Unsupervised Feature Importance

Warith HARCHAOUI

What Makes Paris Look like Paris?

probably Alexei A. Efros *in* SIGGRAPH, 2012

Abstract

Machine learning and pattern recognition requires data analysis: a process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusion and supporting decision-making. Unsupervised learning as a scientific field provide tools for dimensionality reduction, visuzalization procedures, features extraction etc. in order to empower human beings with enough computational power to grasp the environment we live in.

This chapter is more of plea for further investigations towards unsupervised feature importance rather than a scientific contribution. Indeed, we revisit notions such as differenciation for distributions, distribution Wasserstein-based metrics and manifold normal in order to call for both further theoretical and practical research work.

Keywords

Feature importance, relevant features, manifold, foreground/background segmentation, cosegmentation, hypersurface normal

Contents

1 Introduction **3**

2 InWaMaDi: Infinitesimal Wasserstein Maximal Distortion **3**

2.1 A Simplified Case: Empirical Distributions 5

2.2 General Distribution Case 6

3 Optimization **7**

4 Algorithm **10**

5 Possible Computer Vision Applications **11**

6 Future Work and Conclusion **11**

1 Introduction

This chapter aims at exploring possible statistical and algorithmical tools for unsupervised feature importance extraction. Indeed, the problem of extracting the absolute or relative importance of data coordinates is of high interest for understanding data. In supervised learning, revisiting the principle of Occam's razor gave birth to a considerable literature and we suggest at least one thesis to the reader: *Structured Sparsity-Inducing Norms: Statistical and Algorithmic Properties with Applications to Neuroimaging*, the Ph.D. manuscript of Jenatton [2011] which combines coordinate selection, weighting and structure in high-dimensional problems such as neuro-imaging. At the same time, the decision trees literature also carefuly studied the problem in several works by Breiman [2001, 2017] and more recently, there is a will to break the *black box taboo* of neural networks being supposedly non-interpretable with interesting attempts by Knight [2007] and de Sá [2019]. The emerging popularity of add-on toolboxes such as Captum [Kokhlikyan et al., 2019] is a solid proof showing a research trend towards input space interpretability in supervised settings at least.

Back in unsupervised learning, the problem of selecting or weighting coordinates by relevance is probably an ill-posed problem because there is no supervision. As usual in pattern recognition, we assume that data live in an instrisically low-dimensional manifold compared to the whole data space dimensionality. If we are able to find a machine learning procedure able to compute an hyper-surface normal of that manifold at each point of it, then we can interpret that normal direction coordinates as relevant or not for describing the manifold. More precisely, we can ask ourselves *Is it possible to maximally change a manifold of data with an infinitesimally small distortion?* and the distortion would be a function of space giving high amplitudes to coordinates that one should not change in order to preserve the manifold consistency. This question is reminiscent to the notion of gradient and we make the hypothesis that perturbating data in an infinitesimally small fashion can be done with gradient of a Wasserstein distance between the real data distribution and a pertubated version of it.

One possible application of this work could be unsupervised foreground / background segmentation. Indeed, from a dataset of images containing the same high-level semantic category of content (e. g. "wolves") in several outdoor / indoor conditions, the revealed coordinates would select foreground pixels from background non-content-manifold-specific pixels (that can intuitively be changed without breaking the semantic meaning of the image category). Generalizing such an automatic tool would be of great interest in many scientific fields beyond computer vision.

2 InWaMaDi: Infinitesimal Wasserstein Maximal Distortion

In the previous chapter **??**, we reviewed some consequences of the impossibility theorem by Kleinberg [2015]. In particular, the metric invariance is an interesting and difficult subject. Indeed, it seems that choosing a particular metric is a heavy commitment. This is especially true in unsupervised learning probably because determining a metric is choosing the algorithms *lenses* for seeing the data without supervision which is redefining a notion of neighborhood tainted by the curse of dimensionality we mentioned earlier in introduction section **??**.

For a random variable $\mathbf{x}$ coming from distribution p living in a space $\mathcal{X}$ (say $\mathbb{R}^D$), we can consider the distortion function $\mathcal{D} = t \times \mathcal{F}$ mapping $\mathcal{X}$ to $\mathcal{X}$ (and $t \in \mathbb{R}^+$) which creates a second random variable $\mathbf{y}$ defined by:

$$\mathbf{y} = \mathbf{x} + \mathcal{D}(\mathbf{x}) = \mathbf{x} + t \times \mathcal{F}(\mathbf{x}) \tag{1}$$

which defines a new distribution $q_{t,\mathcal{F}}$. For the sake of simplicity, we impose:

$$(\forall \mathbf{x} \in \mathbb{R}^D) \quad \|\mathcal{F}(\mathbf{x})\|_2 = 1 \tag{2}$$

so that the length of the distortion is simply t. Our goal is to measure how different a perturbated distribution $q_{t,\mathcal{F}}$ can be from the original distribution p with an infinitesimal length t and constrained energy $\|\mathcal{F}(\mathbf{x})\|_2 = 1$. Thanks to an already successful probabilistic approach in Machine Learning [Murphy, 2012], we rephrase our question set out in our introduction: *What is the infinitesimal steepest distortion of data?* Indeed, this kind of approach would give a function $\mathcal{F}$ such that when applied to each data point $\mathbf{x}_i \in \mathbb{R}^D$, the computed vector $\mathcal{F}(\mathbf{x}_i) \in \mathbb{R}^D$ would tell which coordinate $(\mathbf{x}_i^{(j)})_{j=1,\ldots,D}$ is relevant i. e. characteristic in an interpretable way for deeper data analysis especially when the dimensionality D is high.

There is a natural mathematical and geometric tool to measure a distortion for distributions: the Wasserstein distance when the associated data space metric is d. Thus, inspired by the optimization idea of *steepest gradient direction*, we can first define such a function measuring a quantity corresponding to the discrepancy $\mathcal{L}(t,\mathcal{F},d)$ induced by the distortion $\mathcal{D} = t \times \mathcal{F}$, namely the Wasserstein distance (based on metric d) between the distributions p and $q_{t,\mathcal{F}}$:

$$\mathcal{L}(t,\mathcal{F},d) = \min_{\gamma \in \Gamma(p,q_{t,\mathcal{F}})} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma} \left[d(\mathbf{x},\mathbf{y}) \right] \tag{3}$$

and because

$$\mathbf{y} \sim q_{t,\mathcal{F}} \iff \mathbf{y} = \mathbf{x}' + t \times \mathcal{F}(\mathbf{x}') \text{ and } \mathbf{x}' \sim p \tag{4}$$

we get (without applying a change of variable formula involving a Jacobian term):

$$\mathcal{L}(t,\mathcal{F},d) = \min_{\gamma' \in \Gamma(p,p)} \mathbb{E}_{(\mathbf{x},\mathbf{x}')\sim\gamma'} \left[d(\mathbf{x},\mathbf{x}' + t \times \mathcal{F}(\mathbf{x}')) \right] \tag{5}$$

with $\Gamma(p,q)$ being the set of coupled distributions with marginals p and q. Please note that without the min operator, this change of variable would not have been valid: between Eq. (3) and Eq. (5), the optimal transport changed from γ to γ' because they are different since Eq. (4).

Now we revisit the notion of *steepest function Wasserstein direction* gives an optimization problem:

$$\mathcal{F}_d^* = \arg\max_{\mathcal{F}} |\nabla_t \mathcal{L}(0,\mathcal{F},d)| \tag{6}$$

where for a given direction $\mathcal{F}$ and metric d, the quantity $\nabla_t \mathcal{L}(0,\mathcal{F},d)$ is the derivative of function $t \mapsto \mathcal{L}(t,\mathcal{F},d)$ on $t = 0^+$ which is:

$$\nabla_t \mathcal{L}(0,\mathcal{F},d) = \lim_{t\to 0^+} \frac{\mathcal{L}(t,\mathcal{F},d) - \mathcal{L}(0,\mathcal{F},d)}{t-0} \tag{7}$$

which simplifies in:

$$\nabla_t \mathcal{L}(0,\mathcal{F},d) = \lim_{t\to 0} \frac{\mathcal{L}(t,\mathcal{F},d)}{t} \tag{8}$$

because $p = q_{0,\mathcal{F}}$ for all direction $\mathcal{F}$ and so, we get:

$$\mathcal{F}_d^* = \arg\max_{\mathcal{F}} \lim_{t\to 0^+} \frac{\mathcal{L}(t,\mathcal{F},d)}{t} \tag{9}$$

At this point, we ignored the distance d operating in the data space $\mathcal{X} = \mathbb{R}^D$ but we can build such a distance in a form that parses a large variety of metrics:

$$\begin{aligned} d_\phi : \quad \mathbb{R}^D \times \mathbb{R}^D &\to \mathbb{R}_+ \\ (\mathbf{x},\mathbf{y}) &\mapsto \|\phi(\mathbf{y}) - \phi(\mathbf{x})\|_2 \end{aligned} \tag{10}$$

and we note that $d_\phi(\mathbf{x},\mathbf{y}) = (L_2 \circ \phi)(\mathbf{x},\mathbf{y}) = L_2(\phi(\mathbf{x}),\phi(\mathbf{y}))$ (L_2 being the euclidean distance).

We make sure that ϕ is a smooth bijection so that we inherit injectivity (and differentiability for optimization reasons we will see later). Indeed, such a bijective ϕ allows the associated function d_ϕ to verify the distinguishability property of a distance, namely:

$$(\forall(\mathbf{x},\mathbf{y}) \in \mathbb{R}^D \times \mathbb{R}^D) \ \ \mathbf{x} = \mathbf{y} \iff \phi(\mathbf{x}) = \phi(\mathbf{y}) \tag{11}$$

$$\iff 0 = L_2(\phi(\mathbf{x}),\phi(\mathbf{y})) = d_\phi(\mathbf{x},\mathbf{y}) \tag{12}$$

and the other required properties for being a distance: positivity, symmetry and triangular inequality are given for free thanks to the euclidean distance.

Moreover, we also want to avoid some *equivalence class explosion* effect due to the fact that there is no practical difference of interpretation between choosing a given metric d and a proportional one $\alpha \times d$ (with $\alpha > 0$) especially for explosively large coefficient α. Avoiding such annoying properties can be obtained by constraining the variations of the function ϕs indexing the distances space made of d_ϕs. Mathematically, the notion of variation for a multivariate bijective function ϕ is studied thanks to the derivative matrix $\nabla\phi(\mathbf{x}) \in \mathbb{R}^{D\times D}$ called the Jacobi matrix at each point $\mathbf{x} \in \mathbb{R}^D$ and the main variation directions are given by its eigen values $(\lambda(\mathbf{x})^{(j)})_{j=1,\ldots,D}$. Thus we can propose two ways to constrain these variations:

Bounding Amplitude (BA) We limit the Jacobi eigenvalues amplitude:

$$(\forall \mathbf{x} \in \mathbb{R}^D)(\forall j \in [\![1, D]\!]) \quad \lambda(\mathbf{x})^{(j)} \in [J_{\min}, J_{\max}] \subset \mathbb{R}_+^* \tag{13}$$

Zeroing Global Log-Amplitude (ZGLA) We enforce a null logarithm of Jacobi determinant mean:

$$0 = \kappa = \mathbb{E}_{\tilde{\mathbf{x}} \sim \mathcal{U}(p, q_{t,\mathcal{F}})} \log |\det \nabla \phi(\tilde{\mathbf{x}})| = \mathbb{E}_{\tilde{\mathbf{x}} \sim \mathcal{U}(p, q_{t,\mathcal{F}})} \left[\sum_{j=1}^{D} \log \lambda(\tilde{\mathbf{x}})^{(j)} \right] \tag{14}$$

where:

$$\begin{aligned} \tilde{\mathbf{x}} \sim \mathcal{U}(p, q_{t,\mathcal{F}}) \quad \text{means} \quad & \tilde{\mathbf{x}} = u \times \mathbf{x} + (1-u) \times \mathbf{y} \\ & \text{with } \; u \sim \mathcal{U}_{\mathbb{R}}(0,1) \\ & \text{and } \; \mathbf{x} \sim p \\ & \text{and } \; \mathbf{y} = \mathbf{x}' + t \times \mathcal{F}(\mathbf{x}') \; \text{ with } \; \mathbf{x}' \sim p \end{aligned} \tag{15}$$

Indeed, we need that "zero overall Jacobian property" being true but only over the convex enveloppe of the original and distorted points and not necessarily on the whole space $\mathbb{R}^D$ (manily because we will not evaluate the functions at hand anywhere else outside that enveloppe). This *convex enveloppe* sampling technique has been used for maintaining a Lipschitzian constraint by Gulrajani et al. [2017].

These two combined constraints over the set of smooth bijections defines the functions set Φ. For more mathematical details, we highly recommend the reader the thorough academic book by Ambrosio et al. [2008] to conduct further and more principled studies.

2.1 A Simplified Case: Empirical Distributions

Before doing some mathematical proposals, we study in this section a simplified case where only empirical distributions are at stake. For that simplified scenario with euclidean distance, we handle $\tilde{p} = \frac{1}{B} \sum_{b=1}^{B} \delta_{\mathbf{x}_{i_b}}$ an empirical distribution from p, we also have $\tilde{q}_{t,\mathcal{F}} = \frac{1}{B} \sum_{b=1}^{B} \delta_{\mathbf{x}_{i_b} + t \times \mathcal{F}(\mathbf{x}_{i_b})}$ from $q_{t,\mathcal{F}}$. For all $t \in \left[0, \frac{1}{2} \min_{i,i'} \|\mathbf{x}_{i'} - \mathbf{x}_i\|_2\right]$ the optimal transport between $\tilde{p}$ and $\tilde{q}_{t,\mathcal{F}}$ is the natural one as the Fig. 1 shows (we will proove the general case later):

$$\begin{aligned} W_{L_2}(\tilde{p}, \tilde{q}_{t,\mathcal{F}}) &= \min_{\gamma \in \Gamma(\tilde{p}, \tilde{q}_{t,\mathcal{F}})} \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim \gamma} \|\mathbf{y} - \mathbf{x}\|_2 \\ &= \min_{\gamma' \in \Gamma(\tilde{p}, \tilde{p})} \mathbb{E}_{(\mathbf{x},\mathbf{x}') \sim \gamma'} \|\mathbf{x}' + t \times \mathcal{F}(\mathbf{x}') - \mathbf{x}\|_2 \end{aligned} \tag{16}$$

but with empirical distributions, the transport γ' is in fact an assignment π of integers i_b parsing the points in the $\tilde{p} = \frac{1}{B} \sum_{b=1}^{B} \delta_{\mathbf{x}_{i_b}}$ sum to the integers $i_{b'}$ parsing the points in the $\tilde{q}_{t,\mathcal{F}} = \frac{1}{B} \sum_{b'=1}^{B} \delta_{\mathbf{x}_{i_{b'}} + t \times \mathcal{F}(\mathbf{x}_{i_{b'}})}$ with $\pi(i) = i'$:

$$\begin{aligned} W_{L_2}(\tilde{p}, \tilde{q}_{t,\mathcal{F}}) &= \min_{\pi} \frac{1}{B} \sum_{b=1}^{B} \|\mathbf{x}_{\pi(i_b)} + t \times \mathcal{F}(\mathbf{x}_{\pi(i_b)}) - \mathbf{x}_{i_b}\|_2 \\ &= \frac{1}{B} \sum_{b=1}^{B} \|\mathbf{x}_{i_b} + t \times \mathcal{F}(\mathbf{x}_{i_b}) - \mathbf{x}_{i_b}\|_2 = t \end{aligned} \tag{17}$$

so that $\lim_{t \to 0^+} \frac{W_{L_2}(\tilde{p}, \tilde{q}_{t,\mathcal{F}})}{t} = 1$ for all directions $\mathcal{F}$.

Unfortunately, this is not useful because it means that all $\mathcal{F}$ directions are equally distorting the original distribution with respect to the euclidean Wasserstein distance in a quantity that does not even depend on the distribution $\tilde{p}$.

Based on that failed study, we decide to also optimize the distance d_ϕ to get non trivial $\mathcal{F}$ directions of distortion because we just saw that the euclidean distance is independent from the manifold at hand and thus *too much* isotropic in a data-independent fashion. Optimizing the distance, gives a new optimization problem:

$$\max_{\mathcal{F}, \phi} \left(\lim_{t \to 0^+} \frac{W_{d_\phi}(p, q_{t,\mathcal{F}})}{t} \right) \tag{18}$$

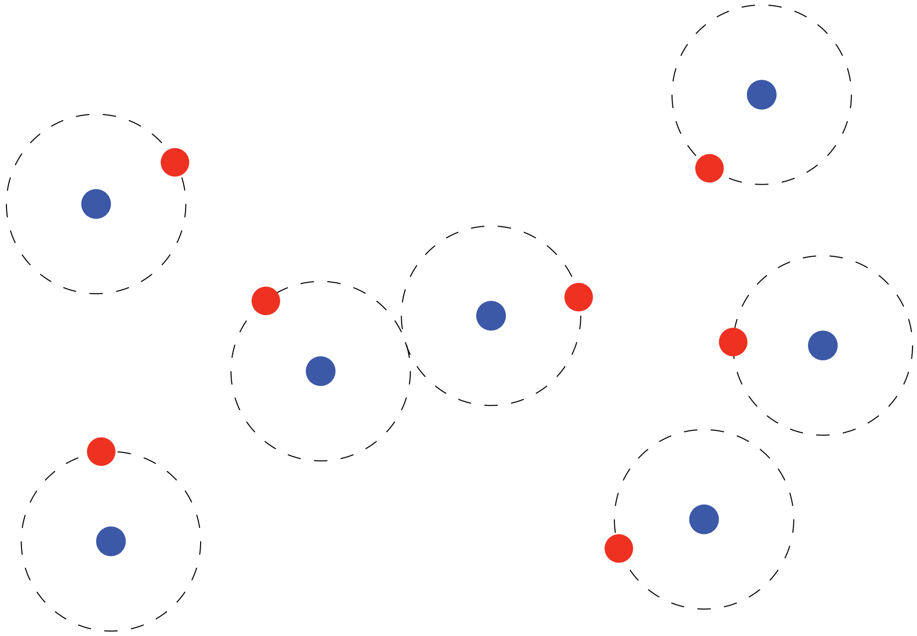

Figure 1: Optimal and Natural Transports are the same in Euclidean Distance Case for close Disttributions

instead of $\max_{\mathcal{F}} \left(\lim_{t \to 0} \frac{W_{L_2}(p, q_{t,\mathcal{F}})}{t} \right) = 1$ that previously left us with no optimization hope. This Eq. (18) is much more powerful because we do not only get the steepest distortion but also the pair of steepest distortion direction $\mathcal{F}$ and optimized associated metric d_ϕ.

Now that we have a better grasp on what is mathematically going on, we tackle the general (smooth) distributions case.

2.2 General Distribution Case

The ideas we just briefly tackled are appealing and now we give a rather theoretical result in order to pursue the optimization side in a general case beyond the euclidean distance.

Proposition. For an infinitesimally small distortion $\mathcal{D} = t \times \mathcal{F}$, the optimal transport between the original distribution p and distorted distribution $q_{t,\mathcal{F}}$ is the natural transport for all smooth metric d_ϕ indexed by bijection ϕ:

$$\begin{aligned}(\exists M \in \mathbb{R}_+^*)(\forall t \in \mathbb{R}_+^*) \;\; t < M \implies W_{d_\phi}(p, q_{t,\mathcal{F}}) &= \min_{\gamma \in \Gamma(p, q_{t,\mathcal{F}})} \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim \gamma} \left[d_\phi(\mathbf{x}, \mathbf{y}) \right] \\ &= \mathbb{E}_{\mathbf{x} \sim p} \left[d_\phi(\mathbf{x}, \mathbf{x} + t \times \mathcal{F}(\mathbf{x})) \right]\end{aligned}$$

Proof. For all pair $(\mathbf{x}, \mathbf{y})$ of points drawn from any transport $\gamma \in \Gamma(p, q_{t,\mathcal{F}})$, there exists a point $\mathbf{x}'$ such that $\mathbf{y} = \mathbf{x}' + t \times \mathcal{F}(\mathbf{x}')$ because this is how $q_{t,\mathcal{F}}$ is built. Applying a Taylor expansion on the function $f_{\mathbf{x},\mathbf{x}'}$ defined by:

$$f_{\mathbf{x},\mathbf{x}'}(t) = d_\phi(\mathbf{x}, \mathbf{x}' + t \times \mathcal{F}(\mathbf{x}')) \tag{19}$$

gives:

$$\begin{aligned} d_\phi(\mathbf{x}, \mathbf{y}) &= d_\phi(\mathbf{x}, \mathbf{x}' + t \times \mathcal{F}(\mathbf{x}')) && (20) \\ &= f_{\mathbf{x},\mathbf{x}'}(t) \\ &= f_{\mathbf{x},\mathbf{x}'}(0) + t \times \nabla_t f_{\mathbf{x},\mathbf{x}'}(0) + o(t) \\ &= d_\phi(\mathbf{x}, \mathbf{x}') + t \times \nabla_t f_{\mathbf{x},\mathbf{x}'}(0) + o(t) \end{aligned}$$
$$\tag{21}$$

Let's focus on the second term $\nabla_t f_{\mathbf{x},\mathbf{x}'}(0)$, we know thanks to a Taylor expansion on ϕ around $\mathbf{x}'$ that:

$$\begin{aligned} f_{\mathbf{x},\mathbf{x}'}(t) &= \|\phi(\mathbf{x}' + t \times \mathcal{F}(\mathbf{x}')) - \phi(\mathbf{x})\|_2 && (22) \\ &= \|\phi(\mathbf{x}') - \phi(\mathbf{x}) + t \times \nabla\phi(\mathbf{x}') \times \mathcal{F}(\mathbf{x}') + o(t)\|_2 \end{aligned}$$

and thus the only remaining non-negligible term depending on t gives:

$$\nabla_t f_{\mathbf{x},\mathbf{x}'}(0) = \|\nabla\phi(\mathbf{x}') \times \mathcal{F}(\mathbf{x}')\|_2 \tag{23}$$

and in the end, for Eq. (20), we obtain:

$$d_\phi(\mathbf{x},\mathbf{y}) = d_\phi(\mathbf{x},\mathbf{x}') + t \times \|\nabla\phi(\mathbf{x}') \times \mathcal{F}(\mathbf{x}')\|_2 + o(t) \tag{24}$$

and from that Eq. (24) as for all transport $\gamma \in \Gamma(p, q_{t,\mathcal{F}})$ there is an associated coupling $\gamma' \in \Gamma(p,p)$ such that each pair $(\mathbf{x},\mathbf{y}) \sim \gamma$ corresponds to the pair $(\mathbf{x},\mathbf{x}') \sim \gamma'$ (as seen earlier) we get:

$$\begin{aligned} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma}\left[d_\phi(\mathbf{x},\mathbf{y})\right] &= \mathbb{E}_{(\mathbf{x},\mathbf{x}')\sim\gamma'}\left[d_\phi(\mathbf{x},\mathbf{x}')\right] \\ &\quad + t \times \mathbb{E}_{\mathbf{x}'\sim p}\left[\|\nabla\phi(\mathbf{x}') \times \mathcal{F}(\mathbf{x}')\|_2\right] \\ &\quad + t \times \epsilon_{\mathcal{F},\gamma}(t) \end{aligned} \tag{25}$$

where $\lim_{t\to 0} \epsilon_{\mathcal{F},\gamma}(t) = 0$ and one can note that the second term decouples the $(\mathbf{x},\mathbf{x}')$ pairing. If we consider the natural transport γ^* (i. e. $(\mathbf{x},\mathbf{y}) \sim \gamma^* \iff (\mathbf{x} \sim p \text{ and } \mathbf{y} = \mathbf{x} + t \times \mathcal{F}(\mathbf{x}))$, then we can measure the difference $D(\gamma,\gamma^*)$ for any other non-natural transport γ:

$$\begin{aligned} D(\gamma,\gamma^*) &= \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma}\left[d_\phi(\mathbf{x},\mathbf{y})\right] - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma^*}\left[d_\phi(\mathbf{x},\mathbf{y})\right] \\ &= \mathbb{E}_{(\mathbf{x},\mathbf{x}')\sim\gamma'}\left[d_\phi(\mathbf{x},\mathbf{x}')\right] + t \times \left(\epsilon_{\mathcal{F},\gamma}(t) - \epsilon_{\mathcal{F},\gamma^*}(t)\right) \quad \text{thanks to Eq. (25)} \end{aligned} \tag{26}$$

which is positive for a $t > 0$ sufficiently small because $\lim_{t\to 0} |\epsilon_{\mathcal{F},\gamma}(t) - \epsilon_{\mathcal{F},\gamma^*}(t)| = 0$ which proves that the natural transport γ^* has the minimal transport cost and thus is the optimal transport. □

When we briefly saw the empirical distribution and euclidean case, it appeared that when the distortion is sufficiently low in amplitude (i. e. with a small t), the optimal transport is the natural one (assigning each original point to its distorted version). With that proposition above in mind and in order to build an optimization objective and an algorithm, we propose a procedure to get a value of such a $t > 0$ satisfying:

$$\max_b \|\phi(\mathbf{x}_{i_b}) - \phi(\mathbf{x}_{i_b} + t \times \mathcal{F}(\mathbf{x}_{i_b}))\|_2 \leq \frac{1}{2} \min_{b,b'} \|\phi(\mathbf{x}_{i_b}) - \phi(\mathbf{x}_{i_{b'}})\|_2 \tag{27}$$

The right term does not depend on t and is easily computed from mini-batches of data. The left term can be approximated to guess the right order of magnitude for t that we can divide later until the inequality Eq. (27) is satisfied in a dichotomic fashion. Indeed:

$$\begin{aligned} \|\phi(\mathbf{x}_{i_b}) - \phi(\mathbf{x}_{i_b} + t \times \mathcal{F}(\mathbf{x}_{i_b}))\|_2 &\simeq t \times \|\nabla\phi(\mathbf{x}_{i_b}) \times \mathcal{F}(\mathbf{x}_{i_b})\|_2 \\ &\leq t \times J_{\max} \;\text{ because } \|\mathcal{F}(\mathbf{x}_{i_b})\|_2 = 1 \end{aligned} \tag{28}$$

and thus

$$t_k = \frac{\min_{b,b'} \|\phi(\mathbf{x}_{i_b}) - \phi(\mathbf{x}_{i_{b'}})\|_2}{2^{k+1} \times J_{\max}} \tag{29}$$

with an increasing $k \in \mathbb{N}^*$ is a good strategy until satisfying Eq. (27).

3 Optimization

In this section, we present a draft of an optimization strategy to summarize the ideas we just presented with neural networks implementations for functions. This is about leveraging the research effort for deep learning in general and GANs in particular for our representation learning and data analysis purposes.

Indeed, we have the function $\mathcal{F}$ implemented by a neural network of parararameters $\theta_{\mathcal{F}}$ mapping $\mathbb{R}^D$ to $\mathbb{R}^D$. We add some layers: a SoftMax layer followed by an element-wise square-root layer with kept sign layer such that we maintain the norm 1 constraint on $\mathcal{F}$.

Thanks to some work accomplished in a different context by Dinh et al. [2017], the implementation of bijection ϕ in a special neural network of parameters θ_ϕ is already done and the constraints needed in section 2 are easily applicable. More specifically, we keep on zeroing the overall jacobian mean in a way that is similar to online k-Means [Bottou and Bengio, 1995] via an intermediate bijection ψ that we divide by scalar κ for control the variation of $\phi = \exp(\frac{-\kappa}{D}) \times \psi$. All these parameters are concatenated in θ that the algorithm 2 optimizes.

Maintaining the constrains that we required Eq. (13) and Eq. (14) in section 2 is facilitated by the structure of our bijective neural network that we took from Dinh et al. [2017]. Indeed, for bijection ψ implemented by L bijective layers $\psi = \text{Layer}_L \circ \cdots \circ \text{Layer}_\ell \circ \cdots \circ \text{Layer}_1$, we have at each layer ℓ among L:

$$\begin{aligned} \mathbf{x}_{[:d]} \mapsto \quad \mathbf{y}_{[:d_\ell]} &= \text{Layer}_\ell(\mathbf{x})_{[:d_\ell]} = \mathbf{x}_{[:d_\ell]} \\ \mathbf{x}_{[d_\ell:]} \mapsto \quad \mathbf{y}_{[d_\ell:]} &= \text{Layer}_\ell(\mathbf{x})_{[d_\ell:]} = \mathbf{x}_{[d_\ell:]} \times \exp(s_\ell(\mathbf{x}_{[:d_\ell]})) + t_\ell(\mathbf{x}_{[:d_\ell]}) \end{aligned} \tag{30}$$

where the functions s_ℓs and t_ℓs are *free* neural networks (with *pythonic* notations for dimensions and without loss of generality in the coordinates order but with an arbitrary pivot d_ℓ) and then we can bound the associated Jacobi matrices:

$$\begin{aligned} \nabla \text{Layer}_\ell(\mathbf{x}) &= \begin{pmatrix} \mathbf{I}_{d_\ell} & \mathbf{0}_{d_\ell \times (D-d_\ell)} \\ - & \text{diag}(\exp(s_\ell(\mathbf{x}_{[:d_\ell]}))) \end{pmatrix} \\ \text{and } \log|\det \nabla \text{Layer}_\ell(\mathbf{x})| &= \sum_{j=1}^{D-d_\ell} s_\ell^{(j)}(\mathbf{x}_{[:d_\ell]}) \end{aligned} \tag{31}$$

Thanks to the chain rule applied to such a bijection ψ as a composition of layers from $\mathbf{x}^0 = \mathbf{x}$ to $\mathbf{x}^\ell = \text{Layer}_\ell(\mathbf{x}^{\ell-1})$ for $\ell \in [\![1, L]\!]$, we can collect through the forward computations of the function ψ output and sum the outputs of s_ℓ in order to get κ:

$$\kappa = \mathbb{E}_{\mathbf{x} \sim \mathcal{U}(p,q_{t,\mathcal{F}})} \left[\log|\det \nabla\psi(\mathbf{x})|\right] = \mathbb{E}_{\mathbf{x} \sim \mathcal{U}(p,q_{t,\mathcal{F}})} \left[\sum_{\ell=1}^{L} \sum_{j=1}^{D-d_\ell} s_\ell^{(j)}(\mathbf{x}^{\ell-1}_{[:d_\ell]})\right] \tag{32}$$

Having an updated estimate of κ allows us to use $\phi = \exp(\frac{-\kappa}{D}) \times \psi$ instead of ψ directly such that $\log|\det \nabla\phi(\mathbf{x})| = \log|\det \nabla\psi(\mathbf{x})| - \kappa$ which as zero mean. For numerical stability reasons, we can impose on the neural networks s_ℓs to have a final element-wise $\frac{3 \times \tanh}{\sum_{\ell=1}^{L}(D-d_\ell)}$ layer so that $\kappa \in [-3, 3]$ is bounded and thus for the local stretching amplitude it gives: $J^\ell_{\text{min/max}} = \exp(\frac{\pm 3}{D-d_\ell})$.

Recently, some Generative Adversarial Networks articles by Pan et al. [2019] and Detlefsen et al. [2019] report some evidence that the Kantorovich-Rubinstein formulation implemented by neural networks in Lipschitz functions [Miyato et al., 2018] has positive regularization effects on the computation and optimization of the Wasserstein-based losses. To benefit from these advances, let's recall what we wanted:

$$\max_{\mathcal{F},\phi} \left(\lim_{t \to 0^+} \frac{W_{d_\phi}(p, q_{t,\mathcal{F}})}{t} \right) \tag{33}$$

Thanks to that Kantorovich-Rubinstein duality (used for example for Wasserstein Generative Adversarial Networks first by Arjovsky et al. [2017]), we know that for sufficiently low value of t (with some approximation):

$$W_{d_\phi}(p, q_{t,\mathcal{F}}) = W_{L_2}(p_\phi, q_{\phi,t,\mathcal{F}}) = \max_{\mathcal{C} \in \text{Lip}_1} \mathbb{E}_{\mathbf{x} \sim p} \left[\mathcal{C}(\phi(\mathbf{x})) - \mathcal{C}(\phi(\mathbf{x} + t \times \mathcal{F}(\mathbf{x}))\right] \tag{34}$$

In fine, we can forget about the limit operator in Eq. (33) for a low value of t because the optimal transport is the natural transport even in the Kantorovich-Rubinstein formulation and thus we have:

$$\max_{\theta_\mathcal{F},\theta_\phi,\theta_\mathcal{C}} \mathcal{L}(\theta_\mathcal{F}, \theta_\phi, \theta_\mathcal{C}) \tag{35}$$

with

$$\mathcal{L}(\theta_\mathcal{F}, \theta_\phi, \theta_\mathcal{C}) = \frac{\mathbb{E}_{\mathbf{x} \sim p} \left[\mathcal{C}(\phi(\mathbf{x})) - \mathcal{C}(\phi(\mathbf{x} + t \times \mathcal{F}(\mathbf{x}))\right]}{t} \tag{36}$$

and in a stochastic gradient descent strategy, the only important quantity is a bias-free estimate of the gradient [Robbins and Monro, 1951]:

$$\hat{\nabla}\mathcal{L}(\theta_\mathcal{F}, \theta_\phi, \theta_\mathcal{C}) = \frac{\nabla \left[\sum_{b=1}^{B} \mathcal{C}(\phi(\mathbf{x}_{i_b})) - \mathcal{C}(\phi(\mathbf{x}_{i_b} + t \times \mathcal{F}(\mathbf{x}_{i_b}))\right]}{t \times B} \tag{37}$$

for a small t and some B random indices $i_b \sim \mathcal{U}_{\mathbb{N}}(1, N)$.

In conclusion of this section, the optimization problem finally gets:

$$\max_{\theta_{\mathcal{F}},\theta_{\phi},\theta_{\mathcal{C}}} \left(\frac{\mathbb{E}_{\mathbf{x}\sim p}\left[\mathcal{C}\left(\phi(\mathbf{x})\right) - \mathcal{C}\left(\phi(\mathbf{x} + t \times \mathcal{F}(\mathbf{x})\right)\right]}{t} \right) \tag{38}$$

such that ϕ is bijective, $\mathcal{C}$ is 1-Lipschitz and $\mathcal{F}$ Jacobian the BA and ZGLA properties in Eq. (13) and Eq. (14) for small t given by Eq. (29).

4 Algorithm

```
**Input:** Data $(\mathbf{x}_i)_{i=1,\ldots,N}$ where $\mathbf{x}_i \in \mathbb{R}^D$, a mini-batch size $B$
**Initialization:**
    $\theta_{\mathcal{F}}$ initialized such that $\mathcal{F}$ is uniform throughout coordinates
    $\theta_{\phi}$ initialized such that $\phi$ corresponds to the identity matrix
    $\theta_{\mathcal{C}}$ for the 1-Lipschitz critic function
    $\theta = \{\theta_{\mathcal{F}}, \theta_{\phi}, \theta_{\mathcal{C}}\}$
$\kappa \leftarrow 0$
$T \leftarrow 0$
**while** $\theta$ has not converged **do**
    Free all gradients accumulators
    Sample a mini-batch of size $B$ from the dataset
        $\mathbf{x}_{i_b} \quad i_b \sim \mathcal{U}_{\mathbb{N}}(1, N) \quad \text{for } b = 1, \ldots, B$
    Sample $B$ values from a pseudo-random generator
        $u_b \sim \mathcal{U}_{\mathbb{R}}(0, 1) \quad \text{for } b = 1, \ldots, B$
    $t \leftarrow \frac{\min_{b,b'} \|\phi(\mathbf{x}_{i_b}) - \phi(\mathbf{x}_{i_{b'}})\|_2}{2 \times J_{\max}}$
    **while** $\max_b \|\phi(\mathbf{x}_{i_b}) - \phi(\mathbf{x}_{i_b} + t \times \mathcal{F}(\mathbf{x}_{i_b}))\|_2 \leq \frac{1}{2} \min_{b,b'} \|\phi(\mathbf{x}_{i_b}) - \phi(\mathbf{x}_{i_{b'}})\|_2$ is not satisfied **do**
        $t \leftarrow \frac{t}{2}$
    **end while**
    $\mathbf{y}_b \leftarrow \mathbf{x}_{i_b} + t \times \mathcal{F}(\mathbf{x}_{i_b}) \quad \text{for } b = 1, \ldots, B$
    $\tilde{\mathbf{x}}_b \leftarrow u_b \times \mathbf{x}_{i_b} + (1 - u_b) \times \mathbf{y}_b \quad \text{for } b = 1, \ldots, B$
    $\tilde{\mathbf{x}}_b^0 = \tilde{\mathbf{x}}_b \text{ and } \tilde{\mathbf{x}}_b^{\ell} = \text{Layer}(\tilde{\mathbf{x}}_b^{\ell-1}) \quad \text{for } b = 1, \ldots, B \text{ and } \ell = 1, \ldots, K$
    $\kappa \leftarrow \kappa + \frac{1}{T+B} \times \left( \Big( \sum_{b=1}^{B} \sum_{\ell=1}^{L} s_{\ell}(\tilde{\mathbf{x}}_{b,[:d]}^{\ell}) \Big) - \kappa) \right) \quad \text{for } b = 1, \ldots, B$
    $T \leftarrow T + B$
    $\mathbf{a}_b \leftarrow \exp(\frac{-\kappa}{D}) \times \psi(\mathbf{x}_{i_b}) \quad \text{for } b = 1, \ldots, B$
    $\mathbf{b}_b \leftarrow \exp(\frac{-\kappa}{D}) \times \psi(\mathbf{y}_b) \quad \text{for } b = 1, \ldots, B$
    $\mathcal{L} \leftarrow \frac{1}{t \times B} \sum_{b=1}^{B} (\mathcal{C}(\mathbf{a}_b) - \mathcal{C}(\mathbf{b}_b))$
    Perform a gradient ascent step with $\mathcal{L}$ over $\theta$
**end while**
```

Figure 2: Unsupervised Feature Importance Algorithm

5 Possible Computer Vision Applications

For the specific computer vision scientific field, pixels are made of one or three coordinates (for gray levels or color images respectively), so showing relevant pixels in images of the same category could lead to an unsupervised foreground/background segmentation where the only supervision is the fact that all images belong to the same semantic category. A similar problem was tackled by Joulin et al. [2010] a few years ago but with a little more supervision: we have access to several images categories labels and they call the problem cosegmentation. As they stated:

> Purely bottom-up, unsupervised segmentation of a single image into foreground and background regions remains a challenging task for computer vision.

This remains true for all kind of data but without different meanings. Fundamentally, if one has a dataset, one could interpret relevance measurements this kind of algorithms could provide. In computer vision, the core idea of co-segmentation is that the availability of multiple images that contain instances of the same "object" classes makes up for the absence of detailed supervisory information. Some research has been efficiently conducted for interactive foreground / background segementation [Rother et al., 2004] but here we would want to avoid user interaction and benefit from a whole dataset: a class of data sharing a common pattern that we want to highlight. The only form of supervision is knowing that data share some information of interest without knowing what precisely.

6 Future Work and Conclusion

Investigating Wasserstein distances with varying metric seems promising for future work. This sketch of contribution can be a stepstone answer to metric invariance pointed out by Kleinberg [2015] for clustering (which is an unsupervised task like feature importance extraction in this work). This shows that this work can be improved in terms of machine learning and optimization and also engineering on real world data.

Indeed, computer vision in general and foreground/background unsupervised segmentation (in the way we present it) in particular are ways to provide a better understanding between unsupervised metric learning and feature importance extraction thanks to large cardinality datasets.

References

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2017. URL `http://proceedings.mlr.press/v70/arjovsky17a.html`.

L. Bottou and Y. Bengio. Convergence Properties of the K-Means Algorithms. In *Advances in Neural Information Processing Systems*, pages 585–592, 1995.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, oct 2001. ISSN 08856125.

L. Breiman. *Classification and regression trees*. Routledge, 2017. ISBN 9781351460491.

C. R. de Sá. Variance-Based Feature Importance in Neural Networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11828 LNAI, pages 306–315. Springer, 2019. ISBN 9783030337773. doi: 10.1007/978-3-030-33778-0_24.

N. S. Detlefsen, M. Jørgensen, and S. Hauberg. Reliable training and estimation of variance networks. In *Advances in Neural Information Processing Systems*, pages 6323–6333, 2019. URL `http://arxiv.org/abs/1906.03260`.

L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp. *International Conference on Learning Representations*, 2017.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.

R. Jenatton. *Structured sparsity-inducing norms: Statistical and algorithmic properties with applications to neuroimaging*. PhD thesis, 2011.

A. Joulin, F. R. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1943–1950. IEEE, 2010. ISBN 9781424469840. URL `https://doi.org/10.1109/CVPR.2010.5539868`.

J. M. Kleinberg. An impossibility theorem for clustering. In *Advances in Neural Information Processing Systems*, pages 463–470, 2015.

R. T. Knight. Neural networks debunk phrenology. *Science*, 316(5831):1578–1579, 2007. ISSN 00368075. doi: 10.1126/science.1144677.

N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, J. Reynolds, A. Melnikov, N. Lunova, and O. Reblitz-Richardson. Pytorch captum. `https://github.com/pytorch/captum`, 2019.

T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

K. Murphy. *Machine Learning, a Probabilistic Perspective*. MIT press, 2012.

Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, and Y. Zheng. Recent Progress on Generative Adversarial Networks (GANs): A Survey. *IEEE Access*, 7:36322–36333, 2019. ISSN 21693536. doi: 10.1109/ACCESS.2019.2905015.

H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

C. Rother, V. Kolmogorov, and A. Blake. "grabcut": interactive foreground extraction using iterated graph cuts. *ACM SIGGRAPH*, 23(3):309–314, 2004.