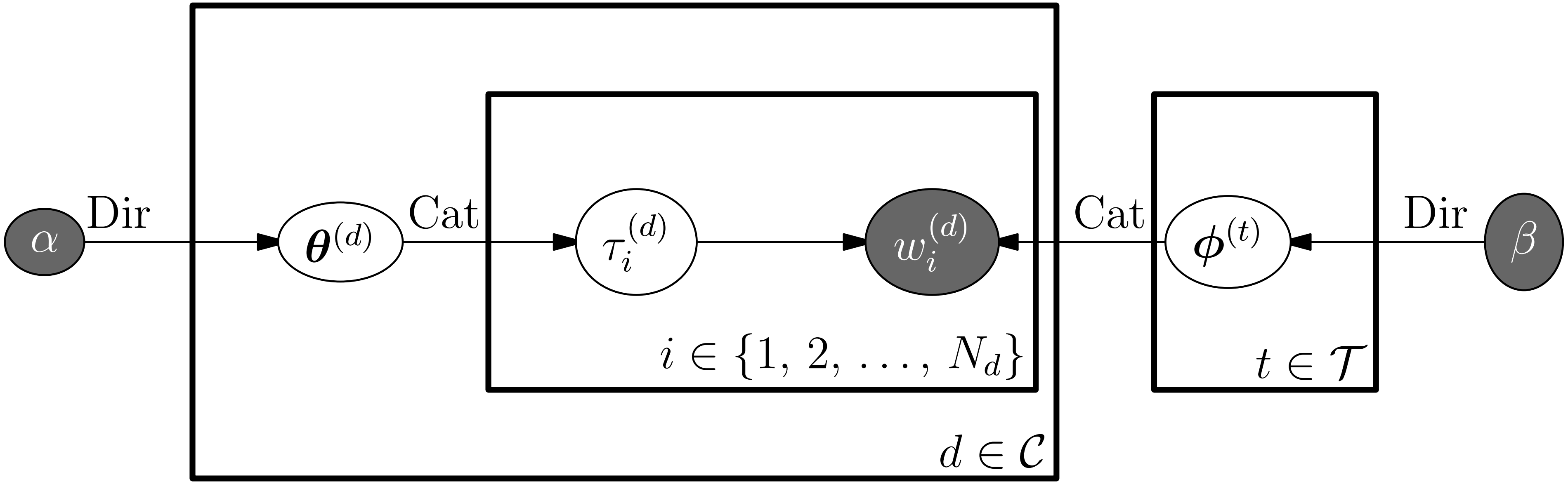
Advanced Machine Learning

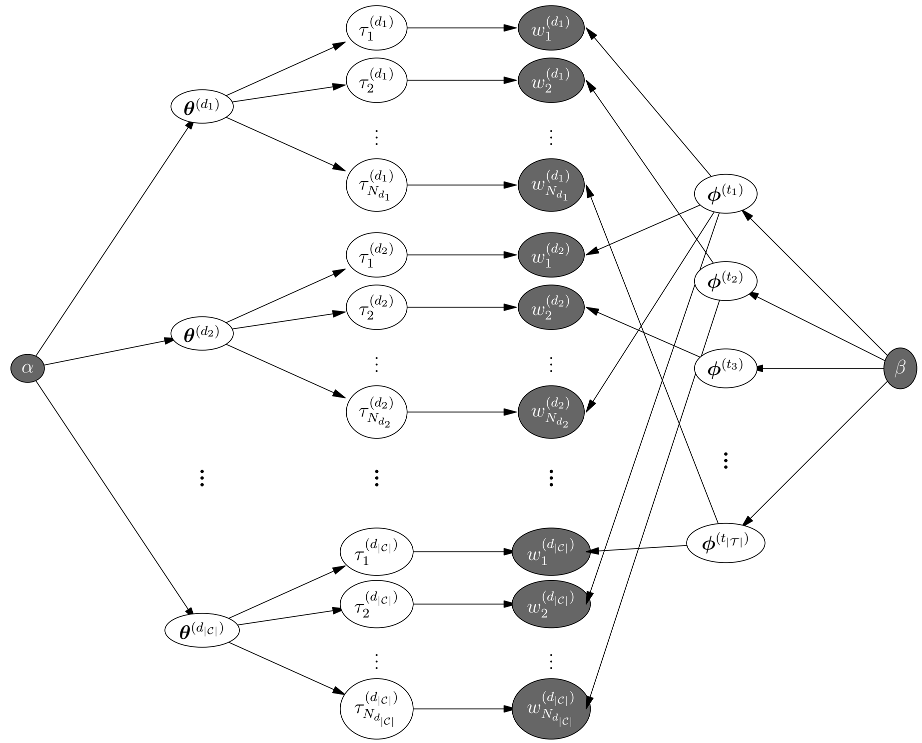
Generative Models



Generative models, graphical models, LDA

Outline

1. **Building Probabilistic Models**
2. Graphical Models
3. Latent Dirichlet Allocation



Building Probabilistic Models

- To describe a system with uncertainty we use random variables, X, Y, Z, etc.
- We use the convention of writing random variables in capitals (this is sometimes confusing as when you observe a random variables it is no longer random)
- The variables are described by probability mass function $\mathbb{P}(X, Y, Z)$ or if our variables are continuous, but probability densities $f_{X,Y,Z}(x, y, z)$
- We build in dependencies in this joint distribution

Discriminative Models

- We often think of our observations as given and the predictions as random variables
- For example we might be given some features $\boldsymbol{x}$ and we wish to predict a class $C \in \mathcal{C}$
- Our objective is then to find the probability $\mathbb{P}\left(C|\boldsymbol{x}\right)$
- This is known as a **discriminative model**
- E.g. in *foundations of machine learning* you learnt how to find the Bayes' optimal discrimination surface

Generative Models

- Sometimes it is easy to think about the joint process of generating the features and outputs together
- This leads to a joint distribution $\mathbb{P}(\boldsymbol{X}, Y)$ where $\boldsymbol{X}$ are your features and Y is your output you are trying to predict
- This is known as a **generative model**
- Generative models are often more natural to think about
- We can use them to do discrimination using

$$\mathbb{P}(Y|\boldsymbol{X}) = \frac{\mathbb{P}(\boldsymbol{X}, Y)}{\mathbb{P}(\boldsymbol{X})} = \frac{\mathbb{P}(\boldsymbol{X}, Y)}{\sum_{Y} \mathbb{P}(\boldsymbol{X}, Y)}$$

Latent Variables

- Sometimes we have models that involve random variables that we don't observe and we don't care about
- These are called **latent variables**
- If we have a latent variable Z and observed variable $\boldsymbol{X}$ and we are predicting a variable Y then we would **marginalise** over the latent variable

$$\mathbb{P}(\boldsymbol{X}, Y) = \sum_{Z} \mathbb{P}(\boldsymbol{X}, Y, Z)$$

Mixture of Gaussians

- Suppose we were observing the decays from two types of short-lived particle
- We observe the half life, X, but not the particle type
- We assume X is normally distributed with unknown means and variances: $\boldsymbol{\Theta} = \{\mu_1,\, \sigma_1^2,\, \mu_2,\, \sigma_2^2\}$
- Let $Z \in \{0, 1\}$ be an indicator that it is particle 1
- The probability of X is given by

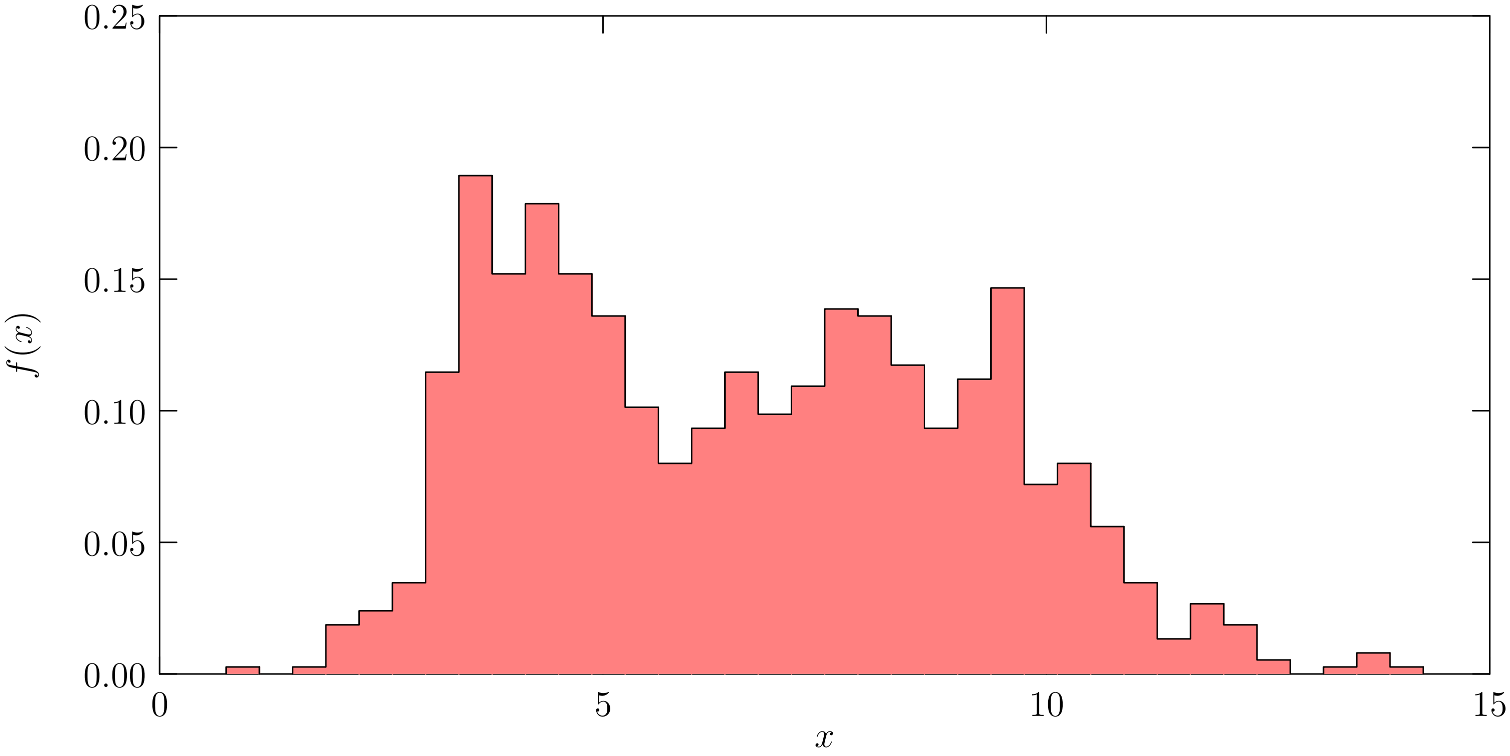
$$f(X|Z, \boldsymbol{\Theta}) = Z\,\mathcal{N}\big(X \big| \mu_1, \sigma_1^2\big) + (1 - Z)\,\mathcal{N}\big(X \big| \mu_2, \sigma_2^2\big)$$

Data

- Note that

$$f(X|\boldsymbol{\Theta}) = \sum_{Z\in\{0,1\}} f(X,Z|\boldsymbol{\Theta}) = \sum_{Z\in\{0,1\}} f(X|Z,\boldsymbol{\Theta})\,\mathbb{P}(Z)$$

$$= \mathbb{E}_Z[f(X|Z,\boldsymbol{\Theta})] = p\,\mathcal{N}\big(X\big|\mu_1,\sigma_1^2\big) + (1-p)\,\mathcal{N}\big(X\big|\mu_2,\sigma_2^2\big)$$



Maximum Likelihood

- To solve the model as a Bayesian we would have to assign priors to our parameters $\boldsymbol{\Theta} = (\mu_1, \sigma_1, \mu_2, \sigma_2, p)$

- This is doable, but complicated (we would also end up with a distribution for our parameters)

- Often we only want a reasonable estimate for some of our parameters (e.g. the half-lives μ_1 and μ_2)

- A reasonable approach is to seek those parameters that maximise the likelihood of our observed data

$$f(\mathcal{D}|\boldsymbol{\Theta}) = \prod_{X \in \mathcal{D}} f(X|\boldsymbol{\Theta})$$

EM Algorithm

- The maximum likelihood is a non-linear function of the parameters so cannot be immediately maximised

- We have a difficulty in that our latent variable Z will depend on the parameter $\boldsymbol{\Theta}$

- And our likelihood will depend on the latent variable

- We therefore proceed iteratively by maximising the expected log-likelihood with respect to the current set of parameters

$$\Theta^{(t+1)} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}} \sum_{\boldsymbol{Z}} \mathbb{P}\left(\boldsymbol{Z}|\mathcal{D}, \boldsymbol{\Theta}^{(t)}\right) \log(f(\mathcal{D}|\boldsymbol{Z}, \boldsymbol{\Theta}))$$

- This is known as the **expectation maximisation algorithm**

EM for Mixture of Gaussians

- Maximise with respect to parameters $\boldsymbol{\theta}$

$$
\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \sum_{\boldsymbol{Z}} \mathbb{P}\left(\boldsymbol{Z}|\mathcal{D}, \boldsymbol{\Theta}^{(t)}\right) \log(f(\mathcal{D}|\boldsymbol{Z}, \boldsymbol{\Theta})) \\
&= \sum_{i=1}^{n} \sum_{Z_i \in \{1,2\}} \mathbb{P}\left(Z_i|X_i, \boldsymbol{\theta}_i\right) \bigg(Z_i \log(p) + (1 - Z_i) \log(1 - p) \\
&\qquad\qquad + \frac{(X_i - \mu_{Z_i})^2}{2\,\sigma_{Z_i}^2} - \log\left(\sqrt{2\,\pi}\,\sigma_{Z_i}\right)\bigg)
\end{aligned}
$$

- Compute update equations

$$
\frac{\partial\, Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial\, \mu_k} = 0, \qquad \frac{\partial\, Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial\, \sigma_k} = 0, \qquad \frac{\partial\, Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial\, p} = 0
$$

Update Equations

- Means

$$\mu_{Z_i}^{(t+1)} = \frac{\sum_{i=1}^{n} \mathbb{P}\left(Z_i | X_i, \boldsymbol{\theta}^{(t)}\right) X_i}{\sum_{i=1}^{n} \mathbb{P}\left(Z_i | X_i, \boldsymbol{\theta}^{(t)}\right)},$$

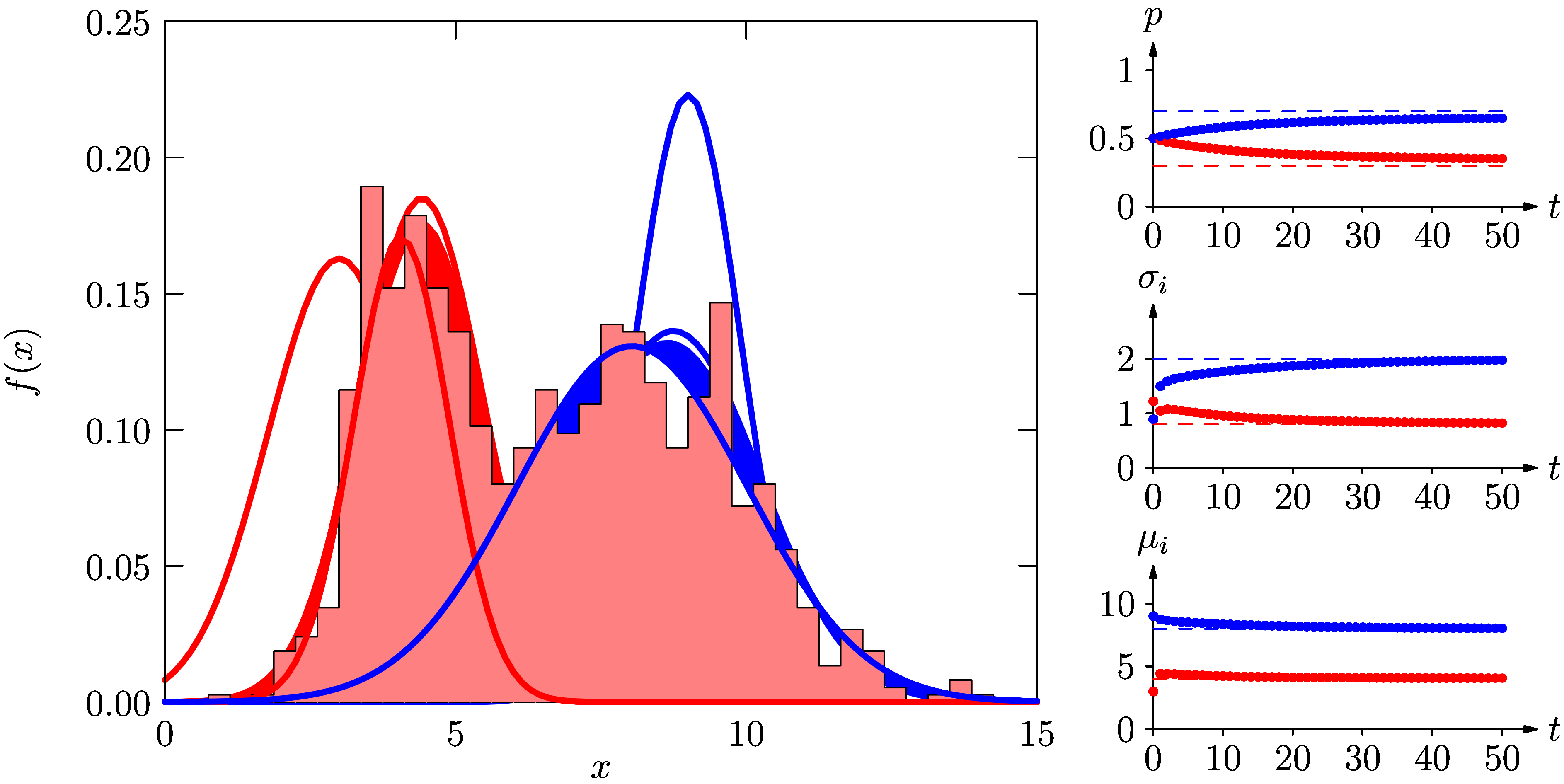
- Variances

$$(\sigma_{Z_i}^{(t+1)})^2 = \frac{\sum_{i=1}^{n} \mathbb{P}\left(Z_i | X_i, \boldsymbol{\theta}^{(t)}\right) (X_i - \mu_{Z_i}^{(t+1)})^2}{\sum_{i=1}^{n} \mathbb{P}\left(Z_i | X_i, \boldsymbol{\theta}^{(t)}\right)}$$

- Probability of being type 1

$$p^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}\left(Z_i | X_i, \boldsymbol{\theta}_i\right)$$

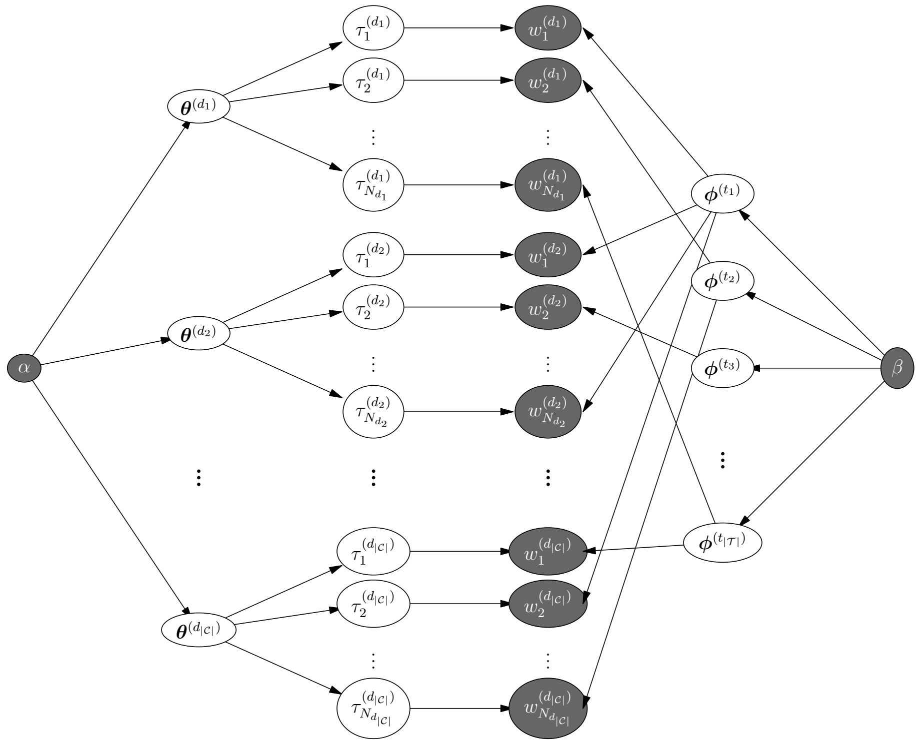
Example



0.25
0.20
0.15
0.10
0.05
0.00
$f(x)$
0
5
10
15
x
p
1
0.5
0
0 10 20 30 40 50
t
σ_i
2
1
0
0 10 20 30 40 50
t
μ_i
10
5
0
0 10 20 30 40 50
t

Outline

1. Building Probabilistic Models
2. **Graphical Models**
3. Latent Dirichlet Allocation



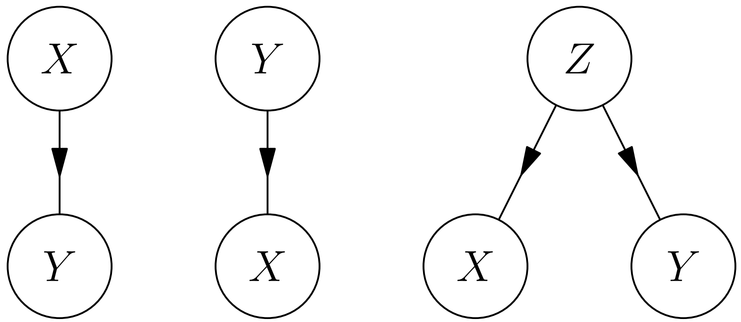
Dependencies Between Variables

- In building a probabilistic model we want to know which random variables depend on each other directly and which don't
- Variables that don't will typically still be correlated
- If two random variables X and Y are correlated then
  - ⋆ X could affect Y
  - ⋆ Y could affect X
  - ⋆ X and Y could not influence each other, but both be affected by another random variable Z

Graphical Models

- Graphical models are directed graphs that show causal relationships between random variables

- We could represent the three conditions described above by



- We can use these graphical representations to work out how to efficiently average over latent variables

Statistical Independence

- Two random variables are statistically independent if

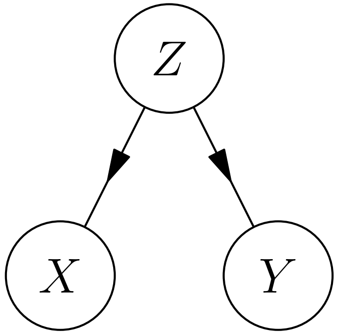
$$\mathbb{P}(X, Y) = \mathbb{P}(X)\, \mathbb{P}(Y)$$

- Equally this implies $\mathbb{P}(X|Y) = \mathbb{P}(X)$ and $\mathbb{P}(Y|X) = \mathbb{P}(Y)$
- Statistically independent variables are uncorrelated
- But statistical independence is often too powerful

Conditional Independence

- A weaker notion is conditional independence

$$\mathbb{P}(X, Y|Z) = \mathbb{P}(X|Z)\,\mathbb{P}(Y|Z)$$



- Conditional independence implies that there is no direct causation

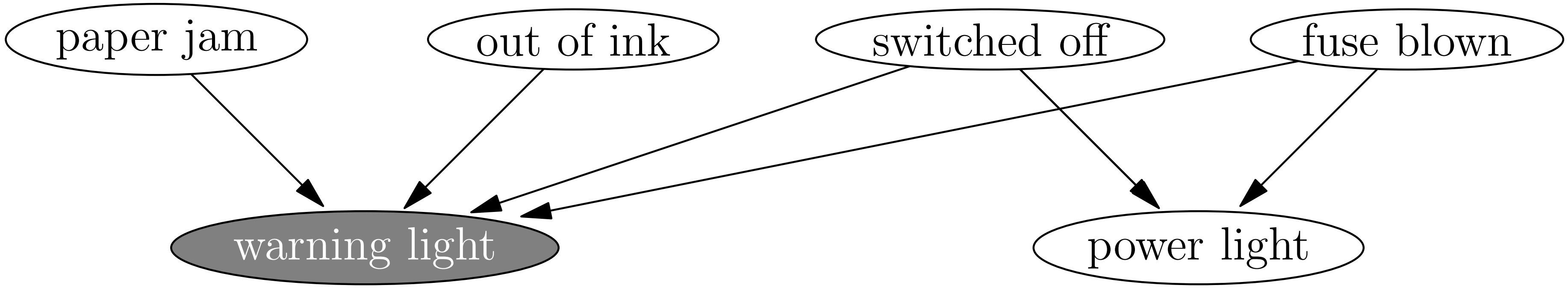
- But it doesn't imply zero correlation

- Conditional independence reduces computational complexity, e.g.

$$\mathbb{E}[X\,Y] = \sum_{X,Y,Z} X\,Y\,\mathbb{P}(X,Y,Z) = \sum_{Z} P(Z)\left(\sum_X X\,P(X|Z)\right)\left(\sum_Y Y\,P(Y|Z)\right)$$

Graphical Models

- Graphical models often provide a quick way to represent the world

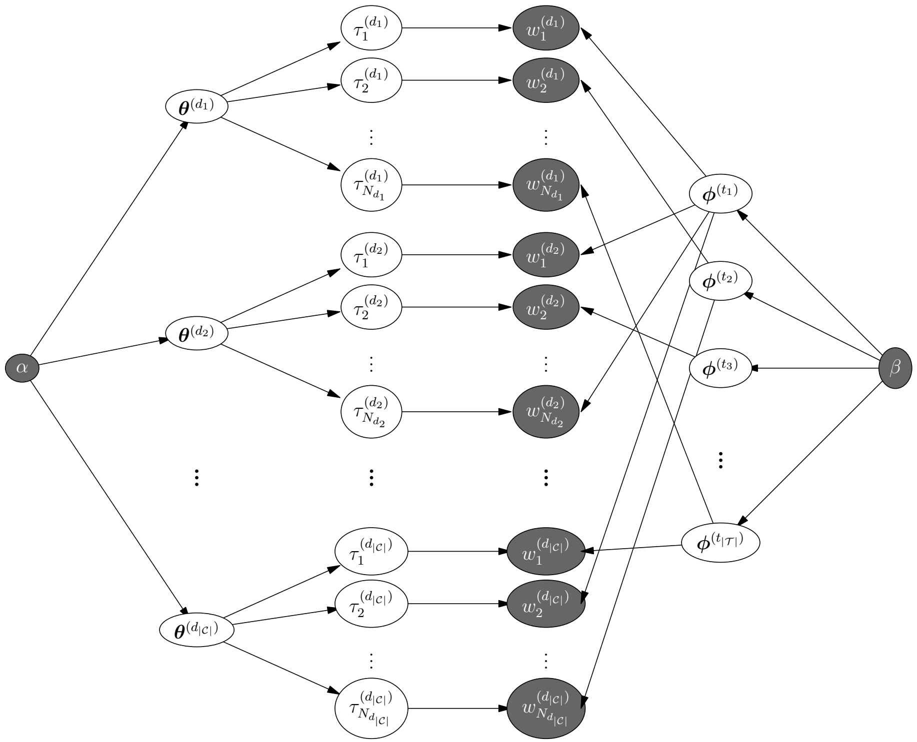


- In graphical models we shade nodes that we observe

- Note that the top events are conditionally independent if we make no observation, but are dependent if we observe a warning light!

Outline

1. Building Probabilistic Models
2. Graphical Models
3. **Latent Dirichlet Allocation**



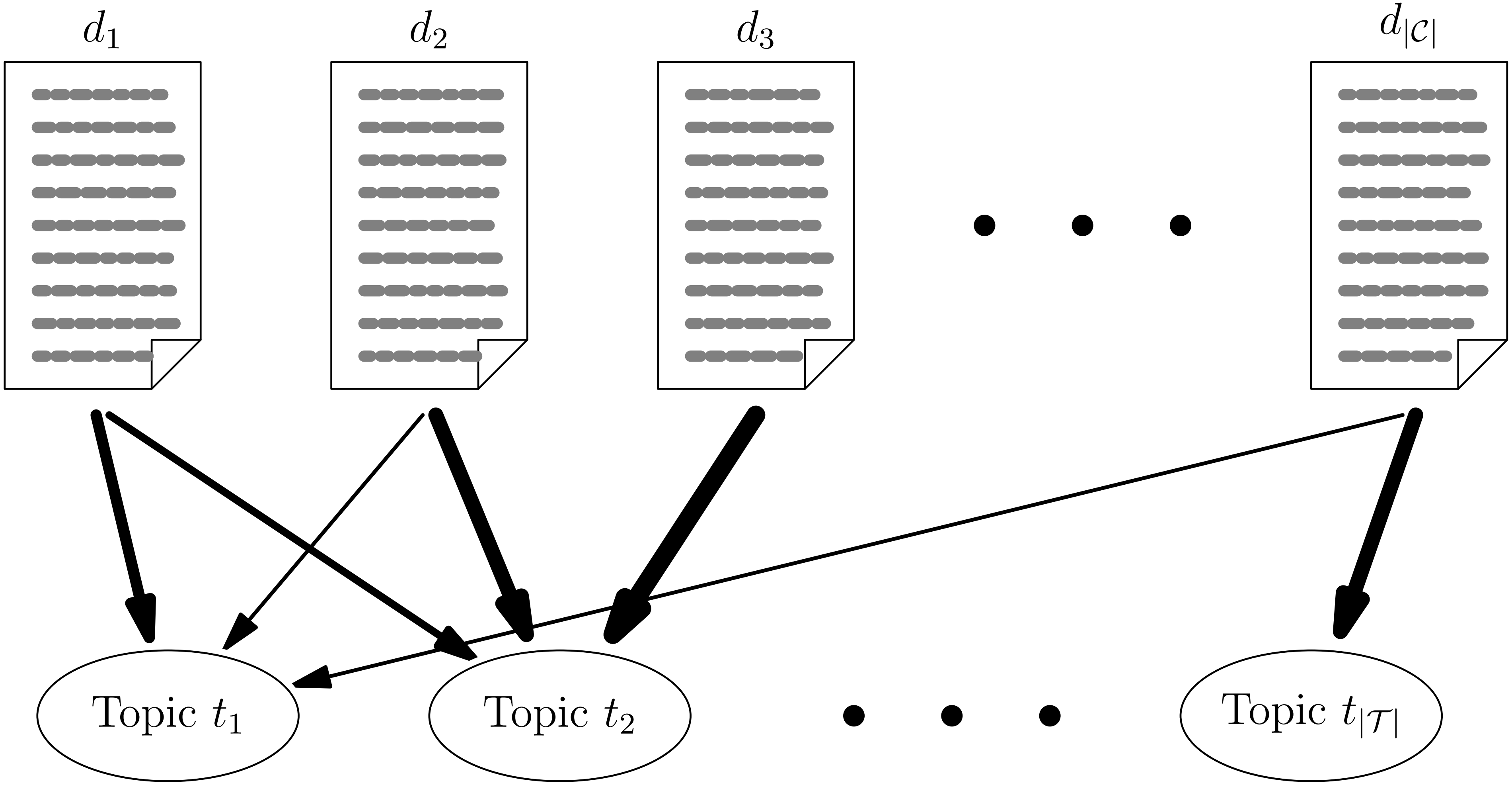
Model for Documents

- We consider a model for the words in a set of documents (we ignore word order)
- We consider a corpus $\mathcal{C} = \{d_i | i = 1, 2, \dots |\mathcal{C}|\}$
- With documents consisting of words

$$d = \left(w_1^{(d)}, w_2^{(d)}, \dots, w_{N_d}^{(d)} \right)$$

- We assume that there is a set of topics $\mathcal{T} = \{t_1, t_2, \dots, t_{|\mathcal{T}|}\}$
- We associate a probability, $\theta_t^{(d)}$, that a word in document d relates to a topic t

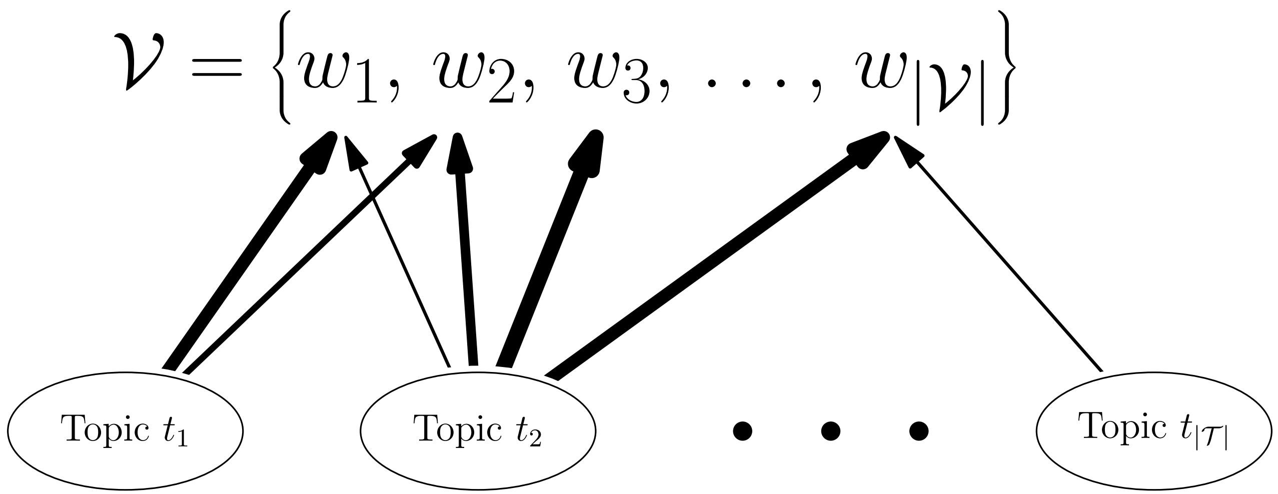
Documents and Topic



$$\boldsymbol{\theta}^{(d)} = (\theta_t^{(d)} | t \in \mathcal{T})$$

Words and Topic

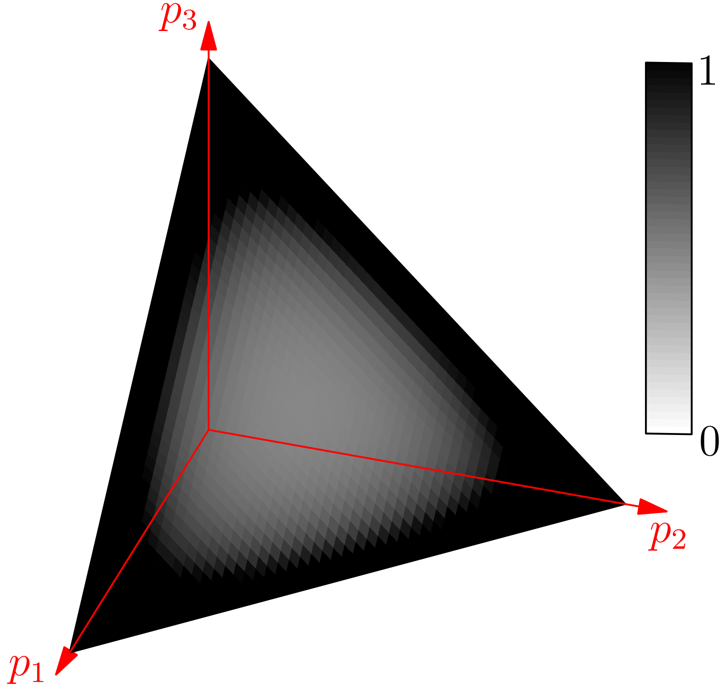
- We associate a probability $\phi_w^{(t)}$ that a word, w, is related to a topic t



$$\boldsymbol{\phi}^{(t)} = (\phi_w^{(t)} | w \in \mathcal{V})$$

Dirichlet Allocation

- Most documents are predominantly about a few topics and most topic have a small number of words associated to them
- We can generate sparse vectors $\boldsymbol{\theta}^{(d)}$ and $\boldsymbol{\phi}^{(t)}$ from a Dirichlet distribution with small parameters $\boldsymbol{\alpha}$



$$\mathrm{Dir}(\boldsymbol{p}|\boldsymbol{\alpha}) = \Gamma\left(\sum_i \alpha_i\right) \prod_{i=1}^{n} \frac{p_i^{\alpha_i - 1}}{\Gamma(\alpha_i)}$$

$$\boldsymbol{\theta}^{(d)} \sim \mathrm{Dir}(\alpha\, \mathbf{1})$$
$$\boldsymbol{\phi}^{(t)} \sim \mathrm{Dir}(\beta\, \mathbf{1})$$

Generating Document

- To generate a document we choose a topic for each word and a word for each topic

$$\forall d \in \mathcal{C} \quad \boldsymbol{\theta}^{(d)} \sim \mathrm{Dir}(\alpha\,\mathbf{1})$$

$$\forall t \in \mathcal{T} \quad \boldsymbol{\phi}^{(t)} \sim \mathrm{Dir}(\beta\,\mathbf{1})$$

$$\forall d \in \mathcal{C} \;\wedge\; \forall i \in \{1, 2, \ldots, N_d\} \quad \tau_i^{(d)} \sim \mathrm{Cat}(\boldsymbol{\theta}^{(d)}),\;\; w_i^{(d)} \sim \mathrm{Cat}(\boldsymbol{\phi}^{(\tau_i^{(d)})})$$

- Where $\mathrm{Cat}(i|\boldsymbol{p}) = p_i$ is the categorical distribution (we choose one of a number of options)

- This model is known as **Latent Dirichlet Allocation**

LDA Graphical Model (version 1)

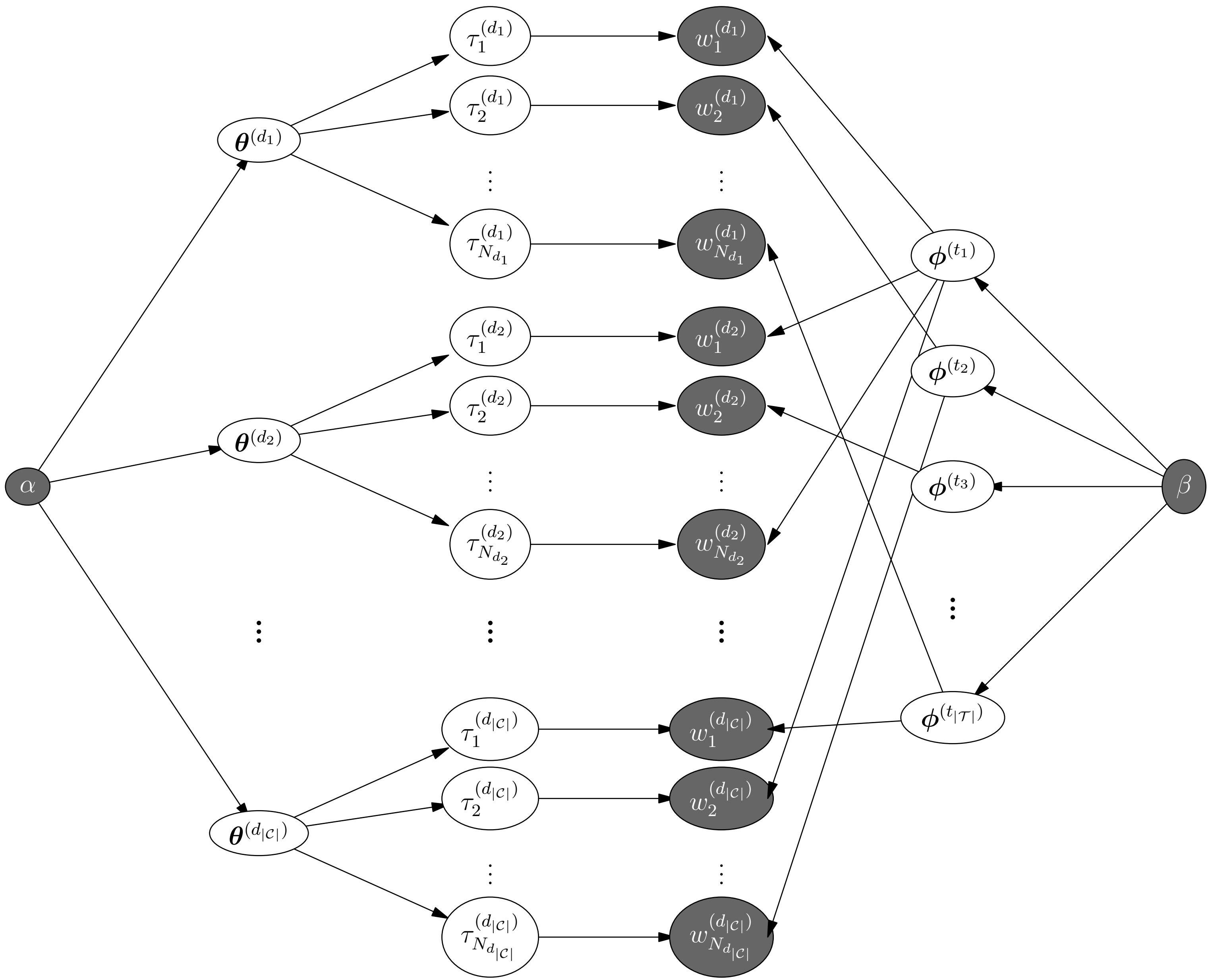
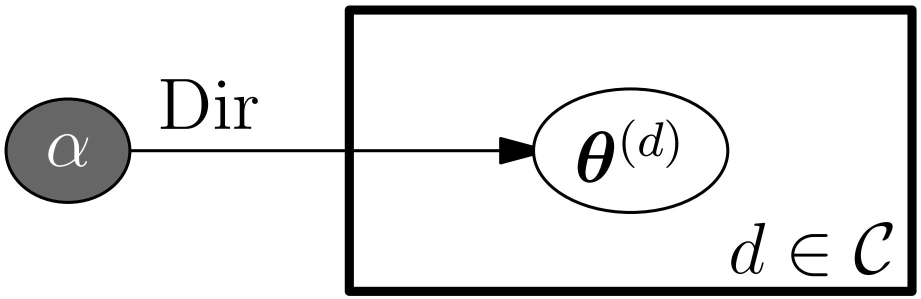


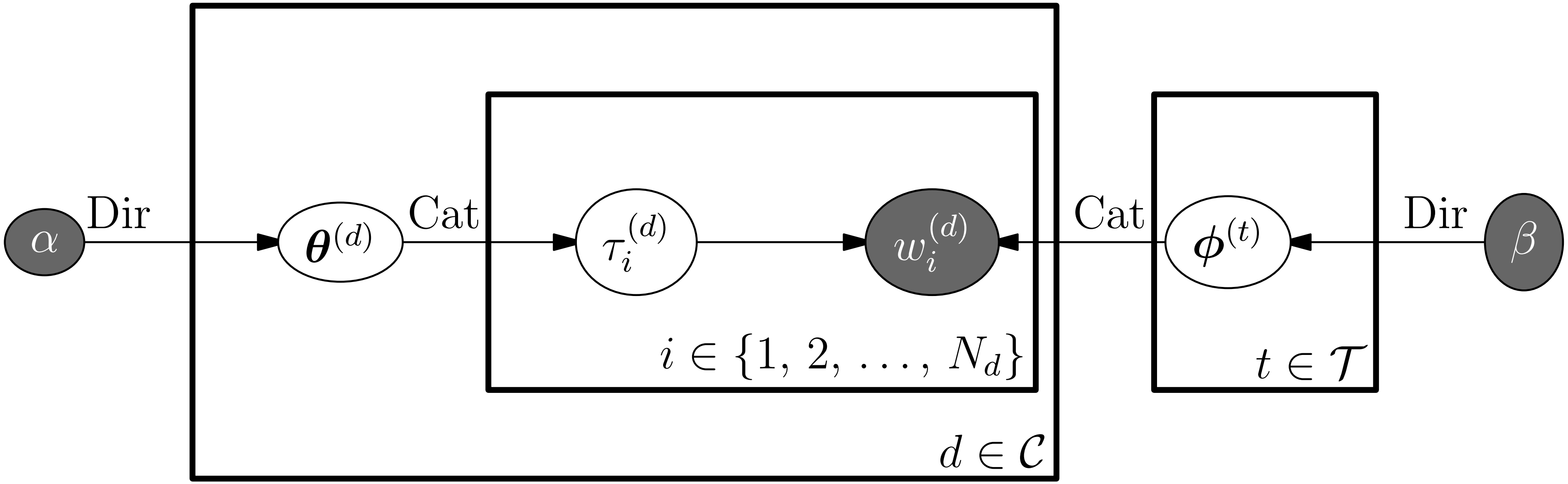
Plate Diagrams

- Drawing every random variable is tedious (and not really possible)
- A short-hand is to draw a box (plate) meaning repeat



- That is we generate vectors $\boldsymbol{\theta}^d$ from a Dirchelet distribution $\mathrm{Dir}\left(\boldsymbol{\theta}|\alpha\mathbf{1}\right)$ for all documents in corpus $\mathcal{C}$

LDA Graphical Model (version 2)



- This is a lot more compact
- Personally, I find it hard to read, but you get used to it

Probabilistic Model

- The graphical Model is shorthand for the variables

$$
\begin{aligned}
\boldsymbol{W} &= (\boldsymbol{w}^{(d)} | d \in \mathcal{C}) \quad \text{with} \quad \boldsymbol{w}^{(d)} = (w_1^{(d)}, w_2^{(d)}, \ldots, w_{N_d}^{(d)}), \quad \text{and} \quad w_i^{(d)} \in \mathcal{V} \\
\boldsymbol{T} &= (\tau_i^{(d)} | d \in \mathcal{C} \ \wedge \ i \in \{1, 2, \ldots, N_d\}) \quad \text{with} \quad \tau_i^{(d)} \in \mathcal{T} \\
\boldsymbol{\Theta} &= (\boldsymbol{\theta}^{(d)} | d \in \mathcal{C}) \quad \text{with} \quad \boldsymbol{\theta}^{(d)} = (\theta_t^{(d)} | t \in \mathcal{T}) \in \Lambda^{|\mathcal{T}|} \\
\boldsymbol{\Phi} &= (\boldsymbol{\phi}^{(t)} | t \in \mathcal{T}) \quad \text{with} \quad \boldsymbol{\phi}^{(t)} = (\phi_w^{(t)} | w \in \mathcal{V}) \in \Lambda^{|\mathcal{V}|}
\end{aligned}
$$

- Distributed according to

$$
\begin{aligned}
\mathbb{P}\left(\boldsymbol{W}, \boldsymbol{T}, \boldsymbol{\Theta}, \boldsymbol{\Phi} \middle| \alpha, \beta\right) = & \left(\prod_{t \in \mathcal{T}} \mathrm{Dir}\left(\boldsymbol{\phi}^{(t)} \middle| \beta \mathbf{1}\right) \right) \\
& \left(\prod_{d \in \mathcal{C}} \mathrm{Dir}\left(\boldsymbol{\theta}^{(d)} \middle| \alpha \mathbf{1}\right) \prod_{i=1}^{N_d} \mathrm{Cat}\left(\tau_i^{(d)} \middle| \boldsymbol{\theta}^{(d)}\right) \mathrm{Cat}\left(w_i^{(d)} \middle| \boldsymbol{\phi}^{(\tau_i^{(d)})}\right) \right)
\end{aligned}
$$

Finding Topics

- We are given the set of words $\boldsymbol{W}$ and don't really care about τ_i^d the topic associated with word i in document d
- But we are interested in the words associated with each topic $\boldsymbol{\phi}^{(t_i)}$
- And the topics associated with each document $\boldsymbol{\theta}^{(d)}$
- To compute them we need to sample the probability distribution
- One way to do this is using Monte Carlo methods (see next lecture)

Summary

- Building probabilistic models is an intricate process
- Identifying random variables that describe the system is the first step
- Graphical models provides a representation showing the causal relationship between random variables
- It is possible to generate very rich models such as Latent Dirchlet Allocation (LDA)