(PDG) Inconsistency: The Universal Loss

May 2021

Abstract

1 Introduction

Many tasks in artificial intelligence can be fruitfully cast as optimization problems, but often the choice of objective is not unique. For instance, a loss function is a key component of a machine learning system, but hinge loss, cross entropy, and mean squared error are all popular loss functions motivated by similar intuitions of 'distance to ground truth', drawing from different formalisms. Each implicitly represents different values, and results in different behavior, and so this choice can be quite important—and yet because the criteria for choosing a "good" loss function are often inscrutible, the choice is usually made by instinct, tradition, or to simplify implementation, rather than because it reflects one's beliefs or values. Furthermore, there is often something to be gained by fiddling with these loss functions: one can add regularization terms, to incentivize desirable behavior. But the process of tinkering with the objective to get desirable behavior, and afterwards spinning a story about why it works, leaves something to be desired: the tinkering itself can be a tedious game without well-defined rules, while results so obtained are vulnerable to overfitting and pedagogically 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 some intuition about reality. This observation motivates our proposition: the use of model inconsistency a "universal" loss function. 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 Probabilistic Dependency Graphs (PDGs), a particularly expressive class graphical models which can can incorporate aribtrary (even inconsistent) probabilistic information in a natural way, and comes equipped with a well-motivated information-theoretic measure of inconsistency [7]. Surprisingly, most standard objective functions (such as cross entropy, KL divergence, Bhattacharyya distance, renyí entropies, the ELBO, L1 and L2 regularizers, log partition functions) arise naturally as inconsistencies of the appropriate underlying PDG. Such a scheme of objective specification is more restrictive, but also more intuitive (since knowledge of the objectives listed above is not necessary), and admits more epistemically grounded support and criticism. This change is analogous to a higher-level programming language, which restricts the set of legal programs but also makes programs easier to reason about, and ultimately, even easier to write.

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. 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 $e(Z|X)$. Despite the fact that they are usually introduced by way of a graphical model (the Bayesian Network comprising the prior $p(Z)$ and decoder $d(X|Z)$), 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 a standard directed graphical

This shouldn't be "but". There's no contrast. Rather, you want to say many loss functions have been considered in the literature.

the choice of loss function

add "simultaneously"

model cannot model them both (after all, they could be inconsistent with one another). The second reason: it is not a VaE's structure, but rather its loss function which makes a VaE tick. A VaE is typically trained by minimizing a function called the Evidence Lower BOund (ELBO), a particularly 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. So PDG semantics capture not only structured objects such as Bayesian Networks and Factor Graphs [7], but also in the same stroke grant VaEs legitimate graphical-model-hood, blurring the line between the structure and the objective.

which -> that

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 to give rise to very simple and intuitive proofs of otherwise unintuitive results, such as the variational inequalitites to which the ELBO owes its name, and the monotonicity of Renyí entropy in its parameter α.

In the coming sections, we will show in more detail how this concept of inconsistency, beyond simply providing a forgiving 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.

1.1 Related Work

This approach naturally extends work in the Bayesian ML community [8, 6, 1, 2] establishing a correspondence between regularizers and priors. We note that a primary benefit of this correspondance is the ability to view (some) regularizers as *prior beliefs*, which are modeling assumptions. We review this in our context in Section 5.

In parallel work focusing on PDG inference, we prove some nice properties of $\langle\!\langle - \rangle\!\rangle$, and argue that many standard algorithms for inference and updating in other probablisitc models can be viewed as algorithms for reducing inconsistency, in the corresponding PDG.

2 Preliminaries

A PDG, like a Bayesian Network (BN), is a directed graph, to which we attatch a conditional probability distributions (cpds). While this data is attached to the *nodes* of a BN, it is attached to the *edges* of PDG. Concretely, while in a BN, the variable X is associated with the distribution $\Pr_X(X \mid \mathbf{Pa}(X))$ on X given its parents—while in a pdg, an edge $X \xrightarrow{L} Y$ is associated with a distribution $\Pr(Y \mid X)$. As a result, a PDG of shape $[X \to Y \leftarrow Z]$ gives a cpd $p(Y \mid X)$ on Y given X and also (separately) a cpd $q(Y \mid Z)$ on Y given Z, while a BN of the same shape has a single cpd $\Pr(Y \mid X, Z)$ on Y given joint values of X, Z. The first interpretation is more expresive, and can encode joint dependence with an extra variable and some syntactic sugar [7]. Formally, the syntax of a PDG is given in Definition 1.

But first, let's fix some notation for the remainder of the paper. We generally use capital letters for variables, and lower case letters for their values. We use letters of both cases for conditional probability distributions (cpds), which we generally conflate their with probability mass functions. Depending on context, we write either $\mathbb{E}_\mu X$ or $\mathbb{E}_{\omega\sim\mu} X(\omega)$ for expectation of $X : \Omega \to \mathbb{R}$ over a distribution μ with outcomes Ω. $\mathrm{H}(\mu)$ and $\mathrm{H}_\mu(X)$, $\mathrm{H}_\mu(Y \mid X)$ refer respectively to the entropy of a distribution μ, the entropy of the variable X with respect to μ, and the conditional entropy of Y given X, with respect to μ. We now proceed with the definition.

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

$\mathcal{N}$ is a finite set of nodes, corresponding to variables;

$\mathcal{E}$ is a set of labeled edges $\{X \xrightarrow{L} Y\}$, each with a source X and target Y in $\mathcal{N}$;

$\mathcal{V}$ associates each variable $N \in \mathcal{N}$ with a set $\mathcal{V}(N)$ of values that the variable N can take;

$\mathbf{p}$ associates to each edge $X \xrightarrow{L} Y \in \mathcal{E}$ a conditional distribution $\mathbf{p}_L(Y \mid x)$ on Y for each $x \in \mathcal{V}(X)$;

α associates to each edge $X \xrightarrow{L} Y$ a non-negative number α_L which, roughly speaking, is the modeler's confidence in the functional dependence of Y on the variables X implicit in L;

β associates to each edge L a real number β_L, 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, and An unconditional distribution $p(X)$ can be given by associating it to a hyper-edge $\emptyset \to \{X\}$, or directly by Definition 1 by taking its source to be a "variable" that can only take on one possible value.

PDGs, like other graphical models, have semantics in terms of joint distributions over all variables. Most directly, a PDG $\mathcal{M}$ determines a family of loss (scoring) functions on joint distributions μ, with a single parameter $\gamma > 0$, given by

$$[\![\mathcal{M}]\!]_\gamma(\mu) = \gamma \mathit{IDef}_{\mathcal{M}}(\mu) + \sum_{X \xrightarrow{L} Y} \mathop{\mathbb{E}}_{x \sim \mu(X)} \beta_L \boldsymbol{D}\Big(\mu(Y \mid x) \,\Big\|\, \mathbf{p}_L(Y \mid x)\Big) \tag{1}$$

where $\boldsymbol{D}(\mu \parallel \nu) = \mathbb{E}_\mu \log \frac{\mu}{\nu}$ is the relative entropy (KL Divergence) of ν with respect to μ, and $\mathit{IDef}_{\mathcal{M}}(\mu) = -\operatorname{H}(\mu) + \sum_{X \xrightarrow{L} Y} \alpha_L \operatorname{H}(Y \mid X)$, called the information deficiency (which plays a minimal role in the present paper), measures the degree to which μ reflects the qualitative features one would expect if it were generated by the causal weighted graph $(\mathcal{N}, \mathcal{E}, \alpha)$. For the examples here, we can select α such that $\mathit{IDef}_{\mathcal{M}}(\mu) = 0$ for all μ, and we can also set $\gamma = 0$ to ignore it explicitly.

can this always be done?

can select or do select?

We are primarily concerned with the second term of (1), which penalizes a candidate distribution μ for every deviation from a cpd $\mathbf{p}_L(Y \mid X)$, in proportion to the number of bits needed to transmit samples from $\mu(Y \mid X)$ with a code optimized for the agent's belief $\mathbf{p}_L(Y \mid X)$, scaled by the confidence of the edge. In this way, violations of high-confidence beliefs are more costly per bit.

add comma

The *best* distribution by this metric (the one with the smallest loss), denoted $[\![\mathcal{M}]\!]^*_\gamma$ exists and is unique for small enough γ, and in the limit as $\gamma \to 0$, which is the intended reading of $[\![\mathcal{M}]\!]^*$, without a subscript. This is how PDGs capture other graphical models. The value of (1) achieved by this optimal distribution is called the *inconsistency* of $\mathcal{M}$, is the best possible incompatibility is denoted by $\langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma$. Because we can essentially set $\mathit{IDef} = 0$ with certain neutral choices of α, and because IDef is not the focus of this paper, we will often drop the subscript γ. Any such formula can be interpreted litterally by setting $\gamma = 0$, and our results can be extended for other values of γ with more careful treatmeant of the qualitative half of the picture.

hard to parse

Now, some short-hand. A cpd p can be regarded as a PDG with a single edge, and sometimes we will implicitly convert a cpd to a PDG. The union of $\mathcal{M}$ and $\mathcal{M}'$ is denoted by $\mathcal{M} + \mathcal{M}'$. A value x of a variable X is identified with the degenerate distribution δ_x that places all mass on x, and hence may be associated to an edge. By default, all cpds are associated with confidence $\beta = 1$, unless specified otherwise, which will be done as a parameter list near the name of the cpd, as in $p^{(\beta:\, 0.7)}$, or $p^{(\beta)}$ as a compressed form of of $p^{(\beta:\, \beta)}$. We will use an exclamation point (e.g., $X \xrightarrow{p!} Y$) to inciate the limit of high confidence ($\beta_p \to \infty$).

2.1 Monotonicity of Inconsistency

It is easy to show that adding edges, or increasing their confidences, cannot decrease the inconsistency of a PDG. Intuitively, believing more things can't make you any less inconsistent. This is captured formally in two different (but related) ways by the following lemma, which is the primary tool we will use to derive relationships between loss functions.

Lemma 1 (monotonicity of inconsistency)**.** *For all pdgs $\mathcal{M}$, $\mathcal{M}'$, we have that*

1. $\langle\!\langle \mathcal{M} + \mathcal{M}' \rangle\!\rangle \geq \langle\!\langle \mathcal{M} \rangle\!\rangle$.
2. *If $\mathcal{M}$ and $\mathcal{M}'$ have respective confdience vectors β and β', and $\beta \succeq \beta'$ (that is, $\beta_L \geq \beta'_L$ for all $L \in \mathcal{E}$), then $\langle\!\langle \mathcal{M} \rangle\!\rangle \geq \langle\!\langle \mathcal{M}' \rangle\!\rangle$.*

Proof. For every μ, adding more edges only adds non-negative terms to (1). In the first case this means, $[\![\mathcal{M} + \mathcal{M}']\!]_\gamma(\mu) \geq [\![\mathcal{M}]\!]_\gamma(\mu)$ for all γ and μ. So it also holds when we take an infemum over μ, giving $\langle\!\langle \mathcal{M} + \mathcal{M}' \rangle\!\rangle_\gamma \geq \langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma$. Analogously, increasing β results in larger scalings for non-negative terms in (1) so $[\![\mathcal{M}]\!]_\gamma(\mu) \geq [\![\mathcal{M}']\!]_\gamma(\mu)$ for all γ and μ. □

3 Simple Information-Theoretic Objectives as Inconsistency

To illustrate how the inconsistency of a PDG can start to generate information theoretic expressions, let's start with a simple example. Suppose you believe that X is distributed according to a probability distribution with mass function p, and also that it (always) equals a certain value x. These beliefs are entirely consistent if $p(X = x) = 1$, and become more inconsistent (in our sense) as $p(X = x)$ decreases. In fact, it is equal to $\mathrm{I}_p(x) = -\log p(x)$, the information content (or surprise) of the event $X = x$, according to p. It also goes by the name "negative log likelihood", and is perhaps the most popular objective for training generative models.

Proposition 2. *Consider a distribution over X with mass function $p(X)$. The surprise $\mathrm{I}_p(x)$ at seeing a sample x is given by the inconsistency of the pdg containing a point mass on x and p, i.e.,*

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

[link to proof]

In some ways, this result is entirely unsurprising, given that PDG scoring rule (1) is a complicated and flexible formula built out of information theoretic primitives. With a little more focus on the context and less on the algebra, perhaps Proposition 2 becomes more curous: the inconsistency of the PDG containing just a distribution $p(X)$ and a sample x just so happens to equal the overlwhelming favorite measure of incompatibility between a distribution p and sample x — and it is known as "surprise", a particular kind of cognitive dissonance.

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".[1] 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

[1] 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.

non-trivial justification (e.g., by direct appeal to information theory), they give us in return simple, intuitive, and uniform answers to many questions — starting with following, simple one. 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 $x \sim p$.

For this argument to be pesuasive, one needs to accept our definition of inconsistency more broadly. Proposition 2 is a very simple special case. One concern is that it does not model the whole state of affairs; we train probabilstic models models with more than one sample. What if we replce x with an emperical distribution on many samples? The first step in the proof of Proposition 2, is no longer valid, but we can get the same effect by being extremely confident in the data distribution.

Proposition 3. *Given a model determining a probability distribution with mass function $p(X)$, and samples $\underline{\mathbf{x}} = \{x_i\}_{i=1}^m$ determining an emperical 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. $\left\langle\!\!\left\langle \xrightarrow{p} X \xleftarrow{\Pr_{\underline{\mathbf{x}}}!} \right\rangle\!\!\right\rangle_\gamma \quad + (1+\gamma)\,\mathrm{H}(\Pr_{\underline{\mathbf{x}}})$

link to proof

Remark 1. *Note that the entropy of the data distribution $\mathrm{H}(\Pr_{\underline{\mathbf{x}}})$ does not depend on p, and so the final grey term in (3, 4) is not relevant for optimizing with respect to p.*

The PDG appearing in the final quantity of Proposition 3 is quite natural. Aside from the certainty weights β, we've simply translated each piece of information into a cpd, and collected them together as a PDG. We argue that the choice of β makes sense as well, in the standard use cases for cross entropy / log likelihood. The cross entropy measures the expected code length per sample, when a (possibly incorrect) distribution p is used to design a code, in place of the true one $\Pr_{\underline{\mathbf{x}}}$. So implicitly, a modeler who chooses a cross-entropy has in some sense implicitly articulated a belief the data distribution is the "true one", by placing infinite certainty in $\Pr_{\underline{\mathbf{x}}}$ than in p. If $p(X)$ represents a probabilistic model before training is complete (say, a neural network with randomly initialized weights), we would be justified in placing dramatically more trust in the data $\Pr_{\underline{\mathbf{x}}}$ than in p, and so this choice seems particularly appropriate. Other choices of β are natural in other contexts, and we will see in Section 4 that they correspond to other losses. (For instance when $\beta_p = \beta_{\Pr_{\underline{\mathbf{x}}}} = 1/2$, the result is the Bhattacharyya distance, rather than the cross entropy.)

We now consider an orthogonal generalization of Proposition 2, in which x is only a partial observation of a joint model of two variables X, Z. In this case, we might hope to recover the *marginal* negative log likelihood, since Z does not interact with the observation.

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*

$$\mathrm{I}_p(x) = \log \frac{1}{p(x)} = \left\langle\!\!\left\langle Z \overset{p}{\swarrow\!\searrow} X \xleftarrow{x} \right\rangle\!\!\right\rangle. \tag{2}$$

link to proof

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.

In much the same way as Proposition 3 generalizes Proposition 2 by considering more than one sample at once, we also can obtain a multi-sample analogue of analog of Proposition 4.

Proposition 5. *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,

$$\ell(p;\underline{\mathbf{x}}) = \left\langle\!\!\left\langle \begin{array}{ccc} & & p \\ & \swarrow\!\!\searrow & \\ Z & & X \end{array} \quad \Pr_{\underline{\mathbf{x}}}! \right\rangle\!\!\right\rangle + \mathrm{H}(\Pr_{\underline{\mathbf{x}}}).$$

[link to proof]

no comma; which -> that

So far we have been considering models of a distribution $p(X)$, which correspond to generative models, or an unsupervised setting. The supervised setting, in which our model predicts Y from X, and is trained from joint samples (X, Y), cross entropy loss is perhaps even more dominant. And it equals the inconsistency of a PDG consisting of the predictor $f(Y|X)$ together with high-confidence data.

Proposition 6. *Consider a probabilistic model f with mass function $f(y \mid x)$. The inconsistency of the PDG containing $p(y \mid x)$ and the emperical distribution $\Pr_{\underline{\mathbf{xy}}}$ of samples $\underline{\mathbf{xy}}$ is equal to the negative log-liklihood (cross-entropy) loss, plus the emperical uncertainty in Y given X (a constant independent of f). That is,*

$$\left\langle\!\!\left\langle \begin{array}{ccc} & \Pr_{\underline{\mathbf{xy}}}! & \\ & \swarrow\!\!\searrow & \\ X & \xrightarrow[f]{} & Y \end{array} \right\rangle\!\!\right\rangle = \frac{1}{|\underline{\mathbf{xy}}|} \sum_{(x,y)\in\underline{\mathbf{xy}}} \left[\log \frac{1}{f(y \mid x)}\right] \quad - \mathrm{H}_{\Pr_{\underline{\mathbf{xy}}}}(Y \mid X).$$

[link to proof]

Remark 2. *Propositions 2 to 6 all require the probability to have a mass function, rather than a density function, which implicitly restricts us to discrete variables. A PDG containing both a continuous distribution function and a discrete empmerical distribution is infinitely consistent, because the probability of any particular sample is zero. Indeed, the analog of surprise $-\log p(X)$ is not well-founded for a density $p(X)$, because $p(X)$ is not dimensionless, but rather in inverse X-units, so depends on an (arbitrary) choice of scale. However, such a rescaling only amounts to adding a constant.*

Furthermore approximating $p(X)$ with a discretization also only contributes a constant, in the sense that the analogues of Propositions 2 to 6 in which a density function is used in place of a mass function, are identical— save for an additive constant depending only discretization size. But this additive constant is irrelevant for optimization, and so in this sense the results here also justify use of the continuous analogues (such as $-\log p(X)$ for a pdf p) as loss functions, even though these functions are not invariant with respect to changes of parameterization.

3.1 Simple Performance Metrics as Inconsistencies

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?

Proposition 7 (log accuracy as inconsistency)**.** *If $h : X \to Y$ is a predictor for an input space X and label space Y, and $f : X \to Y$ generates the correct labels, then the inconsistency of believing f and h (with any degree of confidence), and also that inputs are distributed according to $D(X)$, equals the information content of learning that $f(X) = h(X)$ according to D (which is the log accuracy of the predictor h) times the confidence in D. That is,*

$$\left\langle\!\!\left\langle \begin{array}{ccc} D^{(\beta)} & & h \\ \longrightarrow & X & \rightrightarrows \\ & & f \end{array} \quad Y \right\rangle\!\!\right\rangle = -\beta \log\Big(\text{accuracy}_{f,D}(h)\Big) = \beta\, \mathrm{I}_D[f = h]. \tag{3}$$

[link to proof]

It is common to think of accuracy as a property of a hypothesis h (with some dependence on the true labeling function f and the data distribution D), even though it is symmetric in f and g, in part because we change the predictor more than the labels. The correspondence of Proposition 7 graphically

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 $\mathrm{M}_p^w(\mathbf{r})$

depicts this symmmetry, and also a more subtle phenomenon: a particularly strong dependence on the distribution of inputs. In fact, the inconsistency of the PDG in (3) is scaled by the confidence of D, but does not depend at all on the confidences in h or f.

Why should this be the case? The deterministic nature of f and h encodes extreme confidence (to anthropomorphise: f thinks it impossible that $Y \neq f(X)$, because that event has probability zero), and so the only way to selectively choose samples to make things consistent is to alter the distribution of *inputs* $\Pr(X)$ (so that it is always the case that $f(x) = g(x)$), not to change $\Pr(Y \mid X)$. But it is the direction that we want to pull $\Pr(Y \mid X)$ that informs how we should update our predictor $h(Y \mid X)$, which reflects the non-differentiability of the accuracy (or zero-one loss) with respect to h. Note that this is a the complete opposite of how we captured cross entropy (Proposition 6), in which we are unwilling to budge on the data distribution, but are willing to modify our predictor h.

When Y is a continuous variable rather than a discrete one, the estimator is referred to as a regressor, rather than a classifier, and the topology of the real numbers comes with an intuition that not all mistakes are equally bad; a small deviation from the correct value of Y is close to correct. 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. But before we get there, we first prove a more general result, which is most clearly articulated in terms of a power mean.

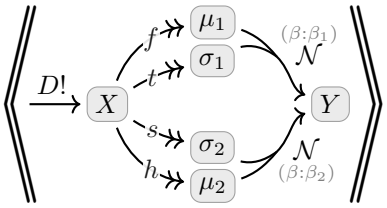
Definition 2. The weighted power mean $\mathrm{M}_p^w(\mathbf{r})$ of the collection of real numbers $\mathbf{r} = r_1, \ldots, r_n$ with respect to the convex weights $w = w_1, \ldots, w_n$ satisfying $\sum_i w_i = 1$, is given by

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

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.

Proposition 8. *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*



$$= \frac{1}{2}\,\mathbb{E}_D\left[\mathrm{HM}(\beta_1, \beta_2)\frac{1}{2}\left(\frac{\mu_1 - \mu_2}{\mathrm{QM}_{\hat{\beta}}(\sigma_1, \sigma_2)}\right)^2 + \mathrm{AM}(\beta_1, \beta_2)\log\frac{\mathrm{QM}_{\hat{\beta}}(\sigma_1, \sigma_2)}{\mathrm{GM}_{\hat{\beta}}(\sigma_1, \sigma_2)}\right] \tag{4}$$

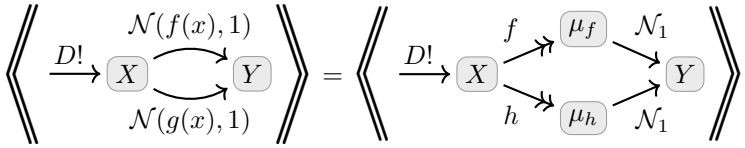
$$= \mathop{\mathbb{E}}_{x\sim D}\left[\frac{\beta_1\beta_2}{2}\frac{\Big(f(x)-h(x)\Big)^2}{\beta_2 s(x)^2+\beta_1 t(x)^2}+\frac{\beta_1+\beta_2}{2}\log\frac{\beta_2 s(x)^2+\beta_1 t(x)^2}{(\beta_1+\beta_2)(s(x)^{\beta_2}t(x)^{\beta_1})^{\frac{1}{\beta_1+\beta_2}}}\right]$$

where $\hat{\beta}=(\frac{\beta_2}{\beta_1+\beta_2},\frac{\beta_1}{\beta_1+\beta_2})$ *represents the normalized and reversed vector of confidences* $\beta=(\beta_1,\beta_2)$ *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.

[link to proof]

Plugging in $s(x)=t(x)=1$ and $\beta_1=\beta_2=1$ gives us the result we had hinted at before.

Corollary 8.1 (Mean Square Error as inconsistency)**.**



$$= \mathbb{E}_D\Big(f(X)-h(X)\Big)^2 =: \mathrm{MSE}(f,h), \tag{5}$$

where $\mathcal{N}_1=\mathcal{N}(-,1)$ *is the normal distribution with unit variance, and mean equal to its argument.*

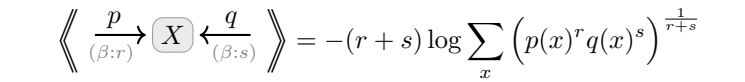
The equality of the two PDG inconsistencies illustrates an orthogonal point: that PDGs handle composition of functions as one would expect, so that it is equivalent to model an entire process as a single arrow, or to break it into stages, ascribing an arrow to each stage, with one step of randomization.

4 Statistical Distances as Inconsistencies

Suppose you are concerned with only a single variable X. One friend has told you that it is distributed according to the probabiltiy distribution $p(X)$; another has told you that it follows $q(X)$. You adopt both beliefs. If $p\neq q$, your mental state will be inconsistent, and moreso, the more p and q differ. As a result, we can view of the inconsistency of a PDG containing only a single varible X and two distributions $p(X)$ and $q(X)$ over it, as a measure of divergence.

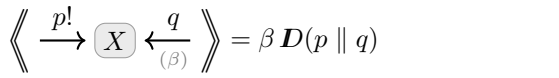
Recall that a PDG also allows us to specify the confidences β_p and β_q of each edge, and so we can form a distinct PDG containing only the distributions p and q for each setting of (β_p,β_q). It turns out that a surprisingly large classs of statistical distances arise as the inconsistency of such a PDG.

Lemma 9.



$$= -(r+s)\log\sum_x\Big(p(x)^r q(x)^s\Big)^{\frac{1}{r+s}}$$

Proposition 10 (KL Divergence as Inconsistency)**.** *The inconsistency of believing* p *with complete certainty, and also* q *with some finite certainty* β*, is* β *times the relative entropy (or KL divergence) of* q *with respect to* p*. That is,*



$$= \beta\, \boldsymbol{D}(p\parallel q)$$

This result gives us an intuitive interpretation of the asymmetry of the KL divergence, the counter-intuitive naming of itsarguments, and a prescription about when it makes most sense to use it. $\boldsymbol{D}(p\parallel q)$ is the inconsistency of a mental state when absolutely certain of p (and not willing to budge on it), which is why it reflects the amount of information required to change q into p, and why we call it the divergence from q to p.

Proposition 11 (Renyí Divergences)**.**

$$\boldsymbol{D}^{\mathrm{PDG}}_{r,s}(p,q)=s\cdot\boldsymbol{D}_{\frac{r}{r+s}}(p\parallel q)\qquad \textit{and}\qquad \boldsymbol{D}_\alpha(p\parallel q)=\boldsymbol{D}^{\mathrm{PDG}}_{(\frac{\alpha}{1-\alpha},1)}(p,q)$$

The Chernoff information

Proposition 12 (Chernoff Divergences)**.**

$$\inf_{\beta\in(0,1)} \left\langle\!\!\left\langle \overset{(\beta)}{p} \longrightarrow \boxed{X} \longleftarrow \overset{(1-\beta)}{q} \right\rangle\!\!\right\rangle = \boldsymbol{D}^{Chernoff}(p,q)$$

5 Regularizers and Priors

The picture presented here is compatible with the correspondence between regularization and Bayesian priors; as has been noted by other authors [8, 6, 2, 3], L2 regularization corresponds to a Gaussian prior [6], and indeed adding such a distribution to the PDG gives an extra term in the consistency, equal to the L2 regularizer. Similarly, the L1 regularizer corresponds to a Leplacian prior [8].

We now review these findings in the context of our framework, as we show how adding a prior distribution in a PDG results in the corresponding regularization term in the inconsistency.

Proposition 13. *Consider a situation where you believe that Y is distributed according to $f_\theta(Y)$ for some parameter θ, and also that you have a prior belief $p(\theta)$ over parameters, and an emperical distribution D over Y which you trust. The inconsistency of also believing that the parameter is some θ_0 is the* regularized-*cross entropy loss, where the regularizer is $\log \frac{1}{p(\theta_0)}$ times the confidence in the prior p. That is,*

$$\left\langle\!\!\left\langle \begin{array}{c} \overset{p^{(\beta)}}{\searrow} \\ \underset{\theta_0}{\searrow\!\!\searrow} \end{array} \boxed{\Theta} \xrightarrow{f} \boxed{Y} \xleftarrow{D!} \right\rangle\!\!\right\rangle = \mathop{\mathbb{E}}_{y\sim D}\left[\log \frac{1}{f(y \mid \theta_0)}\right] + \beta \log \frac{1}{p(\theta_0)} - \mathrm{H}(D) \tag{6}$$

link to proof

Using a (discretized) unit gaussian as a prior, $p(\theta) = \frac{1}{k}\exp(-\frac{1}{2}\theta^2)$ for a normaization constant k, the RHS of (6) becomes

$$\underbrace{\mathop{\mathbb{E}}_{D}\left[\log \frac{1}{f(Y \mid \theta_0)}\right]}_{\substack{\text{Cross entropy loss of } f_\theta \text{ w.r.t. } D \\ \text{(data-fit cost of } \theta_0\text{)}}} + \underbrace{\frac{\beta}{2}\theta_0^2}_{\substack{\ell_2 \text{ regularizer} \\ \text{(complexity cost of } \theta_0\text{)}}} \underbrace{+\beta \log k - \mathrm{H}(D)}_{\substack{\text{constant in } f \text{ and } \theta_0}},$$

which is the ℓ_2 regularized version of Proposition 3. Moreover, the regularization strength corresponds exactly to the confidence β functions as the hyperparameter, a feature unique to our approach.

What about other priors? It is not difficult to see that if we use a (discretized) centered unit Laplacian prior, $p(\theta) = \frac{1}{k}\exp(-|\theta|)$, the second term instead becomes $\beta|\theta_0|$, which is ℓ_1 regularization. More generally, to consider any complexity measure $E(\theta)$, we need only include the Gibbs distribution $\Pr_E(\theta) \propto \exp(-E(\theta))$ into our PDG. And if we lift the restriction that the edge labeled θ_0 is a point mass, the optimal distribution over θ is the posterior distribution after a Baysian update.

Finally, we remark that there is nothing special here about cross entropy here; such a prior may be added to (5) to regularize square loss instead.

6 Varitaional Objectives and Bounds

The fact that the incompatibility of $\mathcal{M}$ with a *specific* joint distribution μ is an upper bound on the inconsistency (i.e., the score of the *best* such joint distribution) is not deep, but it does yield an efficient approximate inference procedure for PDGs: choose a tractable parametric family $\mathcal{P}$ of joint distributions μ, and optimize (1). Making this precise is the focus of a parallel work. Here, we focus on the more

surprising converse: PDG semantics capture many interesting features of variational inference, broadly construed— and moreover, provide a concise graphical language for this kind of reasoning.

We will demonstrate by showing that the the 'Evidence Lower BOund' (ELBO), a common objective for training latent variable models. Suppose we have a joint distribution $p(X, Z)$, but only have access to observations X. Evaluating $p(X)$ requires marginalizing over Z, and could be intractable if Z is a complex space. One workaround is as follows. Fix a family of distributions $\mathcal{Q}$ and assume that Z is distributed according to some $q \in \mathcal{Q}$; to start, we guess $q_0(Z)$. Now it's easy to sample Z (by assumption), but q_0 is likely to be inconsistent with p. But if, while optimizing p to best fit the data, we also optimize q so that it closly mirrors p, (and if $\mathcal{Q}$ is expressive enough), we can avoid computing the "evidence" $\log p(x) = \log \int p(x, z)\, \mathrm{d}z$. But only by being clever. The straightforward divergence-minimizing approach to both optimization problems (the choice of p and q) requires computing this integral [4, §2.2]. Instead, q and p are chosen so as to maximize:

$$\mathrm{ELBO}_{p,\underline{\mathbf{x}}}(q) := \sum_{x \in \underline{\mathbf{x}}} \mathrm{ELBO}_{p,q}(x)\ , \qquad \text{where} \qquad \mathrm{ELBO}_{p,q}(x) := \mathop{\mathbb{E}}_{z \sim q} \log \frac{p(x,z)}{q(z)}, \tag{7}$$

This formula is somewhat difficult to make sense of, especially if p and q are densities, in which case the units don't cancel in the logarithm, making it an ill-defined physical quantity (see Remark 2). If p and q are mass functions, though, it turns out that the ELBO also arises naturally as the inconsistency of the obvious.

Proposition 14. *The negative ELBO is the inconsistency of the PDG containing p, q, and x, with very high confidence in q. That is,*

$$-\mathrm{ELBO}_{p,q}(x) = \left\langle\!\!\left\langle \xrightarrow{q!} \boxed{Z} \overset{p}{\curlyvee} \boxed{X} \xleftarrow{x} \right\rangle\!\!\right\rangle.$$

[link to proof]

The use of the ELBO as an objective is often justified by the fact that it lower-bounds the objective that you "really wanted": the cross entropy. This is often proved by jensen's inequality, or alternatively, but appeal to the non-negativity of relative entropy [**?**]. We now give a very simple, different-looking diagrammatic proof of this second approach, by appealing to the intuitive fact that adding believing more things cannot make you less inconsistent, as captured by Lemma 1.

$$\log \frac{1}{p(x)} = \left\langle\!\!\left\langle \boxed{Z} \overset{p}{\curlyvee} \boxed{X} \xleftarrow{x} \right\rangle\!\!\right\rangle \leq \left\langle\!\!\left\langle \xrightarrow{q!} \boxed{Z} \overset{p}{\curlyvee} \boxed{X} \xleftarrow{x} \right\rangle\!\!\right\rangle = -\mathrm{ELBO}_{p,q}(x),$$

The first and last equalities are Propositions 4 and 14 respectively, and the inequality is an instance of Lemma 1. We submit that this proof is more intuitve and gives a clearer picture of why it should be true. The second PDG has more edges, so it must be at least as inconsistent. Furthermore, this picture shows us that if we simultaneously vary $q(Z)$ within an expressive enough class of distributions, the inequality is also tight, because the distribution that realizes the smallest inconsistency must have some marginal on Z — and taking q to be that distribution will incur no additional inconsistency.

We also have analog analog holds for the entire dataset at once, which is more easily formulated with a slightly different variational form in the next section.

6.1 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 (a compact vector), and is specified by two conditional distributions: an encoder $e(Z \mid X)$, and a decoder $d(X \mid Z)$. Of course, not all pairs of cpds fill this role equally well. Perhaps

most importantly, we would to have low *reconstruction error* (8)—when we decode an encoded image, we would like it to be reasonably similar to the original.

$$\mathrm{Rec}(x) = \mathop{\mathbb{E}}_{z\sim e|x} \underbrace{\mathrm{I}_{d|z}(x)}_{\substack{\text{Additional bits to}\\ \text{specify } x \text{ from } d|z}} = \sum_z e(z \mid x) \log \frac{1}{d(x \mid z)} \tag{8}$$

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 wish in the form of a that Z is distributed according to our favorite $p(Z)$, known as a variational prior.

In the autoencoding setting, e functions as a variational approximation to the latent variable Z, differing from $q(Z)$ of the previous section only in that it can depend on X. Here, the analogue of the ELBO becomes

$$\begin{aligned}\mathrm{ELBO}_{p,e,d}(x) &= \mathop{\mathbb{E}}_{z\sim e|x}\left[\log \frac{p(z)d(x \mid z)}{e(z \mid x)}\right] \\ &= \mathop{\mathbb{E}}_{z\sim e|x}\left[\log \frac{p(z)}{e(z \mid x)}\right] - \mathop{\mathbb{E}}_{z\sim e|x}\left[\log \frac{1}{d(x \mid z)}\right] \\ &= \boldsymbol{D}(e(Z|x) \parallel p) - \mathrm{Rec}(x).\end{aligned}$$

This gives us the following results, analogous to Proposition 14.

Proposition 15. *The conditional ELBO used as a VaE objective is the inconsistency of the PDG containing the encoder e, decoder d prior p, and a sample x. That is,*

$$-\mathrm{ELBO}_{p,e,d}(x) = \left\langle\!\!\left\langle \xrightarrow{p} \boxed{Z} \overset{d}{\underset{e!}{\rightleftarrows}} \boxed{X} \xleftarrow{x} \right\rangle\!\!\right\rangle$$

link to proof

Proposition 16. *The following analog of Proposition 15 for a whole dataset $\underline{\mathbf{x}}$ holds:*

$$-\mathop{\mathbb{E}}_{\Pr_{\underline{\mathbf{x}}}} \mathrm{ELBO}_{p,e,d}(X) = \left\langle\!\!\left\langle \xrightarrow{p} \boxed{Z} \overset{d}{\underset{e!}{\rightleftarrows}} \boxed{X} \xleftarrow{\Pr_{\underline{\mathbf{x}}}!} \right\rangle\!\!\right\rangle + \mathrm{H}(\Pr_{\underline{\mathbf{x}}}).$$

link to proof

6.1.1 Intuitive Proofs of Variational Bounds

Propositions 15 and 16 can be used to derive the variational lower bound. Once again, the addition of the edge e cannot decrease the inconsistency (Lemma 1), but it makes it easier to identify and sample from the best-scoring distributions. This result in the following simple visual proof:

$$-\log \Pr(x) = \left\langle\!\!\left\langle \xrightarrow{p} \boxed{Z} \xrightarrow{d} \boxed{X} \xleftarrow{x} \right\rangle\!\!\right\rangle \leq \left\langle\!\!\left\langle \xrightarrow{p} \boxed{Z} \overset{d}{\underset{e!}{\rightleftarrows}} \boxed{X} \xleftarrow{x} \right\rangle\!\!\right\rangle = -\mathrm{ELBO}_{p,e,d}(x).$$

Here $\Pr(X)$ is the marginal of $p(Z)d(X \mid Z)$ on X. Propositions 3 and 16 lets us do the same thing for many i.i.d. datapoints at once, with only a single application of the inequality:

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

$$\mathrm{H}(\Pr_{\underline{\mathbf{x}}}) + \left\langle\!\!\left\langle\; \xrightarrow{p} \boxed{Z} \overset{d}{\curvearrowright} \boxed{X} \xleftarrow{\Pr_{\underline{\mathbf{x}}}!} \;\right\rangle\!\!\right\rangle \leq \left\langle\!\!\left\langle\; \xrightarrow{p} \boxed{Z} \underset{e!}{\overset{d}{\rightleftarrows}} \boxed{X} \xleftarrow{\Pr_{\underline{\mathbf{x}}}!} \;\right\rangle\!\!\right\rangle + \mathrm{H}(\Pr_{\underline{\mathbf{x}}})$$

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

6.2 β-VaE Objective

The ELBO is not the only objective to train VaEs. For instance, Higgins et. al. [5] have argued that the one might want to weight the reconstruction loss and KL term differently. They suggest an objective of the form

what is this?

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

which for $\beta = 1$ is equivalent to the ELBO from the previous section. The authors view it as a regularization parameter, annd argue that in some cases, you can do better with a stronger prior. Sure enough, this extra parameter corresponds to the confidence of p, which also happens go by the same name.[2]

Proposition 17. *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,*

$$-\beta\text{-ELBO}_{p,e,d}(x) = \left\langle\!\!\left\langle\; \overset{(\beta)}{\xrightarrow{p}} \boxed{Z} \underset{e!}{\overset{d}{\rightleftarrows}} \boxed{X} \xleftarrow{x} \;\right\rangle\!\!\right\rangle$$

link to proof

7 Log Partition Function as Inconsistency

The factors of a factor graph, which in isolation indicate relative probabilities. Factored exponential families form the mathematical backbone of statistical mechanics, in which the normalization constant Z_Ψ for the WFG $\Psi = (\phi_j, \theta_j)_{j \in \mathcal{J}}$, given by

$$Z_\Psi := \sum_{\mathbf{w}} \prod_{j \in \mathcal{J}} \phi_j(\mathbf{w}_j)^{\theta_j},$$

is known as the *partition function*. If every factor is a cpd, and every variable has at least one incoming 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 inconsistency.

link to proof

Proposition 18. *For any weighted factor graph Ψ, we have $\langle\!\langle \boldsymbol{m}_\Psi \rangle\!\rangle_1 = -\log Z_\Psi$.*

Many thermodynamic quantities (e.g., as internal energy, free energy, pressure, volume, and entropy) can be obtained by taking various partial derivatives of Z_Ψ, and calculating the partition function is closely related to infering marginal distributions [].

8 Discussion

Inconsistency, in a different light. Nobody likes building broken things, and so those of who build systems commonly try to eliminate it in models by design. But in doing so, we unwittingly sacrifice tools for dealing with it if (and when) it does arise.

[2] the two terms actually both share an origin in thermodynamic β for inverse temperature.

A Universal Objective. Objective functions in machine learning are often quite opaque to newcomers. Only a few of them are especially common, and they are often presented in an ad-hoc way; there are many loss functions that have the properties we need to train networks. So why do we lean so heavily on a few select choices (dependent on the problem setup), such as the cross entropy, and the ELBO objective? PDGs provide one answer to this question.

There is perhaps a more important argument for using PDGs over other models. By designing a model, rather than an objective function, it is easier to understand what the pieces are, what is and is not relevant to the task, and no longer possible to twiddle with the objective until you get the results you want — you can only twiddle with the model, where hacks are more easily spotted.

Semantics for "Graphical Models". Often, autoencoder tutorials will include diagrams, and falsely claim that they are standard "graphical models". In these cases, among others, people are already reasoning about local probabilistic information in a graph. We have shown here that PDG semantics can sense of these informal diagrams in a way that is deeply connected with the variational reasoning involed.

References

[1] https://towardsdatascience.com/a-bayesian-take-on-model-regularization-9356116b6457.

[2] https://bjlkeng.github.io/posts/probabilistic-interpretation-of-regularization/.

[3] https://www.mit.edu/~9.520/spring09/Classes/class15-bayes.pdf.

[4] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[5] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[6] Jason Rennie. On l2-norm regularization and the gaussian prior. 2003.

[7] Oliver Richardson and Joseph Y Halpern. Probabilistic dependency graphs. *AAAI*, 2021.

[8] Peter M Williams. Bayesian regularization and pruning using a laplace prior. *Neural computation*, 7(1):117–143, 1995.

A Proofs

Proposition 2. *Consider a distribution over X with mass function $p(X)$. The surprise $\mathrm{I}_p(x)$ at seeing a sample x is given by the inconsistency of the pdg containing a point mass on x and p, i.e.,*

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

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

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

□

Proposition 3. *Given a model determining a probability distribution with mass function $p(X)$, and samples $\underline{\mathbf{x}} = \{x_i\}_{i=1}^m$ determining an emperical 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. $\left\langle\!\!\left\langle \xrightarrow{p} \boxed{X} \xleftarrow{\Pr_{\underline{\mathbf{x}}}!} \right\rangle\!\!\right\rangle_\gamma \quad + (1+\gamma)\,\mathrm{H}(\Pr_{\underline{\mathbf{x}}})$

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 $\Pr_{\underline{\mathbf{x}}}$, the optimal distribution must also equal $\Pr_{\underline{\mathbf{x}}}$, so the least inconsistency is attained at this value. Finally it remains to show that the first two and last two are equal:

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

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*

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

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

$$\left\langle\!\!\left\langle \boxed{Z} \overset{p}{\curlywedge} \boxed{X} \overset{x}{\twoheadleftarrow} \right\rangle\!\!\right\rangle = \inf_{\mu(Z)} \left[\!\!\left[\boxed{Z} \overset{p}{\curlywedge} \boxed{X} \overset{x}{\twoheadleftarrow} \right]\!\!\right]\Big(\delta_x(X)\mu(Z)\Big)$$

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

□

Proposition 5. *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,*

$$
\ell(p;\underline{\mathbf{x}}) = \left\langle\!\!\left\langle \; Z \overset{p}{\leftarrow} X \leftarrow \Pr_{\underline{\mathbf{x}}}! \; \right\rangle\!\!\right\rangle + \mathrm{H}(\Pr_{\underline{\mathbf{x}}}).
$$

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

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

□

Proposition 6. *Consider a probabilistic model* f *with mass function* $f(y \mid x)$*. The inconsistency of the PDG containing* $p(y \mid x)$ *and the emperical distribution* $\Pr_{\underline{\mathbf{xy}}}$ *of samples* $\underline{\mathbf{xy}}$ *is equal to the negative*

log-liklihood (cross-entropy) loss, plus the emperical uncertainty in Y given X (a constant independent of f). That is,

$$\left\langle\!\!\left\langle \; \begin{array}{ccc} & \overset{\Pr_{\underline{\mathbf{xy}}}!}{\swarrow\;\searrow} & \\ X & \xrightarrow[f]{} & Y \end{array} \; \right\rangle\!\!\right\rangle = \frac{1}{|\underline{\mathbf{xy}}|} \sum_{(x,y)\in\underline{\mathbf{xy}}} \left[\log \frac{1}{f(y \mid x)}\right] \quad - \mathrm{H}_{\Pr_{\underline{\mathbf{xy}}}}(Y \mid X).$$

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 7. *If $h : X \to Y$ is a predictor for an input space X and label space Y, and $f : X \to Y$ generates the correct labels, then the inconsistency of believing f and h (with any degree of confidence), and also that inputs are distributed according to $D(X)$, equals the information content of learning that $f(X) = h(X)$ according to D (which is the log accuracy of the predictor h) times the confidence in D. That is,*

$$\left\langle\!\!\left\langle \; \begin{array}{ccc} \overset{D^{(\beta)}}{\longrightarrow} & \overset{h}{\nearrow\!\!\searrow} & \\ X & & Y \\ & \underset{f}{\searrow\!\!\nearrow} & \end{array} \; \right\rangle\!\!\right\rangle = -\beta \log\Big(\mathrm{accuracy}_{f,D}(h)\Big) = \beta \, \mathrm{I}_D[f = h]. \tag{3}$$

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

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

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 $\boldsymbol{D}(\mu \parallel D)$, scaled by the confidence β of the emperical distribution D.

So, among those distributions $\mu(X)$ supported on an event $E \subset \mathcal{V}(X)$, which minimizes is the relative entropy of $\boldsymbol{D}(\mu \parallel D)$? 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, [**?**]). Let $f=h$ denote the event that f and h agree. Then we calculate

$$\left\langle\!\!\left\langle \overset{(\beta)}{D} \to X \underset{f}{\overset{h}{\rightrightarrows}} Y \right\rangle\!\!\right\rangle = \inf_{\substack{\mu(X) \text{ s.t.} \\ \operatorname{supp}(\mu)\subseteq[f=h]}} \beta \boldsymbol{D}\Big(\mu(X) \,\Big\|\, D(X)\Big)$$

$$= \beta \boldsymbol{D}\Big(D \mid [f=h] \,\Big\|\, D\Big)$$

$$= \beta \mathop{\mathbb{E}}_{D|f=h} \log \frac{\delta_{f=h}(X)D(X)}{D(f=h)\cdot D(X)}$$

$$= \beta \mathop{\mathbb{E}}_{D|f=h} \log \frac{1}{D(f=h)} \qquad \left[\begin{array}{c}\text{since } \delta_{f=h}(x)=1 \text{ for all } x \text{ that} \\ \text{contribute to the expectation}\end{array}\right]$$

$$= -\beta \log D(f=h) \qquad \left[\text{ since } D(f=h) \text{ is a constant }\right]$$

$$= -\beta \log\Big(\text{accuracy}_{f,D}(h)\Big)$$

$$= \beta \, \mathrm{I}_D[f=h].$$

□

Proposition 8. *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*

$$\left\langle\!\!\left\langle D! \to X \underset{\substack{t \\ s \\ h}}{\overset{f}{\rightrightarrows}} \begin{matrix}\mu_1 \\ \sigma_1 \\ \sigma_2 \\ \mu_2\end{matrix} \overset{(\beta:\beta_1)\ \mathcal{N}}{\underset{\mathcal{N}\ (\beta:\beta_2)}{\rightrightarrows}} Y \right\rangle\!\!\right\rangle = \frac{1}{2}\,\mathbb{E}_D\left[\mathrm{HM}(\beta_1,\beta_2)\frac{1}{2}\left(\frac{\mu_1-\mu_2}{\mathrm{QM}_{\hat{\beta}}(\sigma_1,\sigma_2)}\right)^2 + \mathrm{AM}(\beta_1,\beta_2)\log\frac{\mathrm{QM}_{\hat{\beta}}(\sigma_1,\sigma_2)}{\mathrm{GM}_{\hat{\beta}}(\sigma_1,\sigma_2)}\right] \tag{4}$$

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

where $\hat{\beta} = (\frac{\beta_2}{\beta_1+\beta_2}, \frac{\beta_1}{\beta_1+\beta_2})$ represents the normalized and reversed vector of confidences $\beta=(\beta_1,\beta_2)$ 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,

$$\langle\!\langle \mathcal{M} \rangle\!\rangle_0 = \inf_{\mu} \mathop{\mathbb{E}}_{x\sim D} \mathop{\mathbb{E}}_{y\sim\mu|x}\left[\beta_1 \log\frac{\mu(y\mid x)}{\mathcal{N}(y\mid f(x),t(x))} + \beta_2\log\frac{\mu(y\mid x)}{\mathcal{N}(y\mid h(x),s(x))}\right]$$

$$= \inf_{\mu} \mathop{\mathbb{E}}_{x\sim D} \mathop{\mathbb{E}}_{y\sim\mu|x}\left[\beta_1 \log\frac{\mu(y\mid x)}{\frac{1}{t(x)\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y-f(x)}{t(x)}\right)^2\right)} + \beta_2\log\frac{\mu(y\mid x)}{\frac{1}{s(x)\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y-h(x)}{s(x)}\right)^2\right)}\right]$$

$$= \inf_{\mu} \mathop{\mathbb{E}}_{x\sim D} \mathop{\mathbb{E}}_{y\sim\mu|x}\left[\log\mu(y\mid x)^{\beta_1+\beta_2} \begin{matrix} +\beta_1\log(t(x)\sqrt{2\pi}) + \frac{\beta_1}{2}\left(\frac{y-f(x)}{t(x)}\right)^2 \\ +\beta_2\log(t(x)\sqrt{2\pi}) + \frac{\beta_2}{2}\left(\frac{y-h(x)}{s(x)}\right)^2 \end{matrix}\right]$$

Todo: □

Proposition 14. *The negative ELBO is the inconsistency of the PDG containing p, q, and x, with very high confidence in q. That is,*

$$-\mathrm{ELBO}_{p,q}(x) = \left\langle\!\!\left\langle \xrightarrow{q!} \boxed{Z} \overset{p}{\curvearrowright} \boxed{X} \xleftarrow{x} \right\rangle\!\!\right\rangle.$$

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,

$$\begin{aligned}
\left\langle\!\!\left\langle \xrightarrow{q!} \boxed{Z} \overset{p}{\curvearrowright} \boxed{X} \xleftarrow{x} \right\rangle\!\!\right\rangle &= \inf_{\mu} \left[\!\!\left[\xrightarrow{q!} \boxed{Z} \overset{p}{\curvearrowright} \boxed{X} \xleftarrow{x} \right]\!\!\right](\mu) \\
&= \left[\!\!\left[\xrightarrow{q!} \boxed{Z} \overset{p}{\curvearrowright} \boxed{X} \xleftarrow{x} \right]\!\!\right](\delta_x(X)q(Z)) \\
&= \mathop{\mathbb{E}}_{\substack{x' \sim \delta_x \\ z \sim q}} \log \frac{\delta_x(x')q(z)}{p(x', z)} = -\mathop{\mathbb{E}}_{z \sim q} \frac{p(x, z)}{q(z)} = -\mathrm{ELBO}_{p,q}(x).
\end{aligned}$$

□

We proove both Proposition 15 and Proposition 16 at the same time.

Proposition 15. *The conditional ELBO used as a VaE objective is the inconsistency of the PDG containing the encoder e, decoder d prior p, and a sample x. That is,*

$$-\mathrm{ELBO}_{p,e,d}(x) = \left\langle\!\!\left\langle \xrightarrow{p} \boxed{Z} \underset{e!}{\overset{d}{\rightleftarrows}} \boxed{X} \xleftarrow{x} \right\rangle\!\!\right\rangle$$

Proposition 16. *The following analog of Proposition 15 for a whole dataset $\underline{\mathbf{x}}$ holds:*

$$-\mathop{\mathbb{E}}_{\mathrm{Pr}_{\underline{\mathbf{x}}}} \mathrm{ELBO}_{p,e,d}(X) = \left\langle\!\!\left\langle \xrightarrow{p} \boxed{Z} \underset{e!}{\overset{d}{\rightleftarrows}} \boxed{X} \xleftarrow{\mathrm{Pr}_{\underline{\mathbf{x}}}!} \right\rangle\!\!\right\rangle + \mathrm{H}(\mathrm{Pr}_{\underline{\mathbf{x}}}).$$

Proof. The two proofs are similar. For Proposition 15, the optimal distribution must be $\delta_x(X)e(Z \mid X)$, and for Proposition 16, it must be $\mathrm{Pr}_{\underline{\mathbf{x}}}(X)e(Z \mid X)$, 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

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

in the second. □

Proposition 17. *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,*

$$-\beta\text{-ELBO}_{p,e,d}(x) = \left\langle\!\!\left\langle \quad \overset{(\beta)}{\underset{p}{\longrightarrow}} \boxed{Z} \underset{e!}{\overset{d}{\rightleftarrows}} \boxed{X} \overset{x}{\twoheadleftarrow} \quad \right\rangle\!\!\right\rangle$$

Proof.

$$\begin{aligned}\left\langle\!\!\left\langle \quad \overset{(\beta:\beta')}{\underset{p}{\longrightarrow}} \boxed{Z} \underset{e!}{\overset{d}{\rightleftarrows}} \boxed{X} \overset{x}{\twoheadleftarrow} \quad \right\rangle\!\!\right\rangle &= \inf_{\mu}\left[\!\!\left[\quad \overset{(\beta:\beta')}{\underset{p}{\longrightarrow}} \boxed{Z} \underset{e!}{\overset{d}{\rightleftarrows}} \boxed{X} \overset{x}{\twoheadleftarrow} \quad \right]\!\!\right](\mu) \\ &= \inf_{\mu} \mathop{\mathbb{E}}_{\mu(X,Z)}\left[\beta\log\frac{\mu(Z)}{p(Z)} + \log\frac{\mu(X,Z)}{\mu(Z)d(X \mid Z)}\right]\end{aligned}$$

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

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

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

Recall the statement of Prop 4.6 from the original paper,

$$[\![\boldsymbol{\mathcal{M}}]\!]_\gamma(\mu) = \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu}\left\{\sum_{X\xrightarrow{L}Y}\left[\beta_L\log\frac{1}{\mathbf{p}_L(y^{\mathbf{w}}|x^{\mathbf{w}})} + (\gamma\alpha_L - \beta_L)\log\frac{1}{\mu(y^{\mathbf{w}}|x^{\mathbf{w}})}\right] - \gamma\log\frac{1}{\mu(\mathbf{w})}\right\}, \tag{9}$$

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 Gibbs free energy $\mathit{GFE}(\mu)$. 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 $\langle\!\langle (\mathcal{N}_\Phi, \theta, \theta) \rangle\!\rangle_1$.

$$
\begin{aligned}
\langle\!\langle (\mathcal{N}_\Phi, \theta, \theta) \rangle\!\rangle_1 &= [\![(\mathcal{N}_\Phi, \theta, \theta)]\!]_1 \Big(\Pr_{\Phi,\theta}(\mathbf{w}) \Big) \\
&= \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu} \left\{ \sum_{X \xrightarrow{L} Y} \left[\beta_L \log \frac{1}{\mathbf{p}_L(y^{\mathbf{w}} | x^{\mathbf{w}})} \right] - \log \frac{1}{\Pr_{\Phi,\theta}(\mathbf{w})} \right\} && \Big[\text{ by (9) } \Big] \\
&= \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu} \left\{ \sum_j \left[\theta_j \log \frac{1}{\phi_j(\mathbf{w}_j)} \right] - \log \frac{Z_\Psi}{\prod_j \phi_j(\mathbf{w}_j)^{\theta_j}} \right\} && \Big[\begin{array}{c} \text{cpds } \mathbf{p}_L \text{ correspond} \\ \text{to factors } \phi_j \end{array} \Big] \\
&= \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu} \left\{ \sum_j \left[\theta_j \log \frac{1}{\phi_j(\mathbf{w}_j)} \right] - \sum_j \left[\theta_j \log \frac{1}{\phi_j(\mathbf{w}_j)} \right] - \log Z_\Psi \right\} \\
&= \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu} [-\log Z_\Psi] \\
&= -\log Z_\Psi && \Big[Z_\Psi \text{ is constant in } \mathbf{w} \Big]
\end{aligned}
$$

□