10-708 Probabilistic Graphical Models

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Gaussian Process

+

Causality

+

RL as Inference

Matt Gormley
Lecture 25
May. 7, 2021

Reminders

- Cloud Credits (AWS or GCP)
  - first request deadline: Thu at 11:59pm
- Final Project Milestones
  - Final Poster Session
  - Final Poster submission
  - Final Executive Summary submission

GAUSSIAN PROCESS

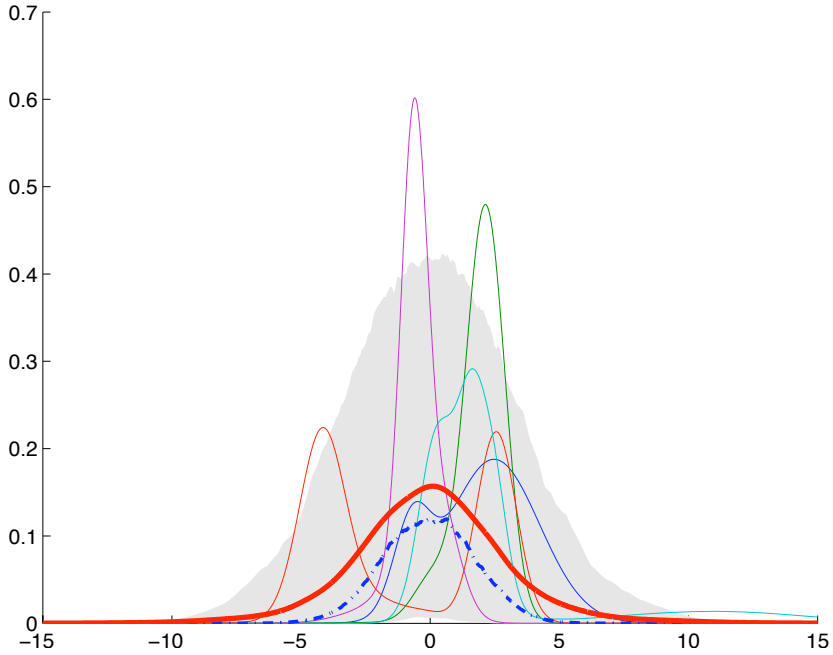
Figure from Teh MLSS 2007

Motivation: Gaussian Process

Density Estimation

- Given data, estimate a probability density function that best explains it
- A nonparametric prior can be placed over an infinite set of distributions

Prior:



Red: mean density. Blue: median density. Grey: 5-95 quantile.
Others: draws.

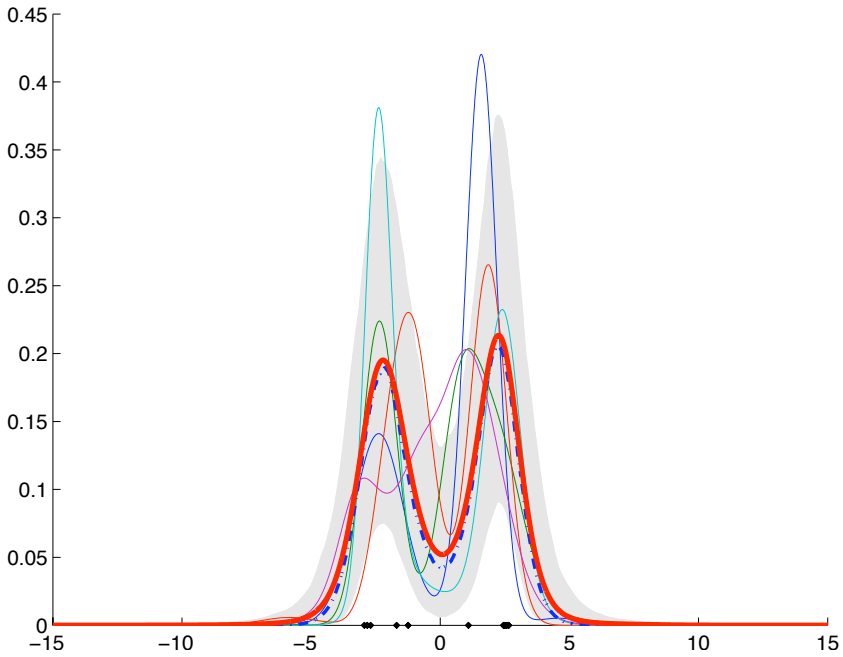
Figure from Teh MLSS 2007

Motivation: Gaussian Process

Density Estimation

- Given data, estimate a probability density function that best explains it
- A nonparametric prior can be placed over an infinite set of distributions

Posterior:



Red: mean density. Blue: median density. Grey: 5-95 quantile.
Black: data. Others: draws.

Gaussian Process

Whiteboard:

- Parametric vs. Nonparametric learning
- High level idea of GP regression
- GP Regression
  - Example prior
  - Strawman inference algorithm
  - Example posterior
- GP Classification
  - approximate inference
  - Example posterior

Multivariate Gaussians

Problem Setup:

Suppose we have a Multivariate Gaussian:

$$p(\mathbf{x}) \sim \mathcal{N}(\mathbf{u}, \boldsymbol{\Sigma})$$

where

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

$$\mathbf{x}_1 \in \mathbb{R}^{D_1} \qquad \mathbf{u}_1 \in \mathbb{R}^{D_1} \qquad \boldsymbol{\Sigma}_{ij} \in \mathbb{R}^{D_i \times D_j}$$

$$\mathbf{x}_2 \in \mathbb{R}^{D_2} \qquad \mathbf{u}_2 \in \mathbb{R}^{D_2}$$

Recall: $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^T$

Question 1: *True or False:* The marginals of the distribution are given by:

$$p(\mathbf{x}_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{ii}), \quad \forall i \in \{1, 2\}$$

Question 2: *True or False:* The conditionals of the distribution are given by:

$$p(\mathbf{x}_i \mid |\mathbf{x}_j) \sim \mathcal{N}\Big(\boldsymbol{\mu}_i + \boldsymbol{\Sigma}_{ij}\boldsymbol{\Sigma}_{jj}^{-1}(\mathbf{x}_j - \mu_j), \boldsymbol{\Sigma}_{jj} - \Sigma_{ij}^T\Sigma_{ii}^{-1}\Sigma_{ij}\Big), \quad \forall i, j \in \{(1, 2), (22, 1)\}$$

Background: Multivariate Gaussians

Whiteboard:

- Marginal of multivariate Gaussian
- Conditional of multivariate Gaussian

Gaussian Process Regression

Whiteboard:

- Function-space view
  - definition of Gaussian Process
  - mean function
  - covariance function
- Example kernels
- Weight-space view
  - linear regression (linear model + Gaussian noise)
  - ridge regression (adding a Gaussian prior)
  - Bayesian linear regression
  - Bayesian kernel regression (aka. GP Regression)

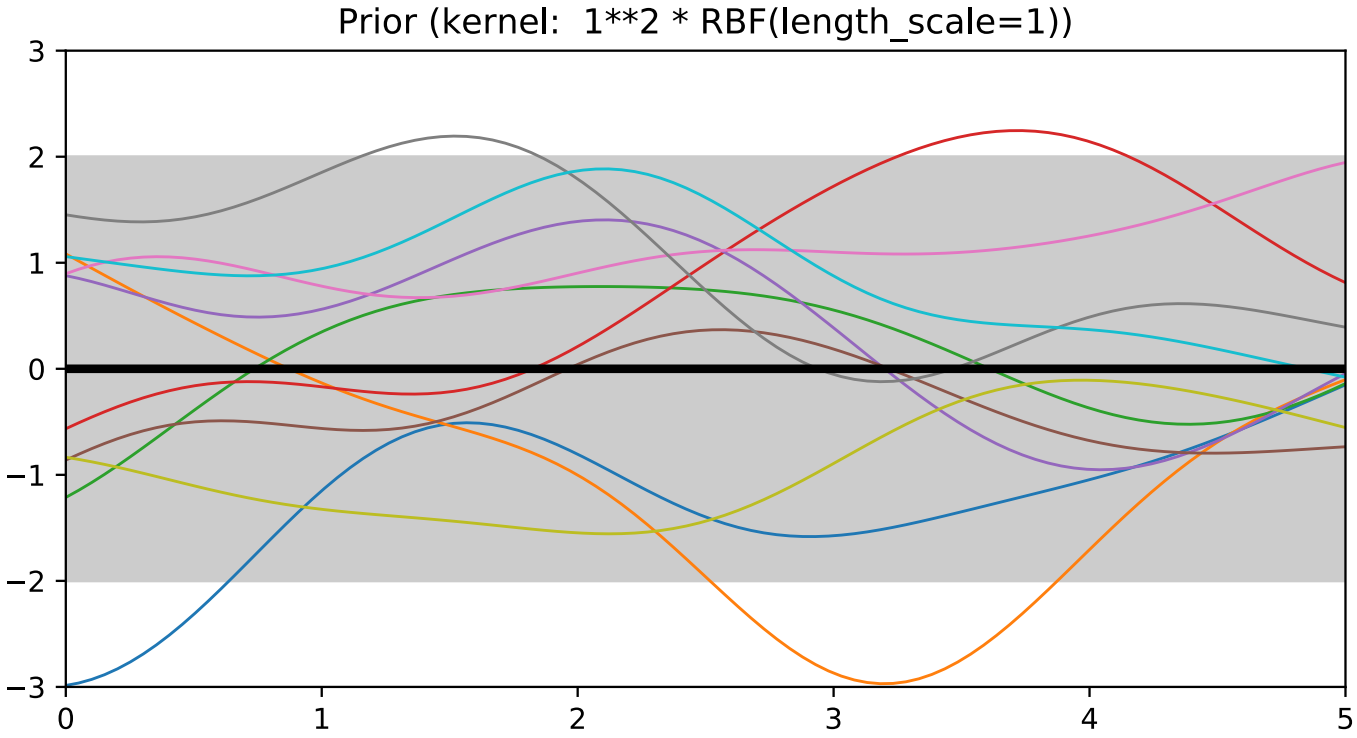
Gaussian Process Regression

Whiteboard:

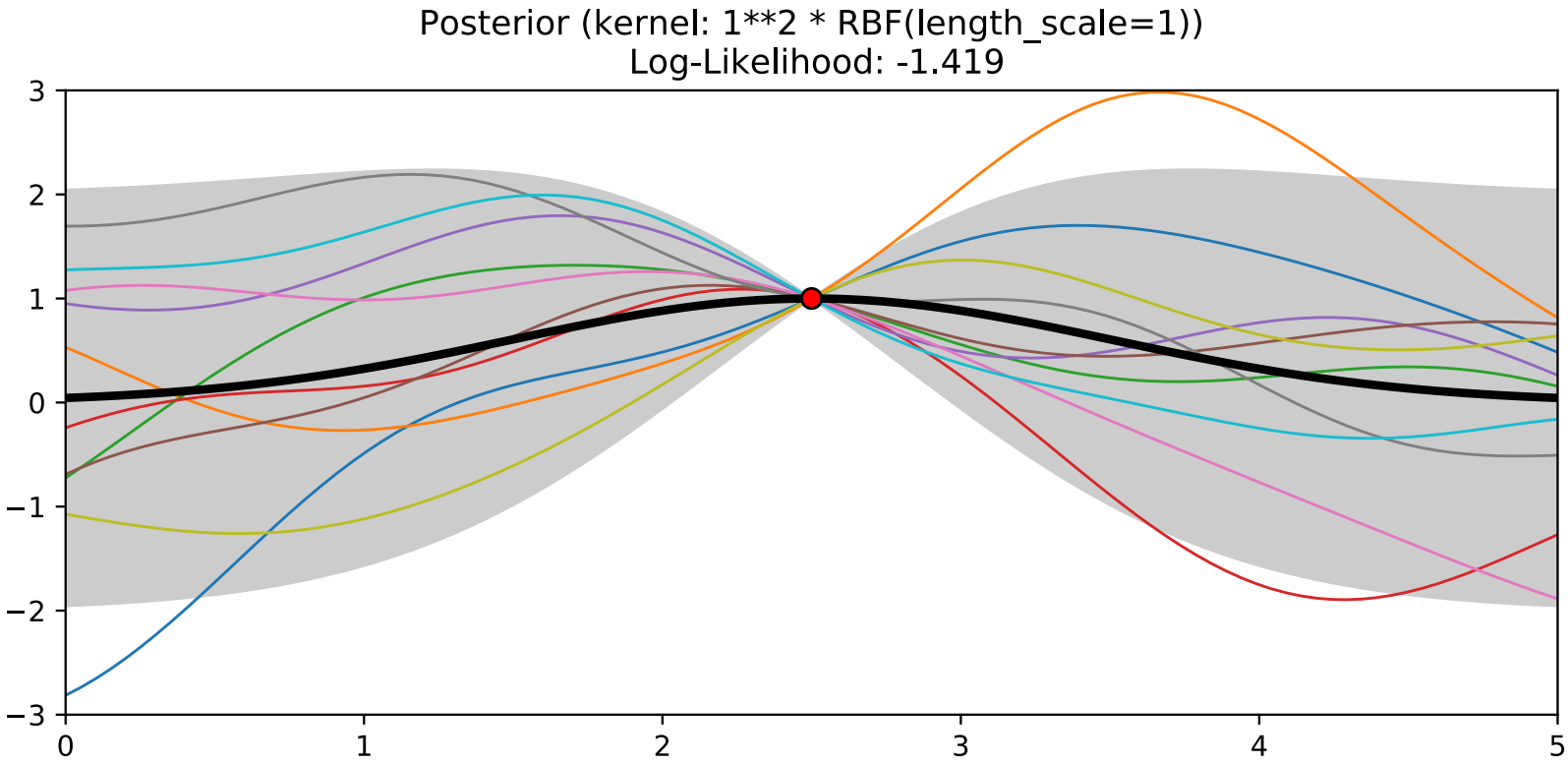
- MBR Decoding
- Computational complexity

GAUSSIAN PROCESS INFERENCE

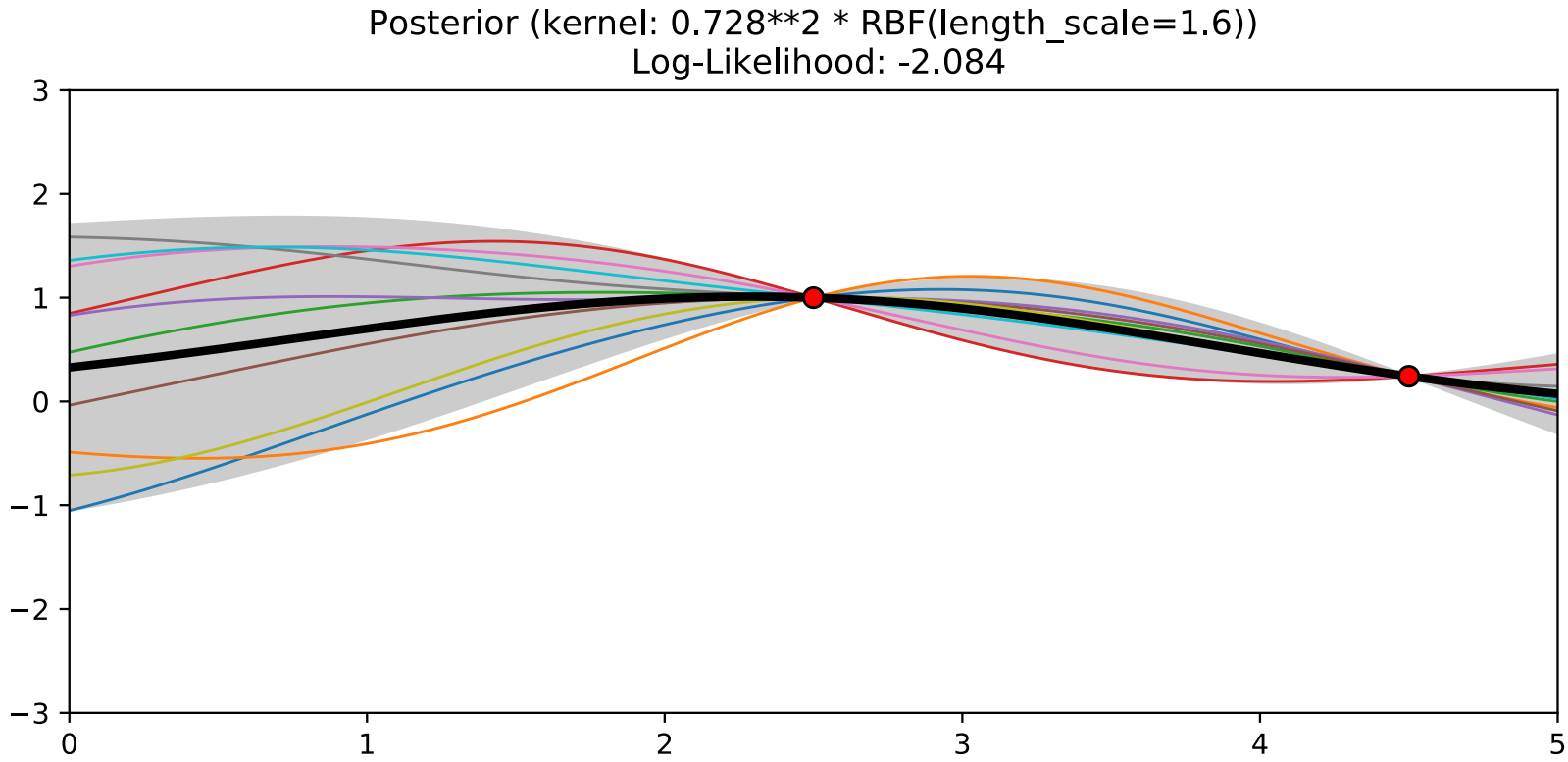
Gaussian Process Example



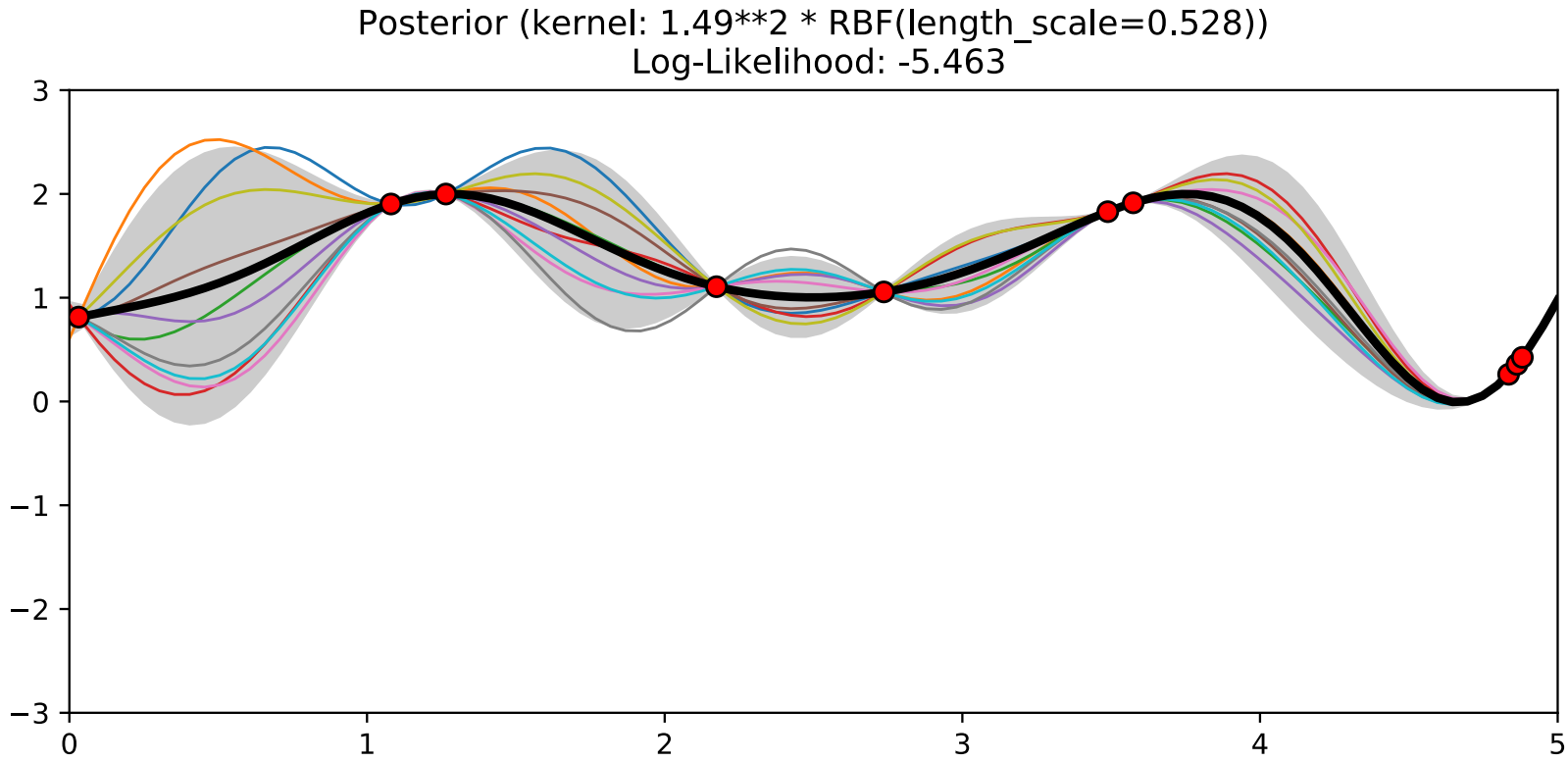
Gaussian Process Example



Gaussian Process Example

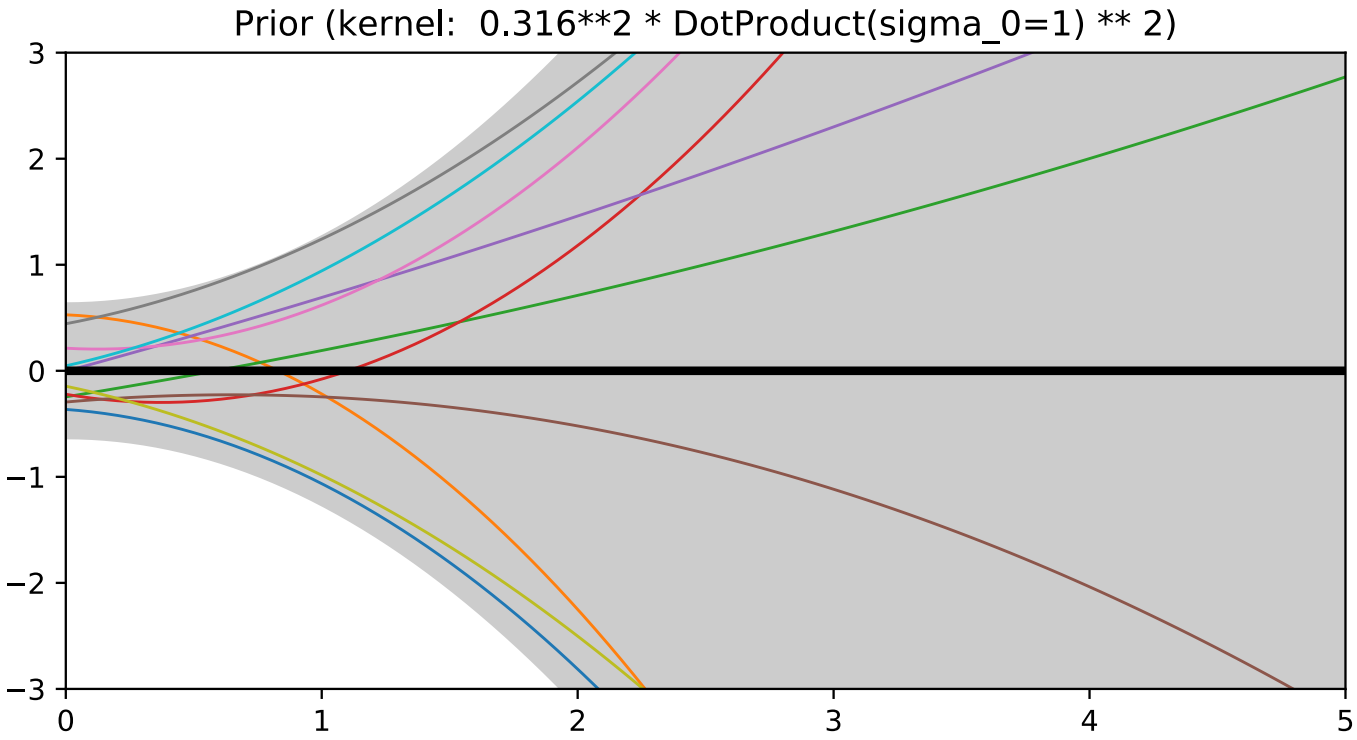


Gaussian Process Example

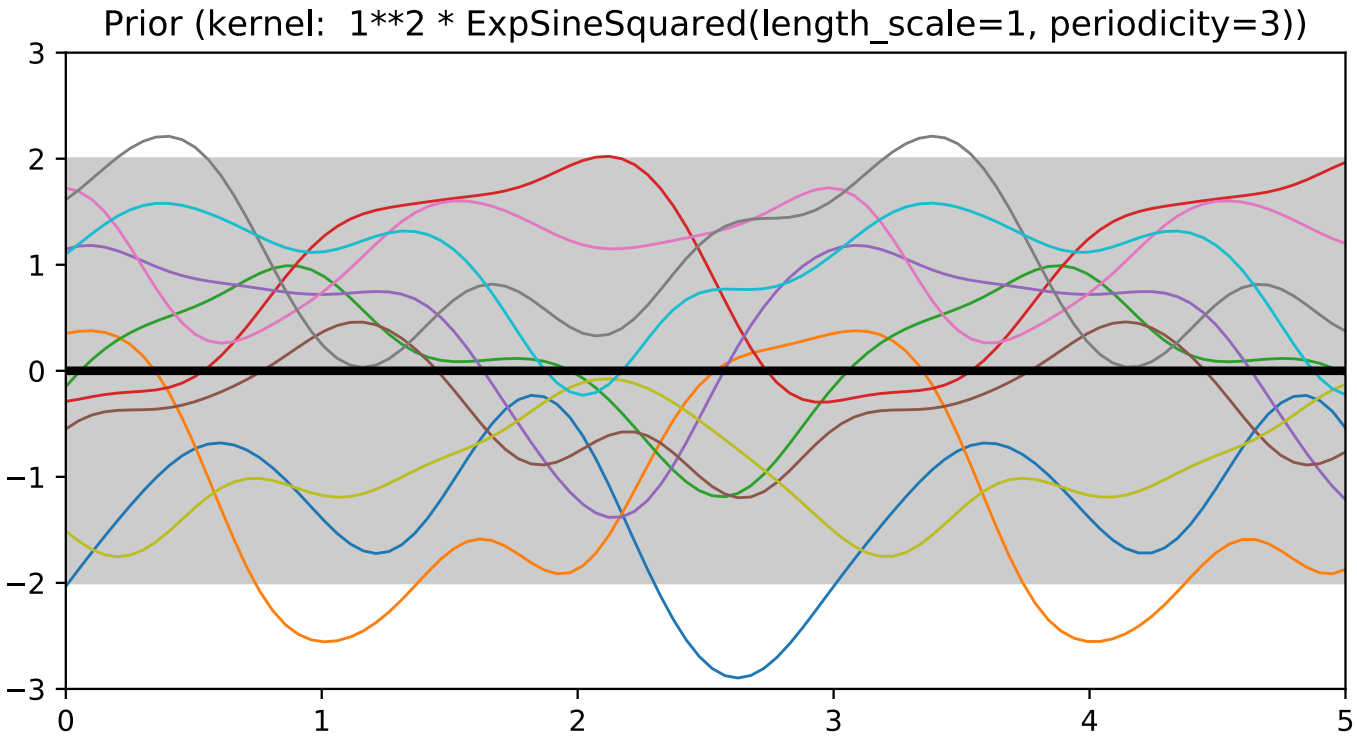


GAUSSIAN PROCESS KERNELS

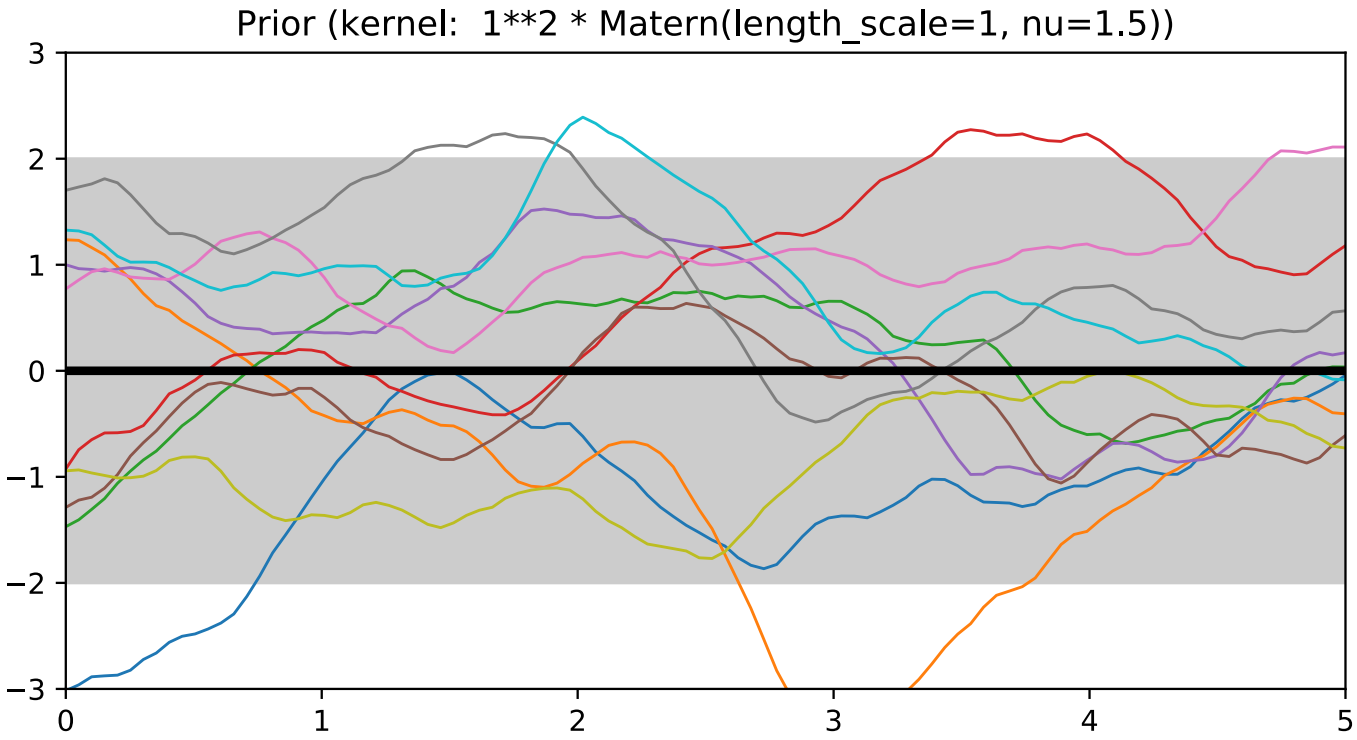
Gaussian Process Example



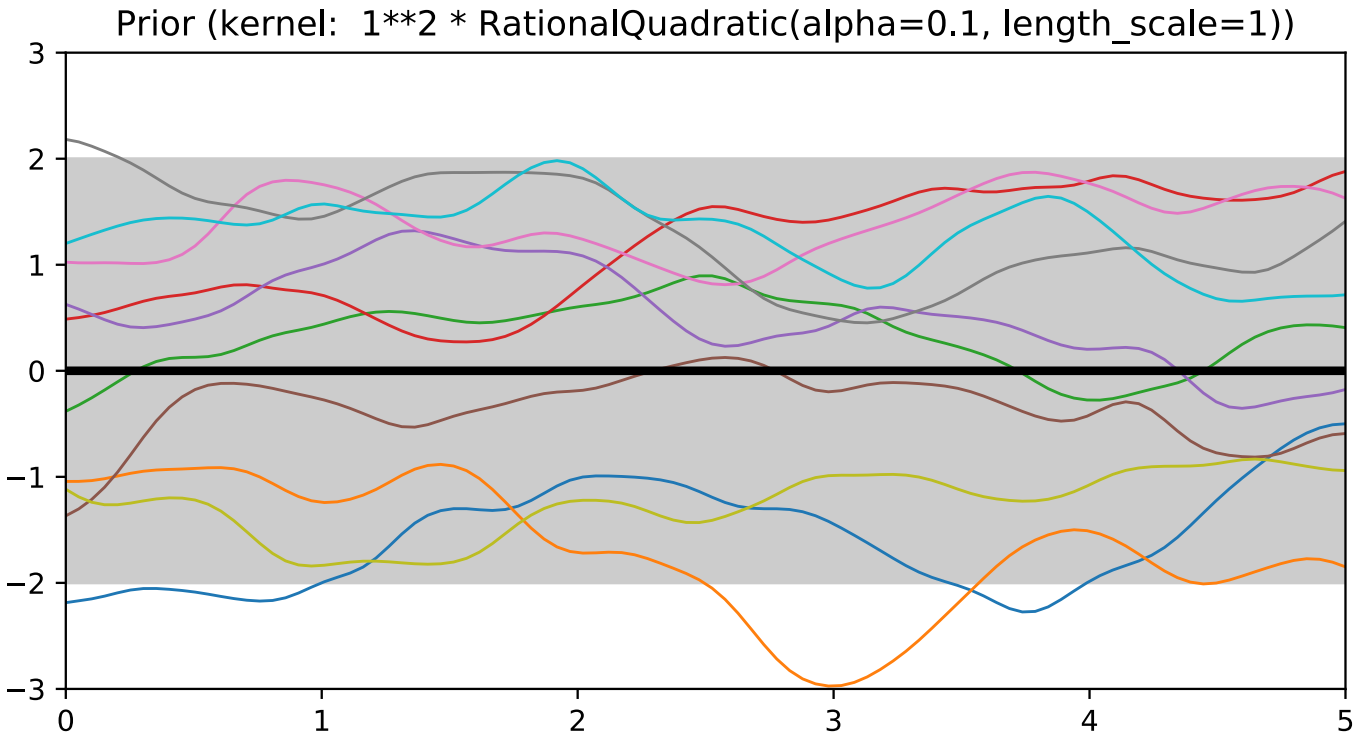
Gaussian Process Example



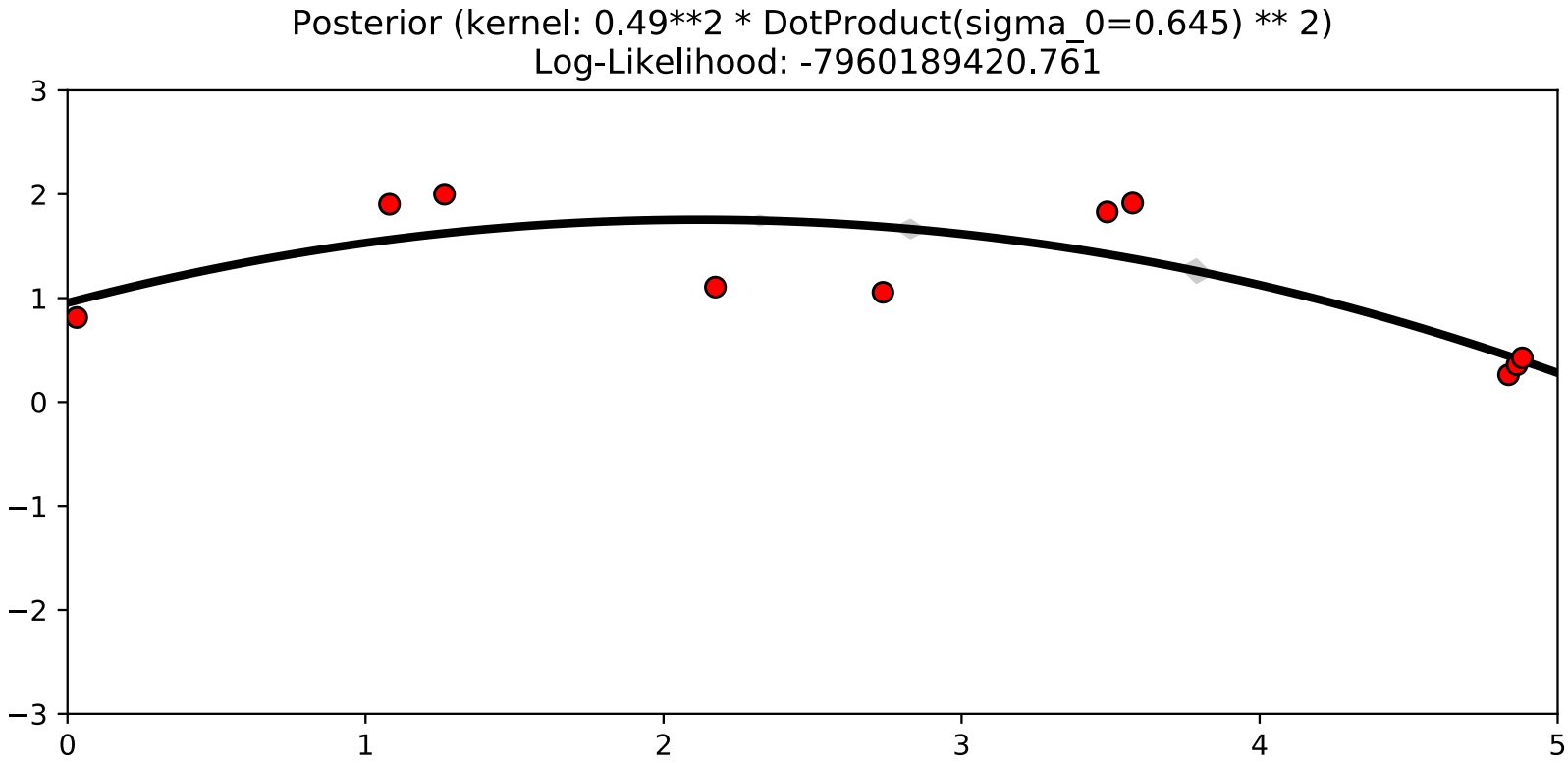
Gaussian Process Example

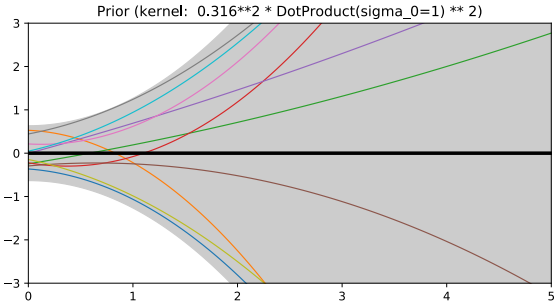


Gaussian Process Example

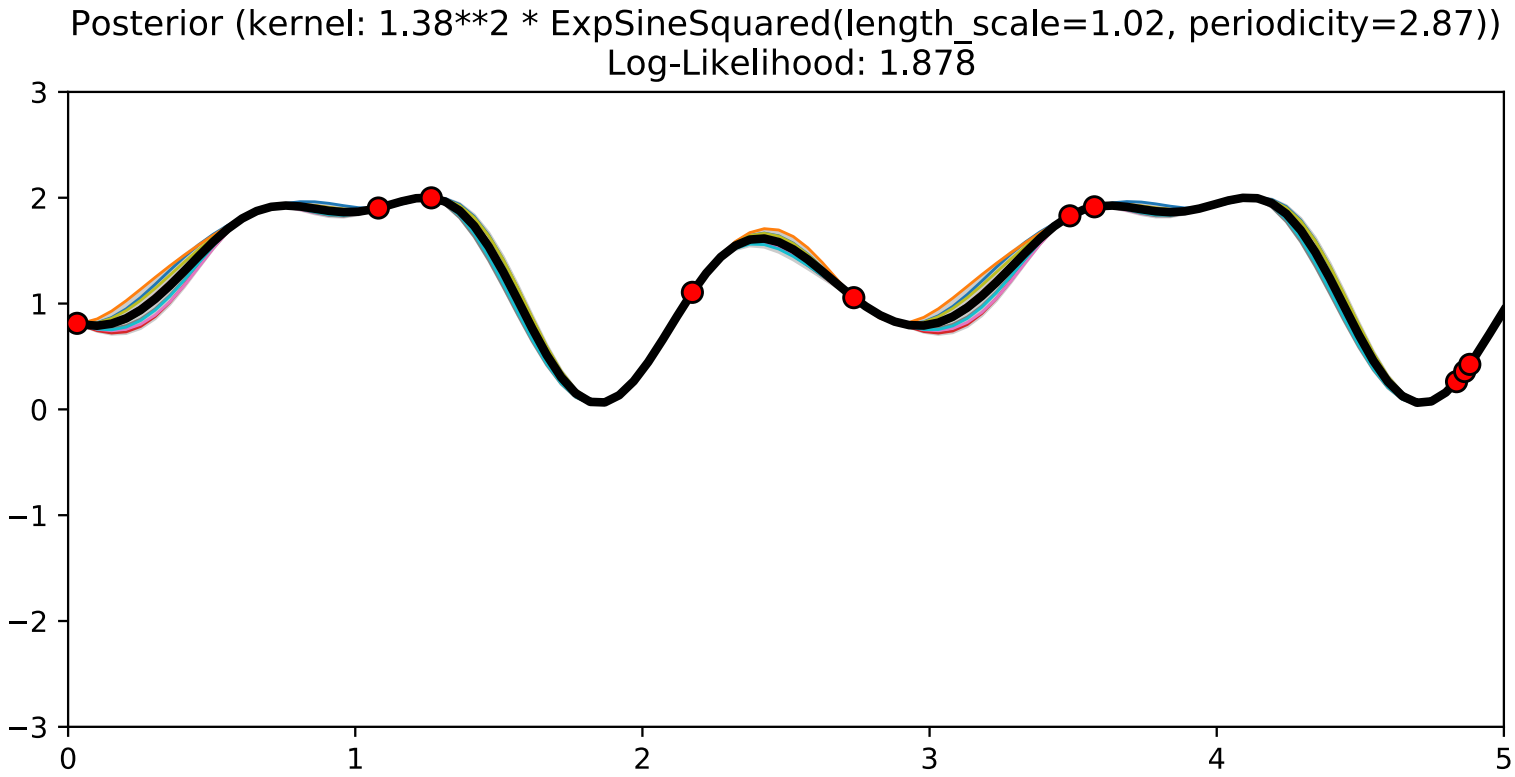


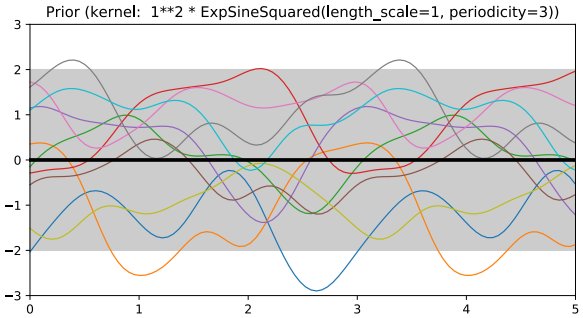
Gaussian Process Example



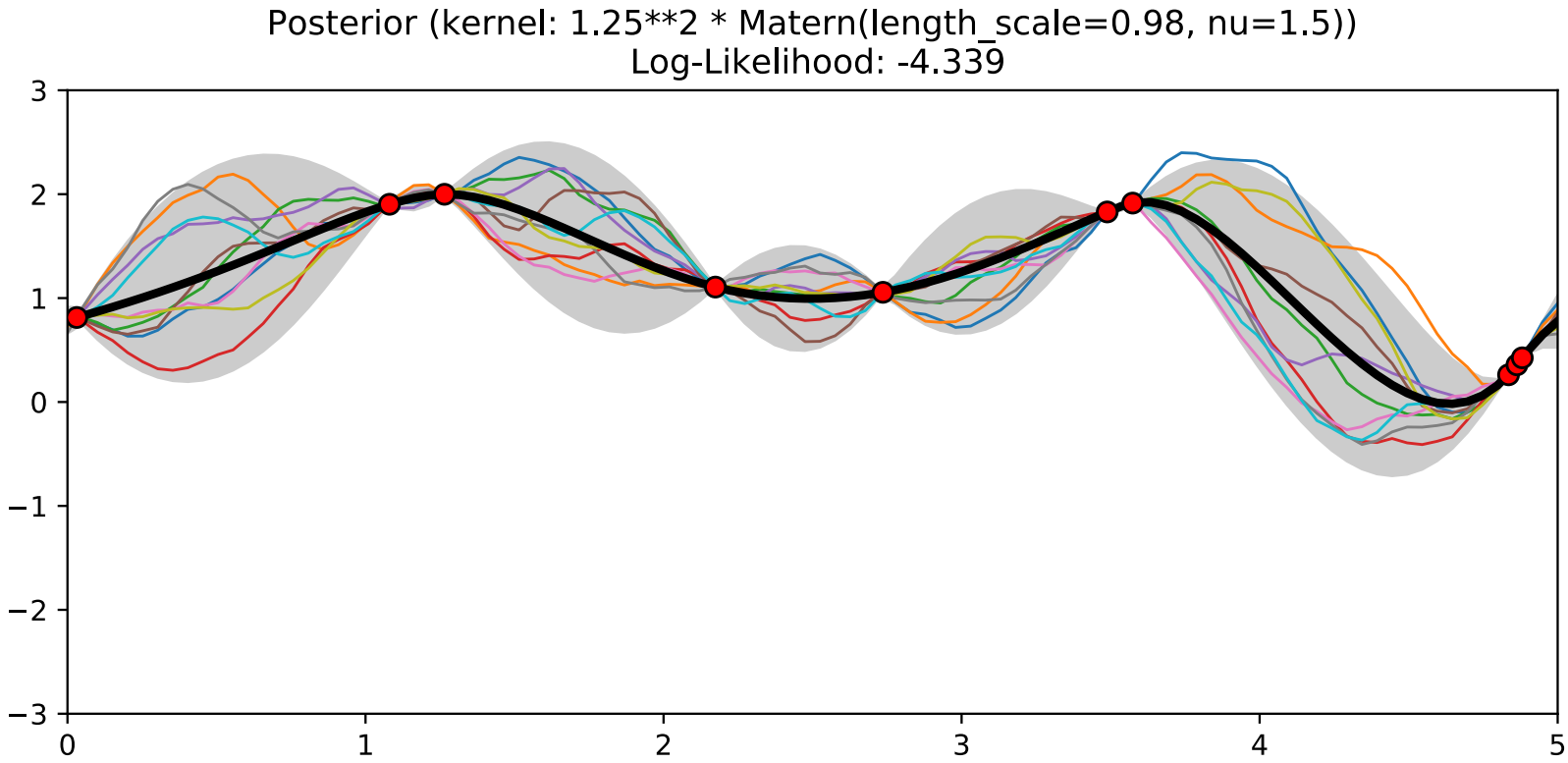


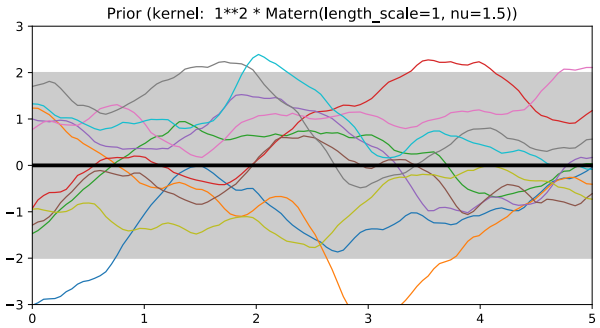
Gaussian Process Example



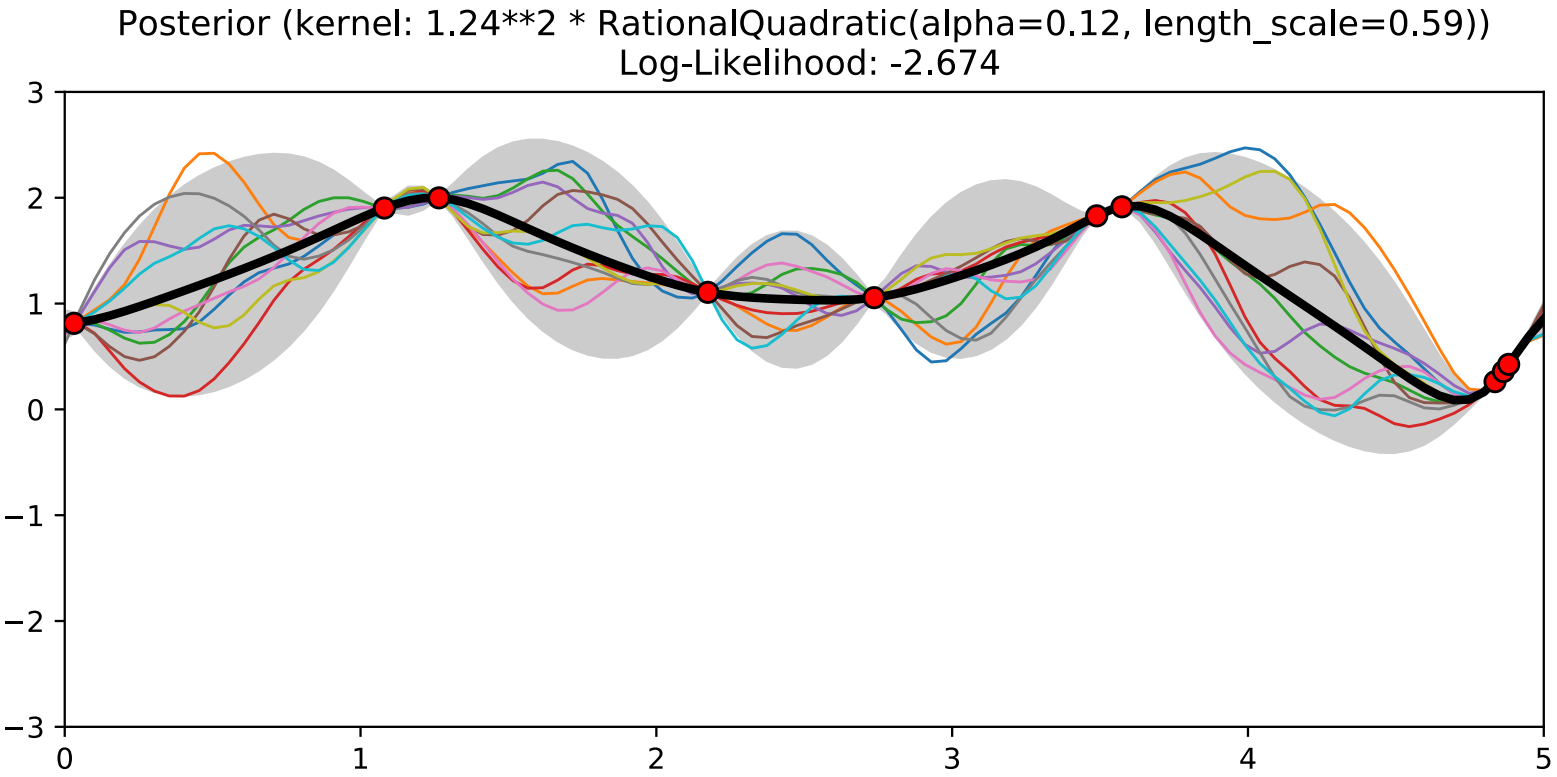


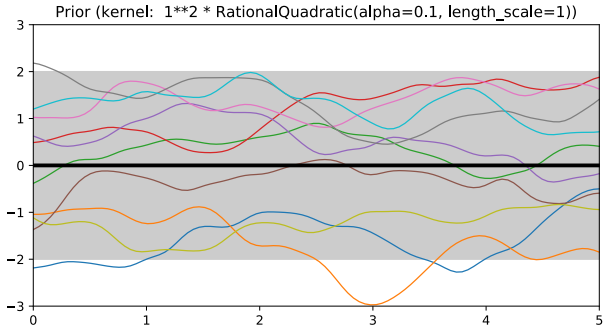
Gaussian Process Example





Gaussian Process Example





CAUSAL INFERENCE

Correlation

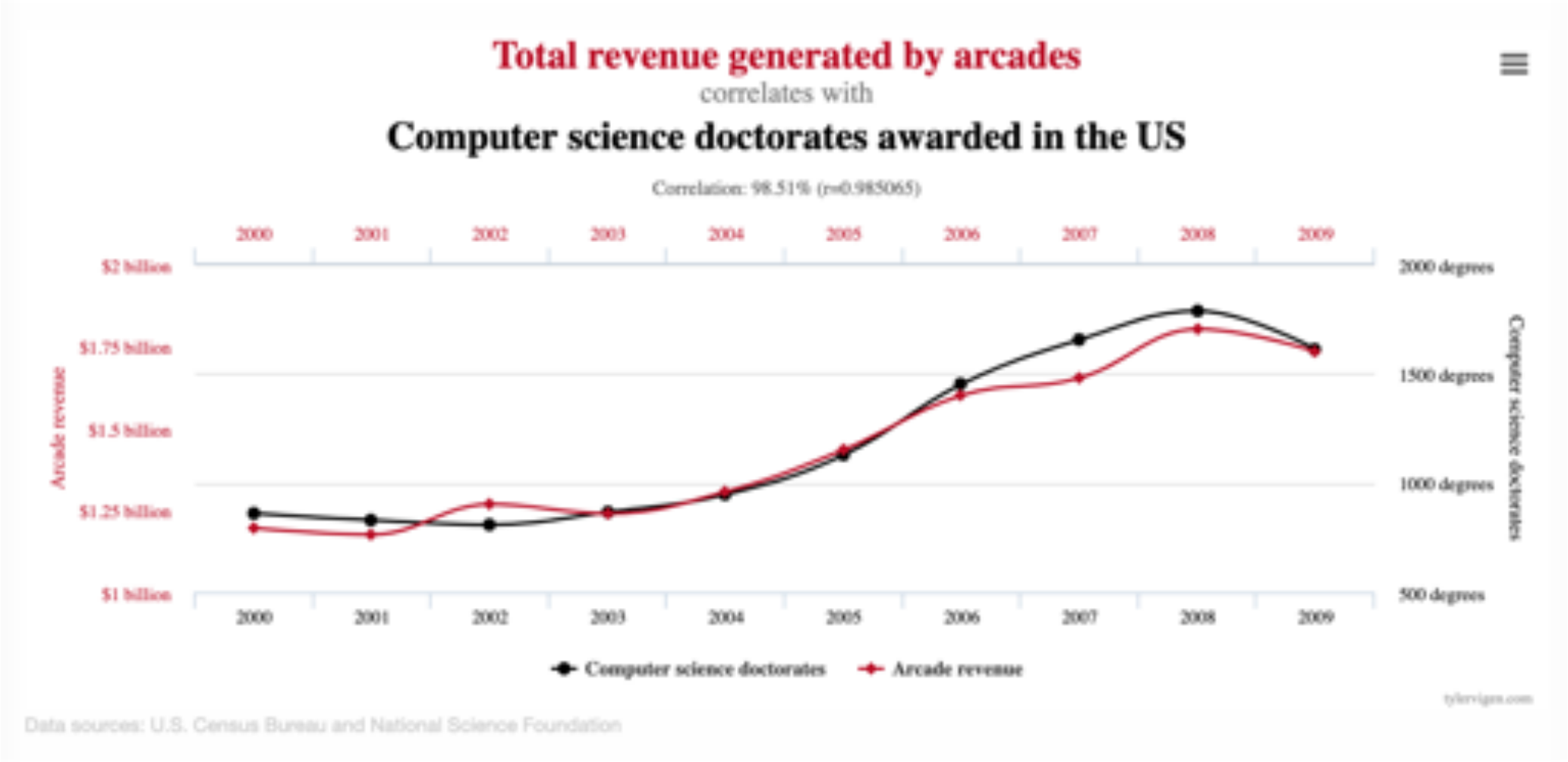


Figure from https://www.tylervigen.com/spurious-correlations

Correlation

Figure from https://www.tylervigen.com/spurious-correlations

Correlation

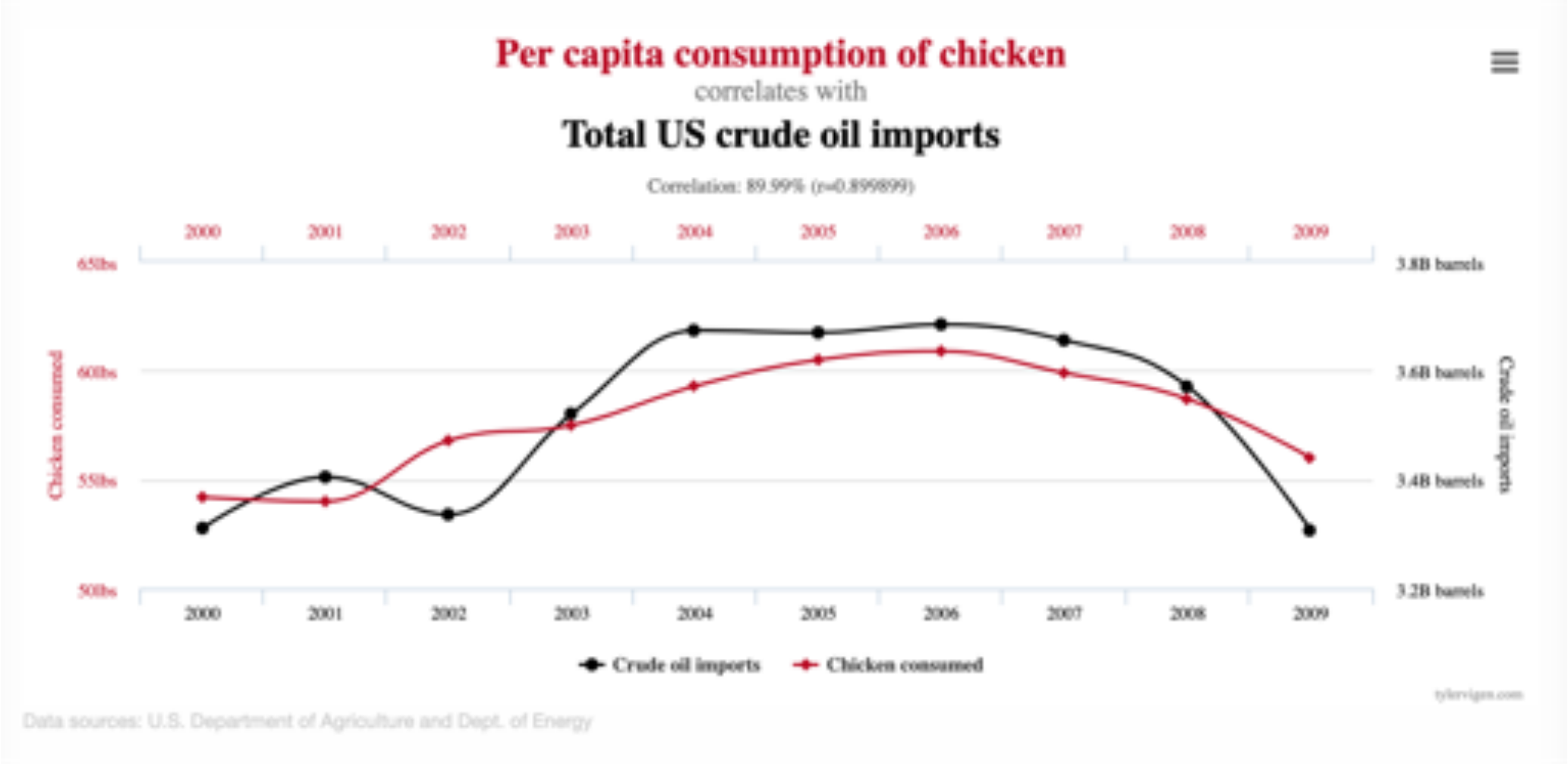


Figure from https://www.tylervigen.com/spurious-correlations

Motivation: Causal Inference

1. What is the efficacy of a given drug in a given population?
2. Whether data can prove an employer guilty of hiring discrimination?
3. What fraction of past crimes could have been avoided by a given policy?
4. What was the cause of death of a given individual, in a specific incident?

Questions from Pearl (2009)

Motivation: Causal Inference

Question:

- Imagine you're a sociologist attempting to understand the factors that have led to the sharp rise in depression amongst U.S. teens in the last half decade.
- Smartphones and social media are purported to be a likely cause, but how could you determine this?
- The gold standard would be data from a randomized control trial (RCT).

What would such an RCT study require of its participants?

Answer:

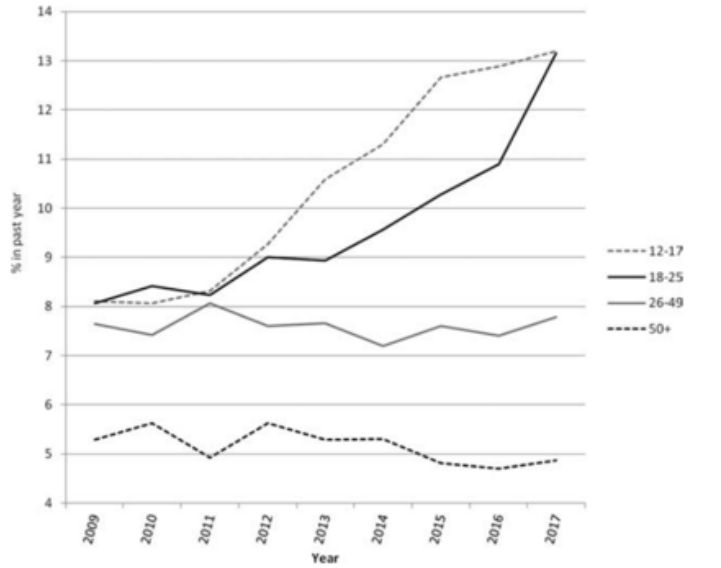


Figure 3. Percent with major depressive episode in the past 12 months, by age group, 2009–2017.

Figure from Twenge et al. (2019)

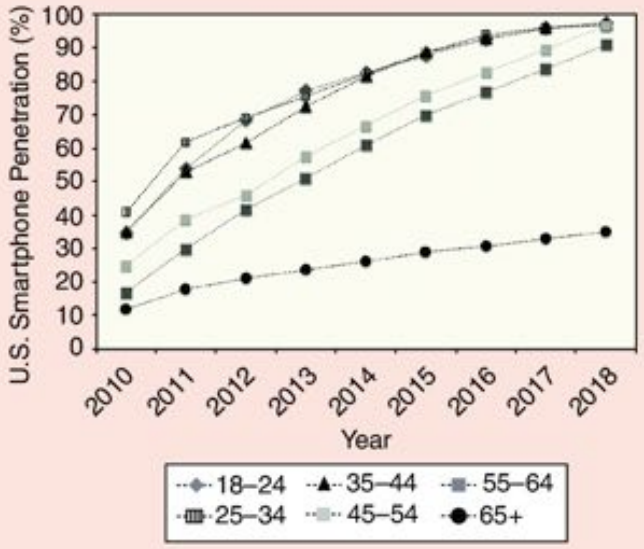


Figure from Berenguer et al. (2017)

Motivation: Causal Inference

Pearl (2018):

> Unlike the rules of geometry, mechanics, optics, or probabilities, the rules of cause and effect have been denied the benefits of mathematical analysis.
>
> To appreciate the extent of this denial readers would likely be stunned to learn that only a few decades ago scientists were unable to write down a mathematical equation for the obvious fact that "Mud does not cause rain." Even today, only the top echelon of the scientific community can write such an equation and formally distinguish "mud causes rain" from "rain causes mud."

Text from Pearl (2018)

Causal Inference

Key Questions:

- What types of questions can causal inference answer?
- What type of data is needed to answer each question?

Statistical vs. Causal Analysis

Whiteboard:

- Statistical vs. Causal Analysis
  - Statistical analysis (goals, assumptions)
  - Causal analysis (goals, assumptions)
- 3-Level Causal Hierarchy
  - Association
  - Intervention
  - Counterfactual

Causal Hierarchy

Figure 1. The causal hierarchy. Questions at level 1 can be answered only if information from level i or higher is available.

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing, Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past two years?

Table from Pearl (2018)

Causal Models

Whiteboard:

- Causal Bayesian Networks
  (i.e. where we are going…)

SCM Inference Engine

Figure 2. How the SCM "inference engine" combines data with a causal model (or assumptions) to produce answers to queries of interest.

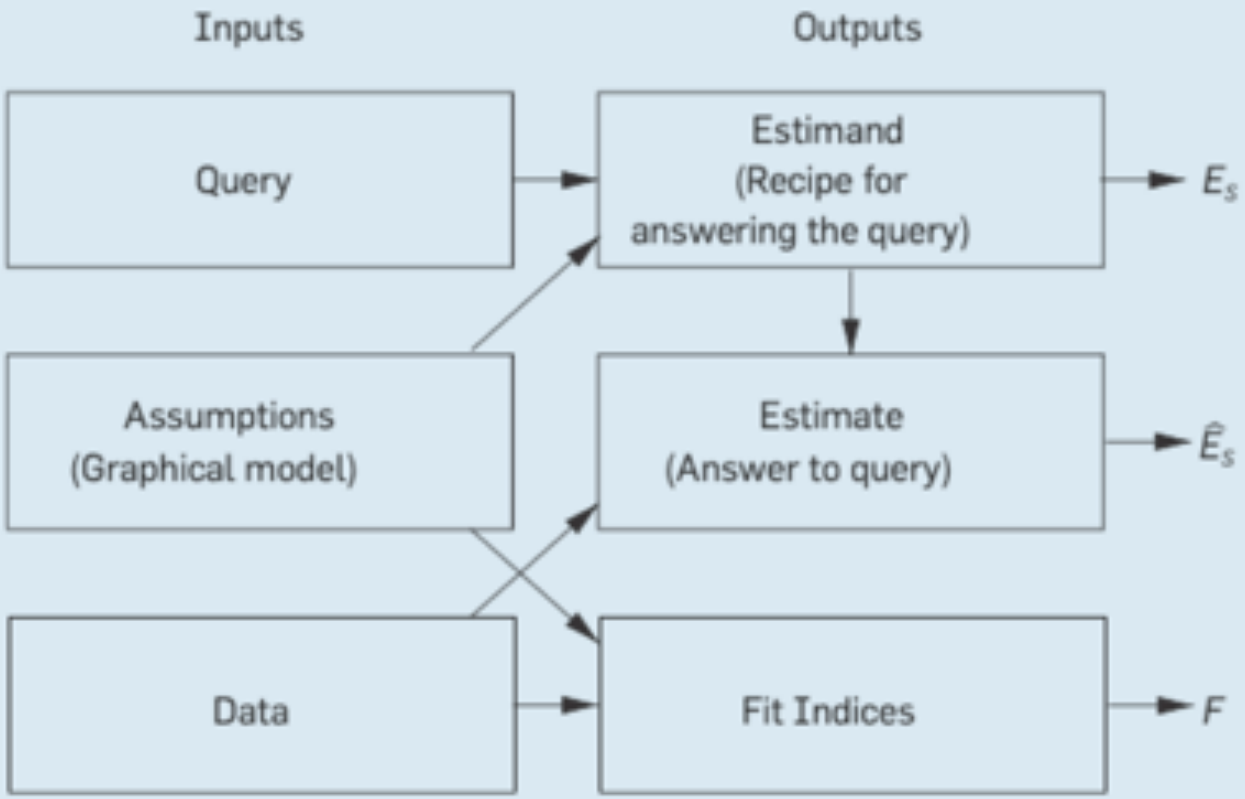


Figure from Pearl (2018)

Causal Models

Whiteboard:

- Structural Causal Models
  - Example: Linear SCM (structural equation model)
  - Example: Nonparametric SCM
  - Intervention
  - Graphical model induced by SCM
- Post-Intervention Distribution vs. Conditional Distribution
- Treatment Efficacy
  - average difference
  - experimental risk ratio

Identification

Identification:

- whether the causal effects are **identifiable**
- **the central question** in analysis of causal effects

Can the post-intervention distribution $p(y \mid do(x_o))$ be estimated by data sampled from the pre-intervention distribution $p(x, y, z)$?

Yes! (Sometimes.)

One very useful case: when the model M is acyclic with all error terms (U_X, U_Y, U_Z) jointly independent, all causal effects are identifiable.

Causal Markov Theorem

Theorem 1 (The Causal Markov Condition). *Any distribution generated by a Markovian model M can be factorized as:*

$$P(v_1, v_2, \ldots, v_n) = \prod_i P(v_i|pa_i) \tag{15}$$

where $V_1, V_2, \ldots, V_n$ are the endogenous variables in M, and pa_i are (values of) the endogenous "parents" of V_i in the causal diagram associated with M.

Corollary 1 (Truncated factorization). *For any Markovian model, the distribution generated by an intervention $do(X = x_0)$ on a set X of endogenous variables is given by the truncated factorization*

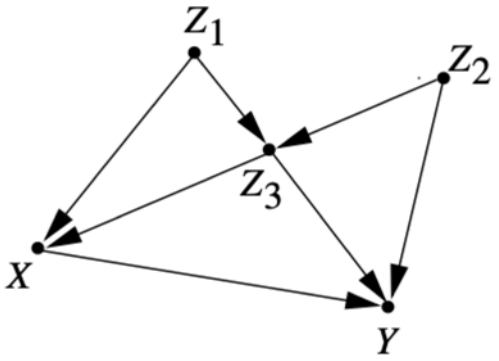
$$P(v_1, v_2, \ldots, v_k|do(x_0)) = \prod_{i|V_i \notin X} P(v_i|pa_i)\,|_{x=x_0} \tag{17}$$

where $P(v_i|pa_i)$ are the pre-intervention conditional probabilities.[8]

Figures from Pearl (2009)

Identification

Example: Model M
(error terms not shown)



1. All of the terms in the post-intervention distribution are from the pre-intervention distribution
2. Those terms could be learned from observational data

Pre-intervention distribution:

$$P(x, z_1, z_2, z_3, y) = P(z_1)P(z_2)P(z_3|z_1, z_2)P(x|z_1, z_3)P(y|z_2, z_3, x)$$

Post-intervention distribution:

$$P(z_1, z_2, z_3, y|do(x_0)) = P(z_1)P(z_2)P(z_3|z_1, z_2)P(y|z_2, z_3, x_0)$$

Causal effect of X on Y:

$$P(y|do(x_0)) = \sum_{z_1, z_2, z_3} P(z_1)P(z_2)P(z_3|z_1, z_2)P(y|z_2, z_3, x_0)$$

Figures from Pearl (2009)

Identification

Identification:

- whether the causal effects are **identifiable**
- **the central question** in analysis of causal effects

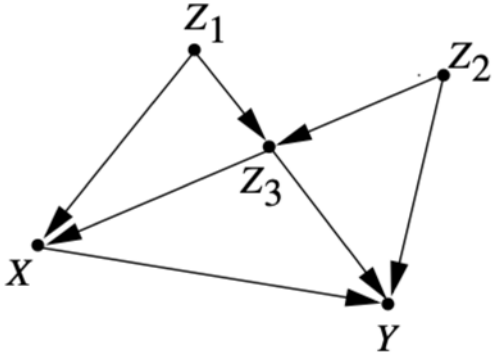
Can the post-intervention distribution $p(y \mid do(x_o))$ be estimated by data sampled from the pre-intervention distribution $p(x, y, z)$?

Yes! (Sometimes.)

One very useful case: when the model M is acyclic with all error terms (U_X, U_Y, U_Z) jointly independent, all causal effects are identifiable.

Unmeasured Confounders

Example: Model M
(error terms not shown)



Suppose in our previous identifiability example, we didn't observe z_2 in our data. Can we still estimate p(y | do(x_0))?

Pre-intervention distribution:

$$P(x, z_1, z_2, z_3, y) = P(z_1)P(z_2)P(z_3|z_1, z_2)P(x|z_1, z_3)P(y|z_2, z_3, x)$$

Post-intervention distribution:

$$P(z_1, z_2, z_3, y|do(x_0)) = P(z_1)P(z_2)P(z_3|z_1, z_2)P(y|z_2, z_3, x_0)$$

Causal effect of X on Y:

$$P(y|do(x_0)) = \sum_{z_1,z_2,z_3} P(z_1)P(z_2)P(z_3|z_1, z_2)P(y|z_2, z_3, x_0)$$

$$P(y|do(x_0)) = \sum_{z_1,z_3} P(z_1)P(z_3|z_1)P(y|z_1, z_3, x_0)$$

Yes! Just marginalize over z_2

Figures from Pearl (2009)

Unmeasured Confounders

- Suppose we wish to measure causal effect of X on Y
- But some *confounding* variables are **unmeasurable** (e.g. genetic trait) and some are **measureable** (e.g. height)
- How to pick an **admissible set** of confounders which, if measured, would enable inference?

Definition 3 (Admissible sets – the back-door criterion). A set S is admissible (or "sufficient") for adjustment if two conditions hold:

1. No element of S is a descendant of X
2. The elements of S "block" all "back-door" paths from X to Y, namely all paths that end with an arrow pointing to X.

Definition 1 (d-separation). A set S of nodes is said to block a path p if either (i) p contains at least one arrow-emitting node that is in S, or (ii) p contains at least one collision node that is outside S and has no descendant in S. If S blocks *all* paths from X to Y, it is said to "d-separate X and Y," and then, X and Y are independent given S, written $X \perp\!\!\!\perp Y|S$.

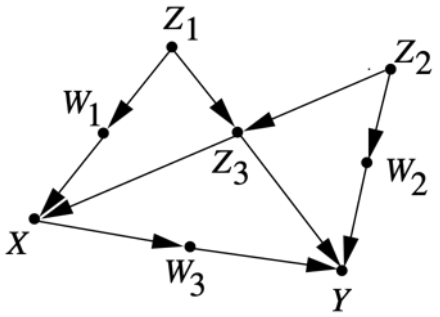
Figures from Pearl (2009)

Unmeasured Confounders

- Suppose we wish to measure causal effect of X on Y
- But some *confounding* variables are **unmeasurable** (e.g. genetic trait) and some are **measureable** (e.g. height)
- How to pick an **admissible set** of confounders which, if measured, would enable inference?

Definition 3 (Admissible sets – the back-door criterion). A set S is admissible (or "sufficient") for adjustment if two conditions hold:

1. No element of S is a descendant of X
2. The elements of S "block" all "back-door" paths from X to Y, namely all paths that end with an arrow pointing to X.



Based on this criterion we see, for example, that the sets $\{Z_1, Z_2, Z_3\}$, $\{Z_1, Z_3\}$, $\{W_1, Z_3\}$, and $\{W_2, Z_3\}$, each is sufficient for adjustment, because each blocks all back-door paths between X and Y. The set $\{Z_3\}$, however, is not sufficient for adjustment because, as explained above, it does not block the path $X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow W_2 \rightarrow Y$.

Figures from Pearl (2009)

EXAMPLE: IDENTIFYING CAUSAL EFFECT

Simpson's Paradox

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

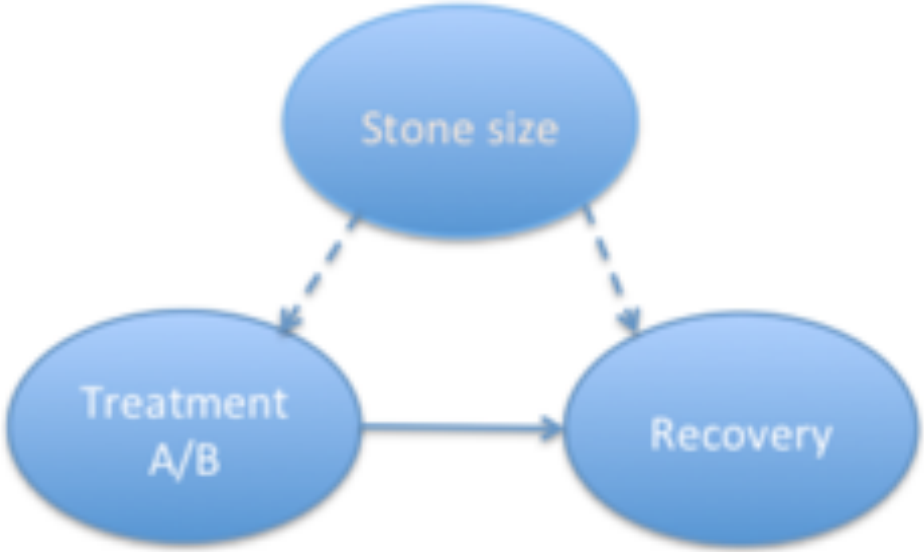
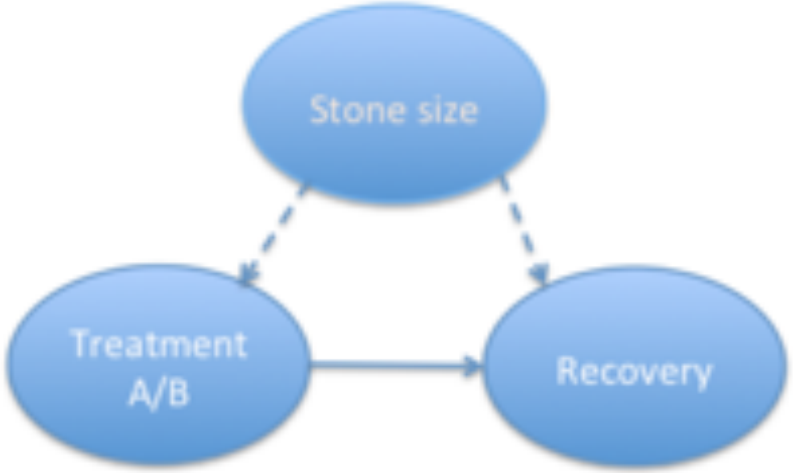


Figure from Kun Zhang's Spring 2019 10-708 Guest Lectures

Identification of Causal Effects

$$P(X3 \mid do\ (X2=1))$$

- "Golden standard": randomized controlled experiments
  - **All the other factors** that influence the outcome variable are either fixed or vary at random, so any changes in the outcome variable must be due to the controlled variable



- Usually expensive or impossible to do!

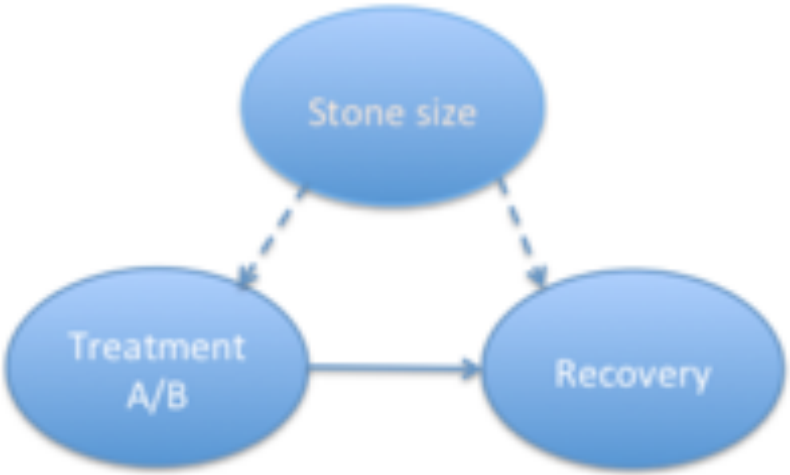
Identification of Causal Effects

Whiteboard:

- Stone-size example:
  - Model 1: path diagram for randomized control trial
  - Model 2: path diagram for observational data
  - Model 3: path diagram for intervention

Identification of Causal Effects: Example

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**



$$P(R|T) = \sum_S P(R|T,S)P(S|T)$$

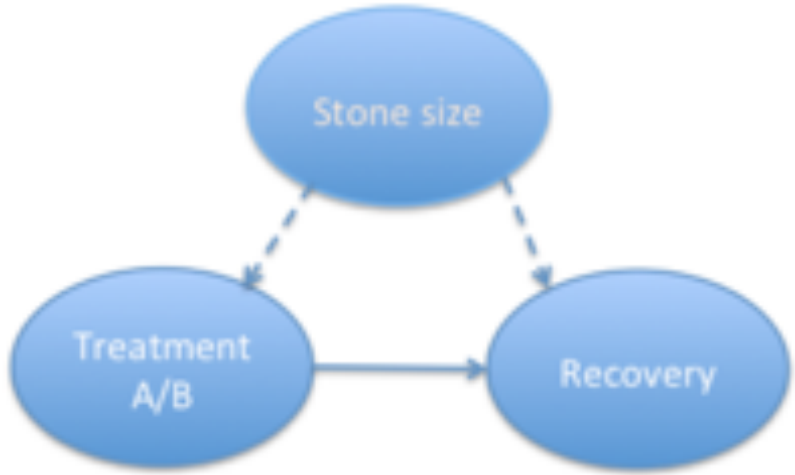
$$P(R\,|\,do(T)) = \sum_S P(R\,|\,T,S)P(S)$$

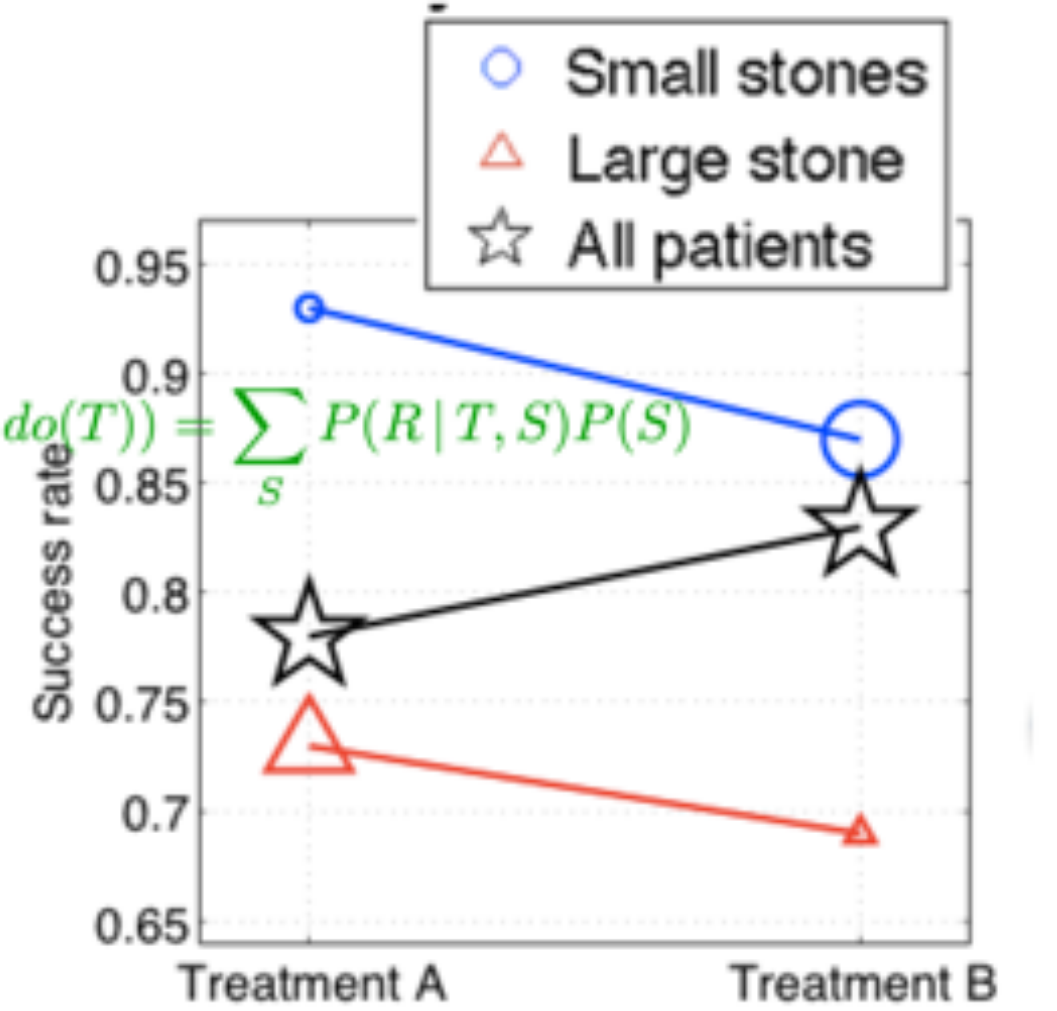
conditioning vs. manipulating

Identification of Causal Effects: Example

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

$$P(R \mid do(T)) = \sum_S P(R \mid T, S) P(S)$$

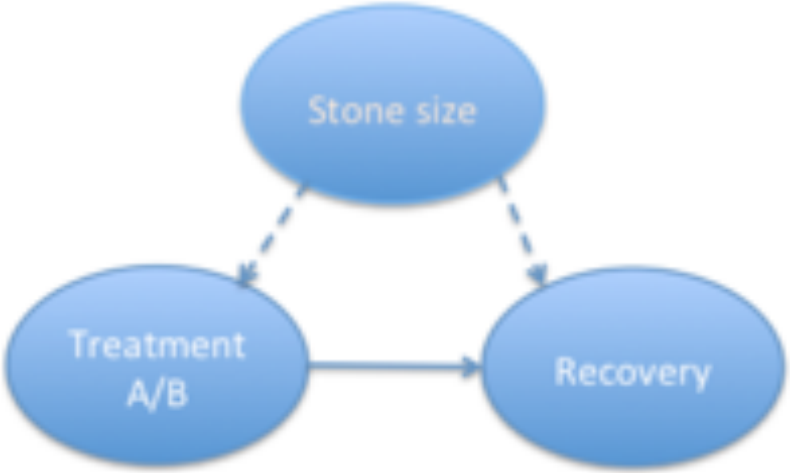




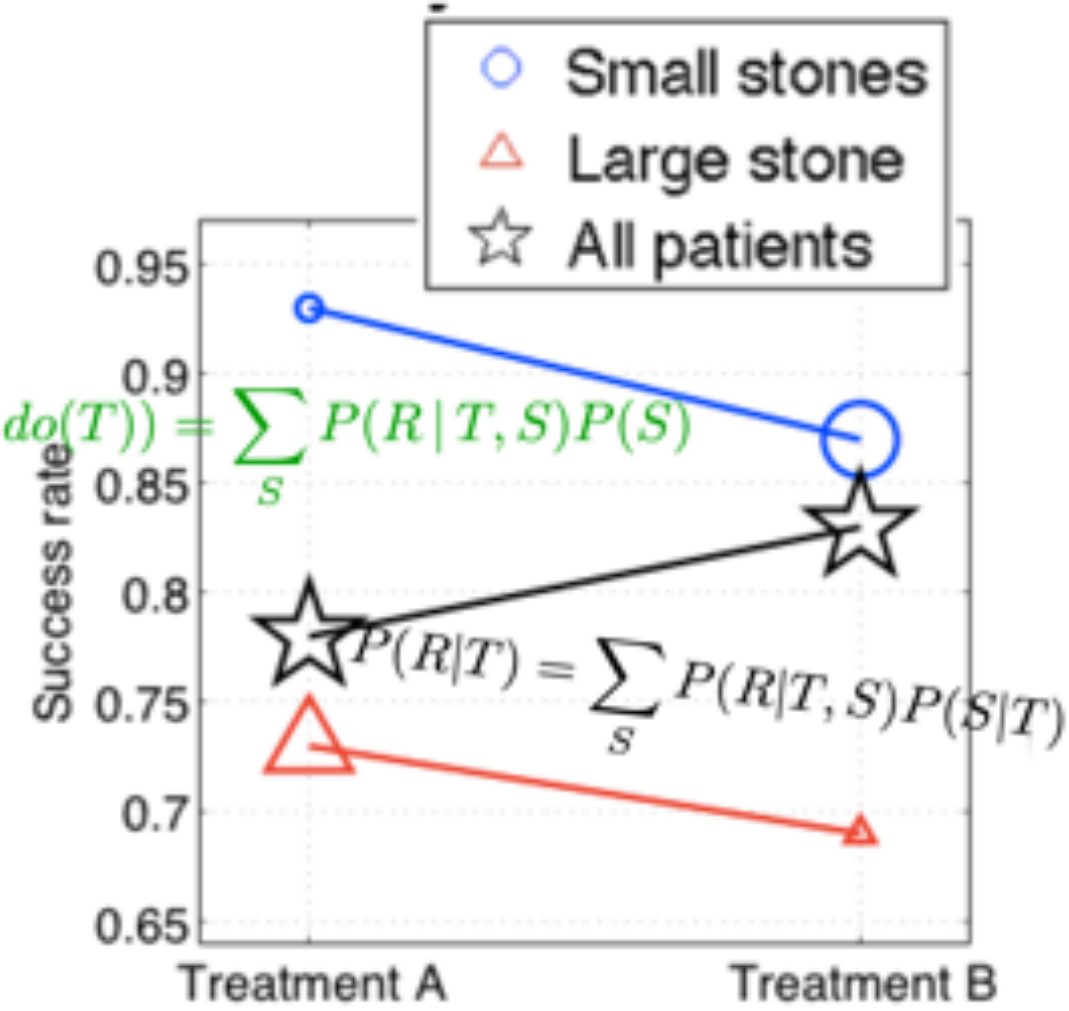
conditioning vs. manipulating

Identification of Causal Effects: Example

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**



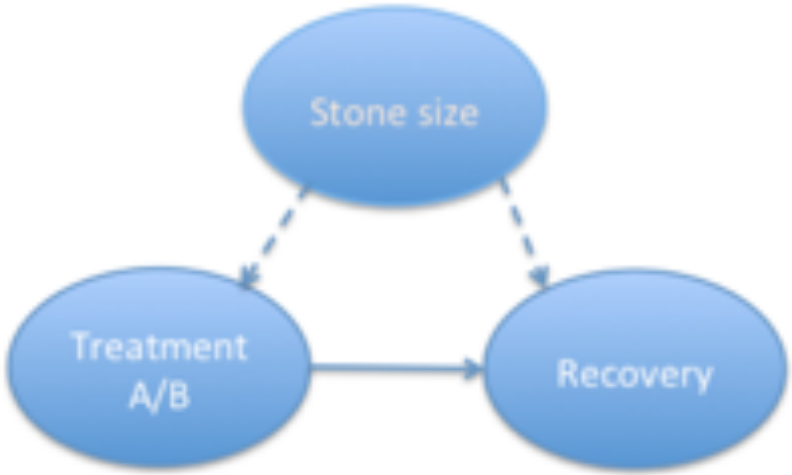
$$P(R \mid do(T)) = \sum_S P(R \mid T, S) P(S)$$



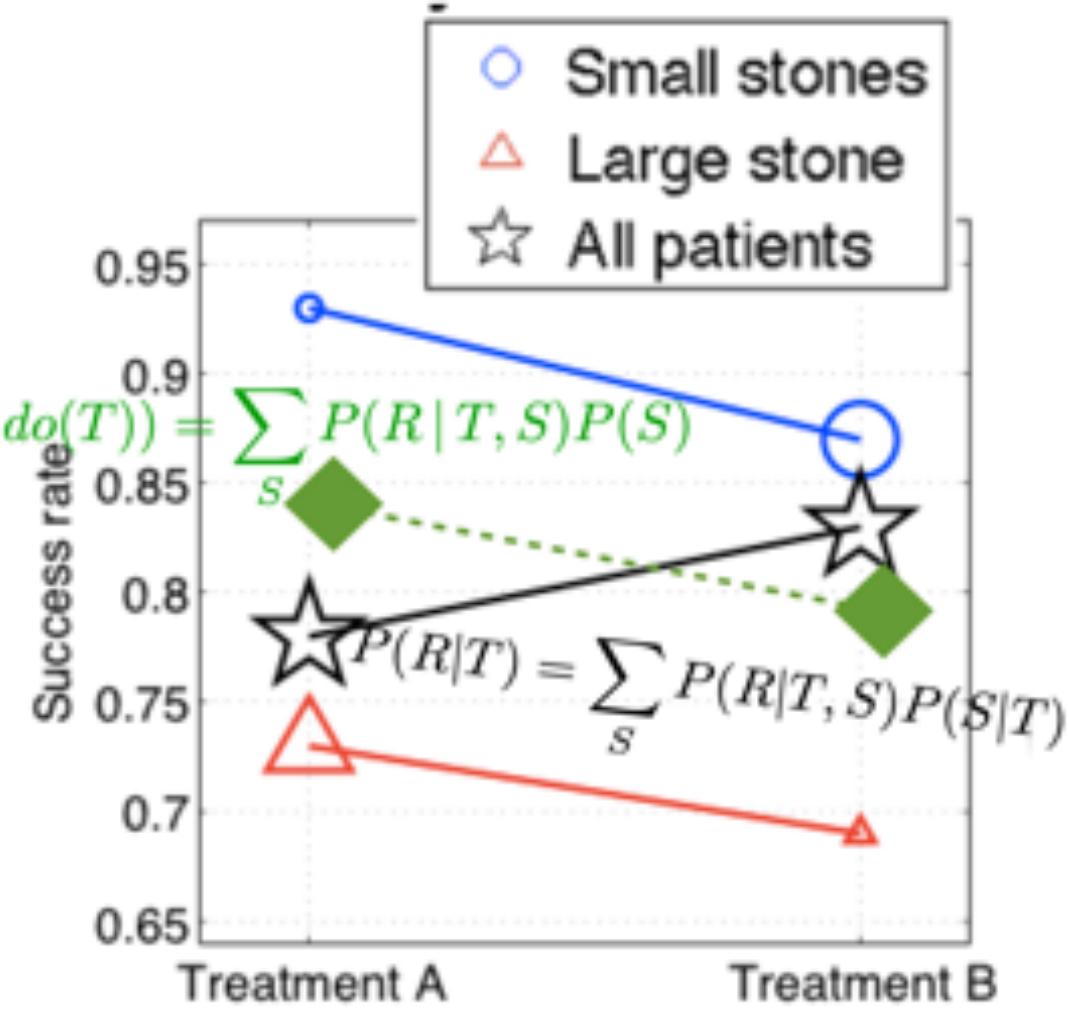
conditioning vs. manipulating

Identification of Causal Effects: Example

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**



$$P(R \mid do(T)) = \sum_S P(R \mid T, S) P(S)$$

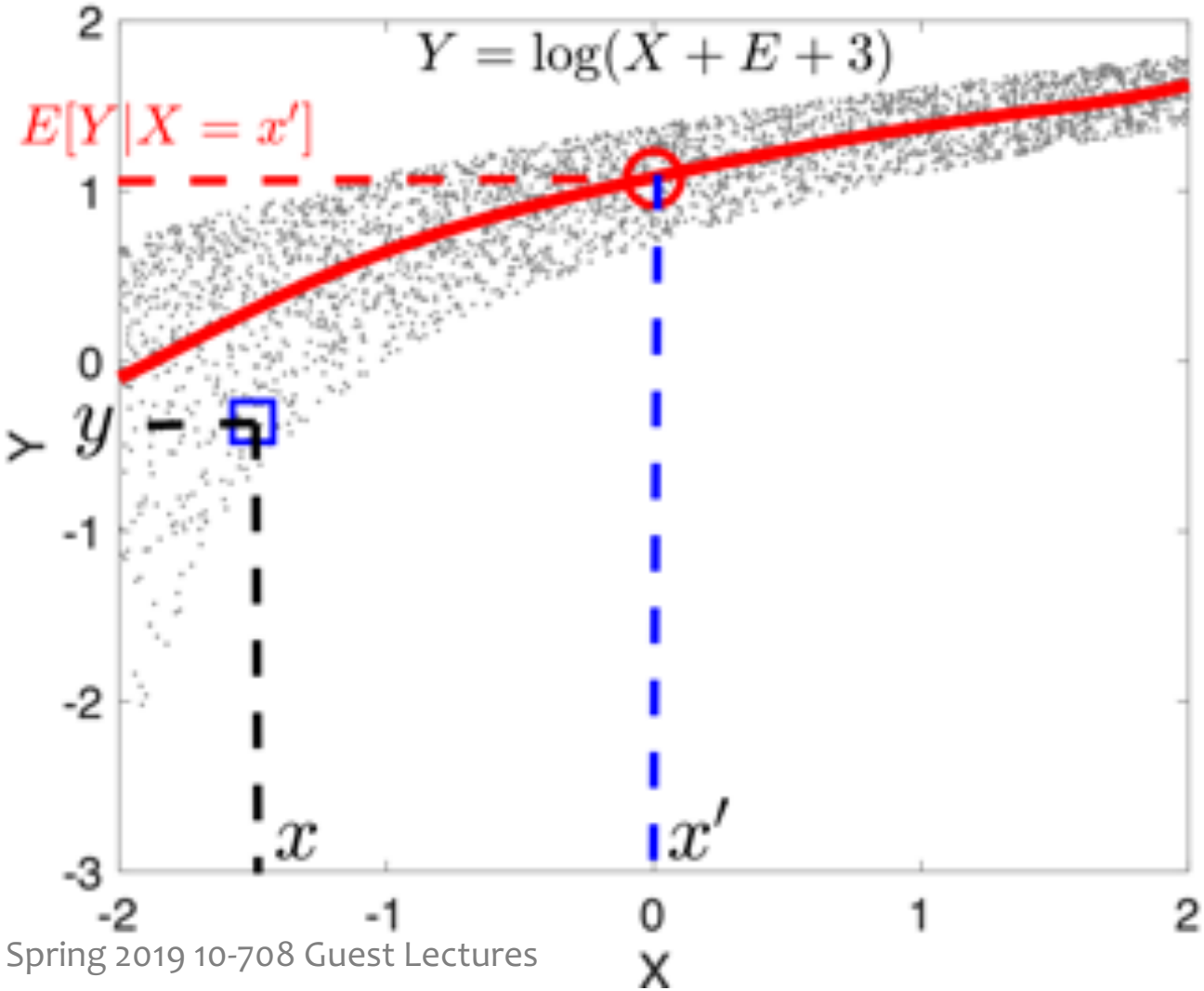


conditioning vs. manipulating

COUNTERFACTUAL INFERENCE

Counterfactual Inference vs. Prediction

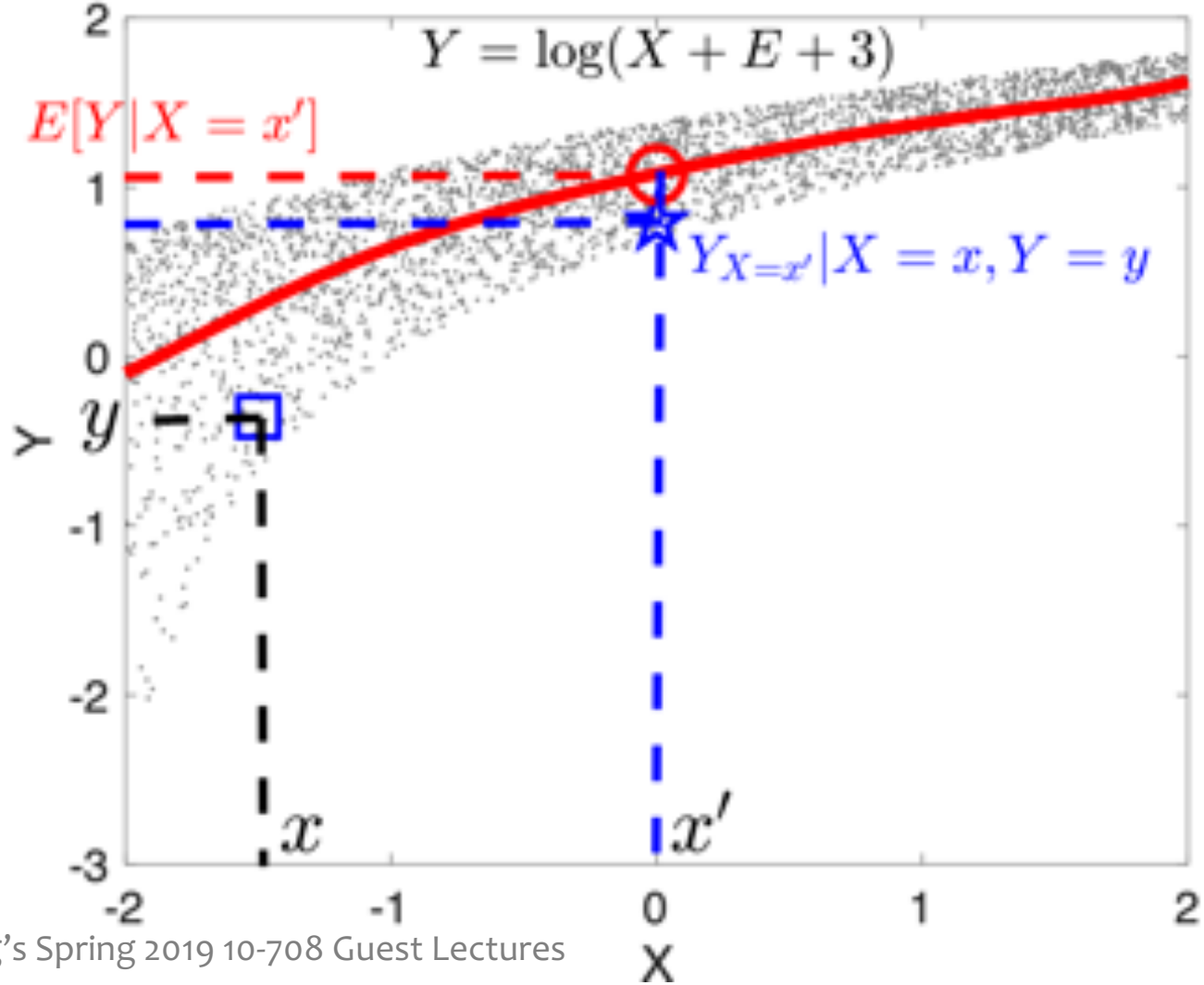
- Suppose X→Y with Y = log(X + E + 3). For an individual with (x,y), what would Y be if X had been x' ?



Slide from Kun Zhang's Spring 2019 10-708 Guest Lectures

Counterfactual Inference vs. Prediction

- Suppose X→Y with Y = log(X + E + 3). For an individual with (x,y), what would Y be if X had been x' ?



Slide from Kun Zhang's Spring 2019 10-708 Guest Lectures

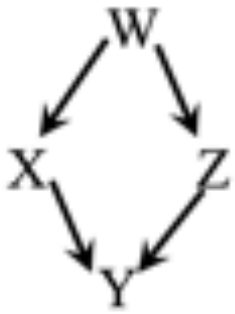
Standard Counterfactual Questions

- We talk about a particular situation (or unit) $U = u$, in which $X = x$ and $Y = y$
- What value would Y be had X been x' in situation u? I.e., we want to know $Y_{X=x'}(u)$, the value of Y in situation u if we do($X=x'$)
- u is not directly observable, so $P(Y_{X=x'} \mid X = x, Y = y)$ instead

For identification of causal effects, U is randomized. It is fixed for counterfactual inference.

Slide from Kun Zhang's Spring 2019 10-708 Guest Lectures

Counterfactual Inference



$$W = U_W$$
$$X = f_X(W, U_X)$$
$$Z = f_Z(W, U_Z)$$
$$Y = f_Y(X, Z, U_Z)$$

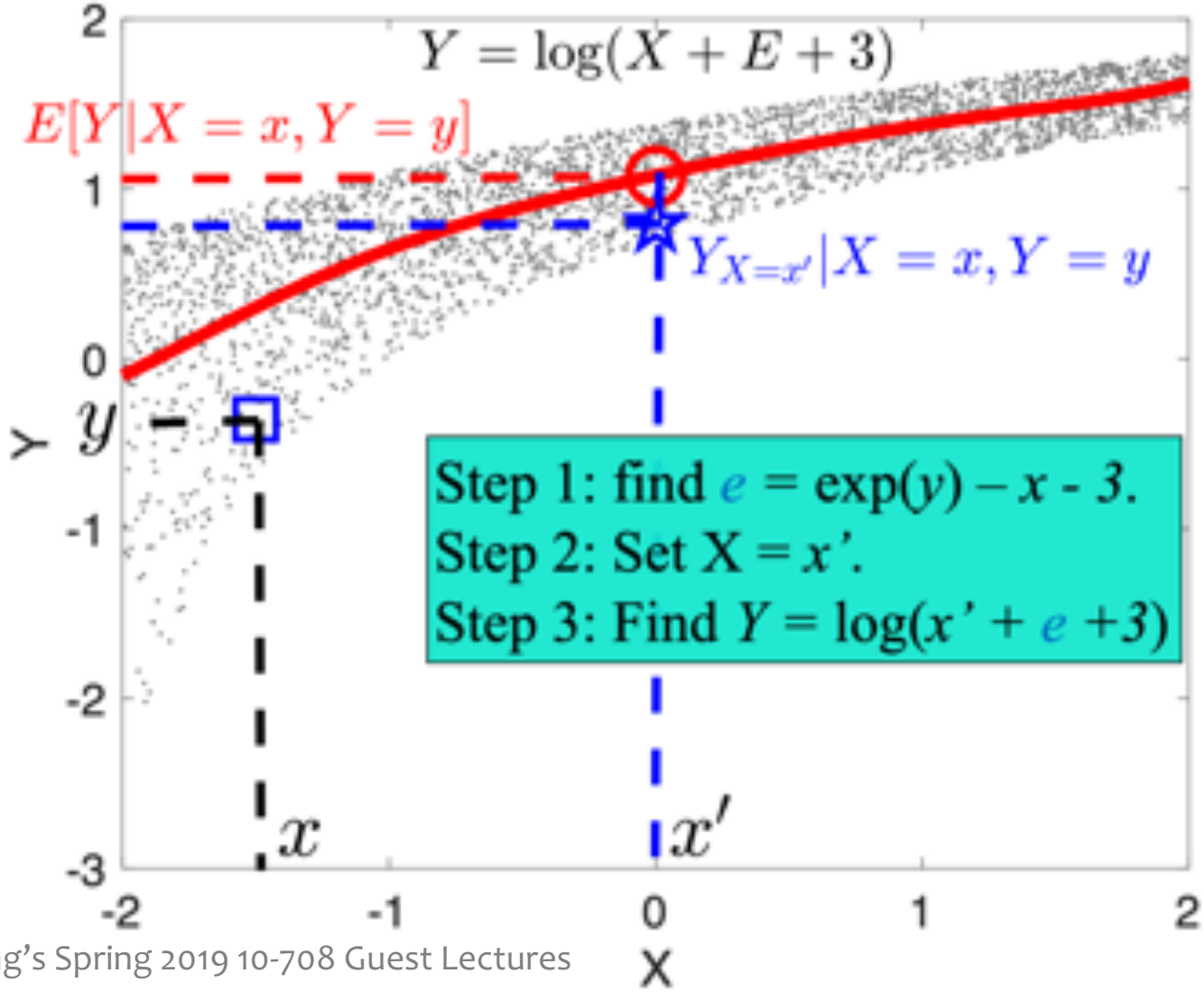
$$P(Y_{X=x'} \mid \underbrace{X = x,\ Y = y,\ W = w}_{\textit{evidence}})$$

- Three steps
  - Abduction: find P(U | *evidence*)
  - Action: Replace the equation for X by $X = x'$
  - Prediction: Use the modified model to predict Y

Slide from Kun Zhang's Spring 2019 10-708 Guest Lectures

Counterfactual Inference vs. Prediction

- Suppose X→Y with Y = log(X + E + 3). For an individual with (x,y), what would Y be if X had been x' ?



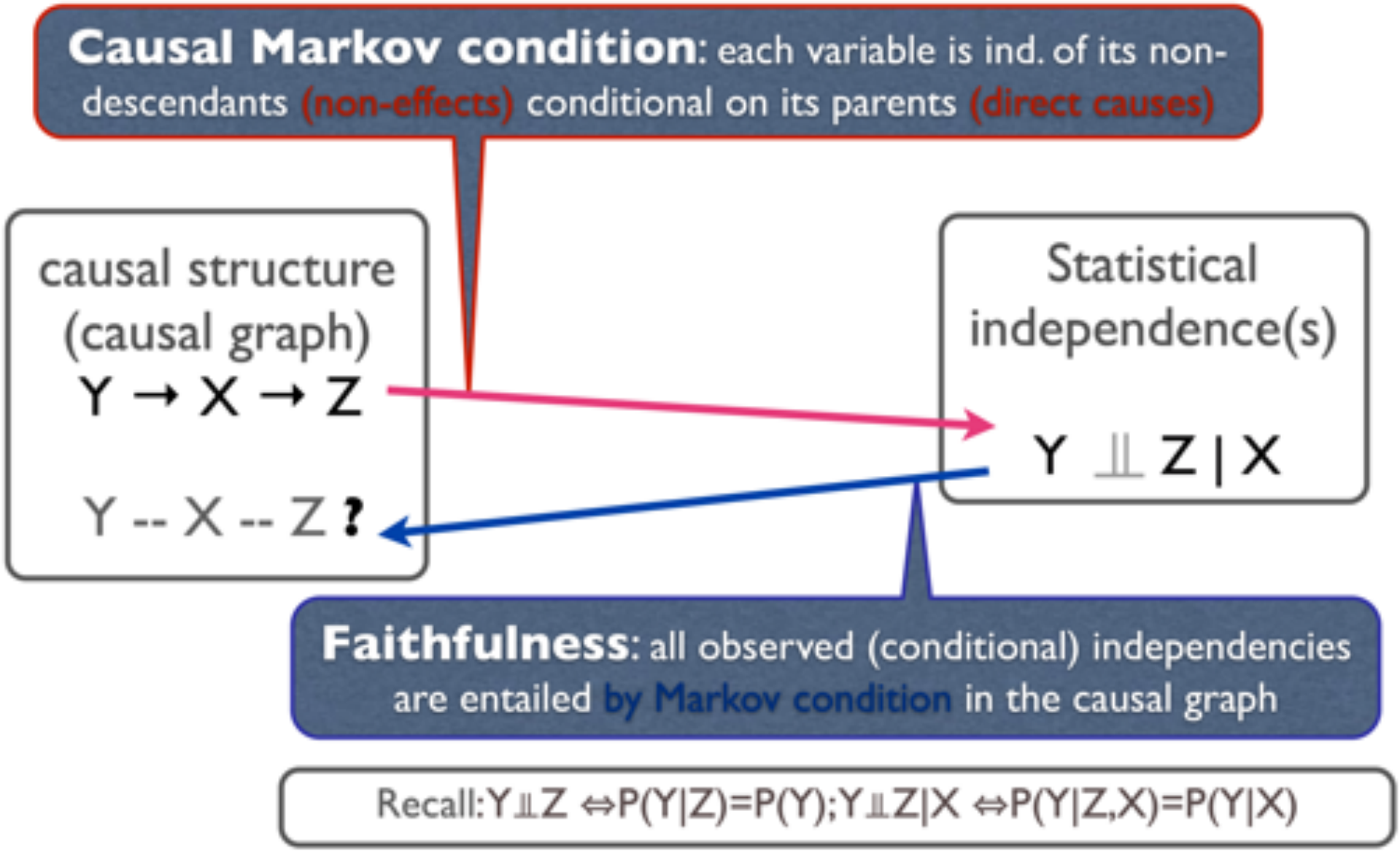
Slide from Kun Zhang's Spring 2019 10-708 Guest Lectures

CAUSAL DISCOVERY

Causal Discovery

- Goal:
  - Find a path diagram (i.e. causal model) that is best supported by the data
- Key Idea:
  - find causal structures that are consistent (in a d-separation sense) with the set of conditional independencies supported by the data
- Where to learn more?
  - Kun Zhang (CMU, Philosophy / ML) guest lectures from Spring 2020 10-708: http://www.cs.cmu.edu/~epxing/Class/10708-20/lectures.html

Causal Structure vs. Statistical Independence (SGS, et al.)

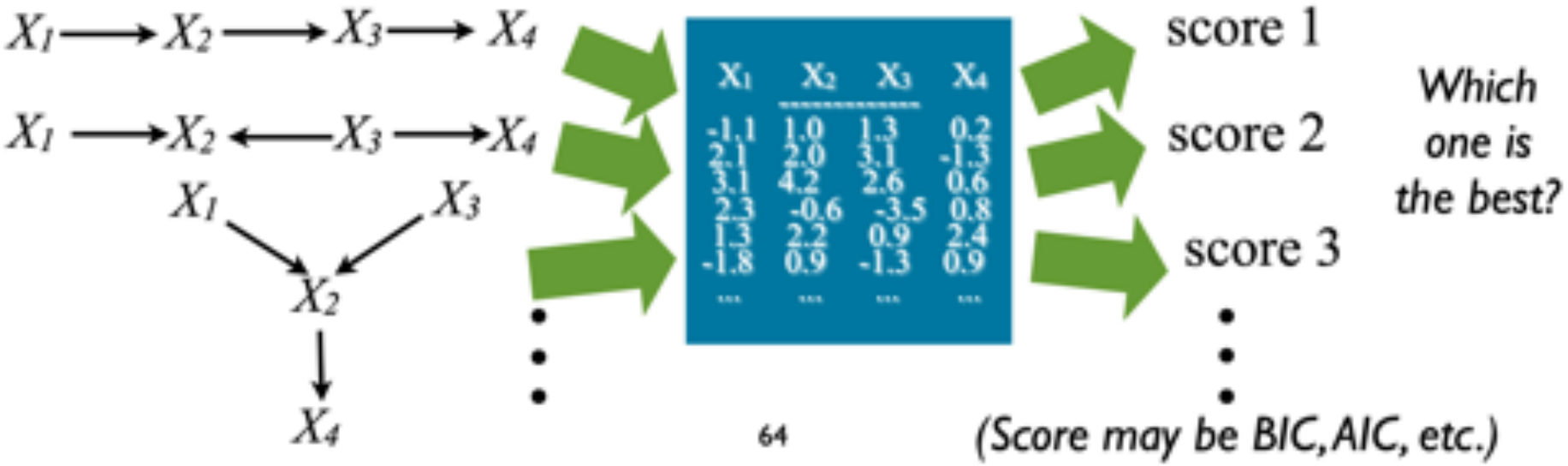


Slide from Kun Zhang's Spring 2019 10-708 Guest Lectures

Constraint-Based vs. Score-Based

- Constraint-based methods

- Score-based methods



Slide from Kun Zhang's Spring 2019 10-708 Guest Lectures

RL AS INFERENCE

A note on materials used in this module

- Sutton & Barto. Reinforcement Learning: An Introduction. 2nd edition.
- David Silver's UCL course on reinforcement learning.
- Materials from UC Berkeley's Deep RL course.
- Sergey Levine's tutorial on RL and control as inference.
- Brian Ziebart's PhD thesis (maximum causal entropy models).

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

Paradigms of machine learning

- Supervised learning

  Given: a collection of data $D = \{(x_i, y_i)\}_{i=1}^N$

  Goal: learn a model that approximates $P(y \mid x)$

- Unsupervised learning

  Given: a collection of data $D = \{(x_1, x_2, \dots, x_d)_i\}_{i=1}^N$

  Goal: learn a model that approximates $P(x_1, x_2, \dots, x_d)$

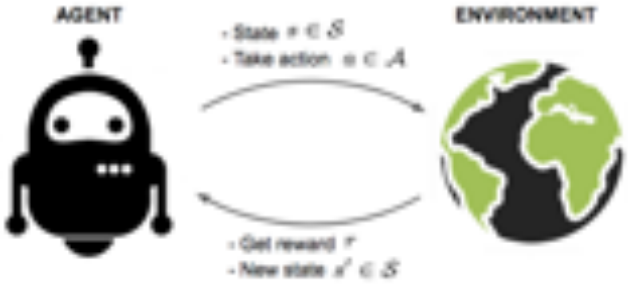
- Reinforcement learning

  Given: an environment that an agent can perceive and interact with

  Goal: learn a controller (policy) that can maximize the utility (reward) in the given environment

GMs allow us to efficiently represent, manipulate, and perform learning and inference on these probabilistic models.

DL gives the tools for learning expressive latent representations that lead to more accurate probabilistic models of the data.

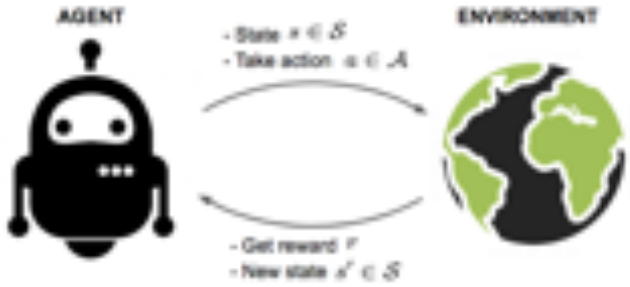


Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

Why sequential decision making and RL?

Ultimately, we want to build autonomous intelligent machines that:

- Can perceive and interact with the world
- Exhibit purposeful goal-directed behavior
- Learn from interactions, adapt to changes, plan and be able to maximize utility functions (specified by humans or inferred from situations)



RL gives us a formal framework for building such autonomous agents.

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

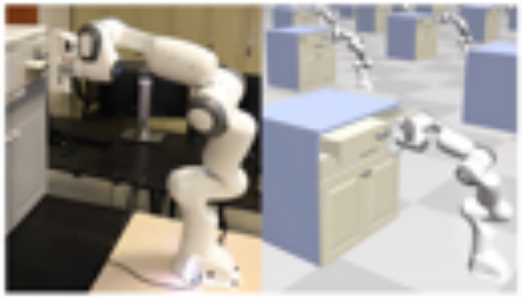
Some recent success stories of RL

Learning to play games

AlphaGo (DeepMind, 2016)

AlphaStar (DeepMind, 2019)

Robotics



Chebotar et al., 2018

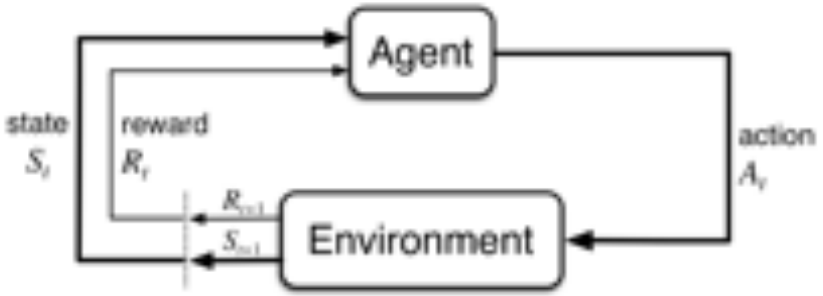
OpenAI, 2018

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

BASIC CONCEPTS IN RL

Markov Decision Processes (MDPs)

Markov Decision Process (MDP):



- Environment has a set of states $\mathcal{S}$
- Agent is given a set of possible actions $\mathcal{A}$
- Environment dynamics: transitions from state s_t into a new state s_{t+1} according to the transition probability $P(s_{t+1}|s_t, a_t)$ after agent takes action a_t
- Reward function: $\mathrm{r}(s, a) = \mathbb{E}[r_{t+1}|s_t = s, a_t = a]$ provides scalar rewards to the agent at each time step
- "Life" of an agent (or trajectory):

$$\tau = (s_1, a_1, r_1, s_2, a_2, r_2, s_3, a_3, r_3, \dots)$$

Sutton & Barto (2018). Reinforcement learning: An introduction. 2nd edition.

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

Markov Decision Processes (MDPs)

What can we do with MDPs:

(1) Policy search: Find a policy $\pi: \mathcal{S} \rightarrow \mathcal{A}$ that outputs actions for each given state such that the cumulative reward along the trajectory is maximized.

(2) Inverse RL: Given a set of optimal trajectories (e.g., generated by a human expert), infer the corresponding MDP.

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

Returns and Episodes

Maximization of the return:

- Return (cumulative reward) starting step t: $G_t = r_{t+1} + r_{t+2} + \cdots + r_T$
- If $T = \infty$, we can use the notion of discounted return:

$$
\begin{aligned}
G_t &= r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots \\
&= \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \\
&= r_{t+1} + \gamma G_{t+1}
\end{aligned}
$$

where $0 \leq \gamma \leq 1$ is called the discount rate

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

Policies and Value Functions

- Value function of a state s:

$$V_\pi(s) := \mathbb{E}_\pi \left[G_t \mid s_t = s \right] = \mathbb{E}_\pi \left[\sum_{k=0}^{T} \gamma^k r_{t+k+1} \mid s_t = s \right]$$

- Value function of the state-action pair (s, a):

$$Q_\pi(s, a) := \mathbb{E}_\pi \left[G_t \mid s_t = s, a_t = a \right]$$

$$= \mathbb{E}_\pi \left[\sum_{k=0}^{T} \gamma^k r_{t+k+1} \mid s_t = s, a_t = a \right]$$

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

Bellman Equation for $Q_\pi(s, a)$

- Bellman equation for the value function of the state-action pair (s, a):

$$Q_\pi(s,a) = r(s,a) + \gamma \sum_{s'} p(s' \mid s,a) \sum_{a'} \pi(a' \mid s') Q_\pi(s',a')$$

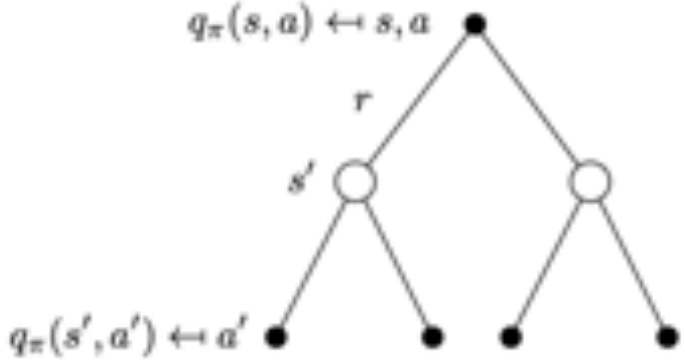


Illustration from David Silver's lecture.

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

Optimal Policies and Value Functions

- Solving an RL task means finding an optimal policy that achieves high reward in the long run.
- Policy π is better or equal to π' ($\pi \geq \pi'$) if its expected return is greater or equal to that of π' for all states:

$$\pi \geq \pi' \Leftrightarrow V_\pi(s) \geq V_{\pi'}(s) \; \forall s \in \mathcal{S}$$

- Optimal value functions (Bellman optimality):

$$V_\star(s) := \max_\pi V_\pi(s) = \max_a \sum_{s'} p(s' \mid s, a) \left[r(s, a) + \gamma V_\star(s') \right]$$

$$Q_\star(s, a) := \max_\pi Q_\pi(s, a) = \sum_{s'} p(s' \mid s, a) \left[r(s, a) + \gamma \max_{a'} Q_\star(s', a') \right]$$

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

How to recover optimal policy and trajectories?

- We can recover an optimal policy from the optimal $Q_\star(s,a)$:

$$\pi_\star(a \mid s) = \delta\left(a = \arg\max_a Q_\star(s,a)\right)$$

- To recover a set of optimal trajectories, just execute the optimal policy:

$$\tau_\star = (s_1^\star, a_1^\star, r_1^\star, s_2^\star, a_2^\star, r_2^\star, \dots)$$
$$s_{t+1}^\star \sim p(s_{t+1} \mid s_t, a_t^\star = \arg\max_a Q_\star(s,a))$$

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

Example: Grid World and an Optimal Policy

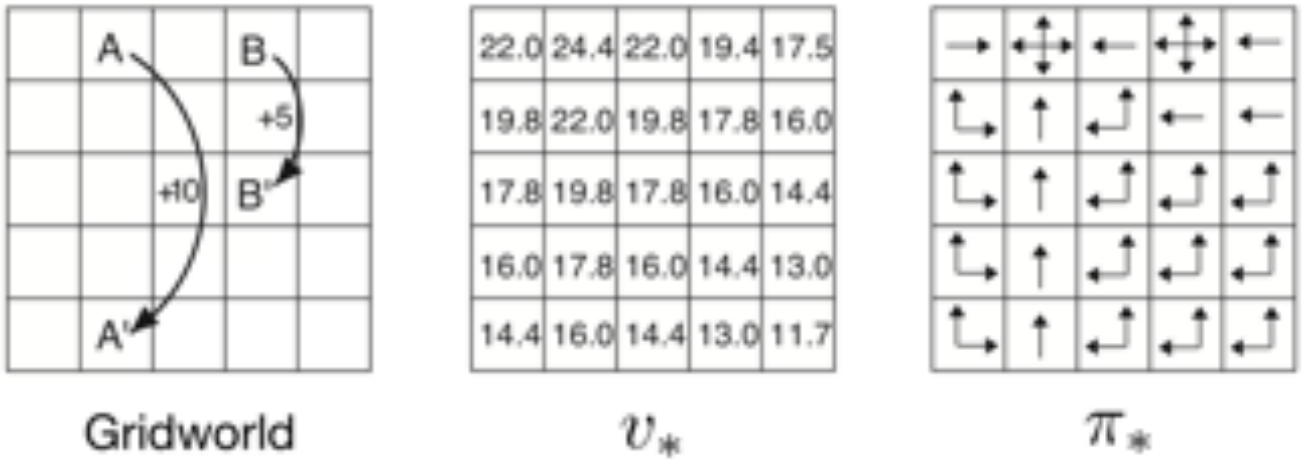


Figure 3.5: Optimal solutions to the gridworld example.

Sutton & Barto (2018). Reinforcement learning: An introduction. 2nd edition.

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

Recap

Initial state	$s_0 \sim p_0(s)$
Transition	$s_{t+1} \sim p(s_{t+1} \mid s_t, a_t)$
Policy	$a_t \sim \pi(a_t \mid s_t)$
Reward	$r_t = r(s_t, a_t)$

- Value functions:

$$V_\pi(s) := \mathbb{E}_\pi \left[\sum_{k=0}^{T} \gamma^k r_{t+k+1} \mid s_t = s \right]$$

$$Q_\pi(s,a) := \mathbb{E}_\pi \left[\sum_{k=0}^{T} \gamma^k r_{t+k+1} \mid s_t = s, a_t = a \right]$$

- Recursive notion of optimality:

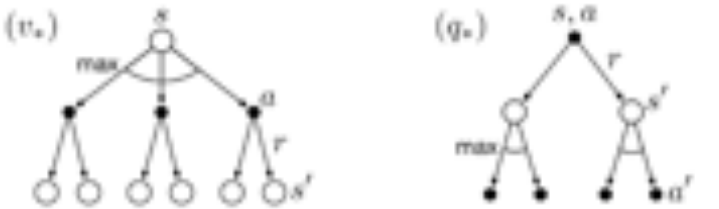
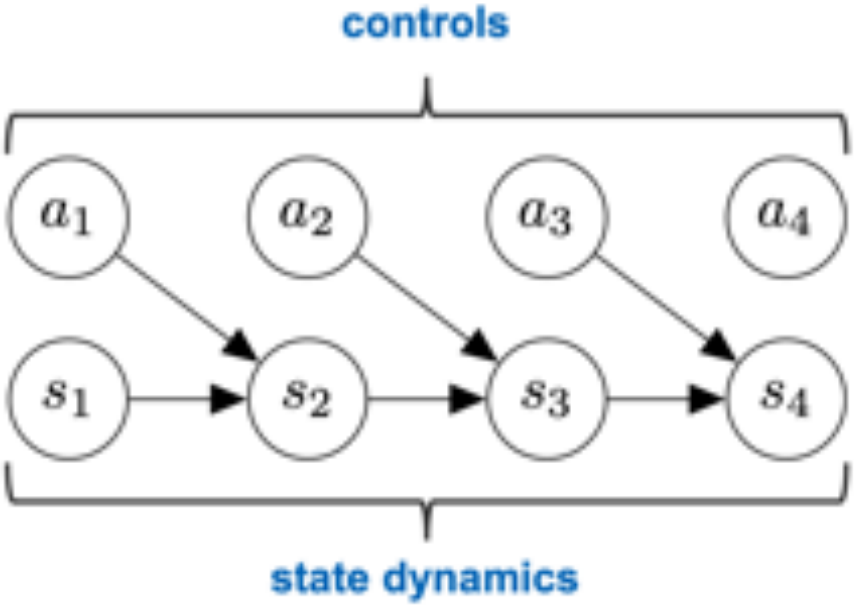


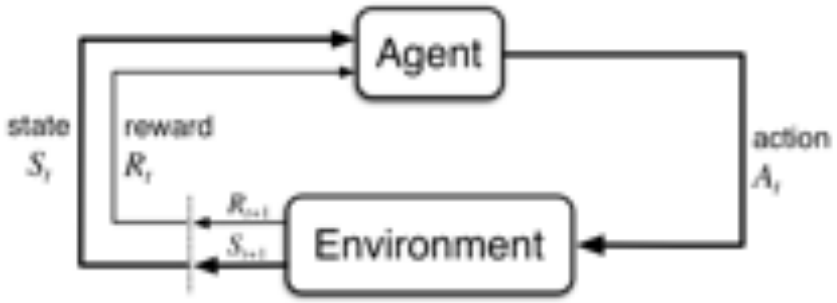
Figure 3.4: Backup diagrams for v_* and q_*

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

RL & CONTROL AS INFERENCE IN A GRAPHICAL MODEL

MDP as a Graphical Model

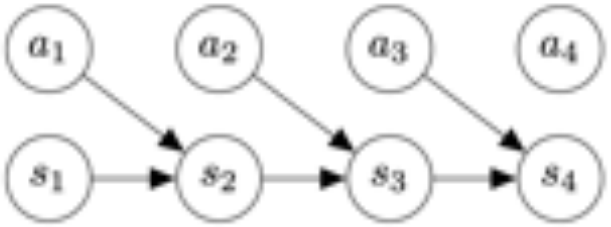




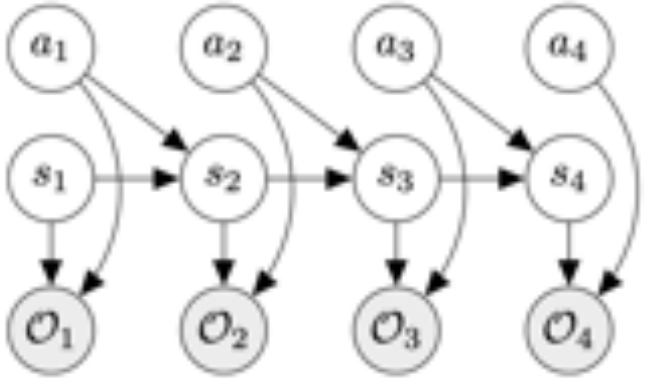
How do we define a distribution over rational/optimal trajectories?

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

MDP as a Graphical Model



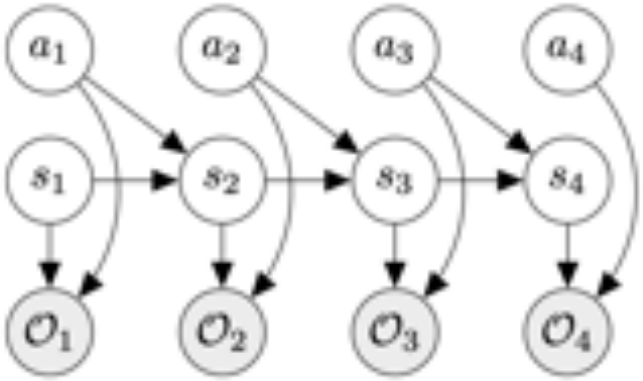
Initial state	$s_0 \sim p_0(s)$
Transition	$s_{t+1} \sim p(s_{t+1} \mid s_t, a_t)$
Policy	$a_t \sim \pi(a_t \mid s_t)$
Reward	$r_t = r(s_t, a_t)$



Initial state	$s_0 \sim p_0(s)$
Transition	$s_{t+1} \sim p(s_{t+1} \mid s_t, a_t)$
Policy	$a_t \sim \pi(a_t \mid s_t)$
Reward	$r_t = r(s_t, a_t)$
Optimality	$p(\mathcal{O}_t = 1 \mid s_t, a_t) = \exp(r(s_t, a_t))$

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

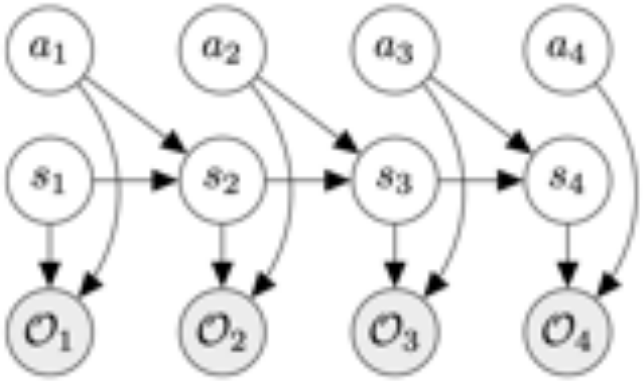
Why is this model interesting?



- Can solve control and planning problems using inference algorithms
- Allows to model suboptimal behavior (important for inverse RL)
- Provides and explanation for why stochastic behavior might be preferred (from the exploration and transfer learning point of view)

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

What can we do with this graphical model?

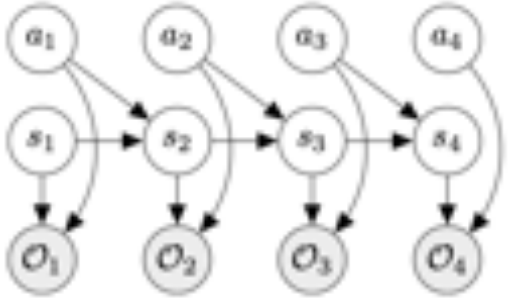


Here is what we can do:

- Given a reward, determine a likely optimal trajectory
- Given a collection of optimal trajectories, infer the reward and priors
- Given a reward, infer the optimal policy

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

Distribution over the optimal trajectories



$$p(\mathcal{O}_t \mid s_t, a_t) = \exp(r(s_t, a_t))$$

$$p(\tau \mid \mathcal{O}_{1:T}) \propto p(s_1) \prod_{t=1}^{T} p(a_t \mid s_t) p(s_{t+1} \mid s_t, a_t) p(\mathcal{O}_t \mid s_t, a_t)$$

Levine (2018). Reinforcement learning and control as probabilistic inference.

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

Inferring the reward & prior that generate trajectories

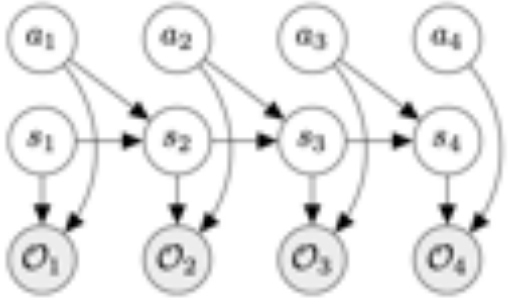
$$p(\tau \mid \mathcal{O}_{1:T}, \theta, \phi) \propto \left[p(s_1) \prod_{t=1}^{T} p(s_{t+1} \mid s_t, a_t) \right] \exp\left(\sum_{t=1}^{T} r_\phi(s_t, a_t) + \log p_\theta(a_t \mid s_t) \right)$$

The model reminds a featurized CRF. (*Note: CRF is undirected and does not preserve the causal structure; this model is more restrictive and known as MEMM.)

Ziebart (2010). Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy. PhD thesis.

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

Optimal policy and planning via inference



- Unroll the dynamics and compute backward messages:

$$\beta_t(s_t, a_t) := p(\mathcal{O}_{t:T} \mid s_t, a_t)$$

- Compute optimal policy:

$$p(a_t \mid s_t, \mathcal{O}_{t:T})$$

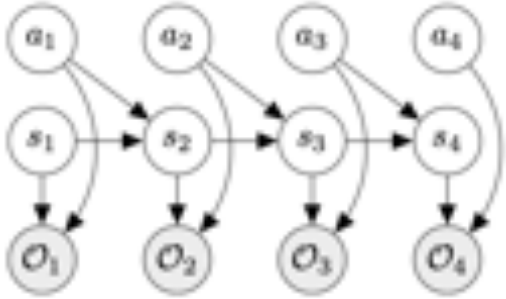
- Compute forward messages (state filtering under optimality constraint):

$$\alpha_t(s_t) = p(s_t \mid \mathcal{O}_{1:t-1})$$

Levine (2018). Reinforcement learning and control as probabilistic inference.

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

Backward messages



Probability that we can be optimal at steps t through T given s_t and a_t

$$\beta_t(s_t, a_t) = p(\mathcal{O}_{t:T} \mid s_t, a_t)$$

$$= \int_{\mathcal{S}} p(\mathcal{O}_{t:T}, s_{t+1} \mid s_t, a_t) ds_{t+1}$$

$$= \int_{\mathcal{S}} p(\mathcal{O}_{t+1:T} \mid s_{t+1}) p(s_{t+1} \mid s_t, a_t) p(\mathcal{O}_t \mid s_t, a_t) ds_{t+1}$$

$$\underbrace{p(\mathcal{O}_{t+1:T} \mid s_{t+1})}_{\beta_{t+1}(s_{t+1})} = \int_{\mathcal{A}} p(\mathcal{O}_{t+1:T} \mid s_{t+1}, a_{t+1}) p(a_{t+1} \mid s_{t+1}) da_{t+1}$$

for $t = T - 1$ to 1 :

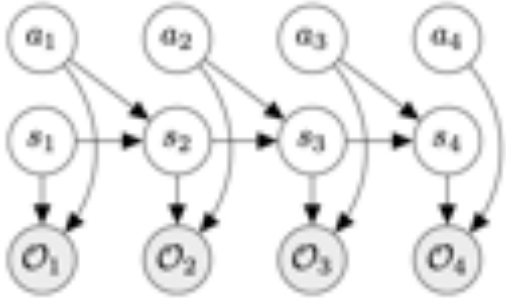
$$\beta_t(s_t, a_t) = p(\mathcal{O}_t \mid s_t, a_t) \mathbb{E}_{s_{t+1} \sim p(s_{t+1} \mid s_t, a_t)} \left[\beta_{t+1}(s_{t+1}) \right]$$

$$\beta_t(s_t) = \mathbb{E}_{a_t \sim p(a_t \mid s_t)} \left[\beta_t(s_t, a_t) \right]$$

Levine (2018). Reinforcement learning and control as probabilistic inference.

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

How are these messages related to RL?



let $V_t(\mathbf{s}_t) = \log \beta_t(\mathbf{s}_t)$

let $Q_t(\mathbf{s}_t, \mathbf{a}_t) = \log \beta_t(\mathbf{s}_t, \mathbf{a}_t)$ $\qquad V(\mathbf{s}_t) = \log \int \exp(Q(\mathbf{s}_t, \mathbf{a}_t) + \log p(\mathbf{a}_t|\mathbf{s}_t))\mathbf{a}_t$

Deterministic dynamics: $Q(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + V(\mathbf{s}_{t+1})$

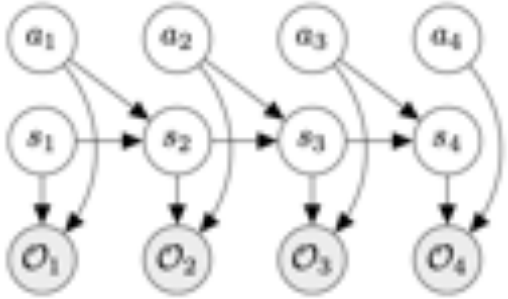
Stochastic dynamics: $Q(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \log E_{\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)}[\exp(V(\mathbf{s}_{t+1}))]$

"optimistic" transition (not good)

Levine (2018). Reinforcement learning and control as probabilistic inference.

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

Optimal policy



$$\pi(\mathbf{a}_t|\mathbf{s}_t) = \frac{\beta_t(\mathbf{s}_t, \mathbf{a}_t)}{\beta_t(\mathbf{s}_t)}$$

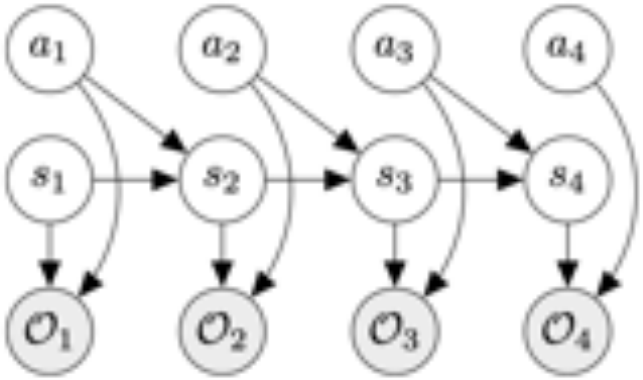
$$\pi(\mathbf{a}_t|\mathbf{s}_t) = \exp(Q_t(\mathbf{s}_t, \mathbf{a}_t) - V_t(\mathbf{s}_t)) = \exp(A_t(\mathbf{s}_t, \mathbf{a}_t))$$

- (Derivation: exercise!)
- Natural interpretation: better actions are more probable + random tie breaking
- Approaches greedy policy as temperature decreases

Levine (2018). Reinforcement learning and control as probabilistic inference.

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures

Summary



- Using auxiliary potentials and/or optimality variables, we reduced optimal control to inference in a graphical model.
- "Solving MDP" becomes very similar to inference in HMM / MEMM / CRF.
- The approach is quite similar to dynamic programming, value iterations, etc.

Slide from Maruan Al-Shedivat's Spring 2020 10-708 Guest Lectures