# Appendix: A Free-Energy Principle for Representation Learning

## A. Details of the experimental setup

**Datasets.** We use the MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky, 2009) datasets for these experiments. The former consists of 28 $\times$28-sized gray-scale images of handwritten digits (60,000 training and 10,000 validation). The latter consists of 32$\times$32-sized RGB images (50,000 training and 10,000 for validation) spread across 10 classes; 4 of these classes (airplane, automobile, ship, truck) are transportation-based while the others are images of animals and birds.

**Architecture and training.** All models in our experiments consist of an encoder-decoder pair along with a classifier that takes in the latent representation as input. For experiments on MNIST, both encoder and decoder are multi-layer perceptrons with 2 fully-connected layers, the decoder uses a mean-square error loss, i.e., a Gaussian reconstruction likelihood and the classifier consists of a single fully-connected layer. For experiments on CIFAR-10, we use a residual network (He et al., 2016) with 18 layers as an encoder and a decoder with one fully-connected layer and 4 deconvolutional layers (Noh et al., 2015). The classifier network for CIFAR-10 is a single fully-connected layer. All models use ReLU non-linearities and batch-normalization (Ioffe & Szegedy, 2015). Further details of the architecture are given in Appendix A. We use Adam (Kingma & Ba, 2014) to train all models with cosine learning rate annealing.

The encoder and decoder for MNIST has 784–256–16 neurons on each layer; the encoding $z$ is thus 16-dimensional which is the input to the decoder. The classifier has one hidden layer with 12 neurons and 10 outputs. The encoder for CIFAR-10 is a 18-layer residual neural network (ResNet-18) and the decoder has 4 deconvolutional layers. We used a slightly larger network for the geodesic transfer learning experiment on MNIST. The encoder and decoder have 784–400–64 neurons in each layer with a dropout of probability 0.1 after the hidden layer. The classifier has a single layer that takes the 64-dimensional encoding and predicts 10 classes.

## B. Proof of Lemma 3

The second statement directly follows by observing that $F$ is a minimum of affine functions in $(\lambda, \gamma)$. To see the first, evaluate the Hessian of $R$ and $F$

$$\mathrm{Hess}(R)\ \mathrm{Hess}(F) = \begin{pmatrix} \frac{\partial^2 R}{\partial D^2} & \frac{\partial^2 R}{\partial D \partial C} \\ \frac{\partial^2 R}{\partial C \partial D} & \frac{\partial^2 R}{\partial C^2} \end{pmatrix} \begin{pmatrix} \frac{\partial^2 F}{\partial \lambda^2} & \frac{\partial^2 F}{\partial \lambda \partial \gamma} \\ \frac{\partial^2 F}{\partial \gamma \partial \lambda} & \frac{\partial^2 F}{\partial \gamma^2} \end{pmatrix}$$

Since we have $F = \min_{e_\theta(z|x), d_\theta(x|z), m_\theta(z)} R + \lambda D + \gamma C$, we obtain

$$\lambda = -\frac{\partial R}{\partial D}, \quad \gamma = -\frac{\partial R}{\partial C}, \quad D = \frac{\partial F}{\partial \lambda}, \quad C = \frac{\partial F}{\partial \gamma}.$$

We then have

$$\begin{aligned} \mathrm{d}\lambda = -\mathrm{d}\left(\frac{\partial R}{\partial D}\right) &= -\frac{\partial^2 R}{\partial D^2}\,\mathrm{d}D - \frac{\partial^2 R}{\partial D \partial C}\,\mathrm{d}C \\ &= -\frac{\partial^2 R}{\partial D^2}\left(\frac{\partial D}{\partial \lambda} d\lambda + \frac{\partial D}{\partial \gamma} d\gamma\right) - \frac{\partial^2 R}{\partial D \partial C}\left(\frac{\partial C}{\partial \lambda} d\lambda + \frac{\partial C}{\partial \gamma} d\gamma\right) \\ &= -\left(\frac{\partial^2 R}{\partial D^2}\frac{\partial^2 F}{\partial \lambda^2} + \frac{\partial^2 R}{\partial D \partial C}\frac{\partial^2 F}{\partial \gamma \partial \lambda}\right)\mathrm{d}\lambda - \left(\frac{\partial^2 R}{\partial D^2}\frac{\partial^2 F}{\partial \lambda \partial \gamma} + \frac{\partial^2 R}{\partial D \partial C}\frac{\partial^2 F}{\partial \gamma^2}\right)\mathrm{d}\gamma; \end{aligned}$$

$$\begin{aligned}
\mathrm{d}\gamma = -\mathrm{d}\left(\frac{\partial R}{\partial C}\right) &= -\frac{\partial^2 R}{\partial C \partial D}\,\mathrm{d}D - \frac{\partial^2 R}{\partial C^2}\,\mathrm{d}C \\
&= -\frac{\partial^2 R}{\partial C \partial D}\,\left(\frac{\partial D}{\partial \lambda}d\lambda + \frac{\partial D}{\partial \gamma}d\gamma\right) - \frac{\partial^2 R}{\partial C^2}\,\left(\frac{\partial C}{\partial \lambda}d\lambda + \frac{\partial C}{\partial \gamma}d\gamma\right) \\
&= -\left(\frac{\partial^2 R}{\partial C \partial D}\frac{\partial^2 F}{\partial \lambda^2} + \frac{\partial^2 R}{\partial C^2}\frac{\partial^2 F}{\partial \gamma \partial \lambda}\right)\,\mathrm{d}\lambda - \left(\frac{\partial^2 R}{\partial C \partial D}\frac{\partial^2 F}{\partial \lambda \partial \gamma} + \frac{\partial^2 R}{\partial C^2}\frac{\partial^2 F}{\partial \gamma^2}\right)\,\mathrm{d}\gamma.
\end{aligned}$$

Compare the coefficients on both sides to get

$$\begin{aligned}
\frac{\partial^2 R}{\partial D^2}\frac{\partial^2 F}{\partial \lambda^2} + \frac{\partial^2 R}{\partial D \partial C}\frac{\partial^2 F}{\partial \gamma \partial \lambda} &= \frac{\partial^2 R}{\partial C \partial D}\frac{\partial^2 F}{\partial \lambda \partial \gamma} + \frac{\partial^2 R}{\partial C^2}\frac{\partial^2 F}{\partial \gamma^2} = -1; \\
\frac{\partial^2 R}{\partial D^2}\frac{\partial^2 F}{\partial \lambda \partial \gamma} + \frac{\partial^2 R}{\partial D \partial C}\frac{\partial^2 F}{\partial \gamma^2} &= \frac{\partial^2 R}{\partial C \partial D}\frac{\partial^2 F}{\partial \lambda^2} + \frac{\partial^2 R}{\partial C^2}\frac{\partial^2 F}{\partial \gamma \partial \lambda} = 0,
\end{aligned}$$

therefore

$$\mathrm{Hess}(R)\,\mathrm{Hess}(F) = -I.$$

Since $0 \succ \mathrm{Hess}(F)$, we have that $\mathrm{Hess}(R) \succ 0$, then the constraint surface $f(R, D, C) = 0$ is convex.

## C. Proof of Lemma 5

Recall the definition of the objective function (14), first we compute the gradient of the objective function as following:

$$\begin{aligned}
\nabla_\theta J(\theta, \lambda, \gamma) &= -\mathop{\mathbb{E}}_{x\sim p(x)} \nabla_\theta \log Z_{\theta,x} \\
&= -\mathop{\mathbb{E}}_{x\sim p(x)} \frac{1}{Z_{\theta,x}} \nabla_\theta Z_{\theta,x} \\
&= -\mathop{\mathbb{E}}_{x\sim p(x)} \frac{1}{Z_{\theta,x}} \int (-\nabla_\theta H)\,\exp(-H)\,\mathrm{d}z \\
&= \mathop{\mathbb{E}}_{x\sim p(x)} \langle \nabla_\theta H \rangle
\end{aligned}$$

Then with some effort of computation, we get

$$\begin{aligned}
A = \nabla_\theta^2 J(\theta, \lambda, \gamma) &= \nabla_\theta \mathop{\mathbb{E}}_{x\sim p(x)} \left[\frac{1}{Z_{\theta,x}} \int \nabla_\theta H\,\exp(-H)\,\mathrm{d}z\right] \\
&= \mathop{\mathbb{E}}_{x\sim p(x)} \left[-\frac{1}{Z_{\theta,x}^2}\left(\int (-\nabla_\theta H)\,\exp(-H)\,\mathrm{d}z\right)\left(\int \nabla_\theta^T H\,\exp(-H)\,\mathrm{d}z\right) + \frac{1}{Z_{\theta,x}}\int \nabla_\theta^2 H\,\exp(-H)\,\mathrm{d}z - \frac{1}{Z_{\theta,x}}\int \nabla_\theta H\,\nabla_\theta^\top H\,\exp(-H)\,\mathrm{d}z\right] \\
&= \mathop{\mathbb{E}}_{x\sim p(x)} \left[\left\langle \nabla_\theta^2 H\right\rangle + \langle \nabla_\theta H\rangle \langle \nabla_\theta H\rangle^\top - \left\langle \nabla_\theta H\,\nabla_\theta^\top H\right\rangle\right]; \\
b_\lambda = -\frac{\partial}{\partial \lambda}\nabla_\theta J &= -\frac{\partial}{\partial \lambda} \mathop{\mathbb{E}}_{x\sim p(x)} \left[\frac{1}{Z_{\theta,x}} \int \nabla_\theta H\,\exp(-H)\,\mathrm{d}z\right] \\
&= -\mathop{\mathbb{E}}_{x\sim p(x)} \left[-\frac{1}{Z_{\theta,x}^2}\left(\int -\frac{\partial H}{\partial \lambda}\,\exp(-H)\,\mathrm{d}z\right)\left(\int \nabla_\theta H\,\exp(-H)\,\mathrm{d}z\right) + \frac{1}{Z_{\theta,x}}\int \frac{\partial}{\partial \lambda}\nabla_\theta H\,\exp(-H)\,\mathrm{d}z - \frac{1}{Z_{\theta,x}}\int \frac{\partial H}{\partial \lambda}\,\nabla_\theta H\,\exp(-H)\,\mathrm{d}z\right] \\
&= -\mathop{\mathbb{E}}_{x\sim p(x)} \left[\left\langle \frac{\partial \nabla_\theta H}{\partial \lambda}\right\rangle - \left\langle \frac{\partial H}{\partial \lambda}\,\nabla_\theta H\right\rangle + \left\langle \frac{\partial H}{\partial \lambda}\right\rangle \langle \nabla_\theta H\rangle\right]; \\
b_\gamma = -\frac{\partial}{\partial \gamma}\nabla_\theta J &= -\frac{\partial}{\partial \gamma} \mathop{\mathbb{E}}_{x\sim p(x)} \left[\frac{1}{Z_{\theta,x}} \int \nabla_\theta H\,\exp(-H)\,\mathrm{d}z\right] \\
&= -\mathop{\mathbb{E}}_{x\sim p(x)} \left[-\frac{1}{Z_{\theta,x}^2}\left(\int -\frac{\partial H}{\partial \gamma}\,\exp(-H)\,\mathrm{d}z\right)\left(\int \nabla_\theta H\,\exp(-H)\,\mathrm{d}z\right) + \frac{1}{Z_{\theta,x}}\int \frac{\partial}{\partial \gamma}\,\nabla_\theta H\,\exp(-H)\,\mathrm{d}z - \frac{1}{Z_{\theta,x}}\int \frac{\partial H}{\partial \gamma}\,\nabla_\theta H\,\exp(-H)\,\mathrm{d}z\right] \\
&= -\mathop{\mathbb{E}}_{x\sim p(x)} \left[\left\langle \frac{\partial \nabla_\theta H}{\partial \gamma}\right\rangle - \left\langle \frac{\partial H}{\partial \gamma}\,\nabla_\theta H\right\rangle + \left\langle \frac{\partial H}{\partial \gamma}\right\rangle \langle \nabla_\theta H\rangle\right].
\end{aligned}$$

According to the quasi-static constraints (16), we have

$$A\dot{\theta} - \dot{\lambda} b_\lambda - \dot{\gamma} b_\gamma = 0,$$

that implies

$$\dot{\theta} = A^{-1} b_\lambda \, \dot{\lambda} + A^{-1} b_\gamma \, \dot{\gamma} = \theta_\lambda \dot{\lambda} + \theta_\gamma \dot{\gamma}. \tag{32}$$

## D. Computation of Iso-classification constraint

We start with computing the gradient of classification loss, clear that $C = \mathbb{E}_{x\sim p(x)} \left[ - \int \mathrm{d}z \; e(z|x) \log c(y|z) \right] = - \mathbb{E}_{x\sim p(x)} \langle \ell \rangle$, where $\ell = \log c_\theta(y_x|z)$ is the logarithm of the classification loss, then

$$\begin{aligned}
\nabla_\theta C &= -\nabla_\theta \mathop{\mathbb{E}}_{x\sim p(x)} \left[ \frac{1}{Z_{\theta,x}} \int \ell \, \exp(-H) \, \mathrm{d}z \right] \\
&= - \mathop{\mathbb{E}}_{x\sim p(x)} \left[ -\frac{1}{Z_{\theta,x}^2} \left( \int (-\nabla_\theta H) \, \exp(-H) \, \mathrm{d}z \right) \left( \int \ell \, \exp(-H) \, \mathrm{d}z \right) + \frac{1}{Z_{\theta,x}} \int \nabla_\theta \, \ell \, \exp(-H) \, \mathrm{d}z - \frac{1}{Z_{\theta,x}} \int \ell \, \nabla_\theta H \, \exp(-H) \, \mathrm{d}z \right] \\
&= - \mathop{\mathbb{E}}_{x\sim p(x)} \left[ \langle \nabla_\theta \, \ell \rangle + \langle \nabla_\theta H \rangle \, \langle \ell \rangle - \langle \ell \, \nabla_\theta H \rangle \right]; \\
\frac{\partial}{\partial \lambda} C &= -\frac{\partial}{\partial \lambda} \mathop{\mathbb{E}}_{x\sim p(x)} \left[ \frac{1}{Z_{\theta,x}} \int \ell \, \exp(-H) \, \mathrm{d}z \right] \\
&= - \mathop{\mathbb{E}}_{x\sim p(x)} \left[ -\frac{1}{Z_{\theta,x}^2} \left( \int -\frac{\partial H}{\partial \lambda} \, \exp(-H) \, \mathrm{d}z \right) \left( \int \ell \, \exp(-H) \, \mathrm{d}z \right) - \frac{1}{Z_{\theta,x}} \int \ell \, \frac{\partial H}{\partial \lambda} \, \exp(-H) \, \mathrm{d}z \right] \\
&= - \mathop{\mathbb{E}}_{x\sim p(x)} \left[ \left\langle \frac{\partial H}{\partial \lambda} \right\rangle \langle \ell \rangle - \left\langle \ell \, \frac{\partial H}{\partial \lambda} \right\rangle \right];
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \gamma} C &= -\frac{\partial}{\partial \gamma} \mathop{\mathbb{E}}_{x\sim p(x)} \left[ \frac{1}{Z_{\theta,x}} \int \ell \, \exp(-H) \, \mathrm{d}z \right] \\
&= - \mathop{\mathbb{E}}_{x\sim p(x)} \left[ -\frac{1}{Z_{\theta,x}^2} \left( \int -\frac{\partial H}{\partial \lambda} \, \exp(-H) \, \mathrm{d}z \right) \left( \int \ell \, \exp(-H) \, \mathrm{d}z \right) - \frac{1}{Z_{\theta,x}} \int \ell \, \frac{\partial H}{\partial \gamma} \, \exp(-H) \, \mathrm{d}z \right] \\
&= - \mathop{\mathbb{E}}_{x\sim p(x)} \left[ \left\langle \frac{\partial H}{\partial \gamma} \right\rangle \langle \ell \rangle - \left\langle \ell \, \frac{\partial H}{\partial \gamma} \right\rangle \right].
\end{aligned}$$

The iso-classification loss constrains together with quasi-static constrains imply that:

$$\begin{aligned}
0 &\equiv \frac{\mathrm{d}}{\mathrm{d}t} C \\
&= \dot{\theta}^\top \, \nabla_\theta C + \dot{\lambda} \frac{\partial C}{\partial \lambda} + \dot{\gamma} \frac{\partial C}{\partial \gamma} \\
&= \dot{\lambda} \left( \theta_\lambda^\top \, \nabla_\theta C + \frac{\partial C}{\partial \lambda} \right) + \dot{\gamma} \left( \theta_\gamma^\top \, \nabla_\theta C + \frac{\partial C}{\partial \gamma} \right) \\
&= -\dot{\lambda} \mathop{\mathbb{E}}_{x\sim p(x)} \left[ \left\langle \frac{\partial H}{\partial \lambda} \right\rangle \langle \ell \rangle - \left\langle \ell \, \frac{\partial H}{\partial \lambda} \right\rangle + \langle \theta_\lambda^\top \nabla_\theta H \rangle \, \langle \ell \rangle - \langle \ell \theta_\lambda^\top \nabla_\theta H \rangle + \langle \theta_\lambda^\top \nabla_\theta \ell \rangle \right] - \dot{\gamma} \mathop{\mathbb{E}}_{x\sim p(x)} \left[ \left\langle \frac{\partial H}{\partial \gamma} \right\rangle \langle \ell \rangle - \left\langle \ell \, \frac{\partial H}{\partial \gamma} \right\rangle + \langle \theta_\gamma^\top \nabla_\theta H \rangle \, \langle \ell \rangle - \langle \ell \theta_\gamma^\top \nabla_\theta H \rangle + \langle \theta_\gamma^\top \nabla_\theta \ell \rangle \right] \\
&= C_\lambda \dot{\lambda} + C_\gamma \dot{\gamma},
\end{aligned}$$

where the third equation is followed by the equilibrium dynamics (17) for parameters $\theta$. So far we developed the constrained dynamics for iso-classification process:

$$\begin{aligned}
0 &= C_\lambda \dot{\lambda} + C_\gamma \dot{\gamma} \\
\dot{\theta} &= \theta_\lambda \dot{\lambda} + \theta_\gamma \dot{\gamma}.
\end{aligned} \tag{33}$$

## E. Iso-classification equations for changing data distribution

In this section we analyze the dynamics for iso-classification loss process when the data distribution evolves with time. $\frac{\partial p(x)}{\partial t}$ will lead to additional terms that represent the partial derivatives with respect to $t$ on both the quasi-static and iso-classification constrains. More precisely, the new terms are

$$b_t = -\frac{\partial}{\partial t}\nabla_\theta J = -\int \frac{\partial p(x)}{\partial t}\langle \nabla_\theta H\rangle \,\mathrm{d}x;$$
$$\frac{\partial}{\partial t}C = -\int \frac{\partial p(x)}{\partial t}\langle \ell\rangle \,\mathrm{d}x,$$

then the quasi-static and iso-classification constraints are ready to be modified as

$$\begin{aligned}
0 \equiv \frac{\mathrm{d}}{\mathrm{d}t}\nabla_\theta J(\theta,\lambda,\gamma) &\iff 0 = \nabla_\theta^2 F\,\dot\theta + \dot\lambda\,\frac{\partial \nabla_\theta F}{\partial \lambda} + \dot\gamma\,\frac{\partial \nabla_\theta F}{\partial \gamma} + \frac{\partial \nabla_\theta F}{\partial t}\\
&\iff \dot\theta = \dot\lambda\, A^{-1}\, b_\lambda + \dot\gamma\, A^{-1}\, b_\gamma + A^{-1}\, b_t\\
&\iff \dot\theta = \dot\lambda\,\theta_\lambda + \dot\gamma\,\theta_\gamma + \theta_t;\\
0 \equiv \frac{\mathrm{d}}{\mathrm{d}t}C &\iff 0 = \dot\theta^\top\,\nabla_\theta C + \dot\lambda\frac{\partial C}{\partial \lambda} + \dot\gamma\frac{\partial C}{\partial \gamma} + \frac{\partial C}{\partial t}\\
&\iff 0 = \dot\lambda\left(\theta_\lambda^\top\,\nabla_\theta C + \frac{\partial C}{\partial \lambda}\right) + \dot\gamma\left(\theta_\gamma^\top\,\nabla_\theta C + \frac{\partial C}{\partial \gamma}\right) + \left(\theta_t^\top\,\nabla_\theta C + \frac{\partial C}{\partial t}\right)\\
&\iff 0 = \dot\lambda\, C_\lambda + \dot\gamma\, C_\gamma + C_t,
\end{aligned}$$

where $A$, $b_\lambda$, $b_\gamma$, $C_\lambda$ and $C_\gamma$ where $C_\lambda$ and $C_\gamma$ are as given in lemma 5 and (21) with the only change being that the outer expectation is taken with respect to $x \sim p(x,t)$. The new terms that depends on time $t$ are

$$C_t = -\int \frac{\partial p(x,t)}{\partial t}\langle \ell\rangle \,\mathrm{d}x - \mathop{\mathbb{E}}_{x\sim p(x,t)}\left[\left\langle \theta_t^\top \nabla_\theta H\right\rangle\langle \ell\rangle - \left\langle \theta_t^\top \nabla_\theta H\,\ell\right\rangle + \left\langle \theta_t^\top \nabla_\theta \ell\right\rangle\right] \tag{34}$$

with $\ell = \log c_\theta(y_{x_t}|z)$. We can combine modified quasi-static and iso-classification constraints to get

$$\begin{aligned}
\dot\theta &= \left(\theta_\lambda - \frac{C_\lambda}{C_\gamma}\,\theta_\gamma\right)\dot\lambda + \left(\theta_t - \frac{C_t}{C_\gamma}\theta_\gamma\right).\\
&=: \hat\theta_\lambda \dot\lambda + \hat\theta_t
\end{aligned} \tag{35}$$

This indicates that $\theta = \theta(\lambda, t)$ is a surface parameterized by $\lambda$ and $t$, equipped with a basis of tangent plane $(\hat\theta_\lambda, \hat\theta_t)$.

## F. Optimally transporting the data distribution

We first give a brief description of the theory of optimal transportation. The optimal transport map between the source task and the target task will be used to define a dynamical process for the task. We only compute the transport for the inputs $x$ between the source and target distributions and use a heuristic to obtain the transport for the labels $y$. This choice is made only to simplify the exposition; it is straightforward to handle the case of transport on the joint distribution $p(x, y)$.

If i.i.d samples from the source task are denoted by $\left\{x_1^s, \ldots, x_{n_s}^s\right\}$ and those of the target distribution are $\left\{x_1^t, \ldots, x_{n_t}^t\right\}$ the empirical source and target distributions can be written as

$$p^s(x) = \frac{1}{n_s}\sum_{i=1}^{n_s}\delta_{x-x_i^s}, \text{and } p^t(x) = \frac{1}{n_t}\sum_{i=1}^{n_t}\delta_{x-x_i^t}$$

respectively; here $\delta_{x-x'}$ is a Dirac delta distribution at $x'$. Since the empirical data distribution is a sum of a finite number of Dirac measures, this is a discrete optimal transport problem and easy to solve. We can use the Kantorovich relaxation to denote by $\mathcal{B}$ the set of probabilistic couplings between the two distributions:

$$\mathcal{B} = \left\{\Gamma \in \mathbb{R}_+^{n_s\times n_t} :\ \Gamma \mathbf{1}_{n_s} = p, \Gamma^\top \mathbf{1}_{n_s} = q\right\}$$

where $\mathbf{1}_n$ is an $n$-dimensional vector of ones. The Kantorovich formulation solves for

$$\Gamma^* = \underset{\Gamma \in \mathcal{B}}{\operatorname{argmin}} \sum_{i=1}^{n_s} \sum_{t=1}^{n_t} \Gamma_{ij}\, \kappa_{ij} \tag{36}$$

where $\kappa \in \mathbb{R}_+^{n_s \times n_t}$ is a cost function that models transporting the datum $x_i^s$ to $x_j^t$. This is the metric of the underlying data domain and one may choose any reasonable metric for $\kappa = \|x_i^s - x_j^t\|_2^2$. The problem in (36) is a convex optimization problem and can be solved easily; in practice we use the Sinkhorn's algorithm (Cuturi, 2013) which adds an entropic regularizer $-h(\Gamma) = \sum_{ij} \Gamma_{ij} \log \Gamma_{ij}$ to the objective in (36).

### F.1. Changing the data distribution

Given the optimal probabilistic coupling $\Gamma^*$ between the source and the target data distributions, we can interpolate between them at any $t \in [0, 1]$ by following the geodesics of the Wasserstein metric

$$p(x, t) = \underset{p}{\operatorname{argmin}}\ (1-t) W_2^2(p^s, p) + t W_2^2(p, p^t).$$

For discrete optimal transport problems, as shown in Villani (2008), the interpolated distribution $p_t$ for the metric $\kappa_{ij} = \|x_i^2 - x_j^t\|_2^2$ is given by

$$p(x, t) = \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} \Gamma_{ij}^*\, \delta_{x-(1-t)x_i^s - t x_j^t}. \tag{37}$$

Observe that the interpolated data distribution equals the source and target distribution at $t = 0$ and $t = 1$ respectively and it consists of linear interpolations of the data in between.

**Remark 11 (Interpolating the labels).** The interpolation in (37) gives the marginal on the input space interpolated between the source and target tasks. To evaluate the functionals in Section 3 for the classification setting, we would also like to interpolate the labels. We do so by setting the true label of the interpolated datum $x = (1-t)x_i^s + t x_j^t$ to be linear interpolation between the source label and the target label.

$$y(x, t) = (1-t)\delta_{y - y_{x_i^s}} + t\delta_{y - y_{x_j^t}}$$

for all $i, j$. Notice that the interpolated distribution $p(x, t)$ is a sum of Dirac delta distributions weighted by the optimal coupling. We therefore only need to evaluate the labels at all the interpolated data.

**Remark 12 (Linear interpolation of data).** Our formulation of optimal transportation leads to a linear interpolation of the data in (23). This may not work well for image-based data where the square metric $\kappa_{ij} = \|x_i^s - x - k^t\|_2^2$ may not be the appropriate metric. We note that this interpolation of data is an artifact of our choice of $\kappa_{ij}$, other choices for the metric also fit into the formulation and should be viable alternatives if they result in efficient computation.