TherML: Thermodynamics of Machine Learning

Alexander A. Alemi [1] **Ian Fischer** [1]

Abstract

In this work we offer a framework for reasoning about a wide class of existing objectives in machine learning. We develop a formal correspondence between this work and thermodynamics and discuss its implications.

The traditional approach to learning involves modelling. A practitioner describes a graphical model they believe generated the data, and then attempts to maximize the likelihood that the observed data was drawn from this model.

> Essentially all models are wrong, but some are useful. — George Box (Box & Draper, 1987)

Ideally we would be able to quantify, control and/or steer the *wrongness* of our model, but to do so requires a formal separation between the real world of our data and the imaginary world of our model. To connect the two we augment the real world with something under our direct control. To that end, we formulate *representation learning*.

1. A Tale of Two Worlds

Let X, Y be some paired data: where X is raw high dimensional data and Y is some relevant low dimensional description or target. For example, a set of images X and their labels Y. We imagine this data comes from some *true* unknown distribution. The data is conditionally independent given some *true* governing parameters ϕ. That is we imagine the joint distribution:

$$p(\{x, y\}, \phi) = p(\phi) \prod_i p(x_i|\phi) p(y_i|x_i, \phi). \qquad (1)$$

We factor the conditional joint distribution $p(x_i, y_i|\phi)$ as $p(y_i|x_i, \phi)p(x_i|\phi)$. Here ϕ stands in for the rest of the universe, insomuch as it is relevant to our data generating procedure. Generally, this distribution is intractable, but we imagine we have or can generate a finite dataset of fair samples from it.

If the data was all we had, we couldn't do much in the way of learning or inference, especially considering we know nothing about the distribution describing our data. Since we want to shy away from directly modelling the data itself, our only choice is to augment the universe with some new random variables under our direct control. We are primarily interested in learning a stochastic *representation* of X, call it Z, defined by some parametric distribution of our own design: $p(z_i|x_i, \theta)$ with its own parameters θ. A *training procedure* is a process that assigns a distribution $p(\theta|\{x, y\})$ (where $\{x, y\}$ is shorthand for the entire set of x_i, y_i) to the parameters conditioned on the observed dataset. With our augmentations, the world now looks like the graphical model in Figure 1a, described by the joint density:

$$p(x, y, \phi, z, \theta) = p(\phi)p(\theta|\{x, y\}) \times \prod_i p(x_i|\phi)p(y_i|x_i, \phi)p(z_i|x_i, \theta) \qquad (2)$$

World P is what we have. It is not necessarily what we want. What we *have* to contend with is an unknown distribution of our data. What we *want* is a world in which Z causally factors X and Y, acting as a latent variable for the pair, leaving no correlations unexplained. Similarly, we would prefer if we could easily marginalize out the dependence on our universal (Φ) and model specific (Θ) parameters. World Q in Figure 1b is the world we *want* [1]. It satisfies the joint density:

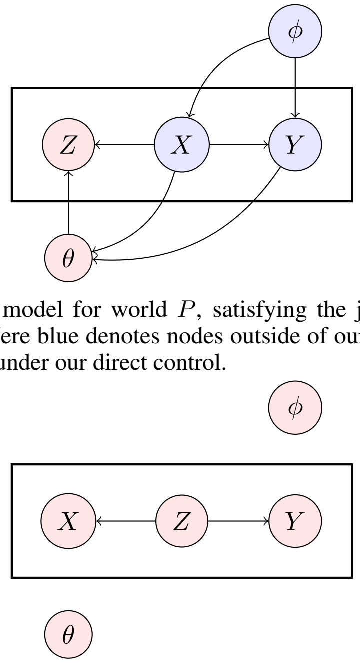
$$q(x, y, \phi, z, \theta) = q(\phi)q(\theta) \prod_i q(z_i)q(x_i|z_i)q(y_i|z_i) \qquad (3)$$

Having expressed both the situation we are in and the one we desire, we can formally express how similar the two worlds are by computing the *relative information* or *relative entropy*

[1] Google. Correspondence to: Alexander A. Alemi <alemi@google.com>.

Presented at the ICML 2018 workshop on Theoretical Foundations and Applications of Deep Generative Models. Copyright 2018 by the author(s).

[1] We could consider different alternatives, deciding to relax some of the constraints we imposed in world Q, or generalizing World P by letting the representation depend on X and Y jointly, for instance. What follows demonstrates a general sort of *calculus* that we can invoke for any specified pair of graphical models. In particular Appendix A discusses a closely related alternative where we desire $q(x, y, z) = q(y|z)q(z|x)$.



(a) Graphical model for world P, satisfying the joint density in Equation 2. Here blue denotes nodes outside of our control, while red nodes are under our direct control.

(b) Graphical model for world Q, the world we desire which satisfies the joint density in Equation 3.

Figure 1. Graphical models.

or KL divergence between them. The relative information between two distributions:

$$\mathrm{KL}(p \mid\mid q) = \int d\chi \, p(\chi) \log \frac{p(\chi)}{q(\chi)} = \left\langle \log \frac{p(\chi)}{q(\chi)} \right\rangle_P \qquad (4)$$

measures how much information (in bits, or nats in the case the natural logarithm is used) you gain when you discover you are in World P when what you expected was World Q (Cover & Thomas, 2012). This is often expressed as the excess bits required to encode a signal from P if you use an entropic code optimized for Q. Here and throughout $\langle \cdot \rangle_A$ is used to denote expectations with respect to distribution A. It is worth noting that while the continuous analog of Shannon entropy $-\int dx \, p(x) \log p(x)$ is neither positive semidefinite nor reparameterization independent (Marsh, 2013), the relative entropy $\int dx \, p(x) \log \frac{p(x)}{q(x)}$ is both positive semidefinite and reparameterization independent, and so Equation 4 is perfectly well defined and behaved for continuous distributions.

Another useful quantity worth defining is the *mutual information*:

$$\begin{aligned} I(X;Y) &\equiv \left\langle \log \frac{p(x,y)}{p(x)p(y)} \right\rangle_P \\ &= \left\langle \log \frac{p(x|y)}{p(x)} \right\rangle_P = \left\langle \log \frac{p(y|x)}{p(y)} \right\rangle_P , \end{aligned} \qquad (5)$$

which measures the relative information between the joint distribution for X and Y and the product of their marginal distributions. That is, how much information you gain about X by observing Y, or vice versa. It is similarly well defined, positive semidefinite and reparameterization independent for continuous distributions (Cover & Thomas, 2012). The mutual information vanishes if and only if the two variables are independent, and diverges for continuous variables if they are related by an invertible transformation.

The relative information between World P and Q is:

$$\begin{aligned} \mathrm{KL}(p \mid\mid q) = \Bigg\langle & \log \frac{p(\phi)}{q(\phi)} + \log \frac{p(\theta|\{x,y\})}{q(\theta)} \\ & + \sum_i \log \frac{p(z_i|x_i,\theta)}{q(z_i)} + \log \frac{p(x_i|\phi)}{q(x_i|z_i)} + \log \frac{p(y_i|x_i,\phi)}{q(y_i|z_i)} \Bigg\rangle_P . \end{aligned} \qquad (6)$$

Here we have split the terms into reparameterization independent components.

In practice, worlds P and Q will often be inconsistent, so that it will not be possible to achieve a vanishing relative information (KL divergence) between the distributions. But amongst all possible distributions Q consistent with the conditional structure in the graphical model, we can ask what the minimum achievable KL is. For this we need to know the optimal Q distribution. The optimal q distributions are simply the corresponding distributions in our true world P, taking whichever marginalizations or conditioning required. For instance, the $q(x_i|z_i)$ that achieves the minimal KL is $p(x_i|z_i)$ the marginal posterior for x_i given z_i integrating out ϕ and θ. Setting all Q distributions to their optimal values, it can be shown (Friedman et al., 2001) that the minimal relative information obtained is given by the difference in multi-informations of each graph (I_P, I_Q). This is the well known *information projection* (Csiszár & Matúš, 2003). The multi-information in the graph is a sum of mutual informations between each node and its parents.

$$\begin{aligned} I_P \equiv \left\langle \log \frac{p(x,y,\phi,z,\theta)}{p(x)p(y)p(\phi)p(z)p(\theta)} \right\rangle_P &= I(\Theta;\{X,Y\}) \\ + \sum_i I(X_i;\Phi) + I(Y_i;X_i,\Phi) + I(Z_i;X_i,\Theta) & \qquad (7) \end{aligned}$$

$$I_Q = \sum_i I(X_i;Z_i) + I(Y_i;Z_i) \qquad (8)$$

In our case:

$$\begin{aligned} \mathrm{KL}(p \mid\mid Q) &\equiv \min_{q \in Q} \mathrm{KL}(p \mid\mid q) = I_P - I_Q \\ &= I(\Theta;\{X,Y\}) + \sum_i I(X_i;\Phi) + I(Y_i;X_i,\Phi) \\ &\quad + \sum_i I(Z_i;X_i,\Theta) - I(X_i;Z_i) - I(Y_i;Z_i). \end{aligned} \qquad (9)$$

This minimal relative information has two terms outside our control and we can take them to be constant: $I(X_i;\Phi)$ and $I(Y_i;X_i,\Phi)$. These terms measure the intrinsic complexity of our data. The remaining four terms are:

- $I(X_i;Z_i)$ - which measures how much information our representation contains about the input (X). This should be maximized to make world P more resemble a world that satisfies the relations in Q.
- $I(Y_i;Z_i)$ - which measures how much information our representation contains about our auxiliary data. This should be maximized as well.
- $I(Z_i;X_i,\Theta)$ - which measures how much information the parameters and input determine about our representation. This should be minimized to ensure consistency between worlds. Notice that this is similar to, but distinct from the first term above.

  $$I(Z_i;X_i,\Theta) = I(Z_i;X_i) + I(Z_i;\Theta|X_i) \tag{10}$$

  by the Chain Rule for mutual information [2].
- $I(\Theta;\{X,Y\})$ - which measures how much information we store about our training data in the parameters of our encoder. This should also be minimized.

These mutual informations are all intractable in general, since we cannot commute the necessary marginals in closed form.

1.1. Functionals

Despite their intractability, we can compute variational bounds on these mutual informations.

- $R \equiv \sum_i \left\langle \log \frac{p(z_i|x_i,\theta)}{q(z_i)} \right\rangle_P \geq \sum_i I(Z_i;X_i,\Theta)$

  The *rate* measures the complexity of our representation. It is the relative information of a sample specific representation $z_i \sim p(z|x_i,\theta)$ with respect to our variational marginal $q(z)$. It measures how many bits we actually encode about each sample.
- $C \equiv -\sum_i \langle \log q(y_i|z_i)\rangle_P \geq \sum_i H(Y_i) - I(Y_i;Z_i) = \sum_i H(Y_i|Z_i)$

  The *classification error* measures the conditional entropy of Y left after conditioning on Z. It is a measure of how much information about Y is left unspecified in our representation. This functional measures our supervised learning performance.
- $D \equiv -\sum_i \langle \log q(x_i|z_i)\rangle_P \geq \sum_i H(X_i) - I(X_i;Z_i) = \sum_i H(X_i|Z_i)$

  The *distortion* measures the conditional entropy of X left after conditioning on Z. It is a measure of how much information about X is left unspecified in our representation. This functional measures our unsupervised learning performance.
- $S \equiv \left\langle \log \frac{p(\theta|\{x,y\})}{q(\theta)} \right\rangle_P \geq I(\Theta;\{X,Y\})$

  The relative entropy in our parameters or just *entropy* for short measures the relative information between the distribution we assign our parameters in world P after learning from the data $\{X,Y\}$, with respect to some data independent $q(\theta)$ *prior* on the parameters. This is an upper bound on the mutual information between the data and our parameters and as such can measure our risk of overfitting.

1.2. Inequalities

Notice that S and R are both themselves upper bounds on mutual informations, and so must be positive semidefinite. Also, if our data is discrete, or if we have discretized it [3], D and C which are both upper bounds on conditional entropies, must be positive as well.

The distributions $p(z|x,\theta), p(\theta|\{x,y\}), q(z), q(x|z), q(y|z)$ can be chosen arbitrarily. Once chosen, the *functionals* R, C, D, S take on well described values. The choice of the five distributional families specifies a single point in a four-dimensional space. This phase space is trivially bounded by the positive semi-definiteness of the functionals mentioned above and, less trivially, by the tighter bounds set by the mutual informations they estimate in the previous section. Furthermore, since these mutual informations are all defined in world P, which itself satisfies a particular set of conditional dependencies, the mutual informations are all interdependent. This means that not even all points in the positive quadrant of the R, C, D, S space can be reached.

For example:

$$\begin{aligned} R + C + D + S &\geq \mathrm{KL}(p \,||\, Q) \\ &+ \sum_i H(X_i) + H(Y_i) - I(X_i;\Phi) - I(Y_i;X_i,\Phi) \\ &= \mathrm{KL}(p \,||\, Q) + \sum_i H(X_i,Y_i) - I(X_i;Y_i;\Phi). \end{aligned} \tag{11}$$

Showing that even if the distributional choices agree perfectly (so that $\mathrm{KL}(p \,||\, Q) = 0$), because the data and its governing distribution are outside our control and can be

[2] Given this relationship, we could actually reduce the total number of functions we consider from 4 to 3, as discussed in Appendix A.

[3] More generally, if we choose some measure $m(x), m(y)$ on both X and Y, we can define D and C in terms of that measure e.g. $D \equiv -\left\langle \log \frac{q(x|z)}{m(x)} \right\rangle_P \geq H_m(X) - I(X;Z) = H_m(X|Z)$

assumed to have non-vanishing entropies and mutual informations, combinations of these functionals will not be able to take on all values. In this case, the inherent complexity in the dataset forbids points too near the origin in the R, C, D, S plane.

Similarly,

$$\begin{aligned} R + D &\geq \sum_i H(X_i) + I(Z_i; X_i, \Theta) - I(X_i; Z_i) \\ &= \sum_i H(X_i) + I(Z_i; \Theta | X_i). \quad (12) \end{aligned}$$

This mirrors the bound given in Alemi et al. (2018) where $R + D \geq H(X)$, which is still true given that all conditional mutual informations are positive semidefinite $(H(X) + I(Z; \Theta|X) \geq H(X))$, but here we obtain a tighter bound that has a term measuring how much information about our encoding is revealed by the parameters after conditioning on the input itself. This term $(I(Z_i; \Theta|X_i)$ is precisely the difference between our two closely related mutual information governing the distortion between worlds (Equation 10) [4].

Assuming we have access to infinitely powerful variational distributions for world Q, we can map out the optimal frontier in terms of R, C, D, S. It forms a convex polytope whose edges, faces and corners are identifiable as well known objectives and their optimal solutions.

1.3. Bounds on True Learning

We can also relate R, C, D and S to mutual informations of interest that are otherwise out of reach. For instance, $I(Z; \Phi)$ which measures how well our Zs approximate the true latent variables for the data Φ. Notice that given the conditional dependencies in world P, we have the following Markov chain:

$$Z_i \leftrightarrow (X_i, Y_i, \Theta) \leftrightarrow \Phi \quad (13)$$

and so by the Data Processing Inequality (Cover & Thomas, 2012):

$$\begin{aligned} I(Z_i; \Phi) &\leq I(Z_i; \Theta, X_i, Y_i) \\ &= I(Z_i; X_i, \Theta) + \cancel{I(Z_i; Y_i|X_i, \Theta)} \leq R. \quad (14) \end{aligned}$$

The rate R forms an upper bound on the mutual information between our encoding Z_i and the *true* governing parameters of our data Φ. Similarly by the Data Processing Inequality we can establish that:

$$I(\Theta; \Phi) \leq I(\Theta; \{X, Y\}) = S. \quad (15)$$

S upper bounds the amount of information our encoder's parameters Θ can contain about the true parameters Φ.

2. Optimal Frontier

As in Alemi et al. (2018), under mild assumptions about the distributional families, it can be argued that the surface is monotonic in all of its arguments. The optimal surface in the infinite family limit can be characterized as a convex polytope. In practice we will be in the realistic setting corresponding to finite parametric families such as neural network approximators. We then expect that there is an irrevocable gap that opens up in the variational bounds. Any failure of the distributional families to model the correct corresponding marginal in P means that the space of all achievable R, C, D, S values will be some convex relaxation of the optimal surface. This surface will be described some function $f(R, C, D, S) = 0$, which means we can identify points on the surface as a function of one functional with respect to the others (e.g. $R = R(C, D, S)$). Finding points on this surface equates to solving a constrained optimization problem, e.g.

$$\begin{aligned} &\min_{q(z)q(x|z)q(y|z)p(z|x,\theta)p(\theta|\{x,y\})} R \\ &\text{such that } D = D_0, S = S_0, C = C_0. \quad (16) \end{aligned}$$

Equivalently, we could solve the unconstrained Lagrange multipliers problem:

$$\min_{q(z)q(x|z)q(y|z)p(z|x,\theta)p(\theta|\{x,y\})} R + \delta D + \gamma C + \sigma S. \quad (17)$$

Here δ, γ, σ are Lagrange multipliers that impose the constraints. They each correspond to the partial derivative of the rate at the solution with respect to their corresponding functional, keeping the others fixed.

Notice that this single objective encompasses a wide range of existing techniques.

- If we retain C alone, we are doing traditional supervised learning and our network will learn to be deterministic in its activations and parameters.
- If $\delta = 0$ we no longer require a variational reconstruction network $q(x|z)$, and are doing some form of supervised learning generally, though a more general form of VIB that similarly regularizes the parameters of our network.
- If $\delta = 0, \sigma = 0$ we exactly recover the Variational Information Bottleneck (VIB) objective of Alemi et al. (2016) (where $\beta = 1/\gamma$), a form of stochastically regularized supervised learning that imposes a bottleneck on how much information our representation can retain about the input, while simultaneously maximizing the amount of information the representation contains about the target.

[4] In Appendix A we consider taking this equality seriously to limit the space only only three functionals, S, C and $V \sim I(Z_i; \Theta|X_i)$

- If $\delta = 0$ and $\sigma, \gamma \to \infty$ but in such a way as to keep the ratio fixed $\beta \equiv \sigma/\gamma$ (that is if we drop the R term and only keep $C + \beta S$ as our objective) we recover the loss of Achille & Soatto (2017), presented as an alternative way to do Information Bottleneck (Tishby et al., 1999) but being stochastic on the parameters rather than the activations as in VIB.

- As a special case, if our objective is set to $C + S$ ($\delta = 0, \sigma, \gamma \to \infty, \sigma/\gamma \to 1$), we obtain the objective for a Bayesian neural network, ala Blundell et al. (2015).

- If we retain only D, we are training an autoencoder.

- If $\sigma = 0, \gamma = 0, \delta = 1$ the objective is equivalent to the ELBO used to train a VAE (Kingma & Welling, 2014).

- If $\sigma = 0, \gamma = 0$ more generally, the objective is equivalent to a β-VAE (Higgins et al., 2017) where $\beta = 1/\delta$.

- If $\gamma = 0$ all terms involving the auxiliary data Y drop out and we are doing some form of unsupervised learning without any variational classifier $q(y|z)$. The presence of the S term makes this more general than a usual β-VAE and should offer better generalization properties and control of overfitting by bottlenecking how much information we allow the parameters of our encoder to extract from the training data.

- $\sigma = 0, \gamma = \alpha, \delta = 1$ recovers the semi-supervised objective of Kingma et al. (2014).

Examples of all of these objectives behavior on a simple toy model is shown in Appendix B.

Notice that all of these previous approaches describe low dimensional sub-surfaces of the optimal three dimensional frontier. These approaches were all interested in different domains, some were focused on supervised prediction accuracy, others on learning a generative model. Depending on your specific problem, and downstream tasks, different points on the optimal frontier will be desirable. However, instead of choosing a single point on the frontier, we can now explore a region on the surface to see what class of solutions are possible within the modeling choices. By simply adjusting the three control parameters δ, γ, σ, we can smoothly move across the entire frontier and smoothly interpolate between all of these objectives and beyond.

2.1. Optimization

So far we've considered explicit forms of the objective in terms of the four functionals. For S this would require some kind of tractable approximation to the posterior over the parameters of our encoding distribution. Alternatively, we can formally describe the exact solution to our minimization problem:

$$\min S \text{ s.t. } R = R_0, C = C_0, D = D_0. \tag{18}$$

Recall that S measures the relative entropy of our parameter distribution with respect to the $q(\theta)$ *prior*. As such, the solution that minimizes the relative entropy subject to some constraints is a generalized Boltzmann distribution (Jaynes, 1957):

$$p^*(\theta|\{x, y\}) = \frac{q(\theta)}{\mathcal{Z}} e^{-(R+\delta D+\gamma C)/\sigma}. \tag{19}$$

Here $\mathcal{Z}$ is the *partition function*, the normalization constant for the distribution

$$\mathcal{Z} = \int d\theta \, q(\theta) \, e^{-(R+\delta D+\gamma C)/\sigma} \tag{20}$$

This suggests an alternative method for finding points on the optimal frontier. We could turn the unconstrained Lagrange optimization problem that required some explicit choice of tractable posterior distribution over parameters into a sampling problem for a richer implicit distribution.

A naive way to draw samples from this posterior would be to use Stochastic Gradient Langevin Dynamics or its cousins (Welling & Teh, 2011; Chen et al., 2014; Ma et al., 2015) which, in practice, would look like ordinary stochastic gradient descent (or its cousins like momentum) for the objective $R + \delta D + \gamma C$, with injected noise. By choosing the magnitude of the noise relative to the learning rate, the effective temperature σ can be controlled.

There is increasing evidence that the stochastic part of stochastic gradient descent itself is enough to turn SGD less into an optimization procedure and more into an approximate posterior sampler (Mandt et al., 2017; Smith & Le, 2017; Achille & Soatto, 2017; Zhang et al., 2018), where hyperparameters such as the learning rate and batch size set the effective temperature. If ordinary stochastic gradient descent is doing something more akin to sampling from a posterior and less like optimizing to some minimum, it would help explain improved performance through ensemble averages of different points along trajectories (Huang et al., 2017).

When viewed in this light, Equation 19 describes the optimal posterior for the parameters so as to ensure the minimal divergence between worlds P and Q. $q(\theta)$ plays the role of the *prior* over parameters, but our overall objective is minimized when

$$q(\theta) = p(\theta) = \langle p(\theta|\{x, y\}) \rangle_{\{x,y\}}. \tag{21}$$

That is, when our *prior* is the marginal of the posteriors over all possible datasets drawn from the true distribution.

A fair draw from this marginal is to take a sample from the posterior obtained on a different but related dataset. Insomuch as ordinary SGD training is an approximate method for drawing a posterior sample, the common practice of fine-tuning a pretrained network on a related dataset is using a sample from the optimal *prior* as our initial parameters. The fact that fine-tuning approximates use of an optimal *prior* presumably helps explain its broad success.

If we identify our true goal not as optimizing some objective but instead directly sampling from Equation 19, we can consider alternative approaches to define our learning dynamics, such as *parallel tempering* or *population annealing* (Machta & Ellis, 2011).

3. Thermodynamics

So far we have described a framework for learning that involves finding points that lie on the surface of a convex three-dimensional surface in terms of four functional coordinates R, C, D, S. Interestingly, this is all that is required to establish a formal connection to thermodynamics, which similarly is little more than the study of exact differentials (Sethna, 2006; Finn, 1993).

Whereas previous approaches connecting thermodynamics and learning (Parrondo et al., 2015; Still, 2017; Still et al., 2012) have focused on describing the thermodynamics and statistical mechanics of physical realizations of learning systems (i.e. the heat bath in these papers is a physical heat bath at finite temperature), in this work we make a formal analogy to the structure of the theory of thermodynamics, without any physical content.

3.1. First Law of Learning

The optimal frontier creates an equivalence class of states, being the set of all states that minimize as much as possible the distortion introduced in projecting world P onto a set of distributions that respect the conditions in Q. The surface satisfies some equation $f(R, C, D, S) = 0$ which we can use to describe any one of these functionals in terms of the rest, e.g. $R = R(C, D, S)$. This function is entire, and so we can equate partial derivatives of the function with differentials of the functionals[5]:

$$dR = \left(\frac{\partial R}{\partial C}\right)_{D,S} dC + \left(\frac{\partial R}{\partial D}\right)_{C,S} dD + \left(\frac{\partial R}{\partial S}\right)_{C,D} dS. \tag{22}$$

Since the function is smooth and convex, instead of identifying the surface of optimal rates in terms of the functionals C, D, S, we could just as well describe the surface in terms of the partial derivatives by applying a Legendre transformation. We will name the partial derivatives:

$$\gamma \equiv -\left(\frac{\partial R}{\partial C}\right)_{D,S} \tag{23}$$

$$\delta \equiv -\left(\frac{\partial R}{\partial D}\right)_{C,S} \tag{24}$$

$$\sigma \equiv -\left(\frac{\partial R}{\partial S}\right)_{C,D}. \tag{25}$$

These measure the exchange rate for turning rate into reduced distortion, reduced classification error, or increased entropy, respectively.

The functionals R, C, D, S are analogous to extensive thermodynamic variables such as volume, entropy, particle number, magnetic field, charge, surface area, length and energy which grow as the system grows, while the named partial derivatives γ, δ, σ are analogous to the intensive, generalized forces in thermodynamics corresponding to their paired state variable, such as pressure, temperature, chemical potential, magnetization, electromotive force, surface tension, elastic force, etc. Just as in thermodynamics, the *extensive* functionals are defined for any state, while the *intensive* partial derivatives are only well defined for *equilibrium states*, which in our language are the states lying on the optimal surface.

Recasting our total differential:

$$dR = -\gamma dC - \delta dD - \sigma dS, \tag{26}$$

we create a law analogous to the *First Law of Thermodynamics*. In thermodynamics the First Law is often taken to be a statement about the conservation of energy, and by analogy here we could think about this *law* as a statement about the conservation of information. Granted, the actual content of the law is fairly vacuous, equivalent only to the statement that there exists a scalar function $R = R(C, D, S)$ defining our surface and its partial derivatives.

3.2. Maxwell Relations

Requiring that Equation 26 be an exact differential has mathematically trivial but intuitively non-obvious implications that relate various partial derivatives of the system to one another, akin to the *Maxwell Relations* in thermodynamics. For example, requiring that mixed second partial derivatives are symmetric establishes that:

$$\left(\frac{\partial^2 R}{\partial D \partial C}\right) = \left(\frac{\partial^2 R}{\partial C \partial D}\right) \implies \left(\frac{\partial \delta}{\partial C}\right)_D = \left(\frac{\partial \gamma}{\partial D}\right)_C. \tag{27}$$

This equates the result of two very different experiments. In the experiment encoded in the partial derivative on the left, one would measure the change in the derivative of the $R - D$ curve (δ) as a function of the classification error (C)

[5] $\left(\frac{\partial X}{\partial Y}\right)_Z$ denotes the partial derivative of X with respect to Y holding Z constant.

at fixed distortion (D). On the right one would measure the change in the derivative of the $R-C$ curve (γ) as a function of the distortion (D) at fixed classification error (C). As different as these scenarios appear, they are mathematically equivalent.

We can also define other potentials analogous to the alternative thermodynamic potentials such as enthalpy, free energy, and Gibb's free energy by performing partial Legendre transformations. For instance, we can define a *free rate*:

$$F(C, D, \sigma) \equiv R - \sigma S \tag{28}$$

$$dF = -\gamma dC - \delta dD - S d\sigma. \tag{29}$$

The free rate measures the rate of our system, not as a function of S (something difficult to keep fixed), but in terms of σ, a parameter in our loss or optimal posterior.

The free rate gives rise to other Maxwell relations such as

$$\left(\frac{\partial S}{\partial C}\right)_\sigma = \left(\frac{\partial \gamma}{\partial \sigma}\right)_C, \tag{30}$$

which equates how much each additional bit of entropy (S) buys you in terms of classification error (C) at fixed effective temperature (σ), to a seemingly very different experiment where you measure the change in the effective supervised tension (γ, the slope on the $R-C$ curve) versus effective temperature (σ) at a fixed classification error (C).

We can additionally take and name higher order partial derivatives, analogous to the susceptibilities of thermodynamics like bulk modulus, the thermal expansion coefficient, or heat capacities. For instance, we can define the analog of heat capacity for our system, a sort of rate capacity at constant distortion:

$$K_D \equiv \left(\frac{\partial R}{\partial \sigma}\right)_D. \tag{31}$$

Just as in thermodynamics, these susceptibilities may offer useful ways to characterize and quantify the systematic differences between model families. Perhaps general scaling laws can be found between susceptibilities and network widths, or depths, or number of parameters or dataset size. Divergences or discontinuities in the susceptibilities are the hallmark of phase transitions in physical systems, and it is reasonable to expect to see similar phenomenon for certain models.

A great deal of first, second and third order partial derivatives in thermodynamics are given unique names. This is because the quantities are particularly useful for comparing different physical systems. We expect a subset of the first, second and higher order partial derivatives of the base functionals will prove similarly useful for comparing, quantifying, and understanding differences between modelling choices.

3.3. Zeroth Law of Learning

A central concept in thermodynamics is a notion of equilibrium. The so called Zeroth Law of thermodynamics defines thermal equilibrium as a sort of reflexive property of systems (Finn, 1993). If system A is in thermal equilibrium with system C, and system B is separately in thermal equilibrium with system C, then system A and B are in thermal equilibrium with each other.

When any subpart of a system is in thermal equilibrium with any other subpart, the system is said to be an equilibrium state.

In our framework, the points on the optimal surface are analogous to the equilibrium states, for which we have well defined partial derivatives. We can demonstrate that this notion of equilibrium agrees with a more intuitive notion of equilibrium between coupled systems. Imagine we have two different models, characterized by their own set of distributions, Model A is defined by $p_A(z|x, \theta), p_A(\theta, \{x, y\}), q_A(z)$, and model B by $p_B(z|x, \theta), p_B(\theta, \{x, y\}), q_B(z)$. Both models will have their own value for each of the functionals: R_A, S_A, D_A, C_A and R_B, S_B, D_B, C_B. Each model defines its own representation Z_A, Z_B. Now imagine coupling the models, by forming the joint representation $Z_C = (Z_A, Z_B)$ formed by concatenating the two representations together. Now the governing distributions over Z are simply the product of the two model's distributions, e.g. $q_C(z_C) = q_A(z_A)q_B(z_B)$. Thus the rate R_C and entropy S_C for the combined model is the sum of the individual models: $R_C = R_A + R_B, S_C = S_A + S_B$.

Now imagine we sample new states for the combined system which are maximally entropic with the constraint that the combined rate stay constant:

$$\min S \text{ s.t. } R = R_C \implies p(\theta|\{x, y\}) = \frac{q(\theta)}{\mathcal{Z}} e^{-R/\sigma}. \tag{32}$$

For the expectation of the two rates to be unchanged after they have been coupled and evolved holding their total rate fixed, we must have,

$$-\frac{1}{\sigma} R_A - \frac{1}{\sigma_B} R_B = -\frac{1}{\sigma_C} R_C = -\frac{1}{\sigma_C}(R_A + R_B)$$
$$\implies \sigma_A = \sigma_B = \sigma_C. \tag{33}$$

Therefore, we can see that σ, the effective temperature, allows us to identify whether two systems are in thermal equilibrium with one another. Just as in thermodynamics, if two systems at different temperatures are coupled, some transfer takes place.

3.4. Second Law of Learning?

Even when doing deterministic training, training is non-invertible (Maclaurin et al., 2015), and we need to contend with and track the entropy (S) term. We set the parameters of our networks initially with a fair draw from some prior distribution $q(\theta)$. The training procedure acts as a Markov process on the distribution of parameters, transforming it from the prior distribution into some modified distribution, the posterior $p(\theta|\{x,y\})$. Optimization is a many-to-one function, that in the ideal limiting case, maps all possible initializations to a single global optimum. In this limiting case S would be divergent, and there is nothing to prevent us from memorizing the training set.

The Second Law of Thermodynamics states that the entropy of an isolated system tends to increase. All systems tend to disorder, and this places limits on the maximum possible efficiency of heat engines.

Formally, there are many statements akin to the Second Law of Thermodynamics that can be made about Markov chains generally (Cover & Thomas, 2012). The central one is that for any for any two distributions p_n, q_n both evolving according to the same Markov process (n marks the time step), the relative entropy $\mathrm{KL}(p_n \,||\, q_n)$ is monotonically decreasing with time. This establishes that for a stationary Markov chain, the relative entropy to the stationary state $\mathrm{KL}(p_n \,||\, p_\infty)$ monotonically decreases [6].

In our language, we can make strong statements about dynamics that target points on the optimal frontier, or dynamics that implement a relaxation towards equilibrium. There is a fundamental distinction between states that live on the frontier and those off of it, analogous to the distinction between equilibrium and non-equilibrium states in thermodynamics.

Any equilibrium distribution can be expressed in the form Equation (19) and identified by its partial derviatives γ, δ, σ. If name the objective in Equation (17):

$$J(\gamma, \delta, \sigma) \equiv R + \delta D + \gamma C + \sigma S, \tag{34}$$

The value this objective takes for any equilibrium distribution can be shown to be given by the log partition function (Equation (20)):

$$\min J(\gamma, \delta, \sigma) = -\sigma \log \mathcal{Z}(\gamma, \delta, \sigma) \tag{35}$$

and the KL divergence between any distribution over parameters $p(\theta)$ and an equilibrium distribution is:

$$\mathrm{KL}(p(\theta) \,||\, p^*(\theta; \gamma, \delta, \sigma)) = \Delta J/\sigma \tag{36}$$

$$\Delta J \equiv J^{\text{noneq}}(p; \gamma, \delta, \sigma) - J(\gamma, \delta, \sigma) \tag{37}$$

Where J^{noneq} is the non-equilibrium objective:

$$J^{\text{noneq}}(p; \gamma, \delta, \sigma) = \langle R + \delta D + \gamma C + \sigma S \rangle_{p(\theta)} \,. \tag{38}$$

For a stationary Markov process whose stationary distribution is an equilibrium distribution the KL divergence to the stationary distribution must monotonically decrease each step. This means the $\Delta J/\sigma$ must decrease monotonically, that is our objective J must decrease monotonically:

$$J_{t=0} \geq J_t \geq J_{t+1} \geq J_{t=\infty}. \tag{39}$$

Furthermore, if we use $q(\theta)$ as our prior over parameters, we know:

$$J_{t=0} = \langle R + \delta D + \gamma C \rangle_{q(\theta)} \tag{40}$$

$$J_{t=\infty} = -\sigma \log Z. \tag{41}$$

4. Conclusion

We have formalized representation learning as the process of minimizing the distortion introduced when we project the real world (World P) onto the world we desire (World Q). The projection is naturally described by a set of four functionals which variationally bound relevant mutual informations in the real world. Relations between the functionals describe an optimal three-dimensional surface in a four dimensional space of *optimal* states. A single learning objective targeting points on this optimal surface can express a wide array of existing learning objectives spanning from unsupervised learning to supervised learning and everywhere in between. The geometry of the optimal frontier suggests a wide array of identities involving the functionals and their partial derivatives. This offers a direct analogy to thermodynamics independent of any physical content. By analogy to thermodynamics, we can begin to develop new quantitative measures and relationships amongst properties of our models that we believe will offer a new class of theoretical understanding of learning behavior.

Acknowledgements

The authors would like to Jascha Sohl-Dickstein, Mallory Alemi, Rif A. Saurous and Kevin Murphy for helpful discussions and feedback on the draft.

References

Achille, A. and Soatto, S. Emergence of Invariance and Disentangling in Deep Representations. *Proceedings of the ICML Workshop on Principled Approaches to Deep Learning*, 2017.

Alemi, Alexander A, Fischer, Ian, Dillon, Joshua V, and Murphy, Kevin. Deep variational information bottleneck. *arXiv:1612.00410*, 2016. URL http://arxiv.org/abs/1612.00410.

Alemi, Alexander A, Poole, Ben, Dillon, Joshua V, Saurous, Rif A, and Murphy, Kevin. Fixing a broken ELBO. *ICML*

[6] For discrete state Markov chains, this implies that if the stationary distribution is uniform, the entropy of the distribution $H(p_n)$ is strictly increasing.

2018, 2018. URL http://arxiv.org/abs/1711.00464.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight Uncertainty in Neural Networks. *arXiv: 1505.05424*, May 2015. URL https//arxiv.org/abs/1505.05424.

Box, George EP and Draper, Norman R. *Empirical model-building and response surfaces*. John Wiley & Sons, 1987.

Chen, T., Fox, E. B., and Guestrin, C. Stochastic Gradient Hamiltonian Monte Carlo. *arXiv:1402.4102*, February 2014. URL https://arxiv.org/abs/1402.4102.

Cover, Thomas M and Thomas, Joy A. *Elements of information theory*. John Wiley & Sons, 2012.

Csiszár, Imre and Matúš, František. Information projections revisited. *IEEE Transactions on Information Theory*, 49 (6):1474–1490, 2003.

Finn, Colin BP. *Thermal physics*. CRC Press, 1993.

Friedman, Nir, Mosenzon, Ori, Slonim, Noam, and Tishby, Naftali. Multivariate information bottleneck. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pp. 152–161. Morgan Kaufmann Publishers Inc., 2001.

Higgins, Irina, Matthey, Loic, Pal, Arka, Burgess, Christopher, Glorot, Xavier, Botvinick, Matthew, Mohamed, Shakir, and Lerchner, Alexander. β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. 2017.

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., and Weinberger, K. Q. Snapshot Ensembles: Train 1, get M for free. *arXiv: 1704.00109*, March 2017. URL https://arxiv.org/abs.1704.00109.

Jaynes, Edwin T. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Kingma, D. P., Rezende, D. J., Mohamed, S., and Welling, M. Semi-Supervised Learning with Deep Generative Models. *arXiv: 1406.5298*, June 2014. URL https://arxiv.org/abs/1406.5298.

Kingma, Diederik P and Welling, Max. Auto-encoding variational Bayes. 2014.

Ma, Y.-A., Chen, T., and Fox, E. B. A Complete Recipe for Stochastic Gradient MCMC. *arXiv:1506.04696*, June 2015. URL https://arxiv.org/abs/1506.04696.

Machta, J. and Ellis, R. S. Monte Carlo Methods for Rough Free Energy Landscapes: Population Annealing and Parallel Tempering. *Journal of Statistical Physics*, 144:541–553, August 2011. doi: 10.1007/s10955-011-0249-0. URL https://arxiv.org/abs/1104.1138.

Maclaurin, Dougal, Duvenaud, David, and Adams, Ryan P. Gradient-based hyperparameter optimization through reversible learning. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 2113–2122. JMLR.org, 2015. URL http://dl.acm.org/citation.cfm?id=3045118.3045343.

Mandt, S., Hoffman, M. D., and Blei, D. M. Stochastic Gradient Descent as Approximate Bayesian Inference. *arXiv: 1704.04289*, April 2017. URL https://arxiv.org/abs/1704.04289.

Marsh, Charles. Introduction to continuous entropy. 2013. URL http://www.crmarsh.com/static/pdf/Charles_Marsh_Continuous_Entropy.pdf.

Parrondo, Juan MR, Horowitz, Jordan M, and Sagawa, Takahiro. Thermodynamics of information. *Nature physics*, 11(2):131–139, 2015. URL http://jordanmhorowitz.mit.edu/sites/default/files/documents/natureInfo.pdf.

Sethna, James. *Statistical mechanics: entropy, order parameters, and complexity*, volume 14. Oxford University Press, 2006. URL http://pages.physics.cornell.edu/~sethna/StatMech/EntropyOrderParametersComplexity.pdf.

Smith, S. L. and Le, Q. V. A Bayesian Perspective on Generalization and Stochastic Gradient Descent. *arXiv:1710.06451*, October 2017. URL https://arxiv.org/abs/1710.06451.

Still, S. Thermodynamic cost and benefit of data representations. *arXiv: 1705.00612*, April 2017. URL https://arxiv.org/abs/1705.00612.

Still, S., Sivak, D. A., Bell, A. J., and Crooks, G. E. Thermodynamics of Prediction. *Physical Review Letters*, 109(12): 120604, September 2012. doi: 10.1103/PhysRevLett.109.120604. URL https://arxiv.org/abs/1203.3271.

Tishby, N., Pereira, F.C., and Biale, W. The information bottleneck method. In *The 37th annual Allerton Conf. on Communication, Control, and Computing*, pp. 368–377, 1999. URL https://arxiv.org/abs/physics/0004057.

Welling, Max and Teh, Yee W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688, 2011.

Zhang, Y., Saxe, A. M., Advani, M. S., and Lee, A. A. Energy-entropy competition and the effectiveness of stochastic gradient descent in machine learning. *arXiv: 1803.01927*, March 2018. URL https://arxiv.org/abs/1803.01927.

A. Reconstruction Free Formulation

We can utilize the Chain Rule of Mutual Information (Equation (10)):

$$I(Z_i; X_i, \Theta) = I(Z_i; X_i) + I(Z_i; \Theta | X_i), \qquad (42)$$

to simplify our expression for the minimum possible KL between worlds (Equation (9)), and consider a reduced set of functionals (compare to Section 1.1):

- $C \equiv -\sum_i \langle \log q(y_i|z_i) \rangle_P \geq \sum_i H(Y_i) - I(Y_i; Z_i) = \sum_i H(Y_i|Z_i)$

  The *classification error*, as before.

- $S \equiv \left\langle \log \frac{p(\theta|\{x,y\})}{q(\theta)} \right\rangle_P \geq I(\Theta; \{X, Y\})$

  The *entropy* as before.

- $V \equiv \left\langle \log \frac{p(z_i|x_i,\theta)}{q(z_i|x_i)} \right\rangle_P \geq I(Z_i; \Theta | X_i)$

  The *volume* of the representation (for lack of a better term), which measures the mutual information between our representation Z and the parameters Θ, conditioned on the input X. That is, this functional bounds how much of the information in our representation can come from the learning algorithm, independent of the actual input.

In principle, these three functionals still fully characterize the distortion introduced in our information projection. Notice that this new functional requires the variational approximation $q(z_i|x_i)$, a variational approximation to the marginal over our parameter distribution. Notice also that we no longer require a variational approximation to $p(x_i|z_i)$. That is, in this formulation we no longer require any form of decoder, or sythesis in our original data space X. While equivalent in its information projection, this more naturally corresponds to the model of our desired world Q:

$$q(x, y, \phi, z, \theta) = q(\phi) q(\theta) \prod_i q(z_i|x_i) q(y_i|z_i), \qquad (43)$$

depicted below in Figure 2. Here we desire, not the joint generative model $X \leftarrow Z \rightarrow Y$, but the predictive model $X \rightarrow Z \rightarrow Y$.

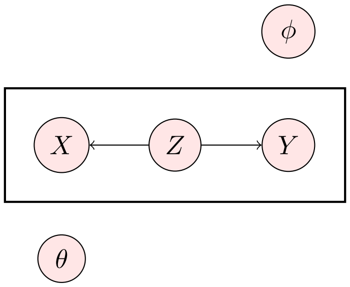


Figure 2. Modified graphical model for world Q, the world we desire which satisfies the joint density in Equation 43.

B. Experiments

We show examples of models trained on a toy dataset for all of the different objectives we define above. The dataset has both an infinite data variant, where overfitting is not a problem, and a finite data variant, where overfitting can be clearly observed for both reconstruction and classification.

Data generation. We follow the toy model from Alemi et al. (2018), but add an additional classification label in order to explore supervised and semi-supervised objectives. The true data generating distribution is as follows. We first sample a latent binary variable, $z \sim \text{Ber}(0.7)$, then sample a latent 1D continuous value from that variable, $h|z \sim \mathcal{N}(h|\mu_z, \sigma_z)$, and finally we observe a discretized value, $x = \text{discretize}(h; \mathcal{B})$, where $\mathcal{B}$ is a set of 30 equally spaced bins, and a discrete label, $y = z$ (so the true label is the latent variable that generated x). We set μ_z and σ_z such that $R^* \equiv \text{I}(x; z) = 0.5$ nats, in the true generative process, representing the ideal rate target for a latent variable model. For the finite dataset, we select 50 examples randomly from the joint $p(x, y, z)$. For the infinite dataset, we directly supply the true full marginal $p(x, y)$ at each iteration during training. When training on the finite dataset, we evaluate model performance against the infinite dataset so that there is no error in the evaluation metrics due to a finite test set.

Model details. We choose to use a discrete latent representation with $K = 30$ values, with an encoder of the form $q(z_i|x_j) \propto -\exp[(w_i^e x_j - b_i^e)^2]$, where z is the one-hot encoding of the latent categorical variable, and x is the one-hot encoding of the observed categorical variable. We use a decoder of the same form, but with different parameters: $q(x_j|z_i) \propto -\exp[(w_i^d x_j - b_i^d)^2]$. We use a classifier of the same form as well: $q(y_j|z_i) \propto -\exp[(w_i^c y_j - b_i^c)^2]$. Finally, we use a variational marginal, $q(z_i) = \pi_i$. Given this, the true joint distribution has the form $p(x, y, z) = p(x)p(z|x)p(y|x)$, with marginal $p(z) = \sum_x p(x, z)$, and conditionals $p(x|z) = p(x, z)/p(z)$ and $p(y|z) = p(y, z)/p(z)$.

The encoder is additionally parameterized following Achille & Soatto (2017) by α, a set of learned parameters for a Log Normal distribution of the form $\log \mathcal{N}(-\alpha_i/2, \alpha_i)$. In total, the model has 184 parameters: 60 weights and biases in the encoder and decoder, 4 weights and biases in the classifier, 30 weights in the marginal, and an additional 30 weights for the α_i parameterizing the stochastic encoder. We initialize the weights so that when $\sigma = 0$, there is no noticeable effect on the encoder during training or testing.

Experiments. In Figure 3, we show the optimal, hand-crafted model for the toy dataset, as well as a selection of parameterizations of the TherML objective that correspond to commonly-used objective functions and a few new objective functions not previously described. In the captions, the parameters are specified with γ, δ, σ as in the main text, as well as ρ, which is a corresponding lagrange multiplier for R, in order to simplify the parameterization. This is still valid, so long as there is always one function whose multiplier is 1. It just parameterizes the optimal surface slightly differently. We train all objectives for 10,000 gradient steps. For all of the objectives described, the model has converged, or come close to convergence, by that point.

Because the model is sufficiently powerful to memorize the dataset, most of the objectives are very susceptible to overfitting. Only the objective variants that are "regularized" by the S term (parameterized by σ) are able to avoid overfitting in the decoder and classifier.

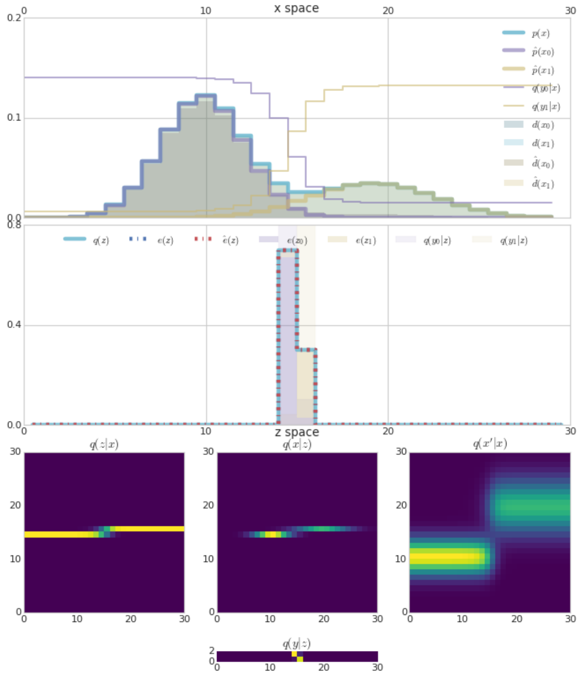
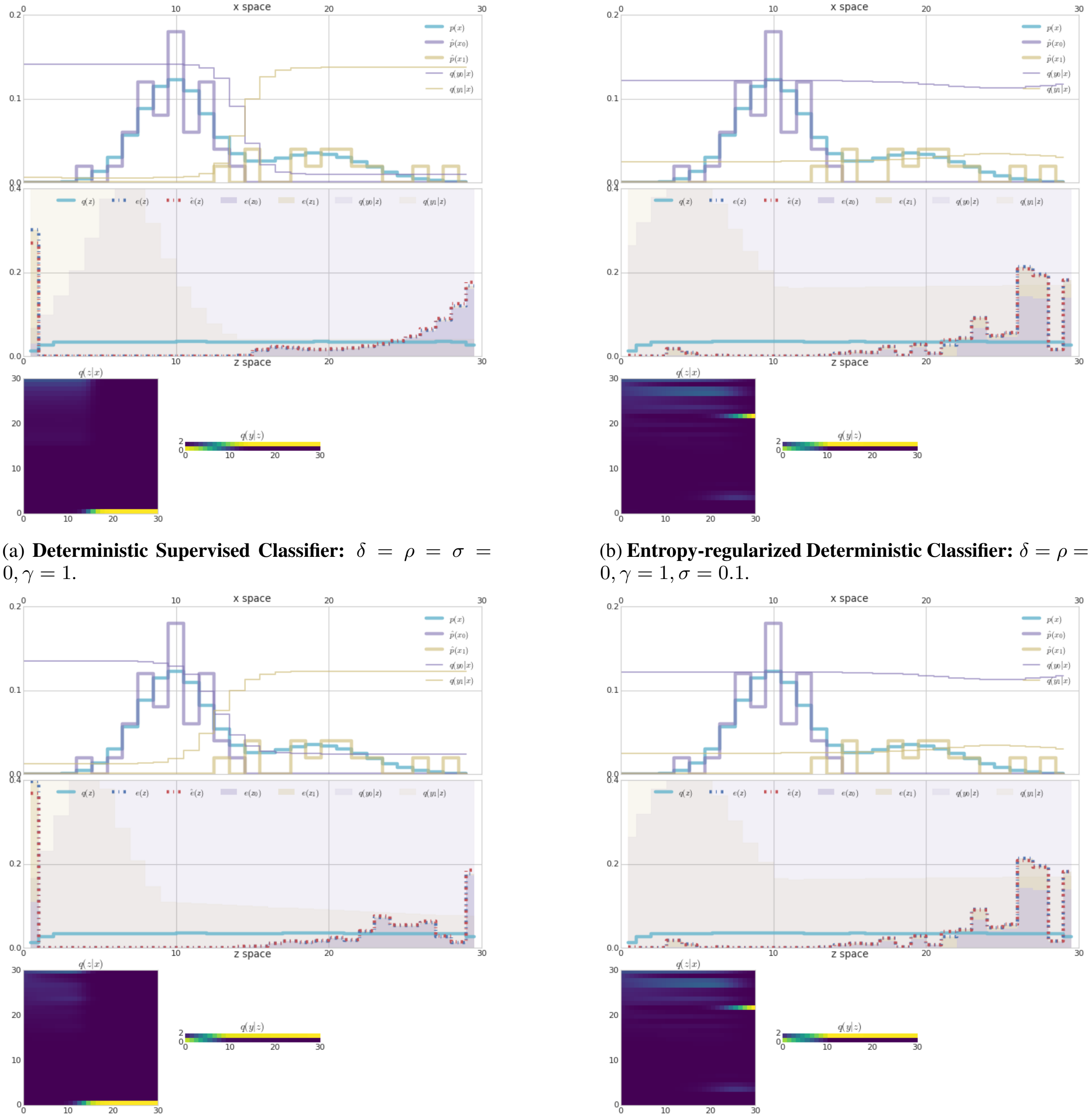


Figure 3. **Hand-crafted optimal model.** Toy Model illustrating the difference between selected points on the three dimensional optimal surface defined by γ, δ, and σ. See Section 2 for more description of the objectives, and Appendix B for details on the experiment setup. **Top (i):** Three distributions in data space: the true data distribution, $p(x)$, the model's generative distribution, $g(x) = \sum_z q(z)q(x|z)$, and the empirical data reconstruction distribution, $d(x) = \sum_{x'} \sum_z p(x')q(z|x')q(x|z)$. **Middle (ii):** Four distributions in latent space: the learned (or computed) marginal $q(z)$, the empirical induced marginal $e(z) = \sum_x p(x)q(z|x)$, the empirical distribution over z values for data vectors in the set $\mathcal{X}_0 = \{x_n : z_n = 0\}$, which we denote by $e(z_0)$ in purple, and the empirical distribution over z values for data vectors in the set $\mathcal{X}_1 = \{x_n : z_n = 1\}$, which we denote by $e(z_1)$ in yellow. **Bottom:** Three $K \times K$ distributions: (iii) $q(z|x)$, (iv) $q(x|z)$ and (v) $q(x'|x) = \sum_z q(z|x)q(x'|z)$.



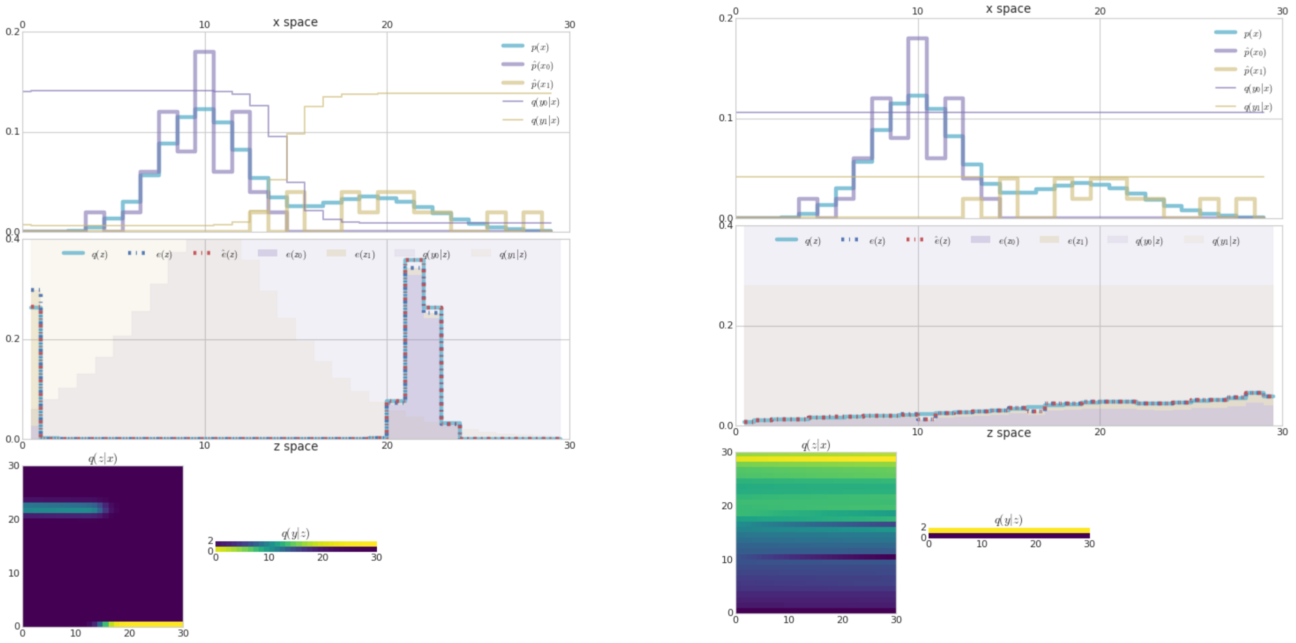
(a) **Deterministic Supervised Classifier:** $\delta = \rho = \sigma = 0, \gamma = 1$.

(b) **Entropy-regularized Deterministic Classifier:** $\delta = \rho = 0, \gamma = 1, \sigma = 0.1$.

(c) **Entropy-regularized IB:** $\delta = 0, \rho = 0, \gamma = 1, \sigma = 0.01$.

(d) **Bayesian Neural Network Classifier:** $\delta = 0, \rho = 0, \sigma = \gamma = 1$.

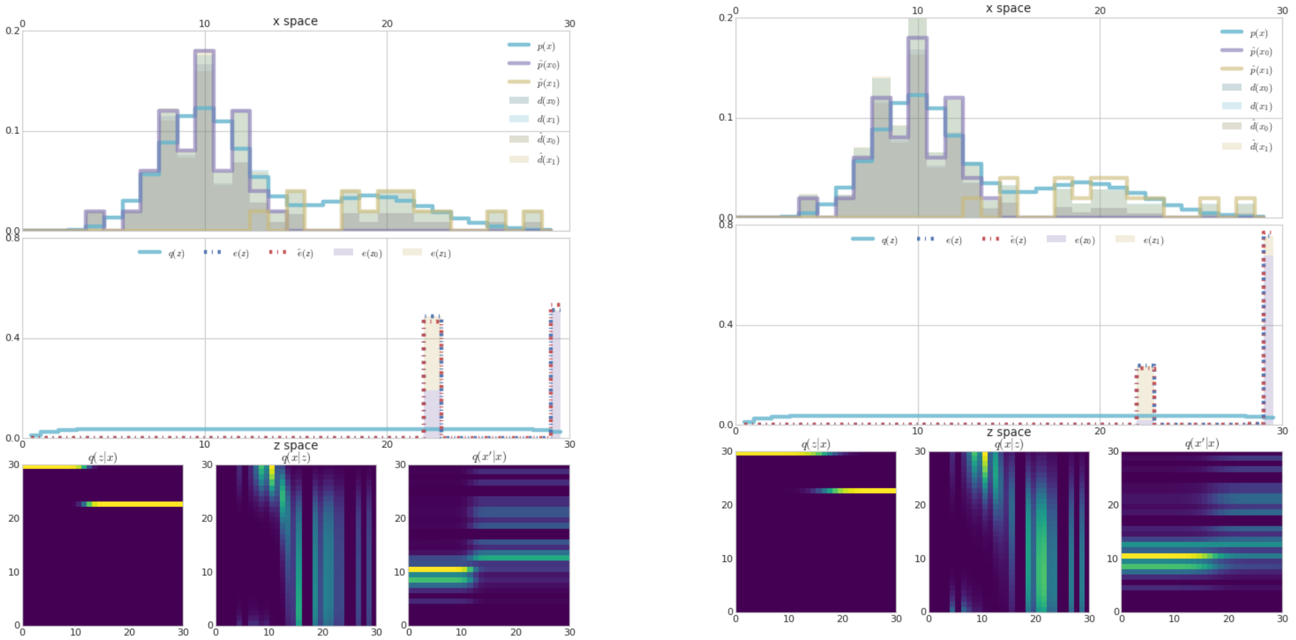
Figure 4. Supervised Learning approaches.



(a) **VIB:** $\delta = 0, \sigma = 0, \gamma = 1, \rho(\beta) = 0.5$.

(b) **Entropy-regularized VIB:** $\delta = 0, \gamma = 1, \rho = 0.9, \sigma = 0.1$.

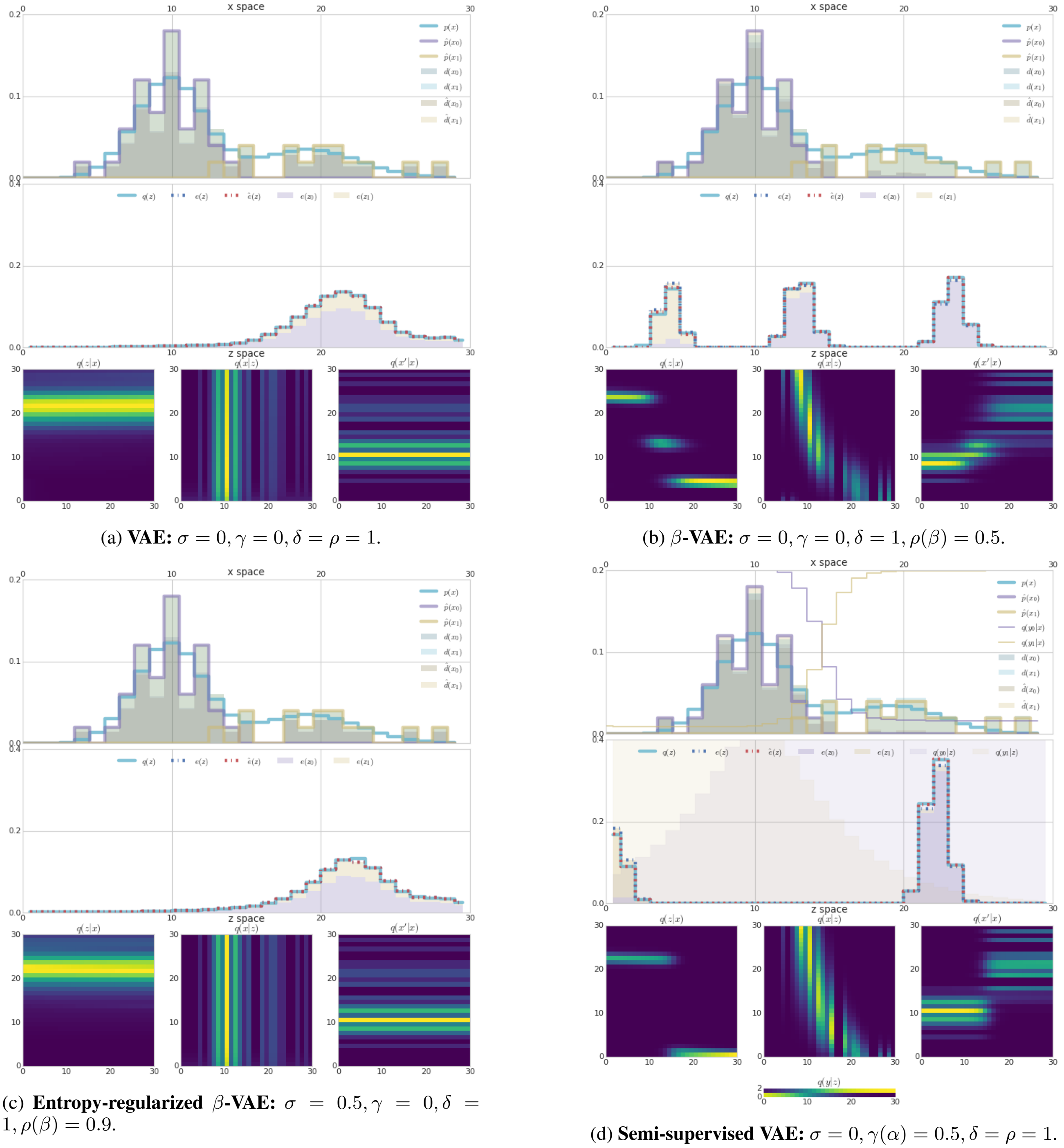
Figure 5. VIB style objectives.



(a) **Deterministic Autoencoder:** $\gamma = \rho = \sigma = 0, \delta = 1$.

(b) **Entropy-regularized Deterministic Autoencoder:** $\gamma = \rho = 0, \delta = 1, \sigma = 0.01$.

Figure 6. Autoencoder objectives.



(a) **VAE:** $\sigma = 0, \gamma = 0, \delta = \rho = 1$.

(b) **β-VAE:** $\sigma = 0, \gamma = 0, \delta = 1, \rho(\beta) = 0.5$.

(c) **Entropy-regularized β-VAE:** $\sigma = 0.5, \gamma = 0, \delta = 1, \rho(\beta) = 0.9$.

(d) **Semi-supervised VAE:** $\sigma = 0, \gamma(\alpha) = 0.5, \delta = \rho = 1$.

Figure 7. VAE style objectives.

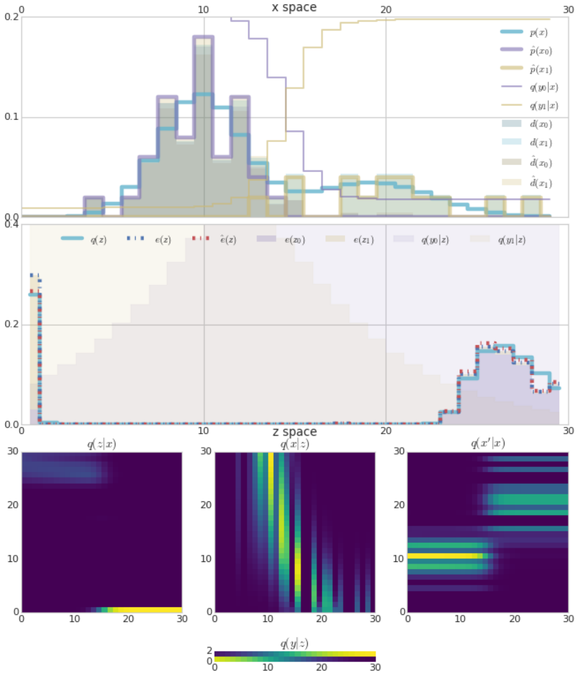


Figure 8. **Full Objective.** $\sigma = 0.5, \gamma = 1000, \delta = 1, \rho = 0.9$. Simple demonstration of the behavior with all terms present in the objective.