Inference, Dynamics, and Inconsistency in PDGs

1 Introduction

In [], we introduced Probabilistic Dependency Graphs (PDGs), argued that they are a particularly natural modeling tool, and proved that their semantics subsumes that of other other common models (Bayesian Networks and factor graphs), but are strictly more expressive. A model has limited usefulness without a query mechanism, and in [] we only gesture briefly at how to draw inferences with them. In the present paper, we give a more complete picture of how to do inference in a PDG, as we explore the rich connection between the inconsistency of a PDG, and standard inference algorithms in other settings.

Outline

We will begin by defining the inconsistency of a PDG (**??**), and proving that it has nice properties which make it amenable to gradient descent (**??**). We then show that many tasks of interest (belief updating and model inference) can be done efficiently with access to an inconsitency oracle (2.4–**??**), although computing the inconsistency is hard (Propositions 2.8 and 2.9). In the remainder of the paper, we lay the foundations for tractable approximations, where we find analogues of belief propogation, variational inference, and particle-based approaches.

Notation and PDGs

Definition 1.1. □

Why bother changing?

Before we get to the central material, we first inroduce some new notation and conventions. We now write $\{\!\{\mathcal{M}\}\!\}$ to denote the set of distributions consistent with a PDG $\mathcal{M}$.[1] We view a cpd $p(Y \mid X)$ as a special case of a PDG with a single edge $X \to Y$ annotated with the cpd p, and taking default weights $\alpha_p = 0$, and $\beta_p = 1$. In cases where we wish to indicate other weights, we indicate the modifications with a parameter assignment. For instance, to give a PDG consisting of the cpd p, attached to a value of α_p equal to α_0, and a value of β_p equal to β_0, we might write $p^{(\alpha:\alpha_0;\beta:\beta_0)}$, or in abbrevaited form as $p^{(\alpha_0;\beta_0)}$. Often we will want to describe limiting PDGs with infinite quantitative confidence β; we will use an explanation point to convey this briefly, so that $p^{!}$ is shorthand for $\lim_{t\to\infty} p^{(\beta:t)}$.

YUCK! Write \beta = \infty.

Why not just write \beta = \beta_0, and leave out \alpha if it takes the default value. (Including = makes it easer).

We call two PDGs $\mathcal{A}$ and $\mathcal{B}$ *value-compatable* if they agree on the values of any variables they have in common—that is, if $\mathcal{V}^{\mathcal{A}}(X) = \mathcal{V}^{\mathcal{B}}(X)$ for all $X \in \mathcal{N}^{\mathcal{A}} \cap \mathcal{N}^{\mathcal{B}}$. If $\mathcal{A}$ and $\mathcal{B}$ are value-compatible PDGs, we write $\mathcal{A} + \mathcal{B}$ to indicate the PDG with the union of the variables and the *disjoint* union of all edges, where each edge L retains its associated parameters $\mathbf{p}_L$, α_L, and β_L.

so an X edge appears twice even if \A and \B agree on the value of X? Then why insist on value compatibility?

2 Inference and Dynamics via Inconsistency

We start by introducing a repackaged version of PDG semantics, in terms of its degree of "inconsistency".

Definition 2.1. If $\mathcal{M}$ is a PDG, let $\langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma$ denote the *degree of inconsistency* of $\mathcal{M}$ at γ, and be given by

$$\langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma := \inf_{\mu \in \Delta[\mathcal{V}(\mathcal{M})]} [\![\mathcal{M}]\!]_\gamma(\mu).$$

□

missing \gamma

So now, for each $\gamma \in [0, \infty)$, $\langle\!\langle \mathcal{M} \rangle\!\rangle$ is a real number; we now justify its name. First, we have already given a sensible notion of what it means for a to be (in)consistent, given by the semantics $\{\!\{-\}\!\}$. Fortunately, this new definition is compatible with it.

?? PDG?

Prop 2.1. *$\mathcal{M}$ is a* consistent *PDG, in that $\{\!\{\mathcal{M}\}\!\} \neq \emptyset$, if and only if $\langle\!\langle \mathcal{M} \rangle\!\rangle_0 = 0$.*

[1] In [?], this set of consistent distributions was denoted $[\![\mathcal{M}]\!]_{\text{sd}}$.

In the full paper, we defined $Inc(\mathcal{M}) := \inf_\mu Inc_{\mathcal{M}}(\mu)$, also a property of a PDG that is motivated to capture inconsistency. Sice $Inc(\mathcal{M})$ is simply $\langle\!\langle \mathcal{M} \rangle\!\rangle$ without the qualitative term, it should come at no surprise that $\langle\!\langle - \rangle\!\rangle$ reduces to $Inc(-)$ in the limit of small γ.

?? You seem to be using your own idiosyncratic terminology, without motivation or explanation.

Prop 2.2. $\lim_{\gamma \to 0} \langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma = Inc(\mathcal{M}) = \langle\!\langle \mathcal{M} \rangle\!\rangle_0$.

[link to proof]

What about for fixed values of γ, or the limit as γ becomes large? These are also worth considering, and correspond to different weightings of the qualitative information (the set of edges and parameters α, indicating the qualitative dependency structure), against the quantitative information (the cpds $\mathbf{p}_L$ and the confidences β in them). We will write $\langle\!\langle \mathcal{M} \rangle\!\rangle$ without a subscript to indicate the function $\gamma \mapsto \langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma$, which we will use when a result does not depend on γ. ??

Although we defined $\langle\!\langle - \rangle\!\rangle$ in terms of $[\![-]\!]$, we note that it is also possible to do the reverse, defining $[\![-]\!]$ in terms of $\langle\!\langle - \rangle\!\rangle$. This is done by adding μ to $\mathcal{M}$ with large β. Intuitively, as our confidence that μ is the right joint distribution becomes large, the best distribution gets closer to μ.

Do you ever use this result?

Prop 2.3. $[\![\mathcal{M}]\!]_\gamma(\mu) = \langle\!\langle \mathcal{M} + \mu! \rangle\!\rangle_\gamma \quad \left(= \lim_{t \to \infty} \left\langle\!\!\left\langle \mathcal{M} + \overset{(\beta:t)}{\mu} \right\rangle\!\!\right\rangle_\gamma \right)$

please avoid this horrible notation

As a result, $\langle\!\langle - \rangle\!\rangle$ may be taken as primitive, and so PDG semantics can be couched in terms of inconsistency. For instance, the PDG $\mathcal{M}_{\mathcal{B}}$ for a Bayesian Network $\mathcal{B}$ is always consistent ($\Pr_{\mathcal{B}} \in \{\!\{\mathcal{M}_{\mathcal{B}}\}\!\}$, and $\langle\!\langle \mathcal{M}_{\mathcal{B}} \rangle\!\rangle = 0$) by construction, because $\mathcal{B}$ represents a distribution $\Pr_{\mathcal{B}}$ that has the appropriate conditional marginals and (in)dependence structure, and more explicitly, one might *define* the semantics of a PDG to be the distribution μ, such that, if combined with the data of $\mathcal{B}$, results in a maximally consistent PDG.

2.1 Inference by Gradiant Descent

We now return to the task at hand. The most important and standard query is a conditional probability query: given a PDG $\mathcal{M}$, how do you compute the probability of Y given X? We use a similar approach as we did in giving PDGs semantics in the first place —rather than giving probabilistic information directly, we instead give a measure the quality of a candidate answer $p(Y \mid X)$. Intuitively, a cpd $p(Y \mid X)$ makes for a good answer to the query if it is consistent with the other cpds in $\mathcal{M}$, and so we propose $\langle\!\langle \mathcal{M} + p \rangle\!\rangle$ as a measure of (dis)quality of the inference p. This is simply a definition, but we now verify that it has nice properties, which we might expect from such a measure of inference quality.

I doubt that disquality is a word …

Perhaps most importantly, the best cpd(s) according to this measure are the conditional marginals $\mu(Y \mid X)$ of the best distributions μ for $\mathcal{M}$.

Prop 2.4. *For all $\mathcal{M}$, $X, Y \in \mathcal{N}^{\mathcal{M}}$, and $\gamma > 0$, we have that* $\arg\min_{p: X \to \Delta Y} \langle\!\langle \mathcal{M} + p \rangle\!\rangle_\gamma = \left\{ \mu(Y \mid X) : \mu \in [\![\mathcal{M}]\!]^*_\gamma \right\}$.

[link to proof]

In the limit, of small γ, since there is only one such distribution, the expression beomes simpler.

Corollary 2.4.1. $[\![\mathcal{M}]\!]^*(Y \mid X)$ *uniquely minimizes* $p(Y \mid X) \mapsto \langle\!\langle \mathcal{M} + p \rangle\!\rangle_0$.

[link to proof]

So $p \mapsto \langle\!\langle \mathcal{M} + p \rangle\!\rangle$ gives the best scores to the marginals of distributions that the scoring semantics of $\mathcal{M}$ view as best. Even supposing we had oracle access to $\langle\!\langle - \rangle\!\rangle$, the prospect of having to enumerate all possible conditional probability distributions p to find the best one sound prohibitively expensive. Fortunately, it is not necessary, as the function $p \mapsto \langle\!\langle \mathcal{M} + p \rangle\!\rangle$ has properties which make for efficient search, given oracle access to $\langle\!\langle \mathcal{M} \rangle\!\rangle$. Most importantly, it is smooth and strictly convex, ensuring that gradient descent converges quickly.

Prop 2.5. *The function $p \mapsto \langle\!\langle \mathcal{M} + p \rangle\!\rangle_\gamma$ is smooth and strictly convex for small enough γ.*

[link to proof]

Conjecture 2.6. *For cpds $p \in \Delta(Y \mid X)$, the function $\langle\!\langle \mathcal{M} + p \rangle\!\rangle_\gamma$ is Lipshitz in p, on any compact region in the interior of $\Delta(Y \mid X)$*

[link to proof]

The lower bound on the convexity of the inconsistency, however, is controlled by γ.

I don't understand what "lower bound on convexity" means, nor why the results/ conjectures say anything about convexity

Conjecture 2.7. $\langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma$ *is continuous as a function of γ, and converges as $\gamma \to 0$.*

As a result, progress towards an optimum of an objective for $\gamma = \gamma_0$ will still be useful for the optimization problem at $(1-\epsilon)\gamma_0$, for small enough ϵ. This suggests using a variant of gradient descent in which in which γ decays to zero during the optimization process,

2.2 Updating via Inconsistency

If, rather than fixing $\mathcal{M}$ and optimizing p, we fix p and optimize $\mathcal{M}$, this same proces corresponds to updating rather than inference. For instance, in the case where $X = \mathbb{1}$, an observation $Y{=}y$ can be added to $\mathcal{M}$ in the form of an edge $\mathbb{1} \xrightarrow{\delta_y} Y$, getting the (possibly inconsistent) PDG $\mathcal{M}'$. One might hope that the distribution $[\![\mathcal{M}']\!]^*$ is the result of conditioning the best distribution $[\![\mathcal{M}]\!]^*$ on $Y{=}y$, as is true for factor graphs, but this is always the case.

> **Example 1.** Consider a PDG $\mathcal{M}$ containing $p(X)$ with confidence r, and $q(Y|X)$ with confidence s. Now condition on $Y = y$ by adding that event to the PDG, to get
>
> $$\mathcal{M} := \xrightarrow[(r)]{p} \boxed{X} \xrightarrow[(s)]{q} \boxed{Y} \qquad \text{and} \qquad \mathcal{M}' := \xrightarrow[(r)]{p} \boxed{X} \xrightarrow[(s)]{q} \boxed{Y} \xleftarrow{y}.$$
>
> We can easily see that $[\![\mathcal{M}]\!]^* = p(X)q(Y|X)$ doesn't depend on the confidences, since we can simultaneously satisfy p and q. After conditioning on $Y = y$, we get a distribution $[\![\mathcal{M}]\!]^*|Y{=}y$ proportional to $p(X)q(y|X)$. But some quick calculation reveals that $[\![\mathcal{M}']\!]^*$ is proportional to $p(X)q(y|X)^{r/s}$, with an additional exponent r/s that was not present before.
>
> So, if $r = s$, so that the data of $\mathcal{M}$ effectively picks out a probability distribution with uniform confidence, we get probabilistic conditioning, and otherwise there is distortion, as the lower-confidence is bent further, absorbing more of the inconsistency. □

So, even though adding the event $Y{=}y$ to the PDG is much like probabilistic conditioning, in that afterwards we only consider distributions μ for which $\mu(Y{=}y) = 1$, it is not always equivalent to probabilistic conditioning. We now argue that the answer given by the PD is more reasonable. which is lost when one throws away the data of the PDG and works with probability distributions directly, Suppose we have a good reasonto trust $q(Y|X)$ much more than $p(X)$. Perhaps $q(Y|X)$ describes a probabilistic program that we wrote ourselves and simply adds a small amount of noise, while $p(X)$ describes the distribution described by a low-quality simulation. Then when we observe $Y{=}y$, we should not ??

What if we only wanted to know $\Pr(Y \mid X = x)$, for a specific value of X, rather than the full cpd $\Pr(Y \mid X)$? To answer this question, recall that we have identified the event $X = x$ with the degenerate distribution $\delta_x(X)$, so it can be included in a PDG as an edge $-x\twoheadrightarrow \boxed{X}$. Why -

If $\mathcal{M}$ is the data of a joint distribution, then this is simply

2.3 Complexity of Computing Inconsistency

Taken together, the results in Section 2.2 show that an oracle for the degree of inconsistency is sufficient to efficiently do learn and draw inferences. Unfortunately this turns out to be a lot to ask for, in general.

Prop 2.8. *Deciding if $\mathcal{M}$ is consistent is NP-hard.* [link to proof]

Prop 2.9. *Computing $\langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma$ is #P-hard, for $\gamma > 0$.* [link to proof]

This is just a lower bound on the complexity of estimating $\langle\!\langle \mathcal{M} \rangle\!\rangle$.

What shall we make of these results? First, the asymptotic compelxity of exact inference is not great, but no worse than for Bayesian Networks or factor graphs.

There is still hope for approximating it, given that it has such nice properties — is convex (albeit in an exponentially large space), smooth, and monotonic (additional edges only increase $\langle\!\langle \mathcal{M} \rangle\!\rangle$).

⟨ INCOMPLETE ⟩

3 Bounded Tree-Width

Conjecture 3.1. *There is an algorithm for inference on PDGs that runs in polynomial time, for the class of PDGs with bounded tree-width.*

Prop 3.2 (Marov Property for PDGs)**.** *Suppose $\mathcal{M}_1$ and $\mathcal{M}_2$ are value-compatible PDGs, with respective sets of nodes $\mathbf{X}_1 := \mathcal{N}^{\mathcal{M}_1}$ and $\mathbf{X}_2 := \mathcal{N}^{\mathcal{M}_2}$. Then for all $\gamma > 0$, we have that*

link to proof

$$[\![\mathcal{M}_1 + \mathcal{M}_2]\!]^*_\gamma \;\models\; \mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2 \mid \mathbf{X}_1 \cap \mathbf{X}_2$$

That is, in every optimal distribution $\mu^* \in [\![\mathcal{M}_1 + \mathcal{M}_2]\!]^*_\gamma$ for some $\gamma > 0$, the variables of $\mathcal{M}_1$ and the variables of $\mathcal{M}_2$ are conditionally independent given the variables they have in common.

In the bounded tree-width setting, to specify a distribution with the Markov Property over the tree-decomposition, it suffices to specify a distribution over each component, subject to the constraint that the distributions have the same marginals. Concretely, a tree decomposition of a PDG $\mathcal{M}$ is a set of subsets of variables $\mathcal{C} \in 2^{2^{\mathcal{N}}}$ called components, such that:

1. Every variable is a member of at least one component.
2. Each edge lies entirely within some component. So for every $X \xrightarrow{L} Y \in \mathcal{E}^{\mathcal{M}}$, there is some $C_L \in \mathcal{C}$ such that $X, Y \in C$.
3. The (undirected) graph $\mathcal{G}(\mathcal{C}) = (\mathcal{C}, \mathcal{E}(\mathcal{C}))$, whose nodes are the components $\mathcal{C}$, and which has an edge between C_1 and C_2 (i.e., $\{C_1, C_2\} \in \mathcal{E}$) iff $C_1 \cap C_2 \neq \emptyset$, is a tree.

Now, because of the Markov property (Prop 3.2), any optimal distribution $\mu^* \in \bigcup_{\gamma>0} [\![\mathcal{M}]\!]^*_\gamma$ for $\mathcal{M}$, must factor over the tree structure, meaning that it can be specified by a collection of marginal probability distributions $\boldsymbol{\nu} = \{\nu_C(C) : C \in \mathcal{C}\}$ that is locally consistent (i.e., overlapping components agree on their shared variables: $C_1 \cap C_2 \neq \emptyset \implies \nu_{C_1}(C_1 \cap C_2) = \nu_{C_2}(C_1 \cap C_2)$). Given such a set of marginals $\boldsymbol{\nu}$, there is precisely one joint distribution that both matches those marginals and also has the independencies required by the Markov property; that distribution is given by

$$\Pr_{\boldsymbol{\nu}}(\cup \mathcal{C}) = \frac{\prod_{C \in \mathcal{C}} \nu_C(C)}{\prod_{\{C_1, C_2\} \in \mathcal{E}(\mathcal{C})} \nu_{C_1}(C_1 \cap C_2)}.$$

Furtheremore, the entropy $\mathrm{H}(\Pr_{\boldsymbol{\nu}})$ of such distributions (which include μ^*), can be analogously computed in terms of the marginals $\boldsymbol{\nu}$, via the Bethe entropy

$$\mathrm{H}(\Pr_{\boldsymbol{\nu}}) = \sum_{C \in \mathcal{C}} \mathrm{H}(\nu_C) - \sum_{\{C_1, C_2\} \in \mathcal{E}(\mathcal{C})} \mathrm{H}(\nu_{C_1}(C_1 \cap C_2)).$$

For a fixed tree decomposition $\mathcal{C}$, note that searching over distributions $\boldsymbol{\nu}$ is a constrained optimization problem, over a strictly (and γ-strongly) convex function, for small enough γ.

⟨ WORRY: smooth, compact, and γ-strongly convex might not be enough: How can we get complexity bounds for this minimization? All bounds I see require Lipshitz assumptions, which don't hold for relative entropy. ⟩

⟨ QUESTION: What about when γ is not small? Now the problem could be non-convex. Can we still say anything about it then? ⟩

⟨ This doesn't seem "fixed-parameter tractable", given that the number of parameters grows with the side of the tree... ⟩

3.1 The Quantitative Limit (small γ)

Our approach relies on the following key fact.

Prop 3.3. *For any PDG $\mathcal{M}$, the highest-compatibility distributions (the minimizers $[\![\mathcal{M}]\!]_0^*$ of $Inc_{\mathcal{M}}$) all have the same conditional probabilities along the edges of $\mathcal{M}$. That is to say, if there is an edge $X \xrightarrow{L} Y \in \mathcal{E}^{\mathcal{M}}$, and $\mu_1, \mu_2 \in [\![\mathcal{M}]\!]_0^*$ are quantitatively optimal distributions, then $\mu_1(Y|X) = \mu_2(Y|X)$.*

Proof. ⟨ TODO: add proof via zeros of the Hessian ⟩

□

Of course, the converse is true as well—$Inc_{\mathcal{M}}(\mu)$ depends only on the conditional marginals $p(Y|X)$ along edges $X \xrightarrow{L} Y \in \mathcal{E}^{\mathcal{M}}$. As a result, the set of distributions $[\![\mathcal{M}]\!]_0^*$ is characterized precisely by the set of linear constraints enforcing that the conditional marginals are the appropriate optimal cpds. And to find these optimal cpds, it suffices to find any qualitatively optimal distribution $\mu_0 \in [\![\mathcal{M}]\!]_0^*$, say by gradient descent as above.

Once we have a solution $\hat{\boldsymbol{\nu}}$ whose corresponding distribution $\Pr_{\hat{\boldsymbol{\nu}}}$ minimizes $Inc_{\mathcal{M}}$, we can perform a second, different minimization with further linear constraints: for each edge $X \xrightarrow{L} Y \in \mathcal{E}^{\mathcal{M}}$, we add $|\mathcal{V}(XY)|$ constraints ensuring that

$$\nu_{C_L}(Y|X) = \nu^0_{C_L}(Y|X) \qquad \iff \sum_{z \in \mathcal{V}(C_L \setminus \{X,Y\})} \nu_{C_L}(X, Y, z) = \hat{\nu}_{C_L}(Y|X) \Big(\sum_{z,y \in \mathcal{V}(C_L \setminus \{X\})} \nu_{C_L}(X, y, z) \Big),$$

for all possible values of X and Y. By Prop 3.3, the distributions that correspond to feasible points are precisely the minimizers of $Inc_{\mathcal{M}}$. We now perform a secondary optimization problem, with these additional constraints, to minimize $IDef_{\mathcal{M}}$, which can also be computed precisely in terms of the marginals $\boldsymbol{\mu}$.

Quick account of complexity. If every variable can take at most V different values, there are at most E edges, the maximum number of variables in any component is K (i.e, the bound on the tree width), and the number of components is M, then we now have a constrained optimization problem with

- at most MV^{2K} paramters, and
- at most $(M-1)V^{2K} + EV^2$ constraints:
  - at most $(M-1)V^{2K}$ marginal constraints to enforce the local polytope (one for each of the $M-1$ junctions),
  - and precisely EV^2 additional constraints in the second stage to ensure an optimizer of $Inc_{\mathcal{M}}$ when searching over $IDef_{\mathcal{M}}$.

⟨ Unlike $Inc_{\mathcal{M}}$, we should be able to get a Lipshitz constant for $IDef_{\mathcal{M}}$, to get a bound on complexity of minimization, but this would require more work. ⟩

4 Variational Inference for PDGs

Variational inference is a standard approximate inference technique for probabilistic models p, in which we use an (arbitrary) auxilary distribution q to directly estimate a feature of p which might be quite difficult to optimize directly, and in the process construct a worst-case bound on the quality of p. Inference can then by performed simultaneously optimizing p to do better with respect to this worst-case bound generated by q, and also optimizing eq so that it more closely matches p. This approach has historically been very fruitful, enabling axpproximations useful approximatate models of physical systems [], and approximate inference in latent variable models [**?**], such as variational autoencoders [].

There is a deep bidirectional connection between PDG semantics and variational techniques. In parallel work [**?**], we argue that PDGs offer a concise visual account of variational inference more generally, that their semantics automate the algebra behind the variational bounds, and we can embue the approach with more

imbue

intuition by doing so. In the present document, we focus on a simpler aspect of this connection: how this variational approach can be applied to draw inferences in PDGs.

In our context, the approach boils down to a very simple observation: the incompatibility $[\![\mathcal{M}]\!]_\gamma(\mu)$ of a PDG $\mathcal{M}$ with a *specific* distribution μ is at least as large as the smallest possible incompatibility of $\mathcal{M}$ with *any* distribution, which is the inconsistency $\langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma$.

So if we choose a nice class of test distributions $\mathcal{M} \subset \Delta(\mathcal{M})$, the variational approch suggests the following procedure:

1. optimize the cpds of $\mathcal{M}$ so to reduce the incompatibility with μ;
2. optimize μ to reduce its incompatibility with $\mathcal{M}$;
3. repeat.

Moreover, since $[\![\mathcal{M}]\!]_\gamma(\mu)$ is strictly convex in $\mathcal{M}$ and γ-strongly convex in μ, this procedure gives a family of inference algorithms for PDGs in line with standard algorithms used to train neural networks.

we note that from Prop 2.3, it follows that this entire procedure is just minimizing the inconsistency of a single PDG $\langle\!\langle \mathcal{M} + \mu \rangle\!\rangle$, where μ has very high confidence. Thus, this approach to inference may be summarized as follows.

> *Adding an edge to a PDG makes it no less inconsistent, but may make the degree of inconsistency easier to calculate — which in turn makes more efficient to minimize, ultimately allowing us to to draw inferences quickly.*

4.1 A Nice Class of Distributions

Recall the Markov Property:

Prop 3.2. *Suppose $\mathcal{M}_1$ and $\mathcal{M}_2$ are value-compatible PDGs, with respective sets of nodes $\mathbf{X}_1 := \mathcal{N}^{\mathcal{M}_1}$ and $\mathbf{X}_2 := \mathcal{N}^{\mathcal{M}_2}$. Then for all $\gamma > 0$, we have that*

$$[\![\mathcal{M}_1 + \mathcal{M}_2]\!]^*_\gamma \;\models\; \mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2 \mid \mathbf{X}_1 \cap \mathbf{X}_2$$

That is, in every optimal distribution $\mu^* \in [\![\mathcal{M}_1 + \mathcal{M}_2]\!]^*_\gamma$ for some $\gamma > 0$, the variables of $\mathcal{M}_1$ and the variables of $\mathcal{M}_2$ are conditionally independent given the variables they have in common.

Moreover, we have a stronger result. Given a dependency graph $\mathcal{G} = (\mathcal{N}, \mathcal{E}, \mathcal{V})$, let $\mathcal{P}_{\mathcal{M}[\mathcal{G}]}$ denote the set of distributions obtainable as optimizers $\mu^* \in \arg\min [\![\mathcal{M}]\!]^*_\gamma$, for some PDG $\mathcal{M} = (\mathcal{G}, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ based on $\mathcal{G}$, by supplying some cpds $\mathbf{p}$, weights $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and $\gamma > 0$. Analogously, let $\mathcal{P}_{\Phi[\mathcal{G}]}$ be the set of distributions represented by some factor graph whose factors touch the same variables as $\mathcal{G}$.

Prop 4.1. $\mathcal{P}_{\mathcal{M}[\mathcal{G}]} = \mathcal{P}_{\Phi[\mathcal{G}]}$.

Proof. ⟨ TODO: fill in Chris's proof from email ⟩

□

Thus, to find for optimal distributions, it suffices to search over the space of factor graphs.

4.2 Experiments

The major drawback of variational inference is that its effectiveness is fundamentally tied to the set of test distributions $\mathcal{M}$ that we consider. A poor choice of $\mathcal{M}$ makes it impossible not to be extremely inconsistent, resulting in unavoidably poor-quality inferencees. For this reason, a variational approach requires emperical validation. In the rest of this section, we preseent some toy simulations that suggest that indeed the approach can be useful for a broad class of PDGs that are otherwise difficult to analyze.

⟨ INCOMPLETE ⟩

5 The Local Inconsistency Reduction (LIR) Algorithm

In Section 2.3, we saw that the cost of computing the inconsistency of a PDG scales poorly with the size of the PDG. But do we really need to know the inconsistency of the *whole* PDG, to effectively estimate the marginal on a variable on the fringe of the graph? For particularly simple (tree-structured) topologies, the answer is no. More generally, might computing the inconsistency of a small neighborhood be enough to do inference?

More concretely, starting with a base pdg $\mathcal{M}$, we perform an iterative updating procedure, where at each step t, we look at some local context $\mathcal{C} \subset \mathcal{M}$, and tweak some subset of the cpds $\mathcal{A} \subset \mathcal{C}$ that hold our attention, to reduce the (local) inconsistency of $\mathcal{C}$. This is done with a gradient descent step, with step size η.

This is the idea behind Local Inconsistency Reduction (Algorithm 1).

Algorithm 1 Local Inconsistency Reduction (LIR) algorithm

Require: A pdg $\mathcal{M}$
Require: A pdg $\mathcal{B}$ of inferred beliefs
Require: A sequence $(A_1 \subset C_1), (A_2 \subset C_2), \ldots$ of subgraphs of $\mathcal{G}(\mathcal{M} + \mathcal{B})$, corresponding to
Require: A tradoff parameter $\gamma \geq 0$

procedure LIR($\mathcal{M}, \mathcal{B}, \mathcal{R}, \gamma$)
  for $t = 1, 2, \ldots$ **do**
    Let $\mathcal{L}^{(t)} := (\mathcal{M} + \mathcal{B})|_{\mathcal{G}_t}$ be the restriction of $\mathcal{M} + \mathcal{B}$ to the subgraph R_t
    for all edges $L \in \mathcal{E}^{\mathcal{B}|_{R_t}}$ **do**
      Update $\mathbf{p}_L \leftarrow \mathbf{p}_L - \eta \vec{\nabla}_{\mathbf{p}_L} \langle\!\langle \mathcal{L} \rangle\!\rangle_\gamma$
  return $\mathcal{B}$.

The specification of the local contexts and attention

⟨*under construction* ⟩

Definition 5.1 (belief propogation)**.** The messages passed are given by:

$$n_{i \to a}(x_i) := \prod_{c \in \partial a \setminus i} m_{c \to i}(x_i) \qquad m_{a \to i}(x) := \sum_{\mathbf{x}_a \setminus x_i} f_a(\mathbf{x}_a) \prod_{j \in \partial a \setminus i} n_{j \to a}(x_j) \tag{1}$$

the message sent from a vertex v on an edge e is the product of the local function at v, with all messages recieved at v (other than from e), summarized for e's variable node. [?]

The variable beliefs are defined as a function of the messages

$$b_X(x) \propto \prod_f m_{f \to x}(x),$$

so that □

In previous work [?], we showed that $\gamma = 1$ essentially regards an unweighted PDG $\mathcal{M}$ (i.e, where $\alpha = \beta$) as a factor graph. We now show that for $\gamma = 1$, the LIR algorithm performs belief propogation. To that end, we now review the presentation of belief propogation given by [?].

Definition 5.2 (cluster graph)**.** A cluster graph $\mathcal{U} = (V, E, \mathbf{C}, \mathbf{S})$ for a colelction of factors $\Phi = \{\phi_J : \mathbf{X}_J \to \mathbb{R}\}_{J \in \mathcal{J}}$ over variables $\mathbf{X}$ is

- an undirected graph (V, E), where the vertex set V is a partition of $\mathcal{J}$, so that there is a unique vertex $V(J) \in V$ for each $J \in \mathcal{J}$;
- a subset $\mathbf{C}_v \subset \mathbf{X}$ of variables (a "cluster") for each vertex $v \in V$, such that $\mathbf{X}_J \subset \mathbf{C}_{V(J)}$;
- a subset $\mathbf{S}_{(u,v)} \subset \mathbf{C}_u \cap \mathbf{C}_v$ of variables (a "sepset") for each edge $(u, v) \in E$.

□

Two extremal cluster graphs for $\Phi = \{\phi\}_{\mathcal{J}}$ include:

$\mathcal{U}_0 :=$ the one-node (no-edge) cluster graph whose lone cluster contains all factors;

$\mathcal{U}_{\text{Bethe}} :=$ the cluster graph where each factor has its own vertex (i.e., with $V = \mathcal{J}$ and $\mathbf{C}_J = \mathbf{X}_J$) containing every edge between any two vertices sharing a variable, so that with $S_{(J,K)} = \mathbf{X}_J \cap \mathbf{X}_K$.

Definition 5.3. □

Prop 5.1. *If Φ is a set of factors, $\mathcal{U}$ is a cluster graph for Φ, and $E_1, E_2, \ldots$ is a sequence of edges of $\mathcal{U}$, then LIR with parameters:*

$$\mathcal{M} = \mathcal{M}_\Phi, \qquad \mathcal{B} = \left\{ \mathbf{p}_X^{(\beta:)} \propto 1 \;\middle|\; X \in \mathcal{N}^{\mathcal{M}} \right\} + \left\{ \mathbf{p}_E \propto \prod_{\phi : v_\phi = v} \phi \;\middle|\; v \in V^{\mathcal{U}} \right\}, \qquad \mathcal{L}_i = E_i, \qquad \gamma = 1$$

is equivalent to belief propogation, in the sense that the cpd $\mathbf{p}_E$ of $\mathcal{B}$ at timestep t is equal to the cluster belief $b_E^{(t)}$ in cluster-graph belief propagation, for every timestep t.

Proof. □

Prop 5.2. *If $\mathcal{N}$ is an unweighted PDG, then* $\mathrm{LIR}^*(\mathcal{N}, \mathcal{U}_{\mathrm{Bethe}}) = \mathrm{BP}(\Phi_{\mathcal{N}}, \mathcal{U}_{\mathrm{Bethe}})$

⟨*under construction* ⟩

Prop 5.3. *If Φ is a factor graph (as a bipartite graph), and $(u_1, v_1), (u_2, v_2), \ldots$ is a sequence of oriented edges of Φ, then*

$$\mathcal{M} = \mathcal{M}_\Phi, \qquad \mathcal{B} = \left\{ \mathbf{p}_X \propto 1 \;\middle|\; X \in \mathcal{N}^{\mathcal{M}} \right\} + \left\{ \mathbf{p}_\phi \propto \phi \;\middle|\; \phi \in \Phi \right\}, \qquad \mathcal{L}_i =, \qquad \gamma = 1$$

is equivalent to belief propogation, in the sense that the cpd $\mathbf{p}_E$ of $\mathcal{B}$ at timestep t is equal to the cluster belief $b_E^{(t)}$ in cluster-graph belief propagation, for every timestep t.

Prop 5.4.

Dampening.

5.1 Message Passing and Divergences

In [], Minsky shows that many message-passing algorithms may be veiwed as local divergence minimization. We now show how this picture fits int ours.

Choose your favorite class of distributions $\mathcal{Q} \subset \Delta\mathcal{M}$, from which you can sample, and for which you can compute the relevant conditional entropies $\mathrm{H}_q(Y \mid X)$ for every edge $X \xrightarrow{L} Y$ in your pdg $\mathcal{M}$.

Initially, start with some $q^0 \in \mathcal{Q}$. As before, we can compute the score of q^0 by

$$[\![\mathcal{M}]\!]_\gamma(q^0) = \mathop{\mathbb{E}}_{q^0}\left[\sum_{X \xrightarrow{L} Y} \beta_L \log \frac{q^0(Y \mid X)}{\mathbf{p}_L(Y \mid X)} \right] + \left[\sum_{X \xrightarrow{L} Y} \alpha_L \, \mathrm{H}_{q^0}(Y \mid X) \right] - \mathrm{H}(q^0)$$

We can define a transformation $\mathrm{T} : \mathcal{Q} \to \mathcal{Q}$ by

$$\mathrm{T}(q) := \arg\min_{q' \in \mathcal{Q}}$$

That is, we assume that it has some of the information-theoretic properties that make

5.2 Generalized Belief Propogation

Definition 5.4. A PDG is said to be *qualitatively exact* if the sum of the qualitative weights into each variable is equal to 1. That is, if

$$\forall Y \in \mathcal{N}. \quad \sum_{X \xrightarrow{L} Y} \alpha_L = 1$$

□

Prop 5.5. *A*

6 Axiom Systems

6.1 MAC axiom system

We begin with three simple inference rules, central to probabilistic reasoning: **M**arginalization, **A**pplication, and **C**onditioning.

M. Given $\Pr(X, Y) = p$, infer $\Pr(X) = \sum_y p(X, Y)$.
A. Given $\Pr(Y \mid X) = p$ and $\Pr(X) = q$, infer $\Pr(X, Y) = p(Y \mid X)q(X)$.
C. Given $\Pr(X, Y) = p$ and $\Pr(X) = q$, infer $\Pr(Y \mid X) = p(X, Y)/q(X)$.

Each axiom corresponds to a transformer on PDGs. For instance, axiom **M** corresponds to the function

$$\mathit{infer}^{\mathbf{M}}_{X,Y}\Big(p(X,Y) + q(X) + \mathscr{R}\mathit{est}\Big) := p + q + (X \xrightarrow{p/q} Y) \; + \mathscr{R}\mathit{est},$$

which takes a PDG containing data about a joint distribution $p(X, Y)$ and a marginal $q(X)$, and adds a new edge with cpd $\Pr(Y \mid X) = p(X, Y)/q(X)$.

Prop 6.1. *The BU algorithm is is repeated application of **M**,*

6.2 Constraint Agglomeration

7 Discussion

One objective of the PDG formalism is to model the internal state of a bounded agent, and such an agent could easily have absorbed more information than they have been able to process. Thus, we view PDG inference algorithms not as ways of getting "the one true answer" out of a modeling tool, but rather as descriptions of what an agent can do in order to:

1. identify and rank inconsistencies in their beliefs by graveness.
2. make progress towards resolving such inconsistencies; and
3. respond to questions without needing to first attain perfect epistemic clarity.

A Proofs and Lemmas

Lemma A.1. *$\boldsymbol{D}(\mu \parallel \nu)$ is a k_μ-strongly in ν, for fixed μ, where $k_\mu > 0$ is a constant that depends on μ.*

Proof.

$$\begin{aligned}\boldsymbol{D}(\mu \parallel \nu) &= \mathop{\mathbb{E}}_{x\sim\mu} \log \frac{\mu(x)}{\nu(x)} \\ &= \mathop{\mathbb{E}}_{x\sim\mu} \log \mu(x) + \mathop{\mathbb{E}}_{x\sim\mu} \log \frac{1}{\nu(x)}\end{aligned}$$

As the first term depends only on μ, it suffices to consider the behavior of the second term, as a function of ν. Let $F(\nu) = -\mathbb{E}_{x\sim\mu} \log \nu(x)$. Then $\nabla F(\nu) = x \mapsto \frac{\mu(x)}{\nu(x)}$. Let $k^\mu_{\nu_1,\nu_2} := \inf_x \frac{\mu(x)}{\nu_1(x)\nu_2(x)}$. Expanding the inner product of the difference in between ν_1 and ν_2, we have

$$\begin{aligned}(\nabla F(\nu_1) - \nabla F(\nu_2)) \cdot (\nu_1 - \nu_2) &= \sum_x \left(\frac{\mu(x)}{\nu_1(x)} - \frac{\mu(x)}{\nu_2(x)}\right)(\nu_2(x) - \nu_1(x)) \\ &= \sum_x \mu(x) \left(\frac{\nu_2(x)}{\nu_1(x)\nu_2(x)} - \frac{\nu_1(x)}{\nu_1(x)\nu_2(x)}\right)(\nu_2(x) - \nu_1(x)) \\ &= \sum_x \mu(x) \left(\frac{1}{\nu_1(x)\nu_2(x)}\right)(\nu_2(x) - \nu_1(x))^2\end{aligned}$$

□

Prop 2.4. *For all $\mathcal{M}$, $X, Y \in \mathcal{N}^{\mathcal{M}}$, and $\gamma > 0$, we have that* $\arg\min_{p:X\to\Delta Y} \langle\!\langle \mathcal{M} + p \rangle\!\rangle_\gamma = \left\{\mu(Y \,|\, X) : \mu \in [\![\mathcal{M}]\!]^*_\gamma\right\}$.

Proof. Because $\alpha_p = 0$, the new cpd p gives the resulting cpd one additional term in its scoring function, equal to the expected divergence from p to the appropriate marginal of μ, giving us

$$[\![\mathcal{M} + p]\!]_\gamma(\mu) = [\![\mathcal{M}]\!]_\gamma(\mu) + \mathop{\mathbb{E}}_{x\sim\mu_X} \boldsymbol{D}\Big(\mu(Y \mid x) \,\Big\|\, p(Y \mid x)\Big)$$

Gibbs inequality tells us that the second term is non-negative, and zero if and only if $\mu(Y \mid X) = p(Y \mid X)$.

If μ minimizes the first term, (i.e., $\mu \in [\![\mathcal{M}]\!]^*_\gamma$), then by definition we have $[\![\mathcal{M}]\!]_\gamma(\mu) = \langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma$ and so by choosing $p := \mu(Y \mid X)$, we get

$$[\![\mathcal{M} + p]\!]_\gamma(\mu) = [\![\mathcal{M} + \mu(Y \,|\, X)]\!]_\gamma(\mu) = [\![\mathcal{M}]\!]_\gamma(\mu) = \langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma$$

but $\langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma \le \langle\!\langle \mathcal{M} + p \rangle\!\rangle_\gamma$ for all p, so $\inf_p \langle\!\langle \mathcal{M} + p \rangle\!\rangle_\gamma = \langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma$, and $\mu(Y \,|\, X)$ minimizes $\langle\!\langle \mathcal{M} + p \rangle\!\rangle_\gamma$. This shows that $\left\{\mu(Y \,|\, X) : \mu \in [\![\mathcal{M}]\!]^*_\gamma\right\} \subseteq \arg\min_p \langle\!\langle \mathcal{M} + p \rangle\!\rangle$.

Conversely, suppose $p(Y \mid X)$ cannot be expressed as a conditional marginal $\mu_0(Y \mid X)$ for any $\mu_0 \in [\![\mathcal{M}]\!]^*_\gamma$. Now, one the one hand, for all μ in $[\![\mathcal{M}]\!]^*_\gamma$, we have

$$[\![\mathcal{M} + p]\!]_\gamma(\mu) = \langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma + \mathop{\mathbb{E}}_{x\sim\mu_X} \boldsymbol{D}\big(\mu(Y \mid x) \,\big\|\, p(Y \mid x)\big) > \langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma,$$

where the strict inequality follows from the fact that $p(Y \mid X) \neq \mu(Y \mid X)$ and Gibbs' inequality. But on the other hand, if $\mu \notin [\![\mathcal{M}]\!]^*_\gamma$, then $[\![\mathcal{M}]\!]_\gamma(\mu) > \langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma$, so we have

$$[\![\mathcal{M} + p]\!]_\gamma(\mu) \ge [\![\mathcal{M}]\!]_\gamma(\mu) > \langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma.$$

Putting the two cases together, we find that for every joint distribution μ, $[\![\mathcal{M} + p]\!]_\gamma(\mu) > \langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma$, and since $\langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma = \inf_p \langle\!\langle \mathcal{M} + p \rangle\!\rangle_\gamma$, we know that p does not minimize $\langle\!\langle \mathcal{M} + p \rangle\!\rangle_\gamma$.

Summarizing the argument, we have shown that

$$p \notin \arg\min_p \langle\!\langle \mathcal{M} + p \rangle\!\rangle \implies p \notin \arg\min_{p'} \langle\!\langle \mathcal{M} + p' \rangle\!\rangle.$$

Taking the contrapositive gives the desired inclusion $\{\mu(Y \mid X) : \mu \in [\![\mathcal{M}]\!]^*_\gamma\} \supseteq \arg\min_p \langle\!\langle \mathcal{M} + p \rangle\!\rangle$.

□

??. *

Proof. The argument used in the proof of Prop 2.4 works uniformly for all γ, and so the limit factors through it (more concretely, the proof remains valid if we insert a limit as $\gamma \to 0$ in front of every $\langle\!\langle - \rangle\!\rangle_\gamma$ or $[\![-]\!]_\gamma$). Since $[\![\mathcal{M}]\!]^*$ the unique element of $\lim_{\gamma\to 0}[\![\mathcal{M}]\!]^*_\gamma$, $[\![\mathcal{M}]\!]^*_\gamma(Y \mid X)$ is the unique element of $\arg\min_p \langle\!\langle \mathcal{M} + p \rangle\!\rangle_0$. □

We will make use of the implicit function theorem in the next result.

Implicit Function Theorem (Dini)**.** *Suppose $Z \subseteq \mathbb{R}^n \times \mathbb{R}^m$ is open, and has coordinates $(y_1, \ldots, y_n, x_1, \ldots, x_n)$. If $\phi : Z \to \mathbb{R}^m$ is a k-times continuously differentiable function, and $(\mathbf{b}, \mathbf{a}) \in Z$ is such that $\phi(\mathbf{b}, \mathbf{a}) = \mathbf{0}$, and if the Jacobian matrix*

$$\mathbf{J}_{\phi,x}(\mathbf{b}, \mathbf{a}) = \left[\frac{\partial \phi_i}{\partial x_j}(\mathbf{b}, \mathbf{a})\right]_{i,j}$$

is an invertible matrix, then there exists an open set $U \subset Y$ containing $\mathbf{b}$ such that there is a unique k-times differentiable function $g : U \to X$ such that $g(\mathbf{b}) = \mathbf{a}$, and $\phi(\mathbf{y}, g(\mathbf{y})) = 0$ for all $\mathbf{y} \in U$.

Prop 2.5. *The function $p \mapsto \langle\!\langle \mathcal{M} + p \rangle\!\rangle_\gamma$ is smooth and strictly convex for small enough γ.*

Proof. We start by expanding the definitions, obtaining

$$\begin{aligned}
\langle\!\langle \mathcal{M} + p \rangle\!\rangle_\gamma &= \inf_\mu \ [\![\mathcal{M} + p]\!]_\gamma(\mu) \\
&= \inf_\mu \left[[\![\mathcal{M}]\!]_\gamma(\mu) + \mathop{\mathbb{E}}_{x \sim \mu_X} \boldsymbol{D}\Big(\mu(Y \mid x) \,\Big\|\, p(Y \mid x)\Big)\right] \\
&= \inf_\mu \left[[\![\mathcal{M}]\!]_\gamma(\mu) + \boldsymbol{D}\Big(\mu(X, Y) \,\Big\|\, p(Y \mid X)\,\mu(X)\Big)\right].
\end{aligned}$$

Fix $\gamma < \min_L \beta_L$. Then we know that $[\![\mathcal{X}]\!]_\gamma(\mu)$ is a γ-strongly convex function for every PDG $\mathcal{X}$, and hence there is a unique joint distribution which minimizes it.

Strict Convexity. Suppose $p_1(Y \mid X)$ and $p_2(Y \mid X)$ are two cpds on Y given X. Fix $\lambda \in [0, 1]$, and set $p_\lambda = (1 - \lambda)p_1 + \lambda p_2$. Let μ_1, μ_2 and μ_λ be the joint distributions that minimze $[\![\mathcal{M} + p_1]\!]_\gamma$, $[\![\mathcal{M} + p_2]\!]_\gamma$ and $[\![\mathcal{M} + p_\lambda]\!]_\gamma$, respectively. Then we have

$$\langle\!\langle \mathcal{M} + p_\lambda \rangle\!\rangle_\gamma = [\![\mathcal{M}]\!]_\gamma(\mu_\lambda) + \boldsymbol{D}\Big(\mu_\lambda(X, Y) \,\Big\|\, p_\lambda(Y \mid X)\mu_\lambda(X)\Big).$$

By convexity of $[\![\mathcal{M}]\!]$ and $\boldsymbol{D}$, we have

$$[\![\mathcal{M}]\!]_\gamma(\mu_\lambda) \le (\lambda - 1)[\![\mathcal{M}]\!]_\gamma(\mu_1) + \lambda[\![\mathcal{M}]\!]_\gamma(\mu_2) \tag{2}$$

and

$$\begin{aligned}
\boldsymbol{D}\Big(\mu_\lambda(XY) \,\Big\|\, p_\lambda(Y \mid X)\mu_\lambda(X)\Big) \le (1 - \lambda)\boldsymbol{D}\Big(\mu_1(XY) \,\Big\|\, p_1(Y \mid X)\mu_1(X)\Big) \\
+ \lambda \ \boldsymbol{D}\Big(\mu_2(XY) \,\Big\|\, p_2(Y \mid X)\mu_2(X)\Big).
\end{aligned} \tag{3}$$

If $\mu_1 \neq \mu_2$ then since $[\![\mathcal{M}]\!]$ is strictly convex, (2) must be a strict inequality. On the other hand, if $\mu_1 = \mu_2$, then since $\mu_\lambda = \mu_1 = \mu_2$ and $\boldsymbol{D}$ is stricly convex in its second argument when its first argument

is fixed (Lemma A.1), (3) must be a strict inequality. In either case, the sum of the two inequalities must be strict, giving us

$$\begin{aligned}\langle\!\langle \mathcal{M} + p_\lambda \rangle\!\rangle_\gamma &= [\![\mathcal{M}]\!]_\gamma(\mu_\lambda) + \boldsymbol{D}\Big(\mu_\lambda(XY) \,\Big\|\, p_\lambda(Y\,|\,X)\mu_\lambda(X)\Big) \\ &< (\lambda - 1)\left[[\![\mathcal{M}]\!]_\gamma(\mu_1) + \boldsymbol{D}\Big(\mu_1(XY) \,\Big\|\, p_1(Y\,|\,X)\mu_1(X)\Big)\right] \\ &\qquad + \lambda\left[[\![\mathcal{M}]\!]_\gamma(\mu_2) + \boldsymbol{D}\Big(\mu_2(XY) \,\Big\|\, p_2(Y\,|\,X)\mu_2(X)\Big)\right] \\ &= (\lambda - 1)\langle\!\langle \mathcal{M} + p_1 \rangle\!\rangle + \lambda\,\langle\!\langle \mathcal{M} + p_2 \rangle\!\rangle,\end{aligned}$$

which shows that $\langle\!\langle \mathcal{M} + p \rangle\!\rangle$ is *strictly* convex in p, as desired.

Smoothness. If $[\![\mathcal{M} + p]\!]^*_\gamma$ is a positive distribution, then by definition $[\![\mathcal{M} + p]\!]$ achieves its minimum on the interior of the probability simplex $\Delta\mathcal{V}(\mathcal{M} + p)$, and so by Lemma A.2, we immediately find that $\langle\!\langle \mathcal{M} + p \rangle\!\rangle_\gamma$ is smooth in p.

Now, suppose that $[\![\mathcal{M} + p]\!]^*_\gamma(\mathbf{w}) = 0$, for some $\mathbf{w} \in \mathcal{V}(\mathcal{M} + p)$.

Applying Lemma A.2 to the function $f = [\![\mathcal{M}]\!]_\gamma$

Now for the second case.

⟨ INCOMPLETE ⟩

If $x^*_b \in \partial X$, then we claim that either

1. There is a subspace $T \subseteq \mathbb{R}^m$ with $\{\!\{\}\!\}$
2. There is a subspace $S \subseteq \mathbb{R}^n$ with $x^*_b \in S \cap \partial X$ such

□

Lemma A.2. *Let X and Y be convex sets, and $f : X \times Y \to \mathbb{R}$ be a smooth (C^∞), convex function. If f is strictly convex in X, and for some $y_0 \in Y$, $f(x, y_0)$ achieves its infemum on the interior of X. then $y \mapsto \inf_x f(x, y)$ is smooth (C^∞) at the point y_0.*

Proof. Let $x^*_0 := \arg\min_x f(x, y_0)$, which is achieved by assumption, and is unique because $f(-, y_0)$ is strictly convex.

We will ultimately apply the implicit function theorem to give us a smooth function which is equal to this infemum, but to do so we must deal with the technicality that it requires an open set; the boundary is the most complicated part of this result. Here we have essentially required that the domain be open by fiat for X, but for Y (which is a possibly non-open subset of $\mathbb{R}^m$), we use the Extension Lemma for smooth functions [**?**, Lemma 2.26]. In our context, it states that for every open set U with $\overline{Y} \subseteq U \subseteq \mathbb{R}^m$, there exists a function $\tilde{f} : X \times \mathbb{R}^m \to \mathbb{R}$, such that $\tilde{f}|_Y = f$ (and $\operatorname{supp} \tilde{f} \subseteq U$). We only need a small fraction of this power: that we can smoothly extend f to *some* open set of $\mathbb{R}^m$, which we fix and call $\tilde{Y}$.

We claim that now all conditions for the Implicit Function Theorem are met if invoked with $\phi(y, x) := \vec{\nabla}_x \tilde{f}(x, y)$ and $(\mathbf{b}, \mathbf{a}) = (y_0, x^*_0)$. Concretely, we have $m = dim\, X$, $n = dim\, Y$, and $Z = (\tilde{Y} \times X)^\circ$, i.e., the interior of $\tilde{Y} \times X$, which is open and contains $(\mathbf{b}, \mathbf{a})$. Becuase ϕ is smooth, it is k-times differentiable for all k. We have $\vec{\nabla}_x \tilde{f}(y_0, x^*_0) = \vec{0}$ because x^*_0 is a local minimum of the smooth function $\tilde{f}(-, y_0)$ which lies on the interior of X.

Moreover, the Jacobian matrix

$$\mathbf{J}_{\nabla \tilde{f}, x}(y_0, x^*_0) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}(x^*_0, y_0)\right]$$

is the Hessian of the strictly convex funtion $f(-, b)$, and therefore positive definite (and in particular non-singular). Therefore, the Implicit Function Theorem guarantees us the existence of a neighborhood

$U \subset \tilde{Y}$ of y_0 for which there is a unique k-times differentiable function $g : U \to X$ such that $g(y_0) = x_0^*$ and $\vec{\nabla}_x \tilde{f}(y, g(y)) = 0$ for all $y \in U$. Of course, this implies $g(y) = \arg\min_x f(x, y)$ at every such point, and $\inf_x f(x, y) = f(g(y), y)$ is a composition of the smooth function f with the k-times differentiable function $g \otimes \mathrm{id}_Y$. Therefore, $\inf_x f(x, y)$ is itself k-times continuously differentiable at y_0 for all k, or in other words, $\inf_x f(x, y)$ is smooth at $y = y_0$. □

Prop 2.2. $\lim_{\gamma \to 0} \langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma = Inc(\mathcal{M}) = \langle\!\langle \mathcal{M} \rangle\!\rangle_0$.

Proof. Unwrapping the definitions, we have

$$\begin{aligned} \langle\!\langle \mathcal{M} \rangle\!\rangle_0 = \lim_{\gamma \to 0^+} \langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma &= \lim_{\gamma \to 0^+} \inf_\mu [\![\mathcal{M}]\!]_\gamma(\mu) \\ &= \lim_{\gamma \to 0^+} \inf_\mu \Big[Inc_{\mathcal{M}}(\mu) + \gamma IDef_{\mathcal{M}}(\mu) \Big] \end{aligned}$$

Since $IDef_{\mathcal{M}}$ is bounded above and below by constants $k \le IDef_{\mathcal{M}} \le K$, we have, for all γ and μ, that

$$Inc_{\mathcal{M}}(\mu) + \gamma k \le Inc_{\mathcal{M}}(\mu) + \gamma IDef_{\mathcal{M}}(\mu) \le Inc_{\mathcal{M}}(\mu) + \gamma K.$$

Since this holds for all μ, $Inc_{\mathcal{M}}$ and $IDef_{\mathcal{M}}$ are bounded below, and set of possible distriutions $\mu \in \Delta\mathcal{V}(\mathcal{M})$ is compact, the infemum $\inf_\mu [Inc_{\mathcal{M}}(\mu) + \gamma IDef_{\mathcal{M}}(\mu)]$ is achieved by some μ^*, for which

$$Inc_{\mathcal{M}}(\mu^*) + \gamma k \le Inc_{\mathcal{M}}(\mu^*) + \gamma\ IDef_{\mathcal{M}}(\mu^*) \le Inc_{\mathcal{M}}(\mu^*) + \gamma K$$

and of course, $Inc_{\mathcal{M}}(\mu^*) = Inc(\mathcal{M})$ by definition of the latter, so

$$Inc(\mathcal{M}) + \gamma k \le \inf_\mu \Big[Inc_{\mathcal{M}}(\mu) + \gamma IDef_{\mathcal{M}}(\mu) \Big] \le Inc(\mathcal{M}) + \gamma K.$$

Taking the limit as $\gamma \to 0$ and using the squeeze theorem, we find that

$$Inc(\mathcal{M}) = \lim_{\gamma \to 0^+} \inf_\mu \Big[Inc_{\mathcal{M}}(\mu) + \gamma IDef_{\mathcal{M}}(\mu) \Big] = \langle\!\langle \mathcal{M} \rangle\!\rangle_0, \qquad \text{as desired.}$$

□

A.1 Conditioning Results

??. *

A.2

Prop 3.2. *Suppose $\mathcal{M}_1$ and $\mathcal{M}_2$ are value-compatible PDGs, with respective sets of nodes $\mathbf{X}_1 := \mathcal{N}^{\mathcal{M}_1}$ and $\mathbf{X}_2 := \mathcal{N}^{\mathcal{M}_2}$. Then for all $\gamma > 0$, we have that*

$$[\![\mathcal{M}_1 + \mathcal{M}_2]\!]^*_\gamma \ \models\ \mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2 \mid \mathbf{X}_1 \cap \mathbf{X}_2$$

That is, in every optimal distribution $\mu^* \in [\![\mathcal{M}_1 + \mathcal{M}_2]\!]^*_\gamma$ for some $\gamma > 0$, the variables of $\mathcal{M}_1$ and the variables of $\mathcal{M}_2$ are conditionally independent given the variables they have in common.

Proof. Choose $\mu \in [\![\mathcal{M}_1 + \mathcal{M}_2]\!]^*_\gamma$. Let $\mu' := \mu(\mathcal{N}_1)\mu(\mathcal{N}_2)$

⟨ Finish Transcribing Proof ⟩

□

A.3 Hardness Results

Prop 2.8. *Deciding if $\mathcal{M}$ is consistent is NP-hard.*

Proof. We can directly encode SAT problems as PDGs. Specifically, let

$$\varphi := \bigwedge_{j \in \mathcal{J}} \bigvee_{i \in \mathcal{I}(j)} (X_{j,i})$$

be a CNF formula over binary variables $\mathbf{X} := \bigcup_{j,i} X_{j,i}$. Let $\mathcal{M}_\varphi$ be the PDG containing every variable $X \in \mathbf{X}$ and a binary variable C_j (taking the value 0 or 1) for each clause $j \in \mathcal{J}$, as well as the following edges, for each $j \in \mathcal{J}$:

- a hyper-edge $\{X_{j,i} : i \in \mathcal{I}(j)\} \twoheadrightarrow C_j$, together with a degenerate cpd encoding the boolean OR function (i.e., the truth of C_j given $\{X_{j,i}\}$);
- an edge $\mathbb{1} \twoheadrightarrow C_j$, together with a cpd asserting C_j be equal to 1.

First, note that the number of nodes, edges, and non-zero entries in the cpds are polynomial in the $|\mathcal{J}|, |\mathbf{X}|$, and the total number of parameters in a simple matrix representation of the cpds is also polynomial if $\mathcal{I}$ is bounded (e.g., if φ is a 3-CNF formula). A satisfying assignment $\mathbf{x} \models \varphi$ of the variables $\mathbf{X}$ can be regarded as a degenerate joint distribution $\delta_{\mathbf{X}=\mathbf{x}}$ on $\mathbf{X}$, and extends uniquely to a full joint distribution $\mu_\mathbf{x} \in \Delta\mathcal{V}(\mathcal{M}_\varphi)$ consistent with all of the edges, by

$$\mu_\mathbf{x} = \delta_\mathbf{x} \otimes \delta_{\{C_j = \vee_i x_{j,i}\}}$$

Conversely, if μ is a joint distribution consistent with the edges above, then any point $\mathbf{x}$ in the support of $\mu(\mathbf{X})$ must be a satisfying assignment, since the two classes of edges respectively ensure that $1 = \mu(C_j = 1 \mid \mathbf{X} = \mathbf{x}) = \bigvee_{i \in \mathcal{I}(j)} \mathbf{x}_{j,i}$ for all $j \in \mathcal{J}$, and so $\mathbf{x} \models \varphi$.

Thus, $\{\!\{\mathcal{M}_\varphi\}\!\} \neq \emptyset$ if and only if φ is satisfiable, so an algorithm for determining if a PDG is consistent can also be adapted (in polynomial space and time) for use as a SAT solver, and so the problem of determining if a PDG consistent is NP-hard. □

Prop 2.9. *Computing $\langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma$ is #P-hard, for $\gamma > 0$.*

Proof. We prove this by reduction to #SAT. Again, let φ be some CNF formula over $\mathbf{X}$, and construct $\mathcal{M}_\varphi$ as in the proof of Prop 2.8. Furthemore, let $[\![\varphi]\!] := \{\mathbf{x} : \mathbf{x} \models \varphi\}$ be the set of assingments to $\mathbf{X}$ satisfying φ, and $\#_\varphi := |[\![\mathcal{M}]\!]|$ denote the number such assignments. We now claim that

$$\#_\varphi = \exp\left[-\frac{1}{\gamma}\langle\!\langle \mathcal{M}_\varphi \rangle\!\rangle_\gamma\right]. \tag{4}$$

If true, we would have a reduced the #P-hard problem of computing $\#_\varphi$ to the problem of computing $\langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma$ for fixed γ. We now proceed with proof (4). By definition, we have

$$\langle\!\langle \mathcal{M}_\varphi \rangle\!\rangle_\gamma = \inf_\mu \Big[Inc_{\mathcal{M}_\varphi}(\mu) + \gamma IDef_{\mathcal{M}_\varphi}(\mu) \Big].$$

We start with a claim about first term.

Claim A.2.1. $Inc_{\mathcal{M}_\varphi}(\mu) = \begin{cases} 0 & \text{if } \operatorname{supp}\mu \subseteq [\![\varphi]\!] \times \{\mathbf{1}\} \\ \infty & \text{otherwise} \end{cases}$.

Proof. Writing out the definition explicitly, the first can be written as

$$Inc_{\mathcal{M}_\varphi}(\mu) = \sum_j \left[\boldsymbol{D}\Big(\mu(C_j) \,\Big\|\, \delta_1\Big) + \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu(\mathbf{X}_j)} \boldsymbol{D}\Big(\mu(C_j \mid \mathbf{X}_j = \mathbf{x}) \,\Big\|\, \delta_{\vee_i \mathbf{x}_{j,i}}\Big)\right], \tag{5}$$

where $\mathbf{X}_j = \{X_{ij} : j \in \mathcal{I}(j)\}$ is the set of variables that appear in clause j, and $\delta_{(-)}$ is the probability distribution placing all mass on the point indicated by its subscript. As a reminder, the relative entropy is given by

$$\boldsymbol{D}\Big(\mu(\Omega) \,\Big\|\, \nu(\Omega)\Big) := \mathop{\mathbb{E}}_{\omega\sim\mu} \log \frac{\mu(\omega)}{\nu(\omega)}, \qquad \begin{array}{c}\text{and in particular,}\\ \text{if } \Omega \text{ is binary,}\end{array} \qquad \boldsymbol{D}\big(\mu(\Omega) \,\|\, \delta_\omega\big) = \begin{cases} 0 & \text{if } \mu(\omega) = 1;\\ \infty & \text{otherwise.}\end{cases}$$

Applying this to (5), we find that either:

1. Every term of (5) is finite (and zero) so $Inc_{\mathcal{M}_\varphi}(\mu) = 0$, which happens when $\mu(C_j = 1) = 1$ and $\mu(C_j = \vee_i\ x_{j,i}) = 1$ for all j. In this case, $\mathbf{c} = \mathbf{1} = \{\vee_i\ x_{j,i}\}_j$ so $\mathbf{x} \models \varphi$ for every $(\mathbf{c}, \mathbf{x}) \in \operatorname{supp}\mu$;
2. Some term of (5) is infinite, so that $Inc_{\mathcal{M}_\varphi}(\mu) = \infty$, which happens if some j, either
   (a) $\mu(C_j \neq 1) > 0$ — in which case there is some $(\mathbf{x}, \mathbf{c}) \in \operatorname{supp}\mu$ with $\mathbf{c} \neq 1$, or
   (b) $\operatorname{supp}\mu(\mathbf{C}) = \{\mathbf{1}\}$, but $\mu(C_j \neq \vee_i\ x_{j,i}) > 0$ — in which case there is some $(\mathbf{x}, \mathbf{1}) \in \operatorname{supp}\mu$ for which $1 = c_j \neq \vee_i\ x_{j,i}$, and so $\mathbf{x} \not\models \varphi$.

Condensing and rearranging slightly, we have shown that

$$Inc_{\mathcal{M}_\varphi}(\mu) = \begin{cases} 0 & \text{if } \mathbf{x} \models \varphi \text{ and } \mathbf{c} = \mathbf{1} \text{ for all } (\mathbf{x}, \mathbf{c}) \in \operatorname{supp}\mu\\ \infty & \text{otherwise}\end{cases}.$$

□

Because *IDef* is bounded, it follows immediately that $\langle\!\langle \mathcal{M}_\varphi \rangle\!\rangle_\gamma$, is finite if and only if there is some distribution $\mu \in \Delta\mathcal{V}(\mathbf{X}, \mathbf{C})$ for which $Inc_{\mathcal{M}_\varphi}(\mu)$ is finite, or equivalently, by Claim A.2.1, iff there exists some $\mu(\mathbf{X}) \in \Delta\mathcal{V}(\mathbf{X})$ for which $\operatorname{supp}\mu(\mathbf{X}) \subseteq [\![\varphi]\!]$, which in turn is true if and only if φ is satisfiable.

In particular, if φ is not satisfiable (i.e., $\#_\varphi = 0$), then $\langle\!\langle \mathcal{M}_\varphi \rangle\!\rangle_\gamma = +\infty$, and

$$\exp\left[-\frac{1}{\gamma}\langle\!\langle \mathcal{M}_\varphi \rangle\!\rangle_\gamma\right] = \exp[-\infty] = 0 = \#_\varphi,$$

so in this case (4) holds as promised. On the other hand, if φ *is* satisfiable, then, again by Claim A.2.1, every μ minimizing $[\![\mathcal{M}_\varphi]\!]_\gamma$, (i.e., every $\mu \in [\![\mathcal{M}_\varphi]\!]^*_\gamma$) must be supported entirely on $[\![\varphi]\!]$ and have $Inc_{\mathcal{M}_\varphi}(\mu) = 0$. As a result, we have

$$\langle\!\langle \mathcal{M}_\varphi \rangle\!\rangle_\gamma = \inf_{\mu\in\Delta\left[[\![\varphi]\!]\times\{\mathbf{1}\}\right]} \gamma\ IDef_{\mathcal{M}_\varphi}(\mu).$$

A priori, by the definition of $IDef_{\mathcal{M}_\varphi}$, we have

$$IDef_{\mathcal{M}_\varphi}(\mu) = -\operatorname{H}(\mu) + \sum_j \Big[\alpha_{j,1} \operatorname{H}_\mu(C_j \mid \mathbf{X}_j) + \alpha_{j,0} \operatorname{H}_\mu(C_j)\Big],$$

where $\alpha_{j,0}$ and $\alpha_{j,1}$ are values of α for the edges of $\mathcal{M}_\varphi$, which we have not specified because they are rendered irrelevant by the fact that their corresponding cpds are deterministic. We now show how this plays out in the present case. Any $\mu \in \Delta\left[[\![\varphi]\!] \times \{\mathbf{1}\}\right]$ we consider has a degenerate marginal on $\mathbf{C}$.

Specifcally, for every j, we have $\mu(C_j) = \delta_1$, and since entropy is non-negative and never increased by conditioning,

$$0 \leq \mathrm{H}_\mu(C_j \mid \mathbf{X}_j) \leq \mathrm{H}_\mu(C_j) = 0.$$

Therefore, $\mathit{IDef}_{\mathcal{M}_\varphi}(\mu)$ reduces to the negative entropy of μ. Finally, making use of the fact that the maximum entropy distribution μ^* supported on a finite set S is the uniform distribution on S, and has $\mathrm{H}(\mu^*) = \log|S|$, we have

$$\begin{aligned}
\langle\!\langle \mathcal{M}_\varphi \rangle\!\rangle_\gamma &= \inf_{\mu \in \Delta\left[[\![\varphi]\!] \times \{\mathbf{1}\}\right]} \gamma\ \mathit{IDef}_{\mathcal{M}_\varphi}(\mu) \\
&= \inf_{\mu \in \Delta\left[[\![\varphi]\!] \times \{\mathbf{1}\}\right]} -\gamma\ \mathrm{H}(\mu) \\
&= -\gamma\ \sup_{\mu \in \Delta\left[[\![\varphi]\!] \times \{\mathbf{1}\}\right]} \mathrm{H}(\mu) \\
&= -\gamma\ \log(\#_\varphi),
\end{aligned}$$

giving us

$$\#_\varphi = \exp\left[-\frac{1}{\gamma}\langle\!\langle \mathcal{M}_\varphi \rangle\!\rangle_\gamma\right],$$

as desired. We have now reduced #SAT to computing $\langle\!\langle \mathcal{M} \rangle\!\rangle_\gamma$, for $\gamma \in \mathbb{R}^{>0}$ and an arbitrary PDG $\mathcal{M}$, which is therefore #P-hard. □

??. *

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

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

Recall the statement of Prop 4.6 from the original paper,

$$[\![\mathcal{M}]\!]_\gamma(\mu) = \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu}\left\{\sum_{X \xrightarrow{L} Y}\left[\beta_L \log\frac{1}{\mathbf{p}_L(y^{\mathbf{w}}|x^{\mathbf{w}})} + (\gamma\alpha_L - \beta_L)\log\frac{1}{\mu(y^{\mathbf{w}}|x^{\mathbf{w}})}\right] - \gamma\log\frac{1}{\mu(\mathbf{w})}\right\}, \tag{6}$$

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

$$\begin{aligned}
\langle\!\langle (\mathcal{N}_\Phi, \theta, \theta) \rangle\!\rangle_1 &= [\![(\mathcal{N}_\Phi, \theta, \theta)]\!]_1\Big(\Pr_{\Phi,\theta}(\mathbf{w})\Big) \\
&= \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu}\left\{\sum_{X \xrightarrow{L} Y}\left[\beta_L \log\frac{1}{\mathbf{p}_L(y^{\mathbf{w}}|x^{\mathbf{w}})}\right] - \log\frac{1}{\Pr_{\Phi,\theta}(\mathbf{w})}\right\} && \left[\text{ by (6) }\right] \\
&= \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu}\left\{\sum_j\left[\theta_j \log\frac{1}{\phi_j(\mathbf{w}_j)}\right] - \log\frac{Z_\Psi}{\prod_j \phi_j(\mathbf{w}_j)^{\theta_j}}\right\} && \left[\begin{array}{c}\text{cpds } \mathbf{p}_L \text{ correspond} \\ \text{to factors } \phi_j\end{array}\right] \\
&= \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu}\left\{\sum_j\left[\theta_j \log\frac{1}{\phi_j(\mathbf{w}_j)}\right] - \sum_j\left[\theta_j \log\frac{1}{\phi_j(\mathbf{w}_j)}\right] - \log Z_\Psi\right\} \\
&= \mathop{\mathbb{E}}_{\mathbf{w}\sim\mu}[-\log Z_\Psi] \\
&= -\log Z_\Psi && \left[\ Z_\Psi \text{ is constant in } \mathbf{w}\ \right]
\end{aligned}$$