A Free-Energy Principle for Representation Learning

Yansong Gao [1] **Pratik Chaudhari** [1]

Abstract

This paper employs a formal connection of machine learning with thermodynamics to characterize the quality of learnt representations for transfer learning. We discuss how information-theoretic functionals such as rate, distortion and classification loss of a model lie on a convex, so-called equilibrium surface. We prescribe dynamical processes to traverse this surface under constraints, e.g., an iso-classification process that trades off rate and distortion to keep the classification loss unchanged. We demonstrate how this process can be used for transferring representations from a source dataset to a target dataset while keeping the classification loss constant. Experimental validation of the theoretical results is provided on standard image-classification datasets.

1. Introduction

A representation is a statistic of the data that is "useful". Classical Information Theory creates a compressed representation and makes it easier to store or transmit data; the goal is always to decode the representation to get the original data back. If we are given images and their labels, we could learn a representation that is useful to predict the correct labels. This representation is thus a statistic of the data *sufficient* for the task of classification. If it is also minimal—say in its size—it would discard information in the data that is not correlated with the labels. Such a representation is unique to the chosen task, it would perform poorly to predict some other labels correlated with the discarded information. If instead the representation were to have lots of redundant information about the data, it could potentially predict other labels correlated with this extra information.

The premise of this paper is our desire to characterize the information discarded in the representation when it is fit on a task. We want to do so in order to learn representations that can be transferred easily to other tasks.

Our main idea is to choose a canonical task—in this paper, we pick reconstruction of the original data—as a way to measure the discarded information. Although one can use any canonical task, reconstruction is special. It is a "capture all" task in the sense that achieving perfect reconstruction entails that the representation is lossless; information discarded by the original task is therefore readily measured as the one that helps solve the canonical task. This leads to the study of the following Lagrangian which is similar to the Information Bottlenck of Tishby et al. (2000)

$$F(\lambda,\gamma) = \min_{\substack{\theta\in\Theta, e_\theta(z|x), m_\theta(z),\\ d_\theta(x|z), c_\theta(y|z)}} R + \lambda D + \gamma C$$

where the rate R is an upper bound on the mutual information of the representation learnt by the encoder $e_\theta(z|x)$ with the input data x, distortion D measures the quality of reconstruction of the decoder $d_\theta(x|z)$ and C measures the classification loss of the classifier $c_\theta(y|z)$. As Alemi & Fischer (2018) show, this Lagrangian can be formally connected to ideas in thermodynamics. We heavily exploit and specialize this point of view, as summarized next.

1.1. Summary of contributions

Our main technical observation is that $F(\lambda,\gamma)$ can be intepreted as a free-energy and a stochastic learning process that minimizes its corresponding Hamiltonian converges to the optimal free-energy. This corresponds to an "equilibrium surface" of information-theoretic functionals R, D and C and a surface $\Theta_{\lambda,\gamma}$ of the model parameters at convergence. We prove that the equilibrium surface is convex and its dual, the free-energy $F(\lambda,\gamma)$, is concave. The free-energy is only a function of Lagrange multipliers (λ,γ), the family of model parameters Θ, and the task, and is therefore invariant of the learning dynamics.

Second, we design a quasi-static stochastic process, akin to an equilibrium process in thermodynamics, to keep the model parameters θ on the equilibrium surface. Such a process allow us to travel to any feasible values of (R, D, C) while ensuring that the parameters θ of the model are on the equilibrium surface. We focus on one process, the "iso-

[1]University of Pennsylvania, USA. Correspondence to: Yansong Gao <gaoyans@sas.upenn.edu>, Pratik Chaudhari <pratikac@seas.upenn.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

classification process" which automatically trades off the rate and distortion to keep the classification loss constant.

We prescribe a quasi-static process that allows for a controlled transfer of learnt representations. It adapts the model parameters as the task is changed from some source dataset to a target dataset while keeping the classification loss constant. Such a process is in stark contrast to current techniques in transfer learning which do not provide any guarantees on the quality of the model on the target dataset.

We provide extensive experimental results which realize the theory developed in this paper.

2. Theroetical setup

This section introduces notation and preliminaries that form the building blocks of our approach.

2.1. Auto-Encoders

Consider an encoder $e(z|x)$ that encodes data x into a latent code z and a decoder $d(x|z)$ that decodes z back into the original data x. If the true distribution of the data is $p(x)$ we may define the following functionals.

$$\begin{aligned} H &= \mathop{\mathbb{E}}_{x\sim p(x)}\left[-\log p(x)\right] \\ D &= \mathop{\mathbb{E}}_{x\sim p(x)}\left[-\int \mathrm{d}z\; e(z|x)\log d(x|z)\right] \\ R &= \mathop{\mathbb{E}}_{x\sim p(x)}\left[\int \mathrm{d}z\; e(z|x)\log \frac{e(z|x)}{m(z)}\right] \end{aligned} \tag{1}$$

We denote expectation over data using the notation $\langle\varphi\rangle_{p(x)} = \int \mathrm{d}x\, p(x)\varphi$. The first functional H is the Shanon entropy of the true data distribution; it quantifies the complexity of the data. The distortion D measures the quality of the reconstruction through its log-likelihood. The rate R is a Kullback-Leibler (KL) divergence; it measures the average excess bits used to encode samples from $e(z|x)$ using a code that was built for our approximation of the true marginal on the latent factors $m(z)$.

2.2. Rate-Distortion curve

The functionals in (1) come together to give the inequality

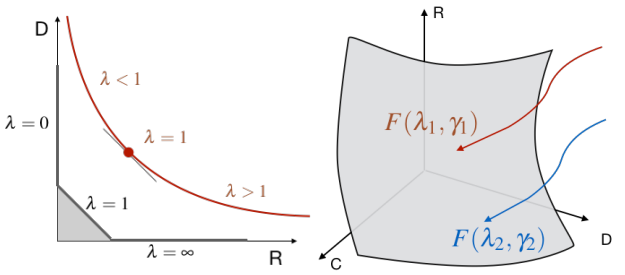
$$H - D \le I_e(x;z) \le R \tag{2}$$

where $I_e = \mathrm{KL}(e(z|x)\;||\;p(z|x))$ is the KL-divergence between the learnt encoder and the true (unknown) conditional of the latent factors. The outer inequality $H \le D + R$ forms the basis for a large body of literature on Evidence Lower Bounds (ELBO, see Kingma & Welling (2013)). Consider Fig. 1a, if the capacity of our candidate distributions $e(z|x), m(z)$ and $d(x|z)$ is infinite, we can obtain the equality $H = R + D$. This is the thick black line in Fig. 1a.

For finite capacity variational families, say parameterized by θ, which we denote by $e_\theta(z|x)$, $d_\theta(x|z)$ and $m_\theta(z)$ respectively, as Alemi et al. (2017) argue, one obtains a convex RD curve (shown in red in Fig. 1a) corresponding to the Lagrangian

$$F(\lambda) = \min_{e_\theta(z|x), m_\theta(z), d_\theta(x|z)} R + \lambda D. \tag{3}$$

This Lagrangian is the relaxation of the idea that given a fixed variational family and data distribution $p(x)$, there exists an optimal value of, say, rate $R = f(D)$ that best sandwiches (2). The optimal Lagrange multiplier is $\lambda = \frac{\partial R}{\partial D}$ evaluated at the desired value of D.



(a) (b)

Figure 1. **Schematic of the equilibrium surface.** Fig. 1a shows that rate (R) and distortion (D) trade off against each other on the equilibrium surface. Similarly in Fig. 1b, the equilibrium surface is a convex constraint that joins rate, distortion and the classification loss. Training objectives with different (λ, γ) (shown in red and blue) reach different parts of the equilibrium surface.

2.3. Incorporating the classification loss

Let us create a classifier that uses the learnt representation z as the input and set the classification loss as the negative log-likelihood of the prediction

$$C = \mathop{\mathbb{E}}_{x\sim p(x)}\left[-\int \mathrm{d}z\; e(z|x)\log c(y|z)\right]. \tag{4}$$

If the parameters of the model—which now consists of the encoder $e(z|x)$, decoder $d(x|z)$ and the classifier $c(y|z)$—are denoted by θ, the training process for the model induces a distribution $p(\theta|\,\{(x,y)\})$ where $\{(x,y)\}$ denotes a finite dataset. In addition to R, D and C, the authors in Alemi & Fischer (2018) define

$$S = \mathop{\mathbb{E}}_{x\sim p(x), y\sim p(y|x)}\left[\log \frac{p(\theta|\,\{x,y\})}{m(\theta)}\right] \tag{5}$$

which is the relative entropy of the distribution on parameters θ after training compared to a prior distribution $m(\theta)$ of our choosing. Using a very similar argument as Section 2.2 the four functionals R, D, C and S form a convex three-dimensional surface in the RDCS phase space. A schematic

is shown in Fig. 1b for $\sigma = 0$. We can again consider a Lagrange relaxation of this surface given by

$$F(\lambda, \gamma, \sigma) = \min_{e(z|x), m(z), d(x|z), c(y|z)} R + \lambda D + \gamma C + \sigma S. \tag{6}$$

Remark 1 ('The 'First Law" of learning). Alemi & Fischer (2018) draw formal connections of the Lagrangian in (6) with the theory of thermodynamics. Just like the first law of thermodynamics is a statement about the conservation of energy in physical processes, the fact that the four functionals are tied together in a smooth constraint $f(R, D, C, S) = 0$ leads to an equation of the form

$$\mathrm{d}R = -\lambda \, \mathrm{d}D - \gamma \, \mathrm{d}C - \sigma \, \mathrm{d}S \tag{7}$$

which indicates that information in learning processes is conserved. The information in the latent representation z is kept either to reconstruct back the original data or to predict the labels. The former is captured by the encoder-decoder pair, the latter is captured by the classifier.

Remark 2 (Setting $\sigma = 0$). The distribution $p(\theta | \{(x, y)\})$ is a posterior on the parameters of the model given the dataset. While this distribution is well-defined under minor technical conditions, e.g., ergodicity, performing computations with this distribution is difficult. **We therefore only consider the case when $\sigma = 0$ in the sequel** and leave the general case for future work.

The following lemma (proved in Appendix B) shows that the constraint surface connecting the information-theoretic functionals R, D and C is convex and its dual, the Lagrangian $F(\lambda, \gamma)$ is concave.

Lemma 3 (The RDC constraint surface is convex). *The constraint surface $f(R, D, C) = 0$ is convex and the Lagrangian $F(\lambda, \gamma)$ is concave.*

We can show using a similar proof that the entire surface joining R, D, C and S is convex by considering the cases $\lambda = 0$ and $\gamma = 0$ separately. Note that the constraint is convex in R, D and C; it need not be convex in the model parameters θ that parameterize $e_\theta(z|x), m_\theta(z)$, etc.

2.4. Equilibrium surface of optimal free-energy

We next elaborate upon the objective in (6). Consider the functionals R, D and C parameterized using parameters $\theta \in \Theta \subseteq \mathbb{R}^N$. First, consider the problem

$$F(\lambda, \gamma) = \min_{e(z|x), \, \theta \in \Theta} R + \lambda D + \gamma C. \tag{8}$$

We can solve this using calculus of variations to get

$$e(z|x) \propto m_\theta(z) d_\theta(x|z)^\lambda \exp\left(\gamma \int \mathrm{d}y \; p(y|x) \log c_\theta(y|z)\right).$$

We assume in this paper that the labels are a deterministic function of the data, i.e., $p(y|x) = \delta(y - y_x)$ where y_x is the true label of the datum x. We therefore have

$$e(z|x) = \frac{m_\theta(z) d_\theta(x|z)^\lambda c_\theta(y_x|z)^\gamma}{Z_{\theta,x}}$$

where the normalization constant is

$$Z_{\theta,x} = \int \mathrm{d}z \; m_\theta(z) d_\theta(x|z)^\lambda c_\theta(y_x|z)^\gamma. \tag{9}$$

The objective $F(\lambda, \gamma)$ can now be rewritten as maximizing the log-partition function, also known as the free-energy in statistical physics (Mezard & Montanari, 2009),

$$F(\lambda, \gamma) = \min_{\theta \in \Theta} \; - \left\langle \log Z_{\theta,x} \right\rangle_{p(x)}. \tag{10}$$

Remark 4 (Why is it called the "equilibrium" surface?). Given a finite dataset $\{(x, y)\}$, one may minimize the objective in (8) using stochastic gradient descent (SGD, Robbins & Monro (1951)) on a Hamiltonian

$$H(z; x, \theta, \lambda, \gamma) \equiv -\log m_\theta(z) - \lambda \log d_\theta(x|z) - \gamma \log c_\theta(y|z) \tag{11}$$

with updates given by

$$\theta^{k+1} = \theta^k - \sigma \nabla_\theta \mathop{\mathbb{E}}_{x \sim p(x)} \left[\int \mathrm{d}z \; e_{\theta^k}(z|x) H(z; x, \theta^k, \lambda, \gamma) \right] \tag{12}$$

where $\sigma > 0$ is the step-size; the gradient ∇_θ is evaluated over samples from $p(x)$ and $e_\theta(z|x)$. Using the same technique as that of Chaudhari & Soatto (2017), one can show that the objective

$$\mathop{\mathbb{E}}_{\theta \sim p(\theta | \{x,y\})} \left[\left\langle -\log Z_{\theta,x} \right\rangle_{p(x)} \right] - \sigma H(p(\theta \mid \{x, y\})).$$

decreases *monotonically*. Observe that our objective in (8) corresponds to the limit $\sigma \to 0$ of this objective along with a uniform non-informative prior $m(\theta)$ in (5). In fact, this result is analogous to the classical result that an ergodic Markov chain makes monotonic improvements in the KL-divergence as it converges to the steady-state, also known as, equilibrium, distribution (Levin & Peres, 2017). The posterior distribution of the model parameters induced by the stochastic updates in (12) is the Gibbs distribution $p^*(\theta \mid \{(x, y)\}) \propto \exp\left(-2(R + \lambda D + \gamma C)/\sigma\right)$.

It is for the above reason that we call the surface in Fig. 1b parameterized by

$$\Theta_{\lambda,\gamma} = \left\{ \theta \in \Theta : - \left\langle \log Z_{\theta,x} \right\rangle_{p(x)} = F(\lambda, \gamma) \right\} \tag{13}$$

as the "equilibrium surface". Learning, in this case minimizing (8), is initialized outside this surface and converges to specific parts of the equilibrium surface depending upon (λ, γ); this is denoted by the red and blue curves in Fig. 1b.

The constraint that ties results in this equilibrium surface is that variational inequalities such as (2) (more are given in Alemi & Fischer (2018)) are tight up to the capacity of the model. This is analogous to the concept of equilibrium in thermodynamics (Sethna, 2006)

3. Dynamical processes on the equilibrium surface

This section prescribes dynamical processes that explore the equilibrium surface. For any parameters $\theta \in \Theta$, not necessarily on the equilibrium surface, let us define

$$J(\theta, \lambda, \gamma) = -\left\langle \log Z_{\theta,x} \right\rangle_{p(x)}. \tag{14}$$

If $\theta \in \Theta_{\lambda,\gamma}$ we have $J(\theta, \lambda, \gamma) = F(\lambda, \gamma)$ which implies

$$\nabla_\theta J(\theta, \lambda, \gamma) = 0 \text{ for all } \theta \in \Theta_{\lambda,\gamma}. \tag{15}$$

Quasi-static process. A quasi-static process in thermodynamics happens slowly enough for a system to remain in equilibrium with its surroundings. In our case, we are interested in evolving Lagrange multipliers (λ, γ) slowly and simultaneously keep the model parameters θ on the equilibrium surface; the constraint (15) thus holds at each time instant. The equilibrium surface is parameterized by R, D and C so changing (λ, γ) adapts the three functionals to track their optimal values corresponding to $F(\lambda, \gamma)$.

Let us choose some values $(\dot\lambda, \dot\gamma)$ and the trivial dynamics $\frac{\mathrm{d}}{\mathrm{d}t}\lambda = \dot\lambda$ and $\frac{\mathrm{d}}{\mathrm{d}t}\gamma = \dot\gamma$. The quasi-static constraint leads to the following partial differential equation (PDE)

$$0 \equiv \frac{\mathrm{d}}{\mathrm{d}t}\nabla_\theta J(\theta, \lambda, \gamma) = \nabla^2_\theta J\, \dot\theta + \dot\lambda \frac{\partial}{\partial\lambda}\nabla_\theta J + \dot\gamma \frac{\partial}{\partial\gamma}\nabla_\theta J \tag{16}$$

valid all $\theta \in \Theta_{\lambda,\gamma}$. At each location $\theta \in \Theta_{\lambda,\gamma}$ the above PDE indicates how the parameters should evolve upon changing the Lagrange multipliers (λ, γ). We can rewrite the PDE using the Hamiltonian H in (11) as shown next.

Lemma 5 (Equilibrium dynamics for parameters θ). *Given $(\dot\lambda, \dot\gamma)$, the parameters $\theta \in \Theta_{\lambda,\gamma}$ evolve as*

$$\begin{aligned}\dot\theta &= A^{-1} b_\lambda\, \dot\lambda + A^{-1} b_\gamma\, \dot\gamma \\ &= \theta_\lambda \dot\lambda + \theta_\gamma \dot\gamma\end{aligned} \tag{17}$$

where H is the Hamiltonian in (11) and

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

All the inner expectations $\langle\cdot\rangle$ above are taken with respect to the Gibbs measure of the Hamiltonian, i.e., $\langle\varphi\rangle = \frac{\int \varphi \exp(-H(z))\,\mathrm{d}z}{\int \exp(-H(z))\,\mathrm{d}z}$. The dynamics for the parameters θ is therefore a function of the two directional derivatives

$$\theta_\lambda = A^{-1}\, b_\lambda, \quad \textit{and} \quad \theta_\gamma = A^{-1}\, b_\gamma \tag{18}$$

with respect to λ and γ. Note that A in (17) is the Hessian of a strictly convex functional.

This lemma allows us to implement dynamical processes for the model parameters θ on the equilibrium surface. As expected, this is an ordinary differential equation (17) that depends on our chosen evolution for $(\dot\lambda, \dot\gamma)$ through the directional derivatives $\theta_\lambda, \theta_\gamma$. The utility of the above lemma therefore lies in the expressions for these directional derivatives. Appendix C gives the proof of the above lemma.

Remark 6 (Implementing the equilibrium dynamics). The equations in Lemma 5 may seem complicated to compute but observe that they can be readily estimated using samples from the dataset $x \sim p(x)$ and those from the encoder $z \sim e_\theta(z|x)$. The key difference between (17) and, say, the ELBO objective is that the gradient in the former depends upon the Hessian of the Hamiltonian H. These equations can be implemented using Hessian-vector products (Pearlmutter, 1994). If the dynamics involves certain constrains among the functionals, as Remark 7 shows, we simplify the implementation of such equations.

3.1. Iso-classification process

An iso-thermal process in thermodynamics is a quasi-static process where a system exchanges energy with its surroundings and remains in thermal equilibrium with the surroundings. We now analogously define an iso-classification process that adapts parameters of the model θ while the free-energy is subject to slow changes in (λ, γ). This adaptation is such that the classification loss is kept constant while the rate and distortion change automatically.

Following the development in Lemma 5, it is easy to create an iso-classification process. We simply add a constraint of

the form

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\nabla_\theta J = 0 \qquad & \text{(Quasi-Static Condition)} \\ \frac{\mathrm{d}}{\mathrm{d}t} C = 0 \qquad & \text{(Iso-classification Condition).} \end{aligned} \tag{19}$$

Using a very similar computation (given in Appendix D) as that in the proof of Lemma 5, this leads to the constrained dynamics

$$\begin{aligned} 0 &= C_\lambda \dot{\lambda} + C_\gamma \dot{\gamma} \\ \dot{\theta} &= \theta_\lambda \dot{\lambda} + \theta_\gamma \dot{\gamma}. \end{aligned} \tag{20}$$

The quantities C_λ and C_γ are given by

$$\begin{aligned} C_\lambda &= -\mathop{\mathbb{E}}_{x\sim p(x)}\left[\left\langle \frac{\partial H}{\partial \lambda}\right\rangle \langle \ell \rangle - \left\langle \frac{\partial H}{\partial \lambda}\,\ell\right\rangle + \langle \theta_\lambda^\top \nabla_\theta H\rangle \langle \ell\rangle - \langle \ell \theta_\lambda^\top \nabla_\theta H\rangle + \langle \theta_\lambda^\top \nabla_\theta \ell\rangle\right] \\ C_\gamma &= -\mathop{\mathbb{E}}_{x\sim p(x)}\left[\left\langle \frac{\partial H}{\partial \gamma}\right\rangle \langle \ell \rangle - \left\langle \frac{\partial H}{\partial \gamma}\,\ell\right\rangle + \langle \theta_\gamma^\top \nabla_\theta H\rangle \langle \ell\rangle - \langle \ell \theta_\gamma^T \nabla_\theta H\rangle + \langle \theta_\gamma^\top \nabla_\theta \ell\rangle\right] \end{aligned} \tag{21}$$

where $\ell = \log c_\theta(y_x|z)$ is the logarithm of the classification loss. Observe that we are not free to pick any values for $(\dot{\lambda}, \dot{\gamma})$ for the iso-classification process anymore, the constraint $\frac{\mathrm{d}C}{\mathrm{d}t} = 0$ ties the two rates together.

Remark 7 (Implementing an iso-classification process). The first constraint in (20) allows us to choose

$$\begin{aligned} \dot{\lambda} &= -\alpha \frac{\partial C}{\partial \gamma} = -\alpha \frac{\partial^2 F}{\partial \gamma^2} \\ \dot{\gamma} &= \alpha \frac{\partial C}{\partial \lambda} = \alpha \frac{\partial^2 F}{\partial \lambda \partial \gamma} \end{aligned} \tag{22}$$

where α is a parameter to scale time. The second equalities in both rows follow because $F(\lambda, \gamma)$ is the optimal free-energy which implies relations like $D = \frac{\partial F}{\partial \lambda}$ and $C = \frac{\partial F}{\partial \gamma}$. We can now compute the two deriatives in (22) using finite differences to implement an iso-classification process. This is equivalent to running the dynamics in (20) using finite-difference approximation for the terms $\frac{\partial H}{\partial \lambda}$, $\frac{\partial H}{\partial \gamma}$, $\frac{\partial \nabla_\theta H}{\partial \lambda}$, $\frac{\partial \nabla_\theta H}{\partial \gamma}$. While approximating all these listed quantities at each update of θ would be cumbersome, exploiting the relations in (20) is efficient even for large neural networks, as our experiments show.

Remark 8 (Other dynamical processes of interest). In this paper, we focus on iso-classification processes. However, following the same program as that of this section, we can also define other processes of interest, e.g., one that keeps $C + \beta^{-1} R$ constant while fine-tuning a model. This is similar to the alternative Information Bottlenck of Achille & Soatto (2017) wherein the rate is defined using the weights of a network as the random variable instead of the latent factors z. This is also easily seen to be the right-hand side of the PAC-Bayes generalization bound (McAllester, 2013). A dynamical process that preserves this functional would be able to control the generalization error which is an interesting prospect for future work.

4. Transferring representations to new tasks

Section 3 demonstrated dynamical processes where the Lagrange multipliers λ, γ change with time and the process adapts the model parameters θ to remain on the equilibrium surface. This section demonstrates the same concept under a different kind of perturbation, namely the one where the underlying task changes. The prototypical example one should keep in mind in this section is that of transfer learning where a classifier trained on a dataset $p^s(x, y)$ is further trained on a new dataset, say $p^t(x, y)$. We will assume that the input domain of the two distributions is the same.

4.1. Changing the data distribution

If i.i.d samples from the source task are denoted by $X^s = \{x_1^s, \ldots, x_{n_s}^s\}$ and those of the target distribution are $X^t = \{x_1^t, \ldots, x_{n_t}^t\}$ 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'. We will consider a transport problem that transports the source distribution $p^s(x)$ to the target distribution $p^t(x)$. For any $t \in [0, 1]$ we interpolate between the two distributions using a mixture

$$p(x, t) = (1 - t)p^s(x) + t p^t(x). \tag{23}$$

Observe that the interpolated data distribution equals the source and target distribution at $t = 0$ and $t = 1$ respectively and it is the mixture of the two distributions for other times. We keep the labels of the data the same and do not interpolate them. As discussed in Appendix F we can also use techniques from optimal transportation (Villani, 2008) to obtain a better transport; the same dynamical equations given below remain valid in that case.

4.2. Iso-classification process with a changing data distribution

The equilibrium surface $\Theta_{\lambda,\gamma}$ in Fig. 1b is a function of the task and also evolves with the task. We now give a dynamical process that keeps the model parameters in equilibrium as the task evolves quasi-statically. We again have the same conditions for the dynamics as those in (19). The following lemma is analogous to Lemma 5.

Lemma 9 (Dynamical process for changing data distribution). *Given $(\dot{\lambda}, \dot{\gamma})$, the evolution of model parameters θ for a changing data distribution given by (23) is*

$$\dot{\theta} = \theta_\lambda \dot{\lambda} + \theta_\gamma \dot{\gamma} + \theta_t \tag{24}$$

where

$$\theta_t = A^{-1}\, b_t =: -A^{-1} \int \frac{\partial p(x,t)}{\partial t}\, \langle \nabla_\theta H \rangle \, \mathrm{d}x \tag{25}$$

and the other quantities are as defined in Lemma 5 with the only change that expectations on data x are taken with respect to $p(x,t)$ instead of $p(x)$. The additional term θ_t arises because the data distribution changes with time.

A similar computation as that of Section 3.1 gives a quasi-static iso-classification process as the task evolves

$$\begin{aligned}\dot\theta &= \theta_\lambda \dot\lambda + \theta_\gamma \dot\gamma + \theta_t \\ 0 &= C_\lambda \dot\lambda + C_\gamma \dot\gamma + C_t\end{aligned} \tag{26}$$

where C_λ and C_γ are as given in (21) with the only change being that the outer expectation is taken with respect to $x \sim p(x,t)$. The new term that depends on time t is

$$C_t = -\int \frac{\partial p(x,t)}{\partial t} \langle \ell \rangle \, \mathrm{d}x - \mathop{\mathbb{E}}_{x\sim p(x,t)} \left[\langle \theta_t^\top \nabla_\theta H \rangle \langle \ell \rangle - \langle \theta_t^\top \nabla_\theta H \, \ell \rangle + \langle \theta_t^\top \nabla_\theta \ell \rangle \right] \tag{27}$$

with $\ell = \log c_\theta(y_{x_t}|z)$. Finally get

$$\begin{aligned}\dot\theta &= \left(\theta_\lambda - \frac{C_\lambda}{C_\gamma}\,\theta_\gamma\right)\dot\lambda + \left(\theta_t - \frac{C_t}{C_\gamma}\theta_\gamma\right). \\ &=: \hat\theta_\lambda \dot\lambda + \hat\theta_t\end{aligned} \tag{28}$$

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

4.3. Geodesic transfer of representations

The dynamics of Lemma 9 is valid for any $(\dot\lambda, \dot\gamma)$. We provide a locally optimal way to change (λ, γ) in this section.

Remark 10 (Rate-distortion tradeoff). Note that

$$\begin{aligned}&\dot C = 0, \\ &\dot D = \frac{\partial D}{\partial \lambda}\dot\lambda + \frac{\partial D}{\partial \gamma}\dot\gamma = -\alpha \left(\frac{\partial^2 F}{\partial \lambda^2}\frac{\partial^2 F}{\partial \gamma^2} - \left(\frac{\partial^2 F}{\partial \lambda \partial \gamma}\right)^2 \right) \\ &\quad = -\alpha \det\left(\mathrm{Hess}(F)\right), \\ &\dot R = \frac{\partial R}{\partial D}\dot D + \frac{\partial R}{\partial C}\dot C = -\lambda \dot D.\end{aligned} \tag{29}$$

The first equality is simply our iso-classification constraint. For $\alpha > 0$, the second one indicates that $\dot D < 0$ using Lemma 3 which shows that $0 \succ \mathrm{Hess}(F)$. This also gives $\dot\lambda > 0$ in (22). The third equality is a powerful observation: it indicates a trade-off between rate and distortion, if $\dot D < 0$ we have $\dot R > 0$. It also shows the geometric structure of the equilibrium surface by connecting $\dot R$ and $\dot D$ together, which we will exploit next.

Computing the functionals R, D and C during the iso-classification transfer presents us with a curve in RDC space. Geodesic transfer implies that the functionals R, D follow the shortest path in this space. But notice that if **we assume that the model capacity is infinite**, the RDC space is Euclidean and therefore the geodesic is simply a straight line. Since we keep the classification loss constant during the transfer, $\dot C = 0$, straight line implies that slope $\mathrm{d}D/\mathrm{d}R$ is a constant, say k. Thus $\dot D = k\dot R$. Observe that $\dot R = \frac{\partial R}{\partial D}\dot D + \frac{\partial R}{\partial C}\dot C + \frac{\partial R}{\partial t} = -\lambda \dot D + \frac{\partial R}{\partial t}$. Combining the iso-classification constraint and the fact that $\dot D = k\dot R = -k\lambda \dot D + k\frac{\partial R}{\partial t}$, gives us a linear system:

$$\begin{aligned}&\frac{\partial D}{\partial t} + \frac{\partial D}{\partial \lambda}\dot\lambda + \frac{\partial D}{\partial \gamma}\dot\gamma = \frac{k\frac{\partial R}{\partial t}}{1+k\lambda}; \\ &\frac{\partial C}{\partial \lambda}\dot\lambda + \frac{\partial C}{\partial \gamma}\dot\gamma + \frac{\partial C}{\partial t} = 0\end{aligned} \tag{30}$$

We solve this system to update (λ, γ) during the transfer.

5. Experimental validation

This section presents experimental validation for the ideas in this paper. We first implement the dynamics in Section 3 that traverses the equilibrium surface and then demonstrate the dynamical process for transfer learning devised in Section 4.

Setup. We use the MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky, 2009) datasets for our experiments. We use a 2-layer fully-connected network (same as that of Kingma & Welling (2013)) as the encoder and decoder for MNIST; the encoder for CIFAR-10 is a ResNet-18 (He et al., 2016) architecture while the decoder is a 4-layer deconvolutional network (Noh et al., 2015). Full details of the pre-processing, network architecture and training are provided in Appendix A.

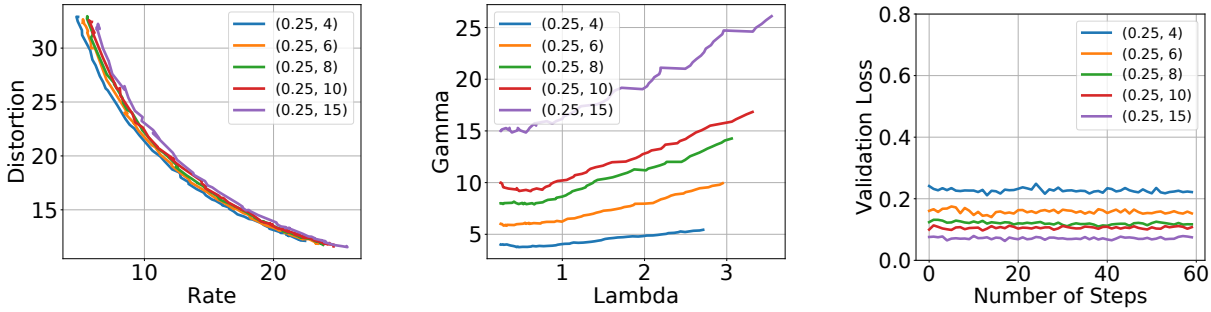
5.1. Iso-classification process on the equilibrium surface

This experiment demonstrates the iso-classification process in Remark 7. As discussed in Remark 4, training a model to minimize the funtional $R + \lambda D + \gamma C$ decreases the free-energy monotonically.

Details. Given a value of the Lagrange multipliers (λ, γ) we first find a model on the equilibrium surface by training from scratch for 120 epochs with the Adam optimizer (Kingma & Ba, 2014); the learning rate is set to 10^{-3} and drops by a factor of 10 every 50 epochs. We then run the iso-classification process for these models in Remark 7 as follows. We modify (λ, γ) according to the equations

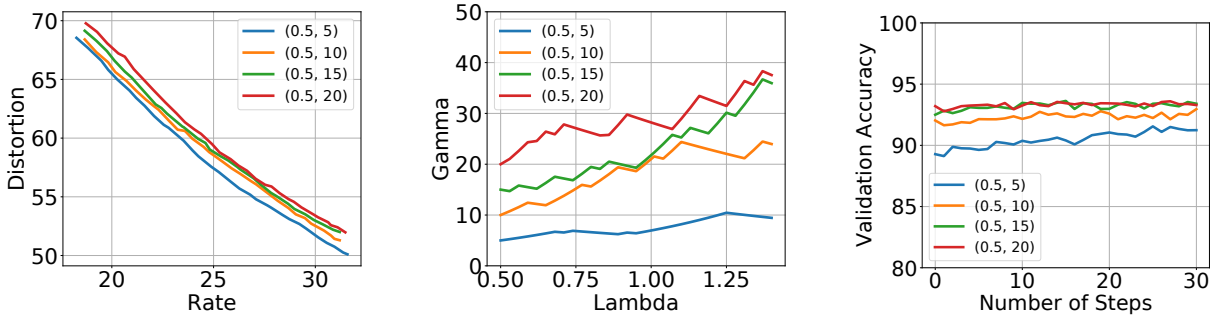
$$\dot\lambda = -\alpha \frac{\partial C}{\partial \gamma} \quad \text{and} \quad \dot\gamma = \alpha \frac{\partial C}{\partial \lambda}. \tag{31}$$

Changes in (λ, γ) cause the equilibrium surface to change, so it is necessary to adapt the model parameters θ so as to keep them on the dynamically changing surface; let us call this proces of adaptation "equilibriation". We achieve this by taking gradient-based updates to minimize $J(\lambda, \gamma)$ with a



(a) (b) (c)

Figure 2. **Iso-classification process for MNIST.** We run 5 different experiments for initial Lagrange multipliers given by $\lambda = 0.25$ and $\gamma \in \{4, 6, 8, 10, 15\}$. During each experiment, we modify these Lagrange multipliers (Fig. 2b) to keep the classification loss constant and plot the rate-distortion curve (Fig. 2a) along with the validation loss (Fig. 2c). The validation accuracy is constant for each experiment; it is between 92–98% for these initial values of (λ, γ). Similarly the training loss is almost unchanged during each experiment and takes values between 0.06–0.2 for different values of (λ, γ).



(a) (b) (c)

Figure 3. **Iso-classification process for CIFAR-10.** We run 4 different experiments for initial Lagrange multipliers $\lambda = 0.5$ and $\gamma \in \{5, 10, 15, 20\}$. During each experiment, we modify the Lagrange multipliers (Fig. 3b) to keep the classification loss constant and plot the rate-distortion curve (Fig. 3a) along with the validation accuracy (Fig. 3c). The validation loss is constant during each experiment; it takes values between 0.5–0.8 for these initial values of (λ, γ). Similarly, the training loss is constant and takes values between 0.02–0.09 for these initial values of (λ, γ). Observe that the rate-distortion curve in Fig. 3a is much flatter than the one in Fig. 2a which indicates that the model family Θ for CIFAR-10 is much more powerful; this corresponds to the straight line in the RD curve for an infinite model capacity is as shown in Fig. 1a.

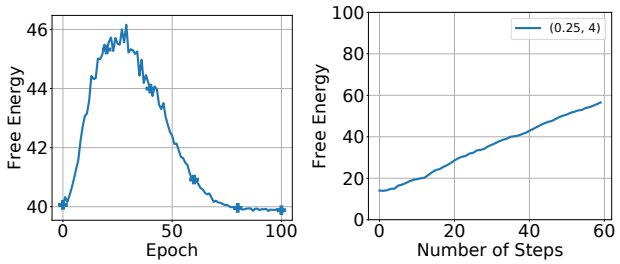
learning rate schedule that looks like a sharp quick increase from zero and then a slow annealing back to zero. The learning rate schedule is given by $\eta(t) = (t/T)^2 (1 - t/T)^5$ where t is the number of mini-batch updates taken since the last change in (λ, γ) and T is total number of mini-batch updates of equilibriation. The maximum value of the learning rate is set to 1.5×10^{-3}. The free-energy should be unchanged if the model parameters are on the equilibrium surface after equilibriation; this is shown in Fig. 4a. Partial derivatives in (31) are computed using finite-differences.

Fig. 2 shows the result for the iso-classification process for MNIST and Fig. 3 shows a similar result for CIFAR-10. We can see that the classification loss remains constant through the process. This experiment shows that we can implement an iso-classification process while keeping the model parameters on the equilibrium surface during it.

5.2. Transfer learning between two subsets of MNIST

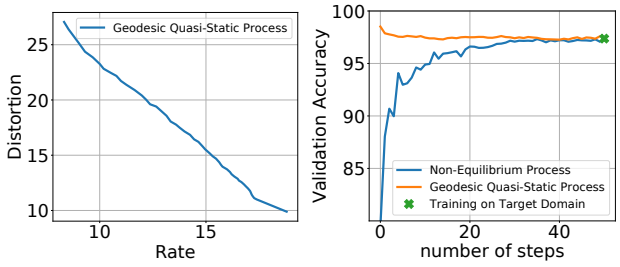
We next present experimental results of an iso-classification process for transferring the learnt representation. We pick the source dataset to be all images corresponding to digits 0–4 in MNIST and the target datast is its complement, images of digits 5–9. Our goal is to adapt a model trained on the source task to the target task while keeping its classification loss constant. We run the geodesic transfer dynamics from Section 4.3 and the results are shown in Fig. 5.

It is evident that the classification accuracy is constant throughout the transfer and is also the same as that of training from scratch on the target. MNIST is an simple dataset and the accuracy gap between iso-classification transfer, fine-tuning from the source and training from scratch is minor. The benefit of running the iso-classification transfer however is that we can be guaranteed about the final accuracy of the model. The gap between these three to



(a) (b)

Figure 4. **Variation of the free-energy $F(\lambda, \gamma)$ across the equilibriation and the iso-classification processes.** Fig. 4a shows the free-energy during equilibriation between small changes of (λ, γ). The initial and final values of the Lagrange multipliers are $(0.5, 1)$ and $(0.51, 1.04)$ respectively and the free-energy is about the same for these values. Fig. 4b shows the free-energy as (λ, γ) undergo a large change from their initial value of $(0.25, 4)$ to $(3.5, 26)$ during the iso-classification process in Fig. 2. Since the rate-distortion change a lot (Fig. 2a), the free-energy also changes a lot even if C is constant (Fig. 2c). Number of steps in Fig. 4b refers to the number of steps of running (31).



(a) (b)

Figure 5. **Transfering from source dataset of MNIST digits 0–4 to the target dataset consisting of digits 5–9.** Fig. 5a shows the variation of rate and distortion during the transfer; as discussed in Section 4.3 we maintain a constant $\mathrm{d}R/\mathrm{d}D$ during the transfer; the rate decreases and the distortion increases. Fig. 5b shows the validation accuracy during the transfer. The orange curve corresponds to geodesic iso-classification transfer; the blue curve is the result of directly fine-tuning the source model on the target data (note the very low accuracy at the start); the green point is the accuracy of training on the target task from scratch.

be significant for more complex datasets in the following section.

5.3. Transfer learning between two subsets of CIFAR10

The iso-classification process is a quasi-static process, i.e., the model parameters θ are lie on the equilibrium surface at all times $t \in [0, 1]$ during the transfer. Note that both the equilibrium surface and the free-energy $F(\lambda, \gamma)$ are functions of the data and change with time. Let us write this explicitly as

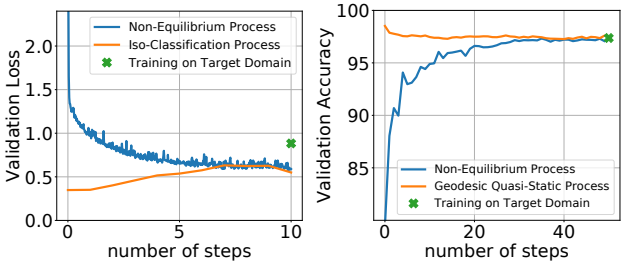
$$F(t) := R(t, \lambda(t), \gamma(t)) + \lambda D(t, \lambda(t), \gamma(t)) + \gamma C_0$$

where C_0 is the classification loss. We prescribed a geodesic transfer above where the Lagrange multipliers λ, γ were adapted simultaneously to confirm to the constraints of the equilibrium surface locally. We can also adapt them using the following heuristic. We let $\dot{\lambda} = k$ for some constant k and use

$$\frac{\partial C}{\partial \lambda}\dot{\lambda} + \frac{\partial C}{\partial \gamma}\dot{\gamma} + \frac{\partial C}{\partial t} = 0, \tag{32}$$

to get the evolution curve of $\gamma(t)$.

Here we present experimental results of an iso-classification process for transferring the learnt representation. We pick the source dataset to be all vehicles (airplane, automobile, ship and truck) in CIFAR-10 and the target dataset consists of four animals (bird, cat, deer and dog). We set the output size of classifier to be four. Our goal is to adapt a model trained on the source task to the target task while keeping its classification loss constant. We run the iso-c transfer dynamics (32) and the results are shown in Fig. 6.



(a) (b)

Figure 6. **Transferring from source dataset of CIFAR-10 vehicles to the target dataset consisting of four animals.** Fig. 6a shows the variation of validation loss during the transfer. Fig. 6b shows the validation accuracy during the transfer. The orange curve corresponds to iso-classification transfer; the blue curve is the result of directly fine-tuning the source model on the target data (note the very low accuracy at the start); the green point is the accuracy of training on the target task from scratch.

It is evident that both the classification accuracy and loss are constant throughout the transfer. CIFAR-10 is a more complex dataset as comparing with MNIST and the accuracy gap between iso-classification transfer, fine-tuning from the source and training from scratch is significant. Observe that the classification loss gap between iso-classification transfer and training from scratch on the target is also significant. The benefit of running the iso-classification transfer is that we can be guaranteed about the final accuracy and validation loss of the model.

Details of the experimental setup for CIFAR-10 transferring. At moment t, parameters λ, γ determine our objective functions. We compute iso-classification loss transfer process by first setting initial states: $(\lambda = 4, \gamma = 100)$. We train on source dataset for 300 epochs with Adam and

a learning rate of 1E-3 that drops by a factor of 10 after every 120 epochs to obtain the initial state. We change λ, γ with respect to time t and then apply the equilibration learning rate schedule of Fig. 4a to achieve the transition between equilibrium states. We compute the partial derivatives $\frac{\partial C}{\partial t}$, $\frac{\partial C}{\partial \lambda}$ and $\frac{\partial C}{\partial \gamma}$ by using finite difference. At each time t, solving (32) with the partial derivatives leads to the solution for $\dot{\gamma}$, where $\dot{\lambda}$ is a constant. In our experiment we set $\dot{\lambda} = -1.5$.

6. Related work

We are motivated by the Information Bottleneck (IB) principle of Shwartz-Ziv & Tishby (2017); Tishby et al. (2000), which has been further explored by Achille & Soatto (2017); Alemi et al. (2016); Higgins et al. (2017). The key difference in our work is that while these papers seek to understand the representation for a given task, we focus on how the representation can be adapted to a new task. Further, the Lagrangian in (8) has connections to PAC-Bayes bounds (Dziugaite & Roy, 2017; McAllester, 2013) and training algorithms that use the free-energy (Chaudhari et al., 2019). Our use of rate-distortion for transfer learning is close to the work on unsupervised learning of Brekelmans et al. (2019); Ver Steeg & Galstyan (2015).

This paper builds upon the work of Alemi & Fischer (2018); Alemi et al. (2017). We refine some results therein, viz., we provide a proof of the convexity of the equilibrium surface and identify it with the equilibrium distribution of SGD (Remark 4). We introduce new ideas such as dynamical processes on the equilibrium surface. Our use of thermodynamics is purely as an inspiration; the work presented here is mathematically rigorous and also provides an immediate algorithmic realization of the ideas.

This paper has strong connections to works that study stochastic processes inspired from statistical physics for machine learning, e.g., approximate Bayesian inference and implicit regularization of SGD (Chaudhari & Soatto, 2017; Mandt et al., 2017), variational inference (Jordan et al., 1998; Kingma & Welling, 2013). The iso-classification process instantiates an "automatic" regularization via the trade-off between rate and distortion; this point-of-view is an exciting prospect for future work. The technical content of the paper also draws from optimal transportation (Villani, 2008).

A large number of applications begin with pre-trained models (Girshick et al., 2014; Sharif Razavian et al., 2014) or models trained on tasks different (Doersch & Zisserman, 2017). Current methods in transfer learning however do not come with guarantees over the performance on the target dataset, although there is a rich body of older work (Baxter, 2000) and ongoing work that studies this (Zamir et al., 2018). The information-theoretic understanding of transfer and the constrained dynamical processes developed in our paper is a first step towards building such guarantees. In this context, our theory can also be used to tackle catastrophic forgetting Kirkpatrick et al. (2017) to "detune" the model post-training and build up redundant features.

7. Discussion

We presented dynamical processes that maintain the parameters of model on an equilibrium surface that arises out of a certain free-energy functional for the encoder-decoder-classifier architecture. The decoder acts as a measure of the information discarded by the encoder-classifier pair while fitting on a given task. We showed how one can develop an iso-classification processs that travels on the equilibrium surface while keeping the classification loss constant. We showed an iso-classification transfer learning process which keeps the classification loss constant while adapting the learnt representation from a source task to a target task.

The information-theoretic point-of-view in this paper is rather abstract but its benefit lies in its exploitation of the equilibrium surface. Relationships between the three functionals, namely rate, distortion and classification, that define this surface, as also other functionals that connect to the capacity of the hypothesis class such as the entropy S may allow us to define invariants of the learning process. For complex models such as deep neural networks, such a program may lead an

References

Achille, A. and Soatto, S. On the emergence of invariance and disentangling in deep representations. *arXiv:1706.01350*, 2017.

Alemi, A. A. and Fischer, I. Therml: Thermodynamics of machine learning. *arXiv preprint arXiv:1807.04162*, 2018.

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. *arXiv:1612.00410*, 2016.

Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., and Murphy, K. Fixing a broken elbo. *arXiv preprint arXiv:1711.00464*, 2017.

Baxter, J. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

Brekelmans, R., Moyer, D., Galstyan, A., and Ver Steeg, G. Exact rate-distortion in autoencoders via echo noise. In *Advances in Neural Information Processing Systems*, pp. 3884–3895, 2019.

Chaudhari, P. and Soatto, S. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. *arXiv preprint arXiv:1710.11029*, 2017.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

Doersch, C. and Zisserman, A. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2051–2060, 2017.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. *arXiv:1603.05027*, 2016.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and A, L. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework . In *ICLR*, 2017.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. In *Learning in graphical models*, pp. 105–161. Springer, 1998.

Kingma, D. and Ba, J. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. *arXiv:1312.6114*, 2013.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Krizhevsky, A. Learning multiple layers of features from tiny images. Master's thesis, Computer Science, University of Toronto, 2009.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Levin, D. A. and Peres, Y. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Mandt, S., Hoffman, M. D., and Blei, D. M. Stochastic Gradient Descent as Approximate Bayesian Inference. *arXiv:1704.04289*, 2017.

McAllester, D. A pac-bayesian tutorial with a dropout bound. *arXiv:1307.2118*, 2013.

Mezard, M. and Montanari, A. *Information, physics, and computation*. Oxford University Press, 2009.

Noh, H., Hong, S., and Han, B. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pp. 1520–1528, 2015.

Pearlmutter, B. A. Fast exact multiplication by the hessian. *Neural computation*, 6(1):147–160, 1994.

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

Sethna, J. *Statistical mechanics: entropy, order parameters, and complexity*, volume 14. Oxford University Press, 2006.

Sharif Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813, 2014.

Shwartz-Ziv, R. and Tishby, N. Opening the black box of deep neural networks via information. *arXiv:1703.00810*, 2017.

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Ver Steeg, G. and Galstyan, A. Maximally informative hierarchical representations of high-dimensional data. In *Artificial Intelligence and Statistics*, pp. 1004–1012, 2015.

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., and Savarese, S. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3712–3722, 2018.