Ockham Efficiency Theorem for Randomized Scientific Methods

Conor Mayo-Wilson and Kevin T. Kelly

Department of Philosophy
Carnegie Mellon University

Formal Epistemology Workshop (FEW) June 19th, 2009

Point of the Talk

- **Ockham efficiency theorem:** Ockham's razor has been explained in terms of minimizing retractions en route to the truth, relative to all *deterministic* scientific strategies (Kevin T. Kelly and Oliver Schulte).

Point of the Talk

- **Ockham efficiency theorem:** Ockham's razor has been explained in terms of minimizing retractions en route to the truth, relative to all *deterministic* scientific strategies (Kevin T. Kelly and Oliver Schulte).
- **Extension:** Ockham's *deterministic* razor minimizes retractions en route to the truth, relative to a broad class of *random* scientific strategies.

Point of the Talk

- **Ockham efficiency theorem:** Ockham's razor has been explained in terms of minimizing retractions en route to the truth, relative to all *deterministic* scientific strategies (Kevin T. Kelly and Oliver Schulte).
- **Extension:** Ockham's *deterministic* razor minimizes retractions en route to the truth, relative to a broad class of *random* scientific strategies.
- **Further significance:** Extending the argument to expected retractions is a necessary step for lifting the idea to a theory of statistical theory choice.

Outline

The Puzzle of Simplicity

- Ockham's Razor is indespensible in scientific inference.
- Inference should be truth-conducive.
- But how could a fixed bias toward simplicity be said to help one find possibly complex truths?

- **Methodological virtues:** Simpler theories are more testable or explanatory or are otherwise more virtuous.

- **Methodological virtues:** Simpler theories are more testable or explanatory or are otherwise more virtuous.
- **Response:** wishful thinking—desiring that the truth be virtuous doesn’t make it so.

Standard Explanations

- **Confirmation:** Simple theories are better confirmed by simple data:

$$\frac{p(T_S \mid E)}{p(T_C \mid E)} > \frac{p(T_S)}{p(T_C)}.$$

Standard Explanations

- **Confirmation:** Simple theories are better confirmed by simple data:

$$\frac{p(T_S \mid E)}{p(T_C \mid E)} > \frac{p(T_S)}{p(T_C)}.$$

- **Responses**
  - Simple data are compatible with complex hypotheses.

Standard Explanations

- **Confirmation:** Simple theories are better confirmed by simple data:

$$\frac{p(T_S \mid E)}{p(T_C \mid E)} > \frac{p(T_S)}{p(T_C)}.$$

- **Responses**
  - Simple data are compatible with complex hypotheses.
  - No clear connection with finding the truth in the short run.

Standard Explanations

- **Over-fitting:** Even if the truth is complex, simple theories improve overall predictive accuracy by trading variance for bias.

Standard Explanations

- **Over-fitting:** Even if the truth is complex, simple theories improve overall predictive accuracy by trading variance for bias.
- **Responses:**
  - The underlying decision theory is unclear—the worst-case solution is the most complex theory.

Standard Explanations

- **Over-fitting:** Even if the truth is complex, simple theories improve overall predictive accuracy by trading variance for bias.
- **Responses:**
  - The underlying decision theory is unclear—the worst-case solution is the most complex theory.
  - The over-fitting account ties Ockham’s razor to choices among stochastic theories.

Standard Explanations

- **Over-fitting:** Even if the truth is complex, simple theories improve overall predictive accuracy by trading variance for bias.
- **Responses:**
  - The underlying decision theory is unclear—the worst-case solution is the most complex theory.
  - The over-fitting account ties Ockham's razor to choices among stochastic theories.
  - "Prediction" must be understood so as to rule out *counterfactual* or *causal* predictions.

Standard Explanations

- **Convergence:** Even if the truth is complex, complexities in the data will eventually over-turn over-simplified theories.

Standard Explanations

- **Convergence:** Even if the truth is complex, complexities in the data will eventually over-turn over-simplified theories.
- **Responses:**
  - Any theory choice in the short run is compatible with finding the true theory in the long run.

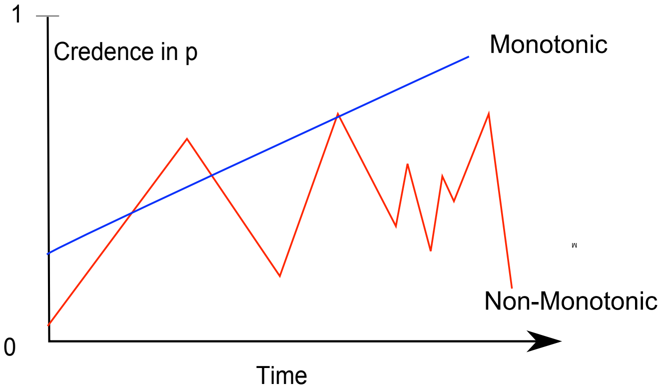
A New Approach

- **Deduction:**
  - Sound;
  - Monotone.

A New Approach

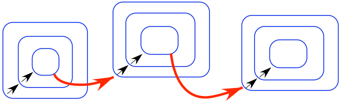
- **Deduction:**
  - Sound;
  - Monotone.
- **Induction:**
  - Approximate Soundness: converge to truth with minimal errors;
  - Approximate Monotonicity: minimize retractions.

Monotonicity for Bayesians



Monotonicity in Belief Revision

- Total number of times $B_{n+1} \not\models B_n$
  $=$ total number of non-expansive belief revisions.



Main Idea

- Ockham's Razor = Closest Inductive Approximation to Deduction

Main Idea

- Ockham's Razor = Closest Inductive Approximation to Deduction
- Ockham's razor converges to the truth with minimal retractions and errors and elapsed time to retractions, elapsed time to errors

Main Idea

- Ockham’s Razor = Closest Inductive Approximation to Deduction
- Ockham’s razor converges to the truth with minimal retractions and errors and elapsed time to retractions, elapsed time to errors
- No other convergent method does

Main Idea

- Ockham's Razor = Closest Inductive Approximation to Deduction
- Ockham's razor converges to the truth with minimal retractions and errors and elapsed time to retractions, elapsed time to errors
- No other convergent method does
- No circular appeal to prior simplicity biases

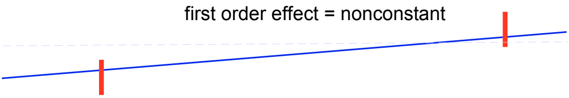
Main Idea

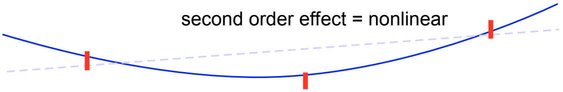
- Ockham's Razor = Closest Inductive Approximation to Deduction
- Ockham's razor converges to the truth with minimal retractions and errors and elapsed time to retractions, elapsed time to errors
- No other convergent method does
- No circular appeal to prior simplicity biases
- No awkward trade-offs between costs are required.

Empirical Problems and Simplicity

Empirical Effects:

- Recognizable eventually.
- Arbitrarily subtle so may take arbitrarily long to be noticed.
- Each theory predicts finitely many.





Empirical Problems and Simplicity

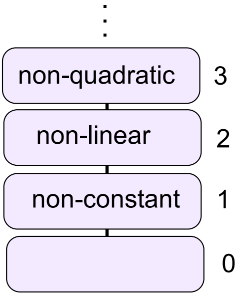
Empirical Problems:

- **Background knowledge** K picks out a set of possible, finite, effect sets.
- **Theory** T_S says that exactly the effects in S will be seen in the unbounded future.
- A **world of experience** w is an infinite sequence that presents some finite set of effects at each stage and that presents some set $S \in K$ in the limit.
- The **empirical problem** corresponding to K is to determine which theory in $\{T_S : S \in K\}$ is true of the actual world of experience w.

Empirical Problems and Simplicity

What simplicity is:

- The empirical complexity of world of experience w is the length of the longest effect path in K to the effect set S_w presented by w.



Empirical Problems and Simplicity

What simplicity isn't:

- notational or computational brevity (MDL),
- a question-begging rescaling of prior probability using $-\ln(x)$ (MML).
- free parameters or dimensionality (AIC).

Three Paradigmatic Examples

- Linearly ordered complexity with refutation: Curve Fitting
  - Kelly [2004]

Three Paradigmatic Examples

- Linearly ordered complexity with refutation: Curve Fitting
    - Kelly [2004]
- Partially ordered complexity with refutation: Causal Inference
    - Schulte, Luo and Greiner [2007],
    - Kelly and Mayo-Wilson [2008]

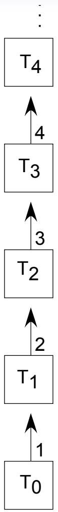
Three Paradigmatic Examples

- Linearly ordered complexity with refutation: Curve Fitting
  - Kelly [2004]
- Partially ordered complexity with refutation: Causal Inference
  - Schulte, Luo and Greiner [2007],
  - Kelly and Mayo-Wilson [2008]
- Partially ordered complexity without refutation: Orientation of Causal Edge
  - Kelly and Mayo-Wilson [2008]

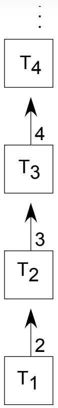
Linearly Ordered Simplicity Structure

- Curve fitting is an instance of a more general type of problem.
- Problem: Choosing amongst theories that are linearly ordered in terms of complexity.
- Evidence: Suppose that any false, simple theory is refuted in some finite amount of time.

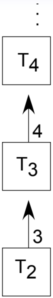
Linearly Ordered Simplicity Structure



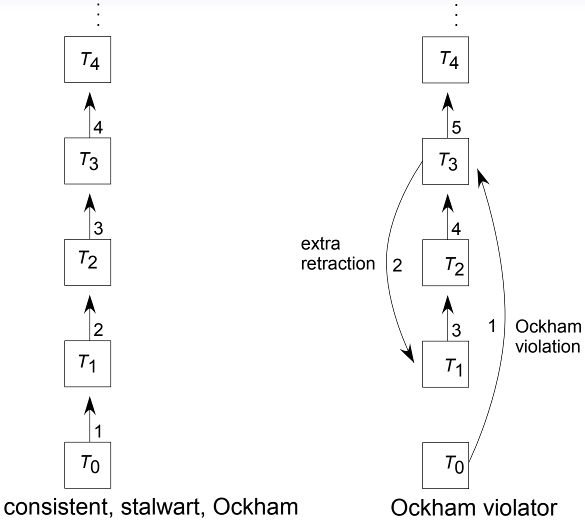
Linearly Ordered Simplicity Structure



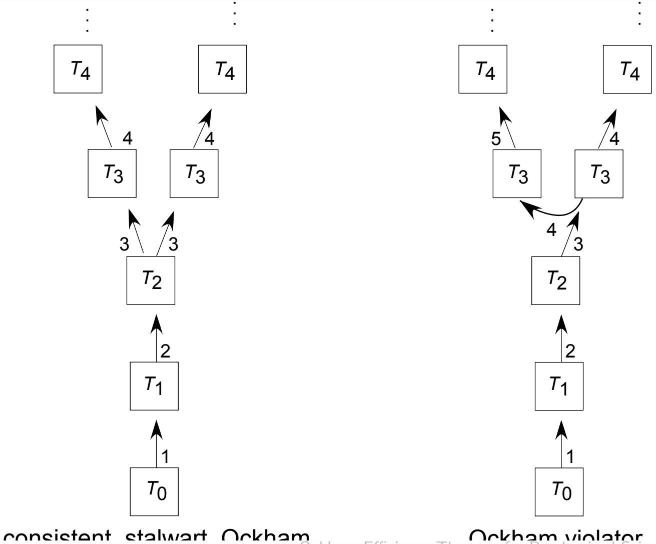
Linearly Ordered Simplicity Structure



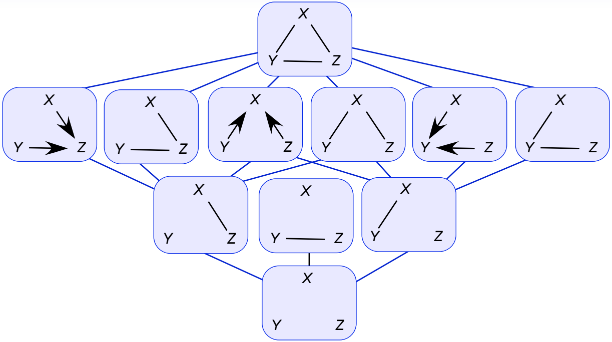
Linearly Ordered Simplicity Structure



Branching Simplicity Structure



Branching Simplicity Structure



Discovering Causal Networks

- Problem: Choose a causal network describing the causal relationships amongst a set of variables.

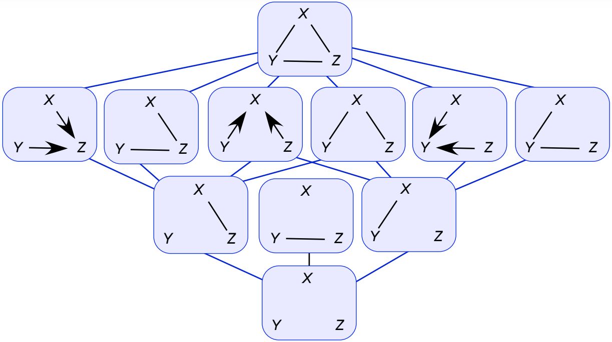
Discovering Causal Networks

- Problem: Choose a causal network describing the causal relationships amongst a set of variables.
- Evidence (Effects): probabilistic dependencies amongst the variables discovered over time.

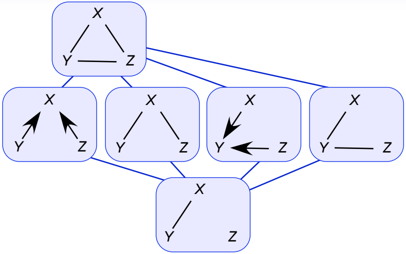
Discovering Causal Networks

- Problem: Choose a causal network describing the causal relationships amongst a set of variables.
- Evidence (Effects): probabilistic dependencies amongst the variables discovered over time.
- Complexity = greater number of edges

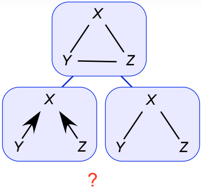
Discovering Causal Networks



Discovering Causal Networks



Discovering Causal Networks



Deterministic Theory Choice Methods

- Given arbitrary, finite initial segment e of a world of experience, a deterministic theory choice method produces a unique theory T_S or '?' indicating refusal to choose.

Methodological Properties

Say a method M is

- **convergent** if, for any world w, M eventually produces the true theory in w.

Methodological Properties

Say a method M is

- **convergent** if, for any world w, M eventually produces the true theory in w.
- **Ockham** if it never says any non-simple theory (relative to evidence).

Methodological Properties

Say a method M is

- **convergent** if, for any world w, M eventually produces the true theory in w.
- **Ockham** if it never says any non-simple theory (relative to evidence).
- **stalwart** if whenever it endorses the simplest theory, it continues to do so until it's no longer simplest.

Methodological Properties

Say a method M is

- **convergent** if, for any world w, M eventually produces the true theory in w.
- **Ockham** if it never says any non-simple theory (relative to evidence).
- **stalwart** if whenever it endorses the simplest theory, it continues to do so until it's no longer simplest.
- **eventually informative** if, in any world, there is some point of inquiry n after which M never says '?' in w.

Ockham's Razor in Probability

Say a method M is a **normally Ockham** if it is Ockham, stalwart, and eventually informative.

Ockham's Razor

Modulo some minor assumptions on the simplicity structure, which all of the above examples satisfy.

Theorem (Efficiency Theorem)

Let M be a normal Ockham method, and let M' be any convergent method. Suppose M and M' agree along some finite initial set of experience e_-, and that M' violates Ockham's razor at e. Then M' commits strictly more retractions (in the worst-case) in every complexity class with respect to e.

Mixed Strategies in Decision and Game Theory

- Randomization in decision theory and games: E.g. Rock, paper, scissors, matching pennies

Mixed Strategies in Decision and Game Theory

- Randomization in decision theory and games: E.g. Rock, paper, scissors, matching pennies
- Can randomization in scientific inquiry improve expected errors and retractions?

Randomized Strategies

- Randomized Methods: Machines (formally, discrete state stochastic processes) for selecting theories from data

Randomized Strategies

- Randomized Methods: Machines (formally, discrete state stochastic processes) for selecting theories from data
- Outputs of machine are a function of (i) its current state and (ii) total input

Randomized Strategies

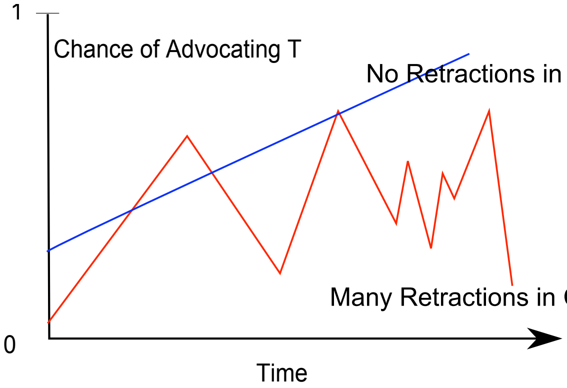
- Randomized Methods: Machines (formally, discrete state stochastic processes) for selecting theories from data
- Outputs of machine are a function of (i) its current state and (ii) total input
- States of machine evolve according to a random process i.e.
  - Future and past states may be correlated to any degree - Independence **not** assumed!
  - No assumptions about process being Markov, etc.

Costs and Randomized Strategies

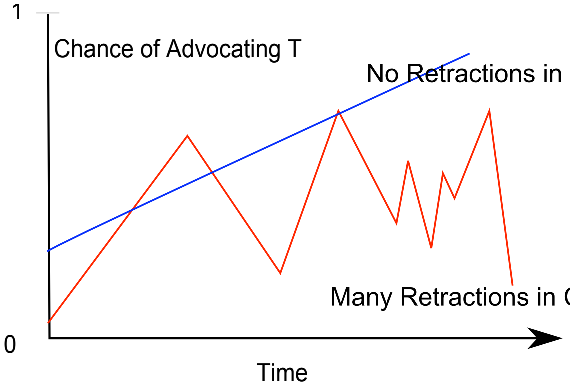
Randomized methods ought to be as deductive as possible:

- Approximate Soundness: convergence **in probability**, minimization of **expected** errors
- Approximate Monotonicity: minimization of **expected** retractions

Retractions in Chance and Expected Retractions



Retractions in Chance and Expected Retractions



Methodological Properties

Say a randomized method M is

- **Ockham** if it never says any non-simple theory with probability greater than zero.

Methodological Properties

Say a randomized method M is

- **Ockham** if it never says any non-simple theory with probability greater than zero.
- **Stalwart** if whenever it endorses the simplest theory (with any positive probability), it continues to do so with unit probability until it is no longer Ockham.

Methodological Properties

Say a randomized method M is

- **Ockham** if it never says any non-simple theory with probability greater than zero.
- **Stalwart** if whenever it endorses the simplest theory (with any positive probability), it continues to do so with unit probability until it is no longer Ockham.
- **Convergent in Probability** if for any world w, the probability that M produces the theory true in w approaches 1 as time elapses.

Ockham's Razor in Probability

Say a method M is a **normally Ockham** if it is Ockham, stalwart, and convergent in probability.

Generalized Efficiency Theorem: Suppose the simplicity structure has no short paths, and let M be a randomized or deterministic method such that

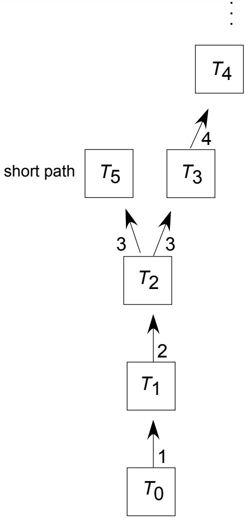
1. M first violates Ockham's razor after some initial segment of evidence e.
2. M is convergent in probability

Then, in comparison to any normal Ockham method (deterministic or not!), M accrues a strictly greater number of retractions in every complexity class with respect to e.

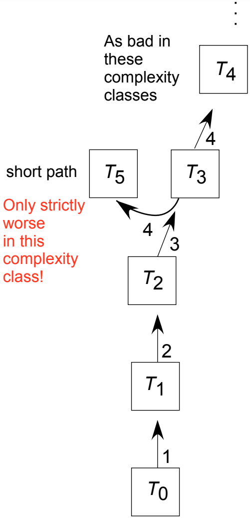
References

- Mark Gold, E. "Language identification in the limit." *Information and Control.* V 10. 447-474.
- Kevin Kelly. *The Logic of Reliable Inquiry.* Oxford University Press, 1996.
- Oliver Schulte. Inferring Conservation Laws in Particle Physics: A Case Study in the Problem of Induction. *The British Journal for the Philosophy of Science.* 2001.
- Oliver Schulte, W. Luo and R. Greiner. "Mind Change Optimal Learning of Bayes Net Structure." In *20th Annual Conference on Learning Theory* (COLT), San Diego, CA, June 12-15, 2007.

Branching Simplicity Structure



consistent stalwart Ockham



Ockham violator

Stochastic Processes

Definition

Let T, Δ, Σ be arbitrary sets. A stochastic process is a quadruple $\mathcal{Q} = (T, (\Delta, \mathcal{D}, p), (\Sigma, \mathcal{S}), X)$, where:

1. T is a set called the *index set* of the process;
2. $(\Delta, \mathcal{D}, p)$ is a (countably additive) probability space;
3. $(\Sigma, \mathcal{S})$ is a measurable space of possible *values* of the process;
4. $X : T \times \Delta \to \Sigma$ is such that for each fixed $t \in T$, the function X_t is $\mathcal{D}/\mathcal{S}$-measurable.

Stochastic Processes

Definition

Let F represent finite, initial segments of data streams, and Ans contain the set of all theories and '?' representing "I don't know. A stochastic empirical method is a triple $\mathcal{M} = (\mathcal{Q}, \alpha, \sigma_0)$ where:

1. $\mathcal{Q} = (F, (\Delta, \mathcal{D}, p), (\Sigma, \mathcal{S}), X)$ is a stochastic process indexed by the set F of all finite, initial segments of data streams.
2. $\sigma_0 \in \Sigma$ is the initial state of our method (i.e. it satisfies: $X_{()}^{-1}(\sigma_0) = \Delta$).
3. $\alpha : F \times \Sigma \to \text{Ans}$ is such that for each $e \in F$, α_e is $\mathcal{G}/2^{\text{Ans}}$-measurable.