Probabilistic Graphical Models

Fundamentals of representation and learning

Murat Can Cobanoglu, Ph.D.

Lyda Hill Department of Bioinformatics
UT Southwestern Medical Center

Table of contents

Introduction

What are probabilistic graphical models?

Formalism to merge our domain knowledge with data.

What are probabilistic graphical models?

Formalism to merge our domain knowledge with data.

Renders large multivariate statistical models feasible.

What are probabilistic graphical models?

Formalism to merge our domain knowledge with data.

Renders large multivariate statistical models feasible.

Probabilistic to accommodate uncertainty (model & data)

What are probabilistic graphical models?

Formalism to merge our domain knowledge with data.

Renders large multivariate statistical models feasible.

Probabilistic to accommodate uncertainty (model & data)

Graph is natural for structural knowledge representation

What are probabilistic graphical models?

Formalism to merge our domain knowledge with data.

Renders large multivariate statistical models feasible.

Probabilistic to accommodate uncertainty (model & data)

Graph is natural for structural knowledge representation

Extensible model formalism. PGMs support:

What are probabilistic graphical models?

Formalism to merge our domain knowledge with data.

Renders large multivariate statistical models feasible.

Probabilistic to accommodate uncertainty (model & data)

Graph is natural for structural knowledge representation

Extensible model formalism. PGMs support:

- Discriminative or generative paradigm

What are probabilistic graphical models?

Formalism to merge our domain knowledge with data.

Renders large multivariate statistical models feasible.

Probabilistic to accommodate uncertainty (model & data)

Graph is natural for structural knowledge representation

Extensible model formalism. PGMs support:

- Discriminative or generative paradigm
- Range of applications: Classification, clustering, density estimation, imputation, dimensionality reduction, ...

What are probabilistic graphical models?

Formalism to merge our domain knowledge with data.

Renders large multivariate statistical models feasible.

Probabilistic to accommodate uncertainty (model & data)

Graph is natural for structural knowledge representation

Extensible model formalism. PGMs support:

- Discriminative or generative paradigm
- Range of applications: Classification, clustering, density estimation, imputation, dimensionality reduction, ...
- Causal or correlative relationships

What are probabilistic graphical models?

Formalism to merge our domain knowledge with data.

Renders large multivariate statistical models feasible.

Probabilistic to accommodate uncertainty (model & data)

Graph is natural for structural knowledge representation

Extensible model formalism. PGMs support:

- Discriminative or generative paradigm
- Range of applications: Classification, clustering, density estimation, imputation, dimensionality reduction, ...
- Causal or correlative relationships

Logistic regression, GMMs, PCA, etc. from yesterday are instances of the PGM formalism

What exactly defines a probabilistic graphical model?

Three essential components:

- Graph
- Model
- Data

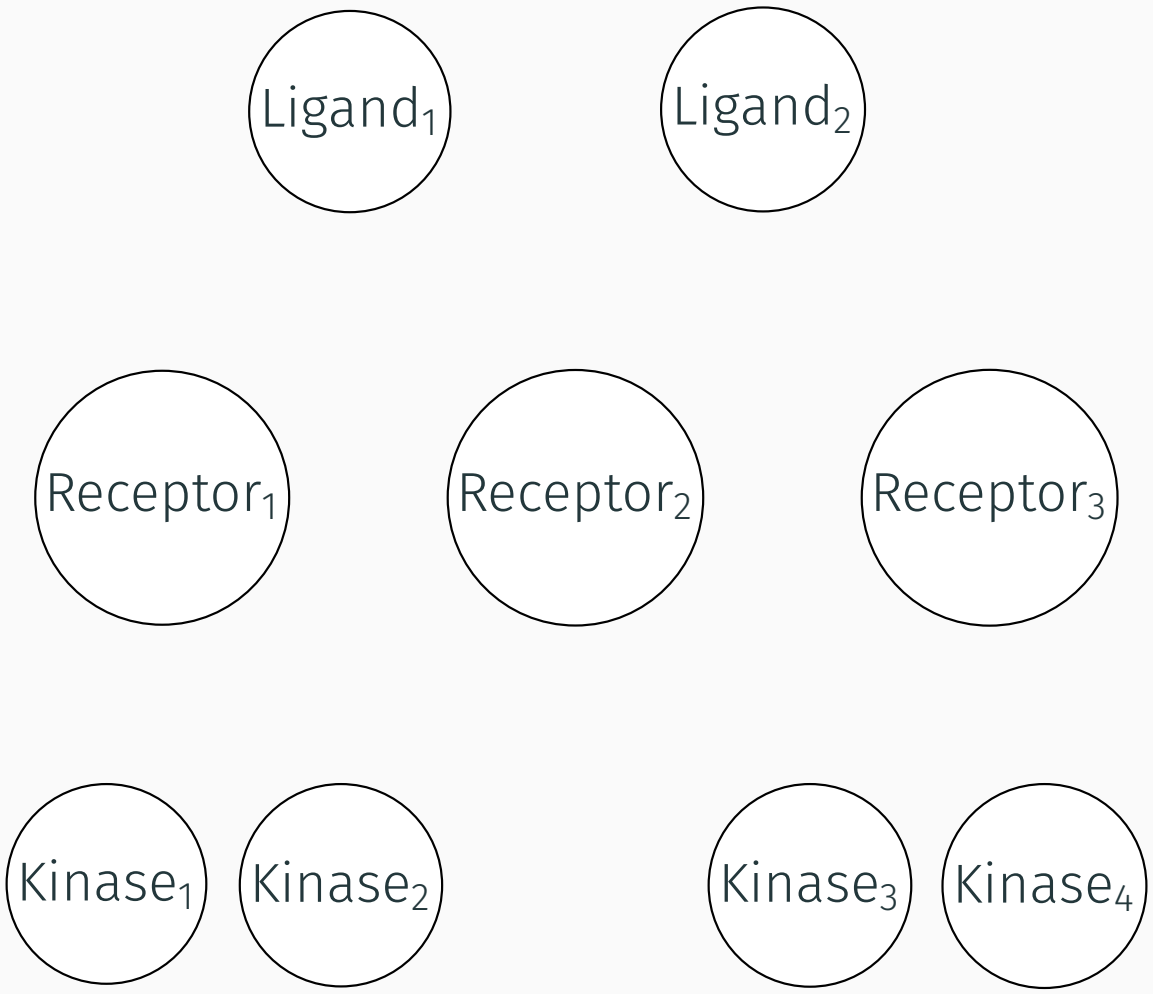
What is the ”graph” in a probabilistic graphical model?

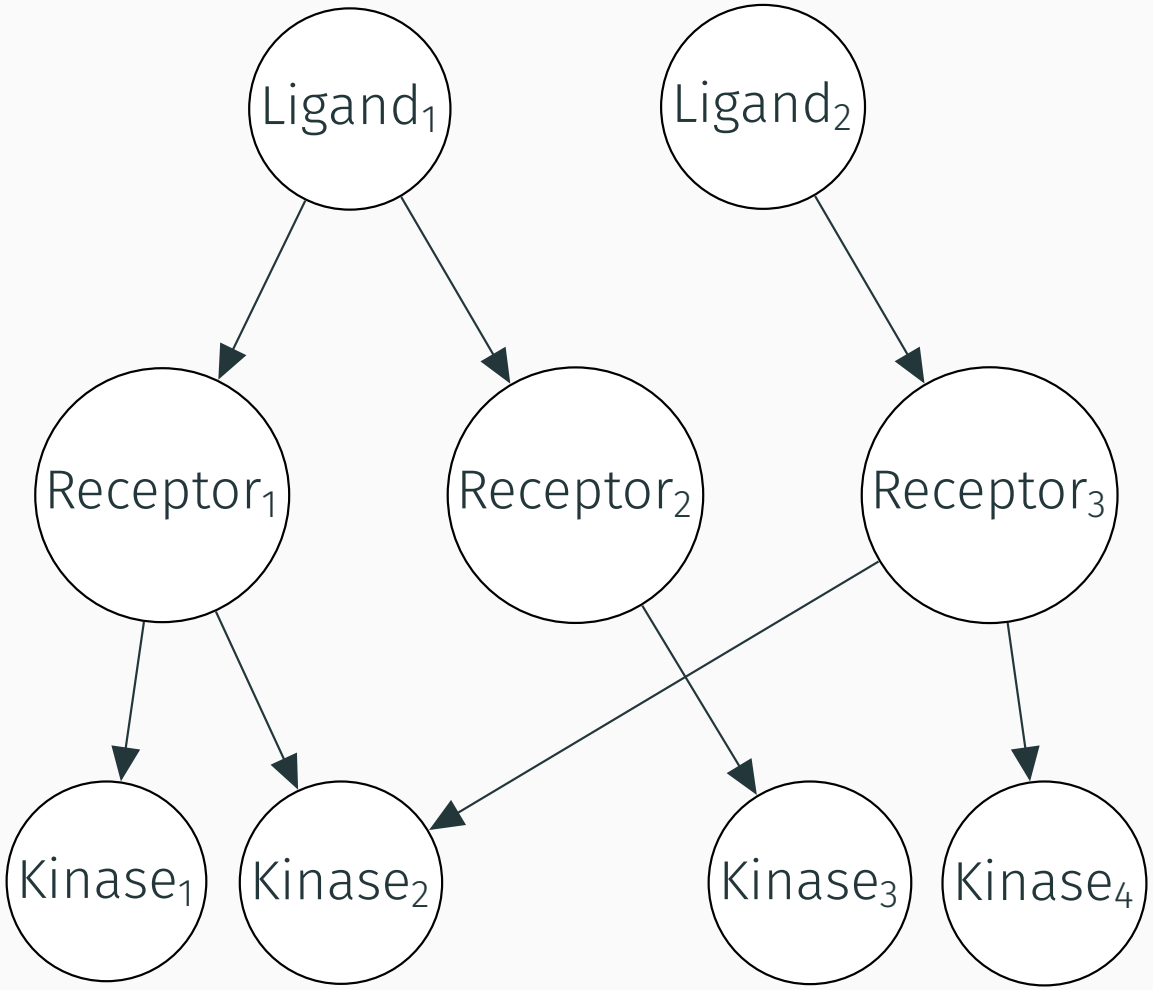
Graphs define the structure of the relationships we know/assume to exist among the constituents of our model.

There are two types of graphs (and thus PGMs):

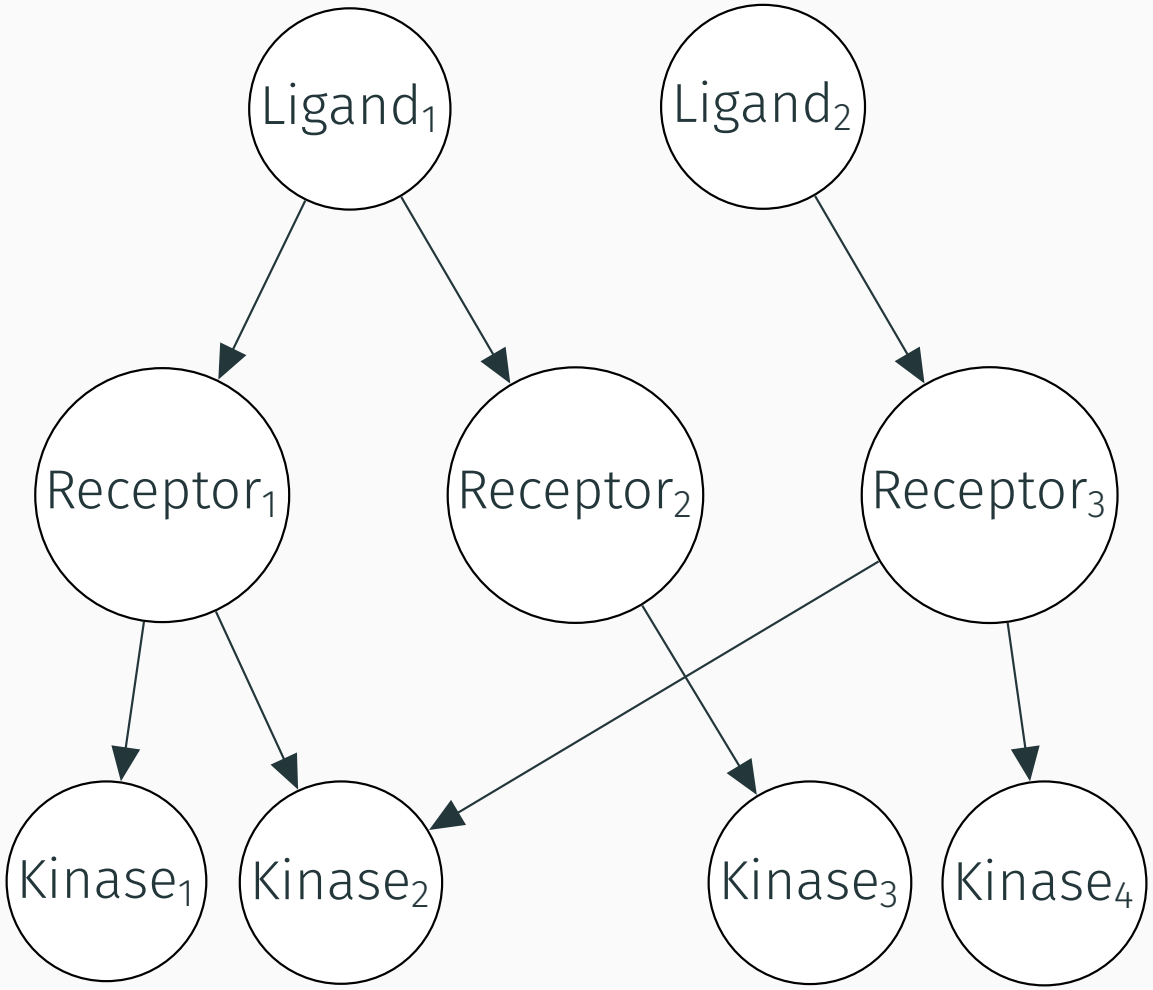
- Directed (Bayesian networks): *causality*
- Undirected (Markov random fields): *correlation*

Possible constituents of a biological model

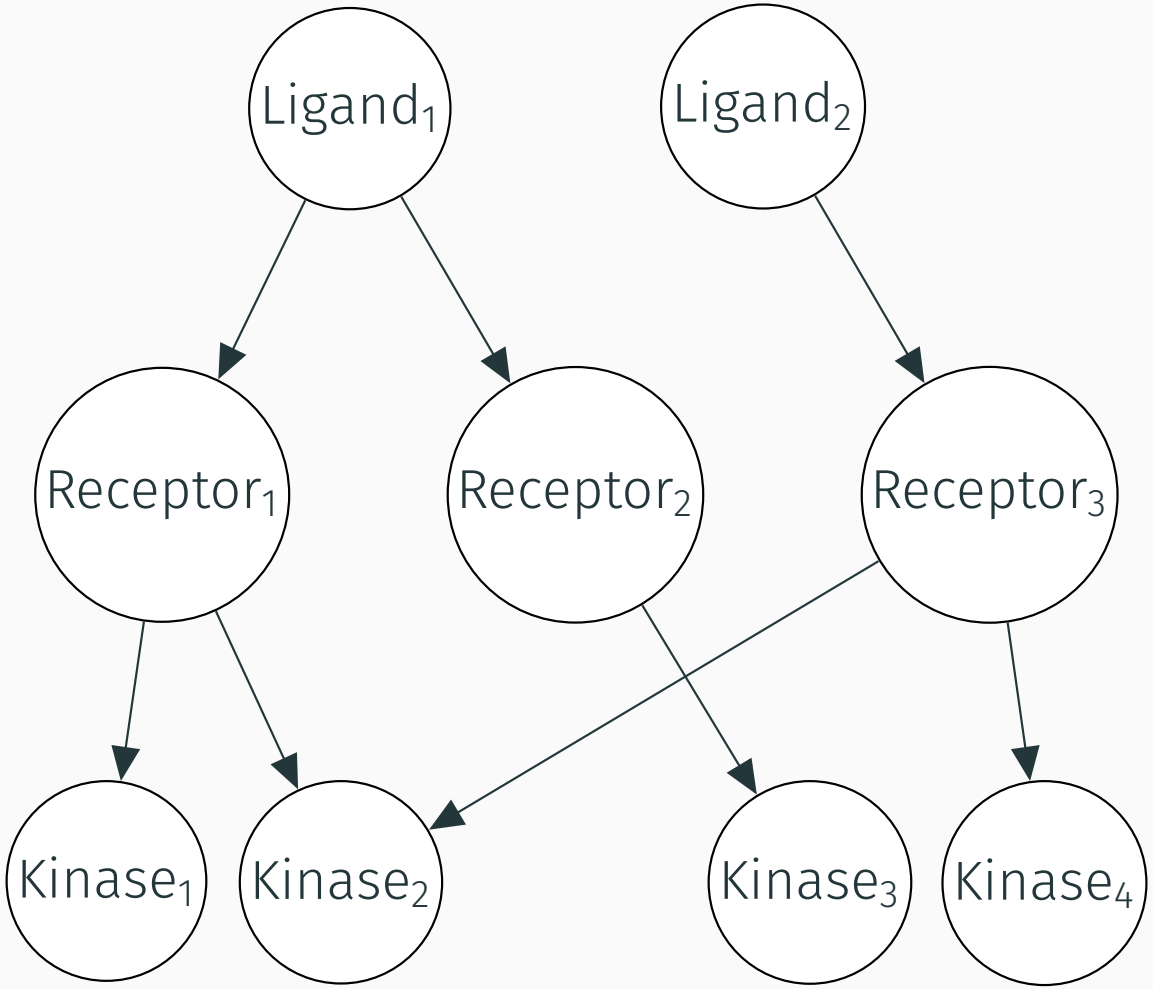




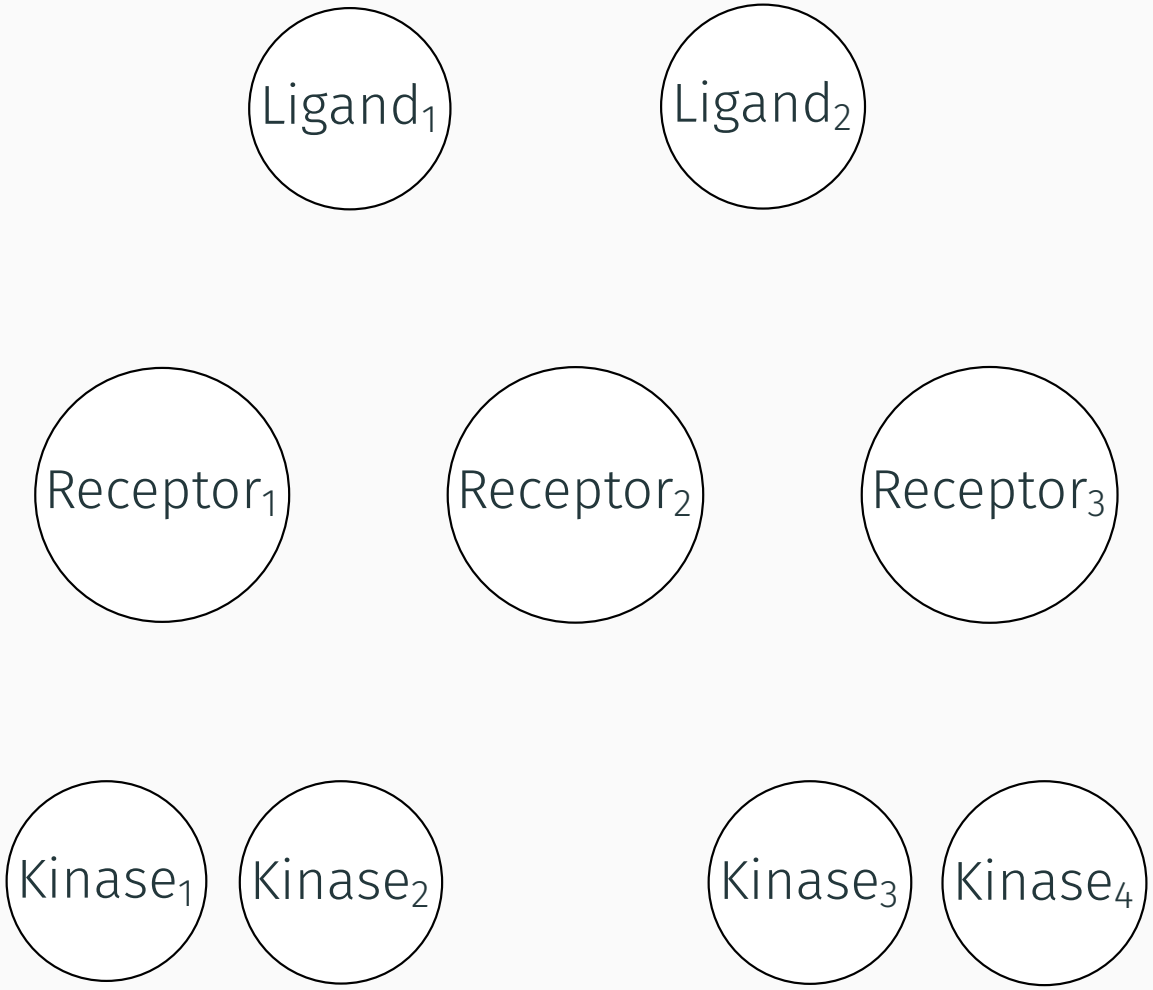
Connections denote our *a priori* knowledge about causal relationships.


Ligand1
Ligand2
Receptor1
Receptor2
Receptor3
Kinase1
Kinase2
Kinase3
Kinase4

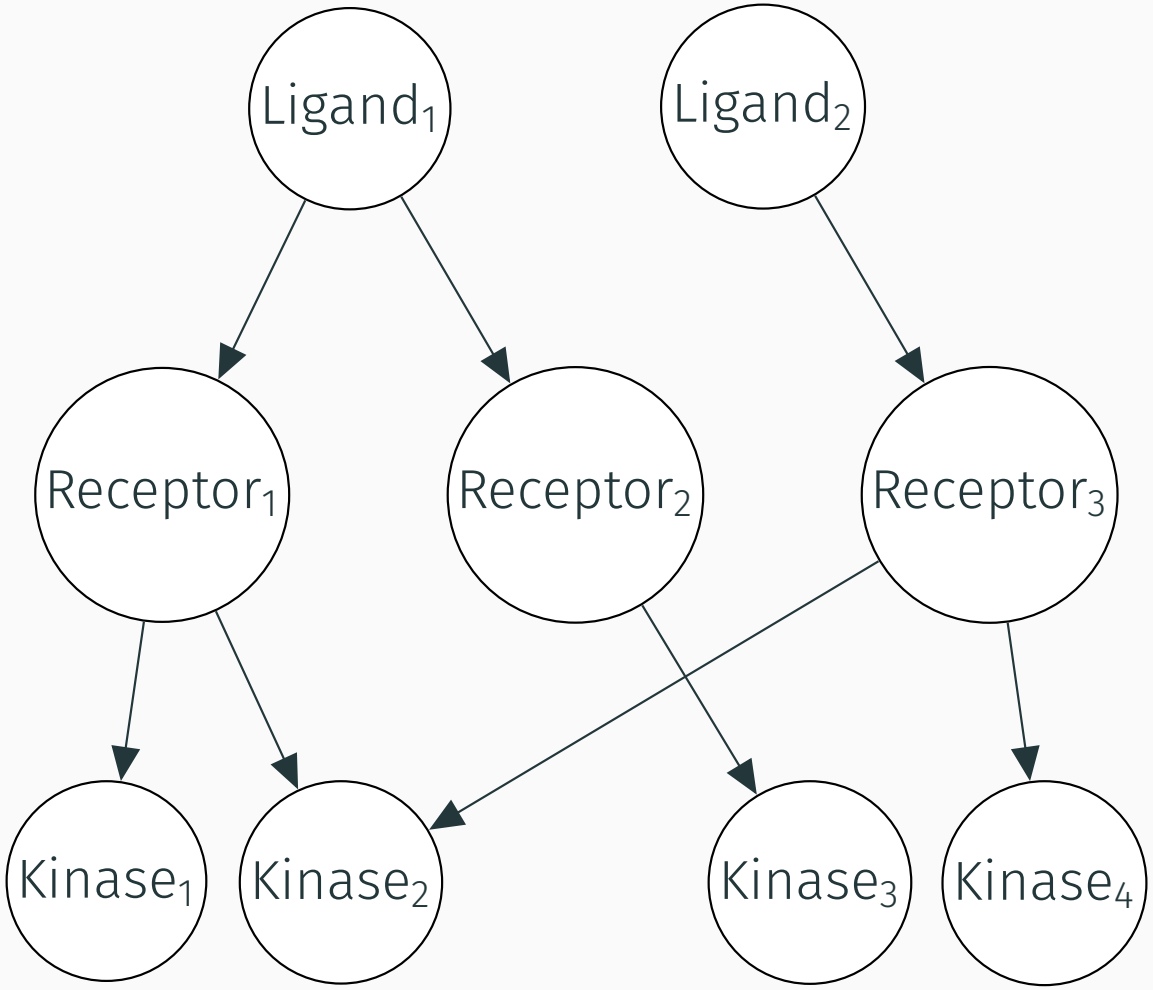
Why bother?



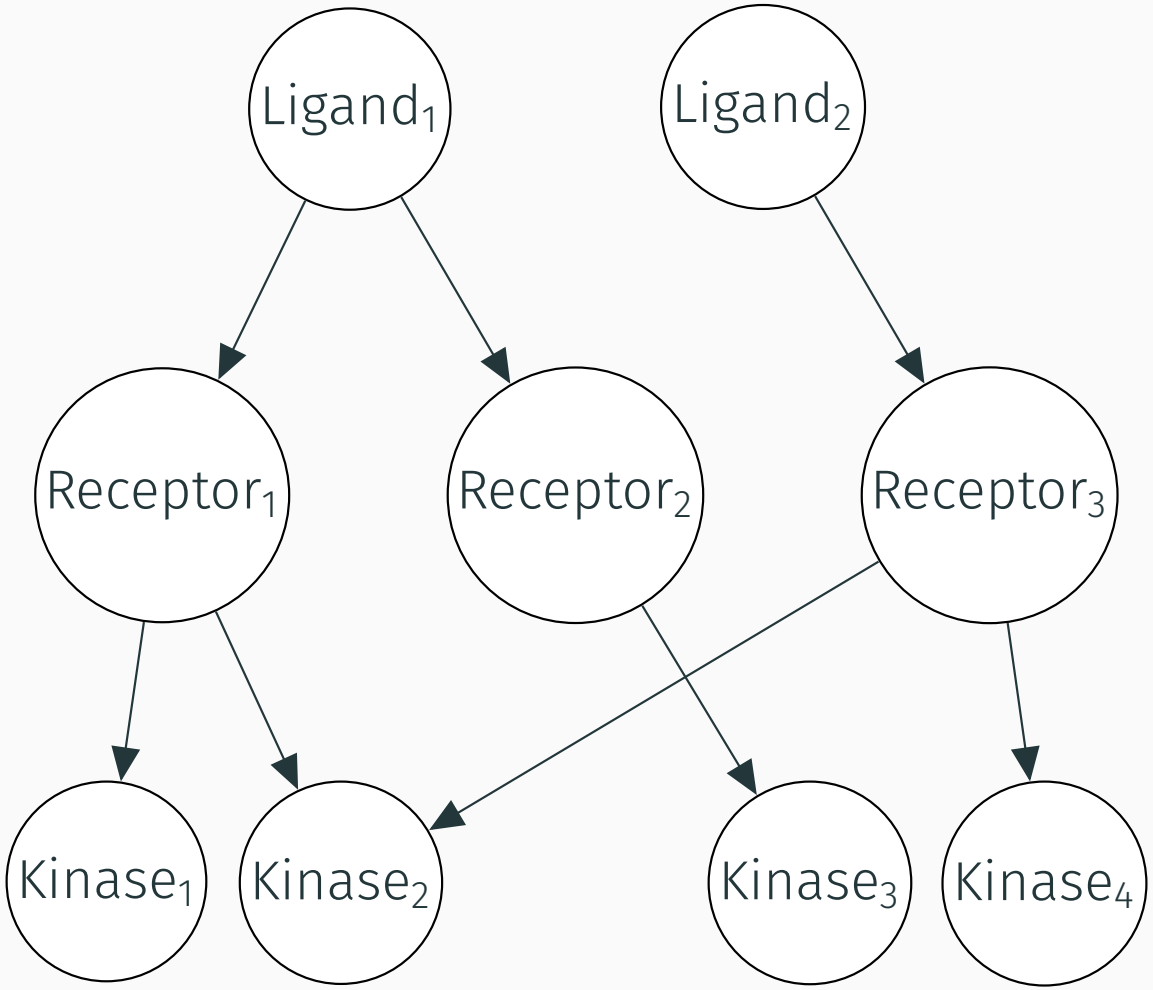
Why bother? Causal relationships induce significant simplification.



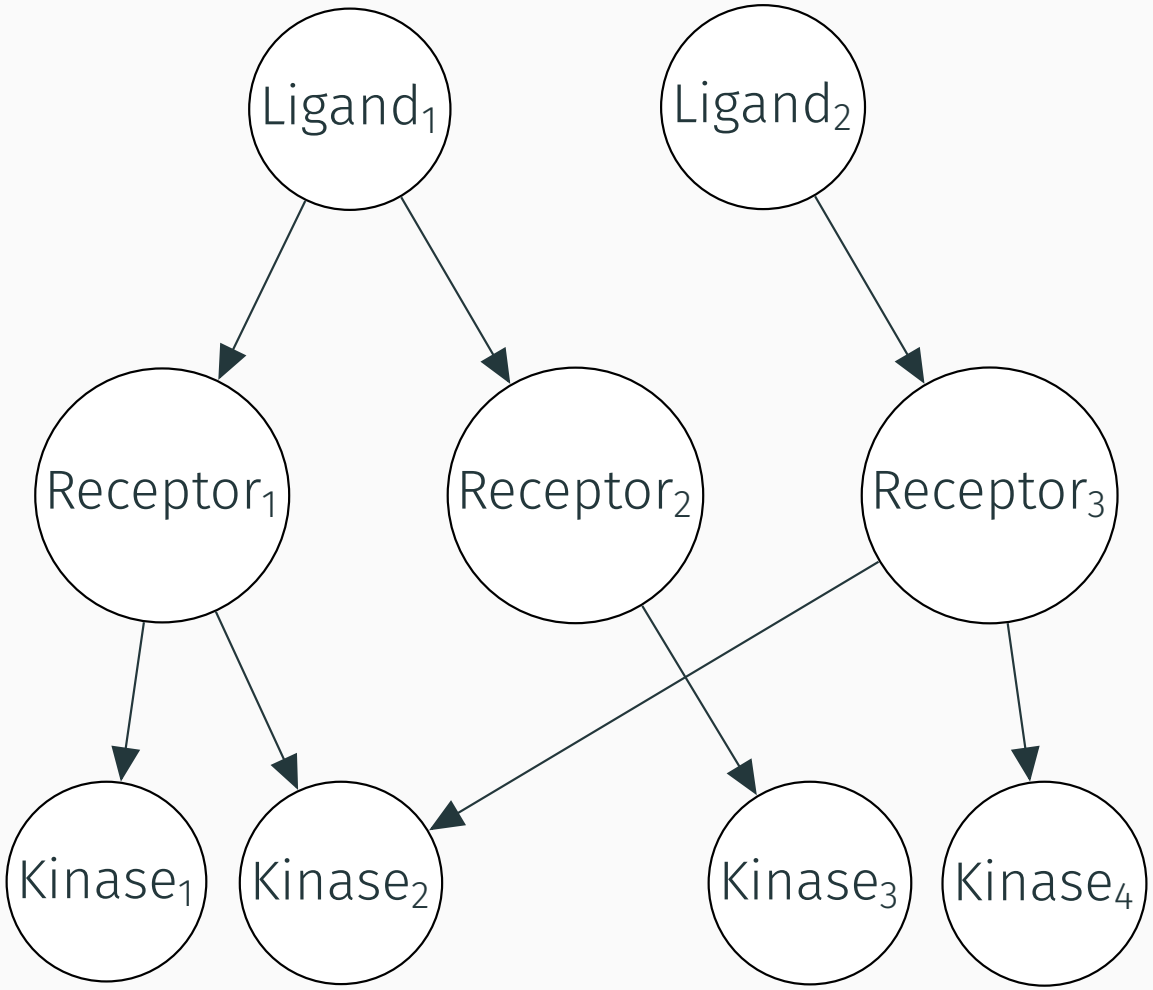
Joint: $P(L1, L2, R1, R2, R3, K1, K2, K3, K4) =$
$P(L1)P(L2|L1)P(R1|L1, L2)P(R2|L1, L2, R1) \cdots P(K4|L1, L2, R1, R2, R3, K1, K2, K3)$



Joint: $P(L1, L2, R1, R2, R3, K1, K2, K3, K4) =$
$P(L1)P(L2)P(R1|L1)P(R2|L1)P(R3|L2)P(K1|R1)P(K2|R1, R3)P(K3|R2)P(K4|R3)$

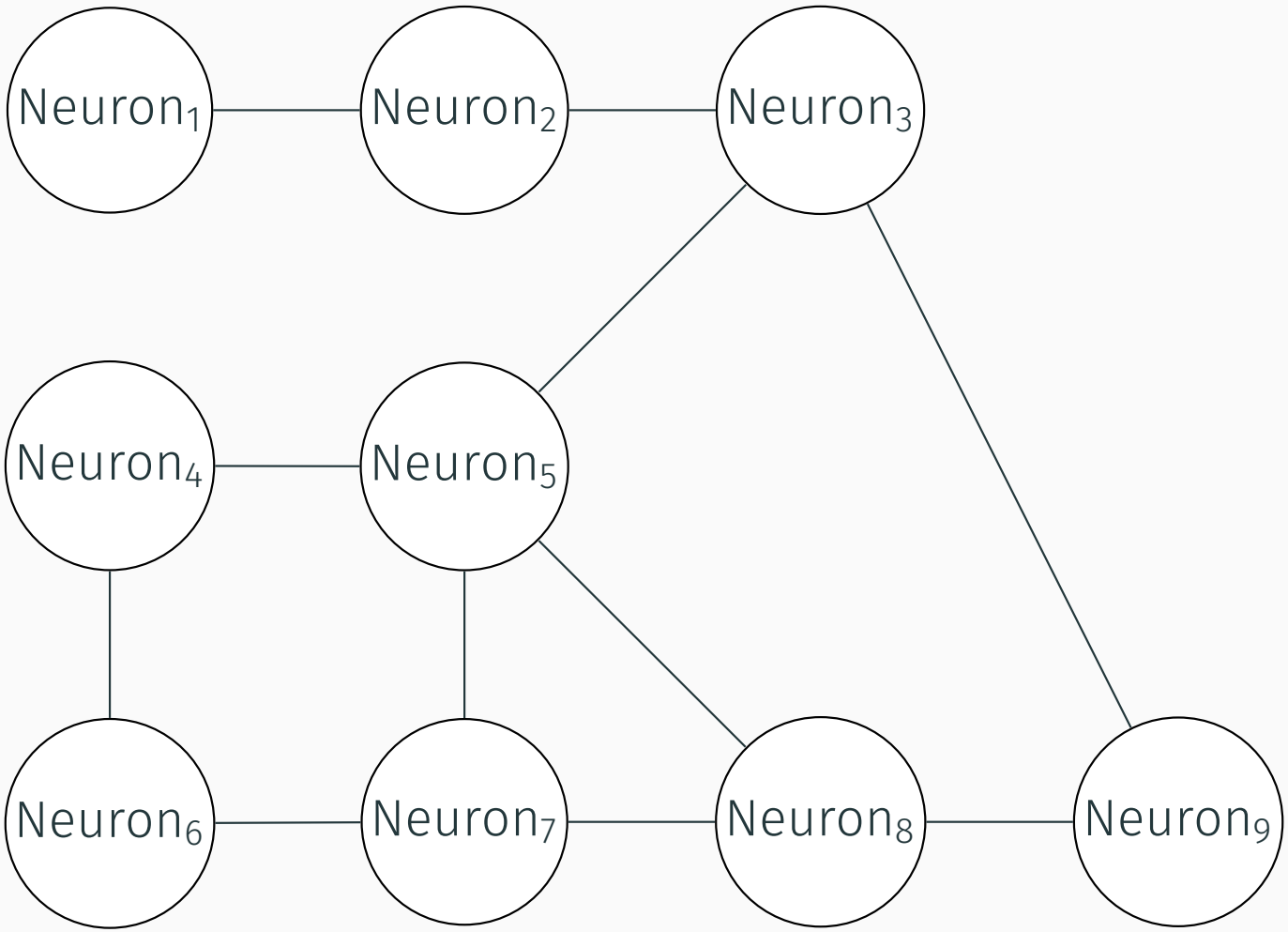


Joint: $P(L1, L2, R1, R2, R3, K1, K2, K3, K4) =$
$P(L1)P(L2)P(R1|L1)P(R2|L1)P(R3|L2)P(K1|R1)P(K2|R1, R3)P(K3|R2)P(K4|R3)$
Exponential growth avoided thanks to structure.


Ligand1
Ligand2
Receptor1
Receptor2
Receptor3
Kinase1
Kinase2
Kinase3
Kinase4

Are we done?

Undirected graph (Markov random field) example



Fundamentals

Utilizing a probabilistic graphical model requires:

- Representation
- Learning
- Inference

Fundamentals

Utilizing a probabilistic graphical model requires:

- Representation
- Learning
- Inference

Fundamentals

Utilizing a probabilistic graphical model requires:

- Representation
- Learning
- Inference

Fundamentals

Utilizing a probabilistic graphical model requires:

- Representation
- Learning
- Inference

PGMs enable utilizing exponentially large probability distributions by exploiting structure to achieve non-exponential cost.

Fundamentals

Utilizing a probabilistic graphical model requires:

- Representation
- Learning
- Inference

PGMs enable utilizing exponentially large probability distributions by exploiting structure to achieve non-exponential cost.

Graph is a natural structure for many domains.

Fundamentals

Utilizing a probabilistic graphical model requires:

- Representation
- Learning
- Inference

PGMs enable utilizing exponentially large probability distributions by exploiting structure to achieve non-exponential cost.

Graph is a natural structure for many domains.

Most biology papers end with a succinct graphical depiction of the "model". PGMs enable reification within a probabilistic framework.

Plate map representation

Semantic graph commonly represented with ”plate map notation”.

Plates are repeated as many times as the number on the lower-right hand side.

Node color/shape conveys information. Common notation:

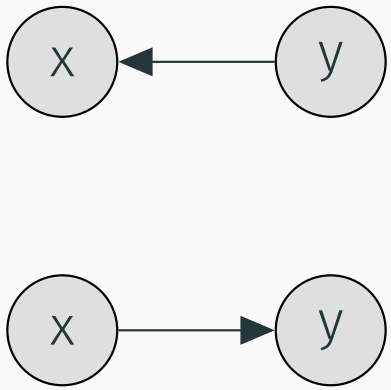
- Gray: observed
- White: latent
- Filled, small: constant / hyperparameter

Full model specification requires probability theory.

ML methods as PGM instances

Classification / Regression (ex: Log. Reg.)

Classification / Regression (ex: Log. Reg.)

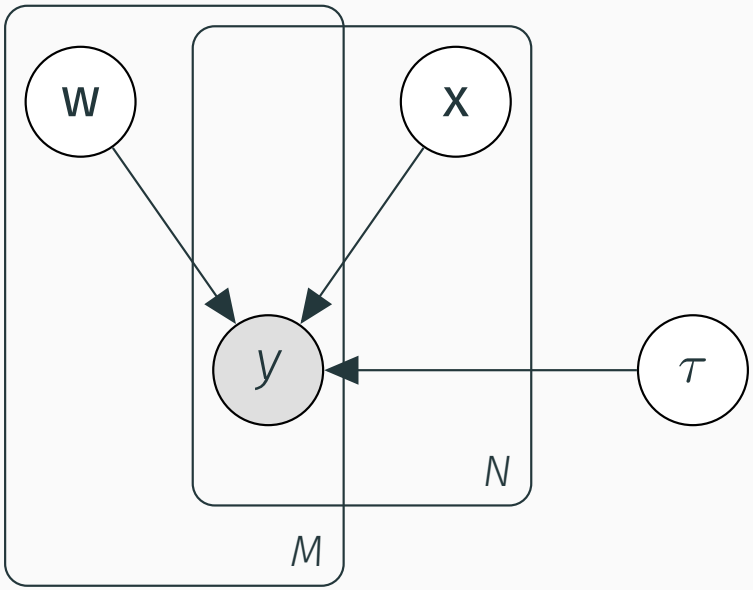


For this simple case, these are equivalent since
$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$.

ML methods as PGM instances

Clustering / Density Estimation (ex: GMMs)

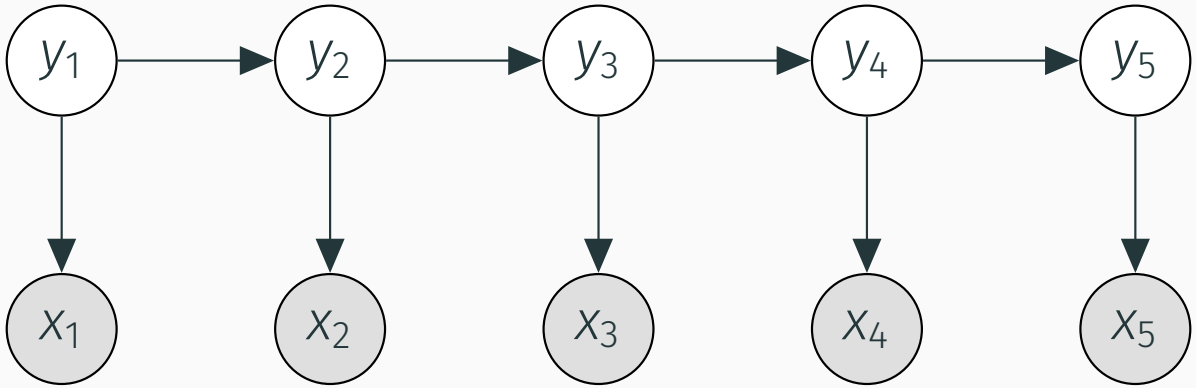
Dimensionality Reduction (PCA)



Model by Laura Dietz

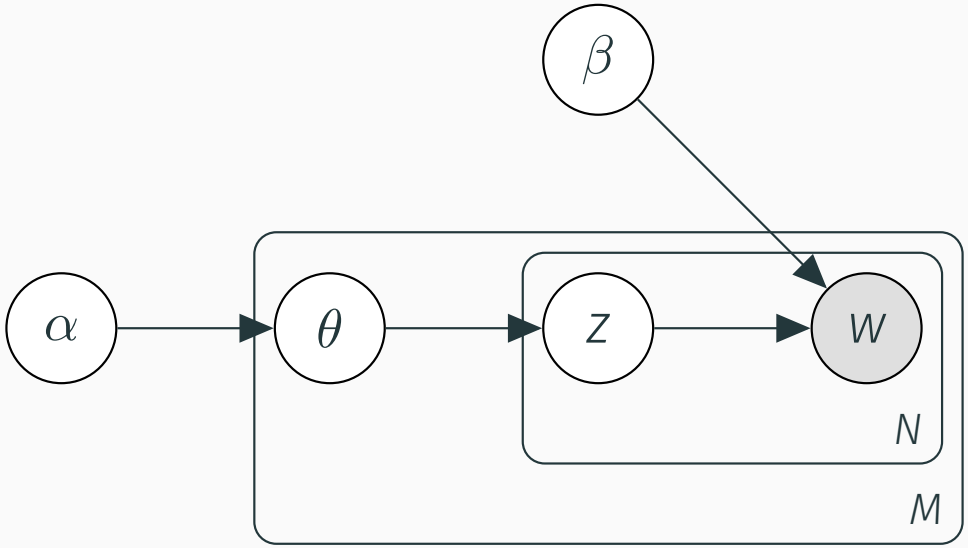
ML methods as PGM instances

Hidden Markov Model (HMM)



Case study: topic modeling

Latent Dirichlet Allocation (LDA)



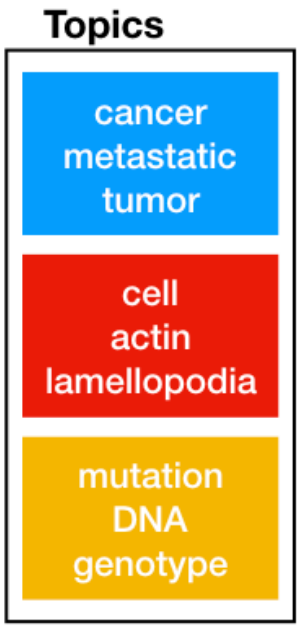
1. Choose $\theta \sim \mathrm{Dirichlet}(\alpha)$
2. For each of the N words w_n:
   (a) Choose a topic $z_n \sim \mathrm{Multinomial}(\theta)$.
   (b) Choose word w_n from $p(w_n|z_n, \beta)$, Mult. conditioned on topic z_n.

$\beta_{i,j}$ is the probability of i^{th} word in topic j.

Topic modeling with latent Dirichlet allocation

Topics

Topic modeling with latent Dirichlet allocation



Topic modeling with latent Dirichlet allocation

Topic modeling with latent Dirichlet allocation

RESEARCH ARTICLE SUMMARY

CELL BIOLOGY

BAK/BAX macropores facilitate mitochondrial herniation and mtDNA efflux during apoptosis

Kate McArthur,* Lachlan W. Whitehead, John M. Heddleston, Lucy Li, Benjamin S. Padman, Viola Oorschot, Niall D. Geoghegan, Stephane Chappaz, Sophia Davidson, Hui San Chin, Rachael M. Lane, Marija Dramicanin, Tahnee L. Saunders, Canny Sugiana, Romina Lessene, Laura D. Osellame, Teng-Leong Chew, Grant Dewson, Michael Lazarou, Georg Ramm, Guillaume Lessene, Michael T. Ryan, Kelly L. Rogers, Mark F. van Delft, Benjamin T. Kile*

CANCER

Detection and localization of surgically resectable cancers with a multi-analyte blood test

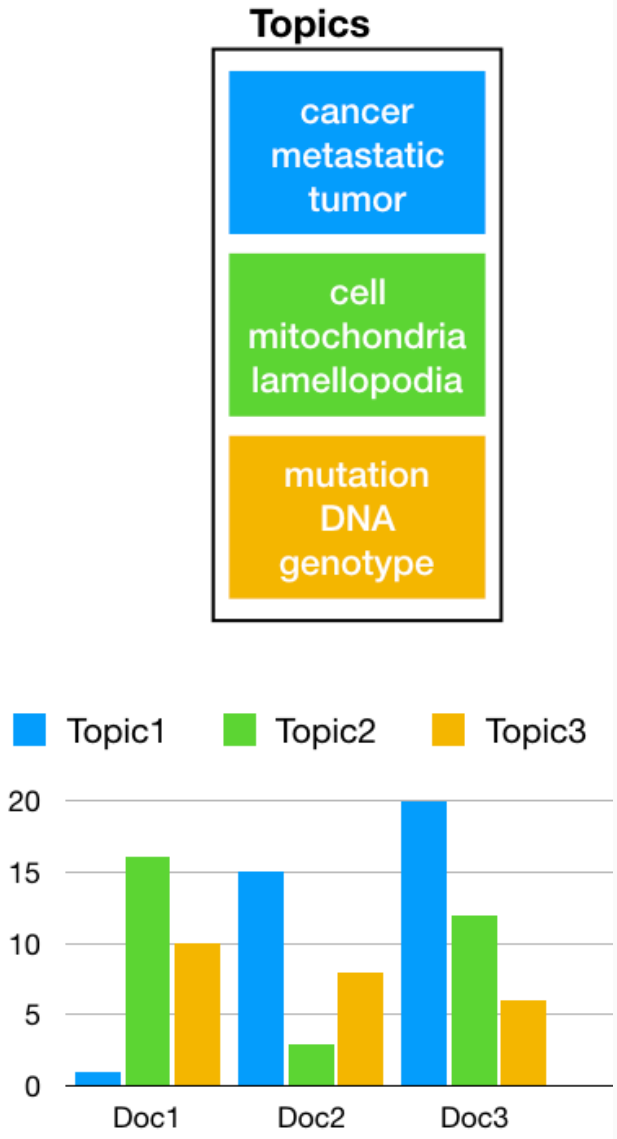
Joshua D. Cohen, Lu Li, Yuxuan Wang, Christopher Thoburn, Bahman Afsari, Ludmila Danilova, Christopher Douville, Ammar A. Javed, Fay Wong, Austin Mattox, Ralph. H. Hruban, Christopher L. Wolfgang, Michael G. Goggins, Alison P. Klein, Natalie Silliman, Robert E. Schoen, Hui-Li Wong, Massimo Falconi, Luis A. Diaz Jr., Bert Vogelstein,

ORGANOIDS

Patient-derived organoids model treatment response of metastatic gastrointestinal cancers

Georgios Vlachogiannis, Somaieh Hedayat, Alexandra Vatsiou, Yann Jamin, Javier Fernández-Mateos, Khurum Khan, Andrea Lampis, Katherine Eason, Ian Huntingford, Rosemary Burke, Mihaela Rata, Dow-Mu Koh, Nina Tunariu, David Collins, Sanna Hulkki-Wilson, Chanthirika Ragulan, Inmaculada Spiteri, Sing Yu Moorcraft, Ian Chau, Sheela Rao, David Watkins, Nicos Fotiadis, Maria Bali, Mahnaz Darvish-Damavandi, Hazel Lote, Zakaria Eltahir, Elizabeth C. Smyth, Ruwaida Begum, Paul A. Clarke, Jens C. Hahne, Mitchell Dowsett, Johann de Bono, Paul Workman, Anguraj Sadanandam, Matteo Fassan, Owen J. Sansom, Suzanne Eccles, Naureen Starling, Chiara Braconi, Andrea Sottoriva, Simon P. Robinson, David Cunningham, Nicola Valeri

Patient-derived organoids (PDOs) have recently emerged as robust preclinical models; however, their potential to predict clinical outcomes in patients has remained unclear. We report on a living biobank of PDOs from metastatic, heavily pretreated colorectal and gastroesophageal cancer patients recruited in phase 1/2 clinical trials. Phenotypic and genotypic profiling of PDOs showed a high degree of similarity to the original patient tumors. Molecular profiling of tumor organoids was matched to drug-screening results, suggesting that PDOs could complement existing approaches in defining cancer vulnerabilities and improving treatment responses. We compared responses to anticancer agents ex vivo in organoids and PDO-based orthotopic mouse tumor xenograft models with the responses of the patients in clinical trials. Our data suggest that PDOs can recapitulate patient responses in the clinic and could be implemented in personalized medicine programs.



Representation

Definition

Probabilistic models can define generative processes for observed data.

We can use graph structures to represent our conditional independence assumptions in probabilistic models.

Definition

Probabilistic models can define generative processes for observed data.

We can use graph structures to represent our conditional independence assumptions in probabilistic models.

Independence: $A \perp B \Leftrightarrow P(A, B) = P(A)P(B)$

Definition

Probabilistic models can define generative processes for observed data.

We can use graph structures to represent our conditional independence assumptions in probabilistic models.

Independence: $A \perp B \Leftrightarrow P(A, B) = P(A)P(B)$
Remember Bayes' rule: $P(A|B)P(B) = P(B|A)P(A) \Leftrightarrow P(A|B) = \frac{P(A,B)}{P(B)}$

Definition

Probabilistic models can define generative processes for observed data.

We can use graph structures to represent our conditional independence assumptions in probabilistic models.

Independence: $A \perp B \Leftrightarrow P(A, B) = P(A)P(B)$
Remember Bayes' rule: $P(A|B)P(B) = P(B|A)P(A) \Leftrightarrow P(A|B) = \frac{P(A,B)}{P(B)}$

Therefore $A \perp B \Rightarrow P(A|B) = \frac{P(A)\cancel{P(B)}}{\cancel{P(B)}} = P(A)$

Definition

Probabilistic models can define generative processes for observed data.

We can use graph structures to represent our conditional independence assumptions in probabilistic models.

Independence: $A \perp B \Leftrightarrow P(A, B) = P(A)P(B)$
Remember Bayes' rule: $P(A|B)P(B) = P(B|A)P(A) \Leftrightarrow P(A|B) = \frac{P(A,B)}{P(B)}$

Therefore $A \perp B \Rightarrow P(A|B) = \frac{P(A)\cancel{P(B)}}{\cancel{P(B)}} = P(A)$

Conditional independence: $A \perp B|C \Leftrightarrow P(A|B, C) = P(A|C)$

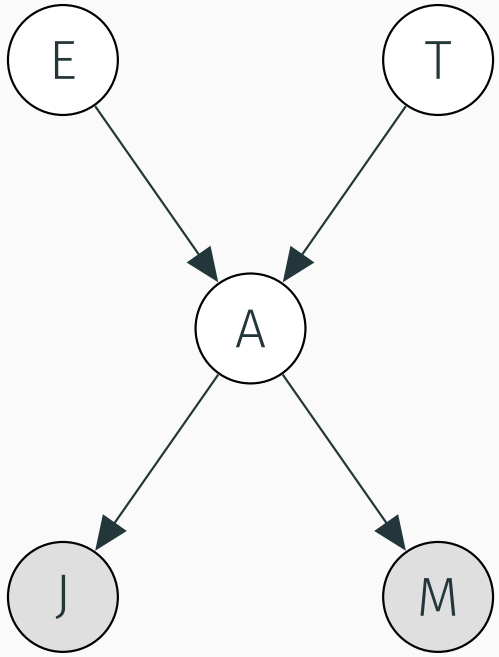
Example:

- Thief break-in causes my alarm to sound.
- Earthquake causes my alarm to sound.
- Mary (neighbor) calls if she hears the alarm.
- John (another neighbor) calls if he hears the alarm.

Example:

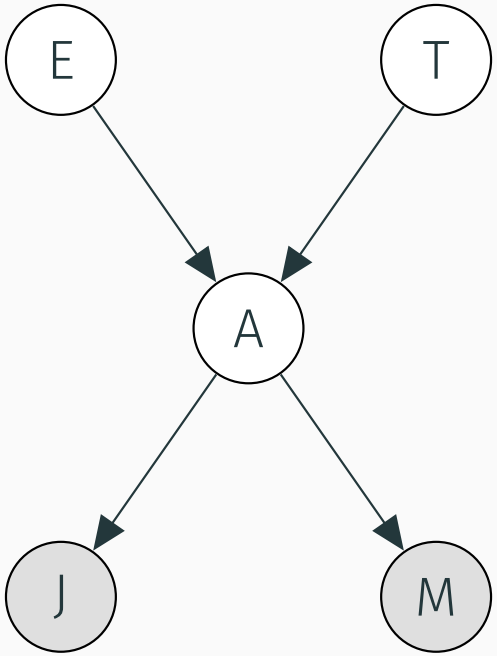
- Thief break-in causes my alarm to sound.
- Earthquake causes my alarm to sound.
- Mary (neighbor) calls if she hears the alarm.
- John (another neighbor) calls if he hears the alarm.

Let's build a causal network.



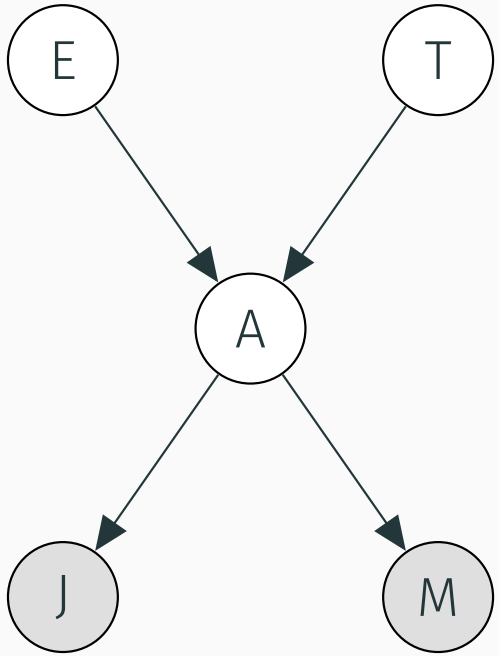
Graph represents our local Markov assumptions.

Example: If we know there is no alarm, the probability of getting a call from Mary is independent of an earthquake (or a thief break-in).

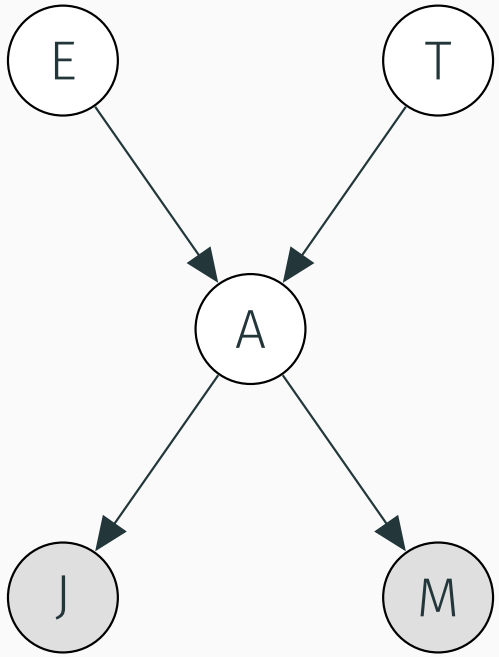


Our local Markov assumptions:

- $E \perp T$
- $J \perp \{E, T, M\} | A$
- $M \perp \{E, T, J\} | A$

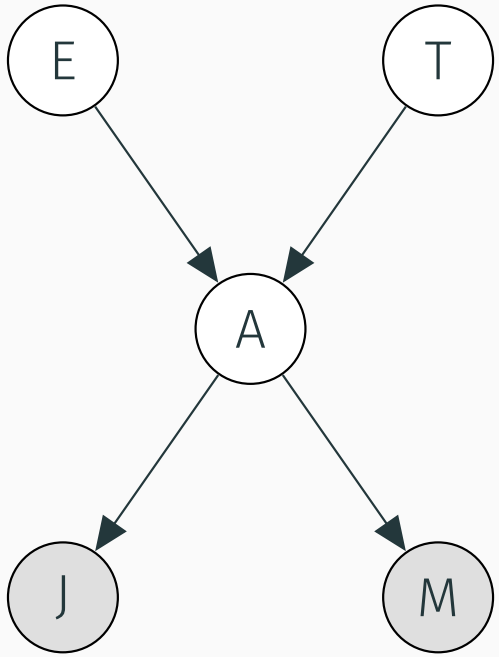


Joint distribution (using chain rule, without any assumptions):
$P(T, E, A, M, J) = P(E)\ P(T|E)\ P(A|E, T)\ P(M|A, E, T)\ P(J|M, A, E, T)$



Joint distribution (using chain rule, without any assumptions):
$P(T, E, A, M, J) = P(E)\ P(T|E)\ P(A|E, T)\ P(M|A, E, T)\ P(J|M, A, E, T)$

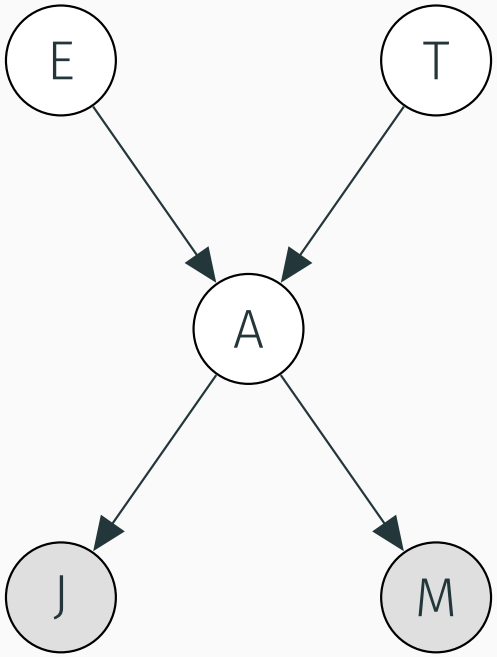
With local Markov assumptions:
$= P(E)\ P(T)\ P(A|E, T)\ P(J|A)\ P(M|A)$



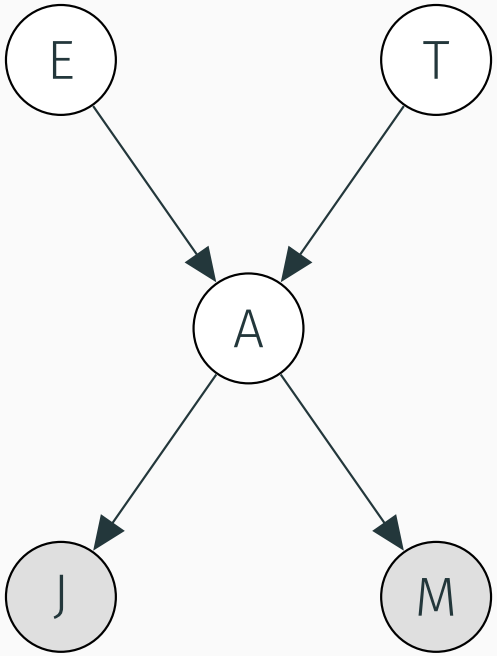
Joint distribution (using chain rule, without any assumptions):
$P(T, E, A, M, J) = P(E)\ P(T|E)\ P(A|E, T)\ P(M|A, E, T)\ P(J|M, A, E, T)$

With local Markov assumptions:
$= P(E)\ P(T)\ P(A|E, T)\ P(J|A)\ P(M|A)$

How many parameters to estimate?



Question: Is $E \perp T \mid A$?



Question: What about $E \perp T \mid J$?

D-separation

X is d-separated from Y by Z $\equiv X \perp Y \mid Z$

Three configurations to think about:

- Causal direction

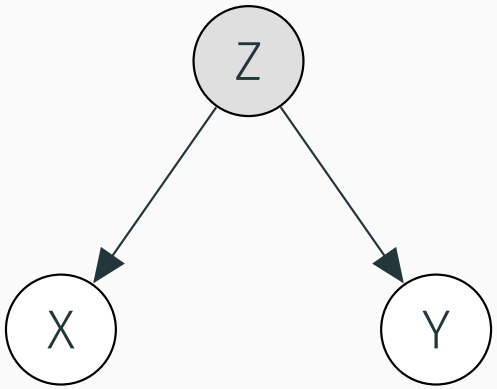
$$X \perp Y \mid Z$$

D-separation

X is d-separated from Y by Z $\equiv X \perp Y \mid Z$

Three configurations to think about:

- Causal direction
- Common cause



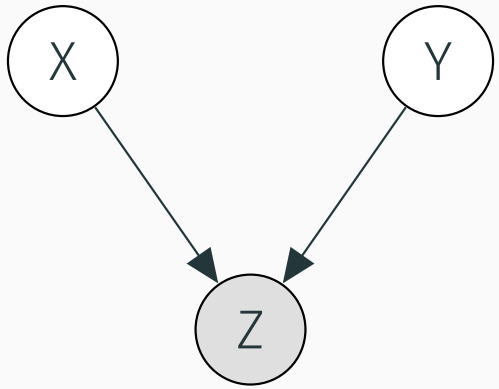
$$X \perp Y \mid Z$$

D-separation

X is d-separated from Y by Z $\equiv X \perp Y \mid Z$

Three configurations to think about:

- Causal direction
- Common cause
- Explaining away



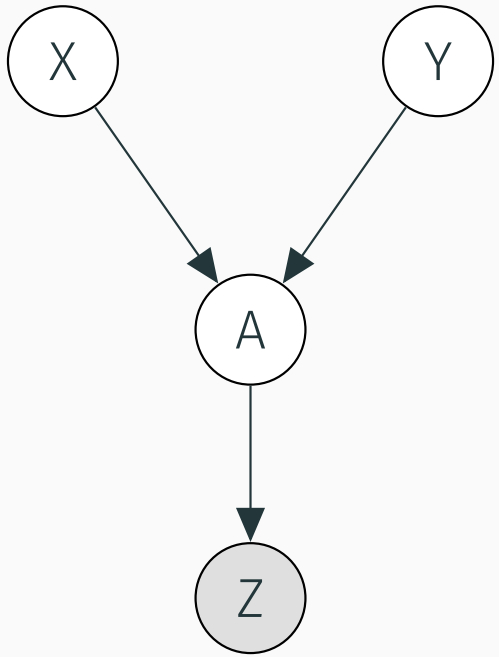
$$X \not\perp Y \mid Z$$

D-separation

X is d-separated from Y by Z $\equiv X \perp Y \mid Z$

Three configurations to think about:

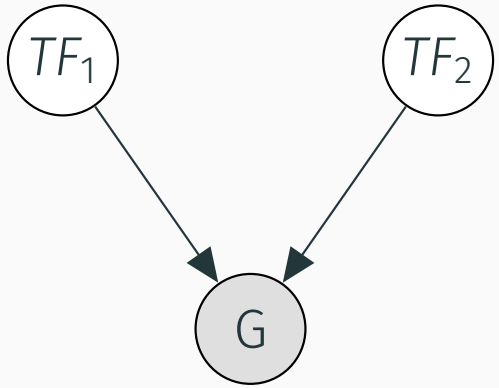
- Causal direction
- Common cause
- Explaining away



$X \not\perp Y \mid Z$

”Explaining away” in action: transcriptional regulation

Assume that TF_1 and TF_2 both regulate the same gene, G.



$$TF_1 \not\perp TF_2 \mid G$$

This happens quite frequently: within the $CD4^+$ T-cell context, using state-of-the-art TF regulatory networks (Marbach *et al.*, Nature Methods, 2016) STAT6 and GATA3 overlap with every other known TF on at least one gene. Across all 394 contexts in the same study, more than 95% of all TF-TF pairs overlap on at least one gene in *every single context*.

Learning

PGM learning has a rich literature, with diverse options for learning:

- Maximum likelihood estimators
- Bayesian inference
- Variational inference
- Sampling based methods

Fundamentals of Learning

Let θ be the parameters of our model, and D be the data.

We can use Bayes' theorem to write the joint as follows:

$$p(\theta, D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}$$

Let θ be the parameters of our model, and D be the data.

We can use Bayes' theorem to write the joint as follows:

$$p(\theta, D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}$$

Maximum likelihood

If we are interested in *maximum likelihood estimate* (MLE) of $\hat{\theta}$:

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}\, p(D|\theta)$$

where $p(D|\theta)$ is called the likelihood. Think of it as quantifying "how likely the data is, given the parameter".

$\hat{\theta}_{MLE}$ would be a point estimate.

Let θ be the parameters of our model, and D be the data.

We can use Bayes' theorem to write the joint as follows:

$$p(\theta, D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}$$

Bayesian inference

If we are interested in *maximum a posteriori* (MAP) estimate of $\hat{\theta}$:

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, p(\theta, D) = \underset{\theta}{\operatorname{argmax}}\, p(D|\theta)p(\theta)$$

where $p(D|\theta)$ is the likelihood and $p(\theta)$ is the prior.

Again, $\hat{\theta}_{MAP}$ would be a point estimate.

Fundamentals of Learning

Let θ be the parameters of our model, and D be the data.

We can use Bayes' theorem to write the joint as follows:

$$p(\theta, D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}$$

What if we are interested in some quantity x given some data D?

$$p(x|D) = \int p(x|\theta, D)p(\theta|D)d\theta = \mathbb{E}_{(\theta|\mathbb{D})}[p(x|\theta, D)]$$

This could be useful when we are interested in some latent variable.

Monte Carlo

Approximate expectations of hard-to-compute integrals with sums from sampling.

$$\mathbb{E}_{(\theta|\mathbb{D})}[p(x|\theta, D)] \approx \frac{1}{S}\sum_{s=1}^{S} p(x|\theta_s, D), \theta_s \sim p(\theta|D)$$

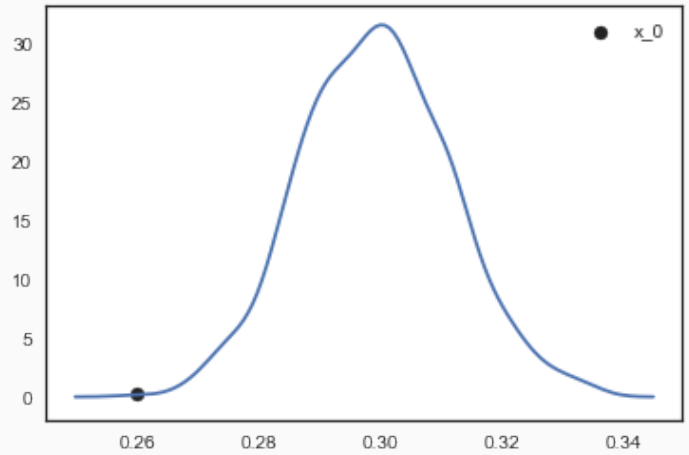
Markov chain Monte Carlo methods involve sampling from a transition probability that, out of all states, depends only on the current state (and none of the history). This is the ”Markov chain” property.

Metropolis algorithm

Metropolis is a classical MCMC algorithm. Briefly:

Initialize with some x_0.

i. Generate x' from proposal distribution (for example, Gaussian: $x' \sim N(x, \sigma^2)$). Initialize with some x_0.
ii. Accept new proposal with probability $min(1, p(x')/p(x))$.
iii. If rejected, $x_{t+1} = x(t)$. If accepted, $x_{t+1} = x'$.

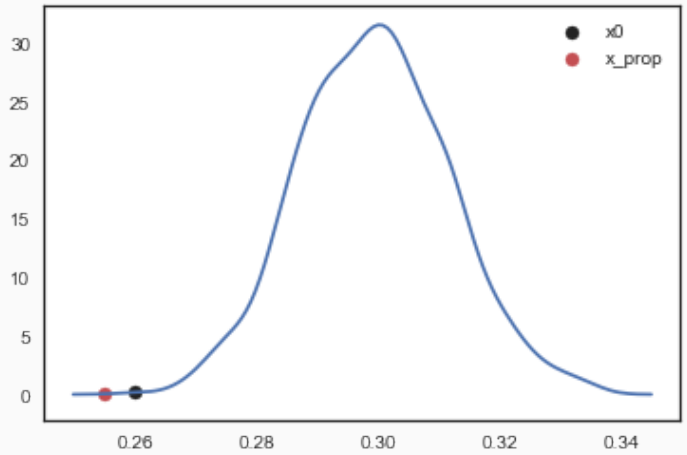


Metropolis algorithm

Metropolis is a classical MCMC algorithm. Briefly:

Initialize with some x_0.

i. Generate x' from proposal distribution (for example, Gaussian: $x' \sim N(x, \sigma^2)$). Initialize with some x_0.
ii. Accept new proposal with probability $min(1, p(x')/p(x))$.
iii. If rejected, $x_{t+1} = x(t)$. If accepted, $x_{t+1} = x'$.



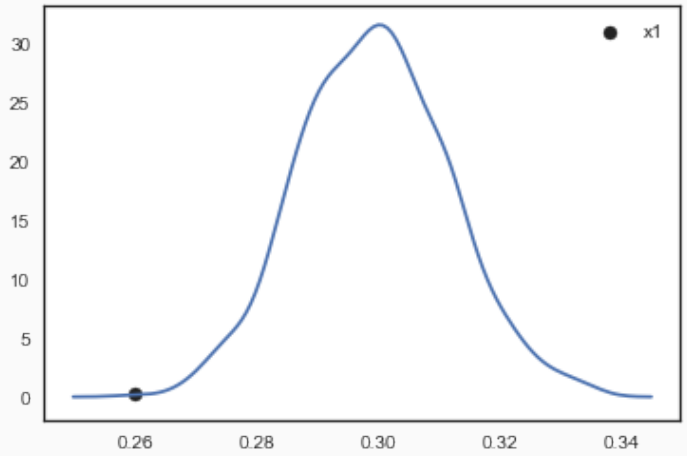
Notice in the first step that proposal depends on current state, hence adaptive.

Metropolis algorithm

Metropolis is a classical MCMC algorithm. Briefly:

Initialize with some x_0.

i. Generate x' from proposal distribution (for example, Gaussian: $x' \sim N(x, \sigma^2)$). Initialize with some x_0.
ii. Accept new proposal with probability $min(1, p(x')/p(x))$.
iii. If rejected, $x_{t+1} = x(t)$. If accepted, $x_{t+1} = x'$.

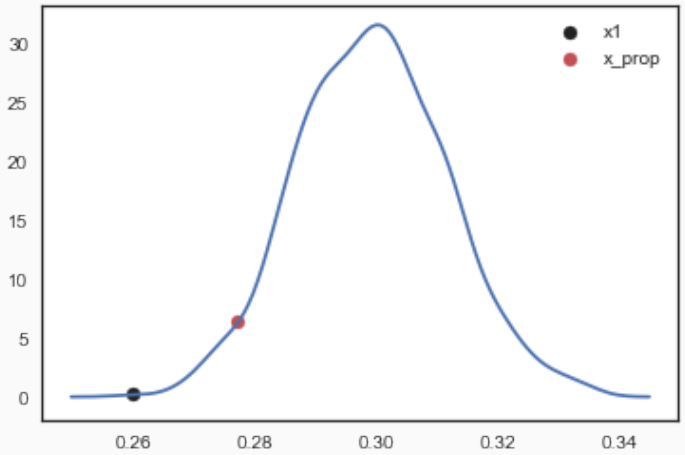


Metropolis algorithm

Metropolis is a classical MCMC algorithm. Briefly:

Initialize with some x_0.

i. Generate x' from proposal distribution (for example, Gaussian: $x' \sim N(x, \sigma^2)$). Initialize with some x_0.
ii. Accept new proposal with probability $min(1, p(x')/p(x))$.
iii. If rejected, $x_{t+1} = x(t)$. If accepted, $x_{t+1} = x'$.



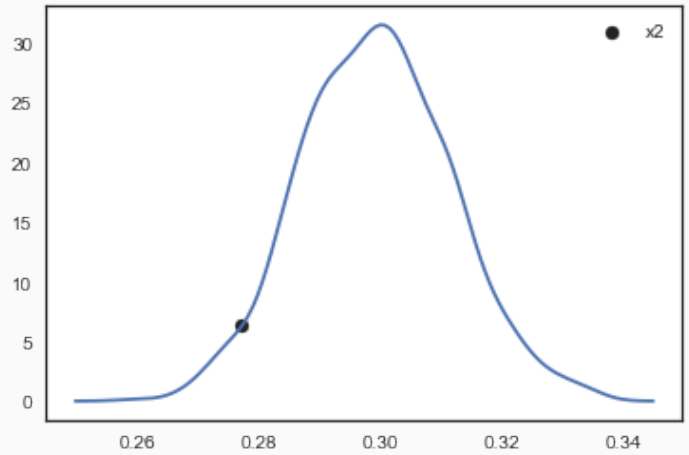
Notice that proposals better than the current state are always accepted. Worse states can be accepted, with lower probability.

Metropolis algorithm

Metropolis is a classical MCMC algorithm. Briefly:

Initialize with some x_0.

i. Generate x' from proposal distribution (for example, Gaussian: $x' \sim N(x, \sigma^2)$). Initialize with some x_0.
ii. Accept new proposal with probability $min(1, p(x')/p(x))$.
iii. If rejected, $x_{t+1} = x(t)$. If accepted, $x_{t+1} = x'$.



Hamiltonian Monte Carlo

Probability distributions can often be rewritten as:

$$p(x) = \frac{1}{Z} exp(-E(x))$$

where $\frac{\delta E(x)}{\delta x}$ can be computed to glean information on where we can find ”greener pastures” (states of higher probability).

Hamiltonian Monte Carlo (HMC) utilizes this information for more efficient learning.

No-U-Turn-Sampler (NUTS) is an HMC sampler that features adaptive step size tuning. State-of-the-art sampler for general purpose packages.

Exercise

Exercise question:

- Gambler sets up street corner bets on coin toss.
- In reality, he has two coins: one is loaded, one is fair.
- To avoid suspicion, uses the loaded coin only on some days.
- We observe the total number of wins on each day.

Inference objectives:

- Coin used on each day
- Probability of success for each coin
- Probability of cheat on any day

Exercise question:

- Gambler sets up street corner bets on coin toss.
- In reality, he has two coins: one is loaded, one is fair.
- To avoid suspicion, uses the loaded coin only on some days.
- We observe the total number of wins on each day.

Inference objectives:

- Coin used on each day
- Probability of success for each coin
- Probability of cheat on any day

What is the corresponding graphical model?

Probability Primer

What is the next toss?

What is the next toss?

Is P(heads) = 1?

How do we reflect our prior beliefs?

How do we reflect our prior beliefs?

What if we think the coin is loaded?

How do we reflect our prior beliefs?

What if we think the coin is loaded?

What if there are two coins alternating?

Random Variables

Random variable: possible values are outcomes of a random phenomenon.

Examples:

- Coin toss

Random Variables

Random variable: possible values are outcomes of a random phenomenon.

Examples:

- Coin toss
- Dice roll

Random Variables

Random variable: possible values are outcomes of a random phenomenon.

Examples:

- Coin toss
- Dice roll
- Read counts in RNA-seq

Random Variables

Random variable: possible values are outcomes of a random phenomenon.

Examples:

- Coin toss
- Dice roll
- Read counts in RNA-seq

What if the coin is loaded?

Random Variables

Random variable: possible values are outcomes of a random phenomenon.

Examples:

- Coin toss
- Dice roll
- Read counts in RNA-seq

... or the dice are not fair?

Random Variables

Random variable: possible values are outcomes of a random phenomenon.

Examples:

- Coin toss
- Dice roll
- Read counts in RNA-seq

How do we reflect different probabilistic assumptions?

Probability Distributions

Probability distributions describe probabilities associated with a random phenomenon.

Probability Distributions

Probability distributions describe probabilities associated with a random phenomenon.

We can define probability distributions in two categories:

- Discrete
- Continuous

Probability Distributions

Probability distributions describe probabilities associated with a random phenomenon.

We can define probability distributions in two categories:

- Discrete
- Continuous

Discrete probability distributions are typically defined by *probability mass functions* that describe the probability of every (possible infinite, but always countable) possible event.

Probability Distributions

Probability distributions describe probabilities associated with a random phenomenon.

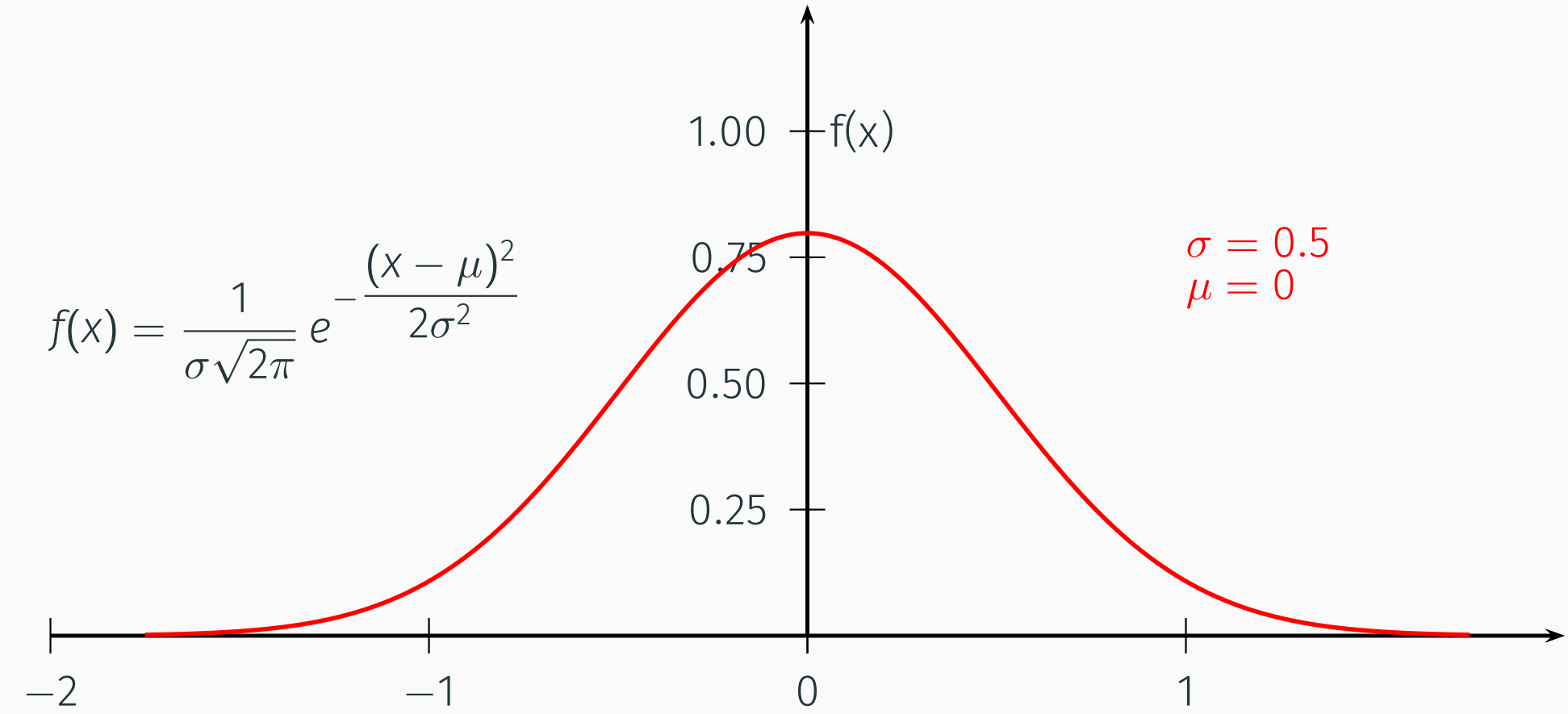
We can define probability distributions in two categories:

- Discrete
- Continuous

Continuous probability distributions are typically defined by *probability density functions* whose integral give the probability.

Example: Gaussian Distribution PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



$\sigma = 0.5$
$\mu = 0$

Bernoulli Distribution

Bernoulli distribution describes a single binary outcome.

Bernoulli Distribution

Bernoulli distribution describes a single binary outcome.

Loaded coin example:

Assume we (arbitrarily) define ”heads” to be success.
Define the outcome as a random variable, C, with probability of success p.

Bernoulli Distribution

Bernoulli distribution describes a single binary outcome.

Loaded coin example:

Assume we (arbitrarily) define ”heads” to be success.
Define the outcome as a random variable, C, with probability of success p.

For example, when $p = 0.5$ this is a fair coin toss.

Bernoulli Distribution

Bernoulli distribution describes a single binary outcome.

Loaded coin example:

Assume we (arbitrarily) define ”heads” to be success.
Define the outcome as a random variable, C, with probability of success p.

$p = 0.2$ would be a ”tails” (i.e. failure) prone coin.

Bernoulli Distribution

Bernoulli distribution describes a single binary outcome.

Loaded coin example:

Assume we (arbitrarily) define ”heads” to be success.
Define the outcome as a random variable, C, with probability of success p.

$p = 0.9$ would be a ”heads” (i.e. success) prone coin.

Bernoulli Distribution

Bernoulli distribution describes a single binary outcome.

Loaded coin example:
Assume we (arbitrarily) define ”heads” to be success.
Define the outcome as a random variable, C, with probability of success p.

We use the following notation to denote that C is a random variable distributed with parameter p:

$C \sim \text{Bernoulli}(p)$

Bernoulli Distribution

Bernoulli distribution describes a single binary outcome.

Loaded coin example:
Assume we (arbitrarily) define ”heads” to be success.
Define the outcome as a random variable, C, with probability of success p.

We use the following notation to denote that C is a random variable distributed with parameter p:

$C \sim \text{Bernoulli}(p)$

Fair coin toss: $C \sim \text{Bernoulli}(p = 0.5)$

Binomial Distribution

Think of repeating n coin tosses, and measuring the number of successes.

How many parameters do we need to think about this, and what are those parameters?

.

Binomial Distribution

Think of repeating n coin tosses, and measuring the number of successes.

How many parameters do we need to think about this, and what are those parameters?

p: probability of success
n: number of trials

.

Binomial Distribution

Think of repeating n coin tosses, and measuring the number of successes.

How many parameters do we need to think about this, and what are those parameters?

p: probability of success
n: number of trials

If we have a random variable, X, that is the number of successes in n independent experiments (or tests, or trials) with each succeeding with probability p, then $X \sim \text{Binomial}(n, p)$.

Binomial Distribution

Think of repeating n coin tosses, and measuring the number of successes.

How many parameters do we need to think about this, and what are those parameters?

p: probability of success
n: number of trials

If we have a random variable, X, that is the number of successes in n independent experiments (or tests, or trials) with each succeeding with probability p, then $X \sim \text{Binomial}(n, p)$.

If $X \sim \text{Binomial}(n, p)$ then for an arbitrary non-negative $k \leq n$, we can calculate the probability of k successes using the Binomial PMF:

$$P(k; n, p) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Beta Distribution

What if we don't know whether if the coin is loaded, and if yes, exactly how it is loaded?

What if the probability of success p is, in turn, another random variable?

Beta Distribution

What if we don't know whether if the coin is loaded, and if yes, exactly how it is loaded?

What if the probability of success p is, in turn, another random variable?

How do we represent our belief in the values that p is likely to assume?

Beta Distribution

Beta distribution is one method for describing the probability density associated with a continuous random variable constrained to the (0,1) interval.

Beta Distribution

Beta distribution is one method for describing the probability density associated with a continuous random variable constrained to the (0,1) interval.

Given the parameters α and β, the beta PDF is:

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and $\Gamma(n) = (n-1)!$ when n is a positive integer.

Beta Distribution

Beta distribution is one method for describing the probability density associated with a continuous random variable constrained to the (0,1) interval.

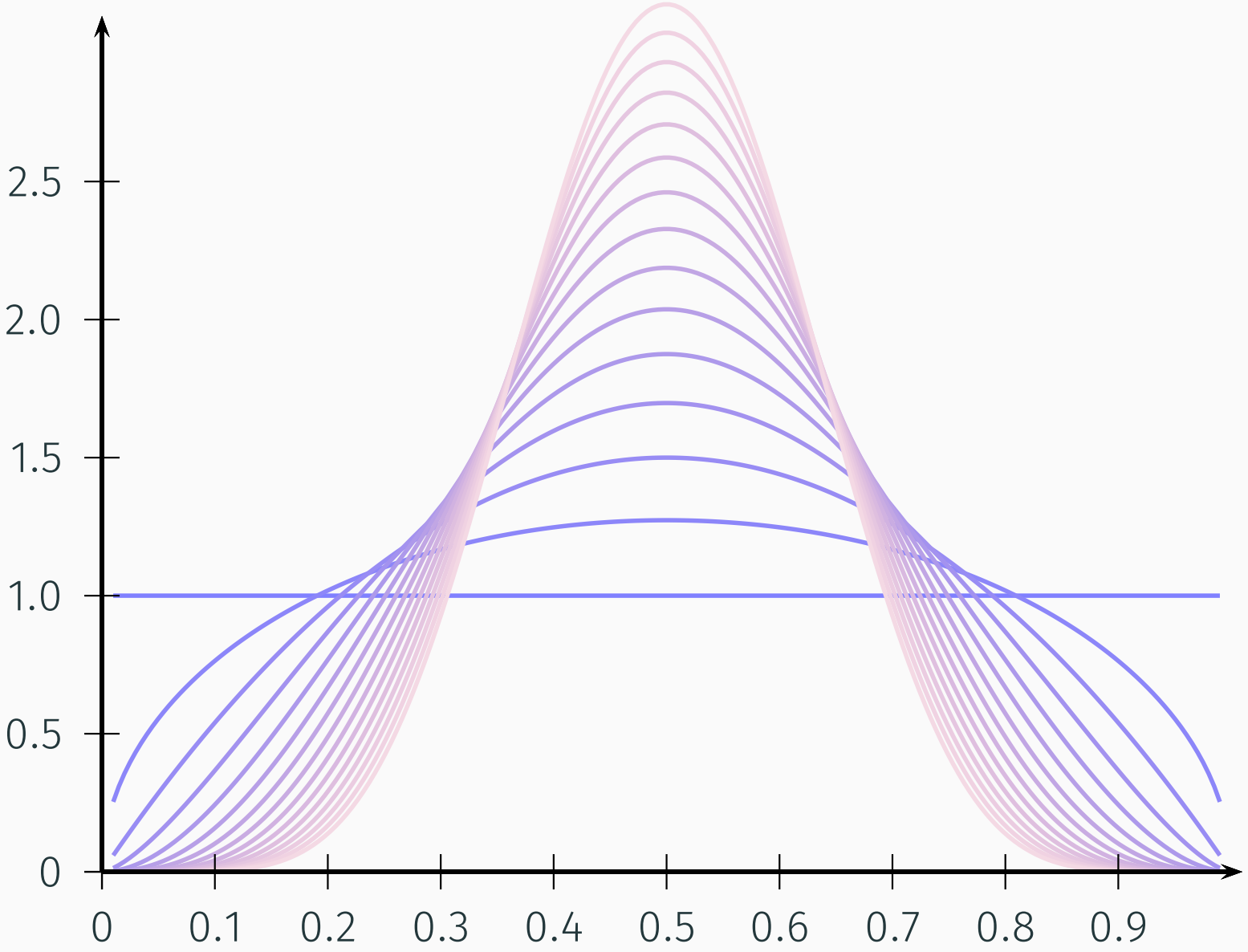
Given the parameters α and β, the beta PDF is:

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

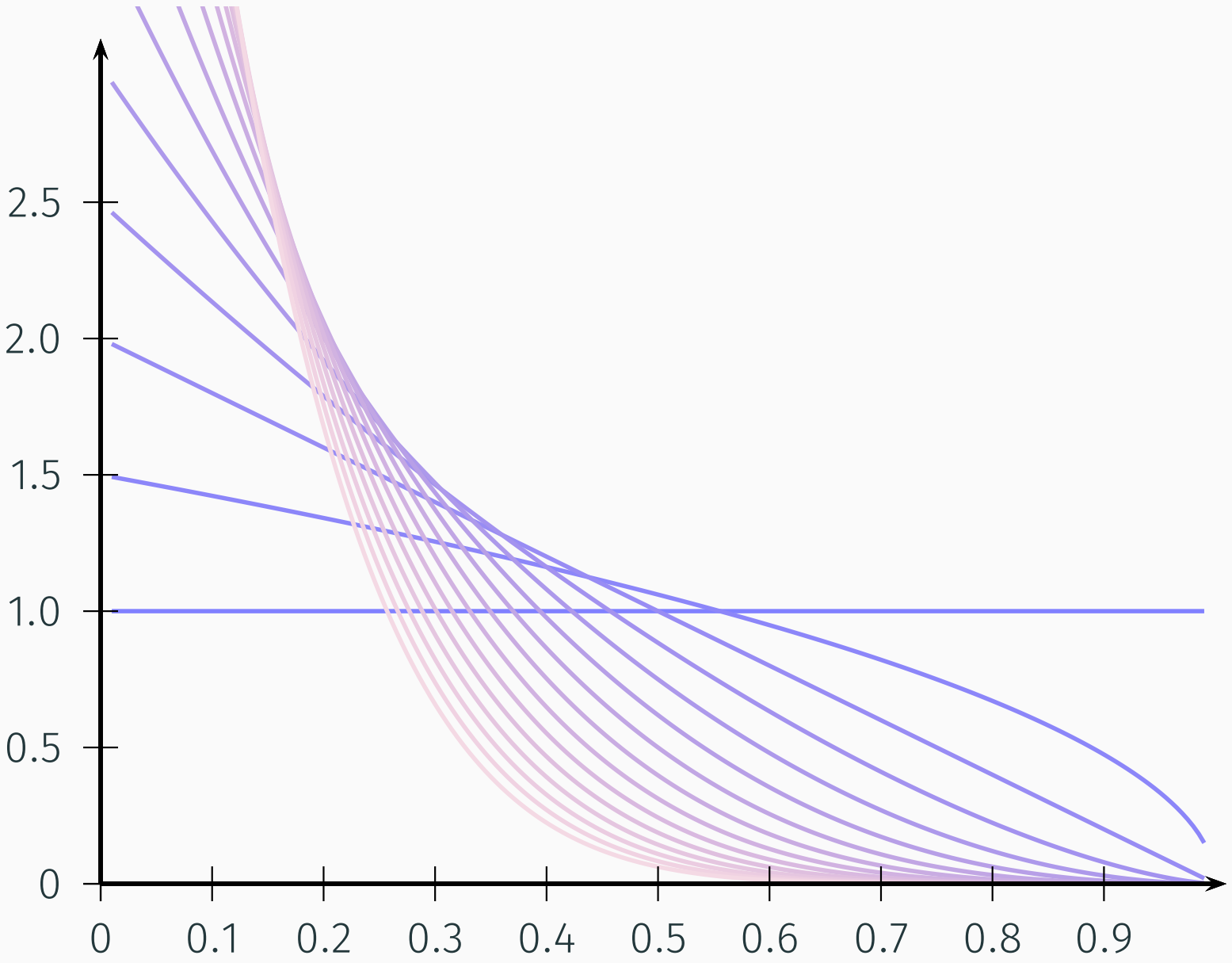
where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and $\Gamma(n) = (n-1)!$ when n is a positive integer.

Beta distribution is actually a particularly good choice: it is the conjugate prior distribution for Bernoulli and binomial distributions.

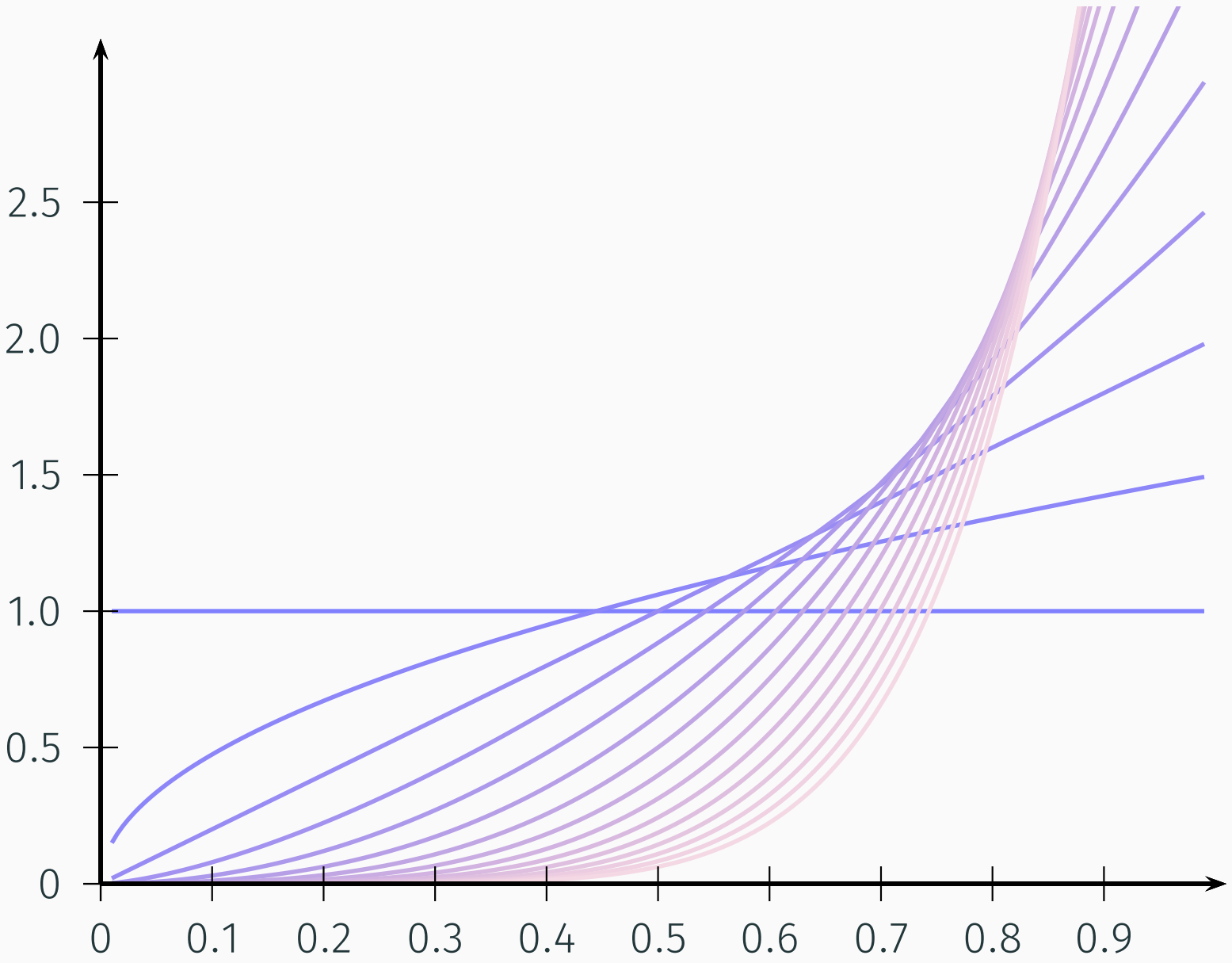
Beta Distribution: $\alpha = \beta = 1 + 0.5 * i, \forall i \in \{0, 1, \cdots, 15\}$



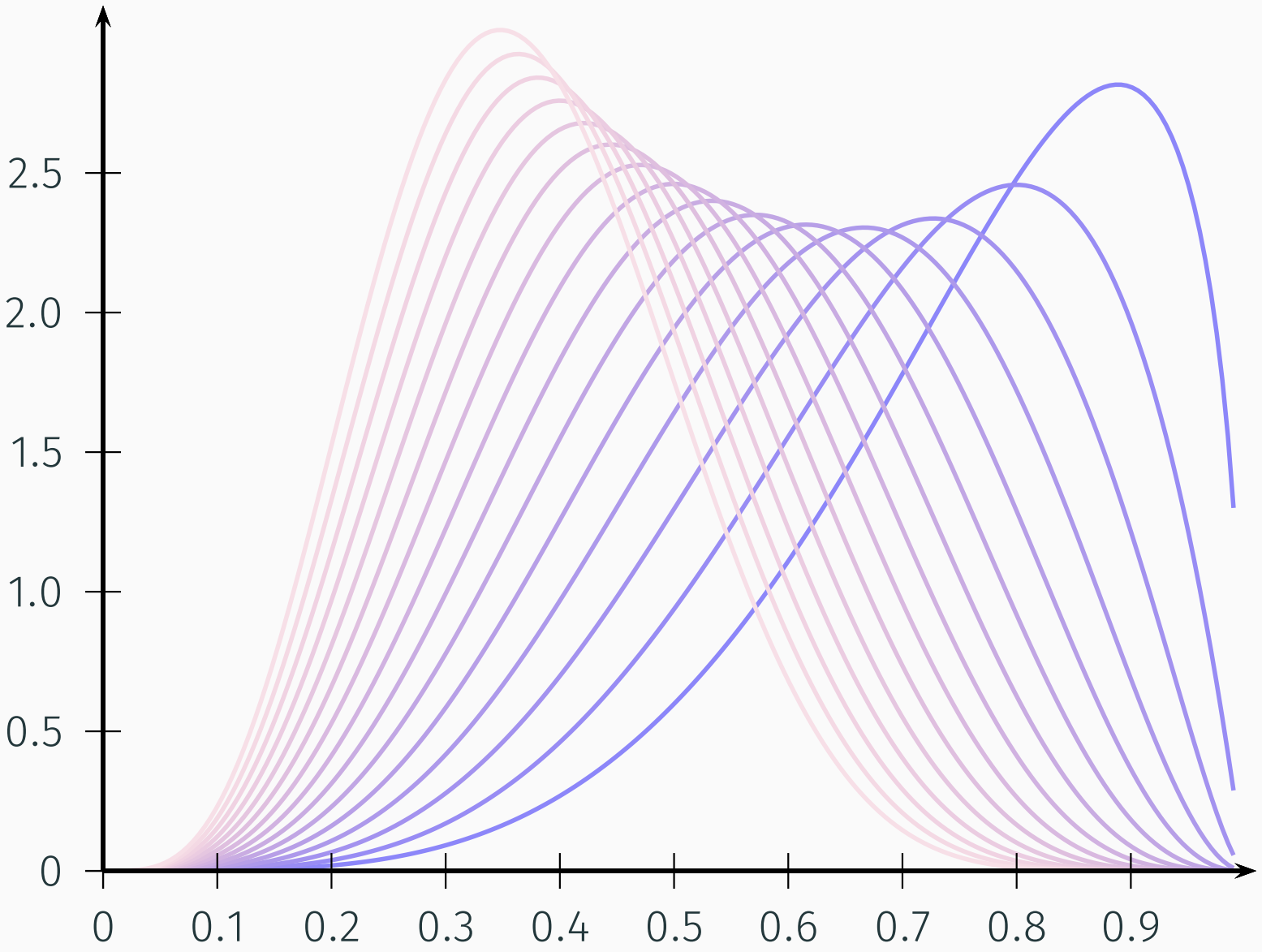
Beta Distribution: $\alpha = 1, \beta = 1 + 0.5 * i, \forall i \in \{0, 1, \cdots, 15\}$



Beta Distribution: $\beta = 1, \alpha = 1 + 0.5 * i, \forall i \in \{0, 1, \cdots, 15\}$



Beta Distribution: $\alpha = 5, \beta = 1 + 0.5 * i, \forall i \in \{1, \cdots, 15\}$



Beta Distribution: $\alpha = \beta * 1.5, \beta = 1 + 0.5 * i, \forall i \in \{1, \cdots, 15\}$