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 or pragmatics, but of model. Probabilistic depencency graphs (PDGs) are probabilistic models that come equipped with a measure of "inconsistency". We prove that many standard loss functions arise as the inconsistency of a natural PDG describing the appropriate scenario, and use the same approach to justify a well-known connection between regularizers and priors. We also show that the PDG inconsistency captures a large class of statistical divergences, and detail benefits of thinking of them in this way, including an intuitive visual language for deriving inequalities between them. In variational inference, we find that the ELBO, a somewhat opaque objective for latent variable models, and variants of it arise for free out of uncontroversial modeling assumptions—as do simple graphical proofs of their corresponding bounds. 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 must minimize. A wide variety of losses are used in pratice. Each implicitly represents different values and results in different behavior, so the choice between them 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 empirics, tradition, and an instinctive calculus acquired through the practice—not by 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 the process of tinkering with the objective until it works is often unsatisfying. It can be a tedious game without clear rules or meaning, while results so obtained are arguably overfitted and difficult to motivate.

By contrast, a choice of *model* admits more principled discussion, in part because models are testable; it makes sense to ask if a model is accurate. This observation motivates our proposal: instead of specifying a loss function directly, one articulates 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 of graphical models that can incorporate arbitrary (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 provide 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 clearer discussion in a more common language.

For a particularly powerful demonstration, consider the variational autoencoder (VAE), an enormously

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

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 $d(X|Z)$, 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, originating in variational calculus. We show that $-\text{ELBO}(x)$ is also precisely the inconsistency of a PDG containing x and the probabilistic information of the autoencoder (p, d, and e). We can form such a PDG precisely because PDGs allow for inconsistency. Thus, PDG semantics simultaneously legitimize the strange structure of 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 *structural* relationships between the underlying models. For instance, we will be able to use this underlying structure to get simple and intuitive visual 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 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 and to measures of statistical distance. We demonstrate that this framework clarifies the relationships between them, by providing simple and 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 $p(Y|X=x)$ or $p(Y|\,x)$ 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 a PDG. For instance, a PDG of shape $X \to Y \leftarrow Z$ contains both a cpd $p(Y|X)$ and (separately) 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 and 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 real number β_L, the modeler's subjective confidence in the reliability of the cpd $\mathbf{p}_L$. □

We conflate a cpd's symbol with its edge label, so we draw the PDG with a single edge attached to $f(Y|X)$ as $X \xrightarrow{f} Y$. Definition 1 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 draw $p(Y|X,Z)$ as $\begin{matrix} Z \\ X \end{matrix} \overset{p}{\rightarrowtail} Y$, and $q(A,B)$ as $\overset{q}{\downarrow}\begin{matrix} A & B \end{matrix}$.

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 $\mathcal{M}$'s cpds, and is given by

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

It is well known that $\boldsymbol{D}(\mu \| p)$ is a measure of divergence between μ and p, that can be viewed as the 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 excess cost of using $\mathcal{M}$'s cpds in proportion to the confidence the modeler has in them, in expectation over μ.

The *inconsistency* of $\mathcal{M}$, denoted $\langle\!\langle \mathcal{M} \rangle\!\rangle$, is the smallest possible incompatibility of $\mathcal{M}$ with any distribution:

$$\langle\!\langle \mathcal{M} \rangle\!\rangle := \inf_{\mu} \mathit{Inc}_{\mathcal{M}}(\mu) \quad \left(= \inf_{\mu} \mathop{\mathbb{E}}_{\mu} \sum_{X \xrightarrow{L} Y} \beta_L \log \frac{\mu(Y|X)}{\mathbf{p}_L(Y|X)} \right).$$

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

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

$\mathit{IDef}_{\mathcal{M}}(\mu)$ can be thought of as the information needed to separately describe the target of each edge L given its source (α_L times), beyond the information needed to fully describe a sample from 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), while the distribution specfied by a factor graph Φ uniquely minimizes the sum $\mathit{Inc}_{\Phi} + \mathit{IDef}_{\Phi}$. In general, for any $\gamma > 0$, 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. 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: $\boxed{X} \text{-}f\!\twoheadrightarrow \boxed{Y}$. We identify an event $X{=}x$ with the degenerate unconditional distribution $\delta_x(X)$ that places all mass on x; hence it may be associated to an edge and drawn simply as $\xrightarrow{x}\!\!\twoheadrightarrow \boxed{X}$. By default, edges have $\alpha = 0$ and $\beta = 1$. To specify a confidence $\beta \neq 1$, we place the value near the edge, lightly colored and parenthesized, as in: $\text{-}\xrightarrow[(\beta)]{p}\, \boxed{X}$. Finally, we write $_{(\infty)}$ 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 cpds or increasing confidences cannot decrease a PDG's inconsistency.

Lemma 1 (Monotonicity of $\langle\!\langle \cdot \rangle\!\rangle$)**.** *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]

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

3 STANDARD METRICS AS INCONSISTENCIES

Suppose you believe that X is distributed according to $p(X)$, and also that it (certainly) equals some value x. These beliefs are consistent if $p(X{=}x) = 1$ but become less so as $p(X{=}x)$ decreases. In fact, this inconsistency is equal to the information content $\mathrm{I}_p[X{=}x] := -\log p(X{=}x)$, or *surprisal* (Tribus 1961), of the event $X{=}x$, according to p.[2] In machine learning, I_p is usually 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 inconsistency of the PDG comprising p and $X = x$ equals the surprisal $\mathrm{I}_p[X{=}x]$. That is,*

$$\mathrm{I}_p[X{=}x] = \Big\langle\!\!\Big\langle \xrightarrow{p} \boxed{X} \twoheadleftarrow\!\!\!\text{-}^{x} \Big\rangle\!\!\Big\rangle.$$

(Recall that $\langle\!\langle \mathcal{M} \rangle\!\rangle$ is the inconsistency of the PDG $\mathcal{M}$.)

In some ways, this result is entirely unsurprising, given that (1) is a flexible formula built out of information theoretic primitives. Even so, note that the inconsistency of believing both a distribution and an event happens to be the standard measure of discrepency between the two—and is even named after "surprise", a particular 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 probabilistic models with more than one sample. What if we replace x with an empirical distribution over many samples?

Proposition 3. *If $p(X)$ is a probabilistic model of X, and $\mathcal{D} = \{x_i\}_{i=1}^m$ is a dataset with empirical distribution $\mathrm{Pr}_{\mathcal{D}}$, then* $\mathrm{CrossEntropy}(\mathrm{Pr}_{\mathcal{D}}, p) =$

$$\frac{1}{m}\sum_{i=1}^{m} \mathrm{I}_p[X{=}x_i] = \Big\langle\!\!\Big\langle \xrightarrow{p} \boxed{X} \xleftarrow[(\infty)]{\mathrm{Pr}_{\mathcal{D}}} \Big\rangle\!\!\Big\rangle + \mathrm{H}(\mathrm{Pr}_{\mathcal{D}}).$$

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

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

[2] This construction requires the event $X{=}x$ to be measurable. One can get similar, but subtler, results for densities, where this is not the case; see Appendix A.

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

Now consider an orthogonal generalization of Proposition 2, in which the sample x is only a partial observation of (x, z) from a joint model $p(X, Z)$.

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

$$\mathrm{I}_p[X{=}x] = \left\langle\!\!\left\langle \begin{array}{ccc} & \overset{p}{\swarrow\!\searrow} & \\ Z & & X \overset{x}{\twoheadleftarrow} \end{array} \right\rangle\!\!\right\rangle. \tag{2}$$

Intuitively, the inconsistency of the PDG on the right side of (2) is localized to X, where the observation x conflicts with $p(X)$; other variables don't make a difference. The multi-sample partial-observation generalization also holds; see Appendix B.3.

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*; the process of training them is called *unsupervised* learning (Hastie, Tibshirani, and Friedman 2009). In the (more common) *supervised* setting, we train *discriminative* models to predict Y from X, via labeled samples $\{(x_i, y_i)\}_i$. There, cross entropy loss is perhaps even more dominant—and it is essentially the inconsistency of a PDG consisting of the predictor $h(Y|X)$ together with high-confidence data.

Proposition 5 (Cross Entropy, Supervised)**.** *The inconsistency of the PDG comprising a probabilistic predictor $h(Y|X)$, and a high-confidence empirical distribution $\Pr_{\mathcal{D}}$ of a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m$ equals the cross-entropy loss (minus the empirical uncertainty in Y given X, a constant depending only on $\mathcal{D}$). That is,*

$$\left\langle\!\!\left\langle \begin{array}{ccc} & \Pr_{\mathcal{D}} \overset{(\infty)}{\swarrow\!\searrow} & \\ X & \xrightarrow{\quad h \quad} & Y \end{array} \right\rangle\!\!\right\rangle = \frac{1}{m}\sum_{i=1}^{m} \log \frac{1}{h(y_i \,|\, x_i)} - \mathrm{H}_{\Pr_{\mathcal{D}}}(Y|X).$$

Simple evaluation metrics, such as the accuracy of a classifier, or the mean squared error of a regressor, also arise naturally as inconsistencies.

Proposition 6 (Log Accuracy as Inconsistency)**.** *Consider functions $f, h : X \to Y$ from inputs to labels, where h is a predictor and f generates the true labels. The inconsistency of believing f and h (with any confidences), and a distribution $D(X)$ with confidence β, is β times the log accuracy of h. That is,*

$$\left\langle\!\!\left\langle \begin{array}{ccccc} D & & \overset{h}{\overset{(r)}{\nearrow}} & & \\ \overset{\longrightarrow}{(\beta)} & X & & Y & \\ & & \underset{(s)}{\underset{f}{\searrow}} & & \end{array} \right\rangle\!\!\right\rangle \begin{array}{l} = -\beta \log \Pr_{x\sim D}(f(x){=}h(x)) \\ = \beta\, \mathrm{I}_D[f = h]. \end{array} \tag{3}$$

One often speaks of the accuracy of a hypothesis h, leaving the true labels f and empirical distribution D implicit. Yet there is a sense in which $D(X)$ 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 same is true for h, so we can only consider μ such that $\mu(f{=}h){=}1$, a restriction which in turn generates inconsistency equal to D's surprisal that h is correct. In other words, the optimal distribution μ^* throws out incorrect samples, so its conditional $\mu^*(Y|x)$ is undefined unless $h(x)$ is already correct. This illustrates why accuracy gives no gradient information for training h (only for D). This is precisely the opposite of how cross entropy played out in Proposition 5: there we were unwilling to budge on either the true labels or input distribution, and the optimal distribution told us how to modify h.

Observe how even properties of these simple metrics—the lack of gradient information, and relationships to other metrics—are clarified by an underlying model.

When $Y \cong \mathbb{R}^n$, an estimator $h(Y|X)$ is referred to as a regressor instead of a classifier. In this setting, most answers are incorrect, but some more so than others. A common way of measuring incorrectness is with mean squared error (MSE): $\mathbb{E}\,|f(X) - Y|^2$. MSE is also the inconsistency of believing that the labels and predictor have Gaussian noise—often a reasonable assumption because of the central limit theorem.

Proposition 7 (MSE as Inconsistency)**.**

$$\left\langle\!\!\left\langle \begin{array}{ccccc} & & \overset{f}{\nearrow} & \mu_f & \overset{\mathcal{N}_1}{\searrow} \\ D & X & & & Y \\ \overset{\longrightarrow}{(\infty)} & & \underset{h}{\searrow} & \mu_h & \underset{\mathcal{N}_1}{\nearrow} \end{array} \right\rangle\!\!\right\rangle \begin{array}{l} = \mathbb{E}_D\big|f(X) - h(X)\big|^2 \\ =: \mathrm{MSE}_D(f, h)\,, \end{array}$$

where $\mathcal{N}_1(Y|\,\mu)$ is a unit Gaussian on Y with mean μ.

In the appendix, we treat general Gaussian predictors, with arbitrary variances and confidences.

4 REGULARIZERS AND PRIORS

Regularizers are extra terms added to loss funtions, which provide a source of inductive bias towards simple

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 the ability to make principled modeling choices about regularizers is a primary benefit of this correspondence. Our approach provides a new justification of it.

Proposition 8. *Suppose you have a parameterized model $p(Y|\Theta)$, a prior $q(\Theta)$, and a trusted 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,*

$$\left\langle\!\!\left\langle \begin{array}{ccccc} & \overset{q}{\searrow} & & & \\ \overset{}{\underset{(\beta)}{}} & & \boxed{\Theta} & \overset{p}{\rightarrow} & \boxed{Y} \\ & \underset{\theta}{\searrow} & & & \uparrow \overset{}{\underset{(\infty)}{D}} \end{array} \right\rangle\!\!\right\rangle = \mathop{\mathbb{E}}_{y\sim D} \log \frac{1}{p(y\,|\,\theta)} + \beta \log \frac{1}{q(\theta)} - \mathrm{H}(D) \tag{4}$$

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

$$\underbrace{\mathbb{E}_D \log \frac{1}{p(Y\,|\,\theta)}}_{\substack{\text{Cross entropy loss} \\ \text{(data-fit cost of } \theta\text{)}}} + \underbrace{\frac{\beta}{2}\theta_0}_{\substack{\text{L2 regularizer} \\ \text{(complexity cost of } \theta\text{)}}} + \underbrace{\beta \log k - \mathrm{H}(D)}_{\text{constant in } p \text{ and } \theta},$$

which is the L2 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) unit Laplacian prior, $q(\theta) \propto \exp(-|\theta|)$, the second term instead becomes $\beta|\theta_0|$, which is L1 regularization. More generally, to consider a complexity measure $U(\theta)$, we need only include the Gibbs distribution $\Pr_U(\theta) \propto \exp(-U(\theta))$ into our PDG. We remark that nothing here is specific to cross entropy; 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, so we can form a PDG divergence $\boldsymbol{D}^{\mathrm{PDG}}_{(r,s)}(p\|q)$ for every setting (r, s) 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.

[3] A full account can be found in the appendix.

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,*

$$\left\langle\!\!\left\langle \overset{p}{\underset{(\infty)}{\rightarrow}} \boxed{X} \overset{q}{\underset{(\beta)}{\leftarrow}} \right\rangle\!\!\right\rangle = \beta\, \boldsymbol{D}(p \parallel q).$$

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 it is usually called 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.

Lemma 10. *The inconsistency $\boldsymbol{D}^{\mathrm{PDG}}_{(r,s)}(p\|q)$ of a PDG comprising $p(X)$ with confidence r and $q(X)$ with confidence s is given in closed form by*

$$\left\langle\!\!\left\langle \overset{p}{\underset{(r)}{\rightarrow}} \boxed{X} \overset{q}{\underset{(s)}{\leftarrow}} \right\rangle\!\!\right\rangle = -(r+s) \log \sum_x \left(p(x)^r q(x)^s\right)^{\frac{1}{r+s}}.$$

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 every Rényi divergence is a PDG divergence, and every (non-limiting) PDG divergence is a (scaled) Rényi divergence.

Corollary 10.1 (Rényi Divergences)**.**

$$\left\langle\!\!\left\langle \overset{p}{\underset{(r)}{\rightarrow}} \boxed{X} \overset{q}{\underset{(s)}{\leftarrow}} \right\rangle\!\!\right\rangle = s \cdot \boldsymbol{D}_{\frac{r}{r+s}}(p \parallel q)$$

$$\textit{and} \qquad \boldsymbol{D}_\alpha(p \parallel q) = \left\langle\!\!\left\langle \overset{p}{\underset{(\frac{\alpha}{1-\alpha})}{\rightarrow}} \boxed{X} \overset{q}{\leftarrow} \right\rangle\!\!\right\rangle$$

However, the two classes are not identical, because the PDG divergences have extra limit points. One big difference is that the reverse KL divergence $\boldsymbol{D}(q \parallel p)$

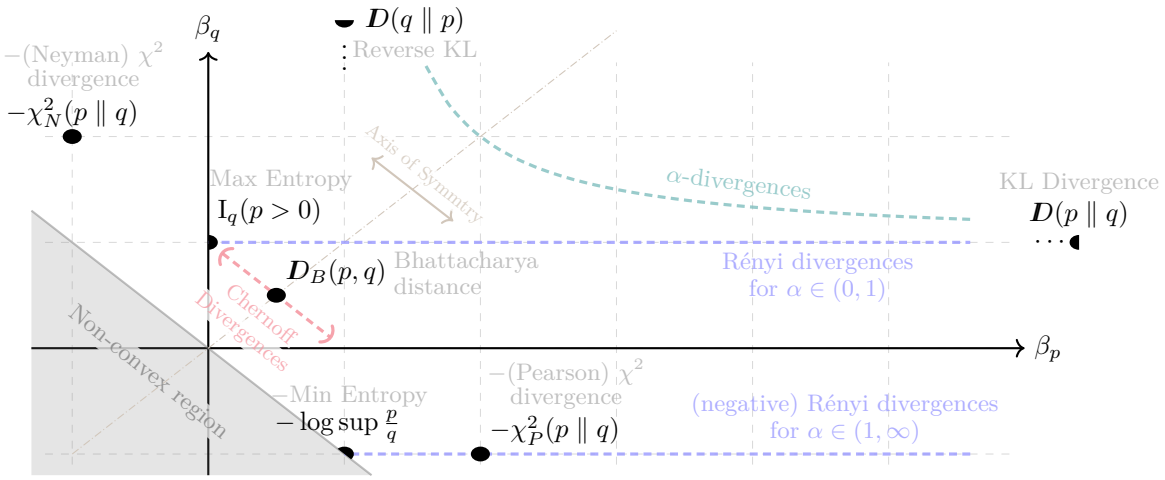


Figure 1: A map of the inconsistency of the PDG comprising $p(X)$ and $q(X)$, as we vary their respective confidences β_p and β_q. Solid circles indicate well-known named measures, semicircles indicate limiting values, and the heavily dashed lines are well-established classes.

is not a Rényi divergence $\boldsymbol{D}_\alpha(p \parallel q)$ for any value (or limit) of α. This lack of symmetry has led others (e.g., Cichocki and Amari 2010) to work instead with a symmetric variant called α-divergence, rescaled by 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 total confidence 1.

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

One significant consequence of representing divergences as inconsistencies is that we can use Lemma 1 to derive relationships between them. The following facts follow directly 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 $\boldsymbol{D}(q \parallel p) = \infty$.*

These divergences correspond to PDGs with only two edges and one variable. What about more complex graphs? For a start, the usual notion of a conditional divergence $\boldsymbol{D}^{\mathrm{PDG}}_{r,s}(p(Y|X) \parallel q(Y|X) \,|\, r(X)) := \mathop{\mathbb{E}}_{x \sim r} \boldsymbol{D}^{\mathrm{PDG}}_{r,s}(p(Y|x) \parallel q(Y|x))$ falls out of PDGs of the form

$$\xrightarrow[(\infty)]{r} X \overset{p}{\underset{q}{\rightrightarrows}} Y \,.$$

Such graphs are also useful intermediates. Lemma 1, plus some structural manipulation, yields visual proofs of many divergence properties; a proof of the data-processing inequality can be seen in Figure 2. In general, PDG inconsistency can be viewed as a vast generalization of divergences to arbitrary structured objects.

6 VARIATIONAL OBJECTIVES AND BOUNDS

The fact that the incompatibility of $\mathcal{M}$ with a *specific* joint distribution μ is an upper bound on the inconsistency is not a deep one, but it is of a variational flavor. Here, we focus on the more surprising converse: PDG semantics capture general aspects of variational inference and provide a graphical proof language for it.

6.1 PDGs and Variational Approximations

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). Suppose we have a model $p(X, Z)$, but only have access to observations of x. In service of adjusting $p(X, Z)$ to make our observations more likely, we would like to maximize $\log p(X{=}x)$, the "evidence" of x (Proposition 4). Unfortunately, computing $p(X) = \sum_z p(X, Z{=}z)$ 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 $\mathrm{ELBO}_{p,q}(x) := \mathbb{E}_{z \sim q} \log \frac{p(x,z)}{q(z)}$. This is something we can estimate, since we can sample from q. By Jensen's inequality,

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

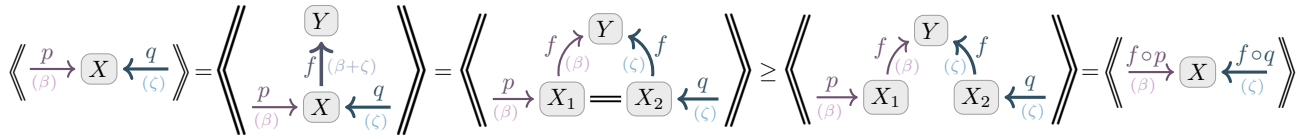


Figure 2: A visual proof of the data-processing inequality: $\boldsymbol{D}^{\mathrm{PDG}}_{(\beta,\zeta)}(p \parallel q) \geq \boldsymbol{D}^{\mathrm{PDG}}_{(\beta,\zeta)}(f \circ p \parallel f \circ q)$. In words: the cpd $f(Y|X)$ can always be satisfied, so adds no inconsistency. It is then equivalent 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 inconsistency. Finally, compose the remaining cpds to give the result. A full justification can be found in the appendix.

with equality if $q(Z) = p(Z)$. So to maximize $p(X)$, it suffices to adjust p and q to maximize $\mathrm{ELBO}_{p,q}(x)$,[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.

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,*

$$-\mathrm{ELBO}_{p,q}(x) = \left\langle\!\!\left\langle \; \xrightarrow[(\infty)]{q} Z \quad \overset{p}{\swarrow\!\searrow} \quad X \xleftarrow{x} \; \right\rangle\!\!\right\rangle.$$

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:

$$\log \frac{1}{p(x)} = \left\langle\!\!\left\langle \; Z \quad \overset{p}{\swarrow\!\searrow} \quad \overset{x}{\swarrow} X \; \right\rangle\!\!\right\rangle \leq \left\langle\!\!\left\langle \; \overset{q}{\circlearrowleft} Z \quad \overset{p}{\swarrow\!\searrow} \quad \overset{x}{\swarrow} X \; \right\rangle\!\!\right\rangle = -\underset{p,q}{\mathrm{ELBO}}(x).$$

The first and last equalities are Propositions 4 and 11 respectively. Now to 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 cost.

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. One important consideration is the *reconstruction error* (6): when we decode an encoded image, we would like it to resemble the original.

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

There are other desiderata as well. Perhaps good latent representations Z have uncorrelated components, and are normally distributed. We encode such 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 analog of the ELBO becomes

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

This gives us the following analog of Proposition 11.

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

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

We now give a visual proof of the analogous variational bound. Let $\Pr_{p,d}(X, Z) := p(Z)d(X|Z)$ be the distribution that arises from decoding the prior. Then:

$$\log \frac{1}{\underset{p,d}{\Pr}(x)} = \left\langle\!\!\left\langle \; \overset{p}{\downarrow} Z \quad \overset{d}{\curvearrowright} \quad \overset{x}{\downarrow} X \; \right\rangle\!\!\right\rangle \leq \left\langle\!\!\left\langle \; \overset{p}{\downarrow} Z \quad \overset{d}{\underset{e}{\curvearrowright}} \quad \overset{x}{\downarrow} X \; \right\rangle\!\!\right\rangle = \underset{p,e,d}{\text{-ELBO}}(x).$$

The first and last equalities are Propositions 4 and 12, and the inequality is Lemma 1. See the appendix for multi-sample analogs of the bound and Proposition 12.

6.3 The β-VAE Objective

The ELBO is not the only objective that has been used to train networks with a VAE structure. In the

[4] For many iid samples: $\max_{p,q} \sum_{x\in\mathcal{D}} \mathrm{ELBO}_{p,q}(x)$.
[5] Especially if p and q are densities. See Appendix A

most common variant, due to Higgins et al. (2016), one weights the reconstruction error (6) and the 'KL term' differently, resulting in a loss function of the form

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

which, when $\beta = 1$, is the ELBO as before. The authors view β as a regularization strength, and argue that it sometimes helps to have a stronger prior. Sure enough, the β-VAE objective is the inconsistency of the same PDG as before, but with confidence β in $p(Z)$.

7 FREE ENERGY AND INCONSISTENCY

A weighted factor graph $\Psi = (\phi_J, \theta_J)_{J \in \mathcal{J}}$, where each θ_J is a real-valued weight, J is associated with a subset of variables $\mathbf{X}_J$, and $\phi_J : \mathcal{V}(\mathbf{X}_J) \to \mathbb{R}$, determines a distribution by

$$\Pr_\Psi(\mathbf{x}) = \frac{1}{Z_\Psi} \prod_{J \in \mathcal{J}} \phi_J(\mathbf{x}_J)^{\theta_J}.$$

Z_Ψ is the constant $\sum_{\mathbf{x}} \prod_{J \in \mathcal{J}} \phi_J(\mathbf{x}_J)^{\theta_J}$ required to normalize the distribution, and is known as the *partition function*. Computing $\log Z_\Psi$ is intimately related to much of probabilistic inference in factor graphs (Ma et al. 2013). Following Richardson and Halpern (2021), let $\mathcal{M}_\Psi$ be the PDG with edges $\{\xrightarrow{J} \mathbf{X}_J\}_{\mathcal{J}}$, cpds $p_J(\mathbf{X}_J) \propto \phi_J(\mathbf{X}_J)$, and weights $\alpha_J, \beta_J := \theta_J$.

If the factors are normalized and all variables are edge targets, then $Z_\Psi \leq 1$, so $\log \frac{1}{Z_\Psi} \geq 0$ measures how far the product of factors is from being a probability distribution. So in a sense, it measures Ψ's inconsistency.

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

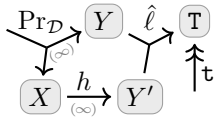
The exponential families generated by weigted factor graphs are a cornerstone of statistical mechanics, where $-\log Z_\Psi$ is known as the (Heimholz) free energy. It is also an especially natural quantity to minimize: the principle of free-energy minimization has been enormously succesful in describing of not only chemical and biological systems (Chipot and Pohorille 2007), but also cognitive ones (Friston 2009).

8 REVERSE-ENGINEERING LOSS?

Given an arbitrary loss function $\ell(X)$, can we find a PDG that gives rise to it? To a first approximation, the answer appears to be yes. Without affecting its semantics, one may add the variable $\mathtt{T}$ that takes values $\{\mathtt{t}, \mathtt{f}\}$, and the event $\mathtt{T}{=}\mathtt{t}$, to any PDG. Now, given a cost function $c : \mathcal{V}(X) \to \mathbb{R}^{\geq 0}$, define the cpd $\hat{c}(\mathtt{T}|X)$ by $\hat{c}(\mathtt{t}|x) := \exp(-c(x))$. By threatening to generate the falsehood $\mathtt{f}$ with probability dependent on the cost of X, $\hat{c}$ ties the value of X to inconsistency.

Proposition 14. $\left\langle\!\!\left\langle \xrightarrow[(\infty)]{p} X \xrightarrow{\hat{c}} \mathtt{T} \xleftarrow{\mathtt{t}} \right\rangle\!\!\right\rangle = \mathop{\mathbb{E}}_{x \sim p} c(x).$

Setting confidence $\beta_p := \infty$ may not be realistic since we're still training the model p, but doing so is necessary to recover $\mathbb{E}_p\, c$.[6] Any mechanism that generates inconsistency based on the value of X (such as this one) also works in reverse: the PDG squirms, contorting the probability of X to disperse the inconsistency. One cannot cannot simply "emit inconsistency" without affecting the probabilistic part of the model, as one does with value in an influence diagram (Howard 1983). Even setting every $\beta := \infty$ may not be enough to prevent the squirming. To illustrate, consider a model of the supervised learning setting, with data $\mathcal{D}$, model h, and an arbitrary loss function ℓ. Define:

$\mathcal{S} :=$ 

and $\quad L := \mathop{\mathbb{E}}_{\substack{(x,y)\sim \Pr_{\mathcal{D}} \\ y' \sim p(Y'|\,x)}} \left[\ell(y, y')\right].$

Given Proposition 14, one might imagine $\langle\!\langle \mathcal{S} \rangle\!\rangle = L$, but this is not so. In some ways, $\langle\!\langle \mathcal{S} \rangle\!\rangle$ is actually preferable. The optimal $h(Y'|X)$ according to L is a degenerate cpd that places all mass on the label y^*_X minimizing expected loss, while the optimal $h(Y'|X)$ according to $\langle\!\langle \mathcal{S} \rangle\!\rangle$ is $\Pr_{\mathcal{D}}(Y|X)$, which means it is *calibrated* (Dawid 1982). If, in addition, we set $\alpha_p, \alpha_{\Pr_{\mathcal{D}}} := 1$ and strictly enforce the qualitative picture, finally no more squirming is possible, as we arrive at $\lim_{\gamma \to \infty} \langle\!\langle \mathcal{S} \rangle\!\rangle_\gamma = L$.

To summarize: working backwards from loss to PDG, although possible, may require reporting absolute certainty in all modeling choices, including some questionable ones. In the end, we must confront our modeling choices; good loss functions come from good models.

9 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.

This "universal loss function", which provides a principled way of choosing a loss function, may be of particular interest to the AI safety community.

[6] If β_p were instead equal to 1, we would have obtained $-\log \mathbb{E}_p \exp(-c(X))$, with optimal distribution $\mu(X) \neq p(X)$.

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.

Chipot, Christophe and Andrew Pohorille (2007). "Free Energy Calculations." In: *Springer Series in Chemical Physics* 86, pp. 159–184.

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.

Dawid, A Philip (1982). "The Well-Calibrated Bayesian." In: *Journal of the American Statistical Association* 77.379, pp. 605–610.

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

Fagin, Ronald et al. (2003). *Reasoning about knowledge*. MIT press.

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

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

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second. Springer.

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

Howard Ronald A., James E. Matheson (1983). "Influence Diagrams." In: *Readings on the Principles and Applications of Decision Analysis*, pp. 719–763.

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 '21*. 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.