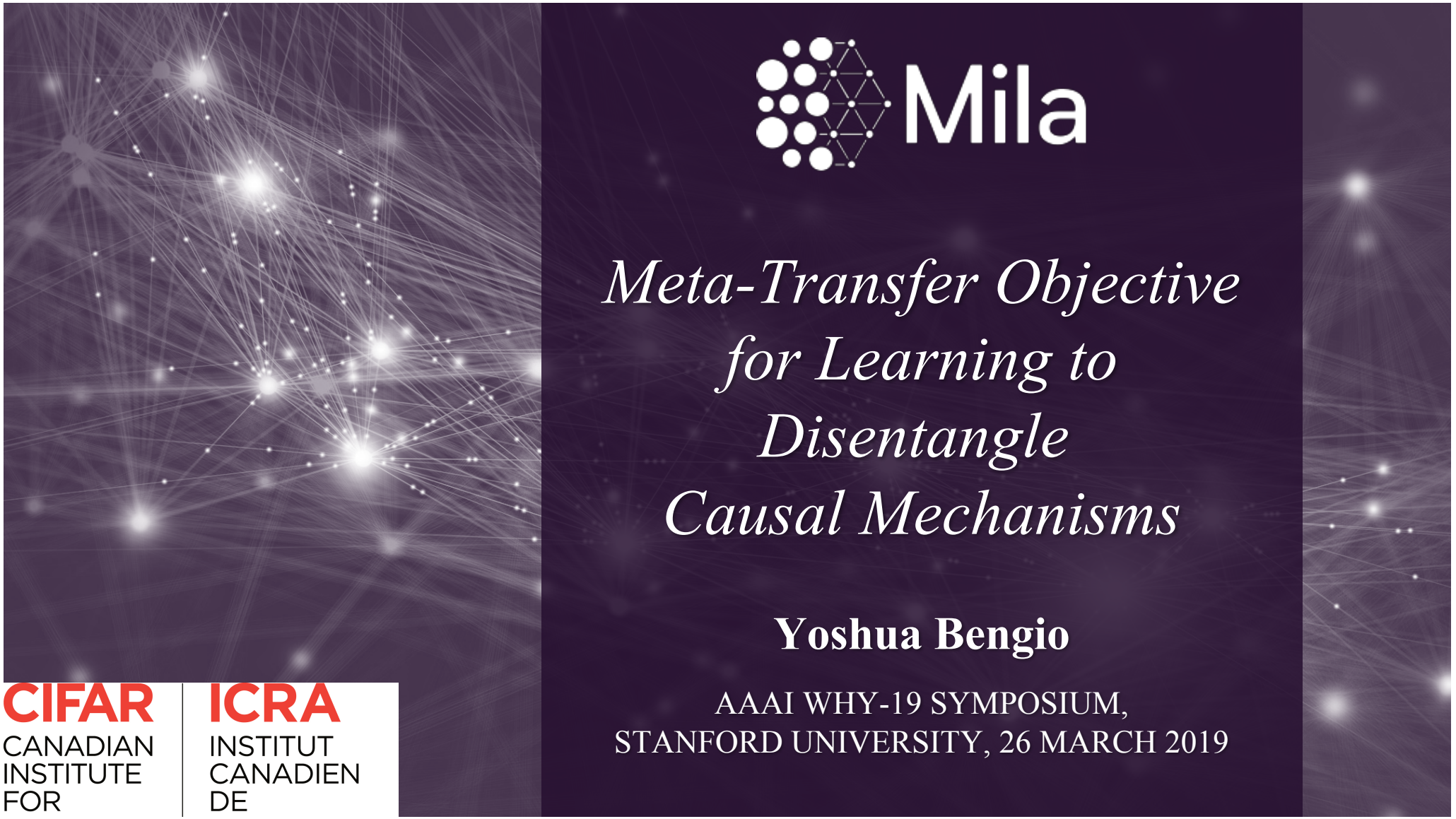

Mila
Meta-Transfer Objective
for Learning to
Disentangle
Causal Mechanisms
Yoshua Bengio
AAAI WHY-19 SYMPOSIUM,
STANFORD UNIVERSITY, 26 MARCH 2019
CIFAR
CANADIAN
INSTITUTE
FOR
ICRA
INSTITUT
CANADIEN
DE

Deep Objective: discover causal representation

- What are the right representations? Causal variables explaining the data
- How to disentangle them?
- How to discover their causal relationship, the causal graph?
- How to modularize knowledge for easier re-use & adaptation, good transfer?

Mila

Beyond iid: Independent Mechanisms and Small Change Hypothesis

- Independent mechanisms:
  - changing one mechanism does not change the others *(Peters, Janzig & Scholkopf 2017)*
- Small change:
  - Non-stationarities, changes in distribution, involve few mechanisms (e.g. the result of a single-variable intervention)

Mila

What if we had the right causal structure?

CLAIM: Under the hypothesis of independent mechanisms and small changes across different distributions:

- smaller sample complexity to recover from a distribution change
  - E.g. for transfer learning, agent learning, domain adaptation, etc.

Mila

Zero Gradient on the Unchanged Mechanisms

Graphical model is parametrized via a set of modules for each P(Variable | pa(Variable))

Proposition 1. *The expected gradient over the transfer distribution of the regret (accumulated negative log-likelihood during the adaptation episode) with respect to the module parameters is zero for the parameters of the modules that (a) were correctly learned in the training phase, and (b) have the correct set of causal parents, corresponding to the ground truth causal graph, if (c) the corresponding ground truth conditional distributions did not change from the training distribution to the transfer distribution.*

Mila

Simple Running Example

- Consider two r.v. A, B, with A cause of B.
- Correct causal model decomposes
  - P(A,B) = P(A) P(B|A)

- Consider 2 distributions P_1 and P_2, only P(A) changes
- If we first train on P_1 and we have the right decomposition, adapting on P_2 is fast because

$$E_{P(B|A)}[\frac{\partial \log P_\theta(B|A)}{\partial \theta}] \approx 0 \text{ when } P_\theta(B|A) \approx P(B|A)$$

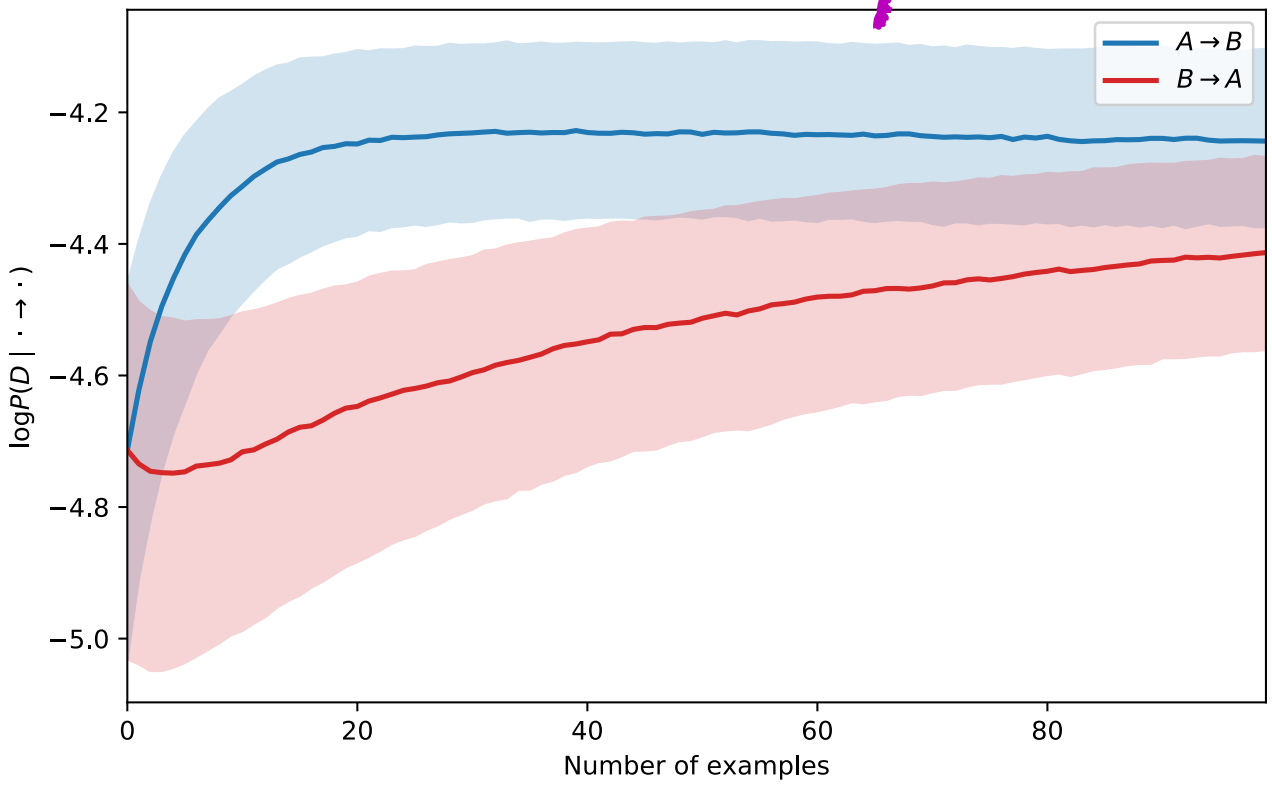
Mila

Wrong Knowledge Factorization Leads to Poor Transfer

- With the wrong factorization P(B) P(A|B), a change in P(A) influences all the modules, all the parameters
  - poor transfer: all the parameters must be adapted
- This is the normal situation with standard neural nets: every parameter participates to every relationship between all the variables
  - this causes ***catastrophic forgetting, poor transfer, difficulties with continual learning or domain adaptation, etc***

Mila

Empirical Confirmation: Correct Causal Structure Leads to Faster Adaptation



A→B is the correct causal structure: faster online adaptation to modified distribution = lower NLL regret

Mila

The Challenge of Systematic Generalization

- See ***'Systematic Generalization: what is required and can it be learned' Bahdanau et al & Courville ICLR 2019***
- Same set of concepts, but combined in different ways in the transfer setting
- Good generalization inside training distribution does not necessarily give good transfer

Mila

Turning a Hindrance into a Useful Signal

- Changes in distribution (nonstationarities in agent learning, transfer scenarios, etc) are seen as a bug in ML, a challenge
- Turn them into a feature, an asset, to help discover causal structure, or more generally to help **factorize knowledge:**
- **Tune knowledge factorization (e.g. causal structure) to maximize fast transfer**

Mila

Simple Training Scenario

- Train on first distribution P_1, then measure online generalization error as we adapt on transfer distribution P_2
- Meta-optimize that online error wrt structural parameters, e.g.
  - the encoder: observations → causal variables
  - the causal graph (which variables are direct causes of which variables)

Mila

Running Example

- A and B are either discrete or continuous
- Separately parametrize modules P(A), P(B|A), P(B), P(A|B)
- First consider only two structural hypotheses (e.g. A→B is ground truth)
  - correct: $P_{A\to B}(A,B) = P_{A\to B}(A)P_{A\to B}(B \mid A)$
  - incorrect: $P_{B\to A}(A,B) = P_{B\to A}(B)P_{B\to A}(A \mid B)$

Mila

Soft Parametrization

- Each transfer adaptation episode of length T
- Regret for episode-wise mixture between 2 hypotheses:

$$\mathcal{R} = -\log\left[\sigma(\gamma)\mathcal{L}_{A\to B} + (1-\sigma(\gamma))\mathcal{L}_{B\to A}\right]$$

$$\mathcal{L}_{A\to B} = \prod_{t=1}^{T} P_{A\to B}(a_t, b_t\,;\,\theta_t)$$

$$\mathcal{L}_{B\to A} = \prod_{t=1}^{T} P_{B\to A}(a_t, b_t\,;\,\theta_t),$$

Mila

Transfer Regret Gradient

Proposition 2. *The gradient of the negative log-likelihood of the transfer data in Equation* (2) *wrt. the structural parameter* $\frac{\partial \mathcal{R}}{\partial \gamma}$ *is given by*

$$\frac{\partial \mathcal{R}}{\partial \gamma} = \sigma(\gamma) - P(A \to B \mid D_2), \tag{3}$$

where D_2 *is the transfer data, and* $P(A \to B \mid D_2)$ *is the posterior probability of the hypothesis* $A \to B$ *(when the alternative is* $B \to A$*). Furthermore, this can be equivalently written as*

$$\frac{\partial \mathcal{R}}{\partial \gamma} = \sigma(\gamma) - \sigma(\gamma + \Delta), \tag{4}$$

where $\Delta = \log \mathcal{L}_{A \to B} - \log \mathcal{L}_{B \to A}$ *is the difference between the log-likelihoods of the two hypotheses on the transfer data* D_2*.*

Mila

Convergence to Correct Causal Hypothesis

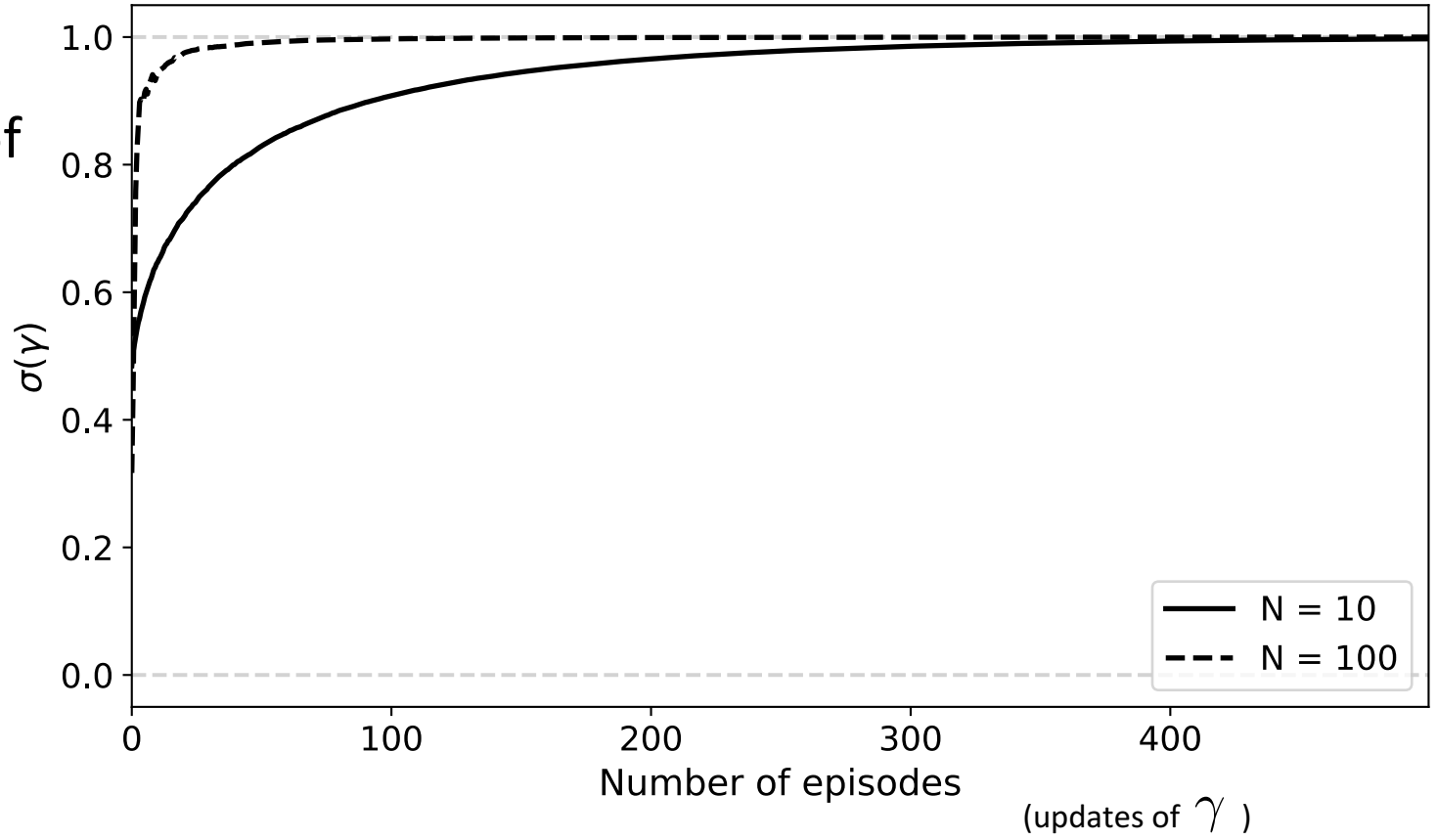
Proposition 3. *Stochastic gradient descent (with appropriately decreasing learning rate) on* $E_{D_2}[\mathcal{R}]$ *with steps from* $\frac{\partial \mathcal{R}}{\partial \gamma}$ *converges towards* $\sigma(\gamma) = 1$ *if* $E_{D_2}[\log \mathcal{L}_{A \to B}] > E_{D_2}[\log \mathcal{L}_{B \to A}]$, *or* $\sigma(\gamma) = 0$ *otherwise.*

Mila

Experimental Validation

Tabular parametrization of marginals and conditionals of bivariate model.

Correct causal graph can be recovered

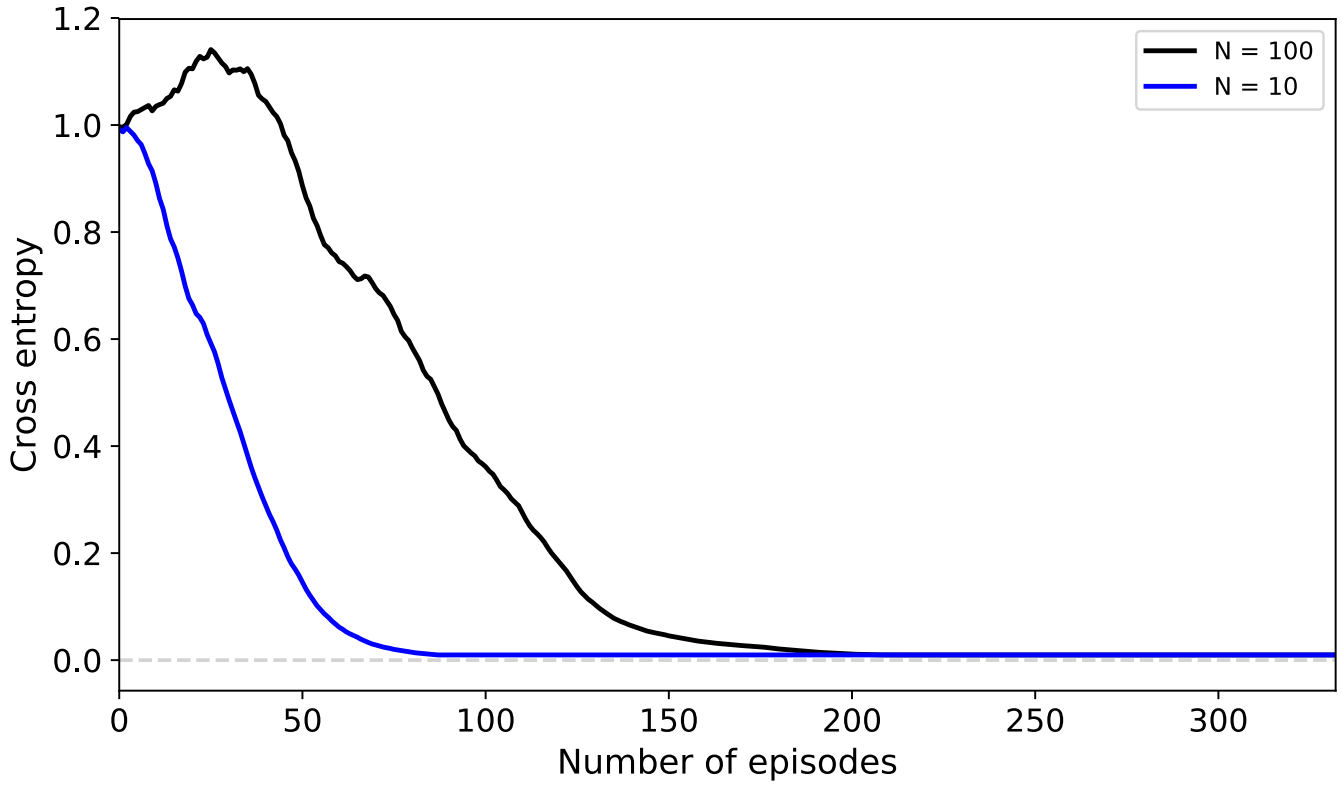


Mila

MLP Conditionals Results

Each conditional is represented by a one-hidden-layer MLP with one-hot inputs, softmax outputs.

It works again better for larger number of values N (N^2 vs N parameters have changed, incorrect vs correct causal graph)

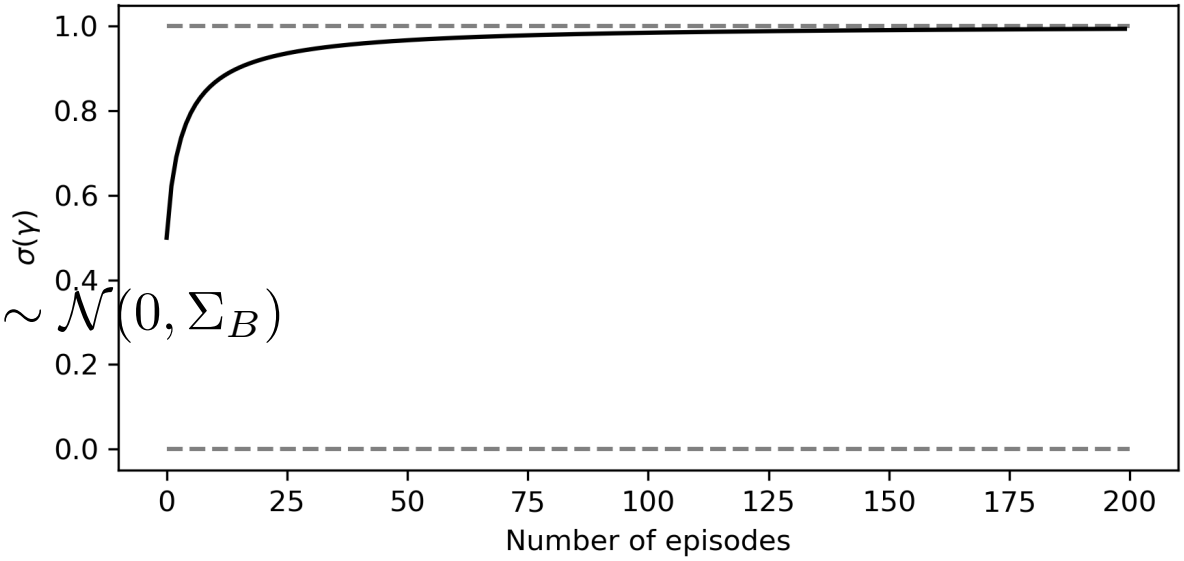


Mila

Linear Gaussian Results

$$A \sim \mathcal{N}(\mu_A, \Sigma_A)$$
$$B := \beta_1 A + \beta_0 + N_B \qquad N_B \sim \mathcal{N}(0, \Sigma_B)$$

$$P_{A\to B}(A) = \mathcal{N}(A; \hat{\mu}_A$$
$$P_{A\to B}(B \mid A = a) = \mathcal{N}(B; \hat{W}_1$$
$$P_{B\to A}(B) = \mathcal{N}(B; \hat{\mu}_B, \hat{\Sigma}_B)$$
$$P_{B\to A}(A \mid B = b) = \mathcal{N}(B; \hat{V}_1 b + \hat{V}_0, \hat{\Sigma}_{B\to A})$$



Quickly recovers the correct structure

Mila

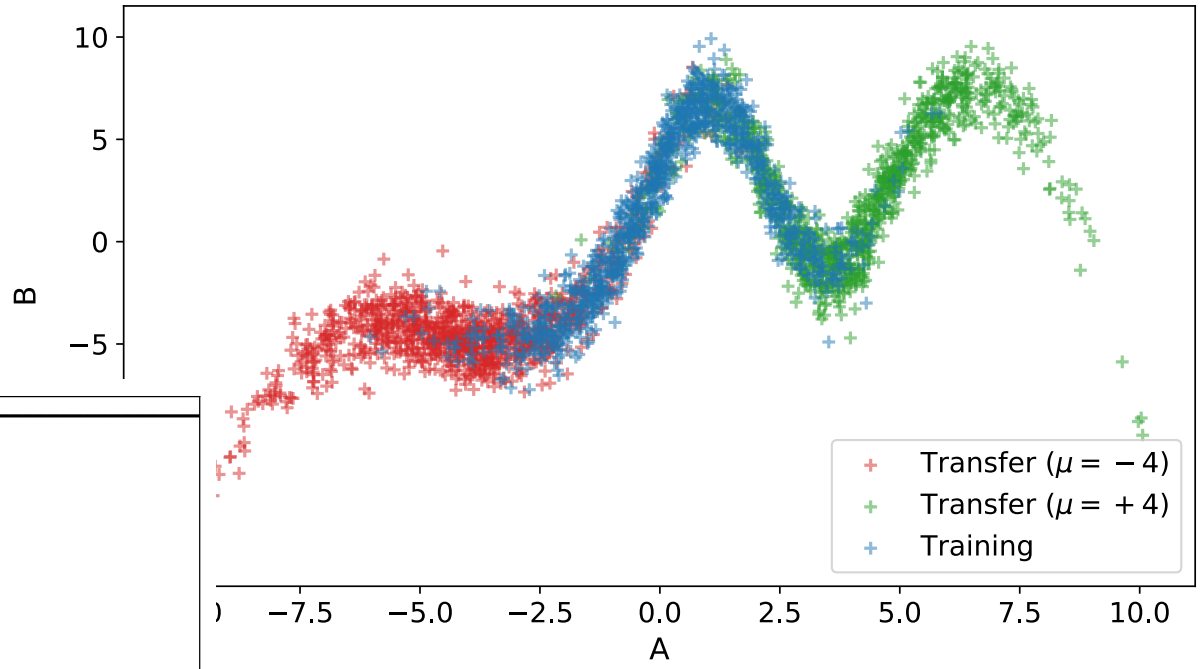
Multimodal Continuous Variables Results

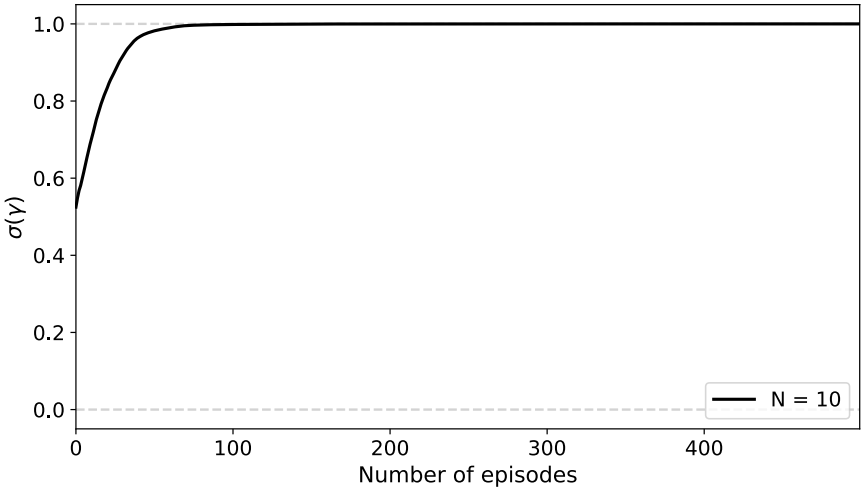
Ground truth data = spline+noise

Vary mean of P(A) to obtain transfer distributions

Models = MLP with GMM output

Quickly recover causal direction.

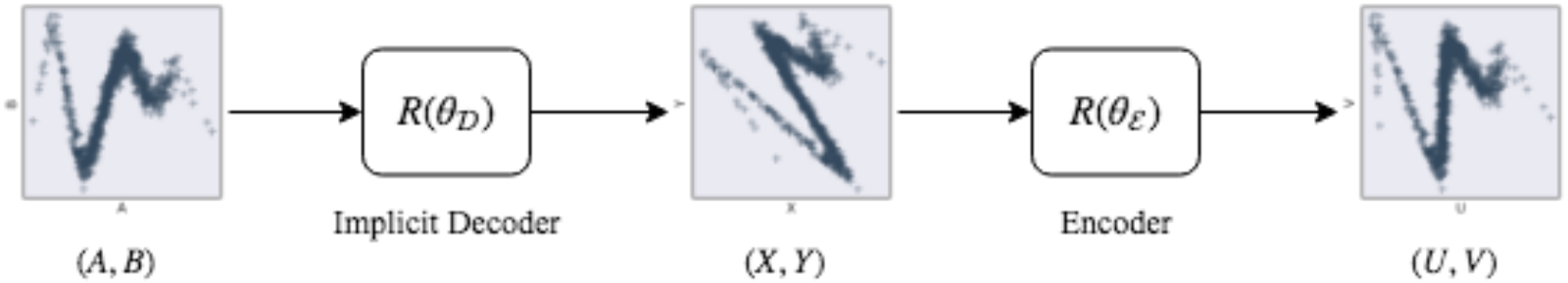




Mila

Disentangling the Causes

- Realistic settings: causal variables are not directly observed
- Need to learn an encoder which maps raw data to causal space
- Consider both the encoder parameters and the causal graph structural parameters as meta-parameters trained together wrt proposed meta-transfer objective



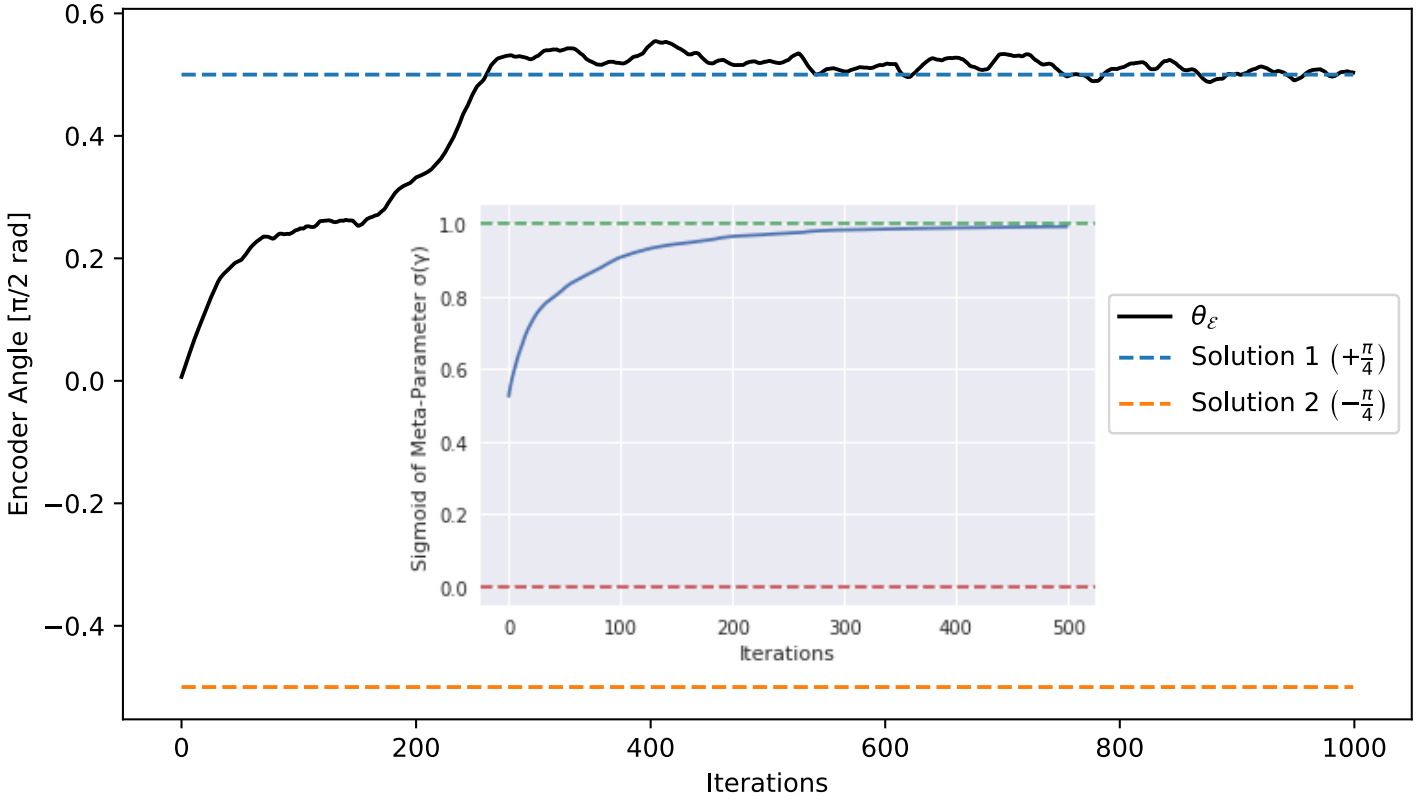
Mila

Simplest possible scenario: linear mixing (rotating decoder) and unmixing (rotating decoder)

Disentangling the Causes

Recovers the correct encoder parameter: disentangles up to permutation.

Simultaneously recovers causal direction (smaller figure inside).



Mila

With More Variables

- Pre-train conditionals with random dropout on the graph structure binomials
- Ground truth changes in distribution: randomly modify one of the conditionals or marginals (generalized intervention)
- Each transfer distribution is a meta-example
- Each variable | parents: modeled by MLP
- Structural choice *B*: which parents? Represented by binomial probability (belief) of dropping that parent.
- Meta-objective:

$$\mathcal{L}_{B_i} = \prod_t P_{B_i}(V_i = v_{ti} \mid \mathrm{pa}(i, v_t, B_i)) \qquad \mathcal{L}_B = \prod_i \mathcal{L}_{B_i}$$

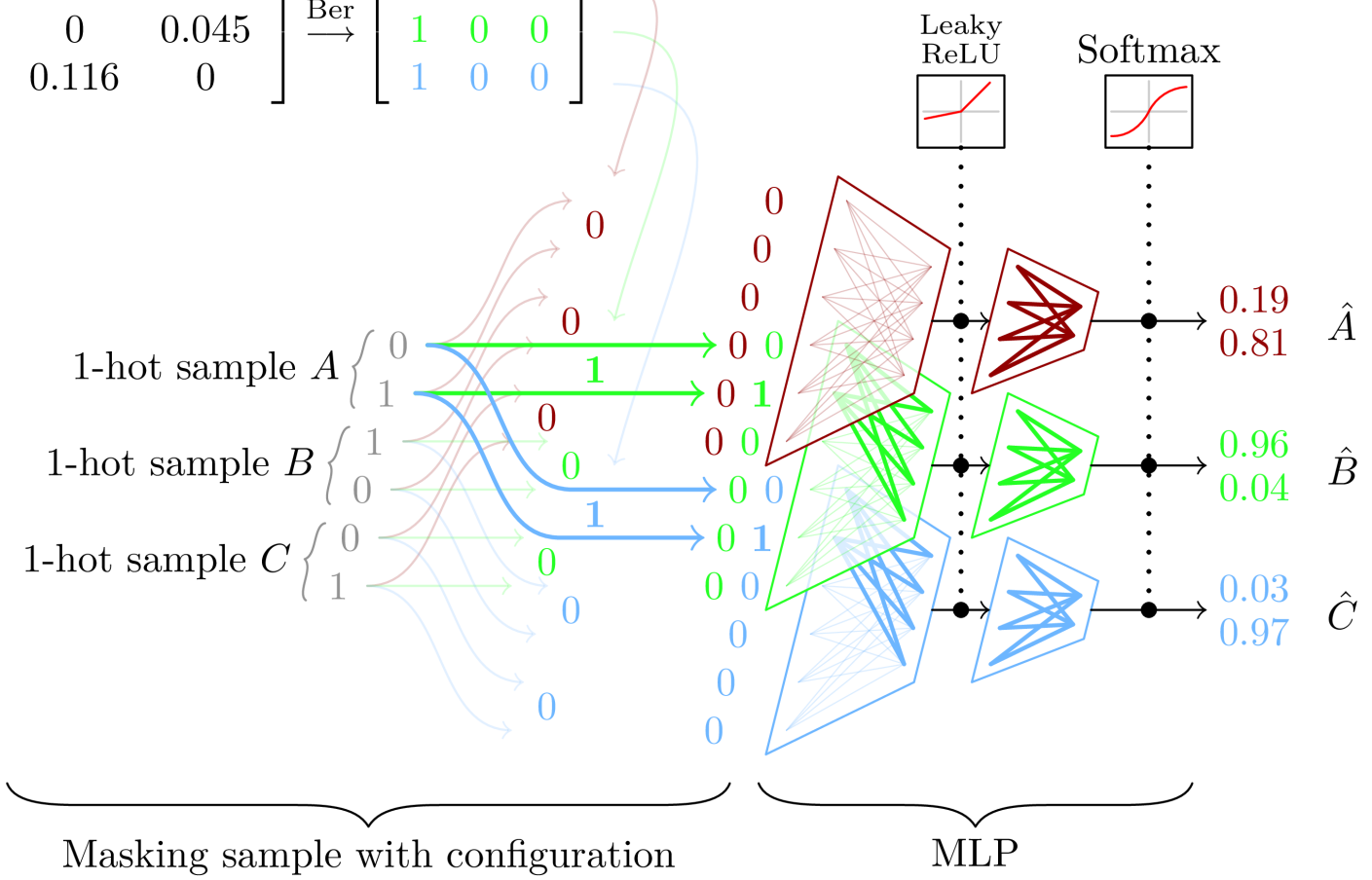
$$\mathcal{R} = -\log E_B[\mathcal{L}_B] \quad \Rightarrow \quad \mathcal{R} = -\sum_i \log \sum_{B_i} P(B_i)\mathcal{L}_{B_i}$$

theorem

Mila

Multivariate Categorical MLP Conditionals

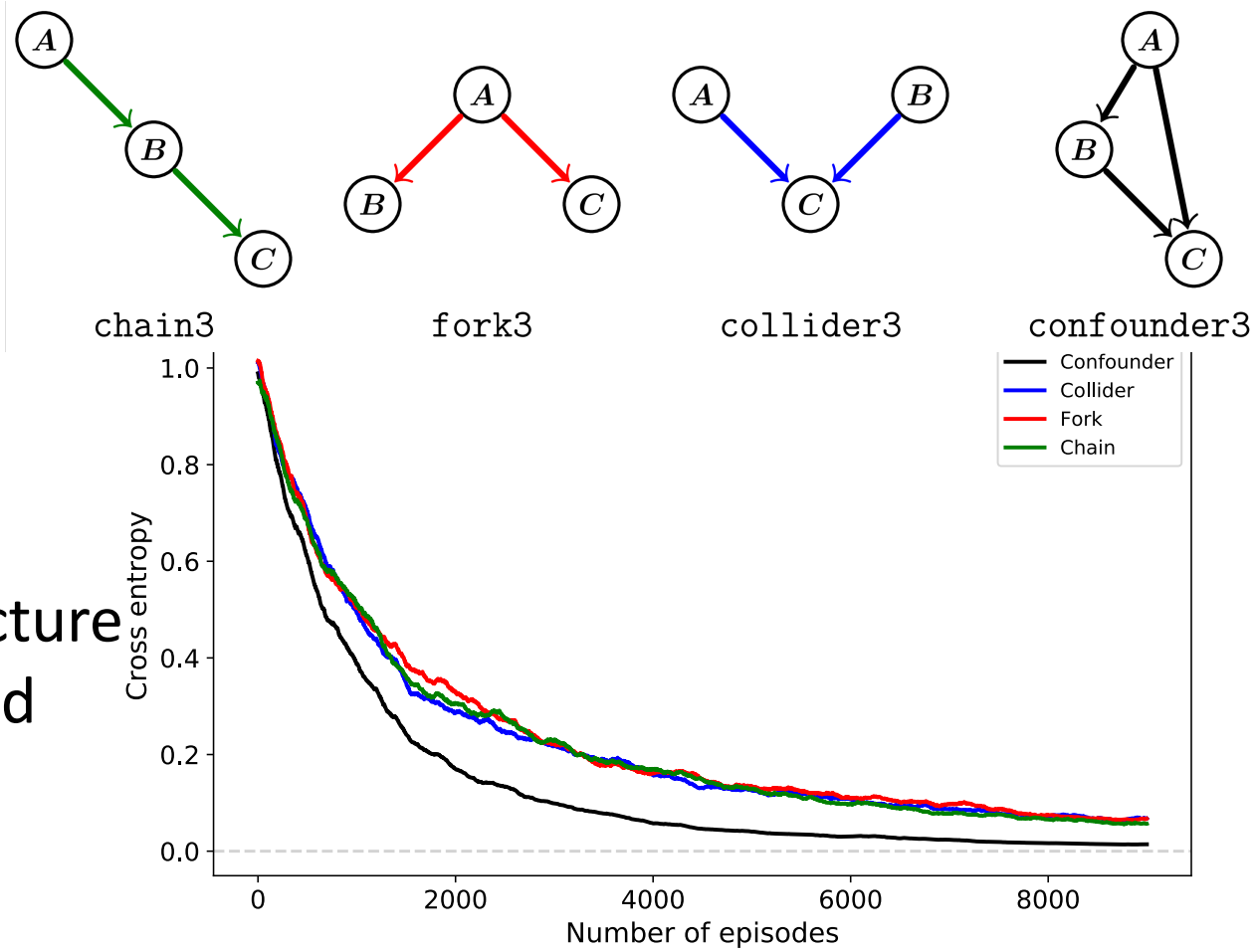
$$\sigma(\gamma) \rightarrow \begin{bmatrix} 0 & 0.088 & 0.090 \\ 0.894 & 0 & 0.045 \\ 0.973 & 0.116 & 0 \end{bmatrix} \xrightarrow{\text{Ber}} \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$



Mila

Results with Three Variables

Evolution during meta-training of cross-entropy between ground truth graph structure and meta-learned beliefs



Mila

Causal Team @ Mila

Yoshua Bengio

Tristan Deleu

Nasim Rahaman

Rosemary Ke

Olexa Bilaniuk

Anirudh Goyal

Chris Pal

Rémi Le Priol