# 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

# Inconsistency...

- Might entertain probabilistic claims that cannot simultaneously hold, *e.g.*,
  - $P(E) = 0.2, \ P(\neg E) = 0.7$
  - multiple models, experimental results, priors, or, or data with (minor) conflicts

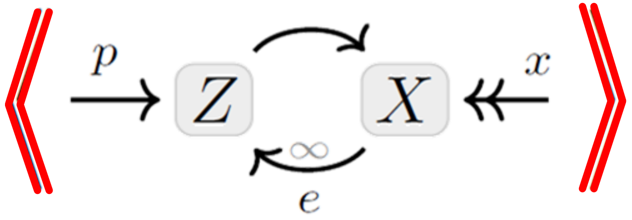
- Such situations can be modeled with *probabilistic dependency graphs* (PDGs)
  - PDGs inconsistency $\langle\!\langle - \rangle\!\rangle$ can be quantified; in most learning settings, it is the standard loss [AISTATS, 2022].

... resolution?

# Inconsistency...

- Might entertain probabilistic claims that cannot simultaneously hold, *e.g.*,
  - $P(E) = 0.2,\ P(\neg E) = 0.7$
  - multiple models, experimental results, priors, or, or data with (minor) conflicts

e.g., variational autoencoders



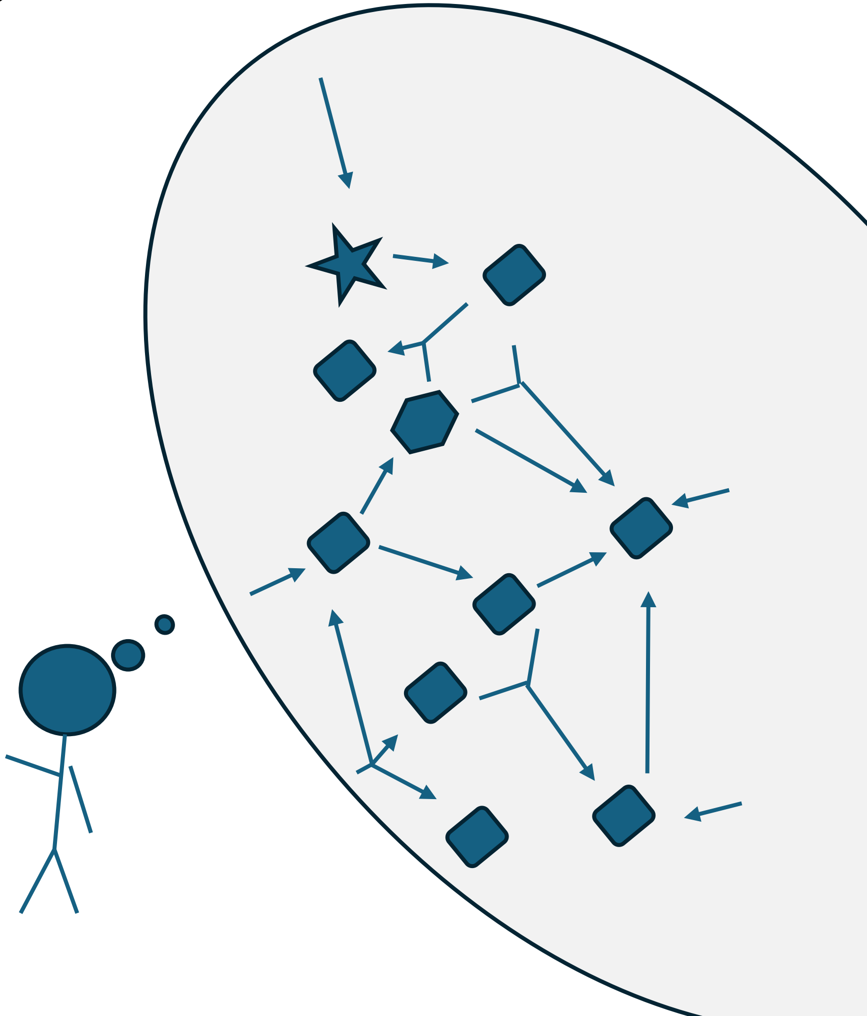

$$= -\underset{p,e,d}{\mathrm{ELBO}}(x)$$

- Such situations can be modeled with *probabilistic dependency graphs* (PDGs)
  - PDGs inconsistency $\langle\!\langle - \rangle\!\rangle$ can be quantified; in most learning settings, it is the standard loss [AISTATS, 2022].

... resolution?

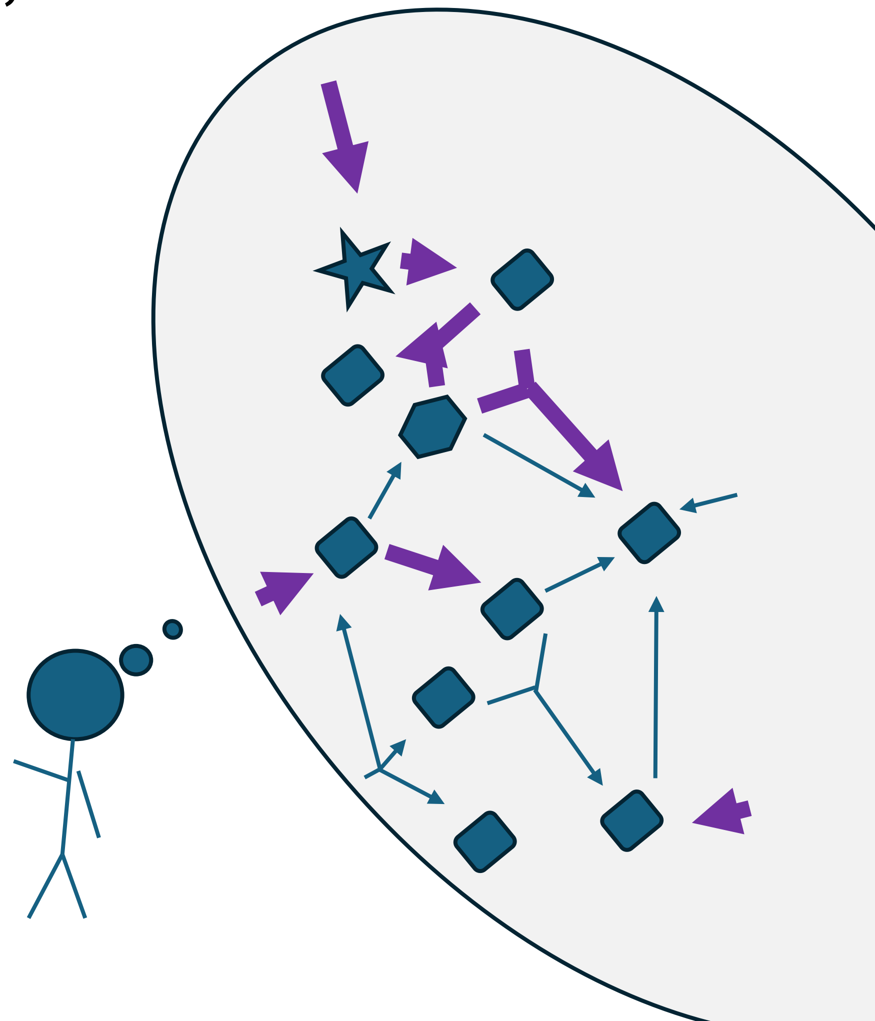
# *Local* Inconsistency Resolution

- In general, calculating inconsistency is #P-hard;



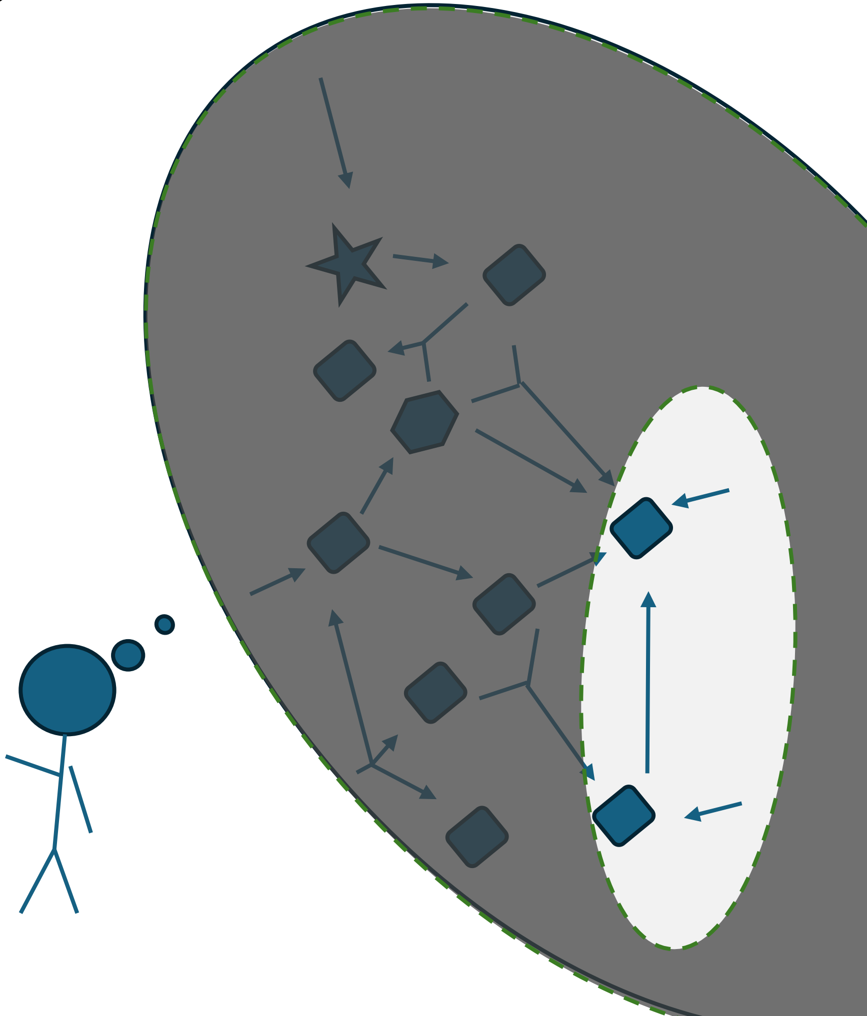
# *Local* Inconsistency Resolution

- In general, calculating inconsistency is #P-hard;
- A heuristic approach to overestimates:
  - *Add extra beliefs and assumptions until the calculation is easy (Variational Inference).*

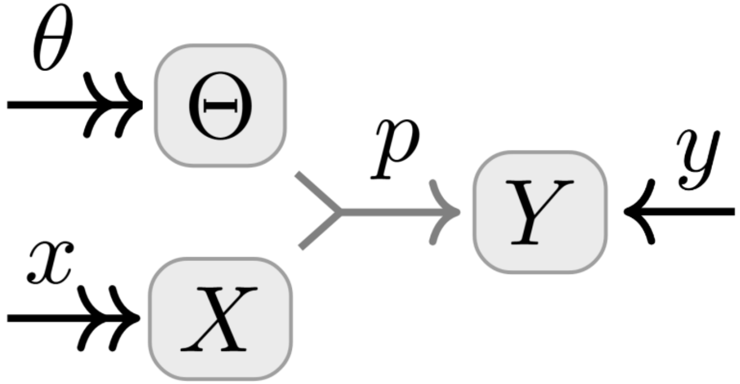


# *Local* Inconsistency Resolution

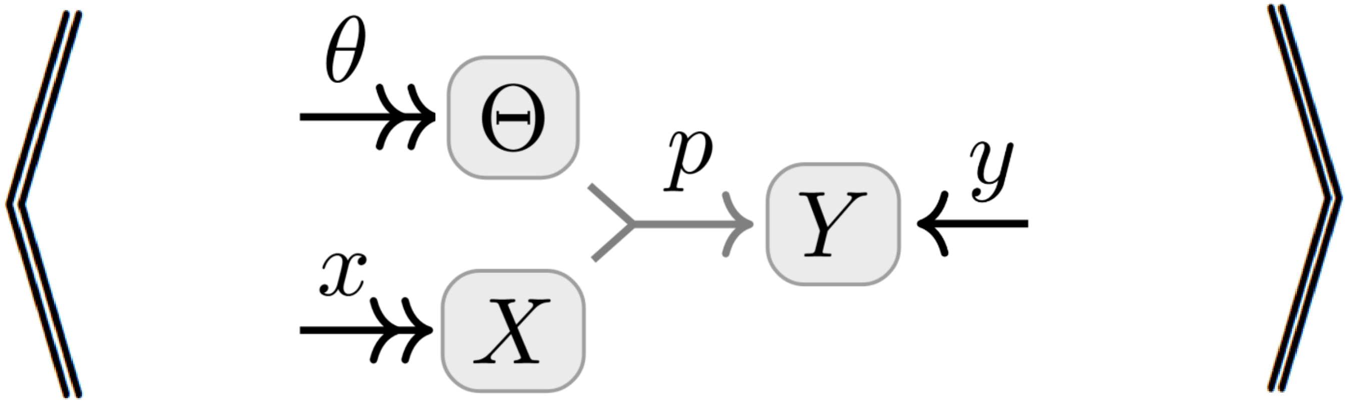
- In general, calculating inconsistency is #P-hard;
- A heuristic approach to overestimates:
  - *Add extra beliefs and assumptions until the calculation is easy (Variational Inference).*
- A heuristic approach to underestimates:
  - *Restrict* attention *to a small part of the picture at a time, and resolve only the local inconsistencies you can see, using the parameters under* control.



# LIR in the classification setting



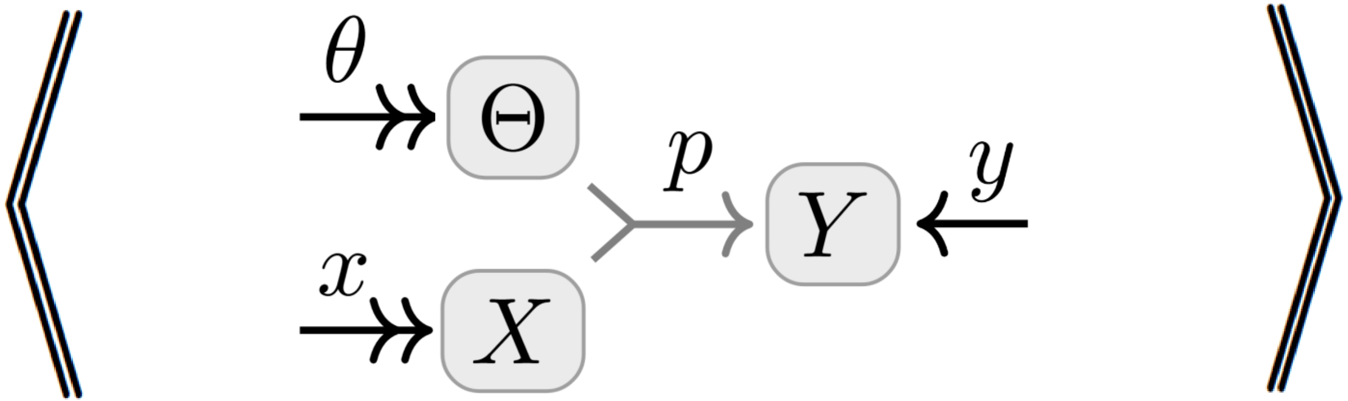
# LIR in the classification setting




$$= -\log p_\theta(y|x)$$

# LIR in the classification setting

an algorithm that resolves inconsistency by...

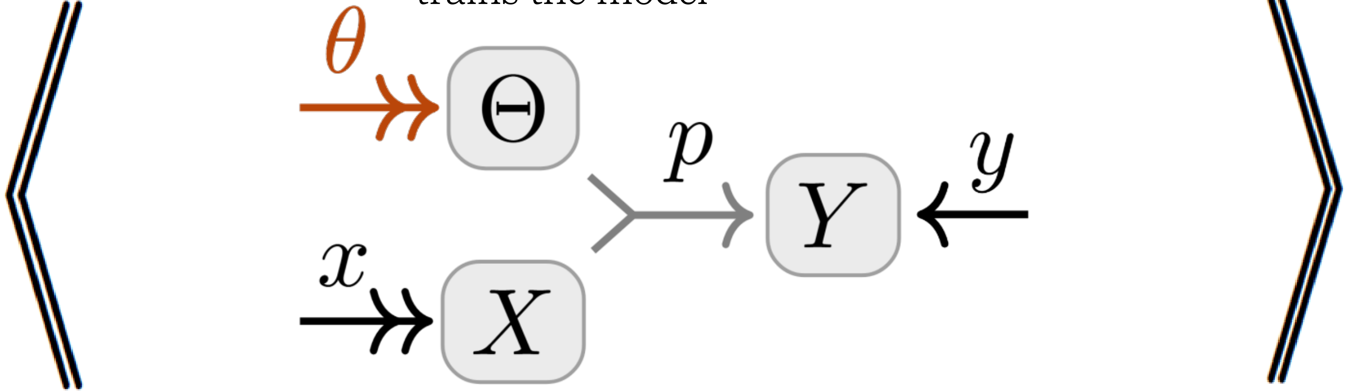



$$= -\log p_\theta(y|x)$$

# LIR in the classification setting

an algorithm that resolves inconsistency by...

... controlling parameters
trains the model

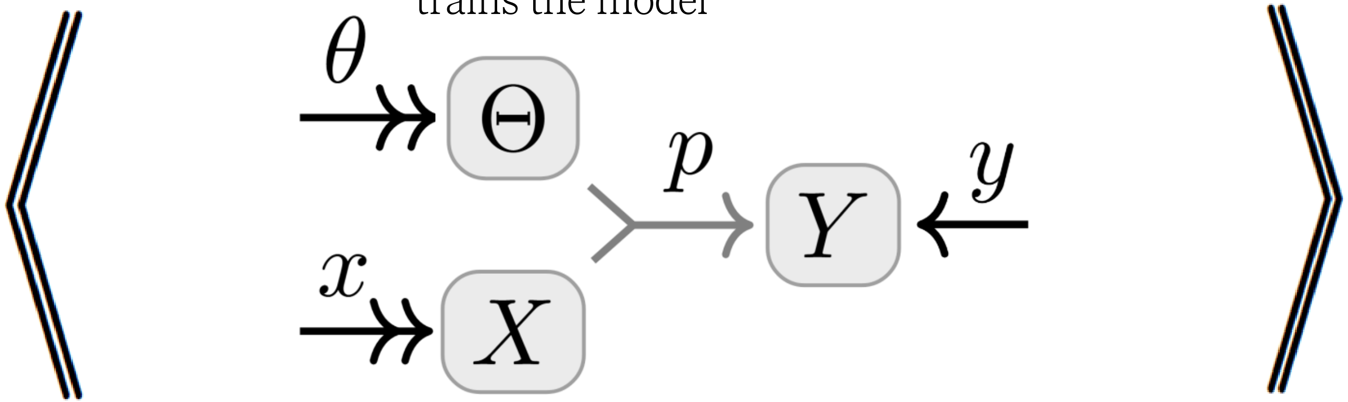



$$= -\log p_\theta(y|x)$$

# LIR in the classification setting

an algorithm that resolves inconsistency by...

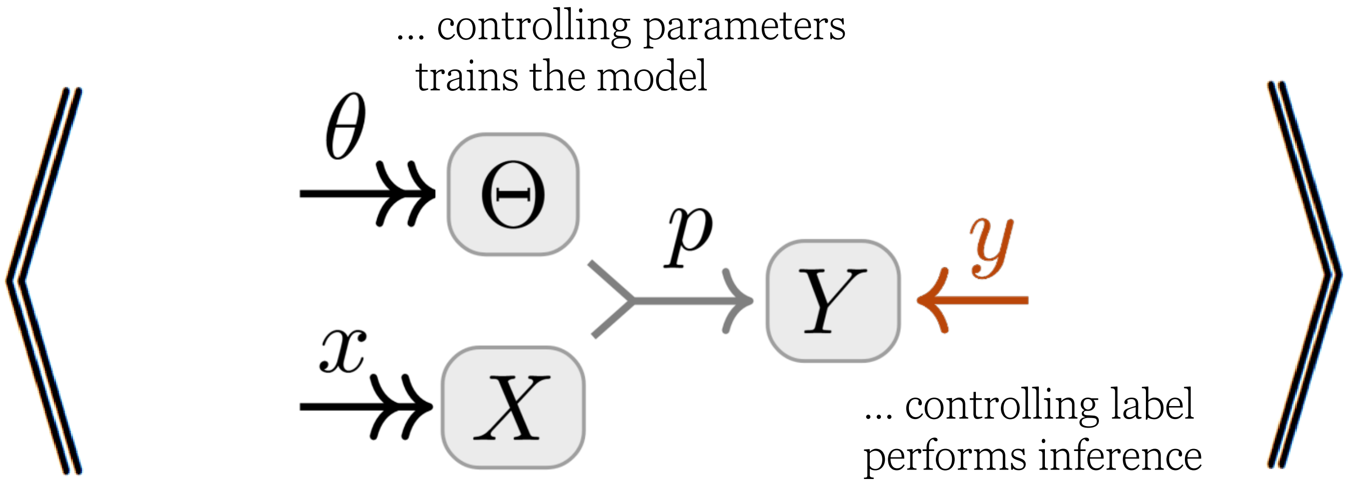
... controlling parameters
trains the model




$$= -\log p_\theta(y|x)$$

# LIR in the classification setting

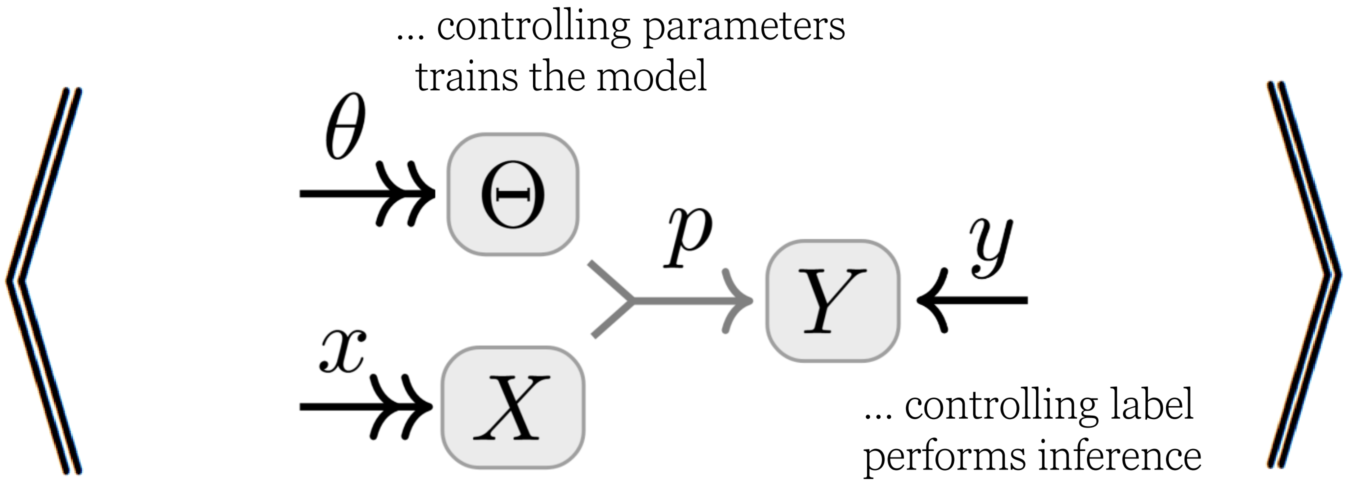
an algorithm that resolves inconsistency by...




$$= -\log p_\theta(y|x)$$

# LIR in the classification setting

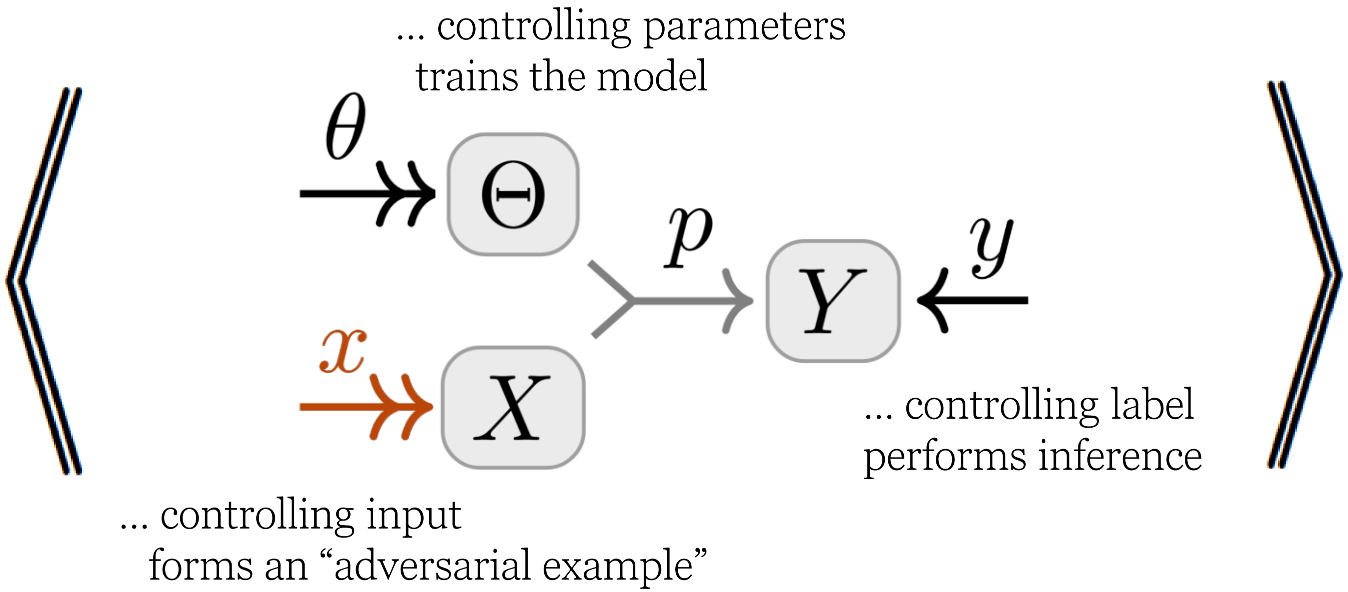
an algorithm that resolves inconsistency by...




$$= -\log p_\theta(y|x)$$

# LIR in the classification setting

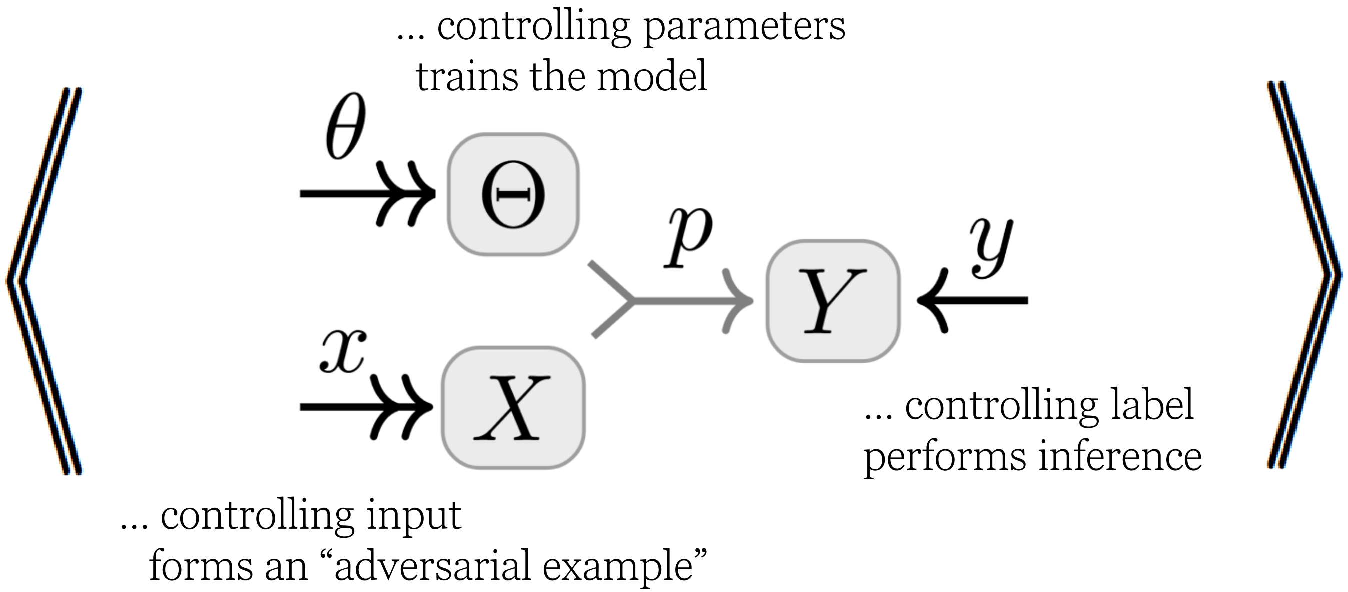
an algorithm that resolves inconsistency by...




$$= -\log p_\theta(y|x)$$

# LIR in the classification setting

an algorithm that resolves inconsistency by…




$$= -\log p_\theta(y|x)$$

**Algorithm**: Local Inconsistency Resolution (LIR)

**for** $t = 0, 1, 2, \ldots$ **do**

$$\varphi, \chi \leftarrow \textsc{Refocus}();$$

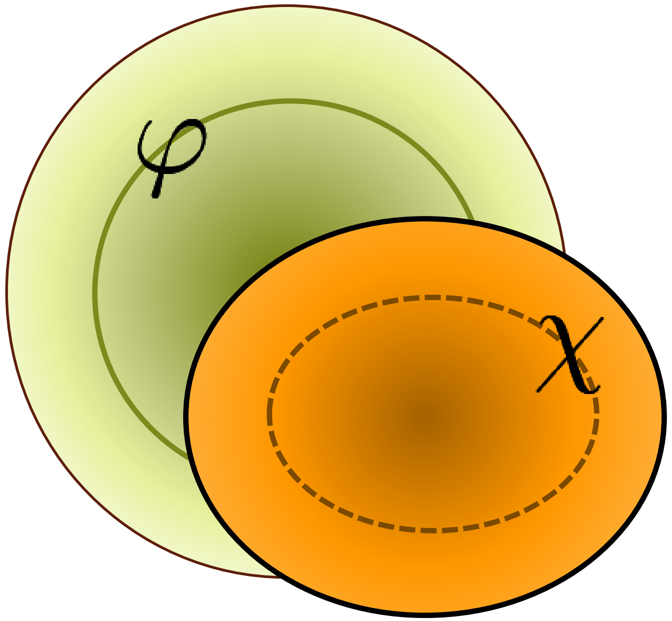
$$\boldsymbol{\theta}^{(t+1)} \leftarrow \exp_{\boldsymbol{\theta}^{(t)}} \Big\{ -\chi \odot \nabla_{\boldsymbol{\theta}} \Big\langle\!\!\Big\langle \varphi \odot \boldsymbol{m}(\boldsymbol{\theta}) \Big\rangle\!\!\Big\rangle \Big\};$$

$$\varphi, \chi \leftarrow \textsc{Refocus}();$$

## Attention and Control

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

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

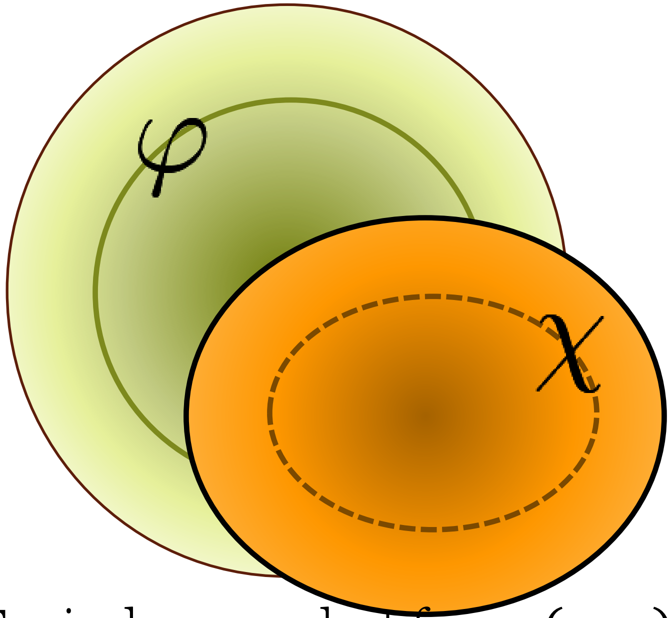


$$\varphi, \chi \leftarrow \textsc{Refocus}();$$

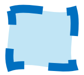
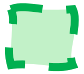
## Attention and Control

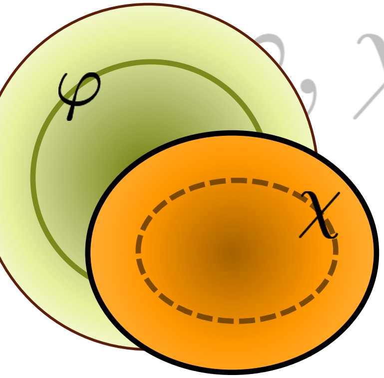
attend only to probabilities of
a subset of arcs $A \subseteq \mathcal{A}$
(or attn mask $\varphi$ )

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



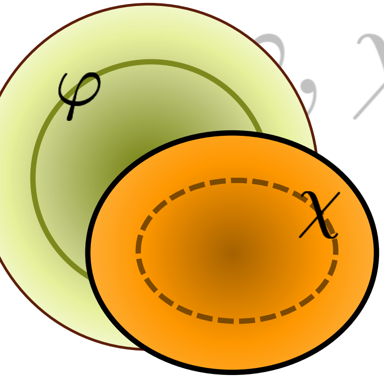

Typical case: select focus $(\varphi, \chi)$ from a fixed set

$\mathbf{F} = \{$  ,  , $\ldots\}$.




$\varphi, \chi \leftarrow \textsc{Refocus}();$

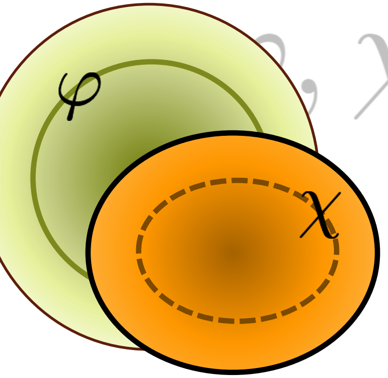
$$\boldsymbol{\theta}^{(t+1)} \leftarrow \exp_{\boldsymbol{\theta}^{(t)}}\left\{-\chi \odot \nabla_{\boldsymbol{\theta}} \left\langle\!\!\left\langle \varphi \odot \boldsymbol{m}(\boldsymbol{\theta}) \right\rangle\!\!\right\rangle\right\};$$




$\varphi, \chi \leftarrow \textsc{Refocus}();$

Calculate the inconsistency
of the current beliefs,
weighted by attention.

$$\left\langle\!\!\left\langle \overbrace{\varphi} \odot \overbrace{\boldsymbol{m}(\boldsymbol{\theta})} \right\rangle\!\!\right\rangle$$



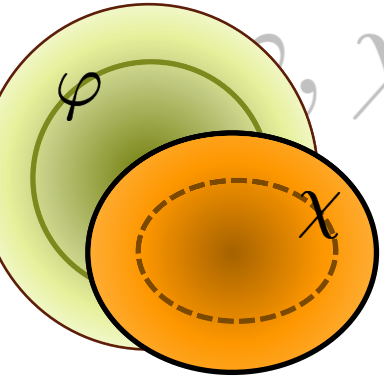

$, \chi \leftarrow \textsc{Refocus}();$

Calculate the inconsistency
of the current beliefs,
weighted by attention.

$$\left\langle\!\!\left\langle\, \varphi \odot \boldsymbol{m}(\boldsymbol{\theta}) \,\right\rangle\!\!\right\rangle$$

$$:= \inf_{\mu} \sum_{S \xrightarrow{a} T \in \mathcal{A}} \varphi(a)\, \boldsymbol{D}_{\mathrm{KL}}\Big(\mu(S,T) \,\Big\|\, \mu(S)\, \mathbb{P}_a(T|S;\theta)\Big)$$

$\varphi, \chi \leftarrow \textsc{Refocus}();$

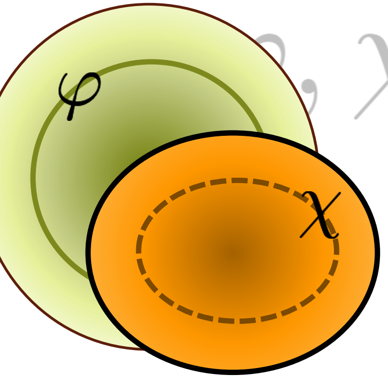



Calculate the inconsistency
of the current beliefs,
weighted by attention.

$$\exp_{\boldsymbol{\theta}^{(t)}}\left\{ \nabla_{\boldsymbol{\theta}} \left\langle\!\!\left\langle \varphi \odot \boldsymbol{m}(\boldsymbol{\theta}) \right\rangle\!\!\right\rangle \right\};$$

Reduce this inconsistency by (an approximation to) gradient flow,
starting at previous state $\theta^{(t)}$,

$, \chi \leftarrow \textsc{Refocus}();$




Calculate the inconsistency
of the current beliefs,
weighted by attention.

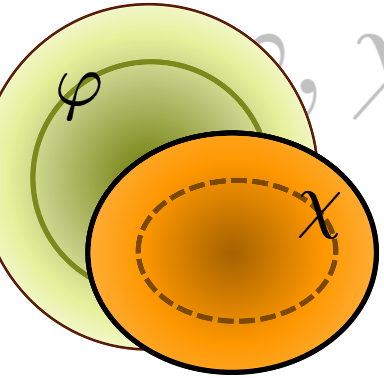
$$\underbrace{\exp}_{}{}_{\underbrace{\boldsymbol{\theta}^{(t)}}_{}}\left\{ \underbrace{\nabla_{\boldsymbol{\theta}}}_{} \left\langle\!\!\left\langle \overbrace{\varphi \odot}^{} \overbrace{\boldsymbol{m}(\boldsymbol{\theta})}^{} \right\rangle\!\!\right\rangle \right\};$$

Reduce this inconsistency by (an approximation to) gradient flow,

starting at previous state $\theta^{(t)}$,

**Gradient Flow** of $f : \Theta \to \mathbb{R}$ starting at $\theta$:

$$t \mapsto \exp_{\theta}(t \nabla_{\Theta} f(\Theta))$$




, $\chi \leftarrow$ REFOCUS();

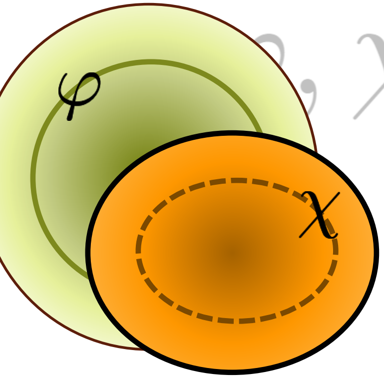
Calculate the inconsistency
of the current beliefs,
weighted by attention.

$$\exp_{\boldsymbol{\theta}^{(t)}}\left\{-\chi \odot \nabla_{\boldsymbol{\theta}} \left\langle\!\!\left\langle \varphi \odot \boldsymbol{m}(\boldsymbol{\theta}) \right\rangle\!\!\right\rangle\right\};$$

Reduce this inconsistency by (an approximation to) gradient flow,
starting at previous state $\theta^{(t)}$,

**Gradient Flow** of $f\colon \Theta \to \mathbb{R}$ starting at $\theta$:
$t \mapsto \exp_\theta(t\nabla_\Theta f(\Theta))$

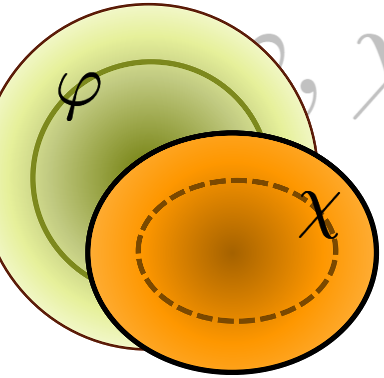
$\ldots, \chi \leftarrow$ REFOCUS();




Calculate the inconsistency
of the current beliefs,
weighted by attention.

$$\exp_{\boldsymbol{\theta}^{(t)}}\left\{-\chi \odot \nabla_{\boldsymbol{\theta}} \left\langle\!\!\left\langle \varphi \odot \boldsymbol{m}(\boldsymbol{\theta}) \right\rangle\!\!\right\rangle\right\};$$

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




$\varphi, \chi \leftarrow \textsc{Refocus}();$

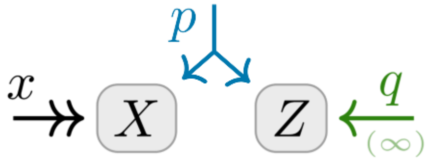
Calculate the inconsistency
of the current beliefs,
weighted by attention.

$$\boldsymbol{\theta}^{(t+1)} \leftarrow \exp_{\boldsymbol{\theta}^{(t)}} \left\{ -\chi \odot \nabla_{\boldsymbol{\theta}} \left\langle\!\!\left\langle \varphi \odot \boldsymbol{m}(\boldsymbol{\theta}) \right\rangle\!\!\right\rangle \right\};$$

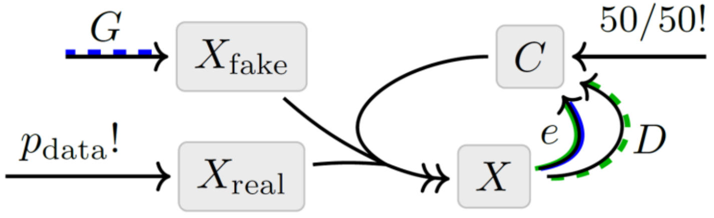
Reduce this inconsistency by (an approximation to) gradient flow,
starting at previous state $\theta^{(t)}$,
changing each parameter in proportion to our control of it.

# Many important algorithms result from a an appropriate selection of focus (attn + ctrl)

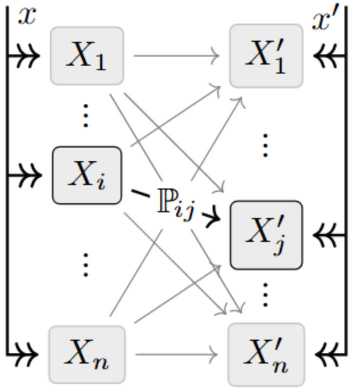
- **The EM algorithm**



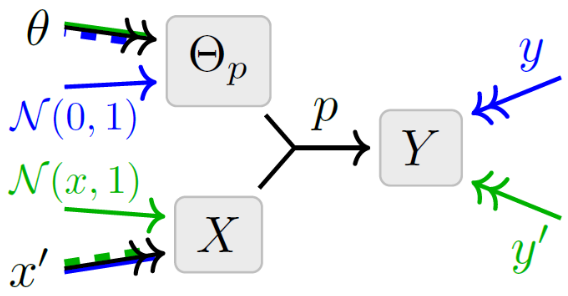

- **Generative Adversarial Networks (GANs)**



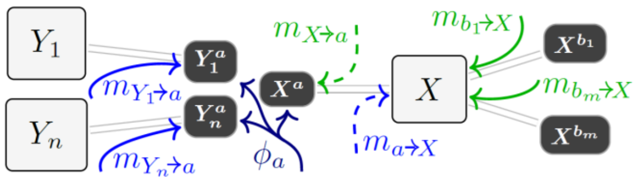

- **Forward pass of Self-Attention Layers in Transformers**



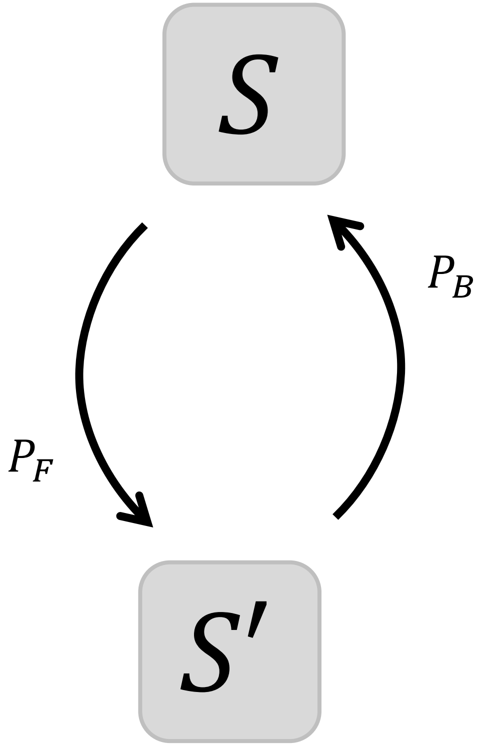

- **Adversarial Training**




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

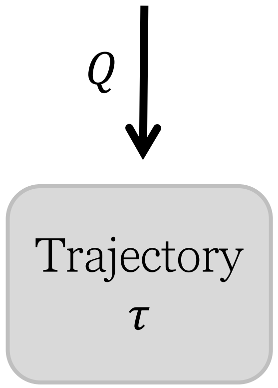


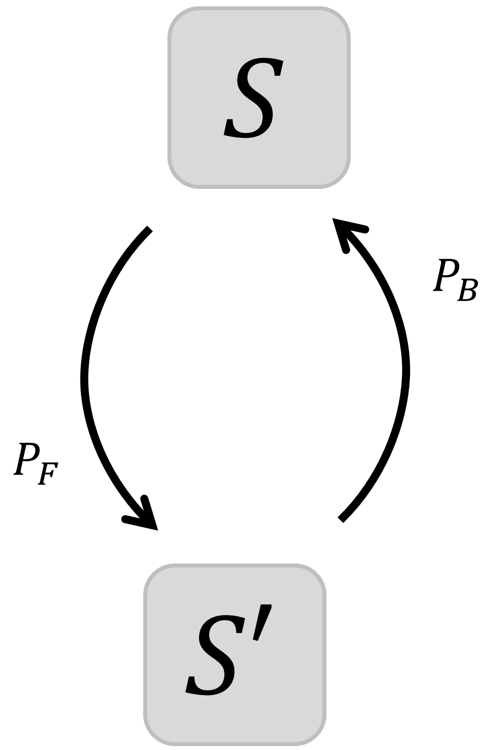
# Modified Trajectory Balance (TB) GFlowNets




[RSSVKB26; AISTATS]

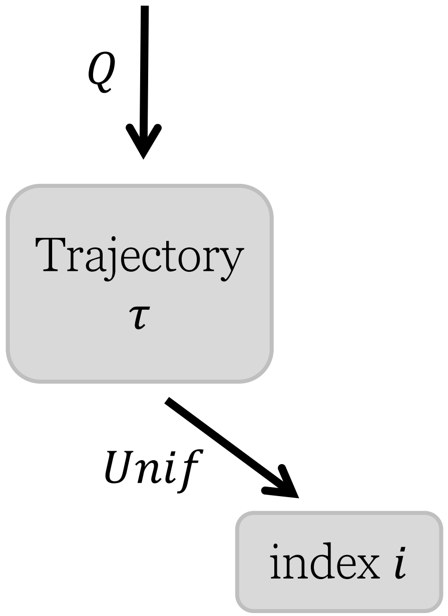
# Modified Trajectory Balance (TB) GFlowNets

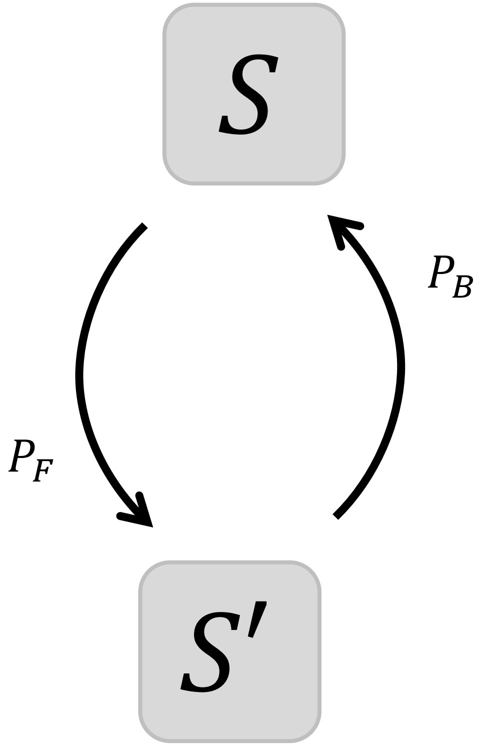



[RSSVKB26; AISTATS]

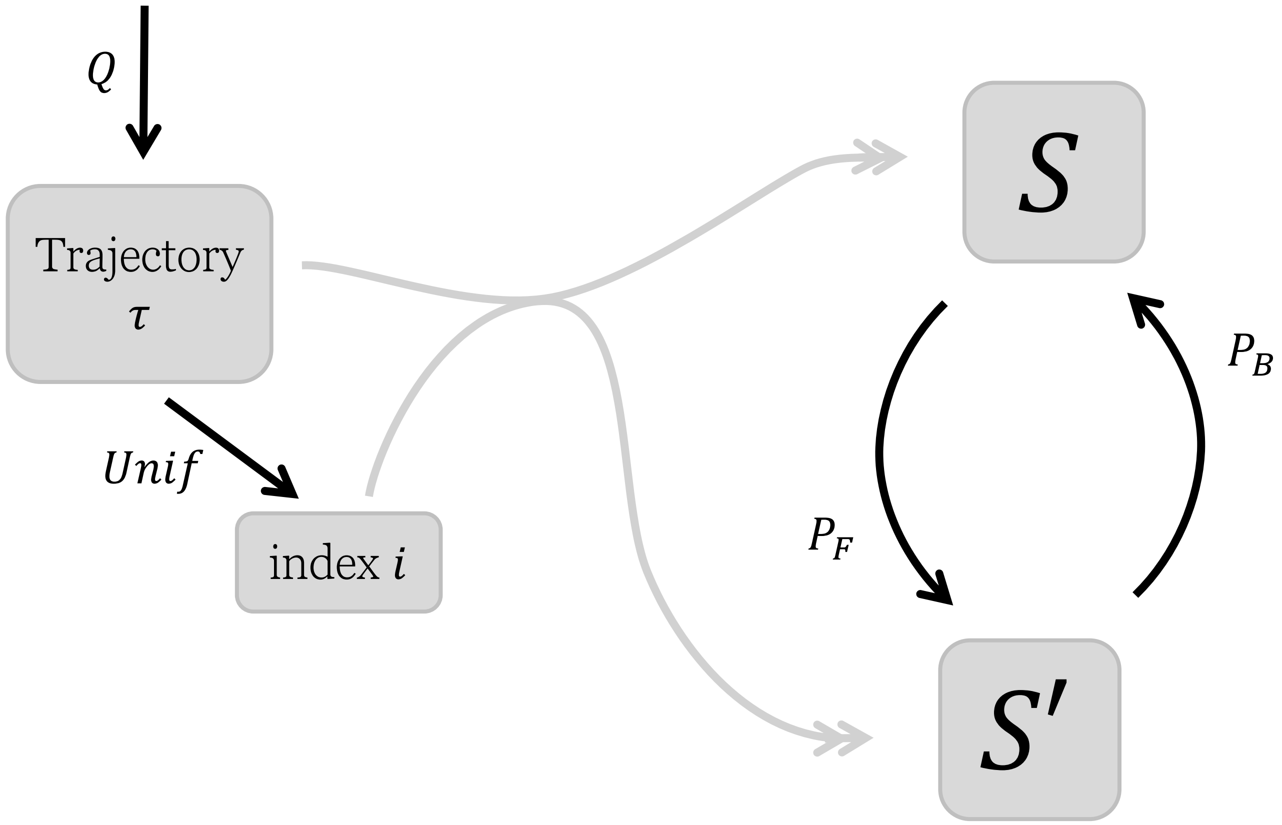
# Modified Trajectory Balance (TB) GFlowNets



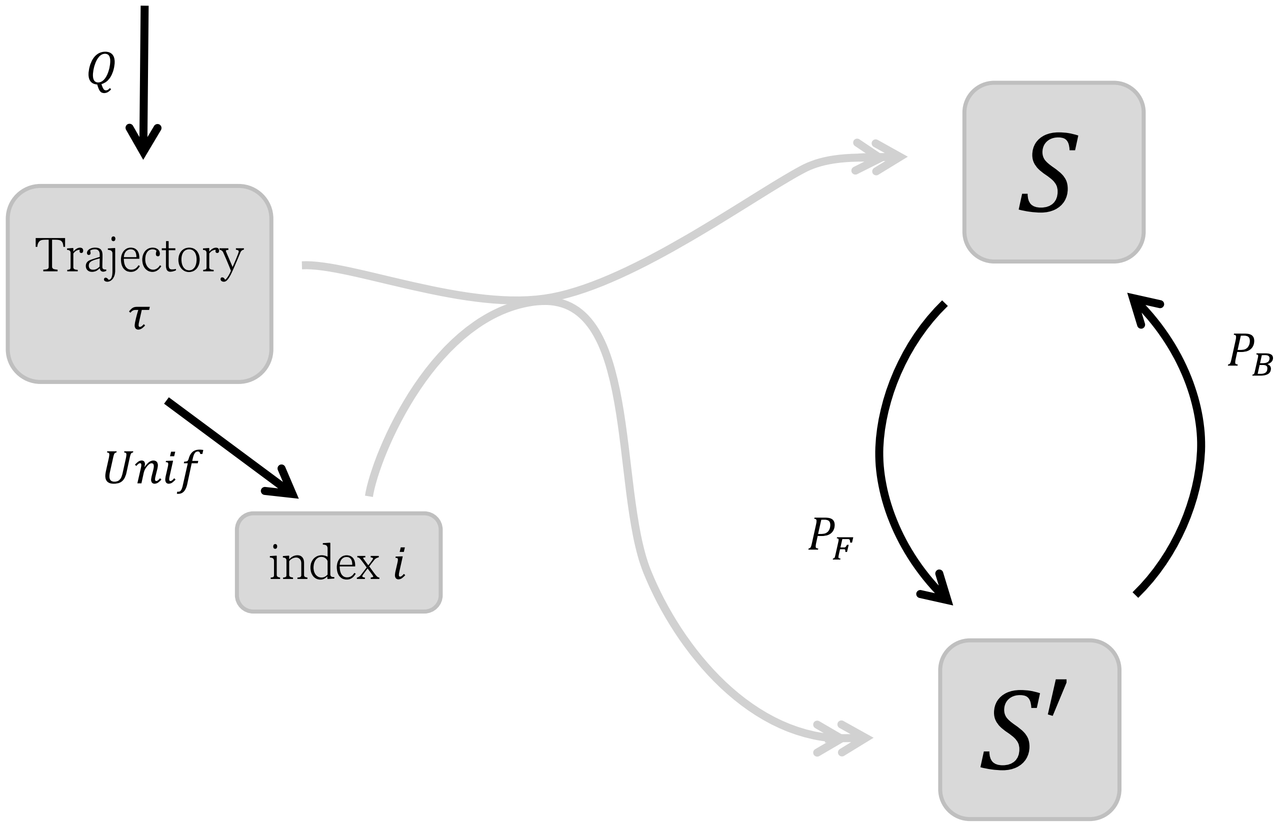

[RSSVKB26; AISTATS]

# Modified Trajectory Balance (TB) GFlowNets




[RSSVKB26; AISTATS]

# Modified Trajectory Balance (TB) GFlowNets



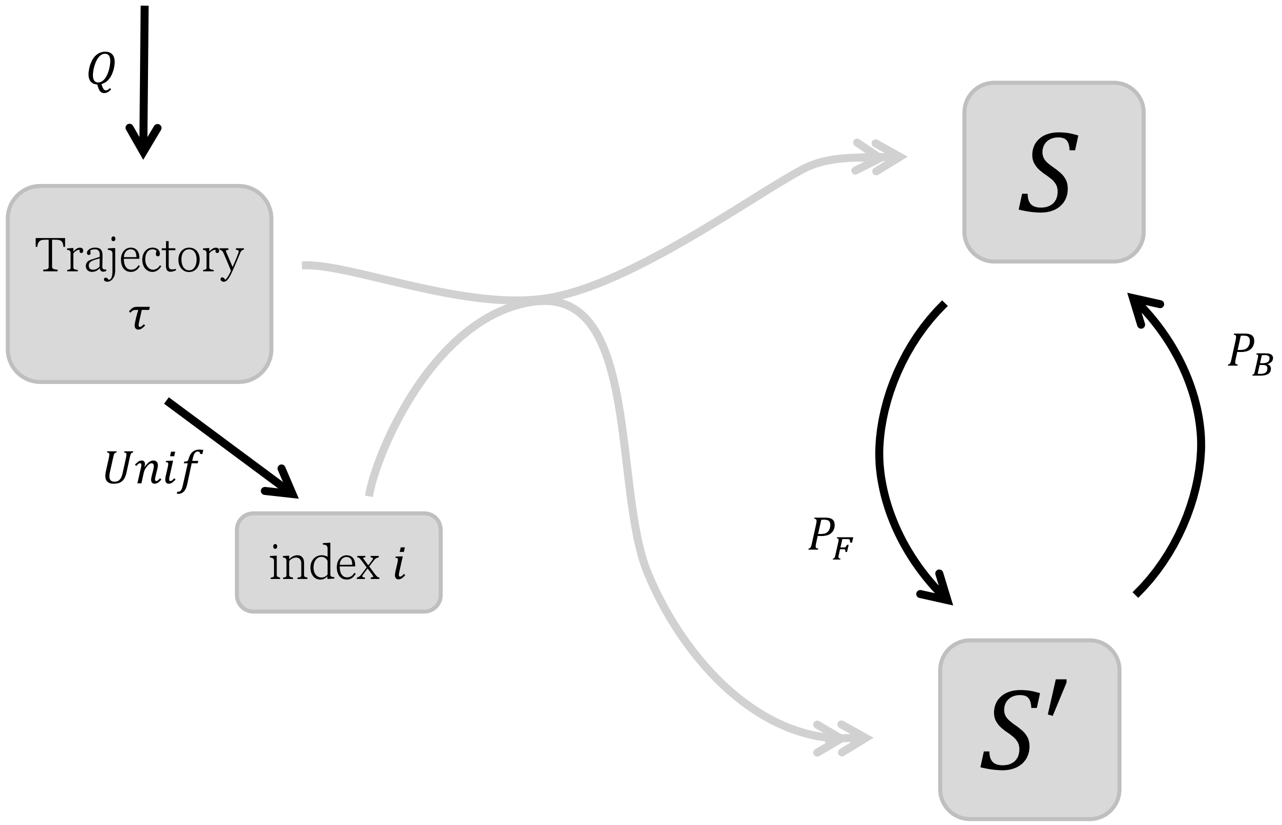

Centered Surprisal-based attention…

$$\varphi(P_F) = I_{P_F}(\tau) - I_{P_B}(\tau) = \log \frac{P_B(\tau)}{P_F(\tau)}$$

$$\varphi(P_B) = I_{P_B}(\tau) - I_{P_F}(\tau) = \log \frac{P_F(\tau)}{P_B(\tau)}$$

[**R**SSVKB26; AISTATS]

# Modified Trajectory Balance (TB) GFlowNets




Centered Surprisal-based attention…

$$\varphi(P_F) = I_{P_F}(\tau) - I_{P_B}(\tau) = \log \frac{P_B(\tau)}{P_F(\tau)}$$

$$\varphi(P_B) = I_{P_B}(\tau) - I_{P_F}(\tau) = \log \frac{P_F(\tau)}{P_B(\tau)}$$

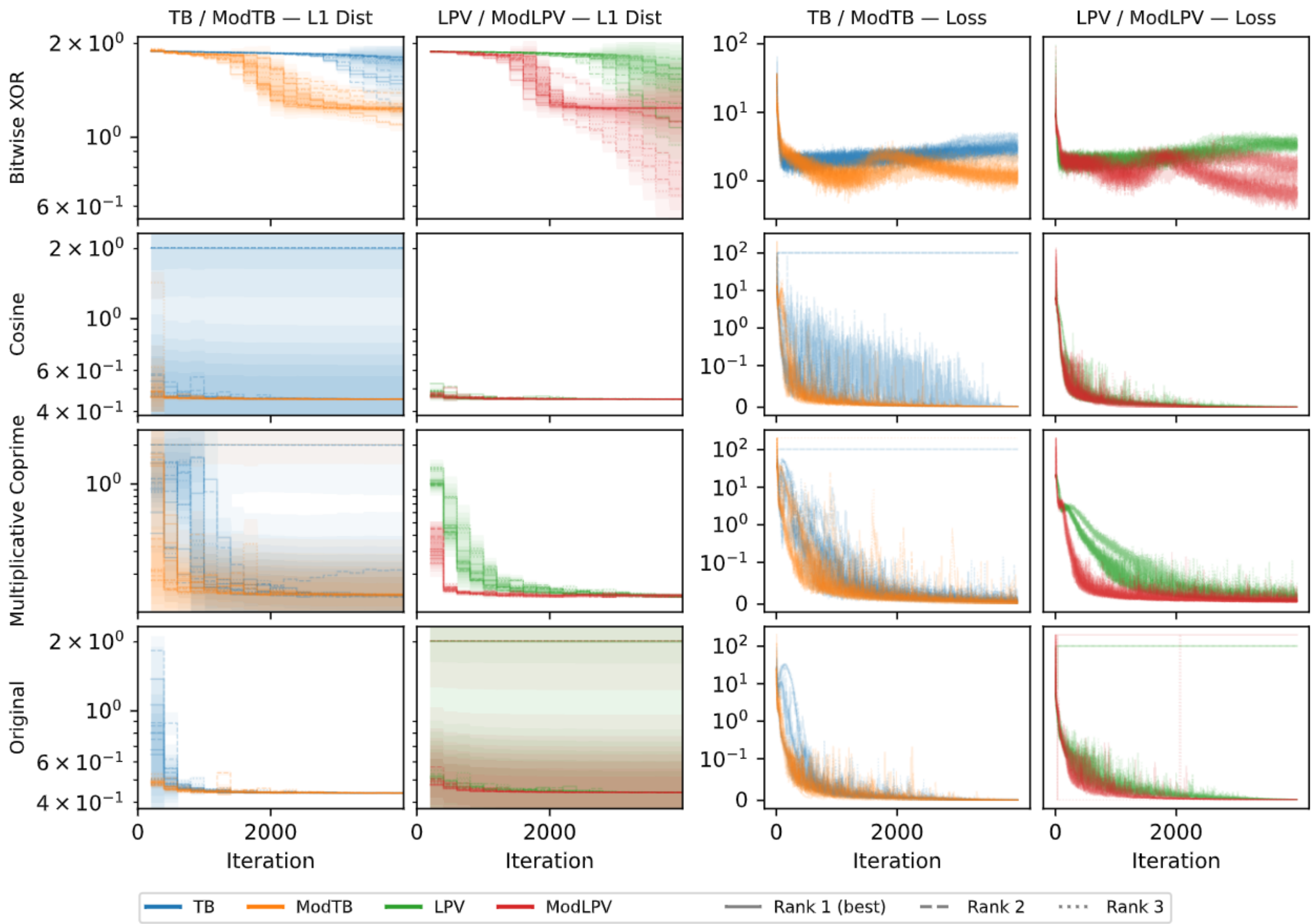
… yields slightly different loss:

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

Can get original TB by scaling attention by $|\tau|$, but might this work better?

[RSSVKB26; AISTATS]

# Modified Trajectory Balance (TB) GFlowNets




Centered Surprisal-based attention…

$$\varphi(P_F) = I_{P_F}(\tau) - I_{P_B}(\tau) = \log \frac{P_B(\tau)}{P_F(\tau)}$$

$$\varphi(P_B) = I_{P_B}(\tau) - I_{P_F}(\tau) = \log \frac{P_F(\tau)}{P_B(\tau)}$$

… yields slightly different loss:

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

Can get original TB by scaling attention by |τ|, but might this work better?

[RSSVKB26; AISTATS]