GFlowNets and System 2 Deep Learning

Yoshua Bengio

Pan-India MSR-IISc Seminar on AI

June 14th, 2022

CIFAR

Université de Montréal

Mila

IVADO

Missing from Current ML: Understanding & Generalization Beyond the Training Distribution

- Learning theory only deals with generalization within the same distribution

- SOTA AI systems learn but do not generalize well (or have high sample complexity when adapting) to modified distributions, non-stationarities, etc.

Missing from Current ML: Understanding, Reasoning & Generalization Beyond the Training **Distribution**

- If not iid, need alternative assumptions, otherwise no reason to expect generalization
    - Humans do a lot better!
    - Inductive biases inspired from brains?

- How do distributions change?
- What knowledge can be re-used?
- Reasoning: how to re-use pieces of knowledge?

Mila

CONSCIOUS PROCESSING HELPS HUMANS REASON & DEAL WITH OOD SETTINGS

Faced with **novel or rare situations**, humans call upon **conscious attention** to combine on-the-fly the appropriate pieces of knowledge, to **reason** with them and **imagine** solutions.

→ we do not follow our habitual routines, we think hard to solve new problems.

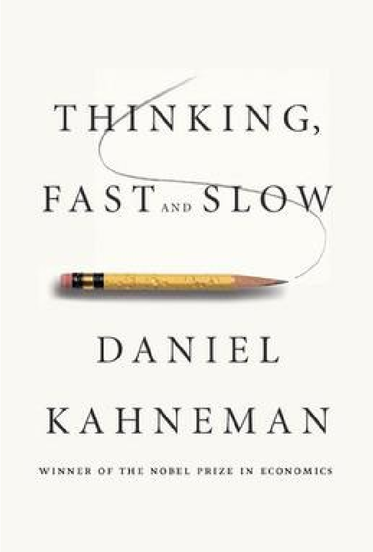
SYSTEM 1 VS. SYSTEM 2 COGNITION

2 systems (and categories of cognitive tasks):

Manipulates high-level / semantic concepts, which can be recombined combinatorially

System 1

- Intuitive, fast, **UNCONSCIOUS**, 1-step parallel, non-linguistic, habitual
- Implicit knowledge
- Current DL



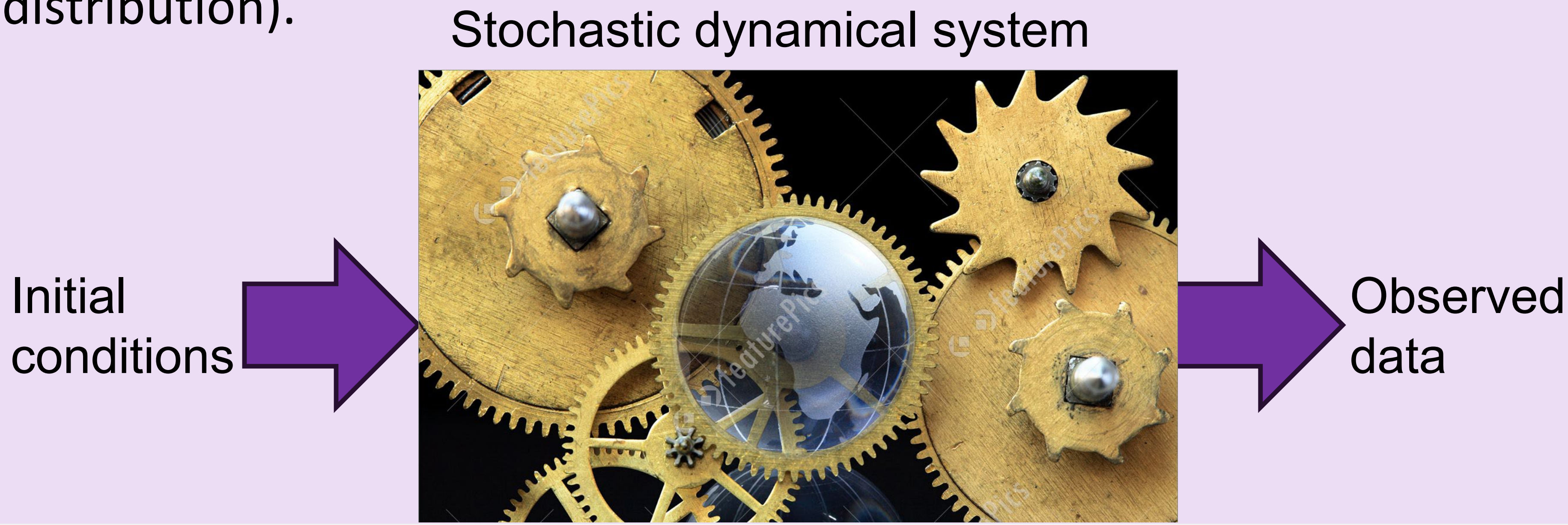
System 2

- Slow, logical, **sequential**, **CONSCIOUS**, linguistic, algorithmic, planning, reasoning
- Explicit knowledge
- DL 2.0

Beyond the iid assumption

- The assumption that the test data is from the same distribution as the training data is too strong, and it is often violated in practice, leading to poor out-of-distribution generalization.

- Consider relaxed assumptions: the test data was generated under the same causal dynamics, but from a different state (initial conditions, interventions, environment), which may be unlikely under the training distribution).

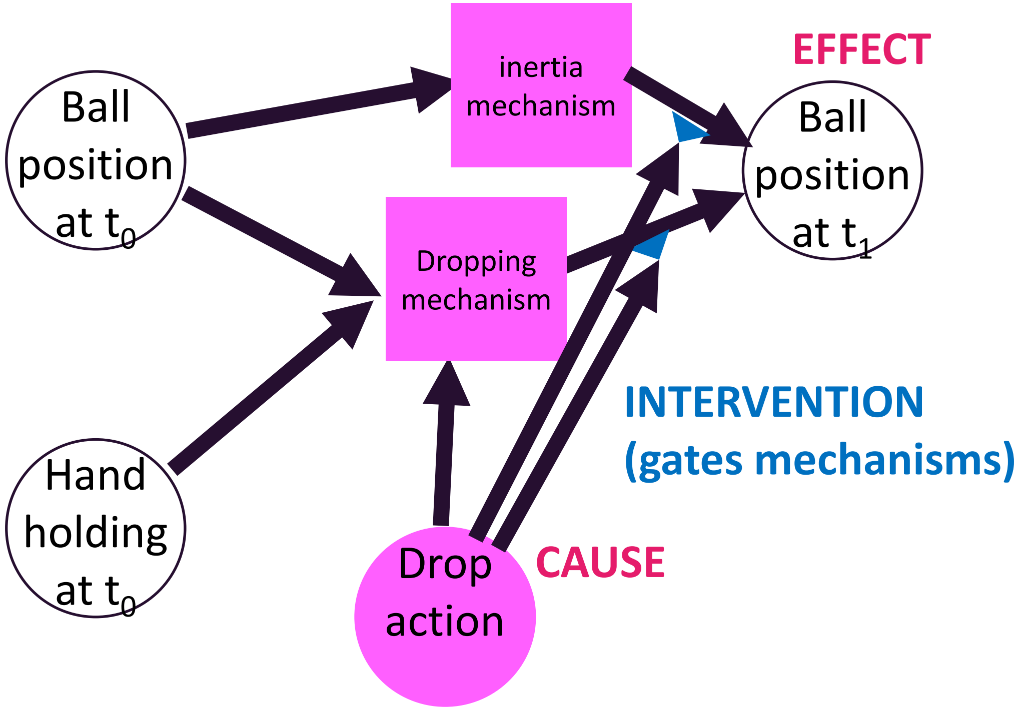


Mila

SPARSE DEPENDENCIES BETWEEN ABSTRACT VARIABLES

Also consistent with Baar's Global Workspace Theory (1997) of conscious processing.

Linguistic example: *"if I drop the ball, it will fall on the ground"*

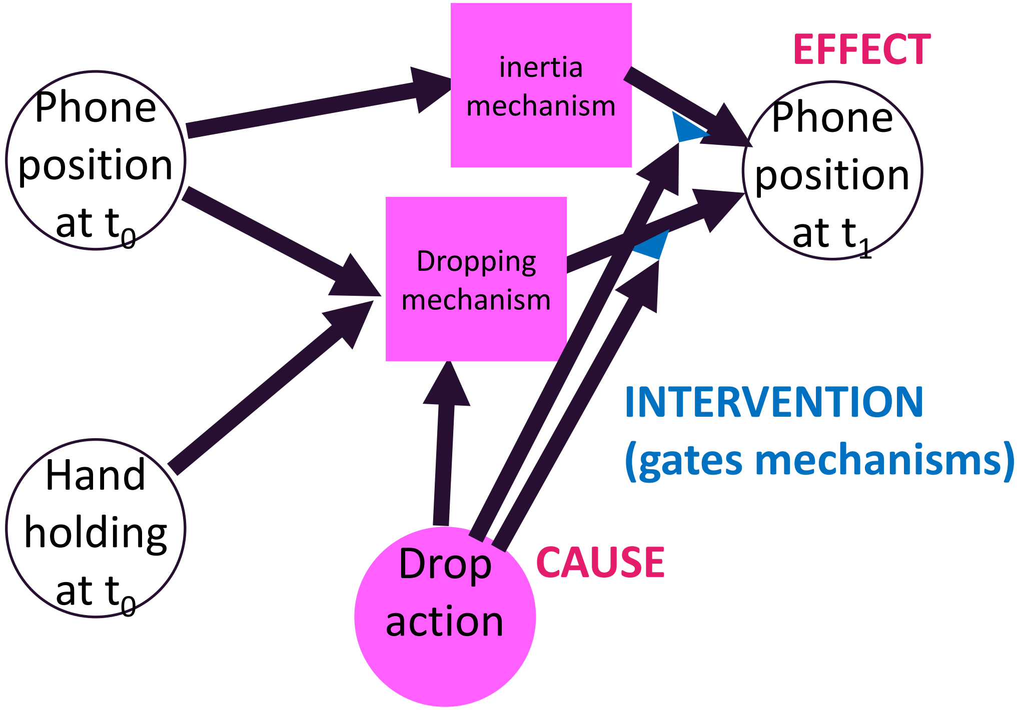


An abstract outcome can be predicted accurately from very few conditioning abstract variables

REUSABLE CAUSAL MECHANISMS

COUNTERFACTUAL

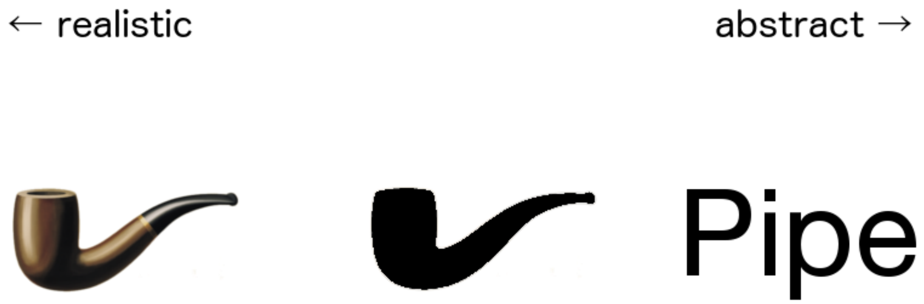
Linguistic example: *"if I had dropped the phone, it would have fallen on the ground"*



The same mechanism can be reused on many instance tuples

DISCRETE, SYMBOLIC, ABSTRACT CONCEPTS

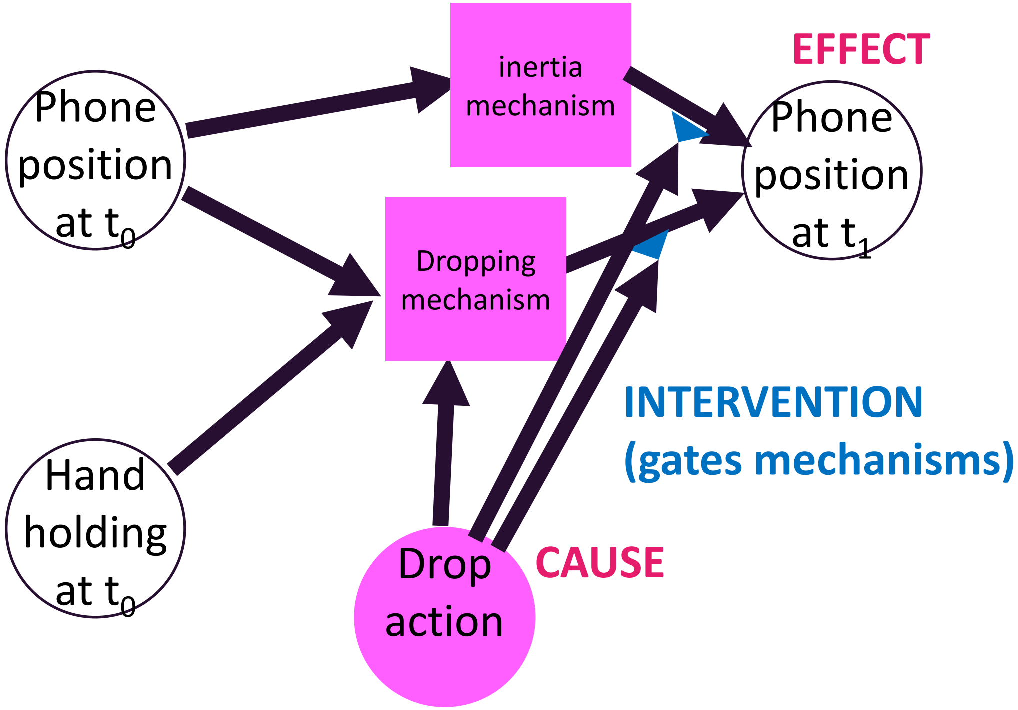
- Language allows communication of simplified, DISCRETE, messages among humans
- Thoughts manipulate such discrete entities
- Evidence that hippocampus represents discrete concepts
- **The bottleneck of discretization in the communication between brain modules may further facilitate systematic generalization, making different brain modules hot-swappable for one another** (e.g. replace a noun by another in a sentence)



SPARSE LOCALIZED INTERVENTIONS

PLANNING

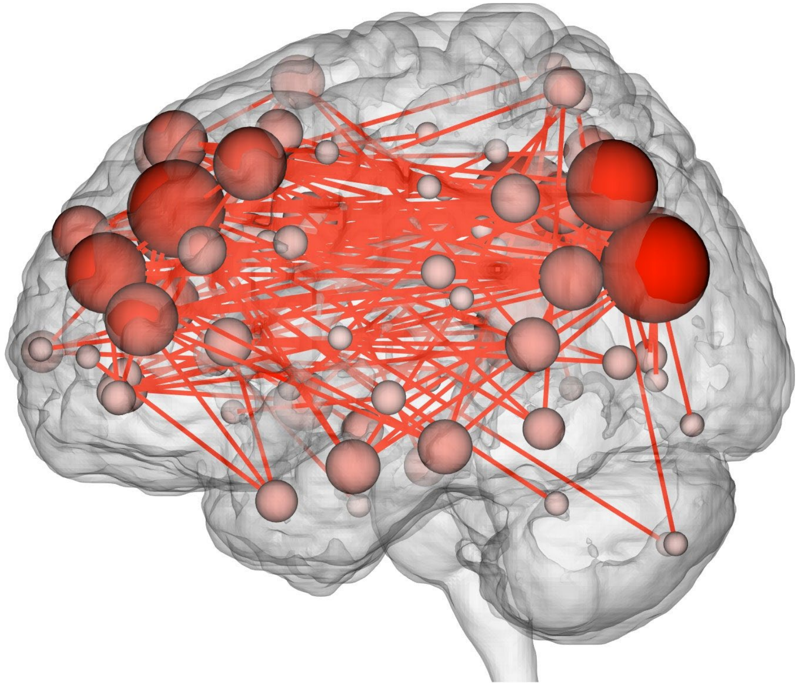
Linguistic example: *"if I decided to drop the phone, it would fall on the ground"*



Only one abstract entity is typically affected by the abstract action = abstract intervention. Typically only one attribute of that entity is directly affected.

LEARNING TO REASON & PLAN

- Reasoning, long-range credit assignment and planning are inference, inherently computationally expensive
- Brains do not use exhaustive search but instead **generate** good candidates
- Conscious processing seems involved in evaluating them for global coherence across the brain's modules
- Attention mechanisms are part of the reasoning policy, converting declarative knowledge into selective computations for inference and decision-making



PROBABILISTIC NEURAL NETS FOR SYSTEM-2 DEEP LEARNING?

- We need neural nets that can sample something like thoughts
- Thoughts are sequentially generated, can be seen to form a hypergraph, each element is a hyperedge linking existing concepts together (a mini-thought)
- Knowledge about the entities thus linked should be modularized (like classical AI rules and facts) and composable (to form these thought-graphs)
- These graphs are stochastic: they represent plausible explanations, counterfactuals or plans, given what we know and the current context

- Can we train neural nets that can implicitly represent such modular knowledge and sample causal / semantic explanations for our experiences?

GFLOWNETS AS THE SWISS ARMY KNIFE OF PROBABILISTIC MODELING

- A hypergraph can be seen as a **set** of hyperedges, each = application of a rule to a small set of variables
- Decompose **sampling of set X according to ANY ORDER** of steps generating each element
- X can have variable size or even be infinite (like all graphs)
- Learn a **sampling policy** to generate X (sampling the order too)
- Intermediate quantities, the **flows**, are implicitly **marginalizing over future choices**
- The flow function **generalizes**: no need to see all the possible sequences
- The distribution is **represented in a structured, compositional way**

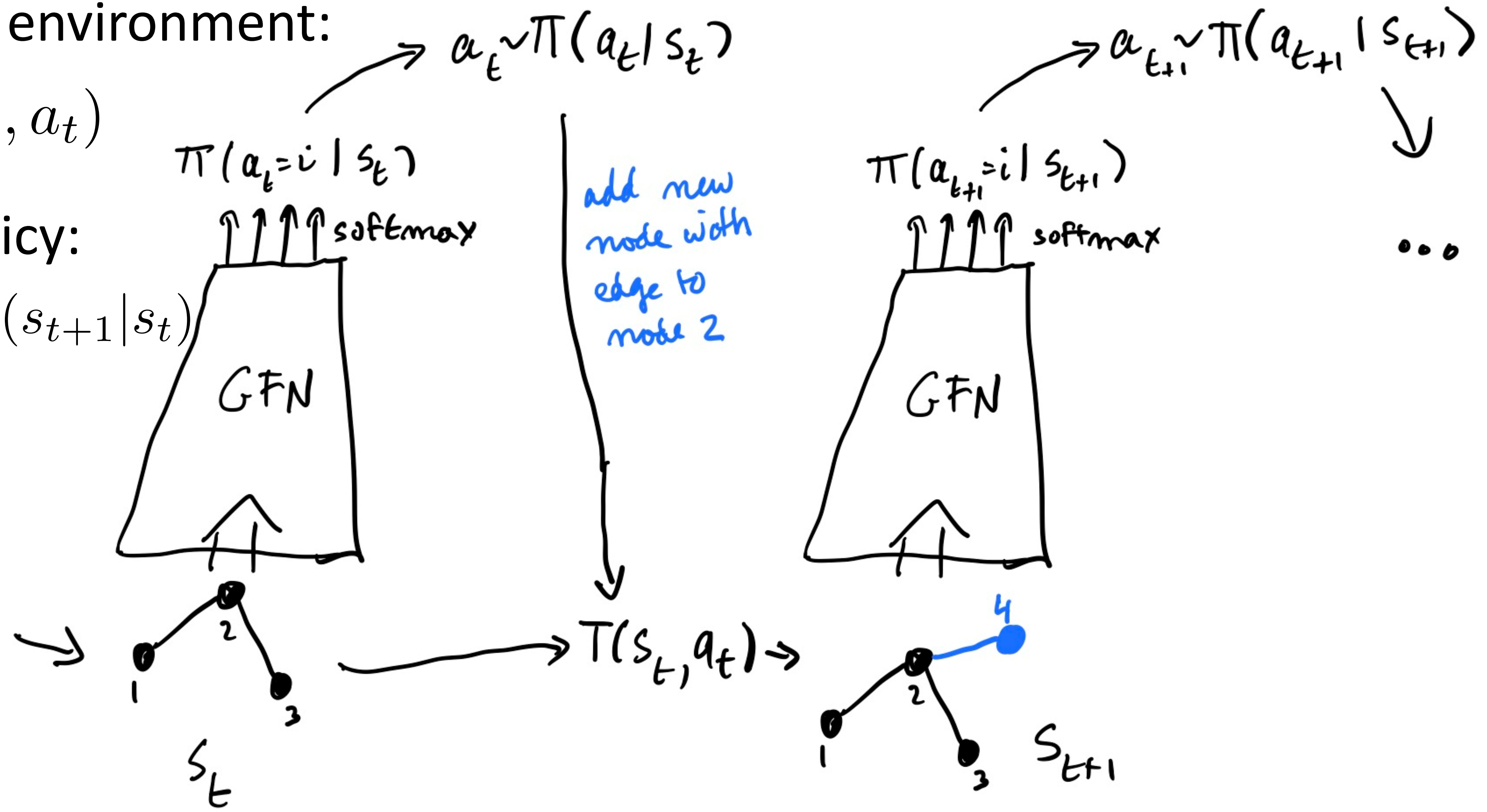
NEURAL NET FLOW ESTIMATOR

Deterministic environment:

$$s_{t+1} = T(s_t, a_t)$$

GFlowNet policy:

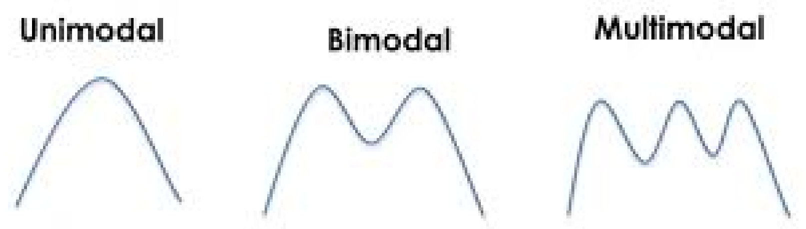
$$\pi(a_t|s_t) = P_F(s_{t+1}|s_t)$$



Sampling from Modes of a Reward Function

Instead of minimizing the loss or maximizing rewards, sample solutions proportionally to the reward. To obtain a diverse set of solutions, we would like cover all the modes a reward function *R(x)* which indicates how desirable a solution *x* is. We could do that by sampling iid from P_T:

$$P_T(x) \approx \frac{R(x)}{Z} = \frac{R(x)}{\sum_x R(x)}$$



Amortized Generation from Energy Function

Given *R(x)=exp(-energy(x))*, train policy sampling *x* with probability

$$P_T(x) \approx \frac{R(x)}{Z} = \frac{R(x)}{\sum_x R(x)}$$

MCMC methods are expensive at generation time and face intractable mode-mixing challenges

Mila

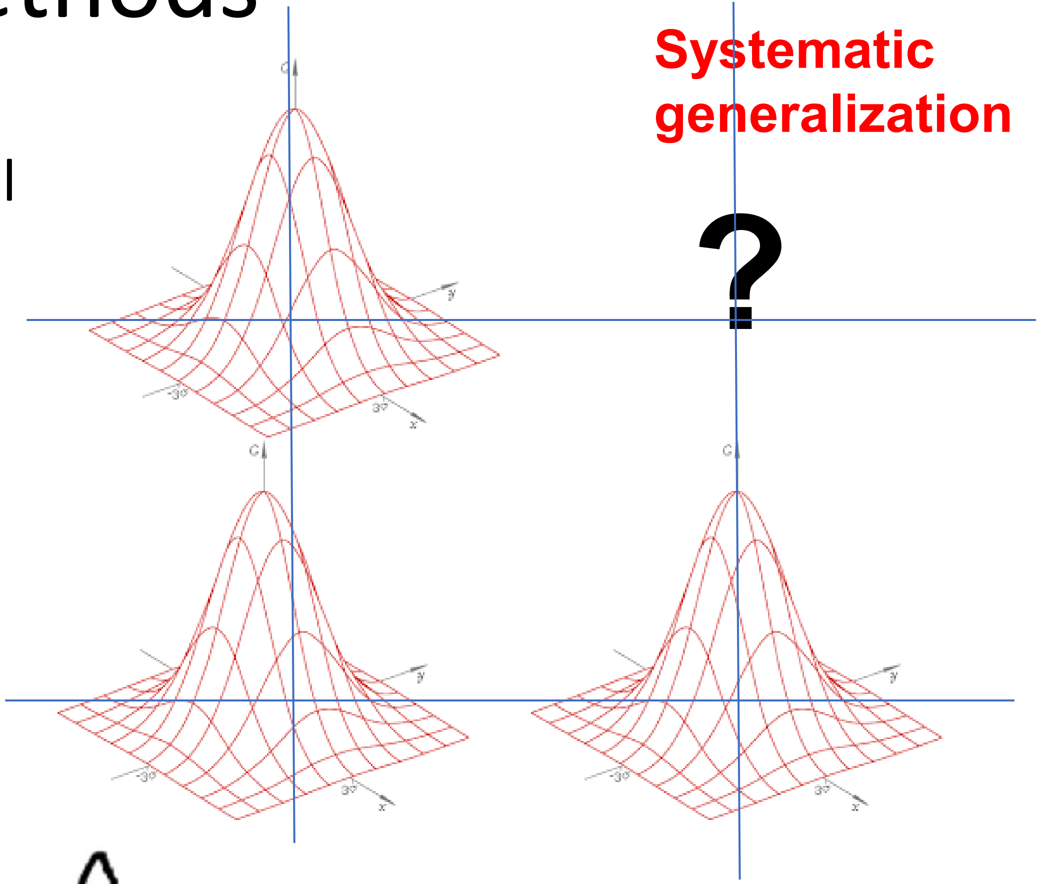
The Limitations of MCMC Methods

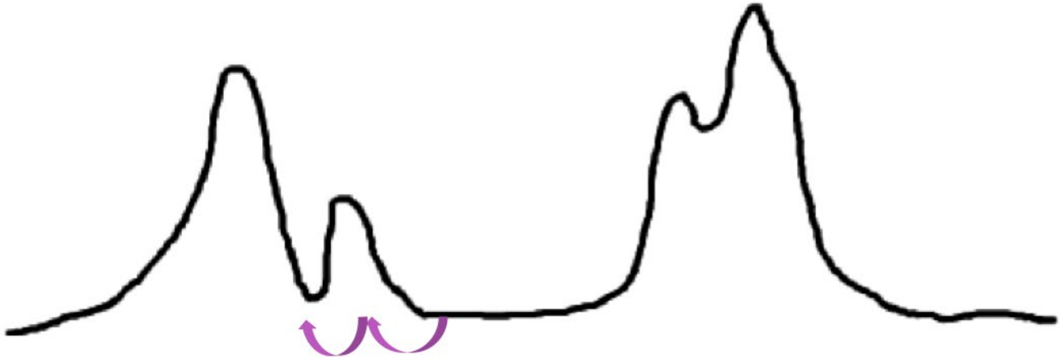
MCMC makes a large number of small, noisy, local moves generally towards higher probability

Asymptotically samples from

$$P(x) \propto \exp(-\text{energy}(x))$$

But exponential time to mix between modes which occupy a small volume and are far from each other.





Mila

NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS'2021
arXiv:2106.04399

Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, Yoshua Bengio



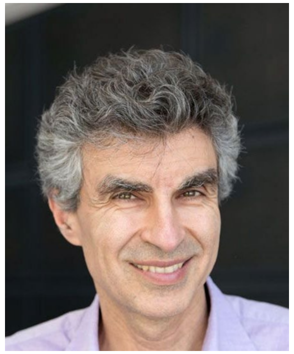
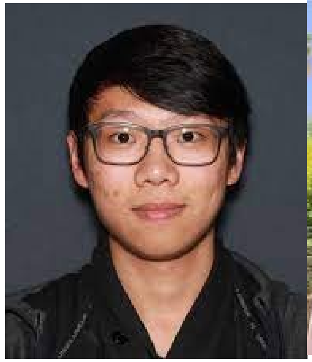


Mila

GFlowNets Foundations

arXiv:2111.09266

Yoshua Bengio, Tristan Deleu, Edward Hu, Mo Tiwari, Salem Lahlou, Emmanuel Bengio

Mila

GFlowNet Objectives = *matching constraints*

Flow Matching:

flow estimator

$$\left(\log\Big(\sum_{s'}\hat{F}(s'\rightarrow s)\Big) - \log\Big(\sum_{s''}\hat{F}(s\rightarrow s'')\Big)\right)^2$$

Reward Matching:
(combined with FM or DB)

$$\left(\log(\hat{F}(x)) - \log(R(x))\right)^2$$

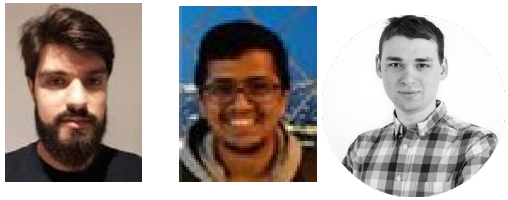
Detailed Balance:

$$\left(\log(\hat{F}(s)P_F(s'|s)) - \log(P_B(s|s')\hat{F}(s'))\right)^2$$

Trajectory Balance:

$$\left(\log\Big(F(s_0)\prod_{t=1}^{n}P_F(s_t|s_{t-1})\Big) - \log\Big(R(s_n)\prod_{t=1}^{n}P_B(s_{t-1}|s_t)\Big)\right)^2$$

Mila **Related to hierarchical variational inference, enables lower-variance estimator & offline training**



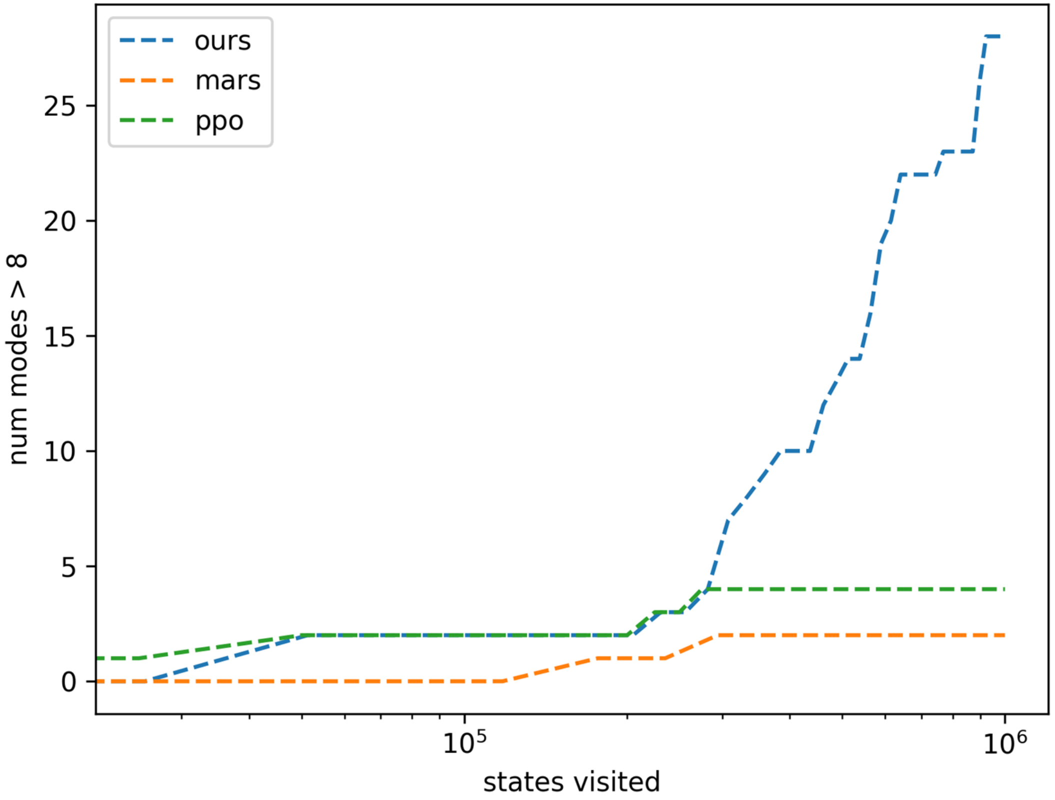
E. Bengio, M. Jain & M. Korablyov @ *Mila*

More Diverse Solutions with GFlowNets (NeurIPS 2021)

GFlowNet is used to generate molecular candidates in LambdaZero pipeline, compared to MCMC (MARS) and RL (PPO) in the same discovery pipeline.

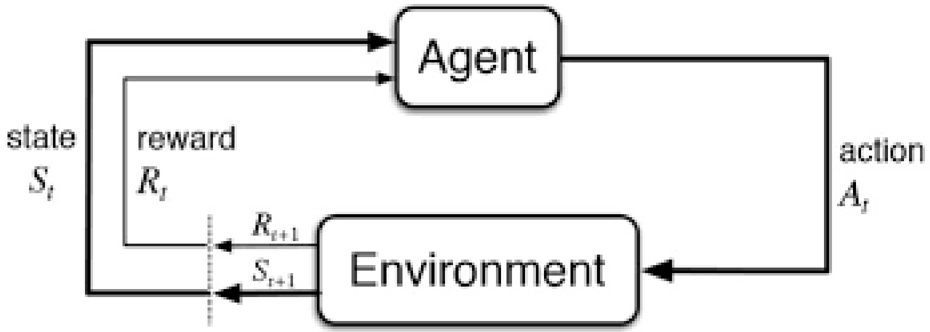
Two potential cluster-center molecules belong to the same "mode" if their Tanimoto similarity >0.7 and we consider only molecules with reward > 8

GFlowNet discovers 10x # of modes!



Mila

Similarities & Differences with Regular RL



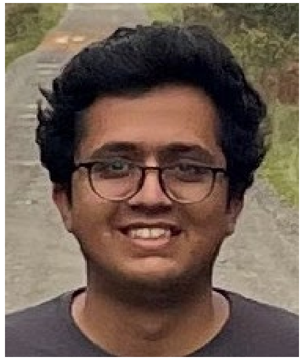
- GFlowNet samples proportionally to reward rather than maximising it → PPO is happy once it finds a mode or a few modes
- Entropy regularization & control as inference are equivalent to GFlowNets if the DAG is a tree, but otherwise they fail to recover the right distribution
- GFlowNet enables estimation of marginalized quantities: partition function, P(subset of variables|other subset), entropy, etc

Mila

Trajectory Balance: Improved Credit Assignment in GFlowNets

arXiv:2201.13259

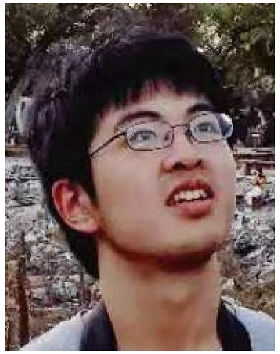
Nikolay Malkin, Moksh, Jain, Emmanuel Bengio, Chen Sun, Yoshua Bengio

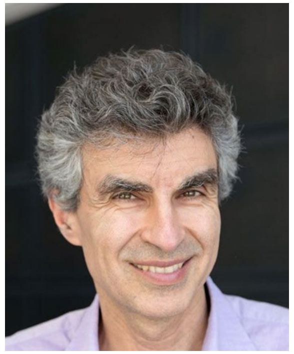

Generative Flow Networks for Discrete Probabilistic Modeling

arXiv:2202.01361

Dinghuai Zhang, Nikolay Malkin, Zhen Liu, Alexandra Volokhova, Aaron Courville, Yoshua Bengio



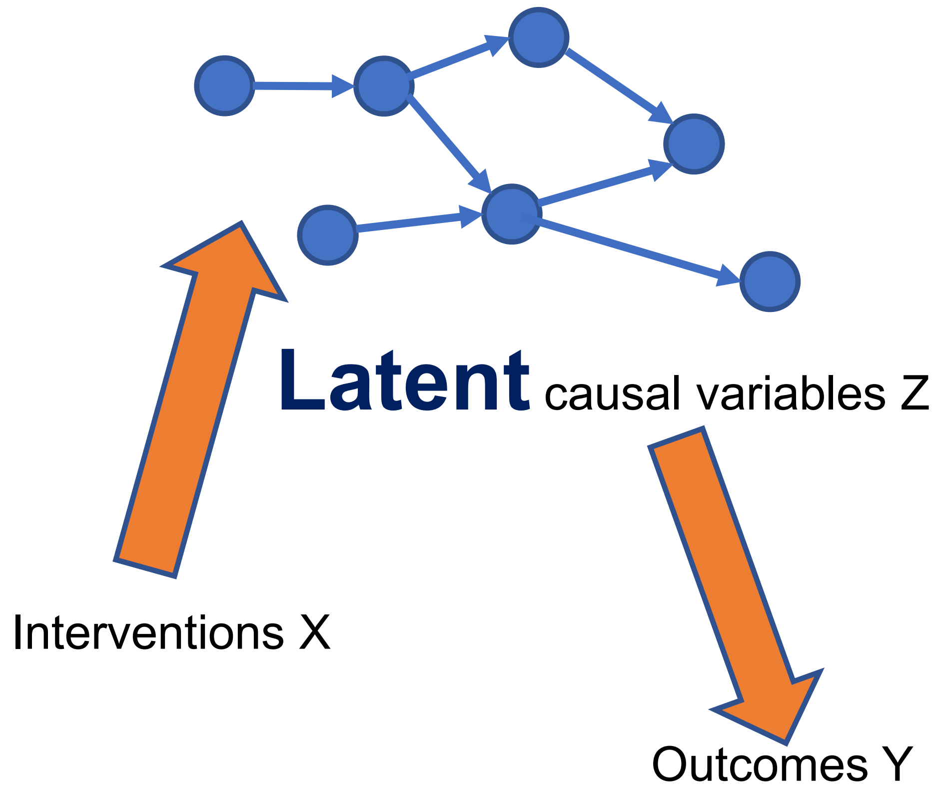


Mila

Posterior over Causal Models

Conjecture: the most complete characterization of ambiguity due to both **non-identifiability** of causal structure and finite data is provided by the **Bayesian posterior over causal models**, given the observational and experimental data.



Mila

Information Gain from Posterior for Experimental Design

The posterior can be used to estimate the mutual information between the model and a candidate experiment.

Another GFlowNet can then be trained to generate experiments that have a high information gain.

Defeats issues with just maximizing error (aleatoric uncertainty) and difficult to predict but deterministic outcomes (curriculum, chaos).

Bayesian Structure Learning with Generative Flow Networks

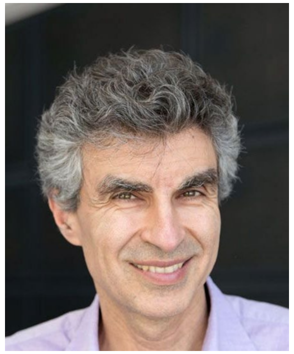
UAI’2022,   arXiv:2202.13903

Tristan Deleu, António Góis, Chris Emezue, Mansi Rankawat, Simon Lacoste-Julien, Stefan Bauer, Yoshua Bengio









Mila

GFN4Bayes: Amortized Sampling from Bayesian Posteriors

- The energy fn parameters = latent variables
- Reward for (Z, θ) given dataset X

$$R(X, Z, \theta) = e^{-(\underbrace{\mathcal{E}_0(\theta)}_{\text{prior energy}} + \sum_t \underbrace{\mathcal{E}(X_t, Z_t, \theta)}_{\text{joint likelihood energy}})}$$

$$\mathcal{E}(X_t, Z_t, \theta) = -\log p(Z_t|\theta) - \log p(X_t|Z_t, \theta)$$

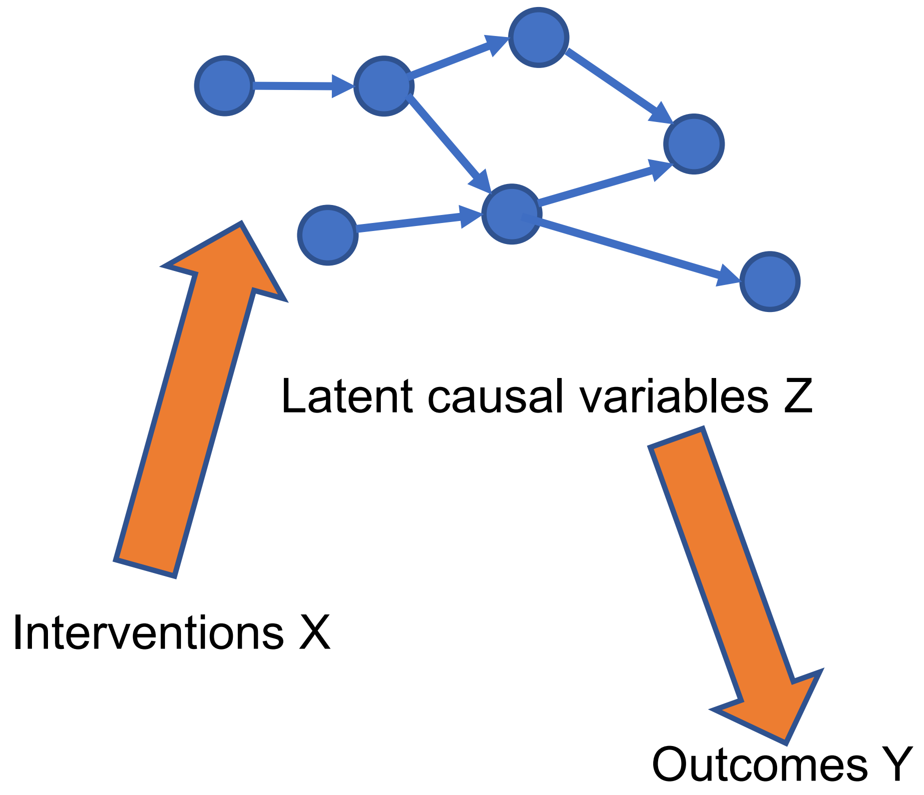
- SGD on the GFlowNet objective: use just a minibatch in the sum and scale by T/B → unbiased gradient!

Mila

Causal Mechanisms, Intervention Model, Outcomes Model

Intervention model: maps actual interventions (e.g. experimental specs) to surgery in the default causal model (e.g. DO-intervention, but generally unrealistic)

Outcomes model: could be high-dimensional phenotype, images, noisy expression data



Mila

MODULAR GFLOWNETS FOR HUMAN-LIKE REASONING

- Each module encapsulates knowledge (with local parameters and expertise on specific concepts)
- The GFlowNet transitions corresponding to the GWT competition between experts sample one of these modules and the content (hyperedge) it proposes as the next piece of thought
- Each module also learns an energy function corresponding to the piece of world-model it embodies
- The same fixed-size neural nets can create new abstract concepts (entities = nodes in these graphs and structural relations between them) as more data is observed.
- The entropy estimation ability of GFlowNets can drive knowledge acquisition

GFLOWNET TUTORIAL

https://tinyurl.com/gflownet-tutorial

Questions?

Université de Montréal

IVADO Mila CIFAR