., and Bernardino, A. (2011). Matrix completion for multi-label image classification. In *Advances in neural information processing systems*, volume 24.

[Chamberlain et al., 2017] Chamberlain, B. P., Clough, J., and Deisenroth, M. P. (2017). Neural embeddings of graphs in hyperbolic space.

[Chami et al., 2020] Chami, I., Gu, A., Chatziafratis, V., and Ré, C. (2020). From trees to continuous embeddings and back: hyperbolic hierarchical clustering. In *Advances in Neural Information Processing Systems*, pages 15065–15076.

[Choudhary et al., 2021] Choudhary, N., Rao, N., Katariya, S., Subbian, K., and Reddy, C. (2021). Probabilistic entity representation model for reasoning over knowledge graphs. *Advances in Neural Information Processing Systems*, 34:23440–23451.

[Deng, 2012] Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142.

[Ganea et al., 2018] Ganea, O.-E., Bécigneul, G., and Hofmann, T. (2018). Hyperbolic entailment cones for learning hierarchical embeddings. In *International Conference on Machine Learning*, pages 1646–1655.

[Gardenfors, 2004] Gardenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT press.

[Geffet and Dagan, 2005] Geffet, M. and Dagan, I. (2005). The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 107–114.

[Goldberg et al., 2010] Goldberg, A., Recht, B., Xu, J., Nowak, R., and Zhu, J. (2010). Transduction with matrix completion: Three birds with one stone. *Advances in neural information processing systems*, 23.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Holland, 1975] Holland, S. S. (1975). *The Current Interest in Orthomodular Lattices*, pages 437–496. Springer Netherlands, Dordrecht.

[Krizhevsky et al., 2009] Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images. Technical report, Toronto, ON, Canada.

[Lecomte et al., 2023] Lecomte, V., Schaeffer, R., Isik, B., Khona, M., LeCun, Y., Koyejo, S., Gromov, A., and Shwartz-Ziv, R. (2023). An information-theoretic understanding of maximum manifold capacity representations. In *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations*.

[Lezama et al., 2018] Lezama, J., Qiu, Q., Musé, P., and Sapiro, G. (2018). Ole: Orthogonal low-rank embedding-a plug and play geometric loss for deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8109–8118.

[Li et al., 2018] Li, X., Vilnis, L., Zhang, D., Boratko, M., and McCallum, A. (2018). Smoothing the geometry of probabilistic box embeddings. In *International Conference on Learning Representations*.

[Liu et al., 2015] Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[Mémoli et al., 2018] Mémoli, F., Sidiropoulos, A., and Sridhar, V. (2018). Quasimetric embeddings and their applications. *Algorithmica*, 80:3803–3824.

[Mittelstaedt, 1972] Mittelstaedt, P. (1972). On the interpretation of the lattice of subspaces of the hilbert space as a propositional calculus. *Zeitschrift für Naturforschung A*, 27(8-9):1358–1362.

[Moreira et al., 2024] Moreira, G., Marques, M., Costeira, J. P., and Hauptmann, A. (2024). Hyperbolic vs euclidean embeddings in few-shot learning: Two sides of the same coin. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2082–2090.

[Nickel and Kiela, 2017] Nickel, M. and Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*.

[Nickel and Kiela, 2018] Nickel, M. and Kiela, D. (2018). Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In *International Conference on Machine Learning*, pages 3779–3788.

[Palel et al., 2022] Palel, D., Dangati, P., Lee, J.-Y., and Boratko, Michaela nd McCallum, A. (2022). Modeling label space interactions in multi-label classification using box embeddings. In *International Conference on Learning Representations*.

[Quantmeyer et al., 2024] Quantmeyer, V., Mosteiro, P., and Gatt, A. (2024). How and where does clip process negation? In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 59–72.

[Radford et al., 2021] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

[Recht et al., 2010] Recht, B., Fazel, M., and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501.

[Ren and Leskovec, 2020] Ren, H. and Leskovec, J. (2020). Beta embeddings for multi-hop logical reasoning in knowledge graphs. *Advances in Neural Information Processing Systems*, 33:19716–19726.

[Singh et al., 2024] Singh, J., Shrivastava, I., Vatsa, M., Singh, R., and Bharati, A. (2024). Learn" no" to say" yes" better: Improving vision-language models via negations. *arXiv preprint arXiv:2403.20312*.

[Steck et al., 2024] Steck, H., Ekanadham, C., and Kallus, N. (2024). Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM on Web Conference 2024*, pages 887–890.

[van Rijsbergen, 2004] van Rijsbergen, C. J. (2004). *The Geometry of Information Retrieval*. Cambridge University Press.

[Vendrov et al., 2015] Vendrov, I., Kiros, R., Fidler, S., and Urtasun, R. (2015). Order-embeddings of images and language. In *arXiv preprint arXiv:1511.06361*.

[Vilnis et al., 2018] Vilnis, L., Li, X., Murty, S., and McCallum, A. (2018). Probabilistic embedding of knowledge graphs with box lattice measures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 263–272.

[Vilnis and McCallum, 2015] Vilnis, L. and McCallum, A. (2015). Word representations via gaussian embedding. In *ICLR*.

[Von Neumann, 2018] Von Neumann, J. (2018). *Mathematical Foundations of Quantum Mechanics*. Princeton University Press.

[Xiao et al., 2017] Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

[Yerxa et al., 2023] Yerxa, T., Kuang, Y., Simoncelli, E., and Chung, S. (2023). Learning efficient coding of natural images with maximum manifold capacity representations. In *Advances in Neural Information Processing Systems*, volume 36.

[Yu et al., 2020] Yu, Y., Chan, K. H. R., You, C., Song, C., and Ma, Y. (2020). Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *Advances in Neural Information Processing Systems*, 33:9422–9434.

[Yuksekgonul et al., 2023] Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., and Zou, J. (2023). When and why vision-language models behave like bags-of-words, and what to do about it? *International Conference on Learning Representations*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes - Sections 3.2]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [No]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes - `https://github.com/gabmoreira/subspaces`]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes - Assumptions presented in all Lemmas and Theorems.]

(b) Complete proofs of all theoretical results. [Yes - Proofs in Appendix B]

(c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes - `https://github.com/gabmoreira/subspaces`.]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes - Section 5 and Appendix C]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [No]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes - Section 5]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [Yes - MNIST, FashionMNIST, CIFAR-10, CIFAR-100 and CelebA]

(b) The license information of the assets, if applicable. [Yes]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable - Publicly available data]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Learning Visual-Semantic Subspace Representations: Supplementary Materials

A Definitions

We consider finite dimensional vector spaces over $\mathbb{R}$ with inner product $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^\top \mathbf{v}$ *i.e.*, the usual dot product. The induced norm is thus $\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. A linear operator is called symmetric (self-adjoint) if $\mathbf{T} = \mathbf{T}^\top$. A symmetric operator $\mathbf{T}$ is said to be positive if $\langle \mathbf{T}\mathbf{x}, \mathbf{x} \rangle > 0, \forall \mathbf{x} \in \mathbb{R}^d$. A projection operator is an idempotent self-adjoint *i.e.*, $\mathbf{T} = \mathbf{T}^\top = \mathbf{T}^2$.

Definition A.1 (Spectral Norm)**.** *The spectral norm is defined as*

$$\|\mathbf{T}\|_2 := \max_{\mathbf{x}} \frac{\|\mathbf{T}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}. \tag{13}$$

If $\sigma_1 \geq \cdots \geq \sigma_d$ are singular values of $\mathbf{T}$ then $\|\mathbf{T}\|_2 = \sigma_1$.

Definition A.2 (Nuclear Norm)**.** *The nuclear norm of an operator, denoted $\|\cdot\|_*$, is the ℓ_1-norm of its singular values. For $\mathbf{T}$ with singular values $\sigma_1 \geq \cdots \geq \sigma_n$, $\|\mathbf{T}\|_* = \sum_{i=1}^n \sigma_i$. It follows that $\|\mathbf{T}\|_* = \operatorname{Tr}\left(\left(\mathbf{T}^\top \mathbf{T}\right)^{\frac{1}{2}}\right)$. The nuclear norm is the convex envelope of the rank.*

Definition A.3 (Partially Ordered Set)**.** *A partially ordered set (poset) is a set where a partial order is defined. Given a set $\mathcal{X}$ a (weak) partial order is a binary relation $\leq$ between certain pairs of elements that is reflexive (every element relates to itself) $a \leq a$, antisymmetric $a \leq b, b \leq a \Rightarrow a = b$ and transitive $a \leq b, b \leq c \Rightarrow a \leq c$. Strong partial orders are not reflexive. A function between posets is called order-preserving or isotone if $x \leq y \Rightarrow f(x) \leq f(y)$. Every poset $(\mathcal{X}, \preceq)$ is isomorphic to the poset $(2^{\mathcal{X}}, \subseteq)$ i.e., the poset defined on the power set of $\mathcal{X}$, with inclusion as the partial order.*

Definition A.4 (Transitive Relation)**.** *A relation R is transitive if $xRy, yRz \implies zRz$. For example, inclusion and composition are transitive relations.*

Definition A.5 (Join and Meet)**.** *An element m of a poset $\mathcal{P}$ is the meet of x, y, denoted $x \wedge y$, if $m \leq x, m \leq y$ and if $w \leq x, w \leq y$ then $w \leq m$. Thus, m is the greatest lower bound (infimum). An element j of $\mathcal{P}$ is the join of x, y denoted $x \vee y$ if is is the lowest upper bound (supremum).*

Definition A.6 (Order Lattice)**.** *A poset is a lattice if every two-element subset has a meet and a join.*

Definition A.7 (Modular Lattice)**.** *A lattice is called modular if for all elements a, b and c the implication*

$$a \leq c \implies a \vee (b \wedge c) = (a \vee b) \wedge c \tag{14}$$

holds. Hence distributivity may not hold.

Definition A.8 (Complemented Lattice)**.** *A bounded lattice, with least and greatest elements denoted 0 and 1, respectively, is called complemented if every element a has a complement b such that $a \wedge b = 0$ and $a \vee b = 1$. Orthocomplementation is the operation that maps a to $a^\perp$ such that $a \wedge a^\perp = 0$, $a \vee a^\perp = 1$ and $a^{\perp^\perp} = 1$.*

Definition A.9 (Boolean Lattice)**.** *A Boolean lattice is a complemented distributive lattice.*

B Proofs

B.1 Proof of Lemma 3.1

The nuclear norm on the left is equal to $\mathrm{Tr}\left(\left(\mathbf{Y}^\top\mathbf{U}_1^\top\mathbf{U}_1\mathbf{Y}+\mathbf{V}^\top\mathbf{X}^\top\mathbf{U}_2^\top\mathbf{U}_2\mathbf{X}\mathbf{V}\right)^{1/2}\right)$ which yields $\mathrm{Tr}\left(\left(\mathbf{Y}^\top\mathbf{Y}+\mathbf{V}^\top\mathbf{X}^\top\mathbf{X}\mathbf{V}\right)^{1/2}\right)$. Through further manipulation we arrive at $\mathrm{Tr}\left(\left((\mathbf{V}\mathbf{Y}^\top\mathbf{Y}\mathbf{V}^\top+\mathbf{X}^\top\mathbf{X}\right)^{1/2}\right)$. The nuclear norm on the right is $\mathrm{Tr}\left(\left(\mathbf{V}\mathbf{Y}^\top\mathbf{Y}\mathbf{V}^\top+\mathbf{X}^\top\mathbf{X}\right)^{1/2}\right)$ and thus equal to that on the left.

B.2 Proof of Theorem 3.2

We prove the case where $d=c$. The other cases follow from this one. Invoking Lemma 3.1, let us remove the gauge freedom and rewrite the problem as

$$\min_{\mathbf{V}_X\in\mathrm{St}_c(\mathbb{R}^n)}\left\|\begin{bmatrix}\boldsymbol{\Sigma}_Y\mathbf{V}_Y^\top\\ \boldsymbol{\Sigma}_X\mathbf{V}_X^\top\end{bmatrix}\right\|_*. \tag{15}$$

Let $f(\mathbf{V}_X):=\left\|\begin{bmatrix}\boldsymbol{\Sigma}_Y\mathbf{V}_Y^\top\\ \boldsymbol{\Sigma}_X\mathbf{V}_X^\top\end{bmatrix}\right\|_*$. Since $\mathbf{V}_Y\in\mathrm{St}_c(\mathbb{R}^n)$, then $\mathbf{V}_Y\mathbf{V}_Y^\top$ is an orthogonal projection and thus

$$f(\mathbf{V}_X)\geq\left\|\begin{bmatrix}\boldsymbol{\Sigma}_Y\mathbf{V}_Y^\top\\ \boldsymbol{\Sigma}_X\mathbf{V}_X^\top\end{bmatrix}\mathbf{V}_Y\mathbf{V}_Y^\top\right\|_*=\left\|\begin{bmatrix}\boldsymbol{\Sigma}_Y\\ \boldsymbol{\Sigma}_X\mathbf{V}_X^\top\mathbf{V}_Y\end{bmatrix}\mathbf{V}_Y^\top\right\|_*=\left\|\begin{bmatrix}\boldsymbol{\Sigma}_Y\\ \boldsymbol{\Sigma}_X\mathbf{V}_X^\top\mathbf{V}_Y\end{bmatrix}\right\|_*. \tag{16}$$

Let $\mathbf{H}\in\mathbb{R}^{c\times c}=\mathbf{V}_X^\top\mathbf{V}_Y$ be the misalignment between $\mathbf{V}_X$ and $\mathbf{V}_Y$ and define the lower bound

$$g(\mathbf{H})=\left\|\begin{bmatrix}\boldsymbol{\Sigma}_Y\\ \boldsymbol{\Sigma}_X\mathbf{H}\end{bmatrix}\right\|_*, \tag{17}$$

which verifies $\max_{\|\mathbf{H}\|_2\leq 1}g(\mathbf{H})\leq\min_{\mathbf{V}_X\in\mathrm{St}_c(\mathbb{R})^n}f(\mathbf{V}_X)$. Noting that $g(\mathbf{H})$ can be written as

$$g(\mathbf{H})=\mathrm{Tr}\left(\sqrt{\boldsymbol{\Sigma}_Y^2+\mathbf{H}^\top\boldsymbol{\Sigma}_X^2\mathbf{H}}\right), \tag{18}$$

we have that it attains its maximum for $\mathbf{H}=\mathbf{I}_c$. We can verify that f attains this lower bound for $\mathbf{V}_X=\mathbf{V}_Y$ *i.e.*,

$$f(\mathbf{V}_Y)=\left\|\begin{bmatrix}\boldsymbol{\Sigma}_Y\mathbf{V}_Y^\top\\ \boldsymbol{\Sigma}_X\mathbf{V}_Y^\top\end{bmatrix}\right\|_*=\left\|\begin{bmatrix}\boldsymbol{\Sigma}_Y\\ \boldsymbol{\Sigma}_X\end{bmatrix}\right\|_*=\mathrm{Tr}\left(\sqrt{\boldsymbol{\Sigma}_Y^2+\boldsymbol{\Sigma}_X^2}\right). \tag{19}$$

B.3 Proof of Theorem 3.3

Writing the SVD of $\mathbf{X}$ as $\mathbf{U}_X\boldsymbol{\Sigma}_X\mathbf{V}_X^\top$, where $\mathbf{U}_X\in O(d)$, $\boldsymbol{\Sigma}_X$ is diagonal and $\mathbf{V}_X\in\mathrm{St}_d(\mathbb{R}^n)$, the terms $\|\mathbf{X}\|_*$ and $\|\mathbf{X}\|_2$ only depend on $\boldsymbol{\Sigma}_X$. Let us rewrite the loss as

$$\left\|\begin{bmatrix}\mathbf{Y}\\ \mathbf{X}\end{bmatrix}\right\|_*-\alpha\|\boldsymbol{\Sigma}_X\|_*+\beta\|\boldsymbol{\Sigma}_X\|_2^2=\left\|\begin{bmatrix}\mathbf{U}_Y\boldsymbol{\Sigma}_Y\mathbf{V}_Y^\top\\ \mathbf{U}_X\boldsymbol{\Sigma}_X\mathbf{V}_X^\top\end{bmatrix}\right\|_*-\alpha\mathrm{Tr}(\boldsymbol{\Sigma}_X)+\beta\|\boldsymbol{\Sigma}_X\|_2^2. \tag{20}$$

Using Lemma 3.1, we have the equivalent problem

$$\min_{\boldsymbol{\Sigma}_X\succeq 0,\mathbf{V}_X\in\mathrm{St}_d(\mathbb{R}^n)}\left\{\left\|\begin{bmatrix}\boldsymbol{\Sigma}_Y\mathbf{V}_Y^\top\\ \boldsymbol{\Sigma}_X\mathbf{V}_X^\top\end{bmatrix}\right\|_*-\alpha\mathrm{Tr}(\boldsymbol{\Sigma}_X)+\beta\|\boldsymbol{\Sigma}_X\|_2^2\right\} \tag{21}$$

or similarly,

$$\min_{\boldsymbol{\Sigma}_X\succeq 0}\left\{\min_{\mathbf{V}_X\in\mathrm{St}_d(\mathbb{R}^n)}\left\{\left\|\begin{bmatrix}\boldsymbol{\Sigma}_Y\mathbf{V}_Y^\top\\ \boldsymbol{\Sigma}_X\mathbf{V}_X^\top\end{bmatrix}\right\|_*\right\}-\alpha\mathrm{Tr}(\boldsymbol{\Sigma}_X)+\beta\|\boldsymbol{\Sigma}_X\|_2^2\right\}. \tag{22}$$

Denote the singular values of $\mathbf{Y}$ by $\mu_1\geq\cdots\geq\mu_c$ and those of $\mathbf{X}$ by $\sigma_1\geq\cdots\geq\sigma_d$. According to Theorem 3.2, the inner minimization yields $\sum_{i=1}^c\sqrt{\mu_i^2+\sigma_i^2}+\sum_{i=c+1}^d\sigma_i$ for $\mathbf{V}_X=\begin{bmatrix}\mathbf{V}_Y & \mathbf{V}\end{bmatrix}$. Thus,

$$\min_{\sigma_1\geq\cdots\geq\sigma_d\geq 0}\left\{\sum_{i=1}^c\sqrt{\mu_i^2+\sigma_i^2}+\sum_{i=c+1}^d\sigma_i-\alpha\sum_{i=1}^d\sigma_i+\beta\sigma_1^2\right\}. \tag{23}$$

This problem is convex. Let us drop the constraint and consider the relaxed problem

$$\begin{aligned}\min_{\sigma,t}\ & \beta t^2 + \sum_{i=1}^{c}\left(\sqrt{\mu_i^2+\sigma_i^2} - \alpha\sigma_i\right) + (1-\alpha)\sum_{i=c+1}^{d}\sigma_i \\ \text{s.t. }\ & t - \sigma_i \geq 0, \quad i \in [c] \\ & \sigma_i \geq 0, \quad i \in \{c+1,\dots,d\}\end{aligned} \tag{24}$$

For dual variables $\lambda_i \geq 0$, $i \in [c]$ and $\nu_i \geq 0, i \in \{c+1,\dots,d\}$ we have the Lagrangian

$$L(\sigma,t;\lambda,\nu) = \sum_{i=1}^{c}\left(\sqrt{\mu_i^2+\sigma_i^2} + (\lambda_i-\alpha)\sigma_i\right) + \beta t^2 - \sum_{i=1}^{c}\lambda_i t + \sum_{i=c+1}^{d}\left((1-\alpha)\sigma_i - \nu_i\sigma_i\right). \tag{25}$$

The KKT conditions sufficient for the optimality of the relaxation are thus

$$\begin{cases} t - \sigma_i \geq 0,\ i \in [c] & \text{(a - primal feasibility)} \\ \sigma_i \geq 0,\ i \in \{c+1,\dots,d\} & \text{(b - primal feasibility)} \\ \lambda_i \geq 0,\ i \in [c] & \text{(c - dual feasibility)} \\ \nu_i \geq 0,\ i \in \{c+1,\dots,d\} & \text{(d - dual feasibility)} \\ \frac{\sigma_i}{\sqrt{\sigma_i^2+\mu_i^2}} + \lambda_i - \alpha = 0,\ i \in [c] & \text{(e - stationarity)} \\ \nu_i = 1-\alpha,\ i \in \{c+1,\dots,d\} & \text{(f - stationarity)} \\ t = \frac{1}{2\beta}\sum_{i=1}^{c}\lambda_i & \text{(g - stationarity)} \\ \sum_{i=1}^{c}\lambda_i(t-\sigma_i) = 0 & \text{(h - complementary slackness)} \\ \sum_{i=c+1}^{d}\nu_i\sigma_i = 0 & \text{(i - complementary slackness).} \end{cases} \tag{26}$$

Let $\sigma_i^* = t^*,\ i \in [c]$ and $\sigma_i^* = 0,\ i \in \{c+1,\dots,d\}$. Primal feasibility (a), (b) and complementary slackness (h), (i) hold. Since $\alpha < 1$, we have $\nu_i^* = 1-\alpha \geq 0,\ i \in \{c+1,\dots,d\}$. Thus, dual feasibility (d) and stationarity (f) hold. Plugging $\sigma_i^* = t^*,\ i \in [c]$ in the stationarity conditions (e) and (g) yields

$$\begin{cases} \frac{t^*}{\sqrt{t^{*2}+\mu_i^2}} + \lambda_i^* - \alpha = 0,\ i \in [c] \\ t^* = \frac{1}{2\beta}\sum_{i=1}^{c}\lambda_i^*. \end{cases} \tag{27}$$

Solving for t^*, we obtain $\alpha c - 2\beta t^* = \sum_{i=1}^{c}\frac{t^*}{\sqrt{\mu_i^2+t^{*2}}}$ and thus $0 < t^* \leq \frac{\alpha c}{2\beta}$. It suffices to verify dual feasibility (c). Plugging this upper bound in the stationarity condition (e) we obtain

$$\lambda_i^* = \alpha - \frac{t^*}{\sqrt{\mu_i^2+t^{*2}}} \geq \alpha - \frac{\alpha c}{2\beta\sqrt{\mu_c^2+(\alpha c/2\beta)^2}} \geq 0. \tag{28}$$

We can thus find α that ensures dual feasibility (c), which yields $\alpha \geq \sqrt{\max\left\{0, 1-\frac{4\beta^2\mu_c^2}{c^2}\right\}}$. All KKT conditions for the relaxation hold. To show that this solution solves the original problem, note that $\sigma_1^* = \dots = \sigma_c^* = t^* > 0$ and $\sigma_{c+1}^* = \dots \sigma_d^* = 0$ imply $\sigma_1^* \geq \dots \geq \sigma_d^* \geq 0$.

Concluding the proof, the solution set is

$$\mathcal{X} = \left\{ \mathbf{U}(t^*\mathbf{I}_c)\begin{bmatrix}\mathbf{V}_Y^\top \\ \mathbf{V}^\top\end{bmatrix} \,\middle|\, \mathbf{U} \in \mathrm{St}_c(\mathbb{R}^d),\ \mathbf{V} \in \mathcal{N}(\mathbf{V}_Y) \right\}, \tag{29}$$

where t^* is the solution to $\alpha c - 2\beta t = \sum_{i=1}^{c}\frac{t}{\sqrt{\mu_i^2+t^2}}$.

B.4 Proof of Lemma 3.5

Proof. Writing the SVD of $\mathbf{X}$ as $\mathbf{U}_X\mathbf{\Sigma}_X\mathbf{V}_X^\top$, where $\mathbf{U}_X \in O(d)$, $\mathbf{\Sigma}_X$ is diagonal and $\mathbf{V}_X \in \mathrm{St}_d(\mathbb{R}^n)$, the terms $\|\mathbf{X}\|_*$ and $\|\mathbf{X}\|_2$ only depend on $\mathbf{\Sigma}_X$. Let us rewrite the loss as

$$\left\|\begin{bmatrix}\mathbf{Y} \\ \mathbf{X}\end{bmatrix}\right\|_* - \|\mathbf{\Sigma}_X\|_* + \beta\|\mathbf{\Sigma}_X\|_2 = \left\|\begin{bmatrix}\mathbf{U}_Y\mathbf{\Sigma}_Y\mathbf{V}_Y^\top \\ \mathbf{U}_X\mathbf{\Sigma}_X\mathbf{V}_X^\top\end{bmatrix}\right\|_* - \mathrm{Tr}(\mathbf{\Sigma}_X) + \beta\|\mathbf{\Sigma}_X\|_2. \tag{30}$$

Using Lemma 3.1, we have

$$\min_{\mathbf{\Sigma}_X \succeq 0, \mathbf{V}_X \in \mathrm{St}_d(\mathbb{R}^n)} \left\{ \left\| \begin{bmatrix} \mathbf{\Sigma}_Y \mathbf{V}_Y^\top \\ \mathbf{\Sigma}_X \mathbf{V}_X^\top \end{bmatrix} \right\|_* - \mathrm{Tr}(\mathbf{\Sigma}_X) + \beta \|\mathbf{\Sigma}_X\|_2 \right\} \tag{31}$$

or equivalently,

$$\min_{\mathbf{\Sigma}_X \succeq 0} \left\{ \min_{\mathbf{V}_X \in \mathrm{St}_d(\mathbb{R}^n)} \left\| \begin{bmatrix} \mathbf{\Sigma}_Y \mathbf{V}_Y^\top \\ \mathbf{\Sigma}_X \mathbf{V}_X^\top \end{bmatrix} \right\|_* - \mathrm{Tr}(\mathbf{\Sigma}_X) + \beta \|\mathbf{\Sigma}_X\|_2 \right\}. \tag{32}$$

Denoting the singular values of $\mathbf{Y}$ by $\mu_1 \geq \cdots \geq \mu_c$ and those of $\mathbf{X}$ by $\sigma_1 \geq \cdots \geq \sigma_d$ then, according to Theorem 3.2, the inner minimization yields $\sum_{i=1}^{c} \sqrt{\mu_i^2 + \sigma_i^2} + \sum_{i=c+1}^{d} \sigma_i$. Thus,

$$\begin{aligned} &\min_{\sigma_1 \geq \cdots \geq \sigma_d \geq 0} \left\{ \sum_{i=1}^{c} \sqrt{\mu_i^2 + \sigma_i^2} + \sum_{i=c+1}^{d} \sigma_i - \sum_{i=1}^{d} \sigma_i + \beta \sigma_1 \right\} \\ = &\min_{\sigma_1 \geq \cdots \geq \sigma_c \geq 0} \left\{ \sum_{i=1}^{c} \left(\sqrt{\mu_i^2 + \sigma_i^2} - \sigma_i \right) + \beta \sigma_1 \right\}. \end{aligned} \tag{33}$$

This problem is convex and independent of the values of σ_i for $i > c$. Let us relax the constraint $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$ and consider the problem

$$\begin{aligned} &\min_{\sigma, t} \ \beta t + \sum_{i=1}^{c} \left(\sqrt{\mu_i^2 + \sigma_i^2} - \sigma_i \right) \\ &\text{s.t.} \ \ t - \sigma_i \geq 0, \quad i \in [c]. \end{aligned} \tag{34}$$

For the dual variables $\lambda_i \geq 0$, $i \in [c]$, we have the Lagrangian

$$L(\sigma, t; \lambda) = \beta t + \sum_{i=1}^{c} \left(\sqrt{\mu_i^2 + \sigma_i^2} - \sigma_i - \lambda_i (t - \sigma_i) \right), \tag{35}$$

with KKT conditions

$$\begin{cases} t - \sigma_i \geq 0, \ i \in [c] & \text{(a - primal feasibility)} \\ \lambda_i \geq 0, \ i \in [c] & \text{(b - dual feasibility)} \\ \frac{\sigma_i}{\sqrt{\mu_i^2 + \sigma_i^2}} + \lambda_i - 1 = 0, \ i \in [c] & \text{(c - stationarity)} \\ \sum_{i=1}^{c} \lambda_i = \beta & \text{(d - stationarity)} \\ \sum_{i=1}^{c} \lambda_i (t - \sigma_i) = 0 & \text{(e - complementary slackness).} \end{cases} \tag{36}$$

Let $\sigma_i^* = t^*$, $i \in [c]$. Then, primal feasibility (a) and complementary slackness (e) hold. In order to have stationarity (c), we set

$$\lambda_i^* = 1 - \frac{t^*}{\sqrt{\mu_i^2 + t^{*2}}}, \ i \in [c] \tag{37}$$

and thus $\lambda_i^* \geq 0$, $i \in [c]$ and dual feasibility holds. Solving (d) yields

$$\sum_{i=1}^{c} \frac{t^*}{\sqrt{\mu_i^2 + t^{*2}}} = c - \beta, \tag{38}$$

from where we find that $t^* > 0$. Thus, the solution of the relaxed problem $\sigma_i^* = t^*$, for $i \in [c]$ is also feasible for the original problem, provided we pick $\sigma_{c+1}, \ldots, \sigma_d$ such that $t^* \geq \sigma_{c+1} \geq \cdots \geq \sigma_d \geq 0$. Concluding the proof, the solution set is

$$\mathcal{X} = \left\{ \mathbf{U} \begin{bmatrix} t^* \mathbf{I}_c & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma} \end{bmatrix} \begin{bmatrix} \mathbf{V}_Y^\top \\ \mathbf{V}^\top \end{bmatrix} \ \middle| \ \mathbf{U} \in O(d), \ t^* \mathbf{I} \succeq \mathbf{\Sigma}, \ \mathbf{V} \in \mathcal{N}(\mathbf{V}_Y) \right\}, \tag{39}$$

where t^* is the solution to $\sum_{i=1}^{c} \frac{t}{\sqrt{\mu_i^2 + t^2}} = c - \beta$. □

B.5 Proof of Lemma 3.4

If we write the SVD of $\mathbf{X}$ as $\mathbf{U}_X\mathbf{\Sigma}_X\mathbf{V}_X^\top$, where $\mathbf{U}_X \in O(d)$, $\mathbf{\Sigma}_X$ is diagonal and $\mathbf{V}_X \in \mathrm{St}_d(\mathbb{R}^n)$ the term $\|\mathbf{X}\|_*$ is simply the trace of $\mathbf{\Sigma}_X$. Let us rewrite the loss as

$$\left\|\begin{bmatrix}\mathbf{Y}\\ \mathbf{X}\end{bmatrix}\right\|_* - \alpha\|\mathbf{\Sigma}_X\|_* = \left\|\begin{bmatrix}\mathbf{U}_Y\mathbf{\Sigma}_Y\mathbf{V}_Y^\top\\ \mathbf{U}_X\mathbf{\Sigma}_X\mathbf{V}_X^\top\end{bmatrix}\right\|_* - \alpha\mathrm{Tr}(\mathbf{\Sigma}_X). \tag{40}$$

Using Lemma 3.1, we have

$$\min_{\mathbf{\Sigma}_X\succeq 0, \mathbf{V}_X\in\mathrm{St}_d(\mathbb{R}^n)} \left\|\begin{bmatrix}\mathbf{\Sigma}_Y\mathbf{V}_Y^\top\\ \mathbf{\Sigma}_X\mathbf{V}_X^\top\end{bmatrix}\right\|_* - \alpha\mathrm{Tr}(\mathbf{\Sigma}_X) \tag{41}$$

or equivalently,

$$\min_{\mathbf{\Sigma}_X\succeq 0}\left\{\min_{\mathbf{V}_X\in\mathrm{St}_d(\mathbb{R}^n)}\left\|\begin{bmatrix}\mathbf{\Sigma}_Y\mathbf{V}_Y^\top\\ \mathbf{\Sigma}_X\mathbf{V}_X^\top\end{bmatrix}\right\|_* - \alpha\mathrm{Tr}(\mathbf{\Sigma}_X)\right\}. \tag{42}$$

Denote the singular values of $\mathbf{Y}$ by $\mu_1 \geq \cdots \geq \mu_c$ and those of $\mathbf{X}$ by $\sigma_1 \geq \cdots \geq \sigma_d$. According to Theorem 3.2, the inner minimization yields $\sum_{i=1}^c \sqrt{\mu_i^2+\sigma_i^2} + \sum_{i=c+1}^d \sigma_i$. Thus,

$$\min_{\sigma_1\geq\cdots\geq\sigma_d\geq 0}\left\{\sum_{i=1}^c\sqrt{\mu_i^2+\sigma_i^2} + \sum_{i=c+1}^d\sigma_i - \alpha\sum_{i=1}^d\sigma_i\right\}. \tag{43}$$

This problem is convex. Let us relax the constraints $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$, replacing them by $\sigma_i \geq 0, i \in [d]$. In this relaxation, the derivative w.r.t. σ_k yields $\frac{\sigma_k}{\sqrt{\mu_k+\sigma_k^2}} - \alpha$ for $k \in [c]$. The first-order stationarity condition puts thus the optimum at $\sigma_i^* = \frac{\alpha}{\sqrt{1-\alpha^2}}\mu_i$ for $i \in [c]$. For $k > c$, the derivative w.r.t. σ_k yields $1-\alpha > 0$. Given the relaxed constraint $\sigma_i \geq 0, i \in [d]$, the minimum is therefore attained for $\sigma_i^* = 0$, $i \in \{c+1,\ldots,d\}$. Note that

$$\sigma_i^* = \begin{cases}\frac{\alpha}{\sqrt{1-\alpha^2}}\mu_i, & i \in [c]\\ 0, & i \in \{c+1,\ldots,d\}\end{cases} \tag{44}$$

is feasible for the original problem and thus optimal as well. The solution set is thus given by

$$\mathcal{X} = \left\{\mathbf{U}\begin{bmatrix}\frac{\alpha}{\sqrt{1-\alpha^2}}\mathbf{\Sigma}_Y & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}_Y^\top\\ \mathbf{V}^\top\end{bmatrix} \;\middle|\; \mathbf{U}\in O(d),\ \mathbf{V}\in\mathcal{N}(\mathbf{V}_Y)\right\}. \tag{45}$$

B.6 Proof of Lemma 4.1

We start by showing that $\forall i,j \in [n]\ \mathbf{y}_i = \mathbf{y}_j \Leftrightarrow \mathbf{v}_i = \mathbf{v}_j$. Each column of $\mathbf{\Sigma}\mathbf{V}^\top$ can be written as an orthogonal transformation of each column of $\mathbf{Y}$ by $\mathbf{U}^\top \in O(c)$. Therefore, $\mathbf{y}_i = \mathbf{y}_j \Leftrightarrow \mathbf{U}^\top\mathbf{y}_i = \mathbf{U}^\top\mathbf{y}_j \Leftrightarrow \mathbf{\Sigma}\mathbf{v}_i = \mathbf{\Sigma}\mathbf{v}_j$. Since $\mathbf{\Sigma}$ is square and full-rank, $\mathbf{\Sigma}\mathbf{v}_i = \mathbf{\Sigma}\mathbf{v}_j \Leftrightarrow \mathbf{v}_i = \mathbf{v}_j$. Together with the definition of $\mathcal{I}$, this implies that $\forall i,j \in \mathcal{I}$ $\mathbf{v}_i \neq \mathbf{v}_j$. For the second part, since $\mathbf{V} \in \mathrm{St}_c(\mathbb{R}^n)$ we have $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_c$. Thus, $\forall \mathbf{x} \in \mathbb{R}^c$, $\mathbf{V}^\top\mathbf{V}\mathbf{x} = \mathbf{x}$. For $i \in \mathcal{I}$, define the sets $\mathcal{J}_i = \{j \in [n] : \mathbf{y}_j = \mathbf{y}_i\}$ and note that $\bigcup_{i\in\mathcal{I}}\mathcal{J}_i = [n]$. We can write $\mathbf{V}^\top\mathbf{V}\mathbf{x}$ as

$$\sum_{i=1}^n \mathbf{v}_i\langle\mathbf{v}_i,\mathbf{x}\rangle = \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}_i}\mathbf{v}_j\langle\mathbf{v}_j,\mathbf{x}\rangle = \sum_{i\in\mathcal{I}}|\mathcal{J}_i|\mathbf{v}_i\langle\mathbf{v}_i,\mathbf{x}\rangle = \sum_{i\in\mathcal{I}}\sqrt{|\mathcal{J}_i|}\mathbf{v}_i\langle\sqrt{|\mathcal{J}_i|}\mathbf{v}_i,\mathbf{x}\rangle, \tag{46}$$

where we used the fact that, by definition, for any two indices $l,k \in \mathcal{J}_i$ we must have $\mathbf{v}_l = \mathbf{v}_k = \mathbf{v}_i$. We can write (46) in matrix notation by defining the matrix $\bar{\mathbf{V}}$ with columns $\{\sqrt{|\mathcal{J}_i|}\mathbf{v}_i\}_{i\in\mathcal{I}}$. The identity $\forall\mathbf{x}\in\mathbb{R}^c\ \mathbf{V}^\top\mathbf{V}\mathbf{x} = \mathbf{x}$ becomes $\bar{\mathbf{V}}^\top\bar{\mathbf{V}}\mathbf{x} = \mathbf{x}$. From the hypothesis that $|\mathcal{I}| = \mathrm{rank}\ \mathbf{Y} = c$, $\bar{\mathbf{V}}$ is square and full-rank. Thus, $\forall\mathbf{x}\in\mathbb{R}^c\ \bar{\mathbf{V}}^\top\bar{\mathbf{V}}\mathbf{x} = \mathbf{x}$ implies that $\bar{\mathbf{V}}^\top\bar{\mathbf{V}} = \mathbf{I}_c$ and we have $\bar{\mathbf{V}} \in O(c)$. Therefore, the columns of $\bar{\mathbf{V}}$, $\{\sqrt{|\mathcal{J}_i|}\mathbf{v}_i\}_{i\in\mathcal{I}}$, form an orthonormal basis for $\mathbb{R}^c$.

B.7 Proof of Lemma 4.2

Since the minterms $\{\mathbf{y}_i\}_{i\in\mathcal{I}}$ correspond to disjoint events, the posterior is given by the probability of the disjunction of all $\mathbf{y}_i$ that imply $\mathbf{q}$. Thus,

$$\begin{aligned} P(\mathbf{q}|\mathbf{x}) &= P\left(\bigvee_{i\in\mathcal{I}\text{ st }\mathbf{y}_i \implies \mathbf{q}} \mathbf{y}_i \middle| \mathbf{x}\right) \overset{(a)}{=} \sum_{i\in\mathcal{I}\text{ st }\mathbf{y}_i \implies \mathbf{q}} P(\mathbf{y}_i|\mathbf{x}) \overset{(b)}{=} \sum_{i\in\mathcal{I}\text{ st }\mathbf{y}_i \implies \mathbf{q}} \left\langle \mathbf{x}^\top \mathbf{e}_i, \mathbf{e}_i^\top \mathbf{x}\right\rangle \\ &= \left\langle \mathbf{x}, \left(\sum_{i\in\mathcal{I}\text{ st }\mathbf{y}_i \implies \mathbf{q}} \mathbf{e}_i\mathbf{e}_i^\top\right)\mathbf{x}\right\rangle \overset{(c)}{=} \langle \mathbf{x}, \mathbf{P}_q\mathbf{x}\rangle, \end{aligned} \tag{47}$$

(a) $\{\mathbf{y}_i\}_{i\in\mathcal{I}}$ correspond to disjoint events; (b) definition of $P(\mathbf{y}_i|\mathbf{x})$; (c) $\mathbf{P}_q$ is a projection operator onto the subspace spanned by $\{\mathbf{e}_i\}_{i\in\mathcal{I}\text{ st }\mathbf{y}_i \implies \mathbf{q}}$.

B.8 Additional theoretical results

If we fix the singular values of $\mathbf{X}$, the nuclear norm of $\mathbf{Z}$ can easily be upper bounded. This stems from the triangle inequality

$$\left\|\begin{bmatrix}\mathbf{Y}\\ \mathbf{X}\end{bmatrix}\right\|_* \leq \left\|\begin{bmatrix}\mathbf{Y}\\ \mathbf{0}_{d\times n}\end{bmatrix}\right\|_* + \left\|\begin{bmatrix}\mathbf{0}_{c\times n}\\ \mathbf{X}\end{bmatrix}\right\|_* \tag{48}$$

being tight if we pick $\mathbf{X}$ such that $\mathbf{Y}\mathbf{X}^\top = \mathbf{Y}^\top\mathbf{X} = 0$ (the nuclear norm is additive if the matrices have orthogonal row and column spaces [Recht et al., 2010]). Hence if $\mathbf{\Sigma}_Y$ denotes the singular values of $\mathbf{Y}$

$$\max_{\mathbf{V}_X\in\mathrm{St}_d(\mathbb{R}^d)} \left\|\begin{bmatrix}\mathbf{Y}\\ \mathbf{U}_X\mathbf{\Sigma}_X\mathbf{V}_X^\top\end{bmatrix}\right\|_* = \mathrm{Tr}\left(\mathbf{\Sigma}_Y + \mathbf{\Sigma}_X\right) \tag{49}$$

Lemma B.1. *Let $\mathbf{X}\in\mathbb{R}^{d\times n}$, with $n > d$, be a matrix with unit ℓ_2-norm columns. Then* $\max_{\mathbf{X}} \|\mathbf{X}\|_* = d\sqrt{n/d}$ *and* $\mathrm{argmax}_{\mathbf{X}} \|\mathbf{X}\|_* = \sqrt{n/d}\,\mathbf{U}\mathbf{V}^\top$, *for some $\mathbf{U}\in O(d)$ and $\mathbf{V}\in\mathrm{St}_d(\mathbb{R}^n)$.*

Proof. Let the SVD be $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. Then, $\mathbf{X}^\top\mathbf{X} = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^\top$ and $\mathrm{Tr}(\mathbf{X}^\top\mathbf{X}) = \sum_{i=1}^d \sigma_i^2$. From the normalization, it follows that $\mathrm{Tr}(\mathbf{X}^\top\mathbf{X}) = \sum_{i=1}^n x_i^\top x_i = n$. The maximization of $\|\mathbf{X}\|_*$ corresponds thus to the optimization problem

$$\begin{aligned} &\max_{\{\sigma_i\}_{i=1}^d} \sum_{i=1}^d \sigma_i \\ &\text{s.t.} \sum_{i=1}^d \sigma_i^2 = n. \end{aligned} \tag{50}$$

The gradient of the cost function is $\mathbf{1}\in\mathbb{R}^d$ and his orthogonal to the feasible set at $\sigma_i = \sqrt{\frac{n}{d}}$, $i = 1,\ldots,d$. Thus, $\max_{\mathbf{X}} \|\mathbf{X}\|_* = d\sqrt{n/d}$. □

Lemma B.2. *For any $\mathbf{X}\in\mathbb{R}^{n\times n}$ with singular values $\sigma_1 \geq \cdots \geq \sigma_n > 0$,*

$$\mathrm{spectrum}\left(\begin{bmatrix}\mathbf{0}_n & \mathbf{X}^\top\\ \mathbf{X} & \mathbf{0}_n\end{bmatrix}\right) = \bigcup_{i=1}^n \{-\sigma_i, \sigma_i\} \tag{51}$$

Proof. Denoting the SVD of $\mathbf{X}$ by $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, it suffices to check that

$$\begin{bmatrix}\mathbf{0}_n & \mathbf{X}^\top\\ \mathbf{X} & \mathbf{0}_n\end{bmatrix} = \begin{bmatrix}\mathbf{V}/\sqrt{2} & -\mathbf{V}/\sqrt{2}\\ \mathbf{U}/\sqrt{2} & \mathbf{U}/\sqrt{2}\end{bmatrix}\begin{bmatrix}\mathbf{\Sigma} & \mathbf{0}_n\\ \mathbf{0}_n & -\mathbf{\Sigma}\end{bmatrix}\begin{bmatrix}\mathbf{V}^\top/\sqrt{2} & \mathbf{U}^\top/\sqrt{2}\\ -\mathbf{V}^\top/\sqrt{2} & \mathbf{U}^\top/\sqrt{2}\end{bmatrix} \tag{52}$$

is the spectral decomposition of the symmetric matrix with $\mathbf{X}$ and $\mathbf{X}^\top$ in the off-diagonals. □

Lemma B.3. *Let $\mathbf{V}_Y, \mathbf{V}_X \in \mathrm{St}_c(\mathbb{R}^n)$, for $n > c$, and denote by $\{\sigma_i\}_{i=1}^{2c}$ the singular values of $\mathbf{Z} \in \mathbb{R}^{2c \times n}$, defined as*

$$\mathbf{Z} := \begin{bmatrix} \mathbf{V}_Y^\top \\ \mathbf{V}_X^\top \end{bmatrix}. \tag{53}$$

Then, $\{\sigma_i\}_{i=1}^{2c} = \bigcup_{i=1}^{c} \left\{ \sqrt{1 + \sigma_i(\mathbf{V}_Y^\top \mathbf{V}_X)}, \sqrt{1 - \sigma_i(\mathbf{V}_Y^\top \mathbf{V}_X)} \right\}$.

Proof. Letting the full SVD of $\mathbf{Z}$ be $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, we have $\mathbf{Z}\mathbf{Z}^\top = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\top$. Thus, its singular values are given by the square roots of the eigenvalues of $\mathbf{Z}\mathbf{Z}^\top$. From $\mathbf{V}_Y^\top \mathbf{V}_Y = \mathbf{V}_X^\top \mathbf{V}_X = \mathbf{I}_c$ we have

$$\sqrt{\lambda_i\left(\mathbf{Z}\mathbf{Z}^\top\right)} = \sqrt{\lambda_i\left(\begin{bmatrix} \mathbf{I}_c & \mathbf{V}_Y^\top \mathbf{V}_X \\ \mathbf{V}_X^\top \mathbf{V}_Y & \mathbf{I}_c \end{bmatrix}\right)} = \sqrt{1 + \lambda_i\left(\begin{bmatrix} \mathbf{0}_c & \mathbf{V}_Y^\top \mathbf{V}_X \\ \mathbf{V}_X^\top \mathbf{V}_Y & \mathbf{0}_c \end{bmatrix}\right)}. \tag{54}$$

Using Lemma B.2, this yields $\bigcup_{i=1}^{c} \left\{ \sqrt{1 + \sigma_i(\mathbf{V}_Y^\top \mathbf{V}_X)}, \sqrt{1 - \sigma_i(\mathbf{V}_Y^\top \mathbf{V}_X)} \right\}$. □

Lemma B.4. *For $\mathbf{V}_Y \in \mathrm{St}_c(\mathbb{R}^n)$, with $n > c$,*

$$\min_{\mathbf{V}_X \in \mathrm{St}_c(\mathbb{R}^n)} \left\| \begin{bmatrix} \mathbf{V}_Y^\top \\ \mathbf{V}_X^\top \end{bmatrix} \right\|_* = \left\| \begin{bmatrix} \mathbf{V}_Y^\top \\ \mathbf{V}_Y^\top \end{bmatrix} \right\|_* = \sqrt{2}c \tag{55}$$

Proof. From Lemma B.3 we can write:

$$\left\| \begin{bmatrix} \mathbf{V}_Y^\top \\ \mathbf{V}_X^\top \end{bmatrix} \right\|_* = \sum_{i=1}^{c} \sqrt{1 + \sigma_i(\mathbf{V}_Y^\top \mathbf{V}_X)} + \sqrt{1 - \sigma_i(\mathbf{V}_Y^\top \mathbf{V}_X)} \tag{56}$$

Note that $0 \leq \sigma_i(\mathbf{V}_Y^\top \mathbf{V}_X) \leq 1$ for $i \in [c]$, and

$$\sqrt{1 + \sigma_i} + \sqrt{1 - \sigma_i} \geq \sqrt{2}, \quad \sigma_i \in [0, 1]. \tag{57}$$

Therefore we have the lower bound

$$\left\| \begin{bmatrix} \mathbf{V}_Y^\top \\ \mathbf{V}_X^\top \end{bmatrix} \right\|_* \geq \sum_{i=1}^{c} \sqrt{2} = \sqrt{2}c. \tag{58}$$

This bound is tight for $\mathbf{V}_X = \mathbf{V}_Y$, since $\sigma_i(\mathbf{V}_Y^\top \mathbf{V}_Y) = \sigma_i(\mathbf{I}_c) = 1, i \in [c]$. □

C Experimental Details

C.1 PyTorch Implementation

Below we present an implementation of the loss function (2) in PyTorch, called `NuclearLoss`, as well as the function for computing the minterm directions entitled `compute_minterms_vec`.

```
class NuclearLoss(nn.Module):
    def __init__(self, alpha : float, beta : float):
        super(Nuclear, self).__init__()
        self.a, self.b = alpha, beta

    def forward(self, x : torch.Tensor, y : torch.Tensor):
        z = torch.cat((y.permute(1,0),
                       x.permute(1,0)), dim=0)

        x_s = torch.linalg.svdvals(x)
        z_s = torch.linalg.svdvals(z)
        loss = z_s.sum() - self.a * x_s.sum() + self.b * x_s.max()**2
        return loss
```

```
def compute_minterms_vec(x : torch.Tensor,
                         y : torch.Tensor,
                         minterms : torch.Tensor):
    minterms_vec = []
    for minterm in minterms:
        mask = (y == minterm).all(dim=-1)
        u, _, _ = torch.linalg.svd(x[mask, :].T)
        minterms_vec.append(u[:,:1].T)
    return torch.cat(minterms_vec, dim=0)
```

The PyTorch code for the image augmentations used with CIFAR-10 and CIFAR-100 is provided below.

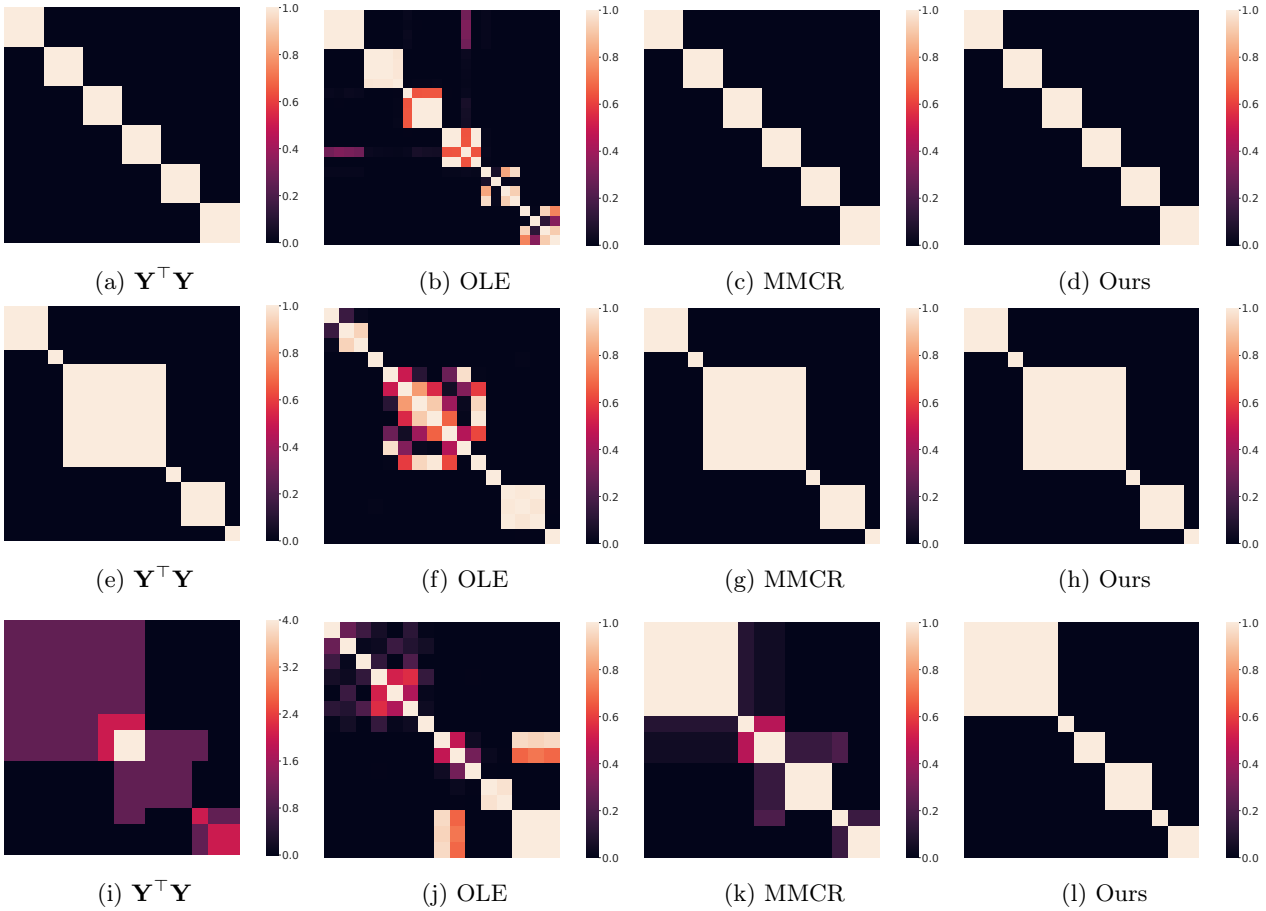
```
class CIFARTransform():
    def __init__(self, split):
        mu  = (0.4914, 0.4822, 0.4465)
        std = (0.2023, 0.1994, 0.2010)

        if split.lower() == "train":
            self.t = Compose([RandomCrop(32, padding=4),
                              RandomHorizontalFlip(),
                              ToTensor(),
                              Normalize(mu,std)])
        else:
            self.t = Compose([ToTensor(),
                              Normalize(mu,std)])
```

The PyTorch code for the image augmentations used with CelebA is provided below.

```
class CELEBATransform():
    def __init__(self, split):
        if split.lower() == "train":
            self.t = Compose([Resize((40, 40)),
                              RandomHorizontalFlip(),
                              ToTensor()])
        else:
            self.t = Compose([Resize((40, 40)),
                              ToTensor()])
```

C.2 Synthetic Experiments

We present in Fig. 5 additional comparisons between our approach, OLE, and MMCR for three binary label matrices $\mathbf{Y}$. Figs. 5a, 5e and 5i show the Gram matrices of the labels $\mathbf{Y}^\top\mathbf{Y}$ while the remaining figures display the



(a) $\mathbf{Y}^\top\mathbf{Y}$ (b) OLE (c) MMCR (d) Ours

(e) $\mathbf{Y}^\top\mathbf{Y}$ (f) OLE (g) MMCR (h) Ours

(i) $\mathbf{Y}^\top\mathbf{Y}$ (j) OLE (k) MMCR (l) Ours

Figure 5: Gram matrix of $\mathbf{Y} \in \{0,1\}^{6\times 15}$ and of the representations optimized with OLE, MCMR and our loss, for three synthetic experiments

Gram matrices of the optimized representations. While MMCR and our approach show the same orthogonalizing behavior in the case of disjoint labels, results are drastically different when co-occurring labels are present. Our loss guarantees orthogonal minterms, thus generalizing the disjoint label setting.

C.3 Classification Experiments

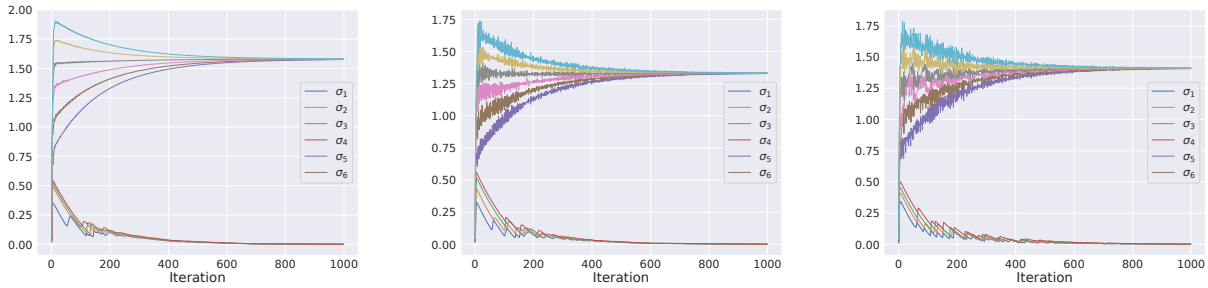
MNIST Dataset of 10,000 grayscale 32x32 images of 10 handwritten digits. 7,000 used for training and 3,000 for testing. No augmentations were used.

FashionMNIST Dataset of 10,000 grayscale 32x32 images of 10 fashion categories. 7,000 used for training and 3,000 for testing. No augmentations were used.

CIFAR-10 Dataset of 60,000 color 32x32 images corresponding to 10 semantic classes. Training set composed of 50,000 images and test set corresponding to the remaining 10,000. The training set was augmented with random 32x32 crops with 4 pixel padding and random horizontal flips.

CIFAR-100 Dataset of 60,000 color 32x32 images corresponding to 100 semantic classes. Training set composed of 50,000 images and test set corresponding to the remaining 10,000. The training set was augmented with random 32x32 crops with 4 pixel padding and random horizontal flips.

Training details for the cross-entropy baselines and for our model are provided in Tables 4 and 5, respectively.



(a) **Y** from Fig. 5a (b) **Y** from Fig. 5e (c) **Y** from Fig. 5i

Figure 6: Convergence of the singular values of **X** during minimization of (2).

Table 4: Standard classification training with **crossentropy loss**.

	MNIST		**FashionMNIST**		**CIFAR10**	**CIFAR100**
Backbone	ConvNet	ResNet-18	ConvNet	ResNet-18	ResNet-18	ResNet-18
Batch size	512	512	512	512	512	512
Epochs	20	20	50	50	200	200
Optimizer	Adam	Adam	Adam	Adam	SGD	SGD
LR	0.01	0.01	0.01	0.01	0.1	0.1
Scheduler	Step(1, 0.7)	Step(1, 0.7)	Step(4, 0.5)	Step(4, 0.5)	CosAnneal(200)	CosAnneal(200)
Accuracy	0.992	0.995	0.930	0.935	0.929	0.705

Table 5: Standard classification training with **our loss**.

	MNIST		**FashionMNIST**		**CIFAR10**	**CIFAR100**
Backbone	ConvNet	ResNet-18	ConvNet	ResNet-18	ResNet-18	ResNet-18
α	0.997	0.997	0.997	0.997	0.999	0.999
β	0.05	0.05	0.05	0.05	0.01	0.01
Batch size	512	512	512	512	512	512
Epochs	50	50	60	60	200	200
Optimizer	SGD	SGD	SGD	SGD	SGD	SGD
LR	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Scheduler	Step(20, 0.5)	Step(20, 0.5)	Step(30, 0.5)	Step(30, 0.5)	Step(100, 0.1)	Step(100, 0.7)
Accuracy	0.992	0.996	0.932	0.935	0.934	0.728

Table 6: Examples of propositional and natural language queries used in the Celeb-A retrieval experiment.

`Bald`	*a bald person*
`Male`	*a male person*
`Eyeglasses`	*a person with eyeglasses*
¬`Bald`	*a person that is not bald*
¬`Wearing Hat`	*a person not wearing a hat*
`Eyeglasses` ∧ `Male`	*a male person with eyeglasses*
`Bald` ∧ `Eyeglasses`	*a bald person with eyeglasses*
`Eyeglasses` ∧ ¬`Wearing Hat`	*a person with eyeglasses and not wearing a hat*
¬`Wearing Necktie` ∧ ¬`Wearing Hat`	*a person not wearing a necktie and not wearing a hat*
`Bald` ∧ `Eyeglasses` ∧ ¬`Wearing Necktie`	*a bald person with eyeglasses and not wearing a necktie*
¬`Wearing Necktie` ∧ ¬`Wearing Hat` ∧ `Male`	*a male person not wearing a necktie and not wearing a hat*
`Eyeglasses` ∧ `Wearing Necktie` ∧ ¬`Wearing Hat`	*a person with eyeglasses wearing a necktie and not wearing a hat*
¬`Bald` ∧ ¬`Eyeglasses` ∧ ¬`Wearing Necktie`	*a person that is not bald, not wearing eyeglasses and not wearing a necktie*
`Eyeglasses` ∧ ¬`Wearing Necktie` ∧ ¬`Wearing Hat` ∧ `Male`	*a male person with eyeglasses, not wearing a hat and not wearing a necktie*
`Eyeglasses` ∧ ¬`Wearing Necktie` ∧ ¬`Wearing Hat` ∧ ¬`Male`	*a person that is not male, with eyeglasses, not wearing a hat and not wearing a necktie*
¬`Bald` ∧ `Eyeglasses` ∧ `Wearing Necktie` ∧ ¬`Wearing Hat` ∧ `Male`	*a male person that is not bald, with eyeglasses, wearing a necktie and not wearing a hat*
¬`Bald` ∧ ¬`Eyeglasses` ∧ ¬`Wearing Necktie` ∧ ¬`Wearing Hat` ∧ `Male`	*a male person that is not bald, without eyeglasses, not wearing a hat and not wearing a necktie*
¬`Bald` ∧ ¬`Eyeglasses` ∧ ¬`Wearing Necktie` ∧ ¬`Wearing Hat` ∧ ¬`Male`	*a person that is not male, not bald, without eyeglasses, not wearing a hat and not wearing a necktie*

C.4 Retrieval Experiments

Table 6 contains additional examples of propositional queries and the corresponding natural language translations, used to query our model and CLIP, respectively.