# A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms

Presenter: Zhe Wang

https://qdata.github.io/deep2Read

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, et.al
**@ Mila**

201909

# Content

# Background

Knowledge in a complex distribution can be represented in a modularized way, and those modules are independent.(Chain Rule)

$$P(A, B) = P(A)P(B|A)$$
$$P(A, B) = P(B)P(A|B)$$

Small intervention (perturbation): The transfer distribution will change only one of few of the modules.

ICM assumption: $P(X)$ and $P(Y|X)$ are not related.

# Introduction

Target: Disentangling causal mechanisms within a joint distribution.

Specifically: $\{A, B\}$, want to tell: $\{A \rightarrow B\}$ or $\{B \rightarrow A\}$.


**Who is cause, who is effect?**

Assumption: Correct causal structural choice leads to faster adaptation to shift distributions.

Main Idea: To use the speed of adaptation to a modified distribution as a meta-learning objective.

# Example

For bivariate model, both random variables are sampled from a categorical distribution:

$$A \sim Cate(\pi_A)$$
$$B|A = b \sim Cate(\pi_{B|a}),$$

where $\pi_A, \pi_{B|a}$ are all probability vector of size $N$.

Marginal

| P(A) | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ |
|---|---|---|---|---|---|---|

Conditiona

| $P(B\|a_1)$ | | | | | | |
|---|---|---|---|---|---|---|
| $\dots$ | | | | | | |

Marginal

| P(B) | $b_1$ | $b_2$ | $b_3$ | $b_4$ | $b_5$ | $b_6$ |
|---|---|---|---|---|---|---|

Condition

| $P(A\|b_1)$ | | | | | | |
|---|---|---|---|---|---|---|
| $\dots$ | | | | | | |

In test environments, $P(A)$ is changed.

Generate the test environments: Sample A and then sample B based on A.

| Learned Causal Structure | # of parameters | nonzero gradients |
|---|---|---|
| $P(A), P(B\|A)$ | $N^2 + N$ | $N$ |
| $P(B), P(A\|B)$ | $N^2 + N$ | $N^2 + N$ |

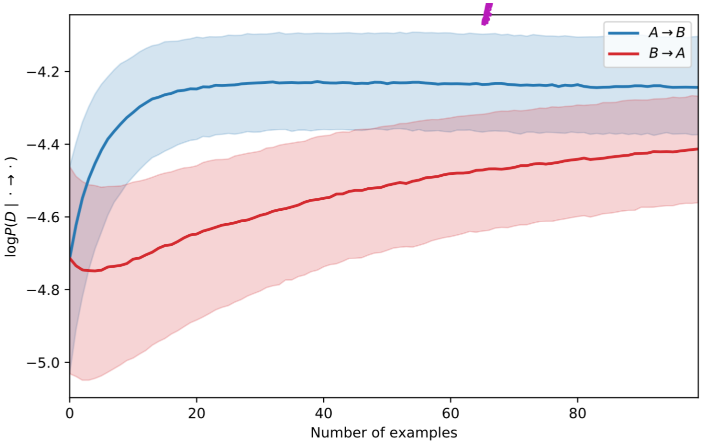
Required number of data to perform adaptation:

$$N_{data} = C_1 \cdot VC$$
$$VC = C_2 \cdot N_{param}$$

Why invariant mechanisms do not need to be relearned?

The gradient w.r.t the invariant module parameter is 0 if:

- correctly learned in the training phase
- have the correct set of causal parents, corresponding to the ground truth causal graph
- the corresponding ground truth conditional distributions is invariant from training distribution to the shifted distribution.




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

# Soft Parametrization

Impossible to enumerate all possible causal structures and compare adaptation speed.

In the simple setting $\{A, B\}$, the transfer objective as a log-likelihood over the mixture of the two explanations:

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

where $L_{A\to B}$ and $L_{B\to A}$ are the online likelihoods on the test data.

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

$$L_{B\to A} = \prod_{t=1}^{T} P_{B\to A}(a_t, b_t; \theta_t) \tag{2}$$

- The $\{(a_t, b_t)\}$ is the set of test examples at time $t$.
- $\theta_t$ are parameters as of time step $t$.
- $P(a, b; \theta)$ is the likelihood of example under model with parameter $\theta$.

It is a meta-learning framework and the inner loop fine-tunes the module parameters, the outer loop updates the structural parameters($\gamma$).

End up where?

**Theorem**

*SGD on* $E_{D_2}[\mathcal{R}]$ *with steps from* $\frac{\partial R}{\partial \gamma}$ *converges towards* $\sigma(\gamma) = 1$ *if* $E_{D_2}[\log L_{A \to} B] > E_{D_2}[\log L_{B \to} A]$ *or* $\sigma(\gamma) = 0$, *otherwise.*

# Meta-Learning

Meta-learning (also known as learning to learn) : Quick leaner.
Pervious learned form a rich base. Quick adaptation with a few data and iterations.

- Optimization based.
  - Learn a good initialization. (MAML: Model-agnostic meta-learning for fast adaptation of deep networks.)
  - Use another NN to update the parameter of the model.
- Model based.
- Distance based.

# MAML

Task distributions:

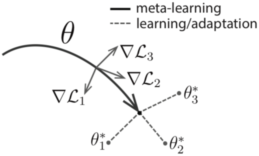
$$\mathcal{T} = \{L(x_1, a_1, \cdots, x_H, a_H), q(x_1), q(x_{t+1}|x_t, a_t), H\}.$$

For simplicity, for classification tasks:

$$\mathcal{T} = \{L(x, y), q(x, y)\}.$$

Suppose the model is $f_\theta$, for each sampled task $\tau_i^0$, the model parameter for each task can be updated via SGD:

$$\theta_i = \theta - \alpha \nabla_\theta L_{\tau_i^0}(f_\theta)$$




Above gradient descent only optimizes one particular task

To generalize: find a $\theta^*$ to guarantee efficient fine-tuning.

We sample new mini-batch from the training tasks, denoted as $\tau_i^1$.

$$\theta^* = \arg\min_{\theta} \sum_{\tau_i \sim p(\tau)} L_{\tau_i^1}(f_{\theta_i})$$

$$\theta = \theta - \beta \nabla_\theta \sum_{\tau_i \sim p(\tau)} L_{\tau_i^1}(f_{\theta - \alpha \nabla_\theta L_{\tau_i^0}(f_\theta)})$$

# Pseudo Code

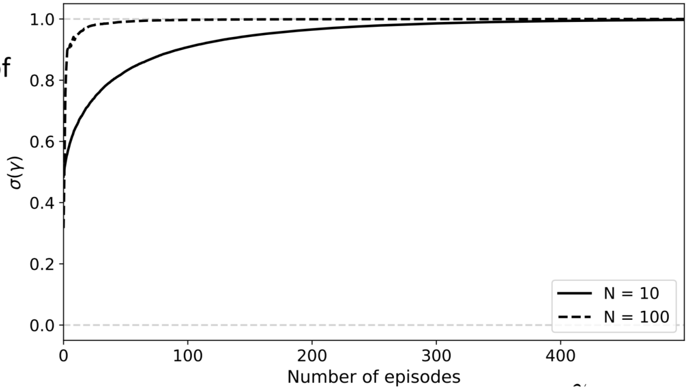
```
Draw initial meta-parameters of learner
Draw a training set from training distr.
Set causal structure to include all edges
Initialize learner parameters for this model
Pre-train the learner's parameters on the training set
Repeat J times
    Draw a transfer distr.
    Draw causal structure(s) according to meta-parameters
    Repeat T times
        Sample minibatch from transfer distribution
        Accumulate online log-likelihood of minibatch
        Update the model parameters accordingly

    Compute the meta-parameters gradient estimator
    Update the meta-parameters by SGD
    Optionally reset parameters to pre-training value
```

**Algorithm 1:** Meta-Transfer Learning of Causal Structure

Tabular parametrization of marginals and conditionals of bivariate model.

Correct causal graph can be recovered



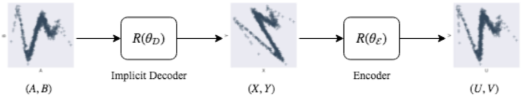
# Disentangling the causes

Realistic settings: no access to the true underlying causal variables $\{A, B\}$.

Example: Sensory-level data like pixels and sounds.

The previous assumption doesn't hold.

Method: Add a encoder to map the observations to hidden space where the assumption holds.




The encoder is trained such that the hidden space helps to optimize the meta-transfer objective described above.
So, encoder's parameters are also regarded as meta-parameters.

# Results

It can recover correct causal variables and recover correct casual direction, simultaneously.