Disentangled Representation Learning through Geometry Preservation with the Gromov-Monge Gap

Anonymous Authors[1]

Abstract

Learning disentangled representations in an unsupervised manner is a fundamental challenge with significant promise for improving generalization, interpretability, and fairness. While impossible in general, recent work has shown that unsupervised disentanglement is provably achievable under assumptions of certain geometrical constraints such as local isometry. Leveraging these insights, we propose a novel perspective on disentangled representation learning through the lens of quadratic optimal transport (OT). We formulate the OT problem in the Gromov-Monge setting to make the alignment of distributions in different spaces possible while preserving their intrinsic geometry. For this, we propose the Gromov-Monge-Gap (GMG), which regularizes a map to learn the most geometry-preserving mapping satisfying a fixed transportation constraint. We demonstrate its effectiveness for disentanglement on four standard benchmarks. Moreover, we show that geometry preservation can even encourage unsupervised disentanglement without the standard reconstruction objective - making the underlying model decoder-free, and promising a more practically viable and scalable perspective on disentanglement.

1. Introduction

Learning low-dimensional representations of high-dimensional data is a fundamental challenge in unsupervised deep learning (Bengio et al., 2014; Locatello et al., 2019b). Emphasis is put on learning robustly generalizing representations that allow for efficient adaptation across a wide range of tasks (Bengio et al., 2014; Higgins et al., 2018; Locatello et al., 2019b). Disentanglement (Bengio et al., 2014; Higgins et al., 2017; 2018; Locatello et al., 2019b; Roth et al., 2023) has shown significant promise in facilitating such generalization (Bengio et al., 2014; Higgins et al., 2017; 2018; Locatello et al., 2019b; 2020; Horan et al., 2021; Roth et al., 2023; Hsu et al., 2023; Barin-Pacela et al., 2024), alongside interpretability and fairness (Locatello et al., 2019a; Träuble et al., 2021). Most works (Bengio et al., 2014; Higgins et al., 2017; Kim and Mnih, 2018; Chen et al., 2018; Locatello et al., 2019b; Roth et al., 2023) regard disentanglement as a one-to-one map between learned representations and ground-truth latent factors, effectively seeking to recover these factors from data alone in an unsupervised fashion.

While unsupervised disentanglement is theoretically impossible (Locatello et al., 2019b), the inductive biases of autoencoder architectures ensure effective disentanglement in practice (Rolinek et al., 2019; Zietlow et al., 2021; Horan et al., 2021). Most approaches operate on Variational autoencoder (VAE) frameworks (Kingma and Welling, 2014), using objectives that match latent VAE posteriors to factorized priors (Higgins et al., 2017; Kim and Mnih, 2018; Kumar et al., 2018; Burgess et al., 2018; Chen et al., 2018). Recent works (Horan et al., 2021; Nakagawa et al., 2023; Huh et al., 2023) provide a new perspective, showing how geometric constraints on representation spaces may enable disentanglement. In particular, Horan et al. (2021) show that unsupervised disentanglement is *always* possible under the assumption of local isometry and non-Gaussianity of generative factors motivating the desiredness of isometry.

In this work, we show how these geometric desiderata can be effectively operationalized through the lens of optimal transport (OT) theory (Santambrogio, 2015; Peyré and Cuturi, 2019), by treating mapping to or from the latent space as transport maps T from or to the data manifold, respectively. However, classic OT puts in correspondence distributions defined *in the same space* $\mathcal{X}$, using an *inter-domain* cost $c(\mathbf{x}, \mathbf{y})$ for any two points $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. Naturally, this is insufficient to map between latent and data space, where both have in parts vastly different dimensionalities, resulting in the absence of a "natural" cost function c between vectors of different sizes. This can be bypassed using the Gromov-Wasserstein (GW) (Sturm, 2020; Mémoli, 2011; Vayer, 2020; Sebbouh et al., 2023) formulation of OT, which

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the SPIGM workshop at ICML 2024. Do not distribute.

instead considers *intra-domain* costs $c_\mathcal{X}, c_\mathcal{Y}$ in each space seeking the most geometry-preserving alignment of two distributions. This mapping minimizes the distortion of the geometries induced by the *intra-domain* costs defined on each space separately. While this distortion itself can be used as a regularization for a given map T as done in Nakagawa et al. (2023), it does not take into account whether a perfect geometry preserving map *exists* between the data and latent. In practice, such a map most likely does not exist, which means that the distortion loss will optimize away from the target of T, in this case the accurate reconstruction of the data. This raises the question of whether one can account for the optimal possible geometry-preserving map.

Motivated by this formalism, we build upon the Monge gap - a regularizer introduced in Uscidda and Cuturi (2023) that measures whether a map T transports a reference distribution at minimal displacement cost - to propose the novel Gromov-Monge gap (GMG) that allows us to measure if T maps points while preserving geometric properties as much as possible - such as (scaled) isometry (distance preserving) or conformity (angle preserving). In contrast to the distortion, the GMG does take the most geometry preserving maping into account. Furthermore, we support the GMG with additional derivations that prove that the GMG and its finite sample version are weakly convex. We lay out how the GMG can serve as an effective regularizer to different geometry-preserving desiderata.

Our experiments on four standard disentangled representation learning benchmarks show that the integration of these geometry-preserving desiderate through the Gromov-Monge Gap (GMG) significantly improves disentanglement performance across various methods, from the standard β-VAE to the combination of β-TCVAE with support factorization (Roth et al., 2023), outperforming a distortion-based regularization. Moreover, we demonstrate that these geometric regularizations can replace the standard reconstruction loss, enabling measurable unsupervised disentanglement even *without a decoder*, which is not feasible in standard frameworks that rely on the decoder-based reconstruction term to prevent collapse. This finding suggests the potential for more scalable unsupervised disentangled representation learning approaches and bridges to popular, weakly- or self-supervised encoder-only representation learning methods (Chen et al., 2020b; Zbontar et al., 2021; Bardes et al., 2022; Garrido et al., 2023).

2. Background and Related Works

2.1. Disentangled Representation Learning

The Disentanglement Formalism. Disentanglement has varying operational definitions (Higgins et al., 2018; Locatello et al., 2019b; Roth et al., 2023). In this work, we follow the common understanding (Locatello et al., 2019b; 2020; Träuble et al., 2021; Roth et al., 2023) where data $\mathbf{x}$ is generated by a process $p(\mathbf{x}|\mathbf{z})$ operating on ground-truth latent factors $\mathbf{z} \sim p(\mathbf{z})$, modeling underlying source of variations (s.a. object shape, color, background...). Given a dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$, unsupervised disentangled representation learning aims to find a mapping e_ϕ s.t. $e_\phi(\mathbf{x}_i) \approx \mathbb{E}[\mathbf{z}|\mathbf{x}_i]$, up to element-wise transformations. This is to be achieved without prior information on $p(\mathbf{z})$ and $p(\mathbf{x}|\mathbf{z})$.

Unsupervised Disentanglement through Prior Matching. Most unsupervised disentanglement methods operate on variational autoencoders (VAEs)(Kingma and Welling, 2014), which define a generative model of the form $p_\theta(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})$. Here, $p_\theta(\mathbf{x}|\mathbf{z})$ is a product of exponential family distributions with parameters computed by a decoder $d_\theta(\mathbf{z})$. The latent prior $p(\mathbf{z})$ is usually chosen to be a normal Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and the probabilistic encoder $q_\phi(\mathbf{z}|\mathbf{x})$ is realized through a neural network $e_\phi(\mathbf{x})$ that predicts the parameters of the latent such that $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|e_\phi(\mathbf{x}))$. The β-VAE(Higgins et al., 2017)

$$\begin{aligned}\mathcal{L}_\beta(\theta, \phi) := \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}, \mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] \\ - \beta \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))]\end{aligned} \tag{1}$$

achieves disentanglement by enforcing stronger, β-weighted prior matching on top of the reconstruction objective, assuming statistical factor independence (Roth et al., 2023). Several follow-ups refine latent prior matching through different objectives or prior choices (Chen et al., 2018; Kumar et al., 2018; Burgess et al., 2018; Rolinek et al., 2019).

Disentanglement through a Geometric Lens. Recent studies (Gropp et al., 2020; Chen et al., 2020a; Lee et al., 2022; Nakagawa et al., 2023; Huh et al., 2023) indicate that disentanglement can arise by encouraging learned representations to preserve meaningful geometric features of the data, such as scaled distances between samples. Notably, Horan et al. (2021) demonstrated that disentanglement is provably feasible when the generative factors are sufficiently non-Gaussian and locally isometric to the data. In this work, we explore how to promote geometry preservation using quadratic OT between the latent and data spaces, which we introduce in the next section.

2.2. Quadratic Optimal Transport

Gromov-{Monge, Wasserstein} Formulations. OT (Peyré and Cuturi, 2019) involves transferring one probability distribution to another while incorporating inductive biases. When these distributions lie on incomparable domains, the task is addressed using the Gromov-Monge and GW problems, also known as OT quadratic formulations. Formally, consider two compact $\mathcal{X} \subset \mathbb{R}^{d_\mathcal{X}}$, $\mathcal{Y} \subset \mathbb{R}^{d_\mathcal{Y}}$, each of them equipped with an *intra-domain* cost $c_\mathcal{X} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $c_\mathcal{Y} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. For $p \in \mathcal{P}(\mathcal{X})$ and $q \in \mathcal{P}(\mathcal{Y})$—two distributions supported on each domain—the Gromov-Monge problem (Mémoli and Needham, 2022) seeks a map $T : \mathcal{X} \to \mathcal{Y}$ that push-forwards p onto q, while minimizing the distortion:

$$\inf_{T:T\sharp p=q} \int_{\mathcal{X}\times\mathcal{X}} d^{c_\mathcal{X},c_\mathcal{Y}}(\mathbf{x},\mathbf{x}',T(\mathbf{x}),T(\mathbf{x}'))\,\mathrm{d}p(\mathbf{x})\,\mathrm{d}p(\mathbf{x}')\,. \tag{GMP}$$

where $d^{c_\mathcal{X},c_\mathcal{Y}}(\mathbf{x},\mathbf{x}',\mathbf{y},\mathbf{y}') = \frac{1}{2}|c_\mathcal{X}(\mathbf{x},\mathbf{x}') - c_\mathcal{Y}(\mathbf{y},\mathbf{y}')|^2$. When it exists, we call a solution $T^\star$ to (GMP) a *Gromov-Monge map* for costs $c_\mathcal{X}, c_\mathcal{Y}$. However, solving this problem is difficult, and existence is not guaranteed in general (Dumont et al., 2022). Moreover, this formulation is ill-suited for discrete distributions p, q, as the constraint set might be empty. Replacing replace maps by coupling $\pi \in \Pi(p,q)$, i.e. probability distributions on $\mathcal{X}\times\mathcal{Y}$ with marginals p and q, we define the Gromov-Wasserstein (GW) metric (Mémoli, 2011; Sturm, 2020) $\mathrm{GW}^{c_\mathcal{X},c_\mathcal{Y}}(p,q) :=$

$$\min_{\pi\in\Pi(p,q)} \int_{(\mathcal{X}\times\mathcal{Y})^2} d^{c_\mathcal{X},c_\mathcal{Y}}(\mathbf{x},\mathbf{x}',\mathbf{y},\mathbf{y}')\,\mathrm{d}\pi(\mathbf{x},\mathbf{y})\,\mathrm{d}\pi(\mathbf{x}',\mathbf{y}')\,. \tag{GWP}$$

A solution $\pi^\star$ of (GWP) always exists, making $\mathrm{GW}^{c_\mathcal{X},c_\mathcal{Y}}(p,q)$ a well-defined quantity. It quantifies the minimal distortion of the geometries induced by $c_\mathcal{X}$ and $c_\mathcal{Y}$ achievable when coupling p and q.

Discrete Solvers. When both p and q are instantiated as samples, the GW Prob. ((GWP)) translates to a quadratic assignment problem, whose objective can be regularized using entropy (Cuturi, 2013). For $p_n = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}$, $q_n = \frac{1}{n}\sum_{j=1}^n \delta_{\mathbf{y}_j}$ and $\varepsilon \geq 0$, we set $\mathrm{GW}_\varepsilon^{c_\mathcal{X},c_\mathcal{Y}}(p_n,q_n) :=$

$$\min_{\mathbf{P}\in U_n} \sum_{i,j,i',j'=1}^n d^{c_\mathcal{X},c_\mathcal{Y}}(\mathbf{x}_i,\mathbf{x}_{i'},\mathbf{y}_j,\mathbf{y}_{j'})\mathbf{P}_{ij}\mathbf{P}_{i'j'} - \varepsilon H(\mathbf{P})\,, \tag{EGWP}$$

where $U_n = \{\mathbf{P} \in \mathbb{R}_+^{n\times n}, \mathbf{P}\mathbf{1}_n = \mathbf{P}^T\mathbf{1}_n = \frac{1}{n}\mathbf{1}_n\}$ and $H(\mathbf{P}) = -\sum_{i,j=1}^n \mathbf{P}_{ij}\log(\mathbf{P}_{ij})$. As $\varepsilon \to 0$, we recover $\mathrm{GW}_0^{c_\mathcal{X},c_\mathcal{Y}} = \mathrm{GW}^{c_\mathcal{X},c_\mathcal{Y}}$. In addition to yielding better statistical (Zhang et al., 2023) and regularity (Rioux et al., 2023) properties, entropic regularization also enhances computational performance. In practice, we can solve (EGWP) using a mirror descent scheme that iterates the Sinkhorn algorithm (Peyré et al., 2016; Scetbon et al., 2022).

Neural Solvers. While for classical OT, numerous neural methods have been proposed (Makkuva et al., 2020; Korotin et al., 2022; Eyring et al., 2022; Uscidda and Cuturi, 2023; Tong et al., 2023), the GW setting has received less attention. To our knowledge, the only neural Gromov-Monge formulation proposed thus far is (Nekrashevich et al., 2023), which involves a min-max-min optimization procedure. On the other hand, Klein et al. (2024) recently proposed an approach to compute neural GW couplings.

3. The Gromov-Monge Gap (GMG)

This section details our novel optimal transport perspective to achieve disentanglement from geometric considerations (see § 2.1), using the VAE framework. To achieve this, we first investigate how one can promote an arbitrary map $T : \mathcal{X} \to \mathcal{Y}$ between two domains $\mathcal{X}$ and $\mathcal{Y}$ to preserve predefined geometric features. In a VAE, T can represent either the encoder e_ϕ, which produces latent codes from the data, or the decoder d_θ, which reconstructs the data from the latent codes. As a result, in the former, the source domain $\mathcal{X}$ is the data, and the target domain $\mathcal{Y}$ is the latent space, with roles swapped in the latter. If we assume that d_θ perfectly reconstructs the data from the latents produced by e_ϕ, it is equivalent whether e_ϕ preserves the geometric features from data to latents or d_θ preserves them from latents to data. Consequently, in what follows, T can refer to either the encoder or the decoder without distinction.

Outline. Leveraging this perspective, this section begins by defining cost functions to encode geometric features and the notion of distortions in §3.1. We leverage this concept in §3.2 to introduce the Gromov-Monge Gap (GMG), a regularizer that measures whether a map moves points while preserving geometric features as much as possible, i.e., minimizing distortion. §3.3 then shows how the GMG can be estimated and computed from samples to be practically applicable in the VAE framework, which transitions into §3.4 studying convexity properties of the GMG. Put together, §3.1-§3.4 define the practical GMG which allows us to learn a latent space that matches, as much as possible, geometrical constraints in the data space. Finally in §3.5, we leverage the GMG with different choices of costs to propose effective disentangled representation learning objectives.

3.1. From the distortion...

We encode the geometric features of interest through two cost functions defined on each domain: $c_\mathcal{X} : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ and $c_\mathcal{Y} : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$. We then want T to preserve these costs, i.e., $c_\mathcal{X}(\mathbf{x},\mathbf{x}') \approx c_\mathcal{Y}(T(\mathbf{x}),T(\mathbf{x}'))$ for $\mathbf{x},\mathbf{x}' \in \mathcal{X}$. Two types of cost functions are particularly meaningful:

- **[i]** **(Scaled) squared Euclidean distance**: $c_\mathcal{X}(\mathbf{x},\mathbf{x}') = \|\mathbf{x}-\mathbf{x}'\|_2^2$ and $c_\mathcal{Y} = \alpha^2\|\mathbf{y}-\mathbf{y}\|_2^2$, with $\alpha > 0$. A map T preserving $c_\mathcal{X}, c_\mathcal{Y}$ preserves the scaled distances between the points, i.e. it is a *scaled isometry*. When $\alpha = 1$, we recover the standard definition of an *isometry*.
- **[ii]** **Cosine similarity**: $c_\mathcal{X}(\mathbf{x},\mathbf{x}') = \text{cos-sim}(\mathbf{x},\mathbf{x}') := \langle \frac{\mathbf{x}}{\|\mathbf{x}\|_2}, \frac{\mathbf{x}'}{\|\mathbf{x}'\|_2}\rangle$ and $c_\mathcal{Y}(\mathbf{y},\mathbf{y}') = \text{cos-sim}(\mathbf{y},\mathbf{y}')$ similarly. On has $\text{cos-sim}(\mathbf{x},\mathbf{x}') = \cos(\theta_{\mathbf{x},\mathbf{x}'})$ where $\theta_{\mathbf{x},\mathbf{x}'}$ is the angle between $\mathbf{x}$ and $\mathbf{x}'$. A map T preserving $c_\mathcal{X}, c_\mathcal{Y}$ then preserves the angles between the points, i.e. it is a *conformal map*. Note that if T is (scaled) isometry (see above), it is automatically a conformal map.

In the following, we say that $c_\mathcal{X}, c_\mathcal{Y}$ are **[i]** or **[ii]** if they belong to these families of costs. Introducing a reference distribution $r \in \mathcal{P}(\mathcal{X})$, weighting the areas of $\mathcal{X}$ where we penalize deviations of $c_\mathcal{X}(\mathbf{x}, \mathbf{x}')$ from $c_\mathcal{Y}(T(\mathbf{x}), T(\mathbf{x}'))$, we can quantify this property using the following criterion.

Definition 3.1 (Distortion)**.** The distortion (DST) of a map T is defined as $\mathcal{D}_r^{c_\mathcal{X}, c_\mathcal{Y}}(T) :=$

$$\int_{\mathcal{X}\times\mathcal{X}} d^{c_\mathcal{X}, c_\mathcal{Y}}(\mathbf{x}, \mathbf{x}', T(\mathbf{x}), T(\mathbf{x}')) \,\mathrm{d}r(\mathbf{x}) \,\mathrm{d}r(\mathbf{x}') \quad \text{(DST)}$$

$\mathcal{D}_r^{c_\mathcal{X}, c_\mathcal{Y}}(T)$ quantifies how much T distorts the geometric features induced by $c_\mathcal{X}$ and $c_\mathcal{Y}$ on the support of r, i.e., when $\mathcal{D}_r^{c_\mathcal{X}, c_\mathcal{Y}}(T) = 0$, one has $c_\mathcal{X}(\mathbf{x}, \mathbf{x}') = c_\mathcal{Y}(T(\mathbf{x}), T(\mathbf{x}'))$ for $\mathbf{x}, \mathbf{x}' \in \mathrm{Spt}(r)$. In disentangled representation learning, it can be desirable to regularize the decoder to be isometric (Horan et al., 2021; Nakagawa et al., 2023; Huh et al., 2023). However, a fully geometry-preserving mapping might not necessarily exist between the latent space and the data distribution. This means there usually exists an inherent trade-off between the accurate reconstruction of the data distribution and this reconstruction being, e.g., a "fully isometric" map. If such a map does not exist, the reconstruction loss and the distortion term cannot be 0 simultaneously. In practice, this means the distortion loss will optimize away from the accurate reconstruction of the data. This raises the question of how to formulate a geometric regularization that takes the most geometry-preserving mapping into account.

3.2. ...to The Gromov-Monge Gap

Recently, Uscidda and Cuturi (2023) introduced the Monge gap, a regularizer that measures whether a map T transports a reference distribution at the minimal displacement cost. In practice, this regularizer is combined with fitting losses to compute Monge maps, which are defined by two main features: (i) they fit a marginal constraint with (ii) minimal displacement cost. Building on this concept, we replace "displacement" with "distortion" to introduce the Gromov-Monge gap, a regularizer that assesses whether a map T transports a reference distribution at the minimal distortion cost. In § 3.5, we use it, alongside fitting losses, to compute Gromov-Monge maps, as defined in Eq. (GMP), which are similarly defined by (i) fitting a marginal constraint with (ii) minimal distortion cost.

Definition 3.2 (Gromov-Monge gap)**.** The Gromov-Monge gap (GMG) of a map T is defined as:

$$\mathcal{GM}_r^{c_\mathcal{X}, c_\mathcal{Y}}(T) := \mathcal{D}_r^{c_\mathcal{X}, c_\mathcal{Y}}(T) - \mathrm{GW}^{c_\mathcal{X}, c_\mathcal{Y}}(r, T\sharp r) \quad \text{(GMG)}$$

From Eq. (GWP), we recall that $\mathrm{GW}^{c_\mathcal{X}, c_\mathcal{Y}}(r, T\sharp r)$ is the minimal distortion achievable when transporting r to $T\sharp r$. Thus, $\mathcal{GM}_r^{c_\mathcal{X}, c_\mathcal{Y}}(T)$ quantifies the difference between the distortion incurred when transporting r to $T\sharp r$ via T, to this *minimal distortion*. More formally, $\mathcal{GM}_r^{c_\mathcal{X}, c_\mathcal{Y}}(T)$ is the optimality gap of T in the Gromov-Monge Prob. (GMP) between r and $T\sharp r$, which is always feasible, even when r is discrete, as T belongs to the constraint set. In light of this, it is a well-defined and

- **The GMG measures how close T is to be a Gromov-Monge map for costs $c_\mathcal{X}, c_\mathcal{Y}$.** Indeed, $\mathcal{GM}_r^{c_\mathcal{X}, c_\mathcal{Y}}(T) \geq 0$ with equality i.f.f. T is a Gromov-Monge map solution of Prob. (GMP) between r and $T\sharp r$, i.e., T moves r with minimal (but eventually non zero) distortion.
- **When transport without distortion is possible, the GMG coincides with the distortion.** When there exists another map $U : \mathcal{X} \to \mathcal{Y}$ transporting r to $T\sharp r$ with zero distortion, i.e., $U\sharp r = T\sharp r$ and $\mathcal{D}_r^{c_\mathcal{X}, c_\mathcal{Y}}(U) = 0$, then $\mathcal{GM}_r^{c_\mathcal{X}, c_\mathcal{Y}}(T) = \mathcal{D}_r^{c_\mathcal{X}, c_\mathcal{Y}}(T)$. Indeed, $\mathrm{GW}^{c_\mathcal{X}, c_\mathcal{Y}}(r, T\sharp r) = 0$ in that case, the coupling $\pi = (\mathrm{Id}, U)\sharp r$ sets the GW objective to zero, thereby minimizing it.

The last point (ii) is fundamental and illustrates how the GMG functions as a debiased distortion. Indeed, it compares the distortion induced by T to a *baseline distortion*, defined as the minimal achievable distortion when transforming the reference distribution into its image under T. Thus, when transformation without any distortion is achievable, the reference distortion becomes zero, and the GMG aligns with the distortion itself, i.e., $\mathcal{GM}_r^{c_\mathcal{X}, c_\mathcal{Y}}(T) = \mathcal{D}_r^{c_\mathcal{X}, c_\mathcal{Y}}(T)$. In this context, the GMG offers the optimal compromise: it avoids the over-penalization induced by the distortion when fully preserving $c_\mathcal{X}, c_\mathcal{Y}$ is not feasible, yet it coincides with it when such preservation is feasible.

The Influence of the Reference Distribution. A crucial property of $\mathcal{D}_r^{c_\mathcal{X}, c_\mathcal{Y}}$ is that if T transforms r without distortion, it will also apply distortion-free to any distribution s whose support is contained within that of r. Formally, if $\mathcal{D}_r^{c_\mathcal{X}, c_\mathcal{Y}}(T) = 0$ and $s \in \mathcal{P}(\mathcal{X})$ with $\mathrm{supp}(s) \subseteq \mathrm{supp}(r)$, then $\mathcal{D}_s^{c_\mathcal{X}, c_\mathcal{Y}}(T) = 0$. This raises a question for the GMG: If T maps r with minimal distortion, does it similarly map s with minimal distortion? We answer this question with Prop. (3.3) when the costs are the (scaled) Euclidean distances or the cosine similarity. This means that if T moves r while preserving (scaled) distances or angles as much as possible, it will also preserve these properties as much as possible when moving any "smaller" distribution within r.

Proposition 3.3. *When $c_\mathcal{X}, c_\mathcal{Y}$ are **[i]** or **[ii]** (see § 3.1), if $\mathcal{GM}_r^{c_\mathcal{X}, c_\mathcal{Y}}(T) = 0$, then for any $s \in \mathcal{P}(\mathcal{X})$ s.t. $\mathrm{Spt}(s) \subseteq \mathrm{Spt}(r)$, one has $\mathcal{GM}_s^{c_\mathcal{X}, c_\mathcal{Y}}(T) = 0$.*

3.3. Estimation and Computation from Samples

Plug-In Estimation. In practice, we estimate Eq. (DST) and Eq. (GMG) using i.i.d. samples $\mathbf{x}_1, ..., \mathbf{x}_n$ from the

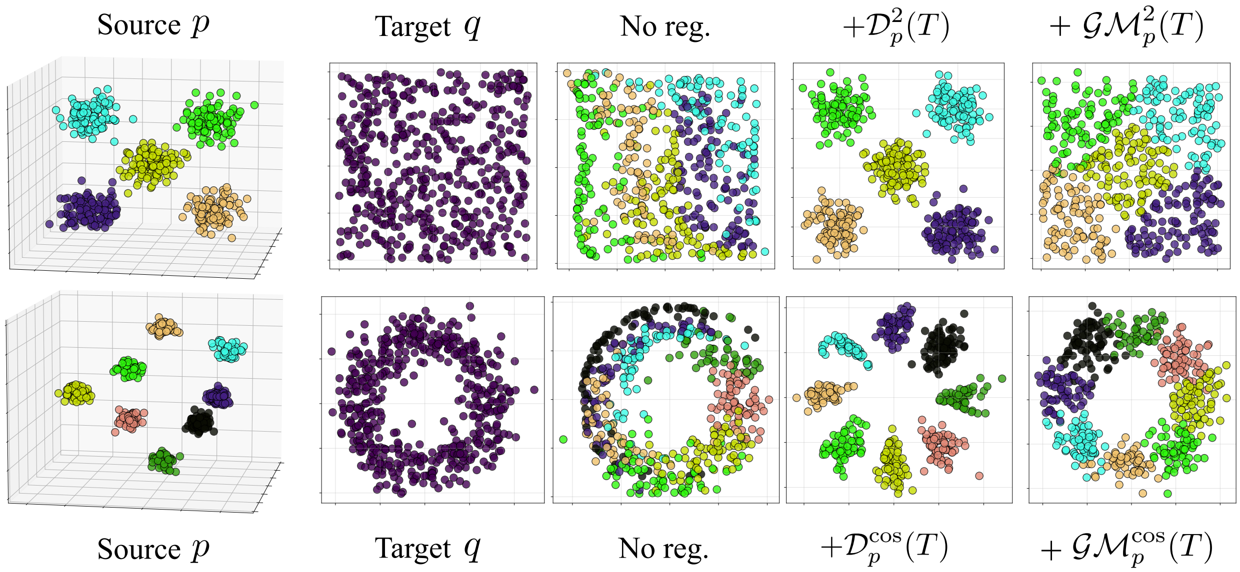

Figure 1: Learning of geometry-preserving maps with the (DST) and the (GMG). Provided a source distribution p, and a target q defining a fitting constraint, we minimize $\mathcal{L}(\theta) := S_\varepsilon(T_\theta \sharp p, q) + \lambda \mathcal{R}(T_\theta)$, where S_ε is the Sinkhorn divergence (Feydy et al., 2019), an OT-based fitting loss. We compare the effect of each regularizer $\mathcal{R} = \mathcal{GM}_p^{c_\mathcal{X}, c_\mathcal{Y}}$ and $\mathcal{R} = \mathcal{D}_p^{c_\mathcal{X}, c_\mathcal{Y}}$, and additionally train a map without regularizer as a baseline. For *all* experiments with regularizer, we use $\lambda = 1$. On the top line, we use **[i]** $c_\mathcal{X} = c_\mathcal{Y} = \|\cdot - \cdot\|_2$, aiming to preserve the distances between the points. On the bottom line, we use **[ii]** $c_\mathcal{X} = c_\mathcal{Y} = \text{cos-sim}(\cdot,\cdot)$, aiming to preserve angles. Without tuning λ, the (GMG) provides the best compromise between preserving geometric features and fitting the marginal constraint.

reference distribution r. We then consider the empirical version $r_n := \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}$ of r and use a plug-in estimator for both cases, i.e., we estimate the distortion via

$$\mathcal{D}_{r_n}^{c_\mathcal{X}, c_\mathcal{Y}}(T) = \tfrac{1}{n^2}\sum_{i,j=1}^n \left(c_\mathcal{X}(\mathbf{x}_i, \mathbf{x}_j) - c_\mathcal{Y}(T(\mathbf{x}_i), T(\mathbf{x}_j))\right)^2,$$

and the GMG via $\mathcal{GM}_{r_n}^{c_\mathcal{X}, c_\mathcal{Y}}(T) = \mathcal{D}_{r_n}^{c_\mathcal{X}, c_\mathcal{Y}}(T) - \mathrm{GW}_{c_\mathcal{X}, c_\mathcal{Y}}(r_n, T\sharp r_n)$, where $T\sharp r_n = \frac{1}{n}\sum_{i=1}^n \delta_{T(\mathbf{x}_i)}$. To better understand what the discrete GMG quantifies, we can reformulate it using the minimal distortion achieved by a permutation $\sigma \in \mathcal{S}_n$ between the $\mathbf{x}_i$ and the $T(\mathbf{x}_i)$.

Proposition 3.4. *When $c_\mathcal{X}, c_\mathcal{Y}$ are **[i]** or **[ii]** , the empirical GMG reads:*

$$\mathcal{GM}_{r_n}^{c_\mathcal{X}, c_\mathcal{Y}}(T) = \mathcal{D}_{r_n}^{c_\mathcal{X}, c_\mathcal{Y}}(T)$$
$$- \min_{\sigma \in \mathcal{S}_n} \tfrac{1}{n^2}\sum_{i,j=1}^n \left(c_\mathcal{X}(\mathbf{x}_i, \mathbf{x}_j) - c_\mathcal{Y}(T(\mathbf{x}_{\sigma(i)}), T(\mathbf{x}_{\sigma(j)}))\right)^2$$

As a Monte Carlo estimator, $\mathcal{D}_{r_n}^{c_\mathcal{X}, c_\mathcal{Y}}(T)$ is naturally consistent. We can ask the same question for $\mathcal{GM}_{r_n}^{c_\mathcal{X}, c_\mathcal{Y}}(T)$, which requires studying the convergence of the empirical GW distance $\mathrm{GW}_{c_\mathcal{X}, c_\mathcal{Y}}(r_n, T\sharp r_n)$. For the costs $c_\mathcal{X}$ and $c_\mathcal{Y}$ of interest, we show that consistency holds.

Proposition 3.5. *When $c_\mathcal{X}, c_\mathcal{Y}$ are **[i]** or **[ii]**, $\mathcal{GM}_{r_n}^{c_\mathcal{X}, c_\mathcal{Y}}(T) \to \mathcal{GM}_{r}^{c_\mathcal{X}, c_\mathcal{Y}}(T)$ almost surely.*

Efficient Computation. Computing $\mathcal{GM}_{r_n}^{c_\mathcal{X}, c_\mathcal{Y}}(T)$ requires solving a discrete GW problem between r_n and $T\sharp r_n$ to obtain $\mathrm{GW}_{c_\mathcal{X}, c_\mathcal{Y}}(r_n, T\sharp r_n)$. To alleviate computational challenges, we estimate this term using an entropic regularization $\varepsilon \geq 0$, as introduced in Eq. (EGWP):

$$\mathcal{GM}_{r_n,\varepsilon}^{c_\mathcal{X}, c_\mathcal{Y}}(T) := \mathcal{D}_{r_n}^{c_\mathcal{X}, c_\mathcal{Y}}(T) - \mathrm{GW}_{r_n,\varepsilon}^{c_\mathcal{X}, c_\mathcal{Y}}(r_n, T\sharp r_n).$$

Choosing $\varepsilon = 0$, we recover the unregularized one $\mathcal{GM}_{r_n,0}^{c_\mathcal{X}, c_\mathcal{Y}} = \mathcal{GM}_{r_n}^{c_\mathcal{X}, c_\mathcal{Y}}$. Moreover, the entropic estimator preserves the positivity, as for $\varepsilon \geq 0$, we have $\mathcal{GM}_{r_n,\varepsilon}^{c_\mathcal{X}, c_\mathcal{Y}} \geq 0$ (see A.1). As described in § 2, we compute $\mathrm{GW}_{r_n,\varepsilon}^{c_\mathcal{X}, c_\mathcal{Y}}(r_n, T\sharp r_n)$ using Peyré et al. (2016)'s solver. While it always has $\mathcal{O}(n^2)$ memory complexity, when $c_\mathcal{X} = c_\mathcal{Y} = \langle\cdot,\cdot\rangle$ or $c_\mathcal{X} = c_\mathcal{Y} = \|\cdot - \cdot\|_2^2$, this solver runs in $\mathcal{O}(n^2 d)$ time (Scetbon et al., 2022, Alg. 2). Since the cosine similarity is equivalent to the inner product, up to pre-normalization of $\mathbf{x}_i$ and $T(\mathbf{x}_i)$, the computation of the GMG for the costs of interest **[i]** or **[ii]** scales as $\mathcal{O}(n^2 d)$ in time. In practice, we use `ott-jax`'s (Cuturi et al., 2022) implementation of this scheme.

3.4. (Weak) Convexity of the Gromov-Monge gap

As laid out, the GMG can be used as a regularization loss to push any model T to be more geometry-preserving. A natural question that arises when defining such a regularizer is: what are its regularity properties, and in particular, is it convex? In the following, we study the convexity of $T \mapsto \mathcal{GM}_r^{c_\mathcal{X},c_\mathcal{Y}}(T)$, and its finite-sample counterpart $T \mapsto \mathcal{GM}_{r_n}^{c_\mathcal{X},c_\mathcal{Y}}(T)$. We focus on the costs of interest **[i]** or **[ii]**. For simplicity, we replace cosine similarity with inner product—i.e., $c_\mathcal{X} = c_\mathcal{Y} = \langle\cdot,\cdot\rangle$—as they are equivalent, up to normalization of r and T. We then study the convexity of the GMG for (i) the (scaled) squared Euclidean distances and (ii) the inner product, denoted respectively by (i) $\mathcal{GM}_r^2$ and (ii) $\mathcal{GM}_r^{\langle\cdot,\cdot\rangle}$. To that end, we introduce a weaker notion of convexity, previously defined for functions on $\mathbb{R}^d$ (Davis et al., 2018), which we extend here to $L_2(r) = \{T \,|\, \|T\|_{L_2(r)}^2 := \int_\mathcal{X} \|T(\mathbf{x})\|_2^2 \,\mathrm{d}r(\mathbf{x}) < +\infty\}$.

Definition 3.6. With $\gamma > 0$, $\mathcal{F} : L_2(r) \to \mathbb{R}$ is γ-weakly convex if $T \mapsto \mathcal{F}(T) + \frac{\gamma}{2}\|T\|_{L_2(r)}^2$ is convex.

A weakly convex functional is convex up to an additive quadratic perturbation. The weak convexity constant γ quantifies the magnitude of this perturbation and indicates a degree of non-convexity of $\mathcal{F}$. A lower γ suggests that $\mathcal{F}$ is closer to being convex, while a higher γ indicates greater non-convexity.

Theorem 3.7. *Both $\mathcal{GM}_r^2$ and $\mathcal{GM}_r^{\langle\cdot,\cdot\rangle}$, as well as their finite sample versions, are weakly convex.*

- ***Finite sample.*** *We note $\mathbf{X} \in \mathbb{R}^{n\times d}$ the matrix that stores the $\mathbf{x}_i$, i.e. the support of r_n, as rows. Then, (i) $\mathcal{GM}_{r_n}^2$ and (ii) $\mathcal{GM}_{r_n}^{\langle\cdot,\cdot\rangle}$ are respectively (i) $\gamma_{2,n}$ and (ii) $\gamma_{\text{inner},n}$-weakly convex, where: $\gamma_{\text{inner},n} = \lambda_{\max}(\frac{1}{n}\mathbf{X}\mathbf{X}^\top) - \lambda_{\min}(\frac{1}{n}\mathbf{X}\mathbf{X}^\top)$ and $\gamma_{2,n} = \gamma_{\text{inner},n} + \max_{i=1\dots n}\|\mathbf{x}_i\|_2^2$.*
- ***Asymptotic.*** *(i) $\mathcal{GM}_r^2$ and (ii) $\mathcal{GM}_r^{\langle\cdot,\cdot\rangle}$ are respectively (i) γ_2 and (ii) γ_{inner}-weakly convex, where: $\gamma_{\text{inner}} = \lambda_{\max}(\mathbb{E}_{\mathbf{x}\sim r}[\mathbf{x}\mathbf{x}^\top])$ and $\gamma_{2,n} = \gamma_{\text{inner}} + \max_{\mathbf{x}\in\mathrm{Spt}(r)}\|\mathbf{x}\|_2^2$.*

From a practitioner's perspective, we analyze the insights provided by Thm. (3.7) in three parts.

- First, we have $\gamma_2 \geq \gamma_{\text{inner}}$. Therefore, $\mathcal{GM}_r^2$ is less convex than $\mathcal{GM}_r^{\langle\cdot,\cdot\rangle}$, making it harder to optimize, and the same holds for their estimator. In other words, we provably recover that, in practice, preserving the (scaled) distances is harder than simply preserving the angles.
- Second, as $\gamma_{\text{inner}} = \lambda_{\max}(\mathbb{E}_{\mathbf{x}\sim r}[\mathbf{x}\mathbf{x}^\top]) \geq \lambda_{\max}(\mathrm{Cov}_{\mathbf{x}\sim r}[\mathbf{x}])$, this exhibits a tradeoff w.r.t. Prop. (3.3): by choosing a bigger reference distribution r, we trade the convexity of the GMG. For γ_2, the dependency in r is even worse. In practice, we then choose r with support as small as possible, precisely where we want T to move points with minimal distortion.
- Third, and probably the most surprising, the finite sample GMG is more convex in high dimension. Indeed, $\gamma_{\text{inner},n}$ is the spectral width of $\frac{1}{n}\mathbf{X}\mathbf{X}^\top$, containing the (rescaled) inner-products between the $\mathbf{x}_i \sim r$. When $n > d$, $\lambda_{\min}(\mathbf{X}\mathbf{X}^\top) = 0$ as $\mathrm{rank}(\mathbf{X}\mathbf{X}^\top) = d$. Then, $\gamma_{\text{inner},n}$ increases, which in turn decreases the GMG's convexity. On the other hand, when $d > n$, $\lambda_{\min}(\mathbf{X}\mathbf{X}^\top) > 0$ if $\mathbf{X}$ is full rank. Intuitively, $\mathcal{GM}_{r_n}^{\langle\cdot,\cdot\rangle}$ is nearly convex when $\mathbf{X}\mathbf{X}^\top$ is well conditioned. Assuming that the $\mathbf{x}_i$ are normalized, this might happen in high dimension, as those points will be orthogonal with high probability. This property suggests that, in practice, and contrary to the insights provided by the statistical OT literature (Weed and Bach, 2017; Genevay et al., 2019; Pooladian and Niles-Weed, 2021; Zhang et al., 2023), the GMG might not benefit a large sample size.

3.5. Learning with the Gromov-Monge gap

General Learning Procedure. Provided a source distribution p and a target q defining a marginal constraint, learning with the GMG remains to optimize a loss of the form

$$\mathcal{L}(\theta) := \Delta(T_\theta, p, q) + \lambda_{\text{GMG}}\mathcal{GM}_r^{c_\mathcal{X},c_\mathcal{Y}}(T_\theta) \tag{2}$$

where Δ is a fitting loss, which can access paired, or unpaired, samples of p and q. In theory, from Prop. (3.3), we can choose any reference r s.t. $\mathrm{Spt}(p) \subset \mathrm{Spt}(r)$. In practice, given the insights of Thm. (3.6), we usually consider $r = p$. Note that replacing $\mathcal{GM}_r^{c_\mathcal{X},c_\mathcal{Y}}$ by $\mathcal{D}_r^{c_\mathcal{X},c_\mathcal{Y}}$ in Eq. (2), we similarly define the learning procedure with the distortion. We compare their effect in Figure 1.

VAE Learning Procedure. In the VAE setting, (i) when we apply the GMG (or the distortion) to the encoder e_ϕ, the fitting loss is defined through the prior matching constraint, as described in § 2.1. Conversely, (ii) when we apply the GMG to the decoder d_ϕ, the fitting loss is defined through the reconstruction loss. Additionally, in both cases, our goal is to promote the latent space to preserve certain geometric features of the data. Therefore, in (i) we use $r = p_{\text{data}}$ the data distribution as reference r, while in (ii) we use the latent distribution $r = q_\phi$. Introducing weightings $\lambda_{\text{enc}}, \lambda_{\text{dec}} \geq 0$, determining which mapping we regularize, this remains to optimize the loss

$$\begin{aligned}\mathcal{L}_{\beta\text{-GMG}}(\theta,\phi) &= \mathcal{L}_\beta(\theta,\phi) \\ &+ \lambda_{\text{enc}}\mathcal{GM}_{p_{\text{data}}}^{c_\mathcal{X},c_\mathcal{Y}}(e_\phi) + \lambda_{\text{dec}}\mathcal{GM}_{q_\phi}^{c_\mathcal{X},c_\mathcal{Y}}(d_\theta),\end{aligned} \tag{3}$$

where $\mathcal{L}_\beta$ is introduce in § 2.1. Note that this loss can easily be extended to β-TCVAE and the combination of other regularization terms. While previous work (Nakagawa et al., 2023; Lee et al., 2022) chooses to apply the geometric regularizations to the decoder, we investigate regularizing both, separately or simultaneously. For completeness, we also derive the VAE-loss when learning with the distortion:

$$\begin{aligned}\mathcal{L}_{\beta\text{-DST}}(\theta,\phi) &= \mathcal{L}_\beta(\theta,\phi) \\ &+ \lambda_{\text{enc}}\mathcal{D}_{p_{\text{data}}}^{c_\mathcal{X},c_\mathcal{Y}}(e_\phi) + \lambda_{\text{dec}}\mathcal{D}_{q_\phi}^{c_\mathcal{X},c_\mathcal{Y}}(d_\theta),\end{aligned} \tag{4}$$

The choice of $c_\mathcal{X}, c_\mathcal{Y}$. Recently, Lee et al. (2022) elucidated that fully isometric regularization—preserving $c_\mathcal{X} = c_\mathcal{Y} = \| \cdot - \cdot \|_2^2$—can be overly restrictive. They introduced a Jacobian-based regularizer to learn scaled isometry–which preserves the costs $c_\mathcal{X} = \|\cdot - \cdot\|_2^2$ and $c_\mathcal{Y} = \alpha^2 \|\cdot - \cdot\|_2^2$ with $\alpha^2 > 0$. Similarly, Nakagawa et al. (2023) proposed using distortion (DST) with these costs and a learnable scaling α^2. In this work, we follow their direction and consider both the distortion and the GMG for all the costs of interest **[i]** and **[ii]** introduced in 3.1, defining a hierarchy of geometric regularization. For $c_\mathcal{X} = \| \cdot - \cdot \|_2^2$ and $c_\mathcal{Y} = \alpha^2 \| \cdot - \cdot \|_2^2$, we refer to this as scaled isometric regularization (**SIR**) for learnable $\alpha > 0$ and isometric regularization (**IR**) with fixed $\alpha = 1$. We refer to it as conformal regularization (**CR**) when $c_\mathcal{X} = c_\mathcal{Y} = \text{cos-sim}(\cdot, \cdot)$. We emphasize that in each setting, using the GMG does not aim to find a map that fully preserves the (scaled) distances (**SIR** and **IR**) or the angles (**CR**), but rather one that preserves them as much as possible while matching the prior when regularizing the encoder or reconstructing the data when regularizing the decoder.

4. Experiments

Experimental setup. We evaluate the effectiveness of the (GMG) as regularizer in disentangled representation learning. We use the standard β-VAE and β-TCVAE as our base models and incorporate the GMG on top of them. Moreover, we consider the recently proposed HFS (Roth et al., 2023) regularization on top of both β-VAE and β-TCVAE, totaling four base models. Our primary goal is to investigate the differences between using the GMG and the (DST) as regularizers, specifically examining whether the (GMG) leads to more disentangled representations compared to the raw distortion. Additionally, we aim to determine which geometric regularization (**IR**, **SIR**, **CR**) is most beneficial for disentanglement and what part of the pipeline should be regularized. Lastly, we investigate whether a geometric regularization can help prevent the collapse of learned representation in the Decoder-free setting. We evaluate the learned latents with **DCI-D** (Eastwood and Williams, 2018) as it was found to be the metric most suitable to measure disentanglement (Locatello et al., 2020; Dittadi et al., 2021). We benchmark over multiple datasets commonly used in disentangled representation learning datasets: Shapes3D (Kim and Mnih, 2018), DSprites (Higgins et al., 2017), SmallNORB (LeCun et al., 2004), and Cars3D (Reed et al., 2015).

4.1. Evaluating Different Geometric Regularizations

Regularizing the Decoder. First, we focus on the difference between optimizing for different geometry-preserving regularizations. We compare between **IR**, **SIR**, and **CR** (Lee et al., 2022) realized through either the (DST), or (GMG). Additionally, we include the Jacobian-based **SIR** as introduced in Lee et al. (2022). We report full results on Shapes3D (Burgess and Kim, 2018) over 5 seeds in Table 1. We observe that the GMG outperforms the sole distortion loss for **all** levels of regularization and baselines. Moreover, we find that a **CR** performs best with respect to disentanglement compared to both **IR** and **SIR**. Note, that employing a **CR** has not been benchmarked for disentangled representation learning before. These results elucidate the clear benefit of using the GMG in its **CR** implementation in terms of learning more disentangled representations significantly improving upon previously proposed regularizations.

Table 1: Effect of different geometric regularization on disentanglement (DCI-D, Shapes3D (Kim and Mnih, 2018)). We highlight the best method per regularization, and the **best**/<u>second best</u> per column.

	β-VAE	β-TCVAE	β-VAE + HFS	β-TCVAE + HFS
Base	65.8 ±15.6	75.0 ±3.4	88.1 ±7.4	90.2 ±7.5
Isometric (IR)				
+ (DST)	71.5 ±3.6	75.8 ±6.6	92.1 ±9.7	90.9 ±7.6
+ (GMG)	72.0 ±12.5	78.9 ±5.0	92.5 ±4.4	91.7 ±6.0
Scaled Isometric (SIR)				
+ Jacobian	61.4 ±12.8	76.7 ±4.5	90.5 ±3.8	91.5 ±5.6
+ (DST)	67.4 ±7.1	77.9 ±4.5	93.2 ±9.7	94.5 ±6.9
+ (GMG)	70.0 ±5.9	81.0 ±3.2	<u>93.3</u> ±8.6	<u>96.1</u> ±3.8
Conformal (CR)				
+ (DST)	<u>76.8</u> ±4.1	<u>81.3</u> ±4.7	87.5 ±3.3	91.9 ±9.4
+ (GMG)	**82.1** ±4.5	**83.7** ±8.8	**95.7** ±5.8	**96.9** ±4.9

Thus, next we benchmark the GMG in its **CR** form against its distortion counterpart across three more datasets again over four different baselines. We report full results in Table 2. Again we observe that the GMG outperforms or performs equally well to its distortion equivalent, confirming the benefits of accounting for the optimal possible mapping in the regularization. Note that for SmallNORB and Cars3D, we found no benefits with respect to DCI-D in adding an HFS regularization and obtained the best results without it. We emphasize that using the GMG as **CR** significantly improves results for all datasets versus not using any isometric regularization. This establishes the GMG as an effective regularization method beneficial for disentangled representation learning.

Regularizing the Encoder. Lastly, we also analyze a **CR** on e_ϕ, as well as regularizing both d_θ and e_ϕ together. We report full results over two datasets in Table 4. Again, the GMG on d_θ achieves best DCI-D over all baselines. This result is expected in the light of Theorem 3.7. Interestingly, regularizing solely d_θ outperforms regularizing both e_ϕ and d_θ. We hypothesize this is due to the regularization of the decoder also offering a stronger signal as its gradients impact both the decoder and the encoder, as in this case, the reference r is the distribution of encoded images.

Table 2: Effect of (GMG) and (DST) leveraged as a conformal regularization (**CR**) on the disentanglement of learned representations as measured by **DCI-D** over four datasets. We highlight the **best**, and second best result for each dataset and method.

CR	β-VAE	β-TCVAE	β-VAE + HFS	β-TCVAE + HFS
		Shapes3D (Kim and Mnih, 2018)		
Base	65.8 ± 15.6	75.0 ± 3.4	88.1 ± 7.4	90.2 ± 7.5
+ (DST)	76.8 ± 4.1	81.3 ± 4.7	87.5 ± 3.3	91.9 ± 9.4
+ (GMG)	**82.1** ± 4.5	**83.7** ± 8.8	**95.7** ± 5.8	**96.9** ± 4.9
		DSprites (Higgins et al., 2017)		
Base	26.2 ± 18.5	32.3 ± 19.3	33.6 ± 17.9	48.7 ± 10.2
+ (DST)	28.6 ± 19.3	32.4 ± 8.5	39.3 ± 18.1	49.0 ± 11.2
+ (GMG)	**39.5** ± 15.2	**42.2** ± 3.6	**46.7** ± 2.0	**50.1** ± 8.5
		SmallNORB (LeCun et al., 2004)		
Base	26.8 ± 0.2	29.8 ± 0.4	26.8 ± 0.2	29.8 ± 0.4
+ (DST)	28.2 ± 0.3	**29.9** ± 0.4	28.2 ± 0.3	**29.9** ± 0.4
+ (GMG)	**28.3** ± 0.6	**29.9** ± 0.5	**28.3** ± 0.6	**29.9** ± 0.5
		Cars3D (Reed et al., 2015)		
Base	29.6 ± 5.7	32.3 ± 4.6	29.6 ± 5.7	32.3 ± 4.6
+ (DST)	26.8 ± 3.6	33.7 ± 4.2	26.8 ± 3.6	33.7 ± 4.2
+ (GMG)	**30.1** ± 5.6	**36.4** ± 5.7	**30.1** ± 5.6	**36.4** ± 5.7

4.2. Towards Decoder-free Disentanglement

Recently, works such as (Burns et al., 2021; von Kügelgen et al., 2021; Eastwood et al., 2023; Matthes et al., 2023; Aitchison and Ganev, 2024) have shown the possibility of disentanglement through self-supervised, contrastive learning objectives in an effort to align with the scalability of encoder-only representation learning (Chen et al., 2020b; Zbontar et al., 2021; Bardes et al., 2022; Garrido et al., 2023). However, these encoder-only approaches still require weak supervision or access to multiple views of an image to encourage meaningful data representations.

As the goal of geometry preservation connects the data manifold and the latent domain through a minimal distortion objective and is applicable to both the encoder and decoder of a VAE (§3, Table 4), we posit that its application may provide sufficient training signal to learn meaningful representations and encourage disentanglement, eliminating the need for a reconstruction loss and decoder. Table 3 shows preliminary results on unsupervised decoder-free disentangled representation learning on the Shapes3D benchmark, where the decoder and associated reconstruction objective have been removed.

Standard approaches such as β-VAE or β-TCVAE collapse and do not achieve measurable disentanglement (DCI-D of 0.0). However, the inclusion of either DST or GMG significantly raises achievable disentanglement and, combined with the β-TCVAE matching objective, can achieve DCI-D scores of up to 53.5 without needing any decoder or reconstruction loss. While these are preliminary insights, we believe they offer promise for more scalable approaches

Table 3: Disentanglement (DCI-D) without a decoder trained with various regularizations on Shapes3D. We highlight the **best**/second best per column.

Decoder-free	β-VAE	β-TCVAE
Base	0.0 ± 0.0	0.0 ± 0.0
Isometric (IR)		
+ (DST)	38.2 ± 0.8	42.7 ± 1.6
+ (GMG)	13.9 ± 0.4	20.5 ± 0.5
Scaled Isometric (SIR)		
+ (DST)	**45.6** ± 1.2	**53.5** ± 1.0
+ (GMG)	15.2 ± 0.3	25.2 ± 0.6
Conformal (CR)		
+ (DST)	37.0 ± 0.4	46.1 ± 1.5
+ (GMG)	37.0 ± 0.9	38.8 ± 1.1

to unsupervised disentangled representation learning and potential bridges to popular and scalable self-supervised representation learning approaches. Note that the distortion loss significantly outperforms the GMG here. This is expected due to the nature of the GMG, as the distortion loss offers a more restrictive and, thus, stronger signal for learning representations, which is necessary in the absence of a reconstruction objective. This highlights that while in most scenarios (§ 4.1, Figure 1), the GMG is preferable over the distortion loss, there also exist settings where a more restrictive optimization signal is desirable.

5. Conclusion

In this work, we introduce an optimal transport (OT) perspective on unsupervised disentangled representation learning to incorporate general latent geometrical constraints. We derive the Gromov-Monge gap (GMG), a provably weakly convex OT regularizer that measures the preservation of geometrical properties by a transport map T. By formulating disentangled representation learning as a transport problem, we integrate the GMG into standard training objectives, allowing for incorporating and studying various geometric constraints on the disentanglement of learned representation spaces. Including these geometry preserving regularization offers significant performance benefits across four standard disentanglement benchmarks when applied to existing disentanglement methods. Moreover, we show promising results on decoder-free unsupervised disentanglement. We demonstrate that optimizing for geometric constraints through the OT lens can provide sufficient training signal and regularization on the model encoder to achieve measurable disentanglement without explicit reconstruction objectives. This opens a possible door towards more scalable unsupervised disentanglement and bridges to weakly- & self-supervised encoder-only representation learning efforts.

References

Laurence Aitchison and Stoil Krasimirov Ganev. InfoNCE is variational inference in a recognition parameterised model. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=chbRsWwjax.

Igor Babuschkin, Kate Baumli, Alison Bell, Surya Bhupatiraju, Jake Bruce, Peter Buchlovsky, David Budden, Trevor Cai, Aidan Clark, Ivo Danihelka, Claudio Fantacci, Jonathan Godwin, Chris Jones, Ross Hemsley, Tom Hennigan, Matteo Hessel, Shaobo Hou, Steven Kapturowski, Thomas Keck, Iurii Kemaev, Michael King, Markus Kunesch, Lena Martens, Hamza Merzic, Vladimir Mikulik, Tamara Norman, John Quan, George Papamakarios, Roman Ring, Francisco Ruiz, Alvaro Sanchez, Rosalia Schneider, Eren Sezener, Stephen Spencer, Srivatsan Srinivasan, Luyu Wang, Wojciech Stokowiec, and Fabio Viola. The DeepMind JAX Ecosystem, 2020. URL http://github.com/deepmind.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Vitória Barin-Pacela, Kartik Ahuja, Simon Lacoste-Julien, and Pascal Vincent. On the identifiability of quantized factors, 2024.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives, 2014.

D. Bertsimas and J.N. Tsitsiklis. *Introduction to linear optimization*. Athena Scientific, 1997.

Garrett Birkhoff. Tres observaciones sobre el algebra lineal. *Universidad Nacional de Tucumán Revista Series A*, 5: 147–151, 1946.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004. ISBN 0521833787. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike-20&path=ASIN/0521833787.

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

Christopher P. Burgess, Irina Higgins, Arka Pal, Loïc Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in beta-vae. *CoRR*, abs/1804.03599, 2018. URL http://arxiv.org/abs/1804.03599.

Andrea Burns, Aaron Sarna, Dilip Krishnan, and Aaron Maschinot. Unsupervised disentanglement without autoencoding: Pitfalls and future directions. *arXiv preprint arXiv:2108.06613*, 2021.

Nutan Chen, Alexej Klushyn, Francesco Ferroni, Justin Bayer, and Patrick van der Smagt. Learning flat latent manifolds with vaes, 2020a.

Ricky T. Q. Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/1ee3dfcd8a0645a25a35977997223d22-Paper.pdf.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020b. URL https://proceedings.mlr.press/v119/chen20j.html.

Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, 2013.

Marco Cuturi, Laetitia Meng-Papaxanthos, Yingtao Tian, Charlotte Bunne, Geoff Davis, and Olivier Teboul. Optimal Transport Tools (OTT): A JAX Toolbox for all things Wasserstein. *arXiv Preprint arXiv:2201.12324*, 2022.

Damek Davis, Dmitriy Drusvyatskiy, Kellie J. MacPhee, and Courtney Paquette. Subgradient methods for sharp weakly convex functions, 2018.

Andrea Dittadi, Frederik Träuble, Francesco Locatello, Manuel Wuthrich, Vaibhav Agrawal, Ole Winther, Stefan Bauer, and Bernhard Schölkopf. On the transfer of disentangled representations in realistic settings. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=8VXvj1QNRl1.

Théo Dumont, Théo Lacombe, and François-Xavier Vialard. On the existence of monge maps for the gromov-wasserstein problem. 2022.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning*

Representations, 2018. URL https://openreview.net/forum?id=By-7dz-AZ.

Cian Eastwood, Julius von Kügelgen, Linus Ericsson, Diane Bouchacourt, Pascal Vincent, Mark Ibrahim, and Bernhard Schölkopf. Self-supervised disentanglement by leveraging structure in data augmentations. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023. URL https://openreview.net/forum?id=JoISqbH8vl.

Luca Vincent Eyring, Dominik Klein, Giovanni Palla, Soeren Becker, Philipp Weiler, Niki Kilbertus, and Fabian J. Theis. Modeling single-cell dynamics using unbalanced parameterized monge maps. *bioRxiv*, 2022. doi: 10.1101/2022.10.04.510766. URL https://www.biorxiv.org/content/early/2022/10/05/2022.10.04.510766.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-Ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 22, 2019.

Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann LeCun. On the duality between contrastive and non-contrastive self-supervised learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kDEL91Dufpa.

Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample Complexity of Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 22, 2019.

Amos Gropp, Matan Atzmon, and Yaron Lipman. Isometric autoencoders, 2020.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations, 2018.

Daniella Horan, Eitan Richardson, and Yair Weiss. When is unsupervised disentanglement possible? In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=XqEF9riB93S.

Kyle Hsu, Will Dorrell, James C. R. Whittington, Jiajun Wu, and Chelsea Finn. Disentanglement via latent quantization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=LLETO26Ga2.

In Huh, changwook jeong, Jae Myung Choe, Young-Gu Kim, and Dae Sin Kim. Isometric quotient variational auto-encoders for structure-preserving representation learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=EdgPb3ngR4.

L Kantorovich. On the transfer of masses (in russian). In *Doklady Akademii Nauk*, volume 37, page 227, 1942.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kim18b.html.

Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2014.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Dominik Klein, Théo Uscidda, Fabian Theis, and Marco Cuturi. Entropic (gromov) wasserstein flow matching with genot, 2024.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. 2022. doi: 10.48550/ARXIV.2201.12220. URL https://arxiv.org/abs/2201.12220.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. VARIATIONAL INFERENCE OF DISENTANGLED LATENT CONCEPTS FROM UNLABELED OBSERVATIONS. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1kG7GZAW.

Jean-François Le Gall. *Intégration, Probabilités et Processus Aléatoires*.

Yann LeCun, Fu Jie Huang, and Léon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:II97–II104, 2004. ISSN 1063-6919. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004 ; Conference date: 27-06-2004 Through 02-07-2004.

Yonghyeon Lee, Sangwoong Yoon, MinJun Son, and Frank C. Park. Regularized autoencoders for isometric representation learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=mQxt8l7JL04.

F. Locatello, G. Abbati, T. Rainforth, S. Bauer, B. Schölkopf, and O. Bachem. On the fairness of disentangled representations. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 14584–14597. Curran Associates, Inc., December 2019a. URL https://papers.nips.cc/paper/9603-on-the-fairness-of-disentangled-representations.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124. PMLR, 09–15 Jun 2019b. URL https://proceedings.mlr.press/v97/locatello19a.html.

Francesco Locatello, Ben Poole, Gunnar Raetsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6348–6359. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/locatello20a.html.

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning (ICML)*, volume 37, 2020.

Tudor Manole and Jonathan Niles-Weed. Sharp convergence rates for empirical optimal transport with smooth costs. *The Annals of Applied Probability*, 34(1B), February 2024. ISSN 1050-5164. doi: 10.1214/23-aap1986. URL http://dx.doi.org/10.1214/23-AAP1986.

Stefan Matthes, Zhiwei Han, and Hao Shen. Towards a unified framework of contrastive learning for disentangled representations. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=QrB38MAAEP.

Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11:417–487, 2011.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, pages 666–704, 1781.

Facundo Mémoli and Tom Needham. Comparison results for gromov-wasserstein and gromov-monge distances, 2022.

Nao Nakagawa, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Gromov-wasserstein autoencoders, 2023.

Maksim Nekrashevich, Alexander Korotin, and Evgeny Burnaev. Neural gromov-wasserstein optimal transport. *arXiv preprint arXiv:2303.05978*, 2023.

K. B. Petersen and M. S. Pedersen. The matrix cookbook, October 2008. URL http://www2.imm.dtu.dk/pubdb/p.php?3274. Version 20081110.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672, 2016.

Gabriel Peyré and Marco Cuturi. Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11(5-6), 2019. ISSN 1935-8245.

Aram-Alexandre Pooladian and Jonathan Niles-Weed. Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*, 2021.

Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/e07413354875be01a996dc560274708e-Paper.pdf.

Gabriel Rioux, Ziv Goldfeld, and Kengo Kato. Entropic gromov-wasserstein distances: Stability, algorithms, and distributional limits, 2023.

Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue pca directions (by accident). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Karsten Roth, Mark Ibrahim, Zeynep Akata, Pascal Vincent, and Diane Bouchacourt. Disentanglement of correlated factors via hausdorff factorized support. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=OKcJhpQiGiX`.

Filippo Santambrogio. Optimal Transport for Applied Mathematicians. *Birkhäuser, NY*, 55(58-63):94, 2015.

Meyer Scetbon, Gabriel Peyré, and Marco Cuturi. Linear-time gromov wasserstein distances using low rank couplings and costs. In *International Conference on Machine Learning*, pages 19347–19365. PMLR, 2022.

Othmane Sebbouh, Marco Cuturi, and Gabriel Peyré. Structured transforms across spaces with cost-regularized optimal transport, 2023.

Karl-Theodor Sturm. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces, 2020.

Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. The unbalanced gromov wasserstein distance: Conic formulation and relaxation, 2023.

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport, 2023.

Frederik Träuble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Schölkopf, and Stefan Bauer. On disentangled representations learned from correlated data. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10401–10412. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/trauble21a.html`.

Théo Uscidda and Marco Cuturi. The monge gap: A regularizer to learn all transport maps, 2023.

Titouan Vayer. A contribution to optimal transport on incomparable spaces, 2020.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, 2021.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance, 2017. URL `https://arxiv.org/abs/1707.00087`.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/zbontar21a.html`.

Zhengxin Zhang, Ziv Goldfeld, Youssef Mroueh, and Bharath K. Sriperumbudur. Gromov-wasserstein distances: Entropic regularization, duality, and sample complexity, 2023.

Dominik Zietlow, Michal Rolinek, and Georg Martius. Demystifying inductive biases for (beta-)vae based architectures. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12945–12954. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/zietlow21a.html`.

Appendix

A. Proofs

A.1. Positivity of the Entropic GMG

Recall that

$$\begin{aligned}\mathcal{GM}_{r_n,\varepsilon}^{c_\mathcal{X},c_\mathcal{Y}}(T) &:= \tfrac{1}{n}\mathcal{D}_{r_n}^{c_\mathcal{X},c_\mathcal{Y}}(T) - \mathrm{GW}_{\varepsilon}^{c_\mathcal{X},c_\mathcal{Y}}(r_n, T\sharp r_n) \\ &= \mathcal{D}_{r_n}^{c_\mathcal{X},c_\mathcal{Y}}(T) - \min_{\mathbf{P}\in U_n} \sum_{i,j,i',j'=1}^{n} (c_\mathcal{X}(\mathbf{x}_i,\mathbf{x}_j) - c_\mathcal{Y}(\mathbf{y}_i,\mathbf{y}_j))^2 \mathbf{P}_{ij}\mathbf{P}_{i'j'} - \varepsilon H(\mathbf{P})\,,\end{aligned}$$

For any coupling $\mathbf{P} \in U_n$, since $-\varepsilon H(\mathbf{P}) = -\varepsilon \sum_{i,j=1}^n \mathbf{P}_{ij}\log(\mathbf{P}_{ij}) < 0$, one has:

$$\sum_{i,j,i',j'=1}^{n} (c_\mathcal{X}(\mathbf{x}_i,\mathbf{x}_j) - c_\mathcal{Y}(\mathbf{y}_i,\mathbf{y}_j))^2 \mathbf{P}_{ij}\mathbf{P}_{i'j'} - \varepsilon H(\mathbf{P}) < \sum_{i,j,i',j'=1}^{n} (c_\mathcal{X}(\mathbf{x}_i,\mathbf{x}_j) - c_\mathcal{Y}(\mathbf{y}_i,\mathbf{y}_j))^2 \mathbf{P}_{ij}\mathbf{P}_{i'j'}$$

As a result, applying minimization on both sides yields that $\mathrm{GW}_{\varepsilon}^{c_\mathcal{X},c_\mathcal{Y}}(r_n, T\sharp r_n) < \mathrm{GW}_{0}^{c_\mathcal{X},c_\mathcal{Y}}(r_n, T\sharp r_n)$, and therefore:

$$\mathrm{GW}_{\varepsilon}^{c_\mathcal{X},c_\mathcal{Y}}(T) > \mathrm{GW}_{0}^{c_\mathcal{X},c_\mathcal{Y}}(T) = \mathrm{GW}^{c_\mathcal{X},c_\mathcal{Y}}(T) \geq 0.$$

A.2. Reminders on Monge and Kantorovich OT

In this section, we recall the Monge and Kantorovich formulations of OT, which we will use to prove various results. These are the classical formulations of OT. Although we introduce them here after discussing the Gromov-Monge and Gromov-Wasserstein formulations, it should be noted that they are generally introduced beforehand. Indeed, the Gromov-Monge and Gromov-Wasserstein formulations were historically developed to derive OT formulations for comparing measures supported on incomparable spaces.

Monge Formulation. Instead of intra-domain cost functions, we consider here an *inter-domain* continuous cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. This assumes that we have a meaningful way to compare elements $\mathbf{x}, \mathbf{y}$ from the source and target domains. The Monge (1781) problem (MP) between $p \in \mathcal{P}(\mathcal{X})$ and $p \in \mathcal{P}(\mathcal{Y})$ consists of finding a map $T : \mathcal{X} \to \mathcal{Y}$ that push-forwards p onto p, while minimizing the average displacement cost quantified by c

$$\inf_{T:T\sharp p=p} \int_\mathcal{X} c(\mathbf{x}, T(\mathbf{x}))\,\mathrm{d}p(\mathbf{x})\,. \tag{MP}$$

We call any solution $T^\star$ to this problem a Monge map between p and q for cost c. Similarly to the Gromov-Monge Problem (GMP), solving the Monge Problem (MP) is difficult, as the constraint set is not convex and might be empty, especially when p, q are discrete.

Kantorovich Formulation. Instead of transport maps, the Kantorovich problem (KP) seeks a couplings $\pi \in \Pi(p, q)$:

$$\mathrm{W}_c(p,q) := \min_{\pi\in\Pi(p,q)} \int_{\mathcal{X}\times\mathcal{Y}} c(\mathbf{x},\mathbf{y})\,\mathrm{d}\pi(\mathbf{x},\mathbf{y})\,. \tag{KP}$$

An optimal coupling $\pi^\star$ solution of (KP), always exists. Studying the equivalence between (MP) and (KP) is easier than in the Gromov-Monge and Gromov-Wasserstein cases. Indeed, when (MP) is feasible, the Monge and Kantorovich formulations coincide and $\pi^\star = (\mathrm{Id}, T^\star)\sharp p$.

A.3. Conditionally Positive Kernels

In this section, we recall the definition of a conditionally positive kernel, which is involved in multiple proofs relying on the linearization of the Gromov-Wasserstein problem as a Kantorovich problem.

Definition A.1. A kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is conditionally positive if it is symmetric and for any $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^d$ and $\mathbf{a} \in \mathbb{R}^n$ s.t. $\mathbf{a}^\top \mathbf{1}_n = 0$, one has

$$\sum_{i,j=1}^{n} \mathbf{a}_i\mathbf{a}_j\, k(\mathbf{x}_i,\mathbf{x}_j) \geq 0$$

Conditionally positive kernels include all positive kernels, such as the inner-product $k(\mathbf{x},\mathbf{y}) = \langle \mathbf{x},\mathbf{y}\rangle$, the cosine similarity $k(\mathbf{x},\mathbf{y}) = \text{cos-sim}(\mathbf{x},\mathbf{y}) = \langle \frac{\mathbf{x}}{\|\mathbf{x}\|_2}, \frac{\mathbf{y}}{\|\mathbf{y}\|_2}\rangle$, but also the negative squared Euclidean distance $k(\mathbf{x},\mathbf{y}) = -\|\mathbf{x}-\mathbf{y}\|_2^2$. Therefore, each of the costs of interest is either a conditionally positive kernel - for the inner product and the cosine distance - or its opposite is - the squared Euclidean distance.

B. Proofs of § 3.2

Proposition 3.3. *When $c_\mathcal{X}, c_\mathcal{Y}$ are **[i]** or **[ii]** (see § 3.1), if $\mathcal{GM}_r^{c_\mathcal{X},c_\mathcal{Y}}(T) = 0$, then for any $s \in \mathcal{P}(\mathcal{X})$ s.t. $\mathrm{Spt}(s) \subseteq \mathrm{Spt}(r)$, one has $\mathcal{GM}_s^{c_\mathcal{X},c_\mathcal{Y}}(T) = 0$.*

Proof. Let T, r, s as described and suppose that $\mathcal{GM}_r^{c}(T) = 0$. Then, $\pi^r := (\mathrm{Id}, T)\sharp r$ is an optimal Gromov-Wasserstein coupling, solution of Problem (GWP) between r and $T\sharp r$ for costs $c_\mathcal{X}$ and $c_\mathcal{Y}$. Therefore, from (Séjourné et al., 2023, Theorem. 3), π^r is an optimal Kantorovich coupling, solution of Problem (KP) between r and $T\sharp r$ for the linearized cost:

$$\tilde{c} : (\mathbf{x},\mathbf{y}) \in \mathcal{X}\times\mathcal{Y} \mapsto \int_{\mathcal{X}\times\mathcal{Y}} \tfrac{1}{2}|c_\mathcal{X}(\mathbf{x},\mathbf{x}') - c_\mathcal{Y}(\mathbf{y},\mathbf{y}')|^2 \,\mathrm{d}\pi^r(\mathbf{x}',\mathbf{y}') \tag{5}$$

Additionally, $\mathcal{X}\times\mathcal{Y}$ is a compact set as a product of compact sets, so since $(\mathbf{x},\mathbf{y}) \mapsto |c_\mathcal{X}(\mathbf{x},\mathbf{x}') - c_\mathcal{Y}(\mathbf{y},\mathbf{y}')|^2$ is continuous as $c_\mathcal{X}$ and $c_\mathcal{Y}$ are continuous, it is bounded on $\mathcal{X}\times\mathcal{Y}$. Afterward, since π^r has finite mass, by Lebesgue's dominated convergence Theorem, it follows that $\tilde{c}$ is continuous, and hence uniformly continuous, again since $\mathcal{X}\times\mathcal{Y}$ is compact.

Afterwards, by virtue of (Santambrogio, 2015, Theorem 1.38), $\mathrm{Spt}\,(\pi^r)$ is a $\tilde{c}$-cyclically monotone (CM) set (see (Santambrogio, 2015, Definition. 1.36)). From the definition of cyclical monotonicity, this property translates to subsets. Then, by defining $\pi^s = (\mathrm{Id}, T)\sharp s$, as $\mathrm{Spt}(p) \subset \mathrm{Spt}(r)$, one has $\mathrm{Spt}(\pi^s) = \mathrm{Spt}((\mathrm{Id},T)\sharp s) \subset \mathrm{Spt}((\mathrm{Id},T)\sharp r) = \mathrm{Spt}(\pi^r)$, so $\mathrm{Spt}(\pi^s)$ is $\tilde{c}$-CM. Finally, since $\mathcal{X}$ and $\mathcal{Y}$ are compact, and $\tilde{c}$ is uniformly continuous, the $\tilde{c}$-cyclical monotonicity of its support implies that the coupling π^p is a Kantorovich optimal coupling between its marginals for cost $\tilde{c}$, thanks to (Santambrogio, 2015, Theorem 1.49). By re-applying (Séjourné et al., 2023, Theorem. 3), we get that π^s solves the Gromov-Wasserstein problem between its marginals for costs $c_\mathcal{X}$ and $c_\mathcal{Y}$. In other words, $\pi^s = (\mathrm{Id}, T)\sharp s$ is Gromov-Wasserstein optimal coupling between s and $T\sharp s$ so T is a Gromov-Monge map between s and $T\sharp s$ and $\mathcal{GM}_s^{c_\mathcal{X},c_\mathcal{Y}}(T) = 0$. □

C. Proofs of § 3.3

Proposition 3.4. *When $c_\mathcal{X}, c_\mathcal{Y}$ are **[i]** or **[ii]** , the empirical GMG reads:*

$$\begin{aligned}&\mathcal{GM}_{r_n}^{c_\mathcal{X},c_\mathcal{Y}}(T) = \mathcal{D}_{r_n}^{c_\mathcal{X},c_\mathcal{Y}}(T)\\ &- \min_{\sigma\in\mathcal{S}_n} \tfrac{1}{n^2}\sum_{i,j=1}^{n}\left(c_\mathcal{X}(\mathbf{x}_i,\mathbf{x}_j) - c_\mathcal{Y}(T(\mathbf{x}_{\sigma(i)}), T(\mathbf{x}_{\sigma(j)}))\right)^2\end{aligned}$$

Proof. We start by showing a more general results, stating that when $c_\mathcal{X}, c_\mathcal{Y}$ are conditionally positive kernels (see A.1), the discrete GW couplings between uniform, empirical distributions supported on the same number of points, ae permutation matrices.

Proposition C.1 (Equivalence between Gromov-Monge and Gromov-Wasserstein problems in the discrete case.)**.** *Let $p_n = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}$ and $q_n = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{y}_i}$ two uniform, empirical measures, supported on the same number of points. We denote by $P_n = \{\mathbf{P}\in\mathbb{R}^{n\times n}, \exists\sigma\in\mathcal{S}_n, \mathbf{P}_{ij} := \delta_{j,\sigma(i)}\}$ the set set of permutation matrices. Assume that $c_\mathcal{X}$ and $c_\mathcal{Y}$ (or $-c_\mathcal{X}$ and $-c_\mathcal{Y}$) are conditionally positive kernels (see A.1). Then, the GM and GW formulations coincide, in the sense that* *we can restrict the GW problem to permutations, namely*

$$\begin{aligned}\mathrm{GW}_{c_\mathcal{X},c_\mathcal{Y}}(p_n,p_n) &= \min_{\mathbf{P}\in U_n}\sum_{i,j,i',j'=1}^{n}(c_\mathcal{X}(\mathbf{x}_i,\mathbf{x}_{i'})-c_\mathcal{Y}(\mathbf{y}_j,\mathbf{y}_{j'}))^2\mathbf{P}_{ij}\mathbf{P}_{i'j'} \\ &= \tfrac{1}{n^2}\min_{\mathbf{P}\in P_n}\sum_{i,j,i',j'=1}^{n}(c_\mathcal{X}(\mathbf{x}_i,\mathbf{x}_{i'})-c_\mathcal{Y}(\mathbf{y}_j,\mathbf{y}_{j'}))^2\mathbf{P}_{ij}\mathbf{P}_{i'j'} \\ &= \tfrac{1}{n^2}\min_{\sigma\in\mathcal{S}_n}\sum_{i,j=1}^{n}(c_\mathcal{X}(\mathbf{x}_i,\mathbf{x}_j)-c_\mathcal{Y}(\mathbf{y}_{\sigma(i)},\mathbf{y}_{\sigma(j)}))^2\end{aligned} \tag{6}$$

Proof. Let $\mathbf{P}^\star\in U_n$ solution of the Gromov-Wasserstein between p_n and p_n, i.e.

$$\mathbf{P}^\star\in\arg\min_{\mathbf{P}\in U_n}\sum_{i,j,i',j'=1}^{n}(c_\mathcal{X}(\mathbf{x}_i,\mathbf{x}_{i'})-c_\mathcal{Y}(\mathbf{y}_j,\mathbf{y}_{j'}))^2\mathbf{P}_{ij}\mathbf{P}_{i'j'}$$

that always exists by continuity of the GW objective function on the compact U_n. We show that $\mathbf{P}^\star$ can be chosen as a (rescaled) permutation matrix without loss of generality.

As we assume that $c_\mathcal{X}$ and $c_\mathcal{Y}$ (or $-c_\mathcal{X}$ and $-c_\mathcal{Y}$) are conditionally positive kernels, from (Séjourné et al., 2023, Theorem. 3), $\mathbf{P}^\star$ also solves:

$$\mathbf{P}^\star\in\arg\min_{\mathbf{Q}\in U_n}\sum_{i,j,i',j'=1}^{n}(c_\mathcal{X}(\mathbf{x}_i,\mathbf{x}_{i'})-c_\mathcal{Y}(\mathbf{y}_j,\mathbf{y}_{j'}))^2\mathbf{P}^\star_{ij}\mathbf{Q}_{i'j'} \tag{7}$$

We then define the linearized cost matrix $\tilde{C}\in\mathbb{R}^{n\times n}$, s.t.

$$\tilde{\mathbf{C}}_{ij}=\sum_{i',j'=1}^{n}(c_\mathcal{X}(\mathbf{x}_i,\mathbf{x}_{i'})-c_\mathcal{Y}(\mathbf{y}_j,\mathbf{y}_{j'}))^2\mathbf{P}^\star_{ij}$$

which allows us to reformulate Eq. (7) as

$$\mathbf{P}^\star\in\arg\min_{\mathbf{Q}\in U_n}\langle\tilde{\mathbf{C}},\mathbf{Q}\rangle \tag{8}$$

Birkhoff's theorem states that the extremal points of U_n are equal to the permutation matrices P_n. Moreover, a seminal theorem of linear programming (Bertsimas and Tsitsiklis, 1997, Theorem 2.7) states that the minimum of a linear objective on a bounded polytope, if finite, is reached at an extremal point of the polyhedron. Therefore, as $\mathbf{P}^\star$ solves Eq. (8), it is an extremal point of U_n, so it can always be chosen as a permutation matrix. Therefore, the equivalence between GW and GM follows. □

To conclude the proof of Prop. 3.4, we simply remark that:

- $r_n=\frac{1}{n}\sum_{i=1}^n\delta_{\mathbf{x}_i}$ and $T\sharp r_n=\frac{1}{n}\sum_{i=1}^n\delta_{T(\mathbf{x}_i)}$ are uniform, empirical distribution, and supported on the same number of points;
- The costs of interests **[i]** or **[ii]** are either conditionally positive, or their opposite is, as detailed below Def (A.1).

□

Proposition C.2. *When $c_\mathcal{X},c_\mathcal{Y}$ are **[i]** or **[ii]**, $\mathcal{GM}_{r_n}^{c_\mathcal{X},c_\mathcal{Y}}(T)\to\mathcal{GM}_{r}^{c_\mathcal{X},c_\mathcal{Y}}(T)$ almost surely.*

Proof. We first note that the empirical estimator of the distortion is consistent, as both costs **[i]** or **[ii]** are continuous, and $\mathcal{X}$ is compact. We then need to study, in both cases, the convergence of $\mathrm{GW}^{c_\mathcal{X},c_\mathcal{Y}}(r_n,T\sharp r_n)$ to $\mathrm{GW}^{c_\mathcal{X},c_\mathcal{Y}}(r_n,T\sharp r)$.

To that end, we first remark that as, almost surely, $r_n\to r$ in distribution, one also has that, almost surely, $T\sharp r_n\to T\sharp r$ in distribution. Indeed, since $\mathcal{Y}$ is compact, T is bounded so for any bounded and continuous $f:\mathcal{Y}\to\mathbb{R}$ and $X\sim r$, $f\circ T(X)$

is well defined and bounded so integrable. Afterwards, one can simply adapt the proof of the almost sure weak convergence of empirical measure based on the strong law of large numbers to show that, almost surely, $T\sharp r_n \to T\sharp r$ in distribution. See for instance (Le Gall, Theorem 10.4.1).

[i] We start by the (scaled) squared euclidean distances. Up to replacing r by $\alpha^2\sharp r$ and T by $T \circ (\frac{1}{\alpha^2})$, and similarly for r_n, we can assume without loss of generality that $\alpha = 1$. As, almost surely, both $r_n \to r$ and $T\sharp r_n \to T\sharp r$ in distribution, the results follows from (Mémoli, 2011, Thm 5.1, (e)).

[ii] We continue with the cosine similarity. To that end, we first consider the inner product, i.e., $c_\mathcal{X} = c_\mathcal{Y} = \langle \cdot, \cdot \rangle$, and show that if $p_n \to p$ and $q_n \to q$ in distribution, then $\mathrm{GW}^{\langle \cdot,\cdot \rangle}(p_n, q_n) \to \mathrm{GW}^{\langle \cdot,\cdot \rangle}(p, q)$. As noticed by Rioux et al. (2023, Lemma 2)–in the first version of the paper– the GW for inner product costs can be reformulated as:

$$\begin{aligned}\mathrm{GW}^{\langle \cdot,\cdot \rangle}(p, q) = &\int_{\mathcal{X}\times\mathcal{X}} \langle \mathbf{x}, \mathbf{x}' \rangle \, \mathrm{d}p(\mathbf{x}) \, \mathrm{d}p(\mathbf{x}') + \int_{\mathcal{Y}\times\mathcal{Y}} \langle \mathbf{y}, \mathbf{y}' \rangle \, \mathrm{d}q(\mathbf{y}) \, \mathrm{d}q(\mathbf{y}') \\ &+ \min_{\mathbf{M}\in\mathcal{M}} \min_{\pi\in\Pi(p,q)} \int_{\mathcal{X}\times\mathcal{Y}} -4\langle \mathbf{M}\mathbf{x}, \mathbf{y} \rangle \, \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) + 4\|\mathbf{M}\|_2^2 \,,\end{aligned} \tag{9}$$

where we define $\mathcal{M} = [-M/2, M/2]^{d_\mathcal{X} \times d_\mathcal{Y}}$ with $M = \sqrt{\int_\mathcal{X} \|\mathbf{x}\|_2^2 \, \mathrm{d}p(\mathbf{x}) \int_\mathcal{Y} \|\mathbf{y}\|_2^2 \, \mathrm{d}q(\mathbf{y})}$. In particular, they show this result for the entropic GW problem with $\varepsilon > 0$, but their proof is also valid for $\varepsilon = 0$. The above terms only involving the marginal, i.e., not involved in the minimization, are naturally stable under convergence in distribution, as $\mathcal{X}$ and $\mathcal{Y}$ are compact, so as $\mathcal{X} \times \mathcal{X}$ and $\mathcal{Y} \times \mathcal{Y}$. As a result, we only need to study the stability of this quantity under the convergence in distribution of the following functional:

$$\mathcal{F}(p, q) = \min_{\mathbf{M}\in\mathcal{M}} \min_{\pi\in\Pi(p,q)} \int_{\mathcal{X}\times\mathcal{Y}} -4\langle \mathbf{M}\mathbf{x}, \mathbf{y} \rangle \, \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) + 4\|\mathbf{M}\|_2^2 \,, \tag{10}$$

We first remark that:

$$\begin{aligned}&|\mathcal{F}(p, q) - \mathcal{F}(p_n, q_n)| \\ &\le \sup_{M\in\mathcal{M}} |\min_{\pi\in\Pi(p,q)} \int_{\mathcal{X}\times\mathcal{Y}} -4\langle \mathbf{M}\mathbf{x}, \mathbf{y} \rangle \, \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) - \min_{\pi\in\Pi(p,q)} \int_{\mathcal{X}\times\mathcal{Y}} -4\langle \mathbf{M}\mathbf{x}, \mathbf{y} \rangle \, \mathrm{d}\pi(\mathbf{x}, \mathbf{y})| \\ &\le \sup_{M\in\mathcal{M}} |\min_{\pi\in\Pi(p,q)} \int_{\mathcal{X}\times\mathcal{Y}} 2\|\mathbf{M}\mathbf{x} - \mathbf{y}\|_2^2 \, \mathrm{d}\pi(\mathbf{x}, \mathbf{y}) - \min_{\pi\in\Pi(p_n,q_n)} \int_{\mathcal{X}\times\mathcal{Y}} 2\|\mathbf{M}\mathbf{x} - \mathbf{y}\|_2^2 \, \mathrm{d}\pi(\mathbf{x}, \mathbf{y})2| \\ &+2 \cdot \sup_{M\in\mathcal{M}} |\int_\mathcal{X} \|\mathbf{M}\mathbf{x}\|_2^2 \, \mathrm{d}p(\mathbf{x}) - \int_\mathcal{X} \|\mathbf{M}\mathbf{x}\|_2^2 \, \mathrm{d}p_n(\mathbf{x})| \\ &+2 \cdot |\int_\mathcal{Y} \|\mathbf{y}\|_2^2 \, \mathrm{d}q(\mathbf{y}) - \int_\mathcal{Y} \|\mathbf{y}\|_2^2 \, \mathrm{d}q_n(\mathbf{y})|\end{aligned} \tag{11}$$

Then, we show the convergence of each term separately.

- For the first term, we remark that (up to a constant factor) it can be reformulated:

  $$\sup_{M\in\mathcal{M}} |\mathrm{W}_2^2(\mathbf{M}\sharp p, q) - \mathrm{W}_2^2(\mathbf{M}\sharp p_n, q_n)|$$

  where we remind that that W_2^2 is the (squared) Wasserstein distance, solution of Eq. (KP) induced by $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$. By virtue of (Manole and Niles-Weed, 2024, Theorem 2), there exists a constant $C > 0$, s.t. we can uniformly bound

  $$\sup_{M\in\mathcal{M}} |\mathrm{W}_2^2(\mathbf{M}\sharp p, q) - \mathrm{W}_2^2(\mathbf{M}\sharp p_n, q_n)| \le Cn^{-1/d}$$

  and the convergence follows.
- For the second one, this follows from from the convergence in distribution of p_n to p along with the Ascoli-Arzela theorem, since both $\mathcal{M}$ and $\mathcal{X}$ are compact sets, so the $\{f_\mathbf{M} \,|\, f_\mathbf{M} : \mathbf{x} \mapsto \|\mathbf{M}\mathbf{x}\|_2^2\}$ are uniformly bounded and equi-continuous.
- For the third one, this follows from the convergence in distribution of q_n to q.

As a result, we finally get $\mathrm{GW}^{\langle\cdot,\cdot\rangle}(p_n, q_n) \to \mathrm{GW}^{\langle\cdot,\cdot\rangle}(p, q)$. Finally, we remark that for any p, q, $\mathrm{GW}^{\text{cos-sim}}(p, q) = \mathrm{GW}^{\langle\cdot,\cdot\rangle}(\mathrm{proj}_{S^{d-1}}\sharp p, \mathrm{proj}_{S^{d-1}}\sharp q)$, where $\mathrm{proj}_{S^{d-1}}(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|_2$. Using similar arguments invoked previously, as $p_n \to p$ in distribution p, $\mathrm{proj}_{S^{d-1}}\sharp p_n \to \mathrm{proj}_{S^{d-1}}\sharp p$ in distribution, and similarly $\mathrm{proj}_{S^{d-1}}\sharp q_n \to \mathrm{proj}_{S^{d-1}}\sharp q$ in distribution. As a result:

$$\begin{aligned}\mathrm{GW}^{\text{cos-sim}}(p_n, q_n) &= \mathrm{GW}^{\langle\cdot,\cdot\rangle}(\mathrm{proj}_{S^{d-1}}\sharp p_n, \mathrm{proj}_{S^{d-1}}\sharp q_n) \\ &\to \mathrm{GW}^{\langle\cdot,\cdot\rangle}(\mathrm{proj}_{S^{d-1}}\sharp p, \mathrm{proj}_{S^{d-1}}\sharp q) \\ &= \mathrm{GW}^{\text{cos-sim}}(p, q)\end{aligned} \tag{12}$$

which yields the desired convergence by using $p_n = r_n$ and $q_n = T\sharp r_n$.

□

D. Proofs of § 3.4

Theorem 3.7. *Both $\mathcal{GM}_r^2$ and $\mathcal{GM}_r^{\langle\cdot,\cdot\rangle}$, as well as their finite sample versions, are weakly convex.*

- ***Finite sample.*** *We note $\mathbf{X} \in \mathbb{R}^{n\times d}$ the matrix that stores the $\mathbf{x}_i$, i.e. the support of r_n, as rows. Then, (i) $\mathcal{GM}_{r_n}^2$ and (ii) $\mathcal{GM}_{r_n}^{\langle\cdot,\cdot\rangle}$ are respectively (i) $\gamma_{2,n}$ and (ii) $\gamma_{\text{inner},n}$-weakly convex, where: $\gamma_{\text{inner},n} = \lambda_{\max}(\frac{1}{n}\mathbf{X}\mathbf{X}^\top) - \lambda_{\min}(\frac{1}{n}\mathbf{X}\mathbf{X}^\top)$ and $\gamma_{2,n} = \gamma_{\text{inner},n} + \max_{i=1\ldots n}\|\mathbf{x}_i\|_2^2$.*
- ***Asymptotic.*** *(i) $\mathcal{GM}_r^2$ and (ii) $\mathcal{GM}_r^{\langle\cdot,\cdot\rangle}$ are respectively (i) γ_2 and (ii) γ_{inner}-weakly convex, where: $\gamma_{\text{inner}} = \lambda_{\max}(\mathbb{E}_{\mathbf{x}\sim r}[\mathbf{x}\mathbf{x}^\top])$ and $\gamma_{2,n} = \gamma_{\text{inner}} + \max_{\mathbf{x}\in\mathrm{Spt}(r)}\|\mathbf{x}\|_2^2$.*

We start by recalling the standard definition of weakly convex function on $\mathbb{R}^d$, along with technical lemmas that we will in the proof of Thm. (3.7).

Definition D.1. A function $f : \mathbb{R}^d \to \mathbb{R}$ is γ-weakly convex if $f + \gamma\|\cdot\|_2^2$ is convex.

Lemma D.2. *Let $\mathbf{A} \in S_d(\mathbb{R})$ a symmetric matrix and define the quadratic form $f_{\mathbf{A}} : \mathbf{x} \in \mathbb{R}^d \mapsto \mathbf{x}^\top\mathbf{A}\mathbf{x}$. Then, $f_{\mathbf{A}}$ is $\max(0, -\lambda_{\min}(\mathbf{A}))$-weakly convex.*

Proof. We use the fact that a twice continuously differentiable function is convex i.f.f. its hessian is positive semi-definite (Boyd and Vandenberghe, 2004, §(3.1.4)). Therefore, $f_{\mathbf{A}}$ is convex i.f.f. $\nabla^2 f_{\mathbf{A}} = \mathbf{A} \geq 0$. If $\lambda_{\min}(\mathbf{A}) \geq 0$, then $\mathbf{A} \geq 0$ so $f_{\mathbf{A}}$ is convex, i.e. 0-weakly convex. Otherwise, $f_{\mathbf{A}} - \frac{1}{2}\lambda_{\min}(\mathbf{A})\|\cdot\|_2^2$ has hessian $A - \lambda_{\min}(\mathbf{A}) \geq 0$, so it is convex, which yields that $f_{\mathbf{A}}$ is $-\lambda_{\min}(\mathbf{A})$-weakly convex. □

Lemma D.3. *Let $(f_i)_{i\in I}$ a family of γ-weakly convex functions, with potentially infinite I. Then, $f : \mathbf{x} \in \mathbb{R}^d \mapsto \sup_{i\in I} f_i(\mathbf{x})$ is γ-weakly convex.*

Proof. As the f_i are γ-weakly convex, $f_i + \frac{1}{2}\gamma$ is convex, so $\mathbf{x} \mapsto \sup_{i\in I} f_i(\mathbf{x}) + \frac{1}{2}\gamma\|x\|_2^2 = (\sup_{i\in I} f_i(\mathbf{x})) + \frac{1}{2}\gamma\|x\|_2^2$ is convex (Boyd and Vandenberghe, 2004, Eq. (3.7)). Therefore, the γ-weak convexity of f follows □

Proof of Thm. (3.7). **Finite sample**. We first study the weak convexity of $\mathcal{GM}_{r_n}^{\langle\cdot,\cdot\rangle}$, i.e. the Gromov-Monge gap for the inner product. For a map $T \in L_2(r)$, it reads

$$\begin{aligned}\mathcal{GM}_{r_n}^{\langle\cdot,\cdot\rangle}(T) = \tfrac{1}{n^2}\sum_{i,j=1}^{n} \tfrac{1}{2}|\langle\mathbf{x}_i, \mathbf{x}_j\rangle - \langle T(\mathbf{x}_i), T(\mathbf{x}_j)\rangle|^2 \\ - \min_{\mathbf{P}\in U_n}\sum_{i,j,i',j'=1}^{n} \tfrac{1}{2}|\langle\mathbf{x}_i, \mathbf{x}_{i'}\rangle - \langle T(\mathbf{x}_j), T(\mathbf{x}_{j'})\rangle|^2\mathbf{P}_{ij}\mathbf{P}_{i'j'}\end{aligned}$$

As r_n and $T\sharp r_n$ are uniform empirical supported on the same number of points, using Prop. C.1, we can reformulate the RHS with permutation matrices, which yields

$$\mathcal{GM}_{r_n}^{\langle\cdot,\cdot\rangle}(T) = \frac{1}{n^2}\sum_{i,j=1}^{n}\frac{1}{2}|\langle\mathbf{x}_i,\mathbf{x}_j\rangle - \langle T(\mathbf{x}_i),T(\mathbf{x}_j)\rangle|^2$$
$$- \frac{1}{n^2}\min_{\mathbf{P}\in P_n}\sum_{i,j,i',j'=1}^{n}\frac{1}{2}|\langle\mathbf{x}_i,\mathbf{x}_{i'}\rangle - \langle T(\mathbf{x}_j),T(\mathbf{x}_{j'})\rangle|^2\mathbf{P}_{ij}\mathbf{P}_{i'j'}$$

From this expression, $\mathcal{GM}_{r_n}^{\langle\cdot,\cdot\rangle}$ can be reformulated as a matrix input function. Indeed, it only depends on the map T via its values on the support of r_n, namely $\mathbf{x}_1,...,\mathbf{x}_n$. Therefore, we write $\mathbf{t}_i := T(\mathbf{x}_i)$, and define $\mathbf{X},\mathbf{T}\in\mathbb{R}^{n\times d}$ which contain observations $\mathbf{x}_i$ and $\mathbf{t}_i$ respectively, stored as rows. Then, studying $\mathcal{GM}_{r_n}^{\langle\cdot,\cdot\rangle}$ remains to study

$$f(\mathbf{T}) := \frac{1}{n^2}\sum_{i,j=1}^{n}\frac{1}{2}|\langle\mathbf{x}_i,\mathbf{x}_j\rangle - \langle\mathbf{t}_i,\mathbf{t}_j\rangle|^2 - \frac{1}{n^2}\min_{\mathbf{P}\in P_n}\sum_{i,j,i',j'=1}^{n}\frac{1}{2}|\langle\mathbf{x}_i,\mathbf{x}_{i'}\rangle - \langle\mathbf{t}_j,\mathbf{t}_{j'}\rangle|^2\mathbf{P}_{ij}\mathbf{P}_{i'j'}$$

By developing each term and exploiting that for any $\mathbf{P}\in P_n$, $\mathbf{P}1_n = \mathbf{P}^\top 1_n = \frac{1}{n}1_n$, we derive

$$\begin{aligned}
f(\mathbf{T}) &= \frac{1}{n^2}\sum_{i,j=1}^{n} -\langle\mathbf{x}_i,\mathbf{x}_j\rangle\cdot\langle\mathbf{t}_i,\mathbf{t}_j\rangle - \min_{\mathbf{P}\in P_n}\frac{1}{n^2}\sum_{i,j,i',j'=1}^{n} -\langle\mathbf{x}_i,\mathbf{x}_{i'}\rangle\cdot\langle\mathbf{t}_j,\mathbf{t}_{j'}\rangle\mathbf{P}_{ij}\mathbf{P}_{i'j'} \\
&= \max_{\mathbf{P}\in P_n}\frac{1}{n^2}\sum_{i,j,i',j'=1}^{n}\langle\mathbf{x}_i,\mathbf{x}_{i'}\rangle\cdot\langle\mathbf{t}_j,\mathbf{t}_{j'}\rangle\mathbf{P}_{ij}\mathbf{P}_{i'j'} - \frac{1}{n^2}\sum_{i,j=1}^{n}\langle\mathbf{x}_i,\mathbf{x}_j\rangle\cdot\langle\mathbf{t}_i,\mathbf{t}_j\rangle \\
&= \max_{\mathbf{P}\in P_n}\langle\tfrac{1}{n^2}\mathbf{P}^\top\mathbf{X}\mathbf{X}^\top\mathbf{P},\mathbf{T}\mathbf{T}^\top\rangle - \langle\tfrac{1}{n^2}\mathbf{X}\mathbf{X}^\top,\mathbf{T}\mathbf{T}^\top\rangle \\
&= \max_{\mathbf{P}\in P_n}\langle\tfrac{1}{n^2}(\mathbf{P}^\top\mathbf{X}\mathbf{X}^\top\mathbf{P} - \mathbf{X}\mathbf{X}^\top),\mathbf{T}\mathbf{T}^\top\rangle \\
&= \max_{\mathbf{P}\in P_n}\langle\tfrac{1}{n^2}(\mathbf{P}^\top\mathbf{X}\mathbf{X}^\top\mathbf{P} - \mathbf{X}\mathbf{X}^\top)\mathbf{T},\mathbf{T}\rangle \\
&= \max_{\mathbf{P}\in P_n}\langle\mathbf{A}_{\mathbf{X},\mathbf{P}}\mathbf{T},\mathbf{T}\rangle
\end{aligned}$$

where we define $\mathbf{A}_{\mathbf{X},\mathbf{P}} := \frac{1}{n^2}(\mathbf{P}^\top\mathbf{X}\mathbf{X}^\top\mathbf{P} - \mathbf{X}\mathbf{X}^\top)\in\mathbb{R}^{n\times n}$. To study the convexity of this matrix input function, we vectorize it. From (Petersen and Pedersen, 2008, Eq. (520)), we note that, for any $\mathbf{M}\in\mathbb{R}^{n\times n}$

$$\langle\mathbf{M}\mathbf{T},\mathbf{T}\rangle = \mathbf{vec}(\mathbf{T})^\top\mathbf{vec}(\mathbf{M}\mathbf{T}) = \mathbf{vec}(\mathbf{T})^\top(\mathbf{M}\otimes I_n)\mathbf{vec}(\mathbf{T})$$

where $\mathbf{vec}$ is the vectorization operator, raveling a matrix along its rows, and $\otimes$ is the Kronecker product. Applying this identity, we reformulate:

$$f(\mathbf{T}) = \max_{\mathbf{P}\in U_n}\mathbf{vec}(\mathbf{T})^\top(\mathbf{A}_{\mathbf{X},\mathbf{P}}\otimes I_n)\mathbf{vec}(\mathbf{T}) \tag{13}$$

To study the convexity of r, we study the convexity of each $r_{\mathbf{A}_{\mathbf{X},\mathbf{P}}}(\mathbf{T}) := \mathbf{vec}(\mathbf{T})^\top(\mathbf{A}_{\mathbf{X},\mathbf{P}}\otimes I_n)\mathbf{vec}(\mathbf{T})$, which are quadratic forms induced by the $\mathbf{A}_{\mathbf{X},\mathbf{P}}\otimes I_n$. This remains to study the (semi-) positive definiteness of the matrices $\mathbf{A}_{\mathbf{X},\mathbf{P}}\otimes I_n$. As each $\mathbf{A}_{\mathbf{X},\mathbf{P}}\in\mathbb{R}^{n\times n}$ is symmetric and square, $\mathbf{A}_{\mathbf{X},\mathbf{P}}\otimes I_n$ is also symmetric and from (Petersen and Pedersen, 2008, Eq. (519)) its eigenvalues are the outer products of the eigenvalues of $\mathbf{A}_{\mathbf{X},\mathbf{P}}$ and I_n, namely

$$\begin{aligned}
\mathbf{eig}(\mathbf{A}_{\mathbf{X},\mathbf{P}}\otimes I_n) &= \{\lambda_i(\mathbf{A}_{\mathbf{X},\mathbf{P}})\cdot\lambda_j(I_n)\}_{1\le i,j\le n} \\
&= \{\underbrace{\lambda_1(\mathbf{A}_{\mathbf{X},\mathbf{P}}),\ldots,\lambda_1(\mathbf{A}_{\mathbf{X},\mathbf{P}})}_{\text{n times}},\ldots,\underbrace{\lambda_n(\mathbf{A}_{\mathbf{X},\mathbf{P}}),\ldots,\lambda_n(\mathbf{A}_{\mathbf{X},\mathbf{P}})}_{\text{n times}}\}
\end{aligned} \tag{14}$$

It follows that the minimal eigenvalue of $\mathbf{A}_{\mathbf{X},\mathbf{P}}\otimes I_n$ is $\lambda_{\min}(\mathbf{A}_{\mathbf{X},\mathbf{P}}\otimes I_n) = \lambda_{\min}(\mathbf{A}_{\mathbf{X},\mathbf{P}})$. Utilizing the expression of $\mathbf{A}_{\mathbf{X},\mathbf{P}}$

$$\begin{aligned}
\lambda_{\min}(\mathbf{A}_{\mathbf{X},\mathbf{P}}) &= \tfrac{1}{n^2}\lambda_{\min}(\mathbf{P}^\top\mathbf{X}\mathbf{X}^\top\mathbf{P} - \mathbf{X}\mathbf{X}^\top) \\
&\ge \tfrac{1}{n^2}(\lambda_{\min}(\mathbf{P}^\top\mathbf{X}\mathbf{X}^\top\mathbf{P}) + \lambda_{\min}(-\mathbf{X}\mathbf{X}^\top)) \\
&= \tfrac{1}{n^2}(\lambda_{\min}(\mathbf{P}^\top\mathbf{X}\mathbf{X}^\top\mathbf{P}) - \lambda_{\max}(\mathbf{X}\mathbf{X}^\top))
\end{aligned} \tag{15}$$

Reminding that $\mathbf{P} \in U_n$, one has $\mathbf{P}^\top = \mathbf{P}^{-1}$, so $\mathbf{P}^\top \mathbf{X}\mathbf{X}^\top$ and $\mathbf{X}\mathbf{X}^\top$ are similar, and they have the same eigenvalues. In particular $\lambda_{\min}(\mathbf{P}^\top \mathbf{X}\mathbf{X}^\top \mathbf{P}) = \lambda_{\min}(\mathbf{X}\mathbf{X}^\top)$. Combining these results, it follows that

$$\lambda_{\min}(\mathbf{A}_{\mathbf{X},\mathbf{P}} \otimes I_n) = \lambda_{\min}(\mathbf{A}_{\mathbf{X},\mathbf{P}}) \geq \tfrac{1}{n^2}(\lambda_{\min}(\mathbf{X}\mathbf{X}^\top) - \lambda_{\max}(\mathbf{X}\mathbf{X}^\top)) \tag{16}$$

We then remind that each $r_{\mathbf{A}_{\mathbf{X},\mathbf{P}}}$ is the quadratic form defined by $\mathbf{A}_{\mathbf{X},\mathbf{P}} \otimes I_n$, so by applying Prop. D.2, it is $\mathbf{A}_{\mathbf{X},\mathbf{P}} \otimes I_n$-weakly convex, and hence $\frac{1}{n^2}(\lambda_{\max}(\mathbf{X}\mathbf{X}^\top) - \lambda_{\min}(\mathbf{X}\mathbf{X}^\top))$-weakly convex. Therefore, applying Prop. (D.3), r is $\frac{1}{n^2}(\lambda_{\max}(\mathbf{X}\mathbf{X}^\top) - \lambda_{\min}(\mathbf{X}\mathbf{X}^\top))$-weakly convex, in $\mathbb{R}^d$. Reminding that $\gamma_{\text{inner}} = \frac{1}{n}(\lambda_{\max}(\mathbf{X}\mathbf{X}^\top) - \lambda_{\min}(\mathbf{X}\mathbf{X}^\top))$, r is $\frac{1}{n}\gamma_{\text{inner}}$ weakly convex. This implies that $\mathbf{T} \mapsto f(\mathbf{T}) + \frac{1}{n}\gamma_{\text{inner}}\|\mathbf{T}\|_2^2$ is convex. By reminding that $\mathbf{T}$ stores the $T(\mathbf{x}_i)$ as rows, $\frac{1}{n}\|\mathbf{T}\|_2^2 = \|T\|_{L_2(r_n)}$. Consequently, $\mathcal{GM}_{r_n}^{\langle\cdot,\cdot\rangle}$ is γ_{inner} in $L_2(r_n)$.

We then study the convexity of $\mathcal{GM}^2_{r_n}$. We follow exactly the same approach. One has:

$$\begin{aligned}\mathcal{GM}^2_{r_n}(T) = \tfrac{1}{n^2}\sum_{i,j=1}^n \tfrac{1}{2}|\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 - \|T(\mathbf{x}_i) - T(\mathbf{x}_j)\|_2^2|^2 \\ - \tfrac{1}{n^2}\min_{\mathbf{P}\in P_n}\sum_{i,j,i',j'=1}^n \tfrac{1}{2}|\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 - \|T(\mathbf{x}_i) - T(\mathbf{x}_j)\|_2^2|^2 \mathbf{P}_{ij}\mathbf{P}_{i'j'}\end{aligned}$$

Similarly, studying the convexity of $\mathcal{GM}^2_{r_n}(T)$ remains to study the convexity of the matrix input function:

$$\begin{aligned}g(\mathbf{T}) := \tfrac{1}{n^2}\sum_{i,j=1}^n \tfrac{1}{2}|\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 - \|\mathbf{t}_i - \mathbf{t}_j\|_2^2|^2 \\ - \tfrac{1}{n^2}\min_{\mathbf{P}\in P_n}\sum_{i,j,i',j'=1}^n \tfrac{1}{2}|\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 - \|\mathbf{t}_i - \mathbf{t}_j\|_2^2|^2 \mathbf{P}_{ij}\mathbf{P}_{i'j'}\end{aligned}$$

As before, by developing each term, one has:

$$\begin{aligned}g(\mathbf{T}) = \max_{\mathbf{P}\in P_n} \tfrac{1}{n^2}\sum_{i,j,i',j'=1}^n \langle \mathbf{x}_i, \mathbf{x}_{i'}\rangle \cdot \langle \mathbf{t}_j, \mathbf{t}_{j'}\rangle \mathbf{P}_{ij}\mathbf{P}_{i'j'} + \tfrac{1}{2n}\sum_{i,j=1}^n \mathbf{P}_{ij}\|\mathbf{x}_i\|_2^2\|\mathbf{t}_i\|_2^2 \\ - \left(\tfrac{1}{n^2}\sum_{i,j=1}^n \langle \mathbf{x}_i, \mathbf{x}_j\rangle \cdot \langle \mathbf{t}_i, \mathbf{t}_j\rangle + \tfrac{1}{2n}\sum_{i,j=1}^n \|\mathbf{x}_i\|_2^2\|\mathbf{t}_i\|_2^2\right)\end{aligned}$$

The quadratic terms in $\mathbf{P}$ can be factorized as before using $\mathbf{A}_{\mathbf{X},\mathbf{P}}$. For the new terms w.r.t. the inner product case, we introduce $\mathbf{D}_\mathbf{X} := \text{diag}(\|\mathbf{x}_1\|_2^2, \ldots, \|\mathbf{x}_n\|_2^2)$, and remark that we can rewrite:

$$\tfrac{1}{2n}\sum_{i,j=1}^n \mathbf{P}_{ij}\|\mathbf{x}_i\|_2^2\|\mathbf{t}_i\|_2^2 - \tfrac{1}{2n}\sum_{i,j=1}^n \|\mathbf{x}_i\|_2^2\|\mathbf{t}_i\|_2^2 = \mathbf{vec}(T)^\top \left(\tfrac{1}{2n}(\mathbf{P}^\top - I_n) \otimes \mathbf{D}_\mathbf{X}\right)\mathbf{vec}(T)$$

As we can always symetrize the matrix when considering its associated quadratic form, we have:

$$\tfrac{1}{2n}\sum_{i,j=1}^n \mathbf{P}_{ij}\|\mathbf{x}_i\|_2^2\|\mathbf{t}_i\|_2^2 - \tfrac{1}{2n}\sum_{i,j=1}^n \|\mathbf{x}_i\|_2^2\|\mathbf{t}_i\|_2^2 = \mathbf{vec}(T)^\top \left(\tfrac{1}{2}(\tfrac{1}{2n}(\mathbf{P}^\top + \mathbf{P}) - I_n) \otimes \mathbf{D}_\mathbf{X}\right)\mathbf{vec}(T)$$

As a result, we denote $\mathbf{B}_{\mathbf{X},\mathbf{P}} = \frac{1}{n}(\frac{1}{2}(\mathbf{P}^\top + \mathbf{P}) - I_n) \otimes \mathbf{D}_\mathbf{X}$ and finally get:

$$g(\mathbf{T}) = \max_{\mathbf{P}\in P_n} \mathbf{vec}(T)^\top (\mathbf{A}_{\mathbf{X},\mathbf{P}} \otimes I_n + \mathbf{B}_{\mathbf{X},\mathbf{P}})\mathbf{vec}(T)$$

As we did for f, studying the weak convexity of f remains to lower bound the minimal eigenvalue of $\mathbf{A}_{\mathbf{X},\mathbf{P}} \otimes I_n + \mathbf{B}_{\mathbf{X},\mathbf{P}}$. First, one remark that:

$$\lambda_{\min}(\mathbf{A}_{\mathbf{X},\mathbf{P}} \otimes I_n + \mathbf{B}_{\mathbf{X},\mathbf{P}}) \geq \lambda_{\min}(\mathbf{A}_{\mathbf{X},\mathbf{P}} \otimes I_n) + \lambda_{\min}(\mathbf{B}_{\mathbf{X},\mathbf{P}})$$

As we we have already lower bounded $\lambda_{\min}(\mathbf{A}_{\mathbf{X},\mathbf{P}} \otimes I_n) \geq \frac{1}{n^2}(\lambda_{\min}(\mathbf{X}\mathbf{X}^\top) - \lambda_{\max}(\mathbf{X}\mathbf{X}^\top))$, we focus on the RHS. Similarly, one has:

$$\begin{aligned}\lambda_{\min}(\mathbf{B}_{\mathbf{X},\mathbf{P}}) &= \lambda_{\min}\left(\tfrac{1}{2n}(\tfrac{1}{2}(\mathbf{P}^\top + \mathbf{P}) - I_n) \otimes \mathbf{D}_{\mathbf{X}}\right) \\ &\geq \lambda_{\min}\left(\tfrac{1}{4n}(\mathbf{P}^\top + \mathbf{P}) \otimes \mathbf{D}_{\mathbf{X}}\right) + \lambda_{\min}\left(-\tfrac{1}{2n} I_n \otimes \mathbf{D}_{\mathbf{X}}\right) \\ &\geq \lambda_{\min}\left(\tfrac{1}{4n}(\mathbf{P}^\top + \mathbf{P}) \otimes \mathbf{D}_{\mathbf{X}}\right) - \lambda_{\max}\left(\tfrac{1}{2n} I_n \otimes \mathbf{D}_{\mathbf{X}}\right)\end{aligned} \tag{17}$$

For both terms, we apply again (Petersen and Pedersen, 2008, Eq. (519)). For the LHS, one has:

$$\mathbf{eig}\left(\tfrac{1}{4n}(\mathbf{P}^\top + \mathbf{P}) \otimes \mathbf{D}_{\mathbf{X}}\right) = \{\lambda_i(\tfrac{1}{4n}(\mathbf{P}^\top + \mathbf{P}))\lambda_j(\mathbf{D}_{\mathbf{X}})\}_{1 \leq i,j \leq n} \tag{18}$$

We remark that $\frac{1}{2}(\mathbf{P}^\top + \mathbf{P})$ is a symetric bi-stochastic matrix, so $\lambda_{\min}(\frac{1}{2}(\mathbf{P}^\top + \mathbf{P})) \geq -1$. Therefore, $\lambda_{\min}(\frac{1}{4n}(\mathbf{P}^\top + \mathbf{P})) \geq -\frac{1}{2n}$. As a result, since the eigenvalues of $\mathbf{D}_{\mathbf{X}}$ are the $\|\mathbf{x}_i\|_2^2$, this yields:

$$\lambda_{\min}\left(\tfrac{1}{4n}(\mathbf{P}^\top + \mathbf{P}) \otimes \mathbf{D}_{\mathbf{X}}\right) \geq -\tfrac{1}{2n} \max_{i=1,\dots,n} \|\mathbf{x}_i\|_2^2$$

Similarly, we have:

$$-\lambda_{\max}\left(\tfrac{1}{2n} I_n \otimes \mathbf{D}_{\mathbf{X}}\right) \geq -\tfrac{1}{2n} \max_{i=1,\dots,n} \|\mathbf{x}_i\|_2^2$$

from which we deduce that:

$$\lambda_{\min}(\mathbf{B}_{\mathbf{X},\mathbf{P}}) \geq -\tfrac{1}{n} \max_{i=1,\dots,n} \|\mathbf{x}_i\|_2^2$$

We can then lower bound:

$$\begin{aligned}\lambda_{\min}(\mathbf{A}_{\mathbf{X},\mathbf{P}} \otimes I_n + \mathbf{B}_{\mathbf{X},\mathbf{P}}) &\geq \tfrac{1}{n^2}(\lambda_{\min}(\mathbf{X}\mathbf{X}^\top) - \lambda_{\max}(\mathbf{X}\mathbf{X}^\top)) - \tfrac{1}{n} \max_{i=1,\dots,n} \|\mathbf{x}_i\|_2^2 \\ &= -\tfrac{1}{n}\gamma_{2,n}\end{aligned} \tag{19}$$

which yields the $\frac{1}{n}\gamma_{2,n}$-weak convexity of g, and finally the $\gamma_{2,n}$-weak convexity of $\mathcal{GM}^2_{r_n}$.

Asymptotic. For any T, we note that, almost surely, $\|T\|^2_{L_2(r_n)} \to \|T\|^2_{L_2(r)}$. As a result, since convexity is preserved under pointwise convergence and by virtue of Prop. (C.2), we study the (almost sure) convergence of $\gamma_{\text{inner},n}$ and $\gamma_{2,n}$.

We start by $\gamma_{\text{inner},n}$. We first remark that $\lambda_{\max}(\frac{1}{n}\mathbf{X}\mathbf{X}^\top) = \lambda_{\max}(\frac{1}{n}\mathbf{X}^\top\mathbf{X})$. Moreover, as $\mathbf{A} \in S_d^+(\mathbb{R}) \mapsto \lambda_{\max}(\mathbf{A})$ is continuous and $\frac{1}{n}\mathbf{X}^\top\mathbf{X} \to \mathbb{E}_{\mathbf{x}\sim r}[\mathbf{x}\mathbf{x}^\top]$ almost surely, one has $\lambda_{\max}(\frac{1}{n}\mathbf{X}\mathbf{X}^\top) \to \lambda_{\max}(\mathbb{E}_{\mathbf{x}\sim r}[\mathbf{x}\mathbf{x}^\top])$ almost surely. Moreover, for any $n > d$, $\lambda_{\min}(\frac{1}{n}\mathbf{X}\mathbf{X}^\top) = 0$. As a result, $\gamma_{\text{inner},n} \to \lambda_{\max}(\mathbb{E}_{\mathbf{x}\sim r}[\mathbf{x}\mathbf{x}^\top])$ almost surely, which provides the desired asymptotic result.

We continue with $\gamma_{2,n}$. We first remark that $\max_{i=1,\dots,n} \|\mathbf{x}_i\|_2^2 \leq \sup_{\mathbf{x}\in\mathrm{Spt}(r)} \|\mathbf{x}\|_2^2$. As a result, by defining $\tilde{\gamma}_{2,n} = \gamma_{\text{inner},n} + \max_{\mathbf{x}\in\mathrm{Spt}(r)} \|\mathbf{x}\|_2^2$, $\mathcal{GM}^2_{r_n}$ is also $\tilde{\gamma}_{2,n}$-weakly convex. Moreover, $\max_{\mathbf{x}\in\mathrm{Spt}(r)} \|\mathbf{x}\|_2^2$ does not depends on n, $\tilde{\gamma}_{2,n} \to \lambda_{\max}(\mathbb{E}_{\mathbf{x}\sim r}[\mathbf{x}\mathbf{x}^\top]) + \max_{\mathbf{x}\in\mathrm{Spt}(r)} \|\mathbf{x}\|_2^2$ almost surely, which also provides the desired asymptotic result.

□

E. Additional Empirical Results

F. Experimental Details

All our experiments build on `python 3` and the `jax`-framework (Babuschkin et al., 2020), alongside `ott-jax` for optimal transport utilities.

Table 4: Disentanglement of regularizing the Encoder and the Encoder and Decoder as measured by **DCI-D** on two different datasets. We highlight **best**, <u>second best</u>, and *third best* results for each method and dataset.

DCI-D	β-VAE	β-TCVAE	β-VAE + HFS	β-TCVAE + HFS
Shapes3D (Kim and Mnih, 2018)				
Base	67.7 ± 7.8	75.6 ± 8.7	88.1 ± 7.4	89.5 ± 7.9
+ Enc-(DST)	69.2 ± 9.1	77.2 ± 7.5	87.7 ± 7.7	90.5 ± 5.9
+ Enc-(GMG)	70.9 ± 9.5	*79.6* ± 6.6	*92.5* ± 5.9	<u>93.5</u> ± 6.9
+ Dec-(DST)	<u>76.8</u> ± 4.1	<u>81.3</u> ± 4.7	87.5 ± 3.3	*91.9* ± 9.4
+ Dec-(GMG)	**82.1** ± 4.5	**83.7** ± 8.8	**95.7** ± 5.8	**96.9** ± 4.9
+ Enc-Dec-(GMG)	*72.8* ± 7.7	79.3 ± 13.9	<u>93.3</u> ± 5.0	91.8 ± 7.3
DSprites (Higgins et al., 2017)				
Base	27.6 ± 13.4	36.0 ± 5.3	*38.7* ± 15.7	48.1 ± 10.8
+ Enc-(DST)	*32.8* ± 15.0	36.5 ± 5.9	33.9 ± 15.9	48.9 ± 11.1
+ Enc-(GMG)	27.5 ± 14.3	*37.4* ± 5.8	31.0 ± 14.3	45.9 ± 10.9
+ Dec-(DST)	28.6 ± 19.3	32.4 ± 8.5	<u>39.3</u> ± 18.1	<u>49.0</u> ± 11.2
+ Dec-(GMG)	**39.5** ± 15.2	**42.2** ± 3.6	**46.7** ± 2.0	**50.1** ± 8.5
+ Enc-Dec-(GMG)	<u>33.1</u> ± 14.9	<u>40.2</u> ± 7.0	28.7 ± 14.6	46.0 ± 11.3

To effectively conduct comprehensive and representative research on disentangled representation learning, we convert the public PyTorch framework proposed in (Roth et al., 2023) to an equivalent `jax` variant. We verify our implementation through replications of baseline and HFS results in Roth et al. (2023), mainting relative performance orderings and close absolute disentanglement scores (as measured using DCI-D, whose implementation directly follows from (Locatello et al., 2019b) and leverages gradient boosted tree implementations from `scikit-learn`).

For exact and fair comparison, we utilize standard hyperparamater choices from Roth et al. (2023) (which leverages hyerparameters directly from (Locatello et al., 2019b), (Locatello et al., 2020) and `https://github.com/google-research/disentanglement_lib`). Consequently, the base VAE architecture utilized across all experiment is the same as the one utilized in (Roth et al., 2023) and (Locatello et al., 2020): With image input sizes of $64 \times 64 \times N_c$ (with N_c the number of input image channels, usually 3). The latent dimensionality, if not otherwise specified, is set to 10. The exact VAE model architecture is as follows:

- **Encoder**: [conv($32, 4 \times 4$, stride 2) + ReLU] $\times$ 2, [conv($64, 4 \times 4$, stride 2) + ReLU] $\times$ 2, MLP(256), MLP(2 $\times$ 10)
- **Decoder**: MLP(256), [upconv($64, 4 \times 4$, stride 2) + ReLU] $\times$ 2, [upconv($32, 4 \times 4$, stride 2) + ReLU], [upconv($n_c, 4 \times 4$, stride 2) + ReLU]

Similar, we retain all training hyperparameters from (Roth et al., 2023) and (Locatello et al., 2020): Using an Adam optimizer ((Kingma and Ba, 2014), $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$) and a learning rate of 10^{-4}. Similarly, we utilize a batch-size of 64, for which we also ablate all baseline methods. The total number of training steps is set to 300000.

The exact hyperparameter grid searches used are highlighted in Tab. 5. All runs run on a RTX 2080TI GPU.

Table 5: Hyperparameter grid searches for different baseline and proposed methods.

Method	**Parameter**	**Values**
β-VAE	β	[2, 4, 6, 8, 10, 16]
β-TCVAE	β	[2, 4, 6, 8, 10, 16]
+ HFS	γ	[1, 10]
+ DST	λ	[0.1, 1, 5, 10, 20]
+ GMG	λ	[0.1, 1, 5, 10, 20]