Toward Causal Learning
Bernhard Schölkopf
Human-level object recognition?

from Perona, 2017;
cf. Lopez-Paz et al., 2016
Machine learning uses correlations rather than causality
Adversarial Vulnerability

“pig”

+ 0.005 x

= “airliner”

Image credit: http://people.csail.mit.edu/madry/lab/blog/adversarial/2018/07/06/adversarial_intro/

Reichenbach’s Common Cause Principle

(i) if $X$ and $Y$ are dependent, then there exists $Z$ causally influencing both;

(ii) $Z$ screens $X$ and $Y$ from each other (given $Z$, $X$ und $Y$ become independent)

\[
\sum_z p(x|z)p(y|z)p(z) \quad p(x)p(y|x) \quad p(x|y)p(y)
\]
Correlation by conditioning on common effects

Berkson’s paradox (1946)
Example: $X, Y, Z$ binary

$X \perp Y$ but $X \not\perp Y \mid Z$

- assumption 1: among the papers submitted to ICML, there is no correlation between having strong theory (X) and good experiments (Y)
- assumption 2: a paper gets in iff it has strong theory or good experiments
- among the accepted papers, there is a negative correlation between strong theory and good experiments
Functional causal models (Pearl, Spirtes, ..)

- Set of observables $X_1, \ldots, X_n$ on a DAG $G$
- Arrows represent direct causal links
- $X_i := f_i(\text{PA}_i, U_i)$ with independent RVs $U_1, \ldots, U_n$.

- Entails observational distribution $p(X_1, \ldots, X_n)$ satisfying the causal Markov condition:

  Conditioned on its parents, $X_j$ is independent of its non-descendants.

Note: $(G, p)$ are a “graphical model” (Pearl, Lauritzen, et al.)

- Interventions are modelled by changing functions; this entails an interventional distribution (analogous for other changes)
**Functional Model and Markov conditions**  
(*Lauritzen 1996, Pearl 2000*)

**Theorem:** the following are equivalent:

- Existence of a functional causal model
- Local Causal Markov condition: $X_j$ statistically independent of non-descendants, given parents (i.e.: every information exchange with its non-descendants involves its parents)
- Global Causal Markov condition: d-separation (characterizes the set of independences implied by local Markov condition)
- Factorization $p(X_1, \ldots, X_n) = \prod_j p(X_j | \text{Parents}_j)$ (conditionals as causal mechanisms generating statistical dependence)

(subject to technical conditions)
Inferring a Causal Model

**Question:** How can we recover $G$ from $p$ (without interventions)?

**Answer:** Can infer a class containing the correct $G$ by conditional independence testing.*

* subject to “faithfulness” *(Spirtes, Glymour, Scheines 2001)*

**Idea:** Independent noises pick up conditional dependences according to the graph topology.
This is independent of the $f_i$. (!)

**Problem:**
- if the $f_i$ are complex, conditional independence testing is hard
- for *two variables* only, there are no conditional independences

**Idea:**
the functions also leave a footprint
Friedrich Nietzsche's
TWILIGHT OF THE IDOLS
or How to Philosophize with a Hammer

The Four Great Errors
What is cause and what is effect?

\[ p(a,t) = p(a|t) \, p(t) \quad T \rightarrow A \]
\[ = p(t|a) \, p(a) \quad A \rightarrow T \]
• **intervention** on $a$: raise the city, find that $t$ changes

• hypothetical intervention on $a$: still expect that $t$ changes, since we can think of a physical mechanism $p(t|a)$ that is **independent** of $p(a)$

• we expect that $p(t|a)$ is **invariant** across, say, different countries in a similar climate zone
The causal generative process is composed of autonomous modules that do not inform or influence each other.
• a “structural” relation not only explains the observed data, it captures a structure connecting the variables; related to autonomy and invariance (Haavelmo 1943, Frisch 1948, ...)

• an equation system becomes structural by virtue of invariance to a domain of modifications (Harwich, 1962)

• “Simon’s invariance criterion:” the true causal order is the one that is invariant under the right sort of intervention (Simon, 1953; Hoover, 2008)

• each parent-child relationship in the network represents a stable and autonomous physical mechanism (Pearl, 2009)

• formalised using algorithmic information theory (Janzing & Schölkopf, 2010)
• Factorization

\[ p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid PA_i) \]

according to the causal graph, with independent causal conditionals

• a change in a real world distribution always comes from a change in a causal conditional / mechanism (i.e., structural assignment/function and/or noise variable)

• by independence: changing one mechanism \( p(X_i \mid PA_i) \) does not change the other \( p(X_j \mid PA_j) \) \((j \neq i)\)

• such changes hence act in a sparse/local way upon the causal factorization

• in non-causal factorizations, changes will not be local

Cf. independence of mechanisms (Janzing & Schölkopf, 2010), independence of cause and mechanism (Janzing et al., 2012), autonomy, (structural) invariance, separability, exogeneity, stability, modularity (Aldrich, 1989; Pearl, 2009)
Independence of input and mechanism, III

- No noise on effect variable
- Assumption: \( y = f(x) \) with invertible \( f \)

\[ \begin{array}{c}
\text{Danuisis, Janzing, Mooij, Zscheischler, Steudel, Zhang, Schölkopf:}
\text{Inferring deterministic causal relations, } UAI \\
\text{2010}
\end{array} \]
Causal independence implies anticausal dependence

Assume that $f$ is a monotonically increasing bijection of $[0, 1]$. View $p_x$ and $\log f'$ as RVs on the prob. space $[0, 1]$ w. Lebesgue measure.

Postulate (independence of mechanism and input):

$$\text{Cov} \left( \log f', p_x \right) = 0$$

Note: this is equivalent to

$$\int_0^1 \log f'(x)p(x)dx = \int_0^1 \log f'(x)dx,$$

since $\text{Cov} \left( \log f', p_x \right) = E \left[ \log f' \cdot p_x \right] - E \left[ \log f' \right] E \left[ p_x \right] = E \left[ \log f' \cdot p_x \right] - E \left[ \log f' \right]$.

Proposition: If $f \neq Id$,

$$\text{Cov} \left( \log f^{-1}', p_y \right) > 0.$$
\(u_x, u_y\) uniform densities for \(x, y\)
\(v_x, v_y\) densities for \(x, y\) induced by transforming \(u_y, u_x\) via \(f^{-1}\) and \(f\)

Equivalent formulations of the postulate:

**Additivity of Entropy:**
\[
S(p_y) - S(p_x) = S(v_y) - S(u_x)
\]

**Orthogonality (information geometric):**
\[
D(p_x \parallel v_x) = D(p_x \parallel u_x) + D(u_x \parallel v_x)
\]
which can be rewritten as
\[
D(p_y \parallel u_y) = D(p_x \parallel u_x) + D(v_y \parallel u_y)
\]

**Interpretation:**
irregularity of \(p_y\) = irregularity of \(p_x\) + irregularity introduced by \(f\)
Algorithmic structural causal model

- for every node $x_j$ there exists a program $u_j$ that computes $x_j$ from its parents $pa_j$

- all $u_j$ are jointly independent

- the program $u_j$ represents the causal mechanism that generates the effect from its causes

- $u_j$ are the analog of the unobserved noise terms in the statistical functional model

Theorem: this model implies the causal Markov condition (replacing Shannon entropy with Kolmogorov complexity).

Gedankenexperiment

Particles scattered at an object

- incoming beam: ‘cause’
- scattering at object: ‘mechanism’
- outgoing beam: ‘effect’, contains information about the object
Independence assumption

- $s$ initial state of a physical system
- $M$ the system dynamics applied for some fixed time

Independence Principle: $s$ and $M$ are algorithmically independent

$$I(s : M) \equiv 0,$$

i.e., knowing $s$ does not enable a shorter description of $M$ and vice versa.
Thermodynamic Arrow of Time

Theorem [non-decrease of entropy]. Let $M$ be a bijective map on the set of states of a system then $I(s : M) \overset{+}{=} 0$ implies

$$K(M(s)) \overset{+}{=} K(s)$$

Proof idea: If $M(s)$ admits a shorter description than $s$, knowing $M$ admits a shorter description of $s$: just describe $M(s)$ and then apply $M^{-1}$.

Cause-Effect Inference with Additive Noise

Forward model: \( y := f(x) + u \) with \( x \perp \!\!\!\perp u \)

Identifiability: when is there a backward model of the same form?

**Theorem.** For an additive model to hold true in both directions, the distributions of \( U, X \) and the function \( f \) need to be matched to each other.*

* Example: \( U, X \) Gaussian and \( f \) linear.

Hoyer et al.: Nonlinear causal discovery with additive noise models. NIPS 21, 2009
Pickup et al.: Seeing the Arrow of Time. CVPR 2014
Benchmark dataset with 106 cause-effect pairs

http://webdav.tuebingen.mpg.de/cause-effect/

Mooij et al: Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks, JMLR 2016
IGCI: Information Geometric Method
AN: Additive Noise Model
(nonlinear)
LINGAM: Shimizu et al., 2006
PNL: AN with post-nonlinearity
GPI: Mooij et al., 2010

Learning cause-effect discrimination, LT results
(Lopez-Paz et al., ICML 2015)

Used the same methods to classify the direction of time:

time series (Peters et al., ICML 2009, Bauer et al., ICML 2016)

videos (Pickup et al., CVPR 2014)
Bernhard Schölkopf

Exoplanet Transits

• Earth: annual 84 ppm signal for ½ day, visible from 0.5% of all directions
• Many planets found, but nothing quite like earth/sun
• Both spacecraft and stars vary, leading to changes that are sometimes much bigger than the signal
Bernhard Schölkopf


ICML 2015
Astrophysical Journal 2015
PNAS 2016

\[ Q - E[Q] = Y - E[Y|X] \]

14 new exoplanets
The Milkman Problem
Half-Sibling Regression

\[ X \perp Q \]
\[ X \text{ and } Y \text{ share information (only) through } N \]

If we try to predict \( Y \) from \( X \), we only pick up the part due to \( N \).

Idea: remove \( E[Y|X] \) from \( Y \) to reconstruct \( Q \).

with David Hogg, Dan Foreman-Mackey, Dun Wang, Dominik Janzing, Jonas Peters, Carl-Johann Simon-Gabriel (ICML 2015)
**Proposition.** $Q, N, Y, X$ random variables, $X \independent Q$, and $f$ measurable. Suppose

- $Y = Q + f(N)$ *(additive noise model)*
- $f(N) = \psi(X)$ for some $\psi$ *(complete information)*.

Then $\hat{Q} := Y - \mathbb{E}[Y|X] = Q - \mathbb{E}[Q]$.

$Q$ can be reconstructed, up to a constant offset, from $Y$ and $\mathbb{E}[Y|X]$. 

*Bernhard Schölkopf*
**Proposition.** $Q, N, Y, X$ random variables, $X \perp Q$, and $f$ measurable. Suppose

- $Y = Q + f(N)$ (additive noise model)

Then $E[(\hat{Q} - (Q - E[Q]))^2] = E[Var[f(N)|X]]$.

*If $f(N)$ can (in principle) be predicted well from $X$, then $Q$ can be reconstructed well.*
14 confirmed exoplanets
Using cause-effect knowledge

- example 1: predict protein from mRNA sequence

- example 2: predict class membership from handwritten digit
Covariate Shift and Semi-Supervised Learning

Goal: learn $X \leftrightarrow Y$, i.e., estimate (properties of) $p(Y|X)$

*Semi-supervised learning:* improve estimate by more data from $p(X)$

*Covariate shift:* $p(X)$ changes between training and test

Causal assumption: $p(C)$ and mechanism $p(E|C)$ "independent"

**Causal learning**

$p(X)$ and $p(Y|X)$ independent

1. semi-supervised learning impossible
2. $p(Y|X)$ invariant under change in $p(X)$

**Anticausal learning**

$p(Y)$ and $p(X|Y)$ independent

hence $p(X)$ and $p(Y|X)$ dependent

1. semi-supervised learning possible
2. $p(Y|X)$ changes with $p(X)$

---

Bareinboim & Pearl, 2012
• Experimental Meta-Analysis confirms prediction
  Schölkopf et al., ICML 2012

• All known SSL assumptions link $p(X)$ to $p(Y|X)$:
  • *Cluster assumption*: points in same cluster of $p(X)$ have the same $Y$
  • *Low density separation assumption*: $p(Y|X)$ should cross 0.5 in an area where $p(X)$ is small
  • *Semi-supervised smoothness assumption*: $E(Y|X)$ should be smooth where $p(X)$ is large
Higher-order Semi-Supervised Learning (with Julius von Kügelgen)

Causal factorization

\[ P(Y, X_E|X_C) = P(Y|X_C)P(X_E|X_C, Y) \]

Non-causal factorization

\[ P(Y, X_E|X_C) = P(X_E|X_C)P(Y|X_C, X_E) \]

\( P(X_E|X_C) \) can be estimated from unlabelled data and contains information about \( P(Y|X_C, X_E) \).

**Conditional cluster assumption**: points in the same cluster of \( p(X_E|X_C) \) have the same \( Y \)
Causal mechanisms in machine learning

• causal conditionals are robust
  \[ p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid PA_i) \]
  \((Schölkopf et al., ICML 2012)\)

• use them for causal inference in multi-task settings
  \((Zhang et al., ICML 2013 & AAAI 2015, Rojas-Carulla et al., JMLR 2018)\)

• related: transportability
  \((Bareinboim & Pearl, 2012)\)

• semi-supervised learning
  \((Schölkopf et al., ICML 2012)\)

• generative models
  \((Parascandolo et al., Besserve et al., Locatello et al.)\)
Causal generative models *(Besserve et al., AISTATS 2018)*

Independence of mechanism principles can be applied to unsupervised learning

\[ P(\text{Latent cause}) \perpendicular\!\!\!\!\!\!\!\!\!\!\!\!\perp P(\text{Data}|\text{Latent cause}) \]

Bernhard Schölkopf
Learning independent mechanisms
(with Parascandolo, Kilbertus, Rojas-Carulla)
ICML 2018

- Data drawn from $p(x)$, transformed by $M$ mechanisms $f_1, \ldots, f_M$
- Goal: learn the independent mechanisms / factors of variation
- Method: generative model with competing mechanisms

Original data

Transformed data
Method

- Mechanisms initialized \( \approx \) identity
- The highest scoring mechanism against the discriminator \( D \) wins the example and is updated to increase the score
- \( D \) is trained on the original data and against the winning outputs
Training progress

- Task: left translation
- Task: down translation
- Task: up left translation
- Task: up right translation
- Task: noise added
- Task: right translation
- Task: down left translation
- Task: down right translation
- Task: invert color
randomly sampled inputs

\begin{verbatim}
3 1 9 1 4 9 3 3 0 9 4 9 1 9 6 4
3 1 9 1 4 9 3 3 0 9 4 9 1 9 6 4
\end{verbatim}

output from the best expert
Accuracy of a CNN trained on MNIST for different test sets
Generalizing to Omniglot characters

<table>
<thead>
<tr>
<th>Exp0</th>
<th>Exp1</th>
<th>Exp2</th>
<th>Exp3</th>
<th>Exp4</th>
<th>Exp5</th>
<th>Exp6</th>
<th>Exp7</th>
<th>Exp8</th>
<th>Exp9</th>
</tr>
</thead>
<tbody>
<tr>
<td>न</td>
<td>न</td>
<td>न</td>
<td>न</td>
<td>न</td>
<td>न</td>
<td>न</td>
<td>न</td>
<td>न</td>
<td>न</td>
</tr>
<tr>
<td>स</td>
<td>स</td>
<td>स</td>
<td>स</td>
<td>स</td>
<td>स</td>
<td>स</td>
<td>स</td>
<td>स</td>
<td>स</td>
</tr>
<tr>
<td>ज</td>
<td>ज</td>
<td>ज</td>
<td>ज</td>
<td>ज</td>
<td>ज</td>
<td>ज</td>
<td>ज</td>
<td>ज</td>
<td>ज</td>
</tr>
<tr>
<td>ल</td>
<td>ल</td>
<td>ल</td>
<td>ल</td>
<td>ल</td>
<td>ल</td>
<td>ल</td>
<td>ल</td>
<td>ल</td>
<td>ल</td>
</tr>
<tr>
<td>त</td>
<td>त</td>
<td>त</td>
<td>त</td>
<td>त</td>
<td>त</td>
<td>त</td>
<td>त</td>
<td>त</td>
<td>त</td>
</tr>
<tr>
<td>ट</td>
<td>ट</td>
<td>ट</td>
<td>ट</td>
<td>ट</td>
<td>ट</td>
<td>ट</td>
<td>ट</td>
<td>ट</td>
<td>ट</td>
</tr>
</tbody>
</table>

Bernhard Schölkopf
## A Modeling Taxonomy

<table>
<thead>
<tr>
<th>Feature</th>
<th>Statistical Model</th>
<th>Causal Model</th>
<th>Differential Equation Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>i.i.d. prediction, pattern recognition, “generalization”</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Predict under shift &amp; intervention, “horizontal generalization”</td>
<td>n</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>Provide physical insight, understand predictions</td>
<td>n</td>
<td>(y)</td>
<td>y</td>
</tr>
<tr>
<td>Think/Reason, “act in an imagined space” (K. Lorenz)</td>
<td>n</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Learn from data</td>
<td>y</td>
<td>(y)</td>
<td>n</td>
</tr>
</tbody>
</table>

**Bernhard Schölkopf**
Towards causal learning

• learn a multi-task-environment SCM
  
  (cf. Schölkopf, Janzing, Lopez-Paz 2016)
  
  • data from multiple tasks in multiple environments
  • functions represent components that are robust across tasks, i.e., causal (independent) mechanisms (-> competitive training?)

• representation learning should move towards representations of causal world models

• “disentanglement” and interventional robustness (Suter et al., ICML 2019; Locatello et al., ICML 2019)

• Konrad Lorenz: ”thinking is acting is an imagined space”

• conjecture: “adversarial vulnerability” does not happen for causal conditionals (ICML 2017)
Towards causal reinforcement learning

(ICML 2017)

• why is RL on high-dimensional ATARI games harder than on low-dim. ones?
  • develop interventional methods to define objects / “sprites”?
  • transformations/interactions (cf. Felix Klein’s *Erlanger Programm*)
• why is RL easier if we permute the replayed data?
  • causal conditionals are robust and can generalize across tasks. Should build RL systems employing causal conditionals
  • exploit temporal structure / nonstationarity
  • invariance as “intrinsic motivation”
Bernhard Schölkopf

Thank You

Niki Kilbertus, Mateo Rojas-Carulla, Paul Rubenstein, Giambattista Parascandolo

Dominik Janzing, Kun Zhang, Michel Besserve, Sebastian Weichwald

David W. Hogg, Dun Wang, Dan Foreman-Mackey, Carl-Johann Simon-Gabriel