Local Inconsistency Resolution: The Interplay Between Attention and Control in Probabilistic Models

Oliver E Richardson, Mandana Samiei, Mehran Shakerinava,
Joseph D Viviano, Abdessamad El Kabid, Ali Parviz, Yoshua Bengio

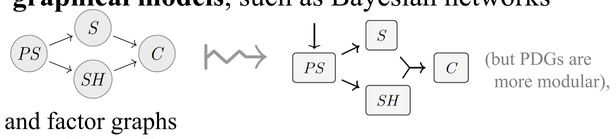
Université de Montréal · LawZero · Institut Polytechnique de Paris · McGill · Mila

A general recipe for learning and (approximate) inference, with an intuitive epistemic interpretation. Unifies many important algorithms.

What causes changes in beliefs? Some say internal conflict. But identifying inconsistencies is difficult. So, in practice, we resolve them locally: looking only at a part of the model at a time and changing only another part.

Key Representation: Probabilistic Dependency Graphs (PDGs) are directed (hyper) graphs with probabilities and confidences attached to edges. PDGs can model conflicting beliefs, come with a natural measurement of inconsistency, and capture:

❖ **graphical models**, such as Bayesian networks



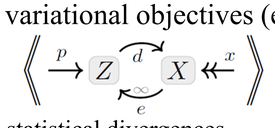
(but PDGs are more modular),

and factor graphs

(but PDGs are better calibrated).

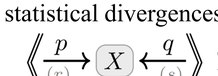
❖ **learning settings** and their **loss functions**, e.g.,

- variational objectives (e.g., VAEs)



$= -\,\mathrm{ELBO}_{p,e,d}(x)$ …. including their standard loss function, as inconsistency

- statistical divergences



Generates Rényi divergences, reverse KL, conditional divergences.

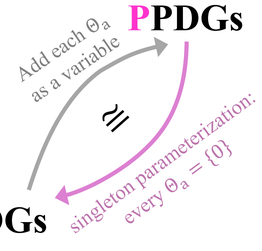
- regularizers as priors, accuracy, MSE, ….

Formalism: (Parameterized) PDGs

$\mathcal{M}(\Theta) =$ a set $\mathcal{X}$ of variables connected by arcs $\mathcal{A}$;
each $(S \xrightarrow{a} T) \in \mathcal{A}$ is associated with:

- a convex parameter space $\Theta_a \subseteq \mathbb{R}^n$
- a conditional probability $\mathbb{P}_a(T|S,\Theta_a)$
- two confidences: β_a (observational) and α_a (structural).

Fix a joint parameter setting $\theta \in \prod_{a\in\mathcal{A}} \Theta_a$ to get an (ordinary) PDG $\mathcal{M}(\theta)$.



Inconsistency semantics.

A joint probability $\mu(\mathcal{X})$ can be incompatible with a PDG in two ways:

Observational Incompatibility with $(\mathbb{P}, \boldsymbol{\beta})$

$$\sum_{S \xrightarrow{a} T \in \mathcal{A}} \beta_a \boldsymbol{D}\Big(\mu(T,S) \,\Big\|\, \mathbb{P}_a(T|S)\mu(S)\Big)$$

Structural Deficiency with $(\mathcal{A}, \boldsymbol{\alpha})$

$$\mathop{\mathbb{E}}_{\mu}\left[\log \frac{\mu(\mathcal{X})}{\lambda(\mathcal{X})} \prod_{S \xrightarrow{a} T} \left(\frac{\lambda(T|S)}{\mu(T|S)}\right)^{\alpha_a}\right]$$

Degree of inconsistency

$$\langle\!\langle \varphi \odot \mathcal{M} \rangle\!\rangle := \inf_{\mu} \Big(\mathit{OInc}_{\mathcal{M}}(\mu) + \gamma\, \mathit{SDef}_{\mathcal{M}}(\mu) \Big)$$

placing weight $\gamma \geq 0$ on the structural information

where $\varphi = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)$,
is the smallest possible incompatibility with any $\mu(\mathcal{X})$.

Algorithm: Local Inconsistency Resolution (LIR)

Input: knowledge base $\mathcal{M}(\Theta)$
Initialze $\theta^{(0)}$;
for $t = 0, 1, 2, \ldots$ **do**

1. $\varphi, \chi, \gamma \leftarrow \textsc{Refocus}()$; (select attn and ctrl masks)

Write $\exp_\theta(t\,X)$ for the path following vector field X for time t, starting at θ.
Gradient Flow of $f\colon \Theta \to \mathbb{R}$ starting at θ: $t \mapsto \exp_\theta(t \nabla_\Theta f(\Theta))$

Calculate the inconsistency masked by attention.

2. $$\boldsymbol{\theta}^{(t+1)} \leftarrow \exp_{\boldsymbol{\theta}^{(t)}}\Big\{-\chi \odot \nabla_{\boldsymbol{\theta}} \Big\langle\!\Big\langle \varphi \odot \mathcal{M}(\boldsymbol{\theta}) \Big\rangle\!\Big\rangle\Big\};$$

Resolve this inconsistency via (an approximation to) gradient flow, starting at previous state $\theta^{(t)}$, changing each parameter in proportion to our control of it.

The Classification Setting

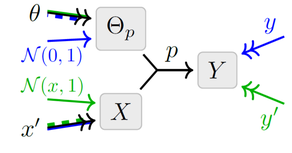
Consider a discriminator $p_\theta(Y|X)$ and sample (x, y).
Together, they have inconsistency

$$\left\langle\!\!\left\langle\; x \twoheadrightarrow X \xrightarrow{p_\theta} Y \xleftarrow{\alpha} y \;\right\rangle\!\!\right\rangle = \log \frac{1}{p_\theta(y|x)}.$$

Can resolve it by modifying:
- θ, to train the discriminator;
- y, resulting in a forward pass;
- x, to form an adversarial example.

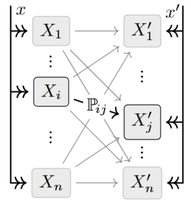
❖ **SGD.** Control over p_θ. Replace (x,y) with empirical distribution over a batch, selected by the new focus. This performs SGD with learning rate $\chi(p) \cdot \varphi(p)$.

❖ **Adversarial Training.**
Add classifier parameters as an explicit variable Θ_p with a Gaussian prior.



Patch p to classify x' as y
Construct attack $x' \approx x$ that p misclassifies as y'

❖ **Self-Attention Transformer Layers**

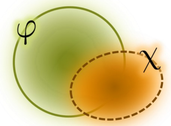


Forward pass resolves inconsistencies between naïve gaussian approximations.
Explains direction of the softmax of the attention mechanism.

$$\mathbb{P}_{ij} = \mathcal{N}(X'_j \mid v_i, I)$$
$$\varphi_{ij} = \exp\langle k_i, q_j \rangle$$

Focus: Attention and Control

attend only to probabilities of a subset of arcs $A \subseteq \mathcal{A}$ (or attn mask φ)



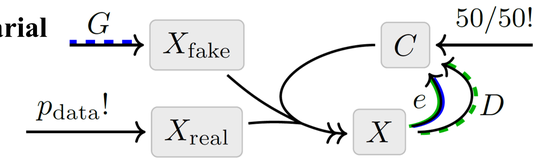
control only parameters of a subset of arcs $C \subseteq \mathcal{A}$ (or ctrl mask χ)

Typically, select focus (φ, χ) from a fixed set of *foci* $\mathbf{F} = \{\blacksquare, \blacksquare, \ldots\}$.
Dashes to indicate control.

Special Cases

❖ **Variational Inference,** the EM algorithm, e.g., VAE training.

❖ **Generative Adversarial Networks (GANs).**
Typically trained with game $\min_G \max_D \mathcal{L}^{GAN}$

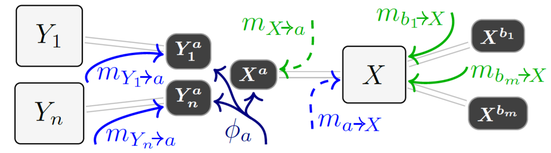


Generator's focus
inconsistency = JSD(G, p_{data});
becomes $+\mathcal{L}^{GAN}$ if disbelieves D

Discriminator's focus
inconsistency = KL(D ‖ D^{opt});
becomes $-\mathcal{L}^{GAN}$ if disbelieves e

❖ **Message Passing algorithms,** e.g., Belief Propagation.

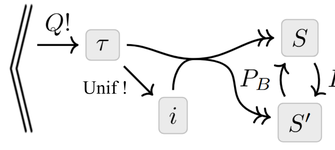
Observation: the message passing equations are sums of products of factors, i.e., correspond to inference in *local* factor graphs.



Send $a \to X$ *Send $X \to a$*

$\mathcal{M}(\theta)$ = collection of messages (BP data structure), plus the original factor graph, as a PDG

❖ **Generative Flow Networks:** the first case where the PDG's inconsistency is not the standard loss:

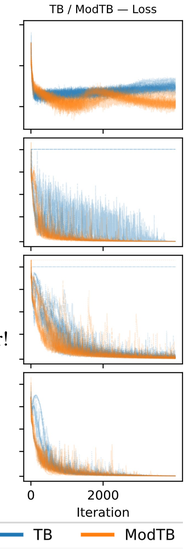


$$= \mathop{\mathbb{E}}_{\tau\sim Q}\left[\frac{1}{|\tau|} \log^2 \frac{P_F(\tau)Z}{R(x)P_B(\tau \mid x)}\right]$$

when we use "centered surprisal-based attention":

$$\varphi(P_F) = \mathrm{I}_{P_F}[\tau] - \mathrm{I}_{P_B}[\tau] = \log(P_B(\tau)/P_F(\tau))$$
$$\varphi(P_B) = \mathrm{I}_{P_B}[\tau] - \mathrm{I}_{P_F}[\tau] = \log(P_F(\tau)/P_B(\tau))$$

The modified version, with $1/|\tau|$ scaling factor, performs better!



What is next?

- Learning adaptive Refocus mechanisms (as in transformers)?
- Can we use it to discover new hybrid algorithms?

Potential for a new kind of intelligent system—more interpretable and potentially safer, driven to resolve conflicts among explicit beliefs rather than rigidly pursuing a fixed goal.