Loss as the Inconsistency of a Probabilistic Dependency Graph: Choose your Model, not your Loss Function

Anonymous Author
Anonymous Institution

Abstract

In a world blessed with a great diversity of loss functions, we argue that that choice between them is not a matter of taste, but a modeling assumption. Probabilistic Dependency Graphs (PDGs) are probabilistic models that come equipped with a measure of inconsistency. We prove that many standard loss functions are the inconsistencies of a natural PDG describing the appropriate scenario. The same approach can be used to justify a well-known connection between regularizers and priors. We also show that the inconsistency of PDGs captures a large class of statistical divergences, and detail some benefits of thinking of them in this way, including an intuitive visual proof language for deriving inequalities between them. In variational inference, we find that the ELBO, a difficult-to-motivate function for training variational models, and variants of it, arise for free out of uncontroversial modeling assumptions. So do intuitive visual proofs of the variational bounds used to justify them. Finally, we observe that inconsistency becomes the log partition function (free energy) in the setting where PDGs are factor graphs.

1 INTRODUCTION

Many tasks in artificial intelligence have been fruitfully cast as optimization problems, but often the choice of objective is not unique. For instance, a key component of a machine learning system is a loss function which the system minimizes, a wide variety of which are used in pratice. Each implicitly represents different values, and results in different behavior, so the choice of loss can be quite important (Wang et al. 2020; Jadon 2020). Yet, because it's unclear how to choose a "good" loss function, the choice is usually made by emperics, tradition, and an instinctive calculus acquired through the practice—not explicitly laying out beliefs. Furthermore, there is something to be gained by fiddling with these loss functions: one can add regularization terms, to (dis)incentivize (un)desirable behavior. But this process—tinkering with the objective and afterwards spinning a story about why it works—is unsatisfying. It can be a tedious game without clear rules or meaning, while results so obtained are arguably overfit and difficult to motivate.

By contrast, a choice of *model* admits more principled discussion, in part because it makes sense to ask if a model is accurate, or of it captures a real phenomenon. In this framework, it is no longer possible to specify an objective directly; rather, one must articulate a situation that gives rise to it, in the (more interpretable) language of probablistic beliefs and certainties. Concretely, we use the machinery of Probabilistic Dependency Graphs (PDGs), a particularly expressive class graphical models that can can incorporate aribtrary (even inconsistent) probabilistic information in a natural way, and comes equipped with a well-motivated measure of inconsistency (Richardson and Halpern 2021).

The goal of this paper is to show that PDGs and their associated inconsistency measure can proide a "universal" model-based loss function. Towards this end, we show that many standard objective functions—cross entropy, square error, many statistical distances, the ELBO, regularizers, and the log partition function—arise naturally as inconsistencies of the appropriate underlying PDG. This is somewhat surprising, since PDGs were not designed with the goal of capturing loss functions at all. Specifying a loss function indirectly like this may be more restrictive, but it is also more intuitive (since it requires no technical familiarity with losses), and admits more epistemically grounded support and criticism.

Preliminary work. Under review by AISTATS 2022. Do not distribute.

For a particularly powerful demonstration, consider the variational autoencoder (VAE), an enormously successful class of generative model that has enabled breakthroughs in image generation, semantic interpolation, and unsupervised feature learning (Kingma and Welling 2014). Structurally, a VAE for a space X consists of a (smaller) latent space Z, a prior distribution $p(Z)$, a decoder $d(Z|X)$, and an encoder $e(Z|X)$. A VAE is not considered a "graphical model" for two reasons. The first is that the encoder $e(Z|X)$ has the same target variable as $p(Z)$, so something like a Bayesian Network cannot simultaneously incorporate them both (besides, they could be inconsistent with one another). The second reason: it is not a VAE's structure, but rather its *loss function* that makes it tick. A VAE is typically trained by maximizing the "ELBO", a somewhat difficult-to-motivate function of a sample x, borrowed from variational analysis. We show that $\mathrm{ELBO}(x)$ is also precisely the inconsistency of a PDG containing the probabilistic information of the autoencoder (p, d, and e) and the sample x. Thus, PDG semantics simultaneously legitimize the strange structure the VAE, and also justify its loss function, which can be thought of as a property of the model itself (its inconsistency), rather than some mysterious construction borrowed from physics.

Representing objectives as model inconsistencies, in addition to providing a principled way of selecting an objective, also has beneficial pedagogical side effects, because of the relationships between the underlying models. For instance, we will be able to use the structure of the PDG to get simple, intuitive proofs of technical results, such as the variational inequalitites that traditionally motivate the ELBO, and the monotonicity of Rényi divergence.

In the coming sections, we will show in more detail how this concept of inconsistency, beyond simply providing a permissive and intuitive modeling framework, reduces exactly to many standard objectives used in machine learning, measures of statistical distance, and also demonstrate that this framework clarifies the relationships between them, by providing simple clear derivations of otherwise opaque inequalities.

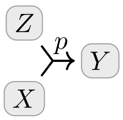
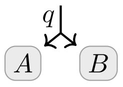
2 PRELIMINARIES

We generally use capital letters for variables, and lower case letters for their values. For variables X and Y, a conditional probability distribution (cpd) p on Y given X, written $p(Y|X)$, consists of a probability distribution on Y (denoted $p(Y \mid X\!=\!x)$ or $p(Y \mid x)$ for short), for each possible value x of X. If μ is a probability on outcomes that determine X and Y, then $\mu(X)$ denotes the marginal of μ on X, and $\mu(Y|X)$ denotes the conditional marginal of μ on Y given X. Depending on which we find clearer in context, we write either $\mathbb{E}_\mu f$ or $\mathbb{E}_{\omega\sim\mu} f(\omega)$ for expectation of $f : \Omega \to \mathbb{R}$ over a distribution μ with outcomes Ω. We write $\boldsymbol{D}(\mu \parallel \nu) = \mathbb{E}_\mu \log \frac{\mu}{\nu}$ for the relative entropy (KL Divergence) of ν with respect to μ, and for finitely supported μ, we write $\mathrm{H}(\mu) := \mathbb{E}_\mu \log \frac{1}{\mu}$ for the entropy of μ, $\mathrm{H}_\mu(X) := \mathrm{H}(\mu(X))$ for the marginal entropy on a variable X, and $\mathrm{H}_\mu(Y \mid X) := \mathbb{E}_\mu \log {}^1\!/_{\mu(Y|X)}$ for the conditional entropy of Y given X.

A *probabilistic dependency graph* (PDG) (Richardson and Halpern 2021), like a Bayesian Network (BN), is a directed graph with cpds attached to it. While this data is attached to the *nodes* of a BN, it is attached to the *edges* of PDG. For instance, a PDG of shape $X \to Y \leftarrow Z$ contains both cpd $p(Y|X)$ and a cpd $q(Y|Z)$, while a BN of the same shape has a single cpd $\Pr(Y|X, Z)$ on Y given joint values of X, Z. The first interpretation is more expresive, and can encode joint dependence with an extra variable and pair of edges. We now restate the formal definition.

Definition 1. A Probabilistic Dependency Graph (PDG) is a tuple $\boldsymbol{\mathcal{m}} = (\mathcal{N}, \mathcal{E}, \mathcal{V}, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, where

- $\mathcal{N}$ is a set of nodes, corresponding to variables;
- $\mathcal{V}$ associates each node $X \in \mathcal{N}$ with a set $\mathcal{V}(X)$ of possible values that the variable X can take;
- $\mathcal{E}$ is a set of labeled edges $\{X \xrightarrow{L} Y\}$, each with a source X and target Y from $\mathcal{N}$;
- $\mathbf{p}$ associates a cpd $\mathbf{p}_{\!_L}(Y|X)$ to each edge $X \xrightarrow{L} Y \in \mathcal{E}$;
- $\boldsymbol{\alpha}$ associates to each edge $X \xrightarrow{L} Y$ a non-negative number α_L representing the modeler's confidence in the functional dependence of Y on X;
- $\boldsymbol{\beta}$ associates to each edge L a number $\beta_L \in \mathbb{R} \cup \{\infty\}$, the modeler's subjective confidence in the reliability of the cpd $\mathbf{p}_{\!_L}$. □

This presentation is equivalent to one in which edge sources and targets are both *sets* of variables. This allows us to indicate joint dependence with multi-tailed arrows, joint distributions with multi-headed arrows, and unconditional distributions with nothing at the tail. For instance, we will draw $p(X)$ as $\xrightarrow{p}\boxed{X}$, $p(Y|X, Z)$ as , and $q(A, B)$ as .

Like other graphical models, PDGs have semantics in terms of joint distributions μ over all variables. Most directly, a PDG $\boldsymbol{\mathcal{m}}$ determines two scoring functions on joint distributions μ. For the purposes of this paper, the more important of the two is the *incompatibility* of μ with respect to $\boldsymbol{\mathcal{m}}$, which measures the *quantitative* discrepency between μ and the cpds of $\mathcal{M}$, is given by

$$\mathit{Inc}_{\mathcal{M}}(\mu) := \sum_{X \xrightarrow{L} Y} \beta_L \cdot \mathop{\mathbb{E}}_{x \sim \mu(X)} \boldsymbol{D}\Big(\mu(Y|x) \,\Big\|\, \mathbf{p}_L(Y|x)\Big). \quad (1)$$

It is well known that $\boldsymbol{D}(\mu \parallel p)$ is a measure of divergence between μ and p, which can be interpreted as overhead (in extra bits per sample) of using codes optimized for p, when in fact samples are distributed according to μ (MacKay 2003). But if one uses edges in proportion to the confidence one has in them, then μ's violations of high-confidence cpds are compounded, and hence more costly. So $\mathit{Inc}_{\mathcal{M}}(\mu)$ measures the total expected excess cost of using the cpds of $\mathcal{M}$ in proportion to the confidence the modeler has in them, in a world drawn from μ. The *inconsistency* of $\mathcal{M}$, denoted $\langle\!\langle \mathcal{M} \rangle\!\rangle := \inf_\mu \mathit{Inc}_{\mathcal{M}}(\mu)$, is the smallest possible incompatibility with any distribution is the smallest possible incompatibility of any distribution with respect to $\mathcal{M}$.

The second scoring function defined by a PDG $\mathcal{M}$, called information deficiency, measures the *qualitative* discrepency between $\mathcal{M}$ and μ, and is given by

$$\mathit{IDef}_{\mathcal{M}}(\mu) := -\operatorname{H}(\mu) + \sum_{X \xrightarrow{L} Y} \alpha_L \operatorname{H}_\mu(Y \mid X).$$

$\mathit{IDef}_{\mathcal{M}}(\mu)$ can be thought of as measuring the (weighted) number of bits needed to describe the target of each edge given its source, beyond the information needed to describe a full sample of μ.

As shown by Richardson and Halpern (2021), it is via these two scoring functions that PDGs capture other graphical models: the distribution specified by a BN $\mathcal{B}$ is the unique one that minimizes both $\mathit{Inc}_{\mathcal{B}}$ and $\mathit{IDef}_{\mathcal{B}}$ (and hence every positive linear combination of the two), and the distribution specfied by a factor graph Φ uniquely minimizes the sum $\mathit{Inc}_\Phi + \mathit{IDef}_\Phi$. In general, for any $\gamma > 0$, the one can consider a weighted combination $[\![\mathcal{M}]\!]_\gamma(\mu) := \mathit{Inc}_{\mathcal{M}}(\mu) + \gamma\, \mathit{IDef}_{\mathcal{M}}(\mu)$, for which there is a corresponding γ-inconsistency $\langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma := \inf_\mu [\![\mathcal{M}]\!]_\gamma(\mu)$. In the limit as $\gamma \to 0$, there is always a unique best distribution whose score is $\langle\!\langle \mathcal{M} \rangle\!\rangle$.

We now present some shorthand to simplify the presentation. We identify an event $X{=}x$ with with the degenerate distribution on X that places all mass on x, and hence may be associated to an edge. We label such an edge simply as 'x'. To emphasize that a cpd $f(Y|X)$ is degenerate (a function $f : X \to Y$), we will draw it with two heads, as in: $\boxed{X} \overset{f}{\twoheadrightarrow} \boxed{Y}$. By default, edges have $\beta = 1$. To specify a confidence $\beta \neq 1$, we place the value near the edge, lightly colored and parenthesized, as in: $\xrightarrow[(\beta)]{p} \boxed{X}$. We use (∞) to denote the limit of high confidence ($\beta \to \infty$).

Intuitively, believing more things can't make you any less inconsistent. Lemma 1 captures this formally: adding information or increasing a confidence cannot decrease a PDG's inconsistency.

Lemma 1. *Suppose PDGs $\mathcal{M}$ and $\mathcal{M}'$ differ only in their edges (resp. $\mathcal{E}$ and $\mathcal{E}'$) and confidences (resp. $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$). If $\mathcal{E} \subseteq \mathcal{E}'$ and $\beta_L \leq \beta'_L$ for all $L \in \mathcal{E}$, then $\langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma \leq \langle\!\langle \mathcal{M}' \rangle\!\rangle_\gamma$ for all γ.*[1]

[link to proof]

This tool is sufficient to derive many interesting relationships between loss functions.

3 STANDARD LOSSES AS INCONSISTENCIES

Let's start with a simple example. Suppose you believe that X is distributed according to $p(X)$, and also that it (always) equals a certain value x. These beliefs are consistent if $p(x) = 1$ and become more inconsistent as $p(x)$ decreases. In fact, this inconsistency is equal to $\operatorname{I}_p(x) = -\log p(x)$, the information content, or *surprisal* (Tribus 1961), of the event $X{=}x$, according to p.[2] In machine learning, I_p is more often called "negative log likelihood", and is perhaps the most popular objective for training generative models (Grover and Ermon 2018; Myung 2003).

Proposition 2. *Consider a distribution $p(X)$. The surprisal $\operatorname{I}_p(x)$ of the event $X{=}x$ equals inconsistency of the PDG containing p and $X{=}x$. That is,*

$$\operatorname{I}_p(x) := \log \frac{1}{p(X{=}x)} = \Big\langle\!\!\Big\langle \xrightarrow{p} \boxed{X} \overset{x}{\twoheadleftarrow} \Big\rangle\!\!\Big\rangle.$$

[link to proof]

In some ways, this result is entirely unsurprising, given that PDG inconsistency (1) is a flexible formula built out of information theoretic primitives. Even so, it is worth noting that the inconsistency of the PDG containing just a distribution $p(X)$ and a sample x happens to be the standard relationship between a distribution p and sample x—and it is even named after "surprise", a certain expression of epistemic conflict.

Still, we have a ways to go before this amounts to any more than a curiosity. One concern is that this picture is incomplete; we train probabilstic models models with more than one sample. What if we replce x with an empirical distribution on many samples?

Proposition 3. *If $p(X)$ is a probabilistic model of X, and $\underline{\mathbf{x}} = \{x_i\}_{i=1}^m$ are samples with empirical distribution $\Pr_{\underline{\mathbf{x}}}$, then* $\operatorname{CrossEntropy}(\Pr_{\underline{\mathbf{x}}}, p) =$

$$\frac{1}{m}\sum_{i=1}^m \operatorname{I}_p(x_i) = \Big\langle\!\!\Big\langle \xrightarrow{p} \boxed{X} \xleftarrow[(\infty)]{\Pr_{\underline{\mathbf{x}}}} \Big\rangle\!\!\Big\rangle + \operatorname{H}(\Pr_{\underline{\mathbf{x}}}).$$

[link to proof]

[1] All proofs can be found in Appendix C.

[2] We have implicitly identified p with its probability mass function. One can get similar, but more mudled results with densities. See Appendix A for details.

Remark 1. *The term $H(\mathrm{Pr}_{\underline{\mathbf{x}}})$ is a constant depending only on the data, so is irrelevant for optimizing p.*

Really the only choices we've made in specifying the PDG of Proposition 3 are the confidences. But $\mathrm{CrossEntropy}(\mathrm{Pr}_{\underline{\mathbf{x}}}, p)$ is the expected code length per sample from $\mathrm{Pr}_{\underline{\mathbf{x}}}$, when using codes optimized for the (incorrect) distribution p. So implicitly, a practitioner using cross-entropy has already articulated a belief the data distribution is the "true one". To get the same effect from a PDG, the modeler must make this belief explicit by placing infinite certainty in $\mathrm{Pr}_{\underline{\mathbf{x}}}$.

We now consider an orthogonal generalization of Proposition 2, in which the sample x is only a partial observation of a joint model $p(X, Z)$. In this case, we might hope to recover the *marginal* surprise, since Z does not interact with the observation.

link to proof

Proposition 4. *If $p(X, Z)$ is a joint distribution, the information content of the partial observation $X = x$, or the marginal negative log likelihood of x, is given by*

$$\mathrm{I}_p(x) = \log \frac{1}{p(x)} = \left\langle\!\!\left\langle \begin{tikzcd} & p \arrow[dl] \arrow[dr] & & \\ \boxed{Z} & & \boxed{X} & x \arrow[l, two heads] \end{tikzcd} \right\rangle\!\!\right\rangle . \tag{2}$$

Intuitively, the inconsistency in the PDG on the right hand side of (2) is localized to x, where the sample conflicts with the distribution; other variables don't make a difference. The natural generalization with both partial observations and multiple samples also holds, and we defer treatment to the appendix.

So far we have considered models of an unconditional distribution $p(X)$. Because they are unconditional, such models must describe how to generate a complete sample X without input, and so are called *generative*, and the process of training them is called *unsupervised* learning. Contrast this with the (perhaps more common) *supervised* setting, in which we train *discriminative* models to predict Y from X, from labeled samples (X, Y). For supervised learning, cross entropy loss is perhaps even more dominant—and it is essentially the inconsistency of a PDG consisting of the predictor $f(Y|X)$ together with high-confidence data.

link to proof

Proposition 5. *Consider a probabilistic predictor $f(Y \mid X)$. The inconsistency of the PDG containing f and the empirical distribution $\mathrm{Pr}_{\underline{\mathbf{xy}}}$ of a sample set $\underline{\mathbf{xy}} = \{(x_i, y_i)\}_{i=1}^m$ is equal to the cross-entropy loss (minus the empirical uncertainty in Y given X, a constant that depends only on the data). That is,*

$$\left\langle\!\!\left\langle \begin{tikzcd} & \mathrm{Pr}_{\underline{\mathbf{xy}}} \arrow[dl] \arrow[dr, "{\scriptstyle(\infty)}"] & \\ \boxed{X} \arrow[rr, "f"'] & & \boxed{Y} \end{tikzcd} \right\rangle\!\!\right\rangle = \frac{1}{m} \sum_{i=1}^{m} \log \frac{1}{f(y_i \,|\, x_i)} - \mathrm{H}_{\mathrm{Pr}_{\underline{\mathbf{xy}}}}(Y|X).$$

There are also simpler scoring metrics used to evaluate the performance of systems on datasets, such as the accuracy of a classifier, or the mean-squared error of a regressor. These too naturally arise as inconsistencies.

link to proof

Proposition 6 (log accuracy as inconsistency)**.** *Consider a distribution $D(X)$ over inputs X, and functions $f, h : X \to Y$, where h is a predictor and f generates the true labels. The inconsistency of believing all three is the log accuracy of h. That is,*

$$\left\langle\!\!\left\langle \begin{tikzcd} D \arrow[r, "{\scriptstyle(\beta)}"'] & \boxed{X} \arrow[r, bend left, two heads, "h"] \arrow[r, bend right, two heads, "f"'] & \boxed{Y} \end{tikzcd} \right\rangle\!\!\right\rangle \begin{aligned} &= -\beta \log \Pr_{x \sim D}(f(x) {=} h(x)) \\ &= \beta \, \mathrm{I}_D[f = h]. \end{aligned} \tag{3}$$

One often speaks of the accuracy of a hypothesis h (leaving the true labels f and empirical distribution D implicit). But in some sense, $D(X)$ plays a more primary role: the inconsistency in (3) is scaled by the confidence in D, and does not depend at all on the confidences in h or f. Why is this the case? Because f is deterministic, codes optimized for it cannot express a sample (x, y) such that $y \neq f(x)$, and so a joint distribution μ incurs infinite cost if $\mu(x, y) > 0$. The only option is to choose $\mu(X)$ so that it only has support where $f(x) = h(x)$, which might generate inconsistency by violating D. In other words, the optimal distribution μ^* throws out samples that don't fit, and its conditional marginal $\mu^*(Y|x)$ says nothing unless $h(x)$ is already correct. This reflects the fact that accuracy (the 0-1 loss) gives no gradient information for training h. Note that this is the *exact opposite* of how we captured cross entropy (Proposition 5): there we were unwilling to budge on either the true labels or input distribution, and the optimal distribution tells us how to modify h.

When Y is continuous rather than discrete, the estimator is referred to as a regressor instead of a classifier, and the newfound topology of Y suggests that some mistakes are worse than others—that we might want to treat a small deviation from the correct value of Y. Perhaps the most common way of measuring this deviation is with mean square error, which corresponds to the inconsistency of believing that f and h control the mean of a unit-variance Gaussian.

link to proof

Proposition 7 (square error as inconsistency)**.**

$$\left\langle\!\!\left\langle \begin{tikzcd} & & \boxed{\mu_f} \arrow[dr, two heads, "\mathcal{N}_1"] & \\ D \arrow[r, "{\scriptstyle(\infty)}"'] & \boxed{X} \arrow[ur, two heads, "f"] \arrow[dr, two heads, "h"'] & & \boxed{Y} \\ & & \boxed{\mu_h} \arrow[ur, "\mathcal{N}_1"'] & \end{tikzcd} \right\rangle\!\!\right\rangle \begin{aligned} &= \mathbb{E}_D \big(f(X) - h(X) \big)^2 \\ &= \mathrm{MSE}(f, h) \ , \end{aligned}$$

where $\mathcal{N}_1$ is the unit-variance normal distribution with the given mean.

We treat the more general case, with arbitrary variances and confidences, in the appendix.

4 REGULARIZERS AND PRIORS

Regularizers are extra terms added to loss funtions, which provide a source of inductive bias on model parameters. There is a well-known correspondence between using a regularizer and doing maximum *a posteriori* inference with a prior,[3] in which L2 regularization corresponds to a Gaussian prior (Rennie 2003), while L1 regularization corresponds to a Laplacian prior (Williams 1995). Note that a primary benefit of this correspondence is the ability to make principled modeling choices about regularizers. Our approach provides another justification of it.

link to proof

Proposition 8. *Suppose you have a parameterized model $p(Y|\Theta)$, a prior $q(\Theta)$, and a trusted empirical distribution $D(Y)$. The inconsistency of also believing $\Theta = \theta$ is the cross entropy loss, plus the regularizer* $\log \frac{1}{q(\theta)}$ *times your confidence in q. That is,*

$$\left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[dr, "q", "(\beta)"'] & & \\ & \boxed{\Theta} \arrow[r, "p"] & \boxed{Y} \\ {} \arrow[ur, "\theta"'] & & D \arrow[u, "(\infty)"'] \end{tikzcd} \right\rangle\!\!\right\rangle = \mathop{\mathbb{E}}_{y\sim D} \log \frac{1}{p(y\,|\,\theta)} + \beta \log \frac{1}{q(\theta)} - \mathrm{H}(D) \tag{4}$$

Now, if our prior is the (discretized) unit gaussian $q(\theta) = \frac{1}{k}\exp(-\frac{1}{2}\theta^2)$ for some constant k, then the right hand side of (4) becomes

$$\underbrace{\mathbb{E}_D \log \frac{1}{p(Y\,|\,\theta)}}_{\substack{\text{Cross entropy loss}\\ \text{(data-fit cost of } \theta)}} + \underbrace{\frac{\beta}{2}\theta_0}_{\substack{\ell_2 \text{ regularizer}\\ \text{(complexity cost of } \theta_0)}} \underbrace{+\beta \log k - \mathrm{H}(D)}_{\text{constant in } p \text{ and } \theta},$$

which is the ℓ_2 regularized version of Proposition 3. Moreover, the regularization strength corresponds exactly to the confidence β. What about other priors? It is not difficult to see that if we use a (discretized) centered unit Laplacian prior, $q(\theta) \propto \exp(-|\theta|)$, the second term instead becomes $\beta|\theta_0|$, which is ℓ_1 regularization. More generally, to consider any complexity measure $U(\theta)$, we need only include the Gibbs distribution $\Pr_U(\theta) \propto \exp(-U(\theta))$ into our PDG. We remark that there is nothing special about cross entropy here; any of the objectives we describe can be regularized in this way.

5 STATISTICAL DISTANCES AS INCONSISTENCIES

Suppose you are concerned with a single variable X. One friend has told you that it is distributed according to $p(X)$; another has told you that it follows $q(X)$. You adopt both beliefs. Your mental state will be inconsistent if (and only if) $p \neq q$, with more inconsistency the more p and q differ. Thus the inconsistency of a PDG comprising p and q is a measure of divergence. Recall that a PDG also allows us to specify the confidences β_p and β_q of each cpd, and so we can form the a 'PDG divergence' in this way for every setting of (β_p, β_q). It turns out that a large class of statistical divergences arise in this way. We start with a familiar one.

Proposition 9 (KL Divergence as Inconsistency)**.** *The inconsistency of believing p with complete certainty, and also q with some finite certainty β, is β times the KL Divergence (or relative entropy) of q with respect to p. That is,*

$$\left\langle\!\!\left\langle \xrightarrow[(\infty)]{p} \boxed{X} \xleftarrow[(\beta)]{q} \right\rangle\!\!\right\rangle = \beta\, \boldsymbol{D}(p \parallel q).$$

This result gives us an intuitive interpretation of the asymmetry of relative entropy / KL divergence, and a prescription about when it makes sense to use it. $\boldsymbol{D}(p \parallel q)$ is the inconsistency of a mental state containing both p and q, when absolutely certain of p (and not willing to budge on it). This concords with the standard intuition that $\boldsymbol{D}(p \parallel q)$ reflects the amount of information required to change q into p, which is why we call it the relative entropy "from q to p".

We now consider the general case of a PDG comprising $p(X)$ and $q(X)$ with arbitrary confidences.

link to proof

Lemma 10. *The inconsistency of a PDG containing $p(X)$ with confidence r and $q(X)$ with confidence s is given in closed form by*

$$\left\langle\!\!\left\langle \xrightarrow[(r)]{p} \boxed{X} \xleftarrow[(s)]{q} \right\rangle\!\!\right\rangle = -(r+s)\log \sum_x \Big(p(x)^r q(x)^s \Big)^{\frac{1}{r+s}}.$$

Of the many generalizations of KL divergence, Rényi divergences, first characterized by Alfréd Rényi 1961 are perhaps the most significant, as few others have found either application or an interpretation in terms of coding theory (Van Erven and Harremos 2014). The Rényi divergence of order α between two distributions $p(X)$ and $q(X)$ is given by

$$\boldsymbol{D}_\alpha(p \parallel q) := \frac{1}{1-\alpha} \log \sum_{x \in \mathcal{V}(X)} p(x)^\alpha q(x)^{1-\alpha}. \tag{5}$$

Rényi introduced this measure in the same paper as the more general class of f-divergences, but directs his attention towards those of the form (5), because they satisfy a natural weakening of standard postulates for Shannon entropy due to Fadeev (1957). Concretely, every symmetric, continuous measure that additively separates over independent events, and with a certain "mean-value property", up to scaling, is of the form (5) for some α (Rényi 1961). It follows from Lemma 10 that varying the confidences of our two-distribution

[3] A more detailed account can be found in the appendix

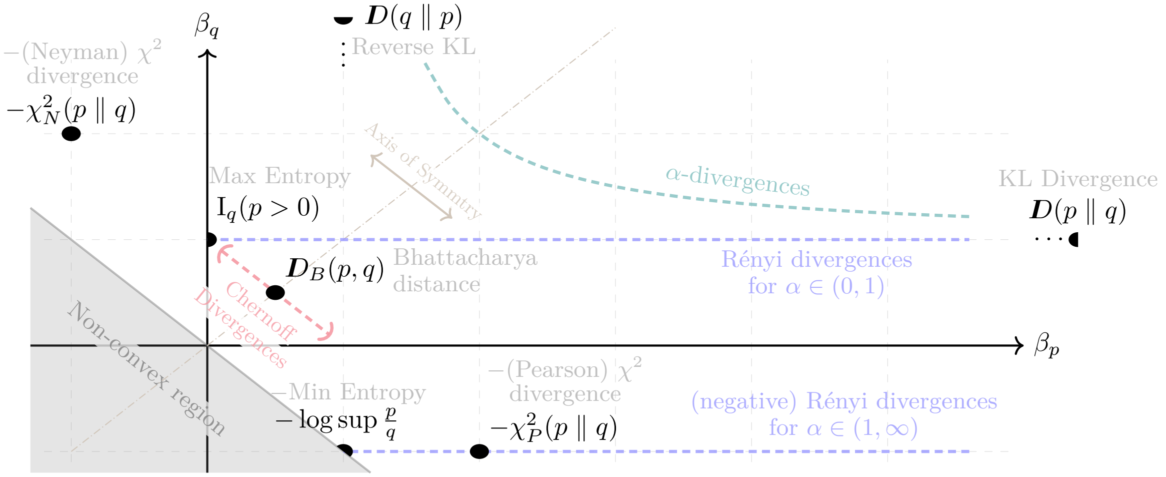


Figure 1: A map of the inconsistency of the PDG comprising $p(X)$ and $q(X)$, as we vary their respective confidences β_p and β_q. Solid circles indicate well-known named measures, semicircles indicate limiting values, and the heavily dashed lines are well-established classes.

PDG carves out essentially the same class: every Rényi divergence is the inconsistency of some PDG of this form, and every PDG divergence is a (scaled) Rényi divergence.

Corollary 10.1 (Rényi Divergences)**.**

$$\left\langle\!\!\left\langle \xrightarrow[(r)]{p} \boxed{X} \xleftarrow[(s)]{q} \right\rangle\!\!\right\rangle = s \cdot \boldsymbol{D}_{\frac{r}{r+s}}(p \parallel q)$$

$$\textit{and} \qquad \boldsymbol{D}_\alpha(p \parallel q) = \left\langle\!\!\left\langle \xrightarrow[(\frac{\alpha}{1-\alpha})]{p} \boxed{X} \xleftarrow{q} \right\rangle\!\!\right\rangle$$

However, the two classes are not identical, because the PDG divergences admit extra limit points. One of the biggest differences is that the reverse KL divergence $\boldsymbol{D}(q \parallel p)$ is not a Renyi entropy of the form $\boldsymbol{D}_\alpha(p \parallel q)$ for any value (or limit) of α. This lack of symmetry has led others (e.g., Cichocki and Amari 2010) to instead work with a re-scaled symmetric version of the Rényi entropy, called α-divergence, which as an additional factor of $\frac{1}{\alpha}$. The relationships between these quantities can be seen in Figure 1.

The Chernoff divergence, measures the tightest possible exponential bound on probability of error (Nielsen 2011) in Bayesian hypothesis testing. It also happens to be the smallest possible inconsistency of simultaneously believing p and q, with combined confidence 1.

Corollary 10.2. *The Chernoff Divergence between p and q equals* $\inf_{\beta \in (0,1)} \left\langle\!\!\left\langle \xrightarrow[(\beta)]{p} \boxed{X} \xleftarrow[(1-\beta)]{q} \right\rangle\!\!\right\rangle$.

One significant consequence of representing divergences as inconsistencies is that we can use Lemma 1 to derive relationships between them. The following facts follow from Figure 1 by inspection.

Corollary 10.3. *1. Rényi entropy is monotonic in its parameter α.*
2. $\boldsymbol{D}(p \parallel q) \geq 2\boldsymbol{D}_B(p, q) \leq \boldsymbol{D}(q \parallel p)$
3. If $q(p > 0) < 1$ (i.e., $q \not\ll p$), then $\boldsymbol{D}(q \parallel p) = \infty$

All of these divergences correspond to PDGs containing only two distributions over one variable. What about other structures? It is not difficult to see that the usual notion of conditional divergences $\boldsymbol{D}(p(Y|X) \| q(Y|X) | r(X))$ falls out of PDGs of the form $\xrightarrow[(\infty)]{r} \boxed{X} \overset{p}{\underset{q}{\rightrightarrows}} \boxed{Y}$, and in general, PDG inconsistency can be viewed as a vast generalization of these divergences to arbitrary structured objects.

Manipulating these structured objects, together with Lemma 10 gives visual proofs of standard properties of divergenes. We give one such proof of the Data Processing Inequality in Figure 2.

6 VARIATIONAL OBJECTIVES AND BOUNDS

6.1 PDGs and Variational Inference

PDG semantics capture interesting features of variational inference and provides a graphical proof language for it. To demonstrate, we begin by recounting the standard development of the 'Evidence Lower BOund' (ELBO), a standard objective for training latent variable models (Blei, Kucukelbir, and McAuliffe 2017, §2.2), and then show that it arises naturally as the inconsistency of the obvious PDG.

Suppose we have a joint distribution $p(X, Z)$, but only have access to observations X. In service of adjust-

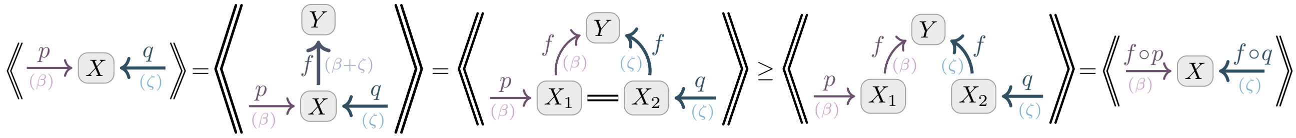


Figure 2: A visual proof of the data-processing inequality: $\boldsymbol{D}^{\text{PDG}}_{(\beta,\zeta)}\left(p \parallel q\right) \geq \boldsymbol{D}^{\text{PDG}}_{(\beta,\zeta)}\left(f \circ p \parallel f \circ q\right)$. In words: the cpd $f(Y|X)$ can always be satisfied, so adds no inconsistency. It is then equivalent to split f and the variable X into X_1 and X_2 with edges enforcing $X_1 = X_2$. But removing such edges can only decrease inconsitency. Finally, compose the remaining cpds to give the result. A full justification can be found in the appendix.

ing $p(X, Z)$ to make our observations more likely, we would like to maximize $\log p(X)$, the "evidence" of X. Unfortunately, computing $p(X) = \sum_z p(X, z)$ requires summing over all of Z, which can be intractable. The variational approach is as follows: fix a family of distributions $\mathcal{Q}$ that is easy to sample from, choose some $q(Z) \in \mathcal{Q}$, and define $\text{ELBO}_{p,q}(x) := \mathbb{E}_{z\sim q} \log \frac{p(x,z)}{q(z)}$. This is something we can estimate, since we can sample from q. By Jensen's inequality,

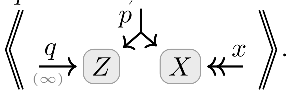
$$\underset{p,q}{\text{ELBO}}(x) = \underset{q}{\mathbb{E}} \log \frac{p(x, Z)}{q(Z)} \leq \log \left[\underset{q}{\mathbb{E}} \frac{p(x, Z)}{q(Z)}\right] = \log p(X),$$

with equality if $q(Z) = p(Z)$. So to maximize $p(X)$, it suffices to adjust p and q to maximize $\text{ELBO}_{p,q}(x)$,[4] provided $\mathcal{Q}$ is expressive enough.

The formula for the ELBO is somewhat difficult to make sense of.[5] Nevertheless, it arises naturally as the inconsistency of the appropriate PDG.

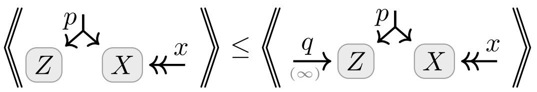
link to proof

Proposition 11. *The negative ELBO of x is the inconsistency of the PDG containing p,q, and $X = x$, with high confidence in q. That is,*

$-\text{ELBO}_{p,q}(x) =$


.

Owing to its structure, a PDG is often more intuitive and easier to work with than the formula for its inconsistency. To illustrate, we now give a simple and visually intuitive proof of the bound traditionally used to motivate the ELBO, via Lemma 1: $\log \frac{1}{p(x)} =$



$= -\text{ELBO}_{p,q}(x)$. The first and last equalities are Propositions 4 and 11 respectively. We now reap some pedagogical benefits. The second PDG has more edges so it is clearly at least as inconsistent. Furthermore, it's easy to see that equality holds when $q(Z) = p(Z)$: the best distribution for the left PDG has marginal $p(Z)$ anyway, so insisting on it incurs no further inconsistency.

6.2 Variational Auto-Encoders and PDGs

An autoencoder is a probabilistic model intended to compress a variable X (e.g., an image) to a compact latent representation Z. Its structure is given by two conditional distributions: an encoder $e(Z|X)$, and a decoder $d(X|Z)$. Of course, not all pairs of cpds fill this role equally well. Perhaps most importantly, we would to have low *reconstruction error* (6)—when we decode an encoded image, we would like it to be reasonably similar to the original.

$$\text{Rec}(x) := \underset{z\sim e(Z|x)}{\mathbb{E}} \underbrace{\text{I}_{d(X|z)}(x)}_{\left(\substack{\text{additional bits required to}\\ \text{decode } x \text{ from its encoding } z}\right)} = \sum_z e(z \mid x) \log \frac{1}{d(x \mid z)} \tag{6}$$

There are other desiderata as well. It would be nice if the distribution on Z had a nice form—perhaps factoring into independent features, which we might use to describe X. We encode this wishful thinking as a belief $p(Z)$, known as a variational prior.

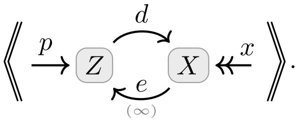
The data of a *variational* auto-encoder (Kingma and Welling 2014) consists of $e(Z|X)$, $d(X|Z)$, and $p(Z)$. The encoder $e(Z|X)$ can be used as a variational approximation of Z, differing from $q(Z)$ of Section 6.1 only in that it can depend on X. Here, the analogue of the ELBO becomes

$$\begin{aligned}\text{ELBO}_{p,e,d}(x) &= \underset{z\sim e|x}{\mathbb{E}} \left[\log \frac{p(z)d(x \mid z)}{e(z \mid x)}\right] \\ &= \boldsymbol{D}(e(Z|x) \parallel p) - \text{Rec}(x).\end{aligned}$$

This gives us the following analog of Proposition 11.

link to proof

Proposition 12. *The VAE objective for a sample x*[†] *is the inconsistency of the PDG containing the encoder e (with high confidence, as it defines the encoding), decoder d prior p, and x. That is,*

$-\text{ELBO}_{p,e,d}(x) =$


.

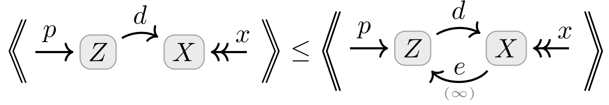
[4] For many iid samples: $\max_{p,q} \sum_{x\in\underline{\mathbf{x}}} \text{ELBO}_{p,q}(x)$.

[5] Especially if p and q are densities. See Appendix A

[†] See appendix for a multi-sample analog.

6.3 Intuitive Proofs of Variational Bounds

Propositions 12 and 15 can be used to derive the variational lower bound. Once again, the addition of the edge e cannot decrease the inconsistency (Lemma 1), but believing it with high confidence does make it possible to generate the encoding for a sample, making inference tractable. This result in the following simple visual proof: $-\log \Pr(x) =$



$= -\text{ELBO}_{p,e,d}(x)$. Here $\Pr(X)$ is the marginal of $p(Z)d(X \mid Z)$ on X. In the appendix, we also give the analogue for many samples, with a single application of the inequality.

6.4 β-VAE Objective

The ELBO is not the only objective that has been used to train networks with the a VAE structure. In the most common variant, Higgins et al. (2016) have argued that the one might want to weight the reconstruction error (6) and KL term differently. They suggest an objective of the form

$$\beta\text{-ELBO}_{p,e,d}(x) := \text{Rec}(x) - \beta \boldsymbol{D}(e(Z|x) \parallel p)$$

which for $\beta = 1$ is equivalent to the ELBO as before. The authors view it as a regularization parameter, annd argue that in some cases, you can do better with a stronger prior. Sure enough, the β-VAE objective is the inconsistency of the same PDG as before, but with confidence β in $p(Z)$.[6]

7 FREE ENERGY AND INCONSISTENCY

The weighted factor graph (WFG) $\Psi = (\phi_j, \theta_j)_{j \in \mathcal{J}}$, where each θ_j is a real weight, j is associated with a subset of variables $\mathbf{X}_j$, and $\phi_j : \mathbf{X}_j \to \mathbb{R}$ determines a distribution by

$$\Pr_\Psi(\mathbf{x}) = \frac{1}{Z_\Psi} \prod_{j \in J} \phi_j(\mathbf{x}_j)^{\theta_j}.$$

$Z_\Psi := \sum_{\mathbf{w}} \prod_{j \in \mathcal{J}} \phi_j(\mathbf{w}_j)^{\theta_j}$, is the constant required to normalize the distribution, and is known as the *partition function*. Computing it is intimately related to many probabilistic inference tasks (Ma et al. 2013).

If every factor is a cpd, and every variable is the target of at least one edge, then Z_Ψ is at most 1, so $-\log Z_\Psi$ is non-negative, and measures how far away the product of factors is from being normalized. Thus, it is in some sense a measure of inconsistency of a factor graph. It turns out that this intuition coincides with our notion of PDG 1-inconsistency.

Proposition 13. *For any weighted factor graph Ψ, we have $\langle\!\langle \mathcal{M}_\Psi \rangle\!\rangle_1 = -\log Z_\Psi$.*

link to proof

Outside of computer science, such factored exponential families form the mathematical backbone of statistical mechanics. In this setting, $-\log Z_\Psi$ is the Heimholtz free energy. The principle of free-energy minimization has been enormously succesful in describing not only the evolution of chemical systems (), but also brains (Friston 2009).

8 FINAL REMARKS

We have now seen that PDG semantics not only capture structured objects such as Bayesian Networks and Factor Graphs as in Richardson and Halpern (2021), but in the same stroke also generate many standard loss functions, including some non-trivial ones. In each case, the appropriate loss is a simple consequence of carefully articulating modeling assumptions. Viewing loss functions in this way also has beneficial side effects, including an intuitive visual proof language for reasoning about the relationships between them.

Taking a step back, we submit that this approach to modeling agents is also more plausible for humans than one that supposes we minimize expectations of some concrete, exogenously given measure of (dis)utility. We also believe that our "universal loss function" which blurs the line between model and objective, and may be of substantial interest to those interested in AI alignment.

Acknowledgements

References

Blei, David M, Alp Kucukelbir, and Jon D McAuliffe (2017). "Variational Inference: A Review for Statisticians." In: *Journal of the American statistical Association* 112.518, pp. 859–877.

Cichocki, Andrzej and Shun-ichi Amari (2010). "Families of Alpha Beta and Gamma Divergences: Flexible and Robust Measures of Similarities." In: *Entropy* 12.6, pp. 1532–1568.

Fadeev, DK (1957). "Zum Begriff der Entropie einer endlichen Wahrscheinlichkeitsschemas." In: *Arbeiten zur Informationstheorie I. Deutscher Verlag der Wissenschaften*, pp. 85–90.

Friston, Karl (2009). "The Free-Energy Principle: a Rough Guide to the Brain?" In: *Trends in Cognitive Sciences* 13.7, pp. 293–301.

[6] The two parameters even share a name, both coming from thermodynamic β (inverse temperature).

Grover, Aditya and Stefano Ermon (2018). *Lecture notes in Deep Generative Models*. deepgenerativemodels.github.io/notes/.

Higgins, Irina et al. (2016). “Beta-VAE: Learning Basic visual concepts with a constrained variational framework.” In:

Jadon, Shruti (2020). “A Survey of Loss Functions for Semantic Segmentation.” In: *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, pp. 1–7.

Kingma, Diederik P and Max Welling (2014). “Auto-Encoding Variational Bayes.” In: *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv: 1312.6114 [stat.ML].

Ma, Jianzhu et al. (2013). “Estimating the Partition Function of Graphical Models using Langevin Importance Sampling.” In: *Artificial Intelligence and Statistics*. PMLR, pp. 433–441.

MacKay, David (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

Myung, In Jae (2003). “Tutorial on Maximum Likelihood Estimation.” In: *Journal of mathematical Psychology* 47.1, pp. 90–100.

Nielsen, Frank (2011). “Chernoff Information of Exponential Families.” In: *arXiv preprint arXiv:1102.2684*.

Rennie, Jason (2003). “On l2-norm regularization and the Gaussian prior.” In:

Rényi, Alfréd (1961). “On Measures of Entropy and Information.” In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. University of California Press, pp. 547–561.

Richardson, Oliver and Joseph Y Halpern (2021). “Probabilistic Dependency Graphs.” In: *AAAI*. arXiv: 2012.10800 [cs.AI].

Tribus, Myron (1961). “Information Theory as the Basis for Thermostatics and Thermodynamics.” In:

Van Erven, Tim and Peter Harremos (2014). “Rényi Divergence and Kullback-Leibler divergence.” In: *IEEE Transactions on Information Theory* 60.7, pp. 3797–3820.

Wang, Qi et al. (2020). “A Comprehensive Survey of Loss Functions in Machine Learning.” In: *Annals of Data Science*, pp. 1–26.

Williams, Peter M (1995). “Bayesian regularization and pruning using a Laplace prior.” In: *Neural computation* 7.1, pp. 117–143.

Name	p	Formula
Harmonic	$(p = -1)$:	$\mathrm{HM}_w(\mathbf{r}) = {}^1\!/\!{}_{(\sum_{i=1}^n {}^{w_i}\!/\!{}_{r_i})}$
Geometric	$(\lim p \to 0)$:	$\mathrm{GM}_w(\mathbf{r}) = \prod_{i=1}^n r_i^{w_i}$
Arithmetic	$(p = 1)$:	$\mathrm{AM}_w(\mathbf{r}) = \sum_{i=1}^n w_i r_i$
Quadratic	$(p = 2)$:	$\mathrm{QM}_w(\mathbf{r}) = \sqrt{\sum_{i=1}^n w_i r_i^2}$

Table 1: special cases of the p-power mean $\mathrm{M}_p^w(\mathbf{r})$

A THE FINE PRINT FOR PROBABILITY DENSITIES

Many of our results (Propositions 2 to 5, 11, 12, 15, 17 and 18) technically require the distribution to be represented with a mass function (not a density function (pdf)). A PDG containg both pdf and a finitely supported distribution on the same variable will typically have infinite inconsistency. But this is not just a quirk of the PDG formalism.

Probability density is not dimensionless (like probability mass), but rather has inverse X-units (e.g., probability per meter), so depends on an arbitrary choice of scale (the pdf for probability per meter and per centimeter will yield different numbers). In places where the objective does not have units that cancel before we take a logarithm, the use of a probability density $p(X)$ becomes sensitive to this arbitrary choice of parameterization. For instance, the analog of surprisal, $-\log p(x)$ for a pdf p, or its expectation, called differential entropy, both suffer from this problem. On the other hand, this choice of scale ultimately amounts to an additive constant.

Moreover, beyond a certain point, decreasing the discretization size k of a discretized approximation $\tilde{p}_k(X)$ *also* contributes a constant that depends only on k. But such constants are irrelevant for optimization, justifying the use of the continuous analogues as loss functions.

The bottom line is that these results for EVERY discretization, but in the limit as the discretization becomes smaller, the quantity in question might diverge to infinity. However, this is just the effect of an additive constant. Using densities in their places results in a morally equivalent loss function.

B MORE DETAILED RESULTS

B.1 Full Characterization of Gaussian Predictors

But before we get there, we first prove a more general result, which is most clearly articulated in terms of a power mean.

Definition 2. The weighted power mean $\mathrm{M}_p^w(\mathbf{r})$ of the collection of real numbers $\mathbf{r} = r_1, \dots, r_n$ with respect to the convex weights $w = w_1, \dots, w_n$ satisfying $\sum_i w_i = 1$, is given by

$$\mathrm{M}_p^w(\mathbf{r}) := \Big(\sum_{i=1}^n w_i (r_i)^p \Big)^{\frac{1}{p}}.$$

We omit the superscript as a shorthand for the uniform weighting $w_i = {}^1\!/\!{}_N$. Most standard means, such as those in Table 1, are special cases. □

It is well known that $\mathrm{M}_p^w(\mathbf{r})$ is increasing in p, and strictly so if not all elements of $\mathbf{r}$ are identical. In particular, $\mathrm{QM}_w(a, b) > \mathrm{GM}_w(a, b)$ for all $a \neq b$ and positive weights w. We now present the result.

[link to proof]

Proposition 14. *Consider a PDG containing two (distinct) conditional Gaussian distributions on a variable*

Y, whose parameters can both depend on a variable X. Its inconsistency takes the form

$$\left\langle\!\!\left\langle \begin{tikzcd} & & \mu_1 \arrow[ddr, "\overset{(\beta:\beta_1)}{\mathcal{N}}"] & \\ & & \sigma_1 \arrow[dr] & \\ {} \arrow[r, "D!"] & X \arrow[uur, two heads, "f"] \arrow[ur, two heads, "t"] \arrow[dr, two heads, "s"] \arrow[ddr, two heads, "h"'] & & Y \\ & & \sigma_2 \arrow[ur] & \\ & & \mu_2 \arrow[uur, "\underset{(\beta:\beta_2)}{\mathcal{N}}"'] & \end{tikzcd} \right\rangle\!\!\right\rangle = \frac{1}{2}\,\mathbb{E}_D\left[\mathrm{HM}(\beta_1,\beta_2)\frac{1}{2}\left(\frac{\mu_1-\mu_2}{\mathrm{QM}_{\hat\beta}(\sigma_1,\sigma_2)}\right)^2 + \mathrm{AM}(\beta_1,\beta_2)\log\frac{\mathrm{QM}_{\hat\beta}(\sigma_1,\sigma_2)}{\mathrm{GM}_{\hat\beta}(\sigma_1,\sigma_2)}\right] \tag{7}$$

$$= \mathop{\mathbb{E}}_{x\sim D}\left[\frac{\beta_1\beta_2}{2}\frac{\Big(f(x)-h(x)\Big)^2}{\beta_2 s(x)^2+\beta_1 t(x)^2} + \frac{\beta_1+\beta_2}{2}\log\frac{\beta_2 s(x)^2+\beta_1 t(x)^2}{(\beta_1+\beta_2)(s(x)^{\beta_2}t(x)^{\beta_1})^{\frac{1}{\beta_1+\beta_2}}}\right]$$

where $\hat\beta = (\frac{\beta_2}{\beta_1+\beta_2}, \frac{\beta_1}{\beta_1+\beta_2})$ represents the normalized and reversed vector of conficences $\beta = (\beta_1, \beta_2)$ for the two distributions, and $\mu_1 = f(X)$, $\mu_2 = g(X)$, $\sigma_1 = s(X)$, $\sigma_2 = t(X)$ are random variables over X.

Plugging in $s(x) = t(x) = 1$ and $\beta_1 = \beta_2 = 1$ proves:

Proposition 7.

$$\left\langle\!\!\left\langle \begin{tikzcd} & & \mu_f \arrow[dr, "\mathcal{N}_1"] & \\ {} \arrow[r, "D", "(\infty)"'] & X \arrow[ur, two heads, "f"] \arrow[dr, two heads, "h"'] & & Y \\ & & \mu_h \arrow[ur, "\mathcal{N}_1"'] & \end{tikzcd} \right\rangle\!\!\right\rangle = \mathbb{E}_D\big(f(X)-h(X)\big)^2 \\ = \mathrm{MSE}(f,h)\ ,$$

where $\mathcal{N}_1$ is the unit-variance normal distribution with the given mean.

This PDG is also equal to

$$\left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "D!"] & X \arrow[r, bend left, "\mathcal{N}(f(x),1)"] \arrow[r, bend right, "\mathcal{N}(g(x),1)"'] & Y \end{tikzcd} \right\rangle\!\!\right\rangle$$

which illustrates an orthogonal point: that PDGs handle composition of functions as one would expect, so that it is equivalent to model an entire process as a single arrow, or to break it into stages, ascribing an arrow to each stage, with one step of randomization.

As a bonus, Proposition 14 also gives a proof of the inequality of the weighted geometric and quadratic means.

Corollary 14.1. *For all $\beta, \sigma_1, \sigma_2$, and D, we have:*

$$\mathop{\mathbb{E}}_{D}\frac{\mathrm{QM}_{\hat\beta}(\sigma_1,\sigma_2)}{\mathrm{GM}_{\hat\beta}(\sigma_1,\sigma_2)} > 0.$$

B.2 Full-Dataset ELBO and Bounds

We now present the promised multi-sample analogs from Section 6.1.

Proposition 15. *The following analog of Proposition 12 for a whole dataset $\underline{\mathbf{x}}$ holds:*

link to proof

$$-\mathop{\mathbb{E}}_{\Pr_{\underline{\mathbf{x}}}}\mathrm{ELBO}_{p,e,d}(X) = \left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "p"] & Z \arrow[r, bend left, "d"] & X \arrow[l, bend left, "e!"] & {} \arrow[l, "\Pr_{\underline{\mathbf{x}}}!"'] \end{tikzcd} \right\rangle\!\!\right\rangle + \mathrm{H}(\Pr_{\underline{\mathbf{x}}}).$$

Propositions 3 and 15 then give us an analog of the visual bounds in the body of the main paper Section 6.3 for many i.i.d. datapoints at once, with only a single application of the inequality:

$$-\log \Pr(\underline{\mathbf{x}}) = -\log \prod_{i=1}^{m} \Big(\Pr(x^{(i)})\Big) = -\frac{1}{m}\sum_{i=1}^{m} \log \Pr(x^{(i)}) =$$

$$\mathrm{H}(\Pr_{\underline{\mathbf{x}}}) + \left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "p"] & \boxed{Z} \arrow[r, bend left, "d"] & \boxed{X} & {} \arrow[l, two heads, "\Pr_{\underline{\mathbf{x}}}!"'] \end{tikzcd} \right\rangle\!\!\right\rangle \leq \left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "p"] & \boxed{Z} \arrow[r, bend left, "d"] & \boxed{X} \arrow[l, bend left, "e!"] & {} \arrow[l, two heads, "\Pr_{\underline{\mathbf{x}}}!"'] \end{tikzcd} \right\rangle\!\!\right\rangle + \mathrm{H}(\Pr_{\underline{\mathbf{x}}})$$

$$= -\mathop{\mathbb{E}}_{\Pr_{\underline{\mathbf{x}}}} \ \mathrm{ELBO}_{p,e,d}(X)$$

We also have the following formal statement of the claim made in **??**.

link to proof

Proposition 16. *The negative β-ELBO objective for a prior $p(X)$, encoder $e(Z \mid X)$, decoder $d(X \mid Z)$, at a sample x, is equal to the inconsistency of the corresponding PDG, where the prior has confidence equal to β. That is,*

$$-\beta\text{-ELBO}_{p,e,d}(x) = \left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "\overset{(\beta)}{p}"] & \boxed{Z} \arrow[r, bend left, "d"] & \boxed{X} \arrow[l, bend left, "e!"] & {} \arrow[l, two heads, "x"'] \end{tikzcd} \right\rangle\!\!\right\rangle$$

As a specific case (i.e., effectively by setting $\beta_p := 0$), we get the reconstruction error as the inconsistency of an autoencoder (without a variational prior):

Corollary 16.1 (reconstruction error as inconsistency)**.**

$$-\mathrm{Rec}_{ed,d}(x) := \mathop{\mathbb{E}}_{z\sim e(Z|x)} \mathrm{I}_{d(X|z)}(x) = \left\langle\!\!\left\langle \begin{tikzcd} \boxed{Z} \arrow[r, bend left, "d"] & \boxed{X} \arrow[l, bend left, "e!"] & {} \arrow[l, two heads, "x"'] \end{tikzcd} \right\rangle\!\!\right\rangle$$

B.3 More Variants of Cross Entropy Results

link to proof

Proposition 17. *Given a model determining a probability distribution with mass function $p(X)$, and samples $\underline{\mathbf{x}} = \{x_i\}_{i=1}^{m}$ determining an empirical distribution $\Pr_{\underline{\mathbf{x}}}$, the following are equal, for all $\gamma \geq 0$:*

1. *The average negative log likelihood $\ell(p; \underline{\mathbf{x}}) = -\frac{1}{m}\sum_{i=1}^{m} \log p(x_i)$*
2. *The cross entropy of p relative to $\Pr_{\underline{\mathbf{x}}}$*
3. $[\![p]\!]_\gamma(\Pr_{\underline{\mathbf{x}}})$ $\quad +(1+\gamma)\,\mathrm{H}(\Pr_{\underline{\mathbf{x}}})$
4. $\left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "p"] & \boxed{X} & {} \arrow[l, "\Pr_{\underline{\mathbf{x}}}!"'] \end{tikzcd} \right\rangle\!\!\right\rangle_\gamma$ $\quad +(1+\gamma)\,\mathrm{H}(\Pr_{\underline{\mathbf{x}}})$

link to proof

Proposition 18. *The average negative log likelihood $\ell(p;x) := -\frac{1}{|\underline{\mathbf{x}}|}\sum_{x\in\underline{\mathbf{x}}} \log p(x)$ (which is also the cross entropy) is the inconsistency of the PDG containing p and the data distribution $\Pr_{\underline{\mathbf{x}}}$, plus the entropy of the data distribution (which is constant in p). That is,*

$$\ell(p;\underline{\mathbf{x}}) = \left\langle\!\!\left\langle \begin{tikzcd} & {} \arrow[dl] \arrow[dr, "p"] & & \\ \boxed{Z} & & \boxed{X} & {} \arrow[l, "\Pr_{\underline{\mathbf{x}}}!"'] \end{tikzcd} \right\rangle\!\!\right\rangle + \mathrm{H}(\Pr_{\underline{\mathbf{x}}}).$$

C PROOFS

Lemma 1. *Suppose PDGs $\mathcal{M}$ and $\mathcal{M}'$ differ only in their edges (resp. $\mathcal{E}$ and $\mathcal{E}'$) and confidences (resp. $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$). If $\mathcal{E} \subseteq \mathcal{E}'$ and $\beta_L \leq \beta'_L$ for all $L \in \mathcal{E}$, then $\langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma \leq \langle\!\langle \mathcal{M}' \rangle\!\rangle_\gamma$ for all γ.*[7]

[7] All proofs can be found in Appendix C.

Proof. For every μ, adding more edges only adds non-negative terms to (1). Thus, $\llbracket m + m' \rrbracket_\gamma(\mu) \geq \llbracket m \rrbracket_\gamma(\mu)$ for all γ and μ. So it also holds when we take an infemum over μ, yielding $\langle\!\langle m + m' \rangle\!\rangle_\gamma \geq \langle\!\langle m \rangle\!\rangle_\gamma$. Analogously, increasing β results in larger coefficients on the (non-negative) terms of (1) so $\llbracket m \rrbracket_\gamma(\mu) \geq \llbracket m' \rrbracket_\gamma(\mu)$ for all γ and μ, so $\langle\!\langle m \rangle\!\rangle \geq \langle\!\langle m' \rangle\!\rangle$. □

Proposition 2. *Consider a distribution $p(X)$. The surprisal $\mathrm{I}_p(x)$ of the event $X{=}x$ equals inconsistency of the PDG containing p and $X{=}x$. That is,*

$$\mathrm{I}_p(x) := \log \frac{1}{p(X{=}x)} = \left\langle\!\!\left\langle \xrightarrow{p} \boxed{X} \overset{x}{\twoheadleftarrow} \right\rangle\!\!\right\rangle.$$

Proof. Any distribution $\mu(X)$ that places mass on some $x' \neq x$ will have infinite KL divergence from the point mass on x. Thus, the only possibility for a finite consistency arises when $\mu = \delta_x$, and so

$$\left\langle\!\!\left\langle \xrightarrow{p} \boxed{X} \overset{x}{\twoheadleftarrow} \right\rangle\!\!\right\rangle = \left\llbracket \xrightarrow{p} \boxed{X} \overset{x}{\twoheadleftarrow} \right\rrbracket(\delta_x) = \boldsymbol{D}(\delta_x \parallel p) = \log \frac{1}{p(x)} = \mathrm{I}_p(x).$$

□

Proposition 17. *Given a model determining a probability distribution with mass function $p(X)$, and samples $\underline{\mathbf{x}} = \{x_i\}_{i=1}^m$ determining an empirical distribution $\mathrm{Pr}_{\underline{\mathbf{x}}}$, the following are equal, for all $\gamma \geq 0$:*

1. *The average negative log likelihood $\ell(p; \underline{\mathbf{x}}) = -\frac{1}{m}\sum_{i=1}^m \log p(x_i)$*
2. *The cross entropy of p relative to $\mathrm{Pr}_{\underline{\mathbf{x}}}$*
3. $\llbracket p \rrbracket_\gamma(\mathrm{Pr}_{\underline{\mathbf{x}}}) \quad + (1+\gamma)\,\mathrm{H}(\mathrm{Pr}_{\underline{\mathbf{x}}})$
4. $\left\langle\!\!\left\langle \xrightarrow{p} \boxed{X} \xleftarrow{\mathrm{Pr}_{\underline{\mathbf{x}}}!} \right\rangle\!\!\right\rangle_\gamma \quad + (1+\gamma)\,\mathrm{H}(\mathrm{Pr}_{\underline{\mathbf{x}}})$

Proof. The equality of 1 and 2 is standard. The equality of 3 and 4 can be seen by the fact that in the limit of infinite confidence on $\mathrm{Pr}_{\underline{\mathbf{x}}}$, the optimal distribution must also equal $\mathrm{Pr}_{\underline{\mathbf{x}}}$, so the least inconsistency is attained at this value. Finally it remains to show that the first two and last two are equal:

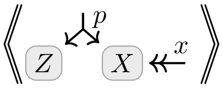
$$\begin{aligned} \llbracket p \rrbracket_\gamma(\mathrm{Pr}_{\underline{\mathbf{x}}}) + (1+\gamma)\,\mathrm{H}(\mathrm{Pr}_{\underline{\mathbf{x}}}) &= \boldsymbol{D}(\mathrm{Pr}_{\underline{\mathbf{x}}} \parallel p) - \gamma\,\mathrm{H}(\mathrm{Pr}_{\underline{\mathbf{x}}}) + (1+\gamma)\,\mathrm{H}(\mathrm{Pr}_{\underline{\mathbf{x}}}) \\ &= \boldsymbol{D}(\mathrm{Pr}_{\underline{\mathbf{x}}} \parallel p) + \mathrm{H}(\mathrm{Pr}_{\underline{\mathbf{x}}}) \\ &= \mathbb{E}_{\mathrm{Pr}_{\underline{\mathbf{x}}}}\left[\log \frac{\mathrm{Pr}_{\underline{\mathbf{x}}}}{p} + \log \frac{1}{\mathrm{Pr}_{\underline{\mathbf{x}}}}\right] = \mathbb{E}_{\mathrm{Pr}_{\underline{\mathbf{x}}}}\left[\log \frac{1}{p}\right], \end{aligned}$$

which is the cross entropy, as desired. □

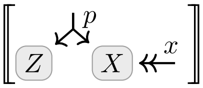
Proposition 4. *If $p(X, Z)$ is a joint distribution, the information content of the partial observation $X = x$, or the marginal negative log likelihood of x, is given by*

$$\mathrm{I}_p(x) = \log \frac{1}{p(x)} = \left\langle\!\!\left\langle \boxed{Z} \leftarrow \overset{p}{\curlyvee} \rightarrow \boxed{X} \overset{x}{\twoheadleftarrow} \right\rangle\!\!\right\rangle. \tag{2}$$

Proof. As before, all mass of μ must be on x for it to have a finite score. Thus it suffices to consider joint distributions of the form $\mu(X, Z) = \delta_x(X)\mu(Z)$. We have



$$= \inf_{\mu(Z)}$$



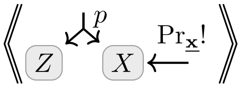
$$\Big(\delta_x(X)\mu(Z)\Big)$$

$$\begin{aligned}
&= \inf_{\mu(Z)} \boldsymbol{D}\Big(\delta_x(X)\mu(Z) \,\Big\|\, p(X, Z)\Big) \\
&= \inf_{\mu(Z)} \mathop{\mathbb{E}}_{z\sim\mu} \log \frac{\mu(z)}{p(x, z)} \quad = \inf_{\mu(Z)} \mathop{\mathbb{E}}_{z\sim\mu} \log \frac{\mu(z)}{p(x, z)} \frac{p(x)}{p(x)} \\
&= \inf_{\mu(Z)} \mathop{\mathbb{E}}_{z\sim\mu} \left[\log \frac{\mu(z)}{p(z \mid x)} + \log \frac{1}{p(x)}\right] \\
&= \inf_{\mu(Z)} \Big[\boldsymbol{D}(\mu(Z) \parallel p(Z \mid x))\Big] + \log \frac{1}{p(x)} \\
&= \log \frac{1}{p(x)} = \mathrm{I}_p(x) && \text{[Gibbs Inequality]}
\end{aligned}$$

□

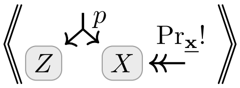
Proposition 18. *The average negative log likelihood* $\ell(p; x) := -\frac{1}{|\underline{\mathbf{x}}|} \sum_{x\in\underline{\mathbf{x}}} \log p(x)$ *(which is also the cross entropy) is the inconsistency of the PDG containing* p *and the data distribution* $\mathrm{Pr}_{\underline{\mathbf{x}}}$*, plus the entropy of the data distribution (which is constant in* p*). That is,*

$$\ell(p; \underline{\mathbf{x}}) =$$

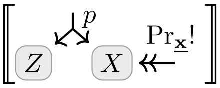


$$+ \mathrm{H}(\mathrm{Pr}_{\underline{\mathbf{x}}}).$$

Proof. The same idea as in Proposition 4, but a little more complicated.



$$= \inf_{\mu(Z|X)}$$



$$\Big(\mathrm{Pr}_{\underline{\mathbf{x}}}(X)\mu(Z \mid X)\Big)$$

$$\begin{aligned}
&= \inf_{\mu(Z|X)} \boldsymbol{D}\Big(\mathrm{Pr}_{\underline{\mathbf{x}}}(X)\mu(Z \mid X) \,\Big\|\, p(X, Z)\Big) \\
&= \inf_{\mu(Z|X)} \mathop{\mathbb{E}}_{\substack{x\sim\mathrm{Pr}_{\underline{\mathbf{x}}} \\ z\sim\mu}} \log \frac{\mu(z \mid x)\,\mathrm{Pr}_{\underline{\mathbf{x}}}(x)}{p(x, z)} \\
&= \frac{1}{|\underline{\mathbf{x}}|} \inf_{\mu(Z|X)} \sum_{x\in\underline{\mathbf{x}}} \mathop{\mathbb{E}}_{z\sim\mu(Z|x)} \log \frac{\mu(z \mid x)\,\mathrm{Pr}_{\underline{\mathbf{x}}}(x)}{p(x, z)} \frac{p(x)}{p(x)} \\
&= \frac{1}{|\underline{\mathbf{x}}|} \inf_{\mu(Z|X)} \sum_{x\in\underline{\mathbf{x}}} \mathop{\mathbb{E}}_{z\sim\mu} \left[\log \frac{\mu(z)}{p(z \mid x)} + \log \frac{1}{p(x)} - \log \frac{1}{\mathrm{Pr}_{\underline{\mathbf{x}}}(x)}\right] \\
&= \frac{1}{|\underline{\mathbf{x}}|} \sum_{x\in\underline{\mathbf{x}}} \left[\inf_{\mu(Z)} \Big[\boldsymbol{D}(\mu(Z) \parallel p(Z \mid x))\Big] + \log \frac{1}{p(x)}\right] - \mathrm{H}(\mathrm{Pr}_{\underline{\mathbf{x}}}) \\
&= \frac{1}{|\underline{\mathbf{x}}|} \sum_{x\in\underline{\mathbf{x}}} \log \frac{1}{p(x)} - \mathrm{H}(\mathrm{Pr}_{\underline{\mathbf{x}}}) = \frac{1}{|\underline{\mathbf{x}}|} \sum_{x\in\underline{\mathbf{x}}} I_p(x) - \mathrm{H}(\mathrm{Pr}_{\underline{\mathbf{x}}}) \\
\Big(\quad &= \boldsymbol{D}(\mathrm{Pr}_{\underline{\mathbf{x}}} \parallel p(X)) \quad\Big)
\end{aligned}$$

□

Proposition 5. *Consider a probabilistic predictor $f(Y \mid X)$. The inconsistency of the PDG containing f and the empirical distribution $\Pr_{\underline{\mathbf{xy}}}$ of a sample set $\underline{\mathbf{xy}} = \{(x_i, y_i)\}_{i=1}^m$ is equal to the cross-entropy loss (minus the empirical uncertainty in Y given X, a constant that depends only on the data). That is,*

$$\left\langle\!\!\left\langle \begin{tikzcd} & \Pr_{\underline{\mathbf{xy}}} \arrow[dl] \arrow[dr, "(\infty)"] & \\ X \arrow[rr, "f"'] & & Y \end{tikzcd} \right\rangle\!\!\right\rangle = \frac{1}{m}\sum_{i=1}^{m} \log \frac{1}{f(y_i \,|\, x_i)} - \mathrm{H}_{\Pr_{\underline{\mathbf{xy}}}}(Y|X).$$

Proof. $\Pr_{\underline{\mathbf{xy}}}$ has high confidence, it is the only joint distribution μ with finite score. Since f is the only other edge, the inconsistency is therefore

$$\begin{aligned} \mathop{\mathbb{E}}_{x\sim\Pr_{\underline{\mathbf{xy}}}} \boldsymbol{D}\Big(\Pr_{\underline{\mathbf{xy}}}(Y \mid x) \,\Big\|\, f(Y \mid x)\Big) &= \mathop{\mathbb{E}}_{x,y\sim\Pr_{\underline{\mathbf{xy}}}} \left[\log \frac{\Pr_{\underline{\mathbf{xy}}}(y \mid x)}{f(y \mid x)}\right] \\ &= \mathop{\mathbb{E}}_{x,y\sim\Pr_{\underline{\mathbf{xy}}}} \left[\log \frac{1}{f(y \mid x)} - \log \frac{1}{\Pr_{\underline{\mathbf{xy}}}(y \mid x)}\right] \\ &= \frac{1}{|\underline{\mathbf{xy}}|} \sum_{(x,y)\in\underline{\mathbf{xy}}} \left[\log \frac{1}{f(y \mid x)}\right] \quad - \mathrm{H}_{\Pr_{\underline{\mathbf{xy}}}}(Y \mid X) \end{aligned}$$

□

Proposition 6. *Consider a distribution $D(X)$ over inputs X, and functions $f, h : X \to Y$, where h is a predictor and f generates the true labels. The inconsistency of believing all three is the log accuracy of h. That is,*

$$\left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "D", "(\beta)"'] & X \arrow[r, bend left, "h"] \arrow[r, bend right, "f"'] & Y \end{tikzcd} \right\rangle\!\!\right\rangle = -\beta \log \Pr_{x\sim D}(f(x) = h(x)) = \beta\, \mathrm{I}_D[f = h]. \tag{3}$$

Proof. Becuase f is deterministic, for every x in the support of a joint distribution μ with finite score, we must have $\mu(Y \mid x) = \delta_f$, since if μ were to place any non-zero mass $\mu(x, y) = \epsilon > 0$ on a pont (x, y) with $y \neq f(x)$ results in an infinite contribution to the KL divergence

$$\boldsymbol{D}(\mu(Y \mid x) \,\|\, \delta_{f(x)}) = \mathop{\mathbb{E}}_{x,y\sim\mu} \log \frac{\mu(y \mid x)}{\delta_{f(x)}} \geq \mu(y, x) \log \frac{\mu(x, y)}{\mu(x) \cdot \delta_{f(x)}(y)} = \epsilon \log \frac{\epsilon}{0} = \infty.$$

The same holds for h. Therefore, for any μ with a finite score, and x with $\mu(x) > 0$, we have $\delta_{f(x)} = \mu(Y \mid x) = \delta_{h(x)}$, meaning that we need only consider μ whose support is a subset of those points on which f and h agree. On all such points, the contribution to the score from the edges associated to f and h will be zero, since μ matches the conditional marginals exactly, and the total incompatibility of such a distribution μ is equal to the relative entropy $\boldsymbol{D}(\mu \,\|\, D)$, scaled by the confidence β of the empirical distribution D.

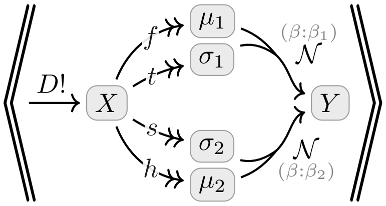
So, among those distributions $\mu(X)$ supported on an event $E \subset \mathcal{V}(X)$, which minimizes is the relative entropy of $\boldsymbol{D}(\mu \,\|\, D)$? It is well known that the conditional distribution $D \mid E \propto \delta_E(X) D(X) = \frac{1}{D(E)}\delta_E(X) D(X)$ satisfies this property uniquely (see, for instance, **halpernRAU**). Let $f = h$ denote the event that f and h agree. Then we calculate

$$\left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "\overset{(\beta)}{D}"] & \boxed{X} \arrow[r, two heads, bend left, "h"] \arrow[r, two heads, bend right, "f"'] & \boxed{Y} \end{tikzcd} \right\rangle\!\!\right\rangle = \inf_{\substack{\mu(X) \text{ s.t.} \\ \text{supp}(\mu) \subseteq [f=h]}} \beta \boldsymbol{D}\Big(\mu(X) \,\Big\|\, D(X)\Big)$$

$$\begin{aligned}
&= \beta \boldsymbol{D}\Big(D \mid [f=h] \,\Big\|\, D\Big) \\
&= \beta \mathop{\mathbb{E}}_{D|f=h} \log \frac{\delta_{f=h}(X) D(X)}{D(f=h) \cdot D(X)} \\
&= \beta \mathop{\mathbb{E}}_{D|f=h} \log \frac{1}{D(f=h)} && \left[\begin{array}{c}\text{since } \delta_{f=h}(x) = 1 \text{ for all } x \text{ that} \\ \text{contribute to the expectation}\end{array}\right] \\
&= -\beta \log D(f=h) && \left[\ \text{since } D(f=h) \text{ is a constant}\ \right] \\
&= -\beta \log \Big(\text{accuracy}_{f,D}(h)\Big) \\
&= \beta \, \mathrm{I}_D[f=h].
\end{aligned}$$

□

Proposition 14. *Consider a PDG containing two (distinct) conditional Gaussian distributions on a variable Y, whose parameters can both depend on a variable X. Its inconsistency takes the form*



$$= \frac{1}{2} \mathbb{E}_D \left[\mathrm{HM}(\beta_1, \beta_2) \frac{1}{2} \left(\frac{\mu_1 - \mu_2}{\mathrm{QM}_{\hat\beta}(\sigma_1, \sigma_2)} \right)^2 + \mathrm{AM}(\beta_1, \beta_2) \log \frac{\mathrm{QM}_{\hat\beta}(\sigma_1, \sigma_2)}{\mathrm{GM}_{\hat\beta}(\sigma_1, \sigma_2)} \right] \tag{7}$$

$$= \mathop{\mathbb{E}}_{x \sim D} \left[\frac{\beta_1 \beta_2}{2} \frac{\Big(f(x) - h(x)\Big)^2}{\beta_2 s(x)^2 + \beta_1 t(x)^2} + \frac{\beta_1 + \beta_2}{2} \log \frac{\beta_2 s(x)^2 + \beta_1 t(x)^2}{(\beta_1 + \beta_2)(s(x)^{\beta_2} t(x)^{\beta_1})^{\frac{1}{\beta_1 + \beta_2}}} \right]$$

where $\hat\beta = (\frac{\beta_2}{\beta_1 + \beta_2}, \frac{\beta_1}{\beta_1 + \beta_2})$ represents the normalized and reversed vector of conficences $\beta = (\beta_1, \beta_2)$ for the two distributions, and $\mu_1 = f(X)$, $\mu_2 = g(X)$, $\sigma_1 = s(X)$, $\sigma_2 = t(X)$ are random variables over X.

Proof. Let $\mathcal{m}$ denote the PDG in question. Since D has high confidence, we know any joint distribution μ with a finite score must have $\mu(X) = D(X)$. Thus,

$$\begin{aligned}
\langle\!\langle \mathcal{m} \rangle\!\rangle_0 &= \inf_\mu \mathop{\mathbb{E}}_{x \sim D} \mathop{\mathbb{E}}_{y \sim \mu|x} \left[\beta_1 \log \frac{\mu(y \mid x)}{\mathcal{N}(y \mid f(x), t(x))} + \beta_2 \log \frac{\mu(y \mid x)}{\mathcal{N}(y \mid h(x), s(x))} \right] \\
&= \inf_\mu \mathop{\mathbb{E}}_{x \sim D} \mathop{\mathbb{E}}_{y \sim \mu|x} \left[\beta_1 \log \frac{\mu(y \mid x)}{\frac{1}{t(x)\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{y - f(x)}{t(x)} \right)^2 \right)} + \beta_2 \log \frac{\mu(y \mid x)}{\frac{1}{s(x)\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{y - h(x)}{s(x)} \right)^2 \right)} \right] \\
&= \inf_\mu \mathop{\mathbb{E}}_{x \sim D} \mathop{\mathbb{E}}_{y \sim \mu|x} \left[\log \mu(y \mid x)^{\beta_1 + \beta_2} \begin{array}{l} + \beta_1 \log(t(x)\sqrt{2\pi}) + \frac{\beta_1}{2} \left(\frac{y - f(x)}{t(x)} \right)^2 \\ + \beta_2 \log(t(x)\sqrt{2\pi}) + \frac{\beta_2}{2} \left(\frac{y - h(x)}{s(x)} \right)^2 \end{array} \right]
\end{aligned}$$

⟨ TODO: finish proof. ⟩

□

Lemma 10. *The inconsistency of a PDG containing $p(X)$ with confidence r and $q(X)$ with confidence s is given in closed form by*

$$\left\langle\!\!\left\langle \xrightarrow[(r)]{p} \boxed{X} \xleftarrow[(s)]{q} \right\rangle\!\!\right\rangle = -(r+s) \log \sum_x \Big(p(x)^r q(x)^s \Big)^{\frac{1}{r+s}} .$$

Proof.

$$\begin{aligned}
\left\langle\!\!\left\langle \xrightarrow[(\beta:r)]{p} \boxed{X} \xleftarrow[(\beta:s)]{q} \right\rangle\!\!\right\rangle &= \inf_\mu \mathop{\mathbb{E}}_\mu \log \frac{\mu(x)^{r+s}}{p(x)^r q(x)^s} \\
&= (r+s) \inf_\mu \mathop{\mathbb{E}}_\mu \left[\log \frac{\mu(x)}{(p(x)^r q(x)^s)^{\frac{1}{r+s}}} \cdot \frac{Z}{Z} \right] \\
&= \inf_\mu (r+s) \boldsymbol{D}\left(\mu \;\middle\|\; \frac{1}{Z} p^{\frac{r}{r+s}} q^{\frac{s}{r+s}} \right) - (r+s) \log Z
\end{aligned}$$

where $Z := (\sum_x p(x)^r q(x)^s)^{\frac{1}{r+s}}$ is the constant required to normalize the denominator as a distribution. Since this is now a relative entropy, it achives its minimum when μ is the other distribution, at which point it contributes zero, so our formula becomes

$$\begin{aligned}
&= -(r+s) \log Z \\
&= -(r+s) \log \sum_x \Big(p(x)^r q(x)^s \Big)^{\frac{1}{r+s}} \qquad \text{as promised.}
\end{aligned}$$

□

Proposition 8. *Suppose you have a parameterized model $p(Y|\Theta)$, a prior $q(\Theta)$, and a trusted empirical distribution $D(Y)$. The inconsistency of also believing $\Theta = \theta$ is the cross entropy loss, plus the regularizer:* $\log \frac{1}{q(\theta)}$ *times your confidence in q. That is,*

$$\left\langle\!\!\left\langle \begin{tikzcd} \arrow[dr, "q", "(\beta)"'] & & \\ & \boxed{\Theta} \arrow[r, "p"] & \boxed{Y} \\ \arrow[ur, two heads, "\theta"'] & & D \arrow[u, "(\infty)"'] \end{tikzcd} \right\rangle\!\!\right\rangle = \mathop{\mathbb{E}}_{y \sim D} \log \frac{1}{p(y \,|\, \theta)} + \beta \log \frac{1}{q(\theta)} \\ - \mathrm{H}(D) \tag{4}$$

Proof. $\langle \quad \rangle$

□

Proposition 11. *The negative ELBO of x is the inconsistency of the PDG containing p,q, and X=x, with high confidence in q. That is,*

$$-\mathrm{ELBO}_{p,q}(x) = \left\langle\!\!\left\langle \begin{tikzcd} & p \arrow[dl] \arrow[dr] & \\ \xrightarrow[(\infty)]{q} \boxed{Z} & & \boxed{X} \arrow[r, two heads, leftarrow, "x"] & {} \end{tikzcd} \right\rangle\!\!\right\rangle.$$

Proof. Every distribution that does marginalize to $q(Z)$ or places any mass on $x' \neq x$ will have infinite score. Thus the only distribution that could have a finite score is $\mu(X, Z)$. Thus,

$$\left\langle\!\!\left\langle \begin{tikzcd}[row sep=small] & & {} \arrow[dl] \arrow[dr, "p"] & & \\ {} \arrow[r, "q!"] & \boxed{Z} & & \boxed{X} & {} \arrow[l, two heads, "x"'] \end{tikzcd} \right\rangle\!\!\right\rangle = \inf_{\mu} \left\llbracket \begin{tikzcd}[row sep=small] & & {} \arrow[dl] \arrow[dr, "p"] & & \\ {} \arrow[r, "q!"] & \boxed{Z} & & \boxed{X} & {} \arrow[l, two heads, "x"'] \end{tikzcd} \right\rrbracket (\mu)$$

$$= \left\llbracket \begin{tikzcd}[row sep=small] & & {} \arrow[dl] \arrow[dr, "p"] & & \\ {} \arrow[r, "q!"] & \boxed{Z} & & \boxed{X} & {} \arrow[l, two heads, "x"'] \end{tikzcd} \right\rrbracket (\delta_x(X)q(Z))$$

$$= \mathop{\mathbb{E}}_{\substack{x'\sim\delta_x \\ z\sim q}} \log \frac{\delta_x(x')q(z)}{p(x',z)} = - \mathop{\mathbb{E}}_{z\sim q} \frac{p(x,z)}{q(z)} = -\mathrm{ELBO}_{p,q}(x).$$

□

We proove both Proposition 12 and Proposition 15 at the same time.

Proposition 12. *The VAE objective for a sample* x[†] *is the inconsistency of the PDG containing the encoder* e *(with high confidence, as it defines the encoding), decoder* d *prior* p*, and* x*. That is,*

$$-\mathrm{ELBO}_{p,e,d}(x) = \left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "p"] & \boxed{Z} \arrow[r, bend left, "d"] & \boxed{X} \arrow[l, bend left, "e", "(\infty)"'] & {} \arrow[l, two heads, "x"'] \end{tikzcd} \right\rangle\!\!\right\rangle.$$

Proposition 15. *The following analog of Proposition 12 for a whole dataset* $\underline{\mathbf{x}}$ *holds:*

$$- \mathop{\mathbb{E}}_{\Pr_{\underline{\mathbf{x}}}} \mathrm{ELBO}_{p,e,d}(X) = \left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "p"] & \boxed{Z} \arrow[r, bend left, "d"] & \boxed{X} \arrow[l, bend left, "e!"] & {} \arrow[l, "\Pr_{\underline{\mathbf{x}}}!"'] \end{tikzcd} \right\rangle\!\!\right\rangle + \mathrm{H}(\Pr_{\underline{\mathbf{x}}}).$$

Proof. The two proofs are similar. For Proposition 12, the optimal distribution must be $\delta_x(X)e(Z \mid X)$, and for Proposition 15, it must be $\Pr_{\underline{\mathbf{x}}}(X)e(Z \mid X)$, because e and the data both have infinite confidence, so any other distribution gets an infinite score. At the same time, d and p define a joint distribution, so the inconsistency in question becomes

$$\boldsymbol{D}\Big(\delta_x(X)e(Z \mid X) \,\Big\|\, p(Z)d(X \mid Z)\Big) = \mathop{\mathbb{E}}_{z\sim e|x} \left[\log \frac{p(z)d(x \mid z)}{e(z \mid x)}\right] = \mathrm{ELBO}_{p,e,d}(x)$$

in the first, case, and

$$\boldsymbol{D}\Big(\Pr_{\underline{\mathbf{x}}}(X)e(Z \mid X) \,\Big\|\, p(Z)d(X \mid Z)\Big) = \frac{1}{|\underline{\mathbf{x}}|} \sum_{x\in\underline{\mathbf{x}}} \mathop{\mathbb{E}}_{z\sim e|x} \left[\log \frac{p(z)d(x \mid z)}{e(z \mid x)} + \log \frac{1}{\Pr_{\underline{\mathbf{x}}}(x)}\right]$$
$$= \mathrm{ELBO}_{p,e,d}(x) - \mathrm{H}(\Pr_{\underline{\mathbf{x}}})$$

in the second. □

Now, we formally state and prove the more general result for β-VAEs.

Proposition 16. *The negative* β*-ELBO objective for a prior* $p(X)$*, encoder* $e(Z \mid X)$*, decoder* $d(X \mid Z)$*, at a sample* x*, is equal to the inconsistency of the corresponding PDG, where the prior has confidence equal to* β*. That is,*

$$-\beta\text{-}\mathrm{ELBO}_{p,e,d}(x) = \left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "\overset{(\beta)}{p}"] & \boxed{Z} \arrow[r, bend left, "d"] & \boxed{X} \arrow[l, bend left, "e!"] & {} \arrow[l, two heads, "x"'] \end{tikzcd} \right\rangle\!\!\right\rangle$$

[†] See appendix for a multi-sample analog.

Proof.

$$\left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "\overset{(\beta:\beta')}{p}"] & \boxed{Z} \arrow[r, bend left, "d"] & \boxed{X} \arrow[l, bend left, "e!"] & {} \arrow[l, two heads, "x"'] \end{tikzcd} \right\rangle\!\!\right\rangle = \inf_{\mu} \left[\!\!\left[\begin{tikzcd} {} \arrow[r, "\overset{(\beta:\beta')}{p}"] & \boxed{Z} \arrow[r, bend left, "d"] & \boxed{X} \arrow[l, bend left, "e!"] & {} \arrow[l, two heads, "x"'] \end{tikzcd} \right]\!\!\right](\mu)$$

$$= \inf_{\mu} \mathop{\mathbb{E}}_{\mu(X,Z)} \left[\beta \log \frac{\mu(Z)}{p(Z)} + \log \frac{\mu(X,Z)}{\mu(Z)d(X \mid Z)} \right]$$

As before, the only candidate for a joint distribution with finite score is $\delta_x(X)e(Z \mid X)$. Note that the marginal on Z for this distribution is itself, since $\int_x \delta_x(X)e(Z \mid X) \,\mathrm{d}x = e(Z \mid x)$. Thus, our equation becomes

$$\begin{aligned} &= \mathop{\mathbb{E}}_{\delta_x(X)e(Z|X)} \left[\beta \log \frac{e(Z \mid x)}{p(z)} + \log \frac{\delta_x(X)e(Z \mid X)}{e(Z \mid x)d(x \mid Z)} \right] \\ &= \mathop{\mathbb{E}}_{e(Z|x)} \left[\beta \log \frac{e(Z \mid x)}{p(Z)} + \log \frac{1}{d(x \mid Z)} \right] \qquad = -\beta\text{-ELBO}_{p,e,d}(x). \end{aligned}$$

□

Proposition 13. *For any weighted factor graph* Ψ*, we have* $\langle\!\langle \mathcal{M}_\Psi \rangle\!\rangle_1 = -\log Z_\Psi$*.*

Proof. Let the$(\{x\}) := x$ be a function that extracts the unique element singleton set. We showed in the orignal paper (Corolary 4.4.1) that

$$\text{the}[\![(\mathcal{N}_\Phi, \theta, \theta)]\!]_1^* = \Pr_{\Phi,\theta}(\mathbf{w}) = \frac{1}{Z_\Psi} \prod_j \phi_j(\mathbf{w}_j)^{\theta_j}.$$

Recall the statement of Prop 4.6 from the original paper,

$$[\![\mathcal{M}]\!]_\gamma(\mu) = \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu} \left\{ \sum_{X \xrightarrow{L} Y} \left[\beta_L \log \frac{1}{\mathbf{p}_L(y^{\mathbf{w}} | x^{\mathbf{w}})} + (\gamma\alpha_L - \beta_L) \log \frac{1}{\mu(y^{\mathbf{w}} | x^{\mathbf{w}})} \right] - \gamma \log \frac{1}{\mu(\mathbf{w})} \right\}, \tag{8}$$

but note that since $\gamma = 1$, and α, β are both equal to θ for our PDG (since $\mathcal{M}_\Psi = \mathcal{M}_{(\Phi,\theta)} = (\mathcal{N}_\Phi, \theta, \theta)$), the middle term disappears, yielding the standard variational free energy *VFE*(μ). Recall also that $\langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma = \inf_\mu [\![\mathcal{M}]\!]_\gamma(\mu)$ and $[\![\mathcal{M}]\!]_\gamma^* = \arg\min [\![\mathcal{M}]\!]_\gamma(\mu)$, so (with a minor abuse of notation), $\langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma = [\![\mathcal{M}]\!]_\gamma([\![\mathcal{M}]\!]_\gamma^*)$. We now compute the value of the inconsistency $\langle\!\langle (\mathcal{N}_\Phi, \theta, \theta) \rangle\!\rangle_1$.

$$\begin{aligned} \langle\!\langle (\mathcal{N}_\Phi, \theta, \theta) \rangle\!\rangle_1 &= [\![(\mathcal{N}_\Phi, \theta, \theta)]\!]_1 \Big(\Pr_{\Phi,\theta}(\mathbf{w}) \Big) \\ &= \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu} \left\{ \sum_{X \xrightarrow{L} Y} \left[\beta_L \log \frac{1}{\mathbf{p}_L(y^{\mathbf{w}} | x^{\mathbf{w}})} \right] - \log \frac{1}{\Pr_{\Phi,\theta}(\mathbf{w})} \right\} && \Big[\text{ by (8) } \Big] \\ &= \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu} \left\{ \sum_j \left[\theta_j \log \frac{1}{\phi_j(\mathbf{w}_j)} \right] - \log \frac{Z_\Psi}{\prod_j \phi_j(\mathbf{w}_j)^{\theta_j}} \right\} && \left[\begin{array}{c} \text{cpds } \mathbf{p}_L \text{ correspond} \\ \text{to factors } \phi_j \end{array} \right] \\ &= \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu} \left\{ \sum_j \left[\theta_j \log \frac{1}{\phi_j(\mathbf{w}_j)} \right] - \sum_j \left[\theta_j \log \frac{1}{\phi_j(\mathbf{w}_j)} \right] - \log Z_\Psi \right\} \\ &= \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu} [-\log Z_\Psi] \\ &= -\log Z_\Psi && \Big[\ Z_\Psi \text{ is constant in } \mathbf{w} \ \Big] \end{aligned}$$

□

D More Notes

D.1 Surprise

A common justification for using $\mathrm{I}_p(x)$ as a cost for updating a probabilistic model $p(x)$ based on an observed sample x, is that by minimizing it, you "maximize the probability of seeing your data".[8] But this explanation applies just as well to $-p(x)$. Why include the logarithm? There are plenty of answers to this question; among them: I_p is convex in p, it decomposes products into arguably simpler sums, is more numerically stable, has a well-defended physical analogue in thermodynamics, and is a primative of information theory.

For those after a quick and rigorous justification (as opposed to handwaving or a thermodynamics textbook), none of these answers are entirely satisfying. They suggest that I_p has certain nice properties, but not that it enjoys them uniquely, or that no other loss function satisfies nicer ones. Pedagogically speaking, the situation is more straightforward for us. Although PDG semantics themselves require non-trivial justification, they give us in return uniform answers to many questions, starting with: Why use the surprise $\mathrm{I}_p(x)$, to measure the loss of a model $p(X)$ on sample x? Because it is the inconsistency of simultanously believing $X = x$ and $X \sim p$.

[8] this justification should not be taken too seriously without constraints on p, because the optimal value of p is δ_x, which does not generalize.