An Analysis of the Adaptation Speed of Causal Models

Rémi Le Priol
Mila, Université de Montreal

Reza Babanezhad Harikandeh
SAIT AI Lab, Montreal

Yoshua Bengio
Mila, Université de Montreal
Canada CIFAR AI Chair

Simon Lacoste-Julien
Mila, Université de Montreal
Canada CIFAR AI Chair
SAIT AI Lab, Montreal

Abstract

Consider a collection of datasets generated by unknown interventions on an unknown structural causal model G. Recently, Bengio et al. (2020) conjectured that among all candidate models, G is the *fastest to adapt* from one dataset to another, along with promising experiments. Indeed, intuitively G has less mechanisms to adapt, but this justification is incomplete. Our contribution is a more thorough analysis of this hypothesis. We investigate the adaptation speed of cause-effect SCMs. Using convergence rates from stochastic optimization, we justify that a relevant proxy for adaptation speed is distance in parameter space after intervention. Applying this proxy to categorical and normal cause-effect models, we show two results. When the intervention is on the cause variable, the SCM with the correct causal direction is advantaged by a large factor. When the intervention is on the effect variable, we characterize the relative adaptation speed. Surprisingly, we find situations where the anticausal model is advantaged, falsifying the initial hypothesis.

1 INTRODUCTION

A learning agent interacting with its environment should be able to answer questions such as "what will happen to Y if I change X". Structural Causal Models (SCM) offer a formalism to answer this kind of questions (Pearl, 2009; Peters et al., 2017). The simplest SCM is the model $X \to Y$ where X is the cause and Y the effect. Modifying X will modify Y but modifying Y will not alter X. In general, SCMs model the distribution of observations with a directed graph where edges represent *independent mechanisms* (Janzing and Scholkopf, 2010).

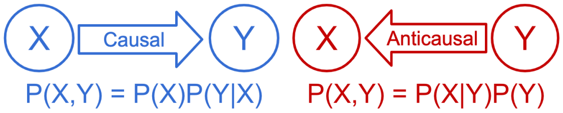


Figure 1: Two models of cause-effect data $X \to Y$.

Modern machine learning methods can fail surprisingly when the test distribution differ from the training distribution (Rosenfeld et al., 2018). A recent line of work describes these distribution shifts as interventions in an underlying causal model (Zhang et al., 2013; Magliacane et al., 2018). If this description is accurate, then an agent endowed with this hypothetical causal model could handle distribution shifts by updating the few mechanisms affected by the intervention. On contrary, an agent endowed with an incorrect model, would have to update many mechanisms. Bengio et al. (2020) infer that the causal agent will be the fastest to adapt to distribution shifts. Conversely, they use the speed of adaptation to unknown interventions as a criterion to learn the true causal model, showing promising empirical results on cause-effect models. Yet they lack a theoretical argument to connect interventions and fast adaptation. Thus we raise the question:

Do causal models adapt faster than non-causal models to distribution shifts induced by interventions?

Contributions. We theoretically and empirically answer this question for cause-effect SCMs with categor-

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

ical variables, and partially for multivariate normal distributions.

- For both settings, we use stochastic optimization convergence rates to show that the adaptation speed mostly depends on the distance in parameter space between the initialization (before intervention) and the optimum (after intervention).
- For categorical variables, we fully characterize this distance. We show that the causal model is faster by a large factor when the intervention is on the cause.
- When the intervention is on the effect, we surprisingly find settings where the anticausal model is systematically faster. As appealing as the fastest-to-adapt hypothesis may sound, it does not hold in every situations.

2 RELATED WORK

Causal relationships are asymmetric. These asymmetries are often visible in observations, so that one can identify which is cause and which is effect under relevant assumptions (Mooij et al., 2016). A common assumption is to constrain the set of functional dependencies between cause and effect. By contrast, in our work, we focus on two families of distributions which are notoriously unidentifiable from observational data: categorical and linear normal variables (Peters et al., 2017, Ch.4). With data coming from a generic directed acyclic graph (DAG), we can only hope to discover the Markov equivalence class of this DAG (Verma and Pearl, 1991). Many methods seek to achieve this goal, whether constraint-based such as the PC algorithm (Spirtes et al., 2000) or score-based methods using greedy search (Chickering, 2002) or more recently continuous optimization (Zheng et al., 2018; Lachapelle et al., 2020). However to discover the exact graph, we need access to interventional data.

Inferring causal links from interventions or experiments is the foundation of science. Inferring causal links from unknown interventions is a much harder and less principled problem. Tian and Pearl (2001) first studied this setting, proposing a constraint based method to infer the interventional equivalence class from a sequence of interventions. Then Eaton and Murphy (2007) proposed an exact Bayesian approach. More recently, Squires et al. (2019); Ke et al. (2019) proposed score based algorithms, improving in scalability and alleviating parametric assumptions. From a machine learning perspective, we are concerned with the predictive power that this structure will give us when faced with new data.

Distribution shifts are a common problem in machine learning, as well as in causal statistics (Zhang et al., 2013; Pearl and Bareinboim, 2014). Schölkopf et al. (2012) first brought up the idea of *invariance* to tackle this problem. Following up on this idea, Peters et al. (2016) designed an algorithm able to identify robust causal features from heterogeneous data. This work has set a fruitful line of research for robust machine learning (Heinze-Deml et al., 2018b,a; Rothenhäusler et al., 2019; Arjovsky et al., 2019). In a way, fast adaptation is the complementary idea of invariance: if most mechanisms are kept invariant, then only a few have to adapt. Schölkopf (2019) shed light on these approaches and the broader scope of causality research for machine learning.

3 BACKGROUND

In this section, we review the formalism of Bengio et al. (2020) on observations, interventions, models and adaptation.

Reference and Transfer Distributions. We assume perfect knowledge of a reference distribution $\boldsymbol{p}$ over the pair (X, Y) sampled from an SCM $X \rightarrow Y$. This distribution is the object of interventions, which results in new *transfer* distributions $\boldsymbol{p}^*$. If the *intervention is on the cause*, X is sampled from a different marginal, then Y is sampled from the reference conditional

$$\boldsymbol{p}^*(x, y) = \boldsymbol{p}^*(x)\boldsymbol{p}(y|x) \ . \tag{1}$$

If the *intervention is on the effect*, X is sampled from the reference marginal, then Y is sampled from another marginal independently of X

$$\boldsymbol{p}^*(x, y) = \boldsymbol{p}(x)\boldsymbol{p}^*(y) \ . \tag{2}$$

For each transfer distribution, we observe a few samples.

Models. We parametrize two generative models of (X, Y) (Fig. 1):

$$\boldsymbol{p}_{\theta_\rightarrow}(x, y) = \boldsymbol{p}_{\theta_X}(x)\boldsymbol{p}_{\theta_{Y|X}}(y|x) \quad \text{– } \textit{causal} \tag{3}$$

$$\boldsymbol{p}_{\theta_\leftarrow}(x, y) = \boldsymbol{p}_{\theta_Y}(y)\boldsymbol{p}_{\theta_{X|Y}}(x|y) \quad \text{– } \textit{anticausal}\,. \tag{4}$$

For each model, we call mechanisms the marginal and conditional models. Each mechanisms has its own set of parameters, e.g. θ_X and $\theta_{Y|X}$. In the following we will use θ to denote interchangeably $\theta_\rightarrow$ and $\theta_\leftarrow$.

Adaptation. Both models are initialized to fit perfectly the reference distribution $\boldsymbol{p}_{\theta^{(0)}_\rightarrow} = \boldsymbol{p}_{\theta^{(0)}_\leftarrow} = \boldsymbol{p}$. They observe fresh samples from $\boldsymbol{p}^*$ one by one and update their parameters $\theta_\rightarrow$ and $\theta_\leftarrow$ to maximize the

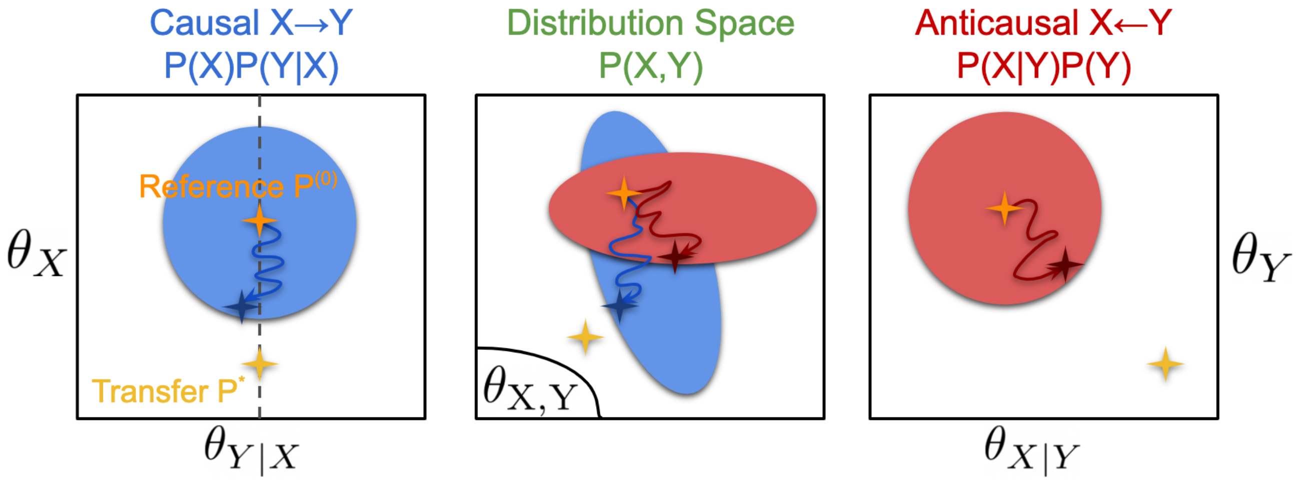


Figure 2: **Intuition behind fast adaptation.** An intervention on X turns the reference distribution $\boldsymbol{p}^{(0)}$ into a transfer distribution $\boldsymbol{p}^*$. The causal model (blue) only has to adapt θ_X, whereas the anticausal model (red) has to adapt both its mechanisms. After adaptation, the causal model ends up the closest from the transfer in terms of KL, as visible in the abstract distribution space. Blue and red balls represent the *proximity prior* induced by taking a few steps of SGD from the reference in each parameter space. Convergence rate analysis reveals that they are spherical functions of the parameter distance, but they get mapped to non-trivial shapes in distribution space – ellipses in this sketch.

log-likelihood with a step of stochastic gradient (SGD). Thanks to the separate parameters, the causal model log-likelihood loss decomposes as

$$\begin{aligned}\mathcal{L}_{\text{causal}}(\theta_\rightarrow) &= \mathbb{E}_{(X,Y)\sim \boldsymbol{p}^*}\left[-\log \boldsymbol{p}_{\theta_\rightarrow}(X,Y)\right] \\ =&\mathbb{E}_{\boldsymbol{p}^*}\left[-\log \boldsymbol{p}_{\theta_X}(X)\right] + \mathbb{E}_{\boldsymbol{p}^*}\left[-\log \boldsymbol{p}_{\theta_{Y|X}}(Y|X)\right] \quad (5)\end{aligned}$$

When $\boldsymbol{p}^*$ comes from an intervention, Bengio et al. (2020) observe that the causal model is often faster to adapt than the anticausal model. Intuitively, this is because the causal model has to adapt only the mechanism which was modified by the intervention. On the other hand, the anticausal model has to adapt both its mechanisms. In Figure 2, we compare these different scenarios and the concept of adaptation figuratively. While appealing, *this reasoning is not rigorous*, as sample complexity bounds of SGD typically do not depend on the number of parameters to update (Bubeck et al., 2015, Th. 6.2 & 6.3). In the next section, we formalize and understand this phenomenon in the light of convergence rates of stochastic optimization methods.

Distribution Families. We study two of the simplest sub-families of the exponential family (Wainwright and Jordan, 2008): categorical and linear normal variables. Their negative log-likelihood is a convex function of their natural parameter. These families are interesting because the direction is not identifiable from observational data (Peters et al., 2017, Ch.4) – e.g. $\boldsymbol{p}_{\theta_\rightarrow}$ and $\boldsymbol{p}_{\theta_\leftarrow}$ can model the same set of distributions – which makes them challenging for causal discovery.

AN OPTIMIZATION PERSPECTIVE One way to formalize adaptation speed is to characterize it via the convergence speed of the stochastic optimization procedure. An appealing aspect of stochastic optimization algorithms such as SGD (when only using fresh samples and running it on the true loss we care about) is that they come with convergence rate guarantees on the *population risk* in machine learning, thus giving us direct sample complexity results to obtain a specific generalization error. The convergence rate is an *upper bound* on the expected suboptimality after a given number of iterations. While these rates are about worst case performance and might also be loose, fortunately, for convex optimization, they tend to correspond well to actual empirical performance (Nesterov, 2004). We can thus use the convergence bounds as theoretical proxy for the convergence speed. In our experiments, we also verify empirically that the bounds correlate well with the observed convergence speed.

Here we provide a classical convergence rate on the expected suboptimality with Average Stochastic Gradient Descent (ASGD) under convexity and bounded gradient assumptions. We re-derive this rate in Appendix B.1 for completeness. This rate applies to log-likelihood maximization for categorical random variables (details in B.2). Since the target distribution is part of the model family, the log-likelihood suboptimality is equal to the KL-divergence – e.g. $\mathcal{L}(\theta) - \mathcal{L}(\theta^*) = D_{\mathrm{KL}}(\boldsymbol{p}^*||\boldsymbol{p}_\theta)$.

ASGD. Assume $\forall \theta, x, \|\nabla \log \boldsymbol{p}_\theta(x)\| \leq B$. After T iterations of SGD on (5),

$$\theta^{(t+1)} = \theta^{(t)} + \gamma \nabla \log \boldsymbol{p}_{\theta^{(t)}}(X_t, Y_t) \tag{6}$$

with learning rate $\gamma := \frac{c}{\sqrt{T}}$, starting from $\theta^{(0)}$, the average parameter's $\bar{\theta}^{(T)} = \frac{1}{T}\sum_{t=0}^{T-1}\theta^{(t)}$ suboptimality is upper bounded by

$$\mathbb{E}\left[D_{\mathrm{KL}}(\boldsymbol{p}^* \| \boldsymbol{p}_{\bar{\theta}^{(T)}})\right] \leq \frac{c^{-1}\|\theta^{(0)} - \theta^*\|^2 + cB^2}{2\sqrt{T}} \tag{7}$$

where the expectation is taken over the sampling of $T-1$ training points X_t, Y_t and θ^* is the closest solution to $\theta^{(0)}$ in the solution set $\operatorname{argmin}_\theta \mathcal{L}(\theta)$.

For categorical models, $B = 2$ (see B.2). Consequently, for a fixed T and with small enough c, the convergence upper bounds for causal and anticausal models differ mainly by $\delta := \|\theta^{(0)} - \theta^*\|^2$.

The bounded gradient assumption of (7) does not apply to the log-likelihood of normal variables. In Section 5.1, we provide an algorithm along with a convergence rate (22) that do apply to this case. Overall both bounds (7) and (22) carry the same message which can be summarized by:

The adaptation speed is dominated by the initial distance

$$\delta_{\text{causal}} = \left\|\theta_{\rightarrow}^{(0)} - \theta_{\rightarrow}^*\right\|^2 \tag{8}$$

$$\delta_{\text{anticausal}} = \left\|\theta_{\leftarrow}^{(0)} - \theta_{\leftarrow}^*\right\|^2 . \tag{9}$$

Other optimization methods. Yang et al. (2016, Theorem 1) provides a unified convergence rate for stochastic heavy ball and Nesterov methods that is similar to (7), where the initial distance is the main difference between causal and anticausal models. Consequently, our theoretical analysis holds for a larger class of algorithms than ASGD. More generally, it applies to any stochastic optimization method whose sample complexity depends on parameter distance.

4 CATEGORICAL VARIABLES

In this section, both cause and effect come from categorical distribution. We provide theoretical bounds on δ_{causal} and $\delta_{\text{anticausal}}$. We consider different scenarios to generate reference and transfer data and explain the consequences of each scenario.

4.1 Definitions

Cause X and effect Y are now two categorical variables taking values in $\{1, \dots, K\}$. Categorical variables are an exponential family with mean parameters $\boldsymbol{p} \in \boldsymbol{\Delta}_K$ the probability vector, and with natural parameter $\boldsymbol{s} \in \mathbb{R}^K$ – the logits or score parameters such that $p_z = \frac{e^{s_z}}{\sum_{z'} e^{s_{z'}}}$. The causal model has parameters $\boldsymbol{s}_X := (s_x)_{x=1\dots K}$ and $\boldsymbol{s}_{Y|X} := (s_{y|x})_{x,y=1\dots K}$. We gather the causal parameters in the variable $\theta_{\rightarrow} = (\boldsymbol{s}_X, \boldsymbol{s}_{Y|X})$ and the anticausal parameters in $\theta_{\leftarrow} = (\boldsymbol{s}_Y, \boldsymbol{s}_{X|Y})$ (Fig. 3). The loss (5) becomes

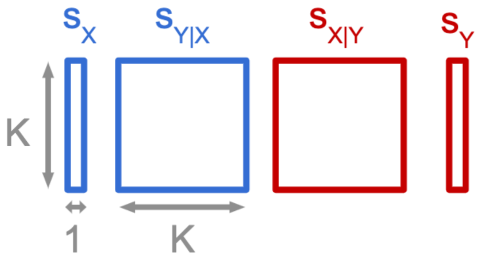


Figure 3: Parametrization of categorical models

$$\mathcal{L}_{\text{causal}}(\theta_{\rightarrow}) = \mathbb{E}_{(X,Y)\sim \boldsymbol{p}^*}\left[-\log \boldsymbol{p}_{\theta_{\rightarrow}}(X,Y)\right] \tag{10}$$

$$= \mathbb{E}_{\boldsymbol{p}^*}\left[-s_X + \log\sum_x e^{s_x} - s_{Y|X} + \log\sum_y e^{s_{y|X}}\right] .$$

Each mechanism's stochastic loss is the sum of a linear function and a softmax function. The softmax function is convex and 1-Lipschitz, so we can apply rate (7). To be self-contained, we include details in Appendix B.2.

4.2 Distance after Intervention

In this section, we prove that interventions on the cause advantage the causal model by a factor K, and we describe when interventions on the effect will advantage one model over another.

Intervention on cause X, $\boldsymbol{s}_X \leftarrow \boldsymbol{s}_X^*$. The causal conditional $\boldsymbol{s}_{Y|X}$ is left unchanged, but the effect marginal $\boldsymbol{s}_Y$ is modified in a non-trivial way. Consequently the initial distances are

$$\delta_{\text{causal}} = \|\boldsymbol{s}_X - \boldsymbol{s}_X^*\|^2 \tag{11}$$

$$\delta_{\text{anticausal}} = \|\boldsymbol{s}_Y - \boldsymbol{s}_Y^*\|^2 + \sum_y \|\boldsymbol{s}_{X|y} - \boldsymbol{s}_{X|y}^*\|^2 . \tag{12}$$

The causal model has to update K parameters, whereas the anticausal model has to adapt $K^2 + K$ parameters. Therefore the causal model seems to be advantaged by a factor K. The following proposition – proved in Appendix C – shows that this is reflected by ℓ_2 distances.

Proposition 1. *When the intervention happens on the cause,*

$$\delta_{anticausal} \geq K \delta_{causal} . \tag{13}$$

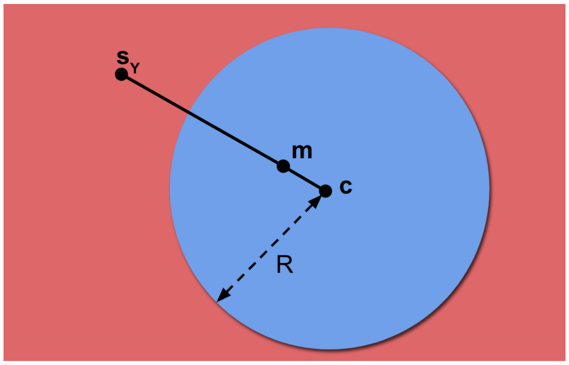


Figure 4: **Illustration of Proposition 2.** $\boldsymbol{c}$ is on the line joining $\boldsymbol{s}_Y$ and $\boldsymbol{m} := \frac{1}{K}\sum_x \boldsymbol{s}_{Y|x}$. When $\boldsymbol{s}_Y^*$ is within the blue ball of radius R centered at $\boldsymbol{c}$, $\Delta \leq 0$ and the causal model is advantaged, otherwise the anticausal model is advantaged (red area). This is a surprising counter-example to the adaptation-speed hypothesis.

Intervention on effect Y, $\forall x, \boldsymbol{s}_{Y|x} \leftarrow \boldsymbol{s}_Y^*$. Cause and effect become independent. The causal model is advantaged only if the intervention $\boldsymbol{s}_Y^*$ is close enough from the previous marginal, as formalized by the following proposition:

Proposition 2. *When the intervention happens on the effect*

$$\begin{aligned}\Delta &:= \delta_{causal} - \delta_{anticausal} \\ &= (K-1)\left(\left\|\boldsymbol{s}_Y^* - \boldsymbol{c}\right\|^2 - R^2\right)\end{aligned} \tag{14}$$

where $R^2 \approx K\widehat{\mathrm{Var}}_X[\log\sum_y e^{s_{y|X}}]$ *and* $\boldsymbol{c} = \frac{(\sum_x \boldsymbol{s}_{Y|x}) - \boldsymbol{s}_Y}{K-1}$.

See Figure 4 for an illustration and Appendix C.3 for the exact formula of R and the proof. When the intervention $\boldsymbol{s}_Y^*$ is close enough to $\boldsymbol{c}$, which depends on the reference, the causal model is advantaged. If $\boldsymbol{s}_Y^*$ is far from $\boldsymbol{c}$ or if R is small then the anticausal model is likely to be advantaged.

4.3 Simulating Reference Distributions

To evaluate the fast adaptation criterion, we are going to work on synthetic data, which raises the question : from which distribution should we sample $\boldsymbol{p} = \boldsymbol{p}_{\theta^{(0)}}$? We call this distribution *prior*. Following the independent mechanism assumption, the marginal on the cause $\boldsymbol{p}_X$ and the conditional of effect given cause $\boldsymbol{p}_{Y|X}$ should not contain any information about each other.

Dense Prior. To sample causal mechanisms, a natural choice is

$$\boldsymbol{p}_X \sim \mathrm{Dir}(\mathbf{1}_K) \quad \text{and} \quad \forall x, \boldsymbol{p}_{Y|x} \sim \mathrm{Dir}(\mathbf{1}_K) \tag{15}$$

where Dir is the Dirichlet distribution and $\mathbf{1}_K$ is the all-one vector of dimension K. $\mathrm{Dir}(\mathbf{1}_K)$ the uniform law over the simplex $\boldsymbol{\Delta}_K$. This prior leads to the K2 score from the Bayesian network literature (Cooper and Herskovits, 1991). We call this choice the *dense prior* by opposition to the sparse prior introduced next. This is the choice made in Bengio et al. (2020), as well as Chalupka et al. (2016). The latter work reports that distributions sampled from this prior exhibit some asymmetry between X and Y. In Appendix D.1, we complement their work, explaining how the effect marginal is likely to be closer from the uniform distribution than the cause marginal. This asymmetry means that *the causal direction is identifiable from observational data.*

Sparse Prior. To fix this issue, we study an alternative prior that is symmetric and ensures that both cause and effect marginals are sampled from a uniform prior over $\boldsymbol{\Delta}_K$. We sample the causal mechanisms as follows

$$\boldsymbol{p}_X \sim \mathrm{Dir}(\mathbf{1}_K) \quad \text{and} \quad \forall x, \boldsymbol{p}_{Y|x} \sim \mathrm{Dir}(\mathbf{1}_K/K)\,. \tag{16}$$

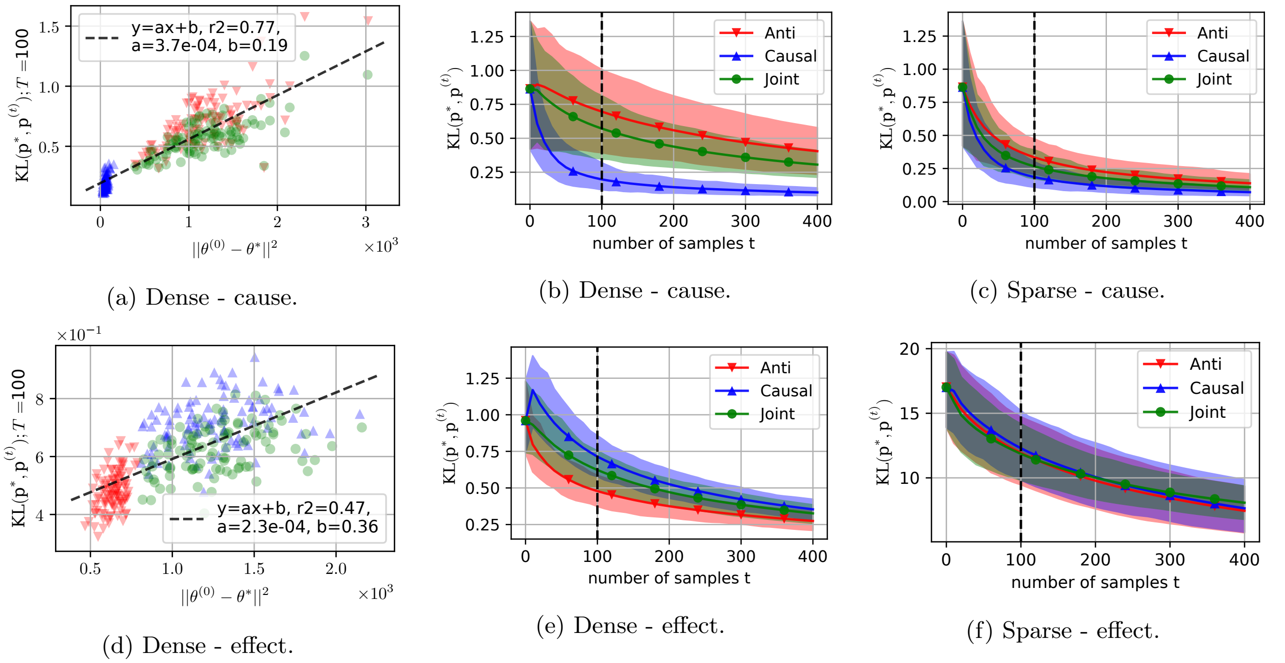
The $\mathbf{1}_K/K$ parameter means that samples will be approximately sparse, hence the name. We show in Appendix D.2 that with this sampling scheme, the joint is sampled from a sparse Dirichlet over $\boldsymbol{\Delta}_{K^2}$: $\boldsymbol{p}_{(X,Y)} \sim \mathrm{Dir}(\mathbf{1}_{K^2}/K)$. This in turns means that we can switch the roles of X and Y in (16). The effect marginal has uniform density over the simplex. In general, *the causal direction is not identifiable from observational data.* In Bayesian Networks literature, this is known as the Bayesian Dirichlet equivalent uniform prior (Heckerman et al., 1995).

4.4 Categorical Variables Experiments

Goal. As discussed in Section 4.3, the prior over the joint distribution on (X, Y) is going to influence the behavior of ASGD. We are seeking answers to two questions:

1. Is the adaptation speed positively correlated with the initial distance, as suggested by the upper bound (7) on the convergence rate of ASGD?
2. Is there a clear difference in adaptation speed between causal and anticausal models?

Data. We consider categorical variables with $K = 20$. For each initialization method, we sample 100 different reference joint distributions. For each of these distributions, we sample an intervention by sampling a probability vector $\boldsymbol{q}$ *uniformly* from $\boldsymbol{\Delta}_K$. If the intervention is on the cause, we plug $\boldsymbol{q}$ instead of $\boldsymbol{p}_X$. If the intervention is on the effect, we redefine $\boldsymbol{p}_{Y|x} = \boldsymbol{q}, \forall x$.



(a) Dense - cause.

(b) Dense - cause.

(c) Sparse - cause.

(d) Dense - effect.

(e) Dense - effect.

(f) Sparse - effect.

Figure 5: **Experimental results on categorical data.** Each plot is captioned with the prior and the intervention considered. **Scatter plots** are showing the positive correlation between the KL after 100 steps of SGD and the initial parameter distance. Each point represent one of 100 synthetic pairs $(\boldsymbol{p}^{(0)}, \boldsymbol{p}^*)$. **Training curves** show the average KL (solid line) and the (5,95) percentiles (shaded) over 100 runs. Remark how all models start from the same initial KL, but they converge at different speeds.

Models. We are comparing causal and anticausal models adaptation speed. We also report results for a model of the joint $\boldsymbol{p}_{X,Y} = \text{softargmax}(\boldsymbol{s}_{X,Y})$ as a reference model. We expect its results to be in between the performance of the causal and anticausal model as it expresses no prior over the direction. We optimize all models with Averaged SGD. In each iteration of SGD we get one fresh sample from the transfer distribution. For each model and each setting, we tune the (constant) learning rate so as to optimize the likelihood after seeing $\frac{K^2}{4} = 100$ samples, to explore the few samples regime. We present results in Figure 5

Dense prior. When the intervention is on the cause, the causal model is much closer from its optimum: in Fig. 5a the blue cluster is on the left of the scatter plots. This is well correlated with faster adaptation (Fig. 5b). On the contrary, *when the intervention is on the effect, the anticausal model starts closer from its optimum* and it converges faster (Fig. 5d, 5e). We can interpret this result in light of Proposition 2. In Appendix C.3, we explain why the radius R is small under the dense prior. As a result, $\boldsymbol{s}_Y^*$ is mostly sampled outside of the ball of radius R, consequently the anticausal model is advantaged. Overall, there is a wider gap between models in Fig. 5b than in Fig. 5e. Consequently, if we take a balanced average of a few interventions on the cause and a few interventions on the effect, the causal model remains faster (details in Appendix C.4).

Sparse prior. When the intervention is on the cause, the causal model has a slight advantage (Fig. 5c). When the intervention is on the effect, no model has a set advantage (Fig. 5f), but the sparsity induces much higher KL values, as explained in Appendix D.3. This KL explosion drowns the signal coming from the cause intervention, calling for further algorithmic developments – such as inferring the intervention, as explored by Ke et al. (2019).

5 MULTIVARIATE NORMAL VARIABLES

In this section, we analyze the case of two multivariate normal variables with a linear relationship. Cause X and effect Y are sampled from the causal model

$$X \sim \mathcal{N}(\mu_X, \Sigma_X) \tag{17}$$
$$Y|X \sim \mathcal{N}(\boldsymbol{A}X + \boldsymbol{a}, \Sigma_{Y|X}) \tag{18}$$

with mean parameters $\mu_X, \boldsymbol{a} \in \mathbb{R}^K$ and $\Sigma_X, \boldsymbol{A}, \Sigma_{Y|X} \in \mathbb{R}^{K\times K}$. This parametrization is the most intuitive but it is unfortunately not appropriate to get convergence rates. We are going to introduce another parametrization along with an algorithm and a convergence rate (Sec. 5.1), before providing empirical results (Sec. 5.2).

5.1 Optimization Analysis

The negative log-likelihood of model (17) is notoriously non-convex. This is problematic for convergence results. For simplicity, we focus in this section on the simple marginal mechanism with mean parameters μ, Σ. We detail the full model in Appendix F. If we use the natural parameters $\eta = \Sigma^{-1}\mu$ and $\Lambda = \Sigma^{-1}$ (precision matrix), the negative log-likelihood is convex

$$\begin{aligned}&\mathbb{E}\left[-\log \boldsymbol{p}_{(\eta,\Lambda)}(X)\right] \\ &= \frac{1}{2}\Big(\mathbb{E}\left[\mathrm{Tr}(XX^\top\Lambda) - 2X^\top\eta\right] + \eta^\top\Lambda^{-1}\eta - \log|\Lambda|\Big) \,.\end{aligned} \tag{19}$$

This objective is composed of a pleasant stochastic linear term, and a difficult deterministic barrier objective which goes to infinity when $\Lambda \to 0$. This barrier is composed of a matrix inverse and a log determinant. The assumptions of Lipschitz or gradient-Lipschitz required to get SGD convergence do not hold for the barrier. While the empirical version of (19) has a close formed formula for its global minimum, quite surprisingly, gradient-based optimization of the normal likelihood is difficult to analyze. Convex optimization typically deals with non-smooth terms by introducing proximal operators (Parikh et al., 2014). However this barrier term is too complex to get an analytic formula for the proximal operator. We transform it into a more convenient form by introducing $\boldsymbol{L}$, the lower triangular Cholesky factor of the precision matrix $\Lambda = \boldsymbol{L}\boldsymbol{L}^T$, and $\zeta = \boldsymbol{L}^{-1}\eta = \boldsymbol{L}^\top\mu$. Then (19) simplifies into

$$\begin{aligned}&\mathbb{E}\left[-\log \boldsymbol{p}_{(\zeta,\boldsymbol{L})}(X)\right] \\ &= \frac{1}{2}\mathbb{E}\left[\left\|\boldsymbol{L}^\top X - \zeta\right\|^2\right] - \sum_i \log \boldsymbol{L}_{i,i} \,.\end{aligned} \tag{20}$$

We will refer to $(\zeta, \boldsymbol{L})$ as *Cholesky parameters*. This objective is more suitable to gradient based optimization with a simple proximal operator, as detailed in the next section. We provide all details about the causal model in Appendix F.

Stochastic Proximal Gradient Algorithm We want to minimize the sum of a stochastic convex smooth function $f_X(\theta) := \frac{1}{2}\left\|\boldsymbol{L}^\top X - \zeta\right\|^2$ and convex non-smooth regularizer $g(\theta) = -\sum_i \log \boldsymbol{L}_{i,i}$. This is exactly the goal of the stochastic proximal gradient (Duchi et al., 2010) update

$$\theta_{t+1} = \operatorname*{argmin}_\theta g(\theta) + \frac{1}{2\gamma_t}\|\theta_t - \gamma_t \nabla f_{X_t}(\theta_t) - \theta\|^2 \tag{21}$$

where γ_t is the step-size and X_t is randomly sampled. For objective (20), the proximal gradient update has a closed form solution that amounts to updating all parameters with the stochastic gradient of the quadratic term, then updating the diagonal elements of $\boldsymbol{L}$ with the mapping $x \mapsto \frac{1}{2}(x + \sqrt{x^2 + 4\gamma})$, thus ensuring that they remain strictly positive (details in Appendix E.2).

Convergence Rate. We assume that stochastic gradients are almost-surely B-Lipschitz. B is known as the smoothness constant. We show in Appendix E.1 that running the stochastic proximal gradient algorithm with step size $\gamma_t = \frac{\gamma}{3B\sqrt{T}}$ where $\gamma \leq 1$, for T iterations guarantees

$$\begin{aligned}&\mathbb{E}\left[D_{\mathrm{KL}}(\boldsymbol{p}^*||\boldsymbol{p}_{\bar{\theta}^{(T)}})\right] \\ &\qquad\leq \frac{3B\|\theta^{(0)} - \theta^*\|^2}{\gamma\sqrt{T}} + \frac{D_{\mathrm{KL}}(p^*||p_{\theta^{(0)}})}{T} \,.\end{aligned} \tag{22}$$

Analysis. The term ${}^{KL(p^*||p_{\theta^{(0)}})}/_{T}$ is equal for causal and anticausal models because we assume $\boldsymbol{p}_{\theta_\rightarrow}^{(0)} = \boldsymbol{p}_{\theta_\leftarrow}^{(0)}$. For normal variables, B depends only on the data and is a priori equal for both models (Appendix E.3). Similarly to (7), both models' rates differ mainly by $\delta = \|\theta^{(0)} - \theta^*\|^2$.

When the intervention is on the cause, we prove in Appendix F.5 that the anticausal model is farther away from its optimum in the natural parametrization

$$\delta^{\text{natural}}_{\text{anticausal}} \geq \delta^{\text{natural}}_{\text{causal}} \,. \tag{23}$$

Unfortunately, in the Cholesky parametrization (Fig. 6, 2nd column), or when the intervention is on the effect (Fig. 6, bottom row),we observe empirically that there is no such hard guarantee, although the causal distance tends to be smaller than the anticausal distance.

5.2 Experiments

Similarly to categorical variables, we need to decide on a prior over reference and transfer distributions. This choice is informed by two criteria. First the independent mechanism principle which states that we should sample θ_X independently of $\theta_{Y|X}$. Second we want θ_Y to have approximately the same distribution as θ_X – e.g. we want the distribution to be approximately symmetric so that we cannot identify the direction from observational data. These considerations lead us to a flavor of normal-Wishart prior (Geiger et al., 2002) described in Appendix G.

We sample 100 random joint distributions from this prior, and for each distribution we sample a random intervention on the cause, and a random intervention on the effect. We then run the stochastic proximal gradient on objective (20). We report results in Figure 6. Similarly to the categorical case, when the intervention

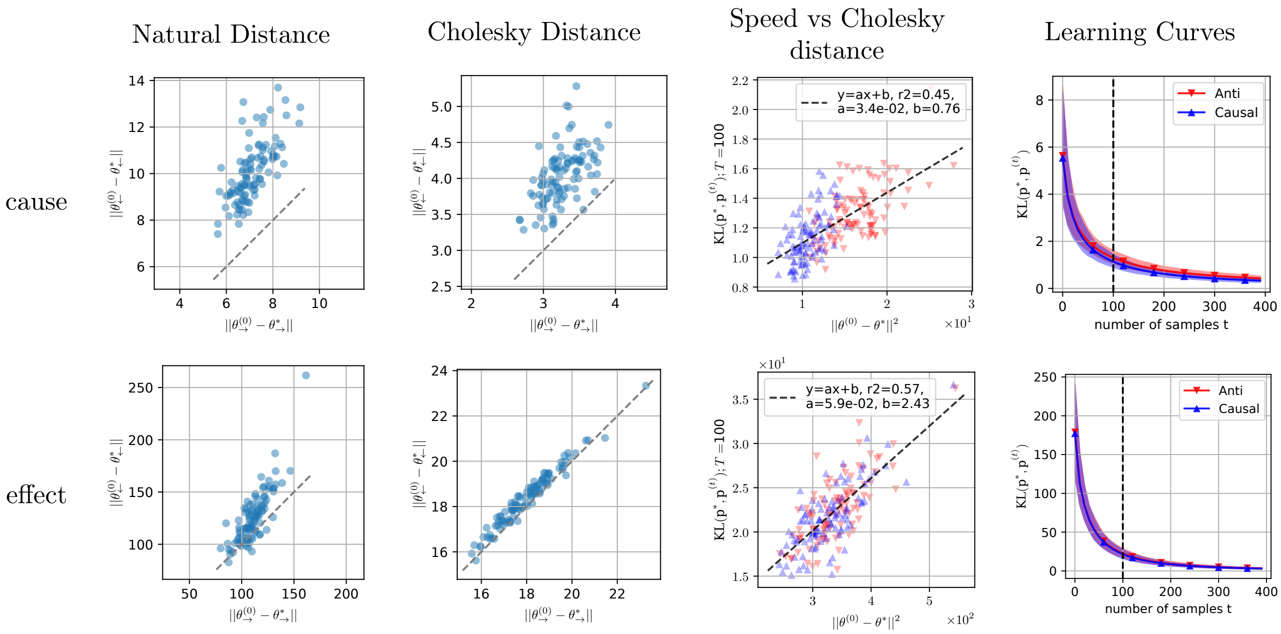


Figure 6: **Multivariate Normal Variables with dimension** $K = 10$. Row 1 and 2 correspond to interventions on cause and effect respectively. *Column 1 & 2:* scatter plot $\delta_{\text{anticausal}}$ vs δ_{causal} respectively in natural and Cholesky parametrization. The grey diagonal is the identity line. We observe a natural tendency for $\delta_{\text{anticausal}} > \delta_{\text{causal}}$ (points above the grey diagonal), but this is systematically true only for the natural distance when the intervention is on the cause. *Column 3 & 4:* same plot as in Figure 5. Once again we observe a correlation between initial distance and optimization speed. When the intervention is on the cause, the causal model is advantaged. When the intervention is on the effect, both curves overlap.

is on the cause, the causal model is advantaged by a slight margin (upper right figure). When the intervention is on the effect both models are learning at the same speed (bottom right figure).

Conclusion

We provided a first theoretical analysis of the adaptation speed in two-variables cause-effect SCMs under localized interventions for categorical and normal data. Convergence guarantees for stochastic optimization on the true population log-likelihood indicates that the adaptation speed is related to the distance between initial point and optimum in parameter space. We verified this correlation empirically. We proved analytically that this distance is lower for the causal model than for the anticausal model when the intervention is on the cause variable. This explains a surprising phenomenon: while both models start with the same suboptimality, one learns faster than the other. When the intervention is on the effect variable, we highlighted examples showing that either model can be advantaged. This observation challenges the intuition that the causal model should be the fastest to adapt, and it raises new questions for the approach of Bengio et al. (2020), such as: are there practical situations where the fastest-to-adapt heuristic is useful ? On a more theoretical note, is it possible to characterize the adaptation speed behavior for more general families of distributions?

Acknowledgments

We thank Sébastien Lachapelle and Tristan Deleu for insightful discussions and useful feedback that led to the main results of this paper. We thank Waïss Azizian for his help in debugging proofs. This work was partially supported by the Canada CIFAR AI Chair Program. Simon Lacoste-Julien is a CIFAR Associate Fellow and Yoshua Bengio is a Program Co-Director in the Learning in Machines & Brains program.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Yves F Atchadé, Gersende Fort, and Eric Moulines. On perturbed proximal gradient algorithms. *Journal of Machine Learning Research*, 18(1):310–342, 2017.

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. In *ICLR*, 2020.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 2015.

Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Estimating causal direction and confounding of two discrete variables. *arXiv preprint arXiv:1611.01504*, 2016.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Gregory F Cooper and Edward Herskovits. A bayesian method for constructing bayesian belief networks from databases. In *Uncertainty Proceedings 1991*. Elsevier, 1991.

John Duchi and Yoram Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10(Dec):2899–2934, 2009.

John C Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite objective mirror descent. In *COLT*, pages 14–26, 2010.

Daniel Eaton and Kevin Murphy. Exact bayesian structure learning from uncertain interventions. In *Artificial intelligence and statistics*, pages 107–114, 2007.

Dan Geiger, David Heckerman, et al. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics*, 30(5):1412–1440, 2002.

David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

Christina Heinze-Deml, Marloes H Maathuis, and Nicolai Meinshausen. Causal structure learning. *Annual Review of Statistics and Its Application*, 2018a.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 2018b.

Dominik Janzing and Bernhard Scholkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10), 2010.

Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural DAG learning. In *ICLR*, 2020.

Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *NeurIPS*, 2018.

Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 2016.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.

Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, 2014.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

Dominik Rothenhäusler, Peter Bühlmann, and Nicolai Meinshausen. Causal dantzig: fast inference in linear structural equation models with hidden variables under additive interventions. *The Annals of Statistics*, 47(3):1688–1722, 2019.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *ICML*, 2012.

Bernhard Schölkopf. Causality for machine learning. *arXiv:1911.10500*, 2019.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-based causal structure learning with unknown intervention targets. *arXiv preprint arXiv:1910.09007*, 2019.

Jin Tian and Judea Pearl. Causal discovery from changes. In *UAI*, 2001.

Thomas Verma and Judea Pearl. *Equivalence and synthesis of causal models*. UCLA, Computer Science Department, 1991.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

Tianbao Yang, Qihang Lin, and Zhe Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. *arXiv preprint arXiv:1604.03257*, 2016.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *ICML*, 2013.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *NeurIPS*, 2018.