A THE FINE PRINT FOR PROBABILITY DENSITIES

Densities and Masses. Many of our results (Propositions 2 to 5, 11, 12, 16, 18 and 19) technically require the distribution to be represented with a mass function (not a density function (pdf)). A PDG containg both pdf and a finitely supported distribution on the same variable will typically have infinite inconsistency. But this is not just a quirk of the PDG formalism.

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

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

The bottom line is that all our results hold in a uniform way for every discretization size — yet in the limit as the discretization becomes smaller, an inconsistency may diverge to infinity. However, this divergence stems from an additive constant that depends only on the discretization size, which is irrelevant to its employment as a loss function. As a result, using one of these "unbalanced" functions involving densities where the units do not work out properly, results in a morally equivalent loss function, except without a diverging constant.

Markov Kernels. In the more general setting of measurable spaces, one may want to adjust the definition of a cpd that we gave, so that one instead works with *Markov Kernels*. This imposes an additional constraint: suppose the variable Y takes values in the measurable space $(\mathcal{V}(Y), \mathcal{B})$. If $p(Y|X)$ is to be a *Markov Kernel*, then for every fixed measurable subset $B \in \mathcal{B}$ of the measure space, the we must require that $x \mapsto \Pr(B|x)$ be a measurable function (with respect to the measure space in which X takes values). This too mostly does not bear on the present discussion, because the σ-algebras for all measure spaces of interest, are fine enough that one can get an arbitrarily close approximation of any cpd with a Markov Kernels. This means that the infemum defining the inconsistency of a PDG does not change.

B FURTHER RESULTS AND GENERALIZATIONS

B.1 Full Characterization of Gaussian Predictors

The inconsistency of a PDG containing two univariate Gaussian regressors of with arbitrary paremeters and confidences, is most cleanly articulated in terms of the geometric and quadratic means.

Definition 2 (Weighted Power Mean)**.** 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

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

We omit the superscript as a shorthand for the uniform weighting $w_i = {}^1\!/\!_N$. □

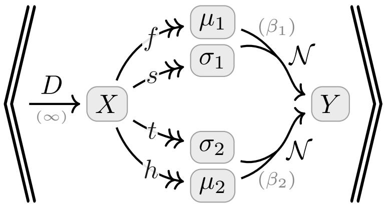
Many standard means, such as those in Table 1, are special cases. It is well known that $\mathrm{M}_p^w(\mathbf{r})$ is increasing in p, and strictly so if not all elements of $\mathbf{r}$ are identical. In particular, $\mathrm{QM}_w(a, b) > \mathrm{GM}_w(a, b)$ for all $a \neq b$ and positive weights w. We now present the result.

Proposition 15. *Consider a PDG containing two (distinct) conditional Gaussian distributions on a variable*

Name	p	Formula
Harmonic	$(p=-1)$:	$\mathrm{HM}_w(\mathbf{r}) = 1 \big/ \left(\sum_{i=1}^n w_i/r_i\right)$
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 $\mathrm{M}_p^w(\mathbf{r})$

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



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

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

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

The PDG on the left is semantically equivalent to (and in particular has the same inconsistency as) the PDG

$$\begin{tikzcd} {} \arrow[r, "D", "(\infty)"'] & X \arrow[r, bend left, "\mathcal{N}(f(x),s(x))"] \arrow[r, bend right, "\mathcal{N}(h(x),t(x))"'] & Y \end{tikzcd}.$$

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

As a bonus, Proposition 15 also gives a proof of the inequality of the weighted geometric and quadratic means.

Corollary 15.1. *For all σ_1 and σ_2, and all weight vectors β, $\mathrm{QM}_{\hat\beta}(\sigma_1,\sigma_2) \geq \mathrm{GM}_{\hat\beta}(\sigma_1,\sigma_2)$.*

B.2 Full-Dataset ELBO and Bounds

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

Proposition 16. *The following analog of Proposition 12 for a whole dataset $\mathcal{D}$ holds:*

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

Propositions 3 and 16 then give us an analog of the visual bounds in the body of the main paper (Section 6.1) for many i.i.d. datapoints at once, with only a single application of the inequality:

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

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

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

We also have the following formal statement of the claim made in Section 6.3.

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,*

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

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

Corollary 17.1 (reconstruction error as inconsistency)**.**

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

B.3 More Variants of Cross Entropy Results

First, we show that our cross entropy results hold for all γ, in the sense that γ contributes only a constant.

Proposition 18. *Given a model determining a probability distribution with mass function $p(X)$, and samples $\mathcal{D} = \{x_i\}_{i=1}^{m}$ determining an empirical distribution $\Pr_{\mathcal{D}}$, the following are equal, for all $\gamma \geq 0$:*

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

As promised, we now give the simultaneous generalization of the surprisal result (Proposition 2) to both multiple samples (like in Proposition 3 and partial observations (as in Proposition 4).

Proposition 19. *The average* marginal *negative log likelihood $\ell(p;x) := -\frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}} \log \sum_{z} p(x,z)$ is the inconsistency of the PDG containing p and the data distribution $\Pr_{\mathcal{D}}$, plus the entropy of the data distribution (which is constant in p). That is,*

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

C PROOFS

Lemma 1. *Suppose PDGs $\mathcal{M}$ and $\mathcal{M}'$ differ only in their edges (resp. $\mathcal{E}$ and $\mathcal{E}'$) and confidences (resp. β and β'). 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.*

Proof. For every μ, adding more edges only adds non-negative terms to (1), while increasing β results in larger coefficients on the existing (non-negative) terms of (1). So for every fixed distribution μ, we have $[\![\mathcal{M}]\!]_\gamma(\mu) \leq [\![\mathcal{M}']\!]_\gamma(\mu)$. So it must also be the case that the infemum over μ, so we find that $\langle\!\langle \mathcal{M} \rangle\!\rangle \leq \langle\!\langle \mathcal{M}' \rangle\!\rangle$. □

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,*

$$\mathrm{I}_p[X{=}x] = \left\langle\!\!\left\langle \xrightarrow{p} \boxed{X} \overset{x}{\twoheadleftarrow} \right\rangle\!\!\right\rangle.$$

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

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 18 is a generalization of Proposition 3, so we prove them at the same time.

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 $\Pr_\mathcal{D}$, then* $\mathrm{CrossEntropy}(\Pr_\mathcal{D}, p) =$

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

Proposition 18. *Given a model determining a probability distribution with mass function $p(X)$, and samples $\mathcal{D} = \{x_i\}_{i=1}^m$ determining an empirical distribution $\Pr_\mathcal{D}$, the following are equal, for all $\gamma \geq 0$:*

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

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

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

which is the cross entropy, as desired. □

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

$$\mathrm{I}_p[X{=}x] = \left\langle\!\!\left\langle \boxed{Z} \overset{p}{\curlywedge} \boxed{X} \overset{x}{\twoheadleftarrow} \right\rangle\!\!\right\rangle. \tag{2}$$

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

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

□

Proposition 19. *The average* marginal *negative log likelihood* $\ell(p; x) := -\frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}} \log \sum_z p(x, z)$ *is the inconsistency of the PDG containing p and the data distribution* $\Pr_\mathcal{D}$*, plus the entropy of the data distribution (which is constant in p). That is,*

$$\ell(p; \mathcal{D}) = \left\langle\!\!\left\langle \boxed{Z} \overset{p}{\curlywedge} \boxed{X} \xleftarrow[(\infty)]{\Pr_\mathcal{D}} \right\rangle\!\!\right\rangle + \mathrm{H}(\Pr_\mathcal{D}).$$

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

$$\begin{aligned}
\left\langle\!\!\left\langle \boxed{Z} \overset{p}{\curlywedge} \boxed{X} \overset{\Pr_\mathcal{D}!}{\twoheadleftarrow} \right\rangle\!\!\right\rangle &= \inf_{\mu(Z|X)} \left[\!\!\left[\boxed{Z} \overset{p}{\curlywedge} \boxed{X} \overset{\Pr_\mathcal{D}!}{\twoheadleftarrow} \right]\!\!\right] \Big(\Pr_\mathcal{D}(X)\mu(Z \mid X)\Big) \\
&= \inf_{\mu(Z|X)} \boldsymbol{D}\Big(\Pr_\mathcal{D}(X)\mu(Z \mid X) \;\Big\|\; p(X, Z)\Big) \\
&= \inf_{\mu(Z|X)} \mathop{\mathbb{E}}_{\substack{x\sim\Pr_\mathcal{D} \\ z\sim\mu}} \log \frac{\mu(z \mid x)\Pr_\mathcal{D}(x)}{p(x, z)} \\
&= \frac{1}{|\mathcal{D}|} \inf_{\mu(Z|X)} \sum_{x\in\mathcal{D}} \mathop{\mathbb{E}}_{z\sim\mu(Z|x)} \log \frac{\mu(z \mid x)\Pr_\mathcal{D}(x)}{p(x, z)} \frac{p(x)}{p(x)} \\
&= \frac{1}{|\mathcal{D}|} \inf_{\mu(Z|X)} \sum_{x\in\mathcal{D}} \mathop{\mathbb{E}}_{z\sim\mu} \left[\log \frac{\mu(z)}{p(z \mid x)} + \log \frac{1}{p(x)} - \log \frac{1}{\Pr_\mathcal{D}(x)}\right]
\end{aligned}$$

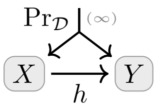
$$= \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \left[\inf_{\mu(Z)} \Big[\boldsymbol{D}(\mu(Z) \parallel p(Z \mid x)) \Big] + \log \frac{1}{p(x)} \right] - \mathrm{H}(\Pr_{\mathcal{D}})$$

$$= \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \log \frac{1}{p(x)} - \mathrm{H}(\Pr_{\mathcal{D}}) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} I_p(x) - \mathrm{H}(\Pr_{\mathcal{D}})$$

$$\Big(\quad = \boldsymbol{D}(\Pr_{\mathcal{D}} \parallel p(X)) \quad \Big)$$

□

Proposition 5. *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,*

$\langle\!\langle$ 

$\rangle\!\rangle$

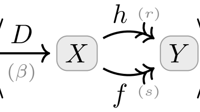
$$= \frac{1}{m} \sum_{i=1}^{m} \log \frac{1}{h(y_i \mid x_i)} - \mathrm{H}_{\Pr_{\mathcal{D}}}(Y|X).$$

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

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

□

Proposition 6. *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,*

$\langle\!\langle$ 

$\rangle\!\rangle$

$$= -\beta \log \Pr_{x \sim D}(f(x) = h(x)) = \beta \; \mathrm{I}_D[f = h]. \tag{3}$$

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

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

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 empirical distribution D.

So, among those distributions $\mu(X)$ supported on an event $E \subset \mathcal{V}(X)$, which minimizes is the relative entropy of $\boldsymbol{D}(\mu \parallel D)$? 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, Fagin et al. 2003). Let $f = h$ denote the event that f and h agree. Then we calculate

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

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

□

Proposition 15. *Consider a PDG containing two (distinct) conditional Gaussian distributions on a variable Y, whose parameters can both depend on a variable X. Its inconsistency takes the form*

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

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

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

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

$$\begin{aligned}
\langle\!\langle \mathcal{m} \rangle\!\rangle &= \inf_{\mu} \mathop{\mathbb{E}}_{x\sim D}\mathop{\mathbb{E}}_{y\sim\mu|x}\left[\beta_1\log\frac{\mu(y\mid x)}{\mathcal{N}(y\mid f(x),s(x))} + \beta_2\log\frac{\mu(y\mid x)}{\mathcal{N}(y\mid h(x),t(x))}\right] \\
&= \inf_{\mu} \mathop{\mathbb{E}}_{x\sim D}\mathop{\mathbb{E}}_{y\sim\mu|x}\left[\beta_1\log\frac{\mu(y\mid x)}{\frac{1}{s(x)\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y-f(x)}{s(x)}\right)^2\right)} + \beta_2\log\frac{\mu(y\mid x)}{\frac{1}{t(x)\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y-h(x)}{t(x)}\right)^2\right)}\right] \\
&= \inf_{\mu} \mathop{\mathbb{E}}_{x\sim D}\mathop{\mathbb{E}}_{y\sim\mu|x}\left[\log\mu(y\mid x)^{\beta_1+\beta_2} \begin{array}{l} +\frac{\beta_1}{2}\left(\frac{y-f(x)}{s(x)}\right)^2 \\ +\beta_1\log(s(x)\sqrt{2\pi})\end{array} \begin{array}{l} +\frac{\beta_2}{2}\left(\frac{y-h(x)}{t(x)}\right)^2 \\ +\beta_2\log(t(x)\sqrt{2\pi})\end{array}\right].
\end{aligned} \tag{8}$$

At this point, we would like make use of the fact that the sum of two parabolas is itself a parabola, so as to combine the two terms on the top right of the previous equation. Concretely, we claim (Claim 1, whose proof is at the end of the present one), that if we define

$$g(x) := \frac{\beta_1 t(x)^2 f(x) + \beta_2 s(x)^2 h(x)}{\beta_1 t(x)^2 + \beta_2 s(x)^2} \quad \text{and} \quad \tilde{\sigma}(x) := \frac{s(x)t(x)}{\sqrt{\beta_1 t(x)^2 + \beta_2 s(x)^2}},$$

then

$$\frac{\beta_1}{s(x)^2}(y-f)^2 + \frac{\beta_2}{t(x)^2}(y-h)^2 = \left(\frac{y-g}{\tilde{\sigma}}\right)^2 + \frac{\beta_1\beta_2}{\beta_1 t(x)^2 + \beta_2 s(x)^2}(f-h)^2.$$

Applying this to (8) leaves us with:

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

Pulling the term on the top right, which does not depend on Y, out of the expectation, and folding the rest of the terms back inside the logarithm (which in particular means first replacing the top middle term φ by $-\log(\exp(-\varphi))$), we obtain:

$$\langle\!\langle m \rangle\!\rangle = \mathop{\mathbb{E}}_{x\sim D} \left[\begin{array}{l} \inf\limits_{\mu(Y)} \mathop{\mathbb{E}}\limits_{y\sim\mu} \left[\log \mu(y)^{\beta_1+\beta_2} - \log\left(\frac{1}{\sqrt{2\pi}^{\beta_1+\beta_2} s(x)^{\beta_1} t(x)^{\beta_2}} \exp\left\{ -\frac{1}{2}\left(\frac{y-g(x)}{\tilde{\sigma}(x)}\right)^2 \right\} \right) \right] \\ \quad + \frac{1}{2}\frac{\beta_1\beta_2}{\beta_1 t(x)^2 + \beta_2 s(x)^2}\Big(f(x)-h(x)\Big)^2 \end{array} \right].$$

To simplify the presentation, let ψ be the term on the top right, and ξ be the term on the bottom. More explicitly, define

$$\psi(x,y) := \frac{1}{2}\frac{1}{\sqrt{2\pi}^{\beta_1+\beta_2} s(x)^{\beta_1} t(x)^{\beta_2}} \exp\left\{ -\frac{1}{2}\left(\frac{y-g(x)}{\tilde{\sigma}(x)}\right)^2 \right\}, \text{ and } \xi(x) := \frac{\beta_1\beta_2}{\beta_1 t(x)^2 + \beta_2 s(x)^2}\Big(f(x) - h(x)\Big)^2,$$

which lets us write the previous expression for $\langle\!\langle m \rangle\!\rangle$ as

$$\langle\!\langle m \rangle\!\rangle = \mathop{\mathbb{E}}_{x\sim D}\left[\inf_{\mu(Y)} \mathop{\mathbb{E}}_{y\sim\mu} \Big[\log \mu(y)^{\beta_1+\beta_2} - \log \psi(x,y) \Big] + \xi(x) \right]. \tag{9}$$

Also, let $\hat{\beta}_1 := \frac{\beta_1}{\beta_1+\beta_2}$, and $\hat{\beta}_2 := \frac{\beta_2}{\beta_1+\beta_2}$. For reasons that will soon become clear, we are actually interested in $\psi^{\frac{1}{\beta_1+\beta_2}}$, which we compute as

$$\begin{aligned} \psi(x,y)^{\frac{1}{\beta_1+\beta_2}} &= (2\pi)^{-\frac{1}{2}}\; s(x)^{\left(\frac{-\beta_1}{\beta_1+\beta_2}\right)} t(x)^{\left(\frac{-\beta_2}{\beta_1+\beta_2}\right)} \exp\left\{ -\frac{1}{2}\left(\frac{y-g(x)}{\tilde{\sigma}(x)}\right)^2 \right\}^{\frac{1}{\beta_1+\beta_2}} \\ &= \frac{1}{\sqrt{2\pi}\, s(x)^{\hat{\beta}_1} t(x)^{\hat{\beta}_2}} \exp\left\{ \frac{-1}{2(\beta_1+\beta_2)}\left(\frac{y-g(x)}{\tilde{\sigma}(x)}\right)^2 \right\}. \end{aligned}$$

Recall that the Gaussian densitity $\mathcal{N}(y|\, g(x),\, \tilde{\sigma}(x)\sqrt{\beta_1+\beta_2})$ of mean $g(x)$ and variance $\tilde{\sigma}(x)^2(\beta_1+\beta_2)$ is given by

$$\mathcal{N}\left(y \,\middle|\, g(x),\, \tilde{\sigma}(x)\sqrt{\beta_1+\beta_2}\right) = \frac{1}{\sqrt{2\pi}\;\tilde{\sigma}(x)\sqrt{\beta_1+\beta_2}} \exp\left\{ \frac{-1}{2(\beta_1+\beta_2)}\left(\frac{y-g(x)}{\tilde{\sigma}(x)}\right)^2 \right\},$$

which is quite similar, and has an identical dependence on y. To facilitate converting one to the other, we explicitly compute the ratio:

$$\frac{\psi(x,y)^{\frac{1}{\beta_1+\beta_2}}}{\mathcal{N}\left(y \,\middle|\, g(x),\, \tilde{\sigma}(x)\sqrt{\beta_1+\beta_2}\right)} = \frac{\tilde{\sigma}\sqrt{2\pi(\beta_1+\beta_2)}}{\sqrt{2\pi}\; s(x)^{\hat{\beta}_1} t(x)^{\hat{\beta}_2}} = \frac{\tilde{\sigma}\sqrt{\beta_1+\beta_2}}{s(x)^{\hat{\beta}_1} t(x)^{\hat{\beta}_2}}$$

$$
\begin{aligned}
&= \left(\frac{s(x)t(x)}{\sqrt{\beta_1 t(x)^2 + \beta_2 s(x)^2}} \right) \frac{\sqrt{\beta_1 + \beta_2}}{s(x)^{\hat\beta_1} t(x)^{\hat\beta_2}} && \text{[expand defn of } \tilde\sigma(x)\text{]} \\
&= s(x)^{1-\hat\beta_1}\, t(x)^{1-\hat\beta_2} \sqrt{\frac{\beta_1 + \beta_2}{\beta_1\, t(x)^2 + \beta_2\, s(x)^2}} \\
&= s(x)^{1-\hat\beta_1}\, t(x)^{1-\hat\beta_2} \sqrt{\frac{1}{\hat\beta_1\, t(x)^2 + \hat\beta_2\, s(x)^2}} && \text{[defn of } \hat\beta_1, \hat\beta_2\text{]} \\
&= \frac{s(x)^{\hat\beta_2}\, t(x)^{\hat\beta_1}}{\sqrt{\hat\beta_1\, t(x)^2 + \hat\beta_2\, s(x)^2}} && \text{[since } \hat\beta_1 + \hat\beta_2 = 1\text{]}
\end{aligned}
$$

Now, picking up from where we left off in (9), we have

$$
\begin{aligned}
\langle\!\langle m \rangle\!\rangle &= \mathop{\mathbb{E}}_{x\sim D} \left[\inf_{\mu(Y)} \mathop{\mathbb{E}}_{y\sim\mu} \left[\log \mu(y)^{\beta_1+\beta_2} - \log \psi(x,y) \right] + \xi(x) \right] \\
&= \mathop{\mathbb{E}}_{x\sim D} \left[\inf_{\mu(Y)} \mathop{\mathbb{E}}_{y\sim\mu} \left[\log \frac{\mu(y)^{\beta_1+\beta_2}}{\psi(x,y)^{\frac{\beta_1+\beta_2}{\beta_1+\beta_2}}} \right] + \xi(x) \right] \\
&= \mathop{\mathbb{E}}_{x\sim D} \left[\inf_{\mu(Y)} \mathop{\mathbb{E}}_{y\sim\mu} \left[(\beta_1+\beta_2) \log \frac{\mu(y)}{\psi(x,y)^{\frac{1}{\beta_1+\beta_2}}} \right] + \xi(x) \right] \\
&= \mathop{\mathbb{E}}_{x\sim D} \left[\inf_{\mu(Y)} \mathop{\mathbb{E}}_{y\sim\mu} \left[(\beta_1+\beta_2) \log \frac{\mu(y)}{\mathcal{N}\left(y \,\middle|\, g(x),\, \tilde\sigma(x)\sqrt{\beta_1+\beta_2}\right) \frac{s(x)^{\hat\beta_2}\, t(x)^{\hat\beta_1}}{\sqrt{\hat\beta_1\, t(x)^2 + \hat\beta_2\, s(x)^2}}} \right] + \xi(x) \right] \\
&= \mathop{\mathbb{E}}_{x\sim D} \left[\inf_{\mu(Y)} \mathop{\mathbb{E}}_{y\sim\mu} \left[(\beta_1{+}\beta_2) \log \frac{\mu(y)}{\mathcal{N}\left(y \,\middle|\, g(x),\, \tilde\sigma(x)\sqrt{\beta_1{+}\beta_2}\right)} \right] + (\beta_1{+}\beta_2) \log \frac{\sqrt{\hat\beta_1\, t(x)^2 + \hat\beta_2\, s(x)^2}}{s(x)^{\hat\beta_2}\, t(x)^{\hat\beta_1}} + \xi(x) \right]
\end{aligned}
$$

but now the entire left term is the infemum of a KL divergence, which is non-negative and equal to zero iff $\mu(y) = \mathcal{N}(y|g(x), \tilde\sigma(x)\sqrt{\beta_1{+}\beta_2})$. So the infemum on the left is equal to zero.

$$
\langle\!\langle m \rangle\!\rangle = \mathop{\mathbb{E}}_{x\sim D} \left[(\beta_1{+}\beta_2) \log \frac{\sqrt{\hat\beta_1\, t(x)^2 + \hat\beta_2\, s(x)^2}}{s(x)^{\hat\beta_2}\, t(x)^{\hat\beta_1}} + \xi(x) \right] \tag{10}
$$

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{x\sim D} \left[(\beta_1{+}\beta_2) \log \sqrt{\hat\beta_1\, t(x)^2 + \hat\beta_2\, s(x)^2} - (\beta_1{+}\beta_2) \log \left(s(x)^{\hat\beta_2}\, t(x)^{\hat\beta_1} \right) + \xi(x) \right] \\
&= \mathop{\mathbb{E}}_{x\sim D} \left[(\beta_1{+}\beta_2) \log \sqrt{\hat\beta_1\, t(x)^2 + \hat\beta_2\, s(x)^2} \begin{array}{l} -\beta_2 \log s(x) \\ -\beta_1 \log t(x) \end{array} + \xi(x) \right] \\
&= \mathop{\mathbb{E}}_{x\sim D} \left[(\beta_1{+}\beta_2) \log \sqrt{\frac{\beta_1\, t(x)^2 + \beta_2\, s(x)^2}{\beta_1 + \beta_2}} \begin{array}{l} -\beta_2 \log s(x) \\ -\beta_1 \log t(x) \end{array} + \frac{1}{2} \frac{\beta_1\beta_2}{\beta_1 t(x)^2 + \beta_2 s(x)^2} \Big(f(x) - h(x) \Big)^2 \right]
\end{aligned} \tag{11}
$$

Whew! Pulling the square root of the logarithm proves complex second half of the proposition. Now, we massage it into into a (slightly) more readable form.

To start, write σ_1 (the random variable) in place of $s(x)$ and σ_2 in place of $t(x)$. Let $\hat\beta$ without the subscript denote the vector $(\hat\beta_2, \hat\beta_1) = (\frac{\beta_2}{\beta_1+\beta_2}, \frac{\beta_1}{\beta_1+\beta_2})$, which we will use for weighted means. The $\hat\beta$-weighted arithmetic, geometric ($p = 0$), and quadratic ($p = 2$) means of σ_1 and σ_2 are:

$$
\mathrm{GM}_{\hat\beta}(\sigma_1, \sigma_2) = (\sigma_1)^{\hat\beta_2} (\sigma_2)^{\hat\beta_1} \qquad \text{and} \qquad \mathrm{QM}_{\hat\beta}(\sigma_1, \sigma_2) = \sqrt{\hat\beta_2 \sigma_1^2 + \hat\beta_1 \sigma_2^2}.
$$

So, now we can write $\xi(x)$ as

$$\begin{aligned}\frac{1}{2}\frac{\beta_1\beta_2}{\beta_1 t(x)^2+\beta_2 s(x)^2}\Big(f(x)-h(x)\Big)^2 &= \frac{1}{2}\frac{\beta_1\beta_2}{\beta_1+\beta_2}\frac{\beta_1+\beta_2}{\beta_1 t(x)^2+\beta_2 s(x)^2}\Big(f(x)-h(x)\Big)^2\\ &= \frac{1}{2}\frac{\beta_1\beta_2}{\beta_1+\beta_2}\left(\frac{1}{\mathrm{QM}_{\hat\beta}(\sigma_1,\sigma_2)}\right)^2\Big(f(x)-h(x)\Big)^2\\ &= \frac{1}{2}\frac{\beta_1\beta_2}{\beta_1+\beta_2}\left(\frac{\mu_1-\mu_2}{\mathrm{QM}_{\hat\beta}(\sigma_1,\sigma_2)}\right)^2;\end{aligned}$$

in the last step, we have replaced $f(x)$ and $g(x)$ with their respective random variables μ_1 and μ_2. As a result, (10) can be written as

$$\langle\!\langle m \rangle\!\rangle = \mathop{\mathbb{E}}_{D}\left[(\beta_1+\beta_2)\log\frac{\mathrm{QM}_{\hat\beta}(\sigma_1,\sigma_2)}{\mathrm{GM}_{\hat\beta}(\sigma_1,\sigma_2)}+\frac{1}{2}\frac{\beta_1\beta_2}{\beta_1+\beta_2}\left(\frac{\mu_1-\mu_2}{\mathrm{QM}_{\hat\beta}(\sigma_1,\sigma_2)}\right)^2\right]$$

. . . which is perhaps more comprehensible, and proves the first half of our proposition. □

Claim 1. *The sum of two functions that are unshifted parabolas as functions of y (i.e., both functions are of of the form $k(y-a)^2$), is itself a (possibly shifted) parabola of y (and of the form $k'(y-a')+b'$). More concretely, and adapted to our usage above, the following algebraic relation holds:*

$$\frac{\beta_1}{\sigma_1^2}(y-f)^2+\frac{\beta_2}{\sigma_2^2}(y-h)^2=\left(\frac{y-g}{\tilde\sigma}\right)^2+\frac{\beta_1\beta_2}{\beta_1\sigma_2^2+\beta_2\sigma_1^2}(f-h)^2,$$

where

$$g := \frac{\beta_1\sigma_2^2 f+\beta_2\sigma_1^2 h}{\beta_1\sigma_2^2+\beta_2\sigma_1^2} \quad \textit{and} \quad \tilde\sigma := \left(\frac{\beta_1}{\sigma_1^2}+\frac{\beta_2}{\sigma_2^2}\right)^{-1/2} = \frac{\sigma_1\sigma_2}{\sqrt{\beta_1\sigma_2^2+\beta_2\sigma_1^2}}.$$

Proof. Expand terms and complete the square. Starting from the left hand side, we have

$$\begin{aligned}&\frac{\beta_1}{\sigma_1^2}(y-f)^2+\frac{\beta_2}{\sigma_2^2}(y-h)^2\\ &=\frac{\beta_1}{\sigma_1^2}(y^2-2yf+f^2)+\frac{\beta_2}{\sigma_2^2}(y^2-2yh+h^2)\\ &=\left(\frac{\beta_1}{\sigma_1^2}+\frac{\beta_2}{\sigma_2^2}\right)y^2-2\left(\frac{\beta_1 f}{\sigma_1^2}+\frac{\beta_2 h}{\sigma_2^2}\right)y+\left(\frac{\beta_1 f^2}{\sigma_1^2}+\frac{\beta_2 h^2}{\sigma_2^2}\right)\\ &=\left(\frac{\beta_1}{\sigma_1^2}+\frac{\beta_2}{\sigma_2^2}\right)y^2-2\left(\frac{\beta_1 f}{\sigma_1^2}+\frac{\beta_2 h}{\sigma_2^2}\right)y+\left(\frac{\beta_1 f^2}{\sigma_1^2}+\frac{\beta_2 h^2}{\sigma_2^2}-\frac{\beta_1\beta_2(f-h)^2}{\beta_1\sigma_2^2+\beta_2\sigma_1^2}\right)+\frac{\beta_1\beta_2(f-h)^2}{\beta_1\sigma_2^2+\beta_2\sigma_1^2} \qquad (12)\end{aligned}$$

where in the last step we added and removed the same term (i.e., the completion of the square, although it is probably still unclear why this quantity will do that). The third parenthesized quantity needs the most work. Isolating it and getting a common denominator gives us:

$$\begin{aligned}&\frac{\beta_1 f^2}{\sigma_1^2}+\frac{\beta_2 h^2}{\sigma_2^2}-\frac{\beta_1\beta_2(f-h)^2}{\beta_1\sigma_2^2+\beta_2\sigma_1^2}\\ &=\frac{\beta_1 f^2(\beta_1\sigma_2^2+\beta_2\sigma_1^2)\sigma_2^2}{\sigma_1^2(\beta_1\sigma_2^2+\beta_2\sigma_1^2)\sigma_2^2}+\frac{\beta_2 h^2(\beta_1\sigma_2^2+\beta_2\sigma_1^2)\sigma_1^2}{\sigma_2^2(\beta_1\sigma_2^2+\beta_2\sigma_1^2)\sigma_1^2}-\frac{\beta_1\beta_2(f^2-2fh+h^2)(\sigma_1^2\sigma_2^2)}{(\beta_1\sigma_2^2+\beta_2\sigma_1^2)(\sigma_1^2\sigma_2^2)}\\ &=\frac{\beta_1^2\sigma_2^4 f^2+\cancel{\beta_1\beta_2\sigma_2^2\sigma_1^2 f^2}+\cancel{\beta_1\beta_2\sigma_1^2\sigma_2^2 h^2}+\beta_2^2\sigma_1^4 h^2-\cancel{\beta_1\beta_2\sigma_1^2\sigma_2^2 f^2}+2\beta_1\beta_2\sigma_1^2\sigma_2^2 fh-\cancel{\beta_1\beta_2\sigma_1^2\sigma_2^2 h^2}}{(\beta_1\sigma_2^2+\beta_2\sigma_1^2)(\sigma_1^2\sigma_2^2)}\\ &=\frac{\beta_1^2\sigma_2^4 f^2+\beta_2^2\sigma_1^4 h^2+2\beta_1\beta_2\sigma_1^2\sigma_2^2 fh}{(\beta_1\sigma_2^2+\beta_2\sigma_1^2)(\sigma_1^2\sigma_2^2)}.\end{aligned}$$

Substituting this expression into the third term of (12), while simultaneously computing common denominators for the first and second terms, yields

$$\left(\frac{\beta_1\sigma_2^2+\beta_2\sigma_1^2}{\sigma_1^2\sigma_2^2}\right)y^2-2\left(\frac{\beta_1\sigma_2^2 f+\beta_2\sigma_1^2 h}{\sigma_1^2\sigma_2^2}\right)y+\frac{\beta_1^2\sigma_2^4 f^2+\beta_2^2\sigma_1^4 h^2+2\beta_1\beta_2\sigma_1^2\sigma_2^2 fh}{(\beta_1\sigma_2^2+\beta_2\sigma_1^2)(\sigma_1^2\sigma_2^2)}+\frac{\beta_1\beta_2(f-h)^2}{\beta_1\sigma_2^2+\beta_2\sigma_1^2}. \tag{13}$$

On the other hand, using the definitions of g and $\tilde\sigma$, we compute:

$$\begin{aligned}\left(\frac{y-g}{\tilde\sigma}\right)^2&=\left(\frac{\beta_1\sigma_2^2+\beta_2\sigma_1^2}{\sigma_1^2\sigma_2^2}\right)\left(y-\frac{\beta_1\sigma_2^2 f+\beta_2\sigma_1^2 h}{\beta_1\sigma_2^2+\beta_2\sigma_1^2}\right)^2\\&=\left(\frac{\beta_1\sigma_2^2+\beta_2\sigma_1^2}{\sigma_1^2\sigma_2^2}\right)\left(y^2-2y\frac{\beta_1\sigma_2^2 f+\beta_2\sigma_1^2 h}{\beta_1\sigma_2^2+\beta_2\sigma_1^2}+\frac{\beta_1^2\sigma_2^4 f^2+\beta_2^2\sigma_1^4 h^2+2\beta_1\beta_2\sigma_2^2\sigma_1^2\, fh}{(\beta_1\sigma_2^2+\beta_2\sigma_1^2)^2}\right)\\&=\left(\frac{\beta_1\sigma_2^2+\beta_2\sigma_1^2}{\sigma_1^2\sigma_2^2}\right)y^2-2\left(\frac{\beta_1\sigma_2^2 f+\beta_2\sigma_1^2 h}{\sigma_1^2\sigma_2^2}\right)y+\frac{\beta_1^2\sigma_2^4 f^2+\beta_2^2\sigma_1^4 h^2+2\beta_1\beta_2\sigma_2^2\sigma_1^2\, fh}{(\beta_1\sigma_2^2+\beta_2\sigma_1^2)(\sigma_1^2\sigma_2^2)}\end{aligned}$$

. . . which is precisely the first 3 terms of (13). Putting it all together, we have shown that

$$\frac{\beta_1}{\sigma_1^2}(y-f)^2+\frac{\beta_2}{\sigma_2^2}(y-h)^2=\left(\frac{y-g}{\tilde\sigma}\right)^2+\frac{\beta_1\beta_2(f-h)^2}{\beta_1\sigma_2^2+\beta_2\sigma_1^2}$$

as desired. □

Proposition 7.

$$\left\langle\!\!\left\langle \begin{tikzcd} & & \mu_f \arrow[dr, "\mathcal{N}_1"] & \\ {} \arrow[r, "D", "(\infty)"'] & X \arrow[ur, two heads, "f"] \arrow[dr, two heads, "h"'] & & Y \\ & & \mu_h \arrow[ur, "\mathcal{N}_1"'] & \end{tikzcd} \right\rangle\!\!\right\rangle = \frac{1}{2}\,\mathbb{E}_D\big|f(X)-h(X)\big|^2 \\ =: \mathrm{MSE}_D(f,h)\,,$$

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

Proof. An immediate corolary of Proposition 15; simply set $s(x)=t(x)=\beta_1=\beta_2=1$ □

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*

$$\left\langle\!\!\left\langle \begin{tikzcd} {} \arrow[r, "p", "(r)"'] & X & {} \arrow[l, "q"', "(s)"] \end{tikzcd} \right\rangle\!\!\right\rangle = -(r+s)\log\sum_x\Big(p(x)^r q(x)^s\Big)^{\frac{1}{r+s}}\,.$$

Proof.

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

where $Z := (\sum_x p(x)^r q(x)^s)^{\frac{1}{r+s}}$ is the constant required to normalize the denominator as a distribution. The first term is now a relative entropy, and the only usage of μ. $\boldsymbol{D}(\mu \parallel \cdots)$ achives its minimum of zero when μ is the second distribution, so our formula becomes

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

□

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

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

Proof. This is another case where there's only one joint distribution $\mu(\Theta, Y)$ that gets a finite score. We must have $\mu(Y) = D(Y)$ since D has infinite confidence, which uniquely extends to the distribution $\mu(\Theta, Y) = D(Y)\delta_\theta(\Theta)$ for which deterministically sets $\Theta = \theta$.

The cpds corresponding to the edges labeled θ and D, then, are satisfied by this μ and contribute nothing to the score. So the two relevant edges that contribute incompatibility with this distribution are p and q. Letting $\mathcal{m}$ denote the PDG in question, we compute:

$$\begin{aligned} \langle\!\langle \mathcal{m} \rangle\!\rangle &= \mathop{\mathbb{E}}_{\mu}\left[\log \frac{\mu(Y|\Theta)}{p(Y|\Theta)} + \beta \log \frac{\mu(\Theta)}{q(\Theta)} \right] \\ &= \mathop{\mathbb{E}}_{y\sim D}\left[\log \frac{D(y)}{p(y|\theta)} + \beta \log \frac{1}{q(\theta)} \right] \\ &= \mathop{\mathbb{E}}_{y\sim D}\left[\log \frac{1}{p(y|\theta)} + \beta \log \frac{1}{q(\theta)} + \log D(y) \right] \\ &= \mathop{\mathbb{E}}_{y\sim D}\left[\log \frac{1}{p(y|\theta)} \right] + \beta \log \frac{1}{q(\theta)} - \mathrm{H}(D) \end{aligned}$$

as desired. □

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

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

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

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

$$= \left[\!\!\left[\begin{tikzcd} {} \arrow[r, "q", "(\infty)"'] & \boxed{Z} & {} \arrow[l, "p"'] \arrow[r, "p"] & \boxed{X} & x \arrow[l, two heads] \end{tikzcd}\right]\!\!\right](\delta_x(X)q(Z))$$

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

□

We proove both Proposition 12 and Proposition 16 at the same time.

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,*

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

Proposition 16. *The following analog of Proposition 12 for a whole dataset $\mathcal{D}$ holds:*

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

Proof. The two proofs are similar. For Proposition 12, the optimal distribution must be $\delta_x(X)e(Z \mid X)$, and for Proposition 16, it must be $\Pr_{\mathcal{D}}(X)e(Z \mid X)$, 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

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

in the second. □

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

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,*

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

Proof.

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

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

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

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

□

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

Proof. In the main text, we defined $\mathcal{M}_\Psi$ to 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 $\alpha_J, \beta_J := \theta_J$. Let $\mathrm{the}(\{x\}) := x$ be a function that extracts the unique element singleton set. It was shown by Richardson and Halpern (2021) (Corolary 4.4.1) that

$$\mathrm{the} [\![\mathcal{M}_\Psi]\!]_1^* = \Pr_{\Phi,\theta}(\mathbf{x}) = \frac{1}{Z_\Psi} \prod_J \phi_J(\mathbf{x}_J)^{\theta_J}.$$

Recall the statement of Prop 4.6 from Richardson and Halpern (2021):

$$[\![\mathcal{M}]\!]_\gamma(\mu) = \mathop{\mathbb{E}}_{\mathbf{w} \sim \mu} \left\{ \sum_{X \xrightarrow{L} Y} \left[\beta_L \log \frac{1}{\mathbf{p}_L(y^{\mathbf{w}} | x^{\mathbf{w}})} + (\gamma \alpha_L - \beta_L) \log \frac{1}{\mu(y^{\mathbf{w}} | x^{\mathbf{w}})} \right] - \gamma \log \frac{1}{\mu(\mathbf{w})} \right\}, \tag{14}$$

where $x^{\mathbf{w}}$ and $y^{\mathbf{w}}$ are the respective values of the variables X and Y in the world $\mathbf{w}$. Note that if $\gamma = 1$, and α, β are both equal to θ in $\mathcal{M}_\Psi$, the middle term (in purple) is zero. So in our case, since the edges are $\{\xrightarrow{J} \mathbf{X}_J\}$ and $\mathbf{p}_J(\mathbf{X}_J) = \phi_J(\mathbf{X}_J)$, (14) reduces to the standard variational free energy

$$\begin{aligned}
\mathit{VFE}_\Psi(\mu) &= \mathop{\mathbb{E}}_{\mu} \left[\sum_{J \in \mathcal{J}} \theta_J \log \frac{1}{\phi_J(\mathbf{X}_J)} \right] - \mathrm{H}(\mu) \qquad (15) \\
&= \mathop{\mathbb{E}}_{\mu} \big\langle \boldsymbol{\varphi}, \boldsymbol{\theta} \big\rangle_{\mathcal{J}} - \mathrm{H}(\mu), \quad \text{where } \varphi_J(\mathbf{X}_J) := \log \frac{1}{\phi_J(\mathbf{X}_J)}.
\end{aligned}$$

By construction, $\Pr_\Psi$ uniquely minimizes *VFE*. The 1-inconsistency, $\langle\!\langle \mathcal{M}_\Psi \rangle\!\rangle$ is the minimum value attained. We calculate:

$$\begin{aligned}
\langle\!\langle \mathcal{M} \rangle\!\rangle_1 &= \mathit{VFE}_\Psi(\Pr_\Psi) \\
&= \mathop{\mathbb{E}}_{\mathbf{x} \sim \mu} \left\{ \sum_{J \in \mathcal{J}} \left[\theta_J \log \frac{1}{\phi_J(\mathbf{x}_J)} \right] - \log \frac{1}{\Pr_{\Phi,\theta}(\mathbf{x})} \right\} && \Big[\text{by } (15) \Big]
\end{aligned}$$

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} \left\{ \sum_{J\in\mathcal{J}} \left[\theta_J \log \frac{1}{\phi_J(\mathbf{x}_J)} \right] - \log \frac{Z_\Psi}{\prod_{J\in\mathcal{J}} \phi_J(\mathbf{x}_J)^{\theta_j}} \right\} && \left[\text{definition of } \Pr_\Psi \right] \\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} \left\{ \sum_{J} \left[\theta_J \log \frac{1}{\phi_J(\mathbf{x}_J)} \right] - \sum_{J\in\mathcal{J}} \left[\theta_J \log \frac{1}{\phi_J(\mathbf{x}_J)} \right] - \log Z_\Psi \right\} \\
&= \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} [-\log Z_\Psi] \\
&= -\log Z_\Psi && \left[Z_\Psi \text{ is constant in } \mathbf{x} \right]
\end{aligned}
$$

□

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

Proof. Since p has high confidence, and $\mathtt{T}$ is always equal to $\mathtt{t}$, the only joint distribution on $(X, \mathtt{T})$ with finite score is $\mu(X, \mathtt{T}) = p(X)\delta_{\mathtt{t}}(\mathtt{T})$. We compute its score directly:

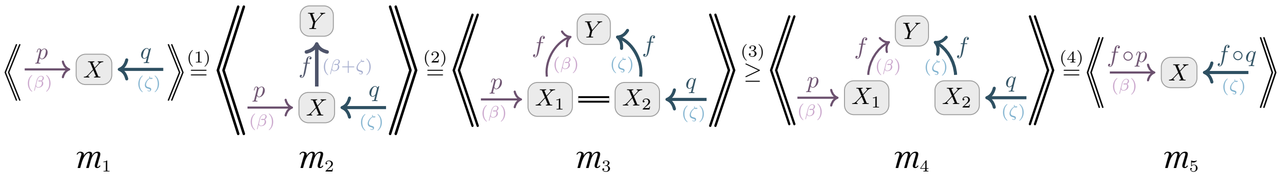
$$
\begin{aligned}
\left\langle\!\!\left\langle \xrightarrow[(\infty)]{p} \boxed{X} \xrightarrow{\hat{c}} \boxed{\mathtt{T}} \overset{\mathtt{t}}{\twoheadleftarrow} \right\rangle\!\!\right\rangle &= \mathop{\mathbb{E}}_{\mu} \frac{\mu(X,\mathtt{T})}{\hat{c}(\mathtt{t}\,|X)} = \mathop{\mathbb{E}}_{p} \log \frac{1}{\hat{c}(\mathtt{t}\,|X)} = \mathop{\mathbb{E}}_{p} \log \frac{1}{\exp(-c(X))} \\
&= \mathop{\mathbb{E}}_{p} \log \exp(c(X)) = \mathop{\mathbb{E}}_{p} c(X) = \mathop{\mathbb{E}}_{x\sim p} c(x).
\end{aligned}
$$

□

C.1 Additional Proofs for Unnumbered Claims

C.1.1 Details on the Data Processing Inequality Proof

We now provide more details on the proof of the Data Processing Equality that appeared in Figure 2 of the main text. We repeat it now for convenience, with labeled PDGs ($\mathcal{m}_1, \ldots, \mathcal{m}_5$) and numbered (in)equalities.



We now enumerate the (in)equalities to prove them.

1. Let $\mu(X)$ denote the (unique) optimal distribution for $\mathcal{m}_1$. Now, the joint distribution $\mu(X,Y) := \mu(X)f(Y|X)$ has incompatibility with $\mathcal{m}_2$ equal to

$$
\begin{aligned}
\mathit{Inc}_{\mathcal{m}_2}(\mu(X,Y)) &= \beta \boldsymbol{D}(\mu(X) \parallel p(X)) + \zeta \boldsymbol{D}(\mu(X) \parallel q(X)) + (\beta{+}\zeta) \mathop{\mathbb{E}}_{x\sim\mu} \left[\boldsymbol{D}(\mu(Y|x) \parallel f(Y|x)) \right] \\
&= \mathit{Inc}_{\mathcal{m}_1}(\mu(X)) + (\beta{+}\zeta) \mathop{\mathbb{E}}_{x\sim\mu} \boldsymbol{D}(\mu(Y|x) \parallel f(Y|x)) \\
&= \langle\!\langle \mathcal{m}_1 \rangle\!\rangle && \left[\begin{array}{c} \text{as } \mu(Y|x) = f(Y|x) \text{ wherever } \mu(x) > 0, \\ \text{and } \mu(X) \text{ minimizes } \mathit{Inc}_{\mathcal{m}_1} \end{array} \right]
\end{aligned}
$$

So $\mu(X,Y)$ witnesses the fact that $\langle\!\langle \mathcal{m}_2 \rangle\!\rangle \leq \langle\!\langle \mathcal{m}_1 \rangle\!\rangle$. Furthermore, every joint distribution ν must have at least this incompatibility, as it must have some marginal $\nu(X)$, which contributes incompatibility $\mathit{Inc}_{\mathcal{m}_1}(\nu(X)) \geq \mathit{Inc}_{\mathcal{m}_1}(\mu(X)) = \langle\!\langle \mathcal{m}_1 \rangle\!\rangle$.

2. The equals sign in $\mathcal{M}_3$ may be equivalently interpreted as a cpd $\mathit{eq}(X_1|X_2) := x_2 \mapsto \delta_{x_2}(X_1)$, a cpd $\mathit{eq}'(X_2|X_1) := x_1 \mapsto \delta_{x_1}(X_2)$, or both at once; in each case, the effect is that a joint distribution μ with support on an outcome for which $X_1 \neq X_2$ gets an infinite penalty, so a minimizer $\mu(X_1, X_2, Y)$ of $\mathit{Inc}\mathcal{M}_3$ must be isomorphic to a distribution $\mu'(X, Y)$.

   Furthermore, it is easy to verify that $\mathit{Inc}_{\mathcal{M}_2}(\mu'(X,Y)) = \mathit{Inc}_{\mathcal{M}_3}(\mu(X,X,Y))$. More formally, we have:

$$\langle\!\langle \mathcal{M}_3 \rangle\!\rangle = \inf_{\mu(X_1,X_2,Y)} \mathop{\mathbb{E}}_{\mu} \left[\beta \log \frac{\mu(X_1)}{p(X_1)} + \zeta \log \frac{\mu(X_2)}{q(X_2)} + \beta \log \frac{\mu(Y|X_1)}{f(Y|X_1)} + \zeta \log \frac{\mu(Y|X_2)}{f(Y|X_2)} + \log \frac{\mu(X_1|X_2)}{\mathit{eq}(X_1, X_2)} \right]$$

   but if X_1 always equals X_2 (which we call simply X), as it must for the optimal μ, this becomes

$$\begin{aligned}
&= \inf_{\mu(X_1=X_2=X,Y)} \mathop{\mathbb{E}}_{\mu} \left[\beta \log \frac{\mu(X)}{p(X)} + \zeta \log \frac{\mu(X)}{q(X)} + \beta \log \frac{\mu(Y|X)}{f(Y|X)} + \zeta \log \frac{\mu(Y|X)}{f(Y|X)} \right] \\
&= \inf_{\mu(X,Y)} \mathop{\mathbb{E}}_{\mu} \left[\beta \log \frac{\mu(X)}{p(X)} + \zeta \log \frac{\mu(X)}{q(X)} + (\beta{+}\zeta) \log \frac{\mu(Y|X)}{f(Y|X)} \right] \\
&= \inf_{\mu(X,Y)} \mathit{Inc}_{\mathcal{M}_2}(\mu) \\
&= \langle\!\langle \mathcal{M}_2 \rangle\!\rangle.
\end{aligned}$$

3. Eliminating the edge or edges enforcing the inequality cannot increase inconsistency, by Lemma 1.

4. Although this final step of composing the edges with shared confidences looks intuitively like it should be true (and it is!), its proof may not be obvious. We now provide a rigorous proof of this equality.

   To ameliorate subscript pains, we henceforth write X for X_1, and Z for X_2. We now compute:

$$\begin{aligned}
\langle\!\langle \mathcal{M}_4 \rangle\!\rangle &= \inf_{\mu(X,Z,Y)} \mathop{\mathbb{E}}_{\mu} \left[\beta \log \frac{\mu(X)\,\mu(Y|X)}{p(X)\,f(Y|X)} + \zeta \log \frac{\mu(Z)\,\mu(Y|Z)}{q(Z)\,f(Y|Z)} \right] \\
&= \inf_{\mu(X,Z,Y)} \mathop{\mathbb{E}}_{\mu} \left[\beta \log \frac{\mu(Y)\,\mu(X|Y)}{p(X)\,f(Y|X)} + \zeta \log \frac{\mu(Y)\,\mu(Z|Y)}{q(Z)\,f(Y|Z)} \right] \qquad \text{[apply Bayes Rule in numerators]}
\end{aligned}$$

   By the chain rule, every distribution $\mu(X, Z, Y)$ may be specified as $\mu(Y)\mu(X|Y)\mu(Z|X, Y)$, so we can rewrite the formula above as

$$\langle\!\langle \mathcal{M}_4 \rangle\!\rangle = \inf_{\mu(Y)} \inf_{\mu(X|Y)} \inf_{\mu(Z|Y,X)} \mathop{\mathbb{E}}_{y\sim\mu(Y)} \mathop{\mathbb{E}}_{x\sim\mu(X|y)} \mathop{\mathbb{E}}_{z\sim\mu(Z|y,x)} \left[\beta \log \frac{\mu(y)\,\mu(x\,|\,y)}{p(x)\,f(y\,|\,x)} + \zeta \log \frac{\mu(y)\,\mu(z\,|\,y)}{q(z)\,f(y\,|\,z)} \right],$$

   where $\mu(Z|Y)$ is the defined in terms of the primitives $\mu(X|Y)$ and $\mu(Z|X, Y)$ as $\mu(Z|Y) := y \mapsto \mathbb{E}_{x\sim\mu(X|y)}\, \mu(Z|y, x)$, and is a valid cpd, since it is a mixture distribution. Since the first term (with β) does not depend on z, we can take it out of the expectation, so

$$\langle\!\langle \mathcal{M}_4 \rangle\!\rangle = \inf_{\mu(Y)} \inf_{\mu(X|Y)} \inf_{\mu(Z|Y,X)} \mathop{\mathbb{E}}_{y\sim\mu(Y)} \mathop{\mathbb{E}}_{x\sim\mu(X|y)} \left[\beta \log \frac{\mu(y)\,\mu(x\,|\,y)}{p(x)\,f(y\,|\,x)} + \zeta \mathop{\mathbb{E}}_{z\sim\mu(Z|y,x)} \left[\log \frac{\mu(y)\,\mu(z\,|\,y)}{q(z)\,f(y\,|\,z)} \right] \right];$$

   we can split up $\mathbb{E}_{\mu(X|y)}$ by linearity of expectation, to get

$$\langle\!\langle \mathcal{M}_4 \rangle\!\rangle = \inf_{\mu(Y)} \inf_{\mu(X|Y)} \inf_{\mu(Z|Y,X)} \mathop{\mathbb{E}}_{y\sim\mu(Y)} \left[\beta \mathop{\mathbb{E}}_{x\sim\mu(X|y)} \left[\log \frac{\mu(y)\,\mu(x\,|\,y)}{p(x)\,f(y\,|\,x)} \right] + \zeta \mathop{\mathbb{E}}_{\substack{x\sim\mu(X|y)\\ z\sim\mu(Z|y,x)}} \left[\log \frac{\mu(y)\,\mu(z\,|\,y)}{q(z)\,f(y\,|\,z)} \right] \right]$$

   Note that the quantity inside the second expectation does not depend on x. Therefore, the second expectation is just an explicit way of sampling z from the mixture distribution $\mathbb{E}_{x\sim\mu(X|y)}\, \mu(Z|x, y)$, which is the definition of $\mu(Z|y)$. Once we make this replacement, it becomes clear that the only feature of $\mu(Z|Y, X)$

that matters is the mixture $\mu(Z|Y)$. Simplifying the second expectation in this way, and replacing the infemum over $\mu(Z|X,Y)$ with one over $\mu(Z|Y)$ yields:

$$\langle\!\langle \mathcal{M}_4 \rangle\!\rangle = \inf_{\mu(Y)} \inf_{\mu(X|Y)} \inf_{\mu(Z|Y)} \mathop{\mathbb{E}}_{y\sim\mu(Y)} \left[\beta \mathop{\mathbb{E}}_{x\sim\mu(X|y)} \left[\log \frac{\mu(y)\,\mu(x\,|\,y)}{p(x)\,f(y\,|\,x)} \right] + \zeta \mathop{\mathbb{E}}_{z\sim\mu(Z|y)} \left[\log \frac{\mu(y)\,\mu(z\,|\,y)}{q(z)\,f(y\,|\,z)} \right] \right]$$

Now, a cpd $\mu(X|Y)$ is just[7] a (possibly different) distribution $\nu_y(X)$ for every value of Y. Observe that, inside the expectation over $\mu(Y)$, the cpds $\mu(X|Y)$ and $\mu(Z|Y)$ are used only for the *present* value of y, and do not reference, say, $\mu(X|y')$ for $y' \neq y$. Because there is no interaction between the choice of cpd $\mu(X|y)$ and $\mu(X|y')$, it is not necessary to jointly optimize over entire cpds $\mu(X|Y)$ all at once. Rather, it is equivalent to to take the infemum over $\nu(X)$, separately for each y. Symmetrically, we may as well take the infemum over $\lambda(Z)$ separately for each y, rather than jointly finding the optimal $\mu(Z|Y)$ all at once. Operationallly, this means we can pull the infema inside the expectation over Y. And since the first term doesn't depend on Z and the second doesn't depend on X, we get:

$$\langle\!\langle \mathcal{M}_4 \rangle\!\rangle = \inf_{\mu(Y)} \mathop{\mathbb{E}}_{y\sim\mu(Y)} \left[\inf_{\nu(X)} \beta \mathop{\mathbb{E}}_{\nu(X)} \left[\log \frac{\mu(y)\,\nu(X)}{p(X)\,f(y\,|X)} \right] + \inf_{\lambda(Z)} \zeta \mathop{\mathbb{E}}_{\lambda(Z)} \left[\log \frac{\mu(y)\,\lambda(Z)}{q(Z)\,f(y\,|Z)} \right] \right]$$

Next, we pull the same trick we've used over and over: find constants so that we can regard the dependence as a relative entropy with respect to the quantity being optimized. Grouping the quantities apart from $\nu(X)$ on the left term and normalizing them (and analogously for $\lambda(Z)$ on the right), we find that

$$\langle\!\langle \mathcal{M}_4 \rangle\!\rangle = \inf_{\mu(Y)} \mathop{\mathbb{E}}_{y\sim\mu(Y)} \left[\begin{array}{l} \beta \inf_{\nu(X)} \boldsymbol{D}\Big(\nu(X) \,\Big\|\, \frac{1}{C_1(y)} p(X) \frac{f(y|X)}{\mu(y)}\Big) - \beta \log C_1(y) \\ +\zeta \inf_{\lambda(Z)} \boldsymbol{D}\Big(\lambda(Z) \,\Big\|\, \frac{1}{C_2(y)} q(Z) \frac{f(y|Z)}{\mu(y)}\Big) - \zeta \log C_2(y) \end{array} \right],$$

where

$$C_1(y) = \sum_x p(x) \frac{f(y|x)}{\mu(y)} = \frac{1}{\mu(y)} \mathop{\mathbb{E}}_{p(X)} f(y|X) \qquad \text{and} \qquad C_2(y) = \sum_z q(z) \frac{f(y|z)}{\mu(y)} = \frac{1}{\mu(y)} \mathop{\mathbb{E}}_{q(Z)} f(y|Z)$$

are the constants required to normalize the distributions. Both relative entropies are minimized when their arguments match, at which point they contribute zero, so we have

$$\begin{aligned} \langle\!\langle \mathcal{M}_4 \rangle\!\rangle &= \inf_{\mu(Y)} \mathop{\mathbb{E}}_{y\sim\mu(Y)} \left[\beta \log \frac{1}{C_1(y)} + \zeta \log \frac{1}{C_2(y)} \right] \\ &= \inf_{\mu(Y)} \mathop{\mathbb{E}}_{y\sim\mu(Y)} \left[\beta \log \frac{\mu(y)}{\mathbb{E}_{p(X)} f(y|X)} + \zeta \log \frac{\mu(y)}{\mathbb{E}_{q(Z)} f(y|Z)} \right] \\ &= \inf_{\mu(Y)} \mathop{\mathbb{E}}_{\mu} \Big[\beta \boldsymbol{D}(\mu \parallel f \circ p) + \zeta \boldsymbol{D}(\mu \parallel f \circ q) \Big] \\ &= \langle\!\langle \mathcal{M}_5 \rangle\!\rangle. \end{aligned}$$

C.1.2 Details for Claims made in Section 8

Distortion Due to Inconsistency. In the footnote on Page 8, we claimed that if the model confidence β_p were 1 rather than ∞, we would have obtained an incconsistency of $\log \mathbb{E}_{x\sim p} \exp(c(x))$, and that the optimal distribution would not have been $p(X)$.

$$\left\langle\!\!\left\langle \xrightarrow{p} \boxed{X} \xrightarrow{\hat{c}} \boxed{\mathtt{T}} \twoheadleftarrow{\mathtt{t}} \right\rangle\!\!\right\rangle = \inf_{\mu}(X) \mathop{\mathbb{E}}_{x\sim\mu} \left[\log \frac{\mu(x)}{p(x)} + \log \frac{\mu(\mathtt{t}\,|\,x)}{\hat{c}(\mathtt{t}\,|\,x)} \right]$$

$$= \inf_{\mu}(X) \mathop{\mathbb{E}}_{x\sim\mu} \left[\log \frac{\mu(x)}{p(x)} + \log \frac{1}{\hat{c}(\mathtt{t}\,|\,x)} \right]$$

[7] modulo measurability concerns that do not affect the infemum; see Appendix A

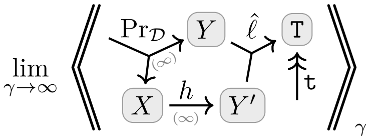
$$= \inf_{\mu}(X) \mathop{\mathbb{E}}_{x\sim\mu} \left[\log \frac{\mu(x)}{p(x)\exp(-c(x))} \cdot \frac{Z}{Z} \right]$$

where $Z = \sum_x p(x)\exp(-c(x)) = \mathbb{E}_p \exp(-c(X))$ is the constant required to normalize the distribution

$$\begin{aligned} &= \inf_{\mu(X)} \boldsymbol{D}\left(\mu \;\middle\|\; \frac{1}{Z}\, p(X)\exp(-c(X))\right) - \log Z \\ &= -\log Z \\ &= -\log \mathop{\mathbb{E}}_{x\sim p} \exp(-c(x)) \end{aligned}$$

as promised. Note also that in the proof, we showed that the optimal distribution is proportional to $p(X)\exp(-c(X))$ which means that it equals $p(X)$ if and only if $c(X)$ is constant in X.

Enforcing the Qualitative Picture. We also claimed without careful proof in Section 8 that, if $\alpha_h = \alpha_{\Pr_\mathcal{D}} = 1$, then



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

Why is this? For such a setting of α, which intuitively articulates a causal picture where X, Y is generated from $\Pr_\mathcal{D}$, and Y' generated by $h(Y'|X)$, the information deficiency $\mathit{IDef}_\mathcal{S}(\mu(X,Y,Y'))$ of a distribution μ is

$$\begin{aligned} \mathit{IDef}_\mathcal{S}(\mu(X,Y,Y')) &= -\operatorname{H}_\mu(X,Y,Y') + \operatorname{H}(X,Y) + \operatorname{H}(Y'|X) \\ &= \operatorname{H}_\mu(Y'|X) - \operatorname{H}_\mu(Y'|X,Y) \\ &= \operatorname{I}_\mu(Y;Y'|X). \end{aligned}$$

Both equalities of the derivation above standard information theoretic identities (See, for instance, MacKay 2003), and the final quantity $\operatorname{I}_\mu(Y;Y'|X)$ is the *conditional mutual information* between Y and Y' given X, and is a non-negative number that equals zero if and only if Y and Y' are conditionally independent given X.

As a result, as $\gamma \to \infty$ any distribution that for which Y' and Y are not independent given X will incur infinite cost. Since the confidences in h and $\Pr_\mathcal{D}$ are also infinite, so will a violation of either cpd. There is only one distribution that has both cpds and also this independence; that distribution is $\mu(X,Y,Y') := \Pr_\mathcal{D}(X,Y)h(Y'|X)$. Now the argument of Proposition 14 applies: all other cpds must be matched, and the inconsistency is the expected incompatibility of $\hat{l}$, which equals

$$\mathop{\mathbb{E}}_{\substack{(x,y)\sim\Pr_\mathcal{D} \\ y'\sim p(Y'|\,x)}} \log \frac{1}{\hat{\ell}(\mathtt{t}\,|y,y')} = \mathop{\mathbb{E}}_{\substack{(x,y)\sim\Pr_\mathcal{D} \\ y'\sim p(Y'|\,x)}} \log \frac{1}{\exp(-\ell(y,y'))} = \mathop{\mathbb{E}}_{\substack{(x,y)\sim\Pr_\mathcal{D} \\ y'\sim p(Y'|\,x)}} \left[\log\exp(\ell(y,y'))\right] = \mathop{\mathbb{E}}_{\substack{(x,y)\sim\Pr_\mathcal{D} \\ y'\sim p(Y'|\,x)}} \left[\ell(y,y')\right].$$

D More Notes

D.1 Maximum A Posteriori and Priors

The usual telling of the correspondence between regularizers and priors is something like the following. Suppose you have a parameterized family of distributions $\Pr(X|\Theta)$ and have observed evidence X, but do not know the parameter Θ. The maximum-likelihood estimate of Θ is then

$$\theta^{\text{MLE}}(X) := \arg\max_{\theta\in\Theta} \Pr(X|\theta) = \arg\max_{\theta\in\Theta} \log\Pr(X|\theta).$$

The logarithm is a monotonic transformation, so it does not change the argmax, but it has nicer properties, so that function is generally used instead. (Many of the loss functions in main body of the paper are log-likelihoods also.)

In some sense, better than estimating the maximum likelihood, is to perform a Bayesian update with the new information, to get a *distribution* over Θ. If that's too expensive, we could simply take the estimate with the highest posterior probability, which is called the Maximum A Posteriori (MAP) estimate. For any given θ, the Bayesian reading of Bayes rule states that

$$\text{posterior } \Pr(\Theta|X) = \frac{\text{likelihood } \Pr(X|\Theta) \cdot \text{prior } \Pr(\Theta)}{\text{evidence } \Pr(X) = \sum_{\theta'} \Pr(X|\theta') \Pr(\theta')}.$$

So taking a logarithm,

$$\text{log-posterior } \log \Pr(\Theta|X) = \text{log-likelihood } \log \Pr(X|\Theta) \;+\; \text{log-prior } \log \Pr(\Theta) - \text{log-evidence } \log \Pr(X).$$

The final term does not depend on θ, so it is not relevant for finding the optimal θ by this metric. Swapping the signs so that we are taking a minimum rather than a maximum, the MAP estimate is then given by

$$\theta^{\text{MAP}}(X) := \arg\min_{\theta \in \Theta} \left\{ \log \frac{1}{\Pr(X|\theta)} + \log \frac{1}{\Pr(\theta)} \right\}.$$

Note that if negative log likelihood (or surprisal, $-\log \Pr(X|\theta)$) was our original loss function, we have now added an arbitrary extra term, as a function of Θ, to our loss function. It is in this sense that priors classically correspond to regularizers.