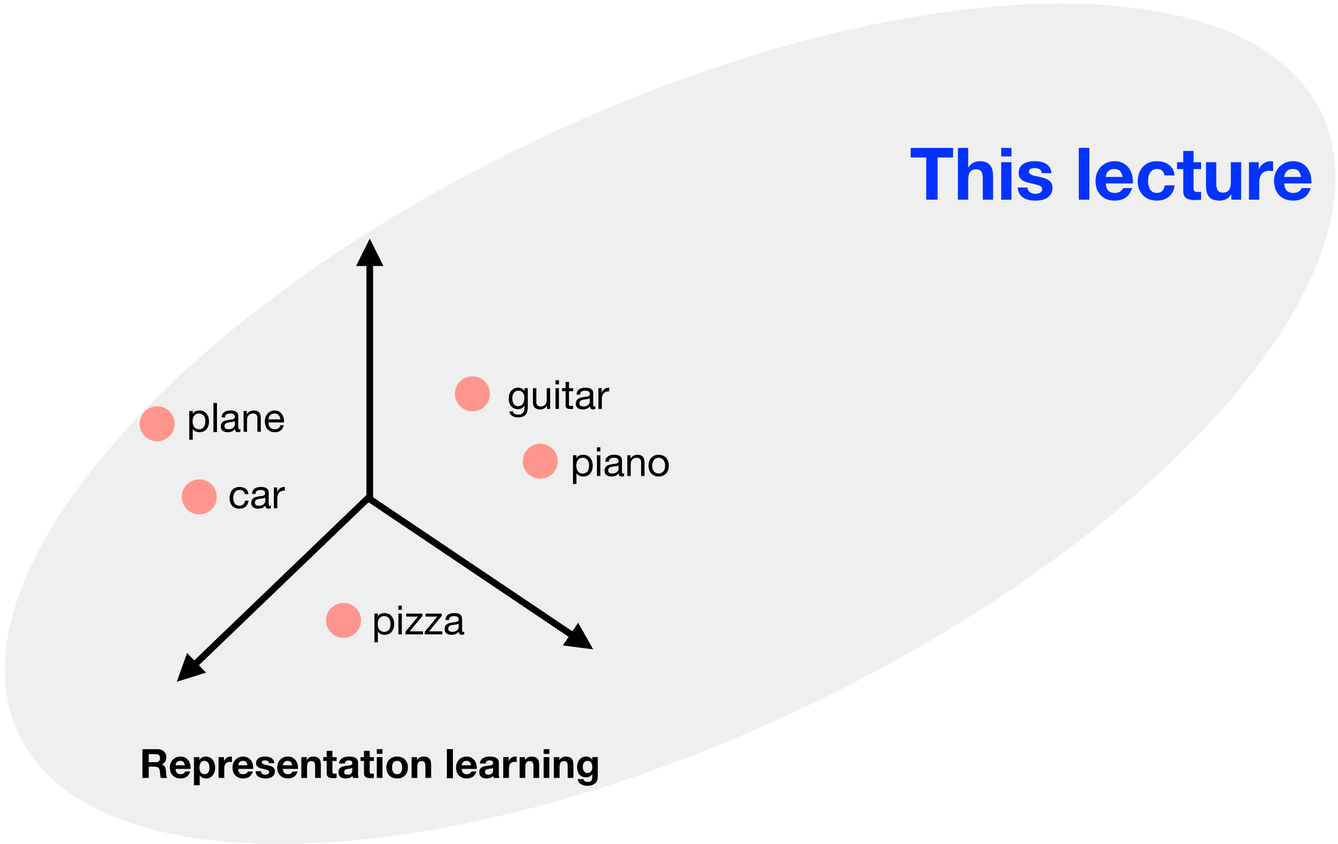
Adaptive Skip-gram

Sergey Bartunov, DeepMind

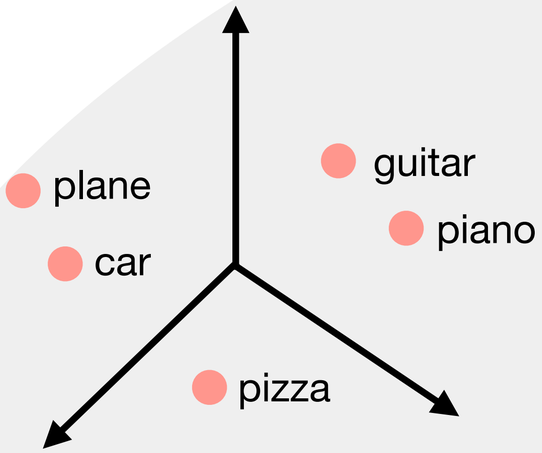
(joint work with Dmitry Kondrashkin, Anton Osokin and Dmitry Vetrov)

This lecture

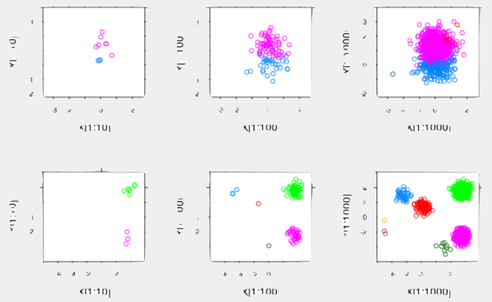
This lecture



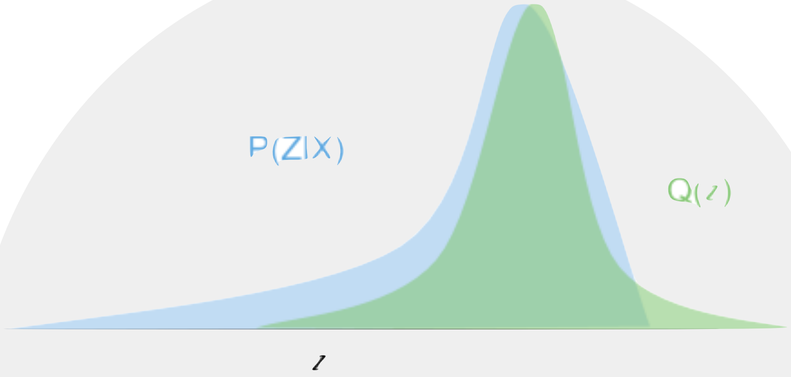
This lecture



Representation learning

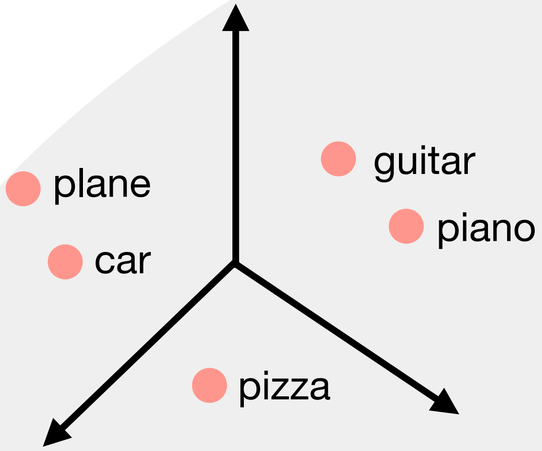


Bayesian nonparametrics

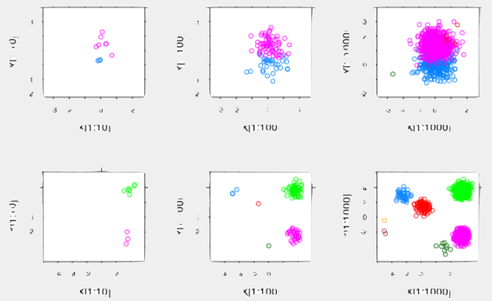


Variational inference

This lecture



Representation learning



Bayesian nonparametrics

Representation learning

Representation learning

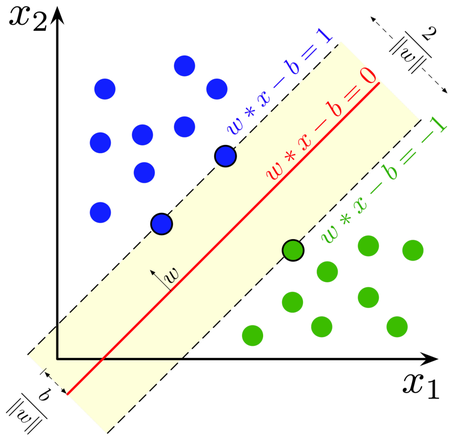
Elephants, the largest existing land animals, are mammals of the family Elephantidae. Three species are currently recognised: the African bush elephant, the African forest elephant, and the Asian elephant.

Data

Representation learning

Elephants, the largest existing land animals, are mammals of the family Elephantidae. Three species are currently recognised: the African bush elephant, the African forest elephant, and the Asian elephant.

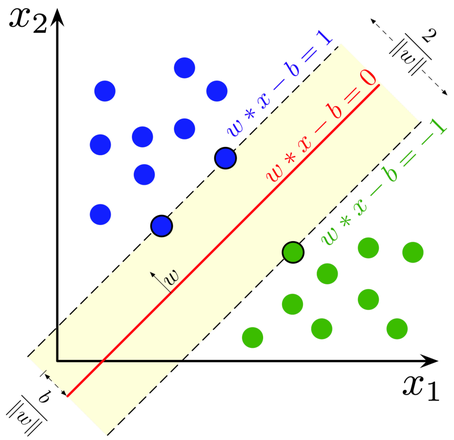
Data

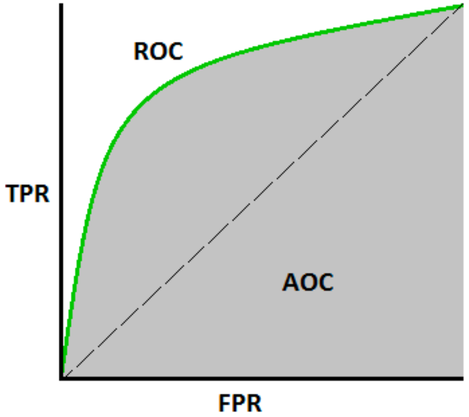


ML algorithm

Representation learning

Elephants, the largest existing land animals, are mammals of the family Elephantidae. Three species are currently recognised: the African bush elephant, the African forest elephant, and the Asian elephant.





Data

ML algorithm

Supervision / analysis

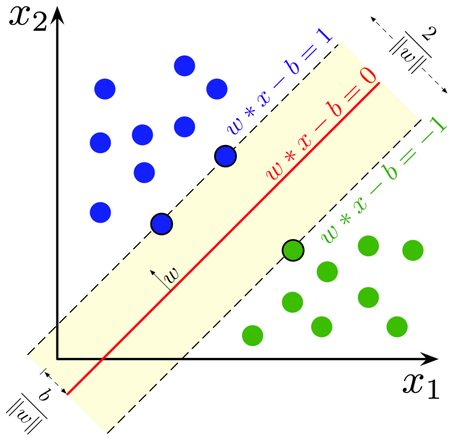
Representation learning

Elephants, the largest existing land animals, are mammals of the family Elephantidae. Three species are currently recognised: the African bush elephant, the African forest elephant, and the Asian elephant.

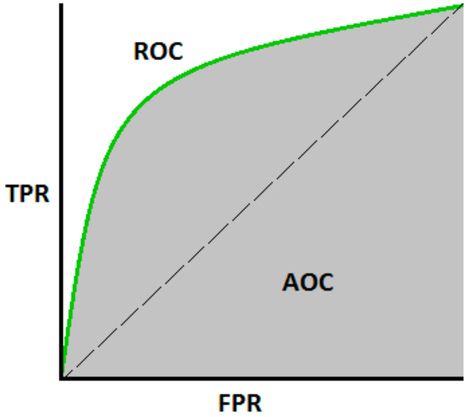
Data

0.34
-1.29
0.0
1.7
-2.34
3.3
0.0

Feature extraction



ML algorithm



Supervision / analysis

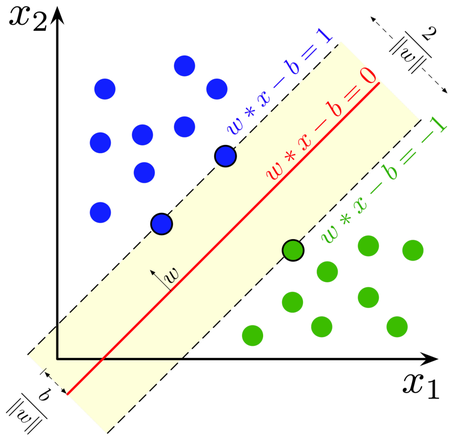
Representation learning

Elephants, the largest existing land animals, are mammals of the family Elephantidae. Three species are currently recognised: the African bush elephant, the African forest elephant, and the Asian elephant.

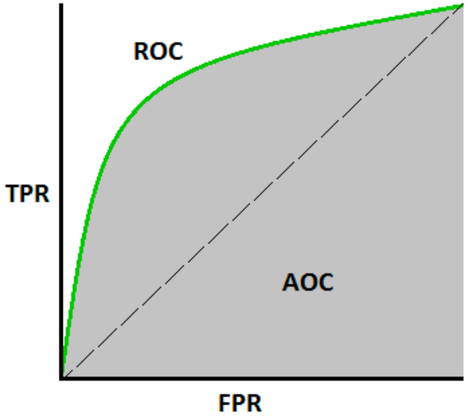
Data

0.34
-1.29
0.0
1.7
-2.34
3.3
0.0

Feature extraction



ML algorithm

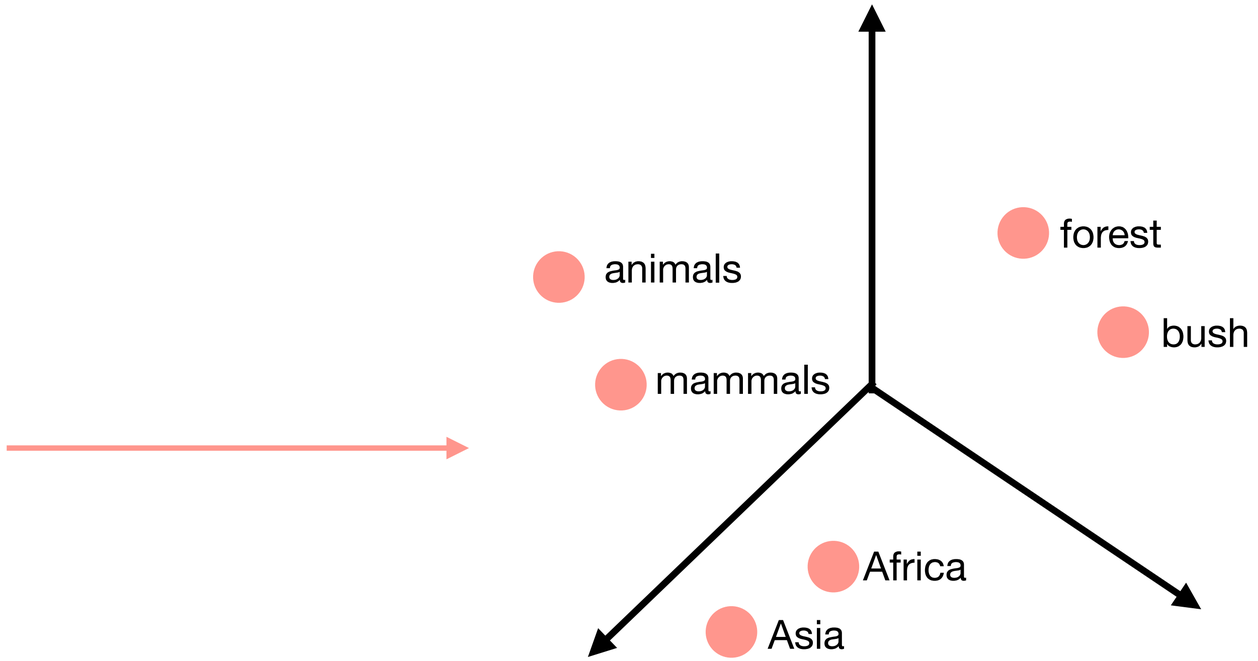


Supervision / analysis

Word embeddings

Elephants, the largest existing land animals, are mammals of the family Elephantidae. Three species are currently recognised: the African bush elephant, the African forest elephant, and the Asian elephant.

Text data



Goal: learn close embeddings for semantically similar words

Word embeddings in Skip-gram

Elephants, the largest existing land animals, are mammals of the family Elephantidae. Three species are currently recognised: the African bush elephant, the African forest elephant, and the Asian elephant.

Text data

Word	Input embedding	Output embedding
mammals	[0.5, 3.45, -1.26]	[0.5, 3.45, -1.26]
animals	[0.45, -0.5, -2.6]	[-0.93, 1.2, 0.]
land	[0.18, 0.45, 0.6]	[-1.52, -0.39, 0.89]
…		

Skip-gram model

…largest existing land animals, are **mammals** of the family Elephantidae...

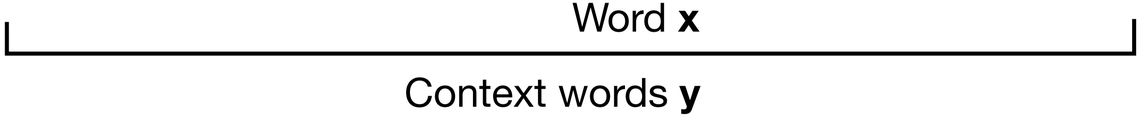
Skip-gram model

…largest existing land animals, are **mammals** of the family Elephantidae...

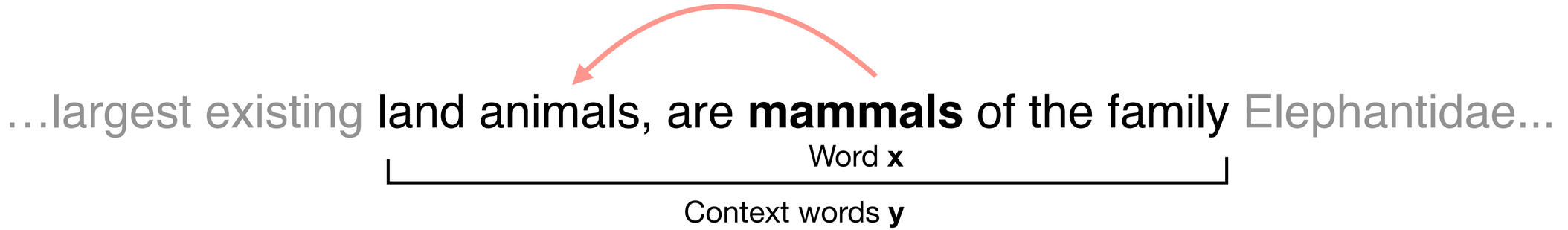
Word **x**

Skip-gram model

…largest existing land animals, are **mammals** of the family Elephantidae...

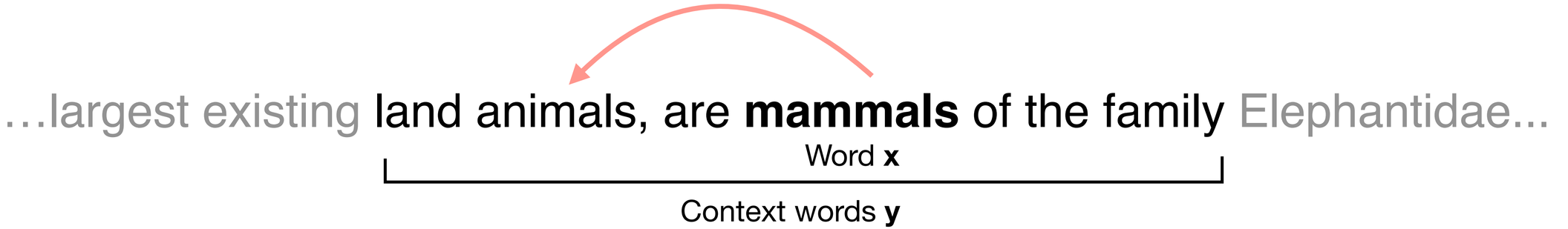


Skip-gram model



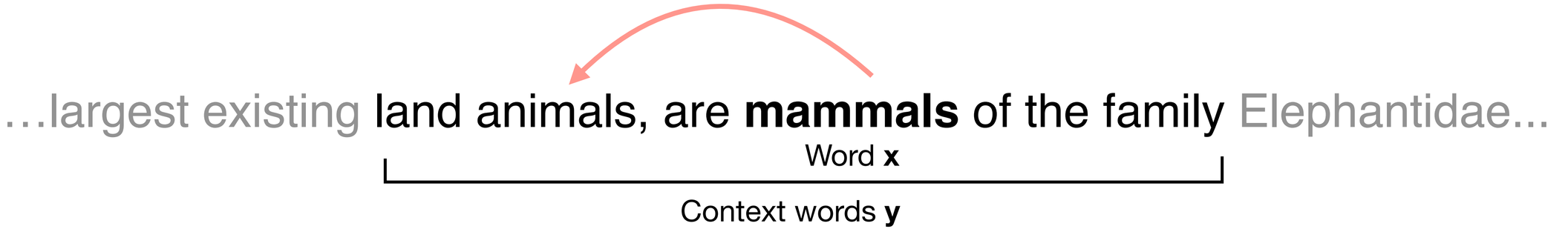
$$p(\mathbf{y}|x) = \prod_j p(y_j|x)$$

Skip-gram model



$$p(\mathbf{y}|x) = \prod_j p(y_j|x) \qquad p(v|w) = \frac{\exp(\text{in}_w^T \text{out}_v)}{\sum_{v'} \exp(\text{in}_w^T \text{out}_{v'})}$$

Skip-gram model



$$p(\mathbf{y}|x) = \prod_j p(y_j|x) \qquad p(v|w) = \frac{\exp(\text{in}_w^T \text{out}_v)}{\sum_{v'} \exp(\text{in}_w^T \text{out}_{v'})}$$

Distributional hypothesis: similar words appear in similar contexts

Training Skip-gram

Training Skip-gram

…largest existing land **animals**, are mammals of the family Elephantidae...

Training Skip-gram

…largest existing land **animals**, are mammals of the family Elephantidae...

Gradient update: $\theta \leftarrow \theta + \eta \nabla \log p(\mathbf{y}_i | x_i)$

Training Skip-gram

…largest existing land **animals**, are mammals of the family Elephantidae...

Gradient update: $\theta \leftarrow \theta + \eta \nabla \log p(\mathbf{y}_i | x_i)$

…largest existing land animals, **are** mammals of the family Elephantidae...

Training Skip-gram

…largest existing land **animals**, are mammals of the family Elephantidae...

Gradient update: $\theta \leftarrow \theta + \eta \nabla \log p(\mathbf{y}_i | x_i)$

…largest existing land animals, **are** mammals of the family Elephantidae...

Gradient update: $\theta \leftarrow \theta + \eta \nabla \log p(\mathbf{y}_i | x_i)$

Training Skip-gram

…largest existing land **animals**, are mammals of the family Elephantidae...

Gradient update: $\theta \leftarrow \theta + \eta \nabla \log p(\mathbf{y}_i | x_i)$

…largest existing land animals, **are** mammals of the family Elephantidae...

Gradient update: $\theta \leftarrow \theta + \eta \nabla \log p(\mathbf{y}_i | x_i)$

…largest existing land animals, are **mammals** of the family Elephantidae...

…

Summary

Summary

- Learns high-quality semantically rich embeddings

Summary

- Learns high-quality semantically rich embeddings
- Sparse gradients (not explaining why in this lecture)

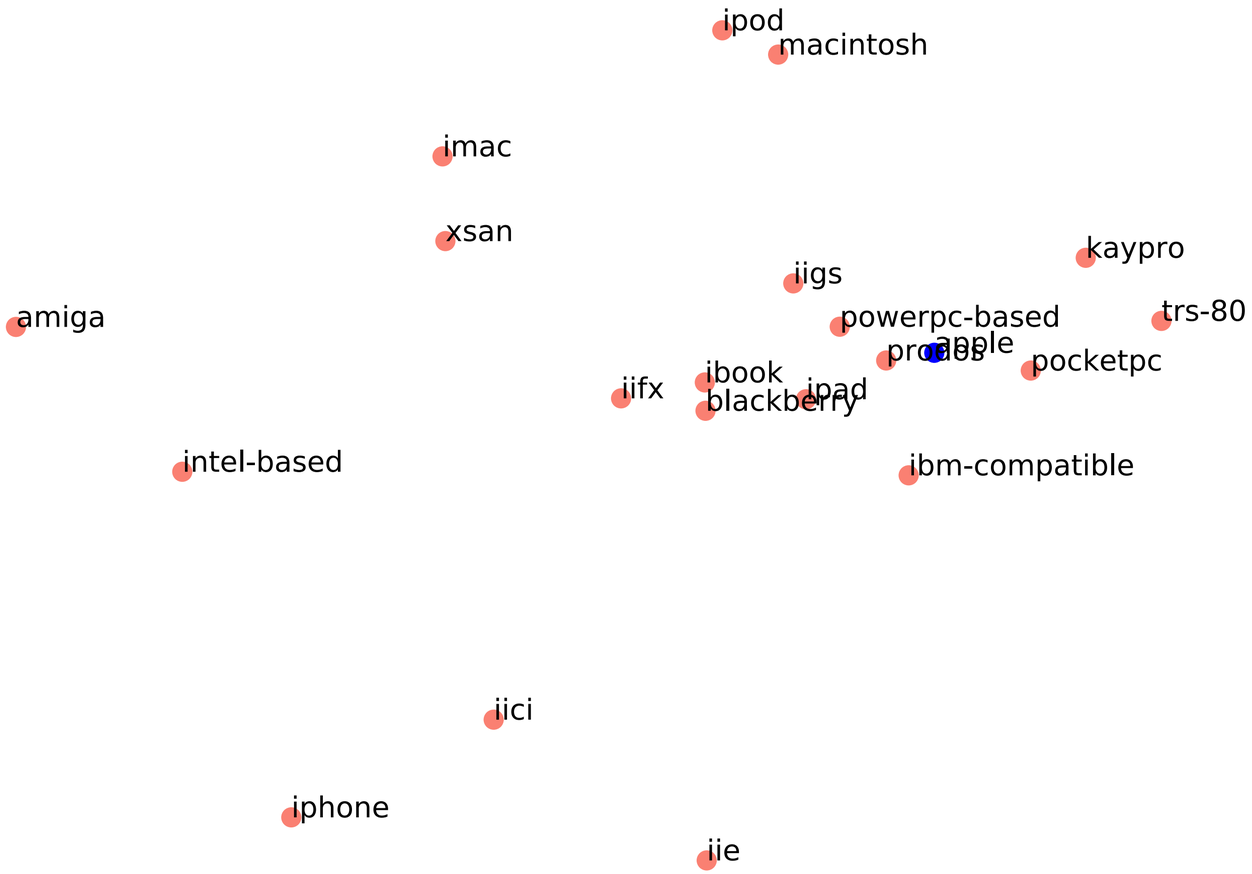
Summary

- Learns high-quality semantically rich embeddings
- Sparse gradients (not explaining why in this lecture)
- Very efficient parallel training

Summary

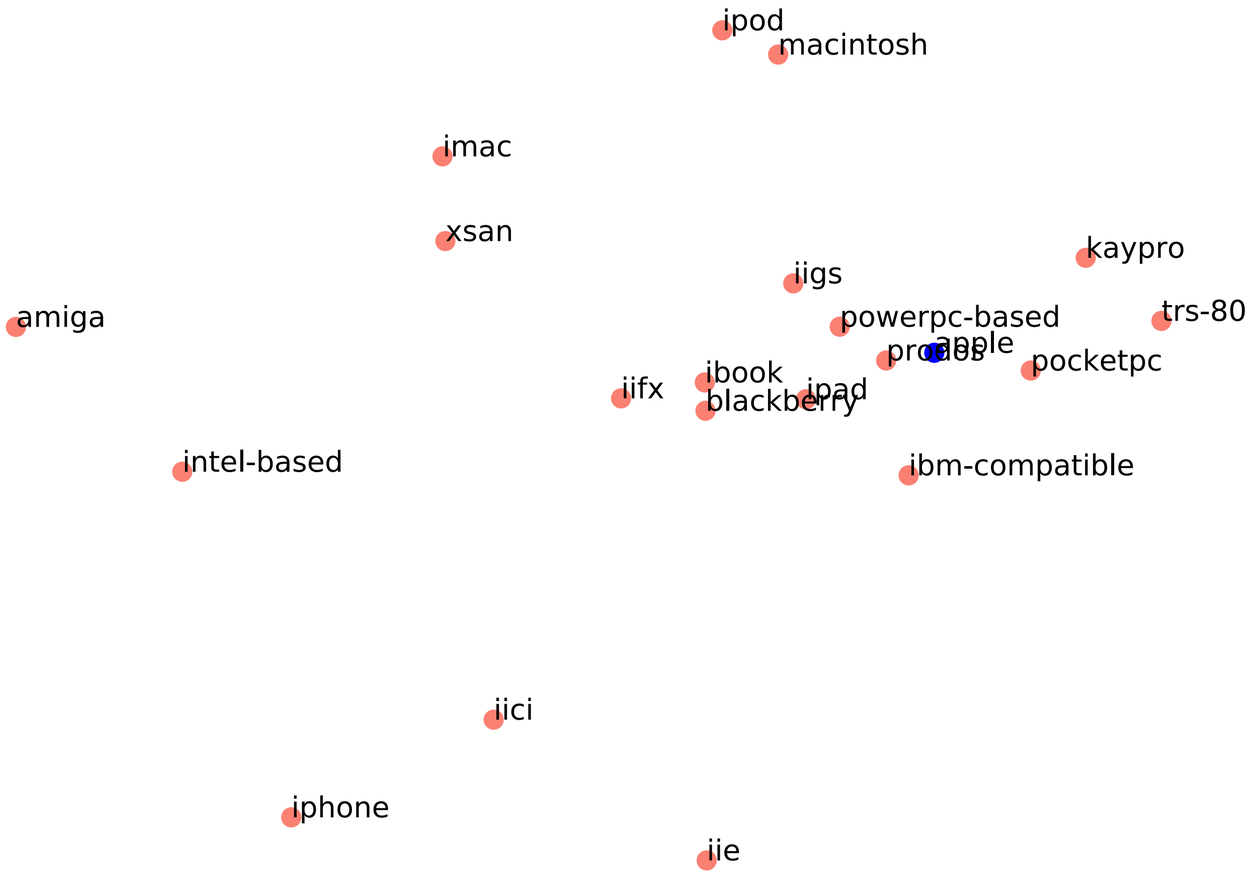
- Learns high-quality semantically rich embeddings
- Sparse gradients (not explaining why in this lecture)
- Very efficient parallel training
- Are we done now?

Skip-gram: embeddings



T-SNE visualisation of 20 nearest neighbours of word “apple”.
300D Skip-gram trained on wikipedia

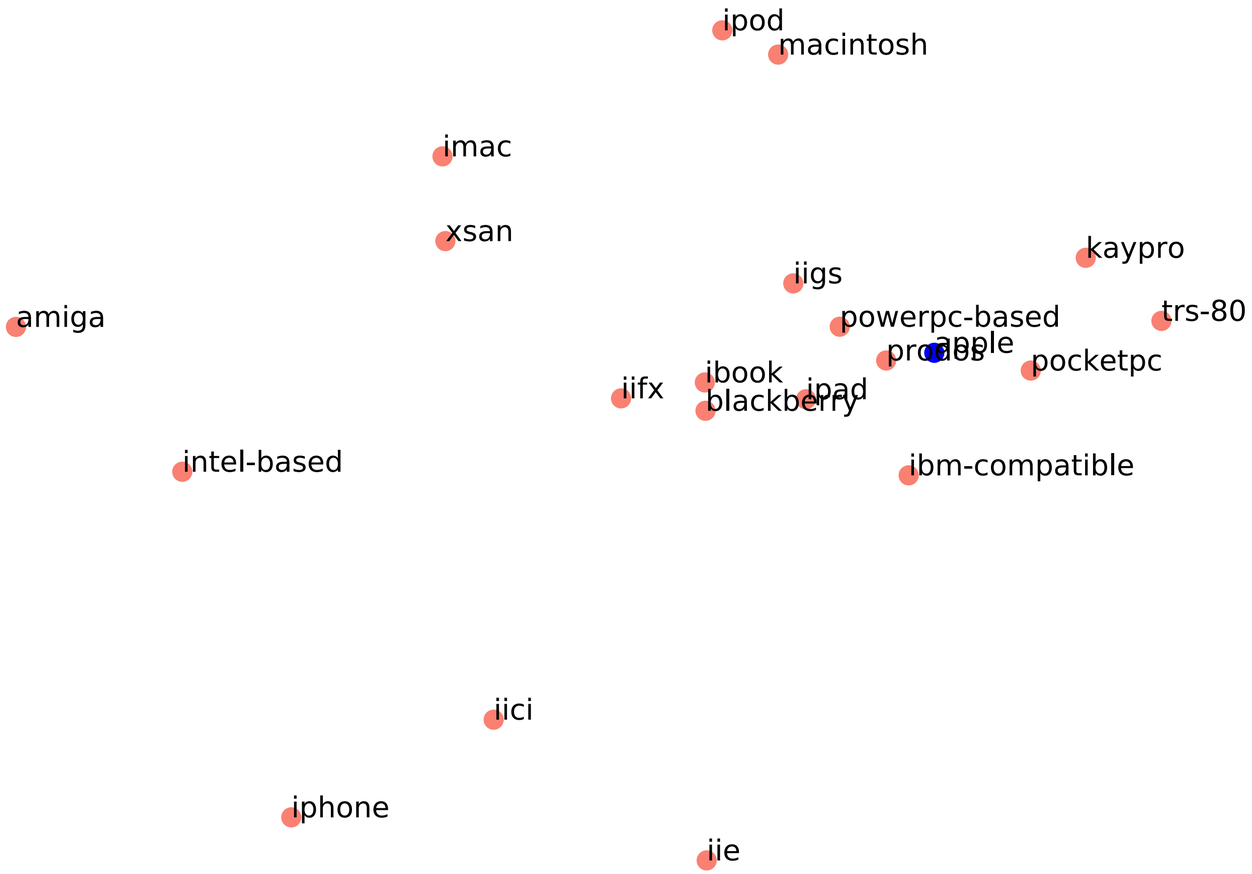
Skip-gram: embeddings



T-SNE visualisation of 20 nearest neighbours of word “apple”.
300D Skip-gram trained on wikipedia

- For some words only one *meaning* is captured

Skip-gram: embeddings



T-SNE visualisation of 20 nearest neighbours of word “apple”.
300D Skip-gram trained on wikipedia

- For some words only one *meaning* is captured
- For others meanings get uncontrollably mixed up

Word ambiguity

Mammal (disambiguation)

From Wikipedia, the free encyclopedia

A **mammal** is a member of a class of vertebrates.

Mammal, **mammals** or **mammalia** may also refer to:

- Mammal (band), an Australian band
- *Mammal* (film), a 2016 Irish film
- The Mammals, an American folk rock band
- *Mammals* (play), a play by Amelia Bullmore
- "Mammal", a song by They Might Be Giants on the album *Apollo 18*
- *Mammal* (EP), the debut EP by Mammal
- *Mammal* (album), an album by Irish black metal band Altar of Plagues
- *Mammalia*, the debut album by drum & bass band Comparative Anatomy

Look up ***bank*** in Wiktionary, the free dictionary.

A **bank** is a financial institution and a financial intermediary that accepts deposits and channels those deposits into lending activities.

Bank or **banking** may also refer to:

Geography

- Bánk, a village and municipality in the comitat of Nógrád, Hungary
- Bank, Iran, a city in Bushehr Province, Iran
- Bank junction, a major road junction in the City of London
- Bank Station (OC Transpo), a bus stop in Ottawa
- Bank Street (Ottawa), Ontario, Canada
- Bank–Monument station, a tube station in London
- Bankə, Azerbaijan
- East Bank (disambiguation)
- Left Bank (disambiguation)
- North Bank (disambiguation)
- Promysel Narimanova or Bank, Azerbaijan
- Right Bank (disambiguation)
- South Bank (disambiguation)
- West Bank (disambiguation)

People

- Bank (surname)

Arts, entertainment, and media

Films

- *DCI Banks*, a British television crime drama, aired by ITV 2010-2016
- *Overdrawn at the Memory Bank*, a 1983 telemovie
- *The Bank* (1915 film), starring Charlie Chaplin
- *The Bank* (2001 film), starring David Wenham

Music

- "Bank", a 2017 song by Brockhampton from *Saturation*
- "Bank", a 2018 song by Lil Baby from *Harder Than Ever*

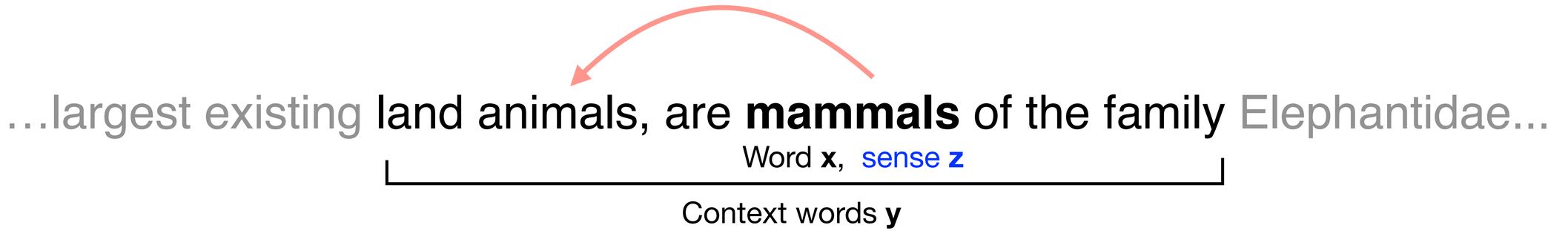
Other uses in arts, entertainment, and media

- *Memory Bank* (UK game show), a daytime game show which was shown on Five in the UK
- *Memory Banks* (comic strip), a British comic strip

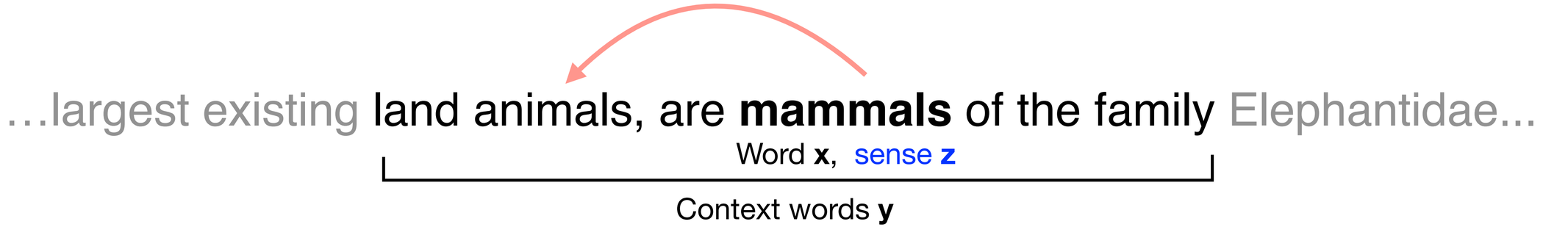
Computing and technology

- Data bank, a storage area for information in telecommunications
- Memory bank, a logical unit of storage

Solution: latent-variable model

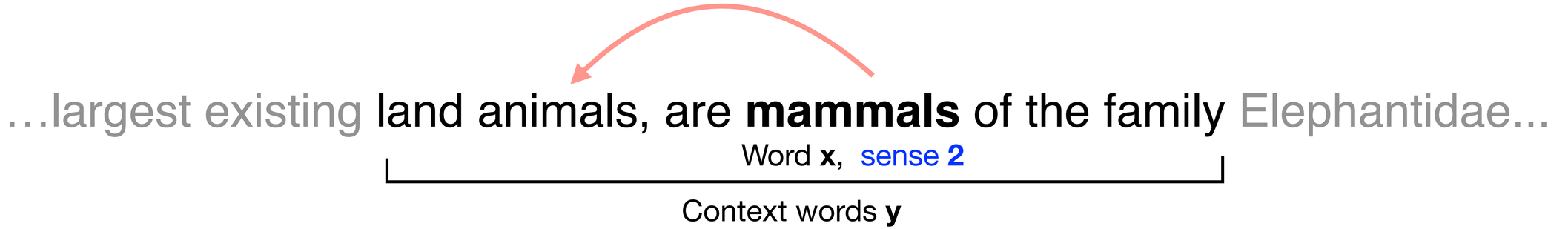


Solution: latent-variable model



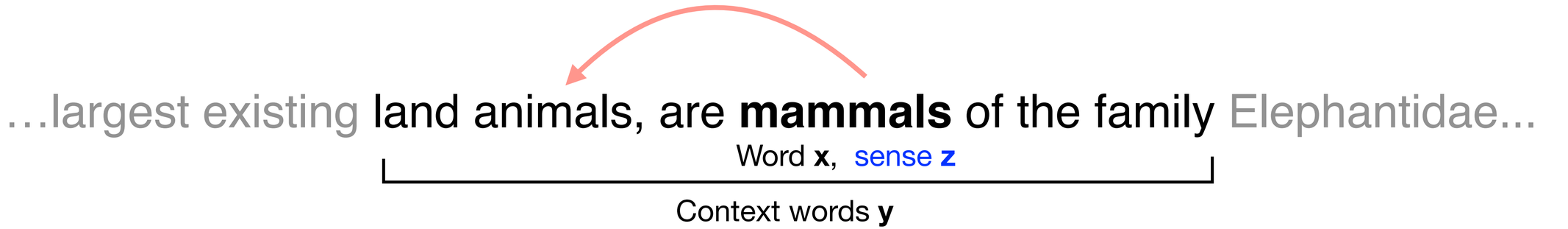
Word	Sense1 embedding	Sense2 embedding	Sense3 embedding	Output embedding
mammals	[-0.73, -0.62, 0.91]	[1.0, 0.43, 0.04]	[-0.39, 0.59, -1.0]	[0.5, 3.45, -1.26]
animals	[0.76, 0.8, -1.1]	[0.3, 1.2, -1.7]	[0.52, -0.24, -0.073]	[-0.93, 1.2, 0.]
land	[0.1, -0.23, 0.58]	[2.0, 0.21, 2.0]	[0.49, -0.022, -0.71]	[-1.52, -0.39, 0.89]
	…	…	…	

Solution: latent-variable model



Word	Sense 1 embedding	Sense 2 embedding	…	Sense K embedding	Output embedding
mammals	[-0.73, -0.62, 0.91]	[1.0, 0.43, 0.04]		[-0.39, 0.59, -1.0]	[0.5, 3.45, -1.26]
animals	[0.76, 0.8, -1.1]	[0.3, 1.2, -1.7]		[0.52, -0.24, -0.073]	[-0.93, 1.2, 0.]
land	[0.1, -0.23, 0.58]	[2.0, 0.21, 2.0]		[0.49, -0.022, -0.71]	[-1.52, -0.39, 0.89]
	…	…		…	

Solution: latent-variable model

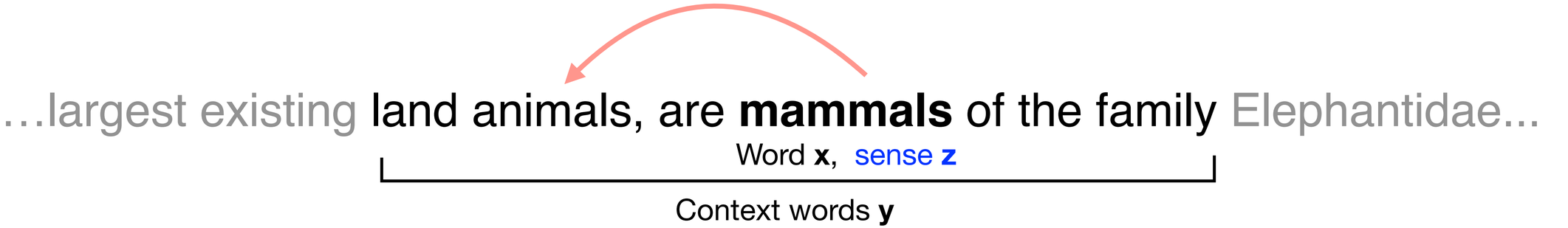


Latent-variable Skip-gram

$$p(v|w,z) = \frac{\exp(\text{in}_{w,z}^T \text{out}_v)}{\sum_{v'} \exp(\text{in}_{w,z}^T \text{out}_{v'})}$$

$$p(\mathbf{y}|x,z) = \prod_j p(y_j|x,z)$$

Solution: latent-variable model



Latent-variable Skip-gram

$$p(v|w,z) = \frac{\exp(\text{in}_{w,z}^T \text{out}_v)}{\sum_{v'} \exp(\text{in}_{w,z}^T \text{out}_{v'})}$$

$$p(\mathbf{y}|x,z) = \prod_j p(y_j|x,z)$$

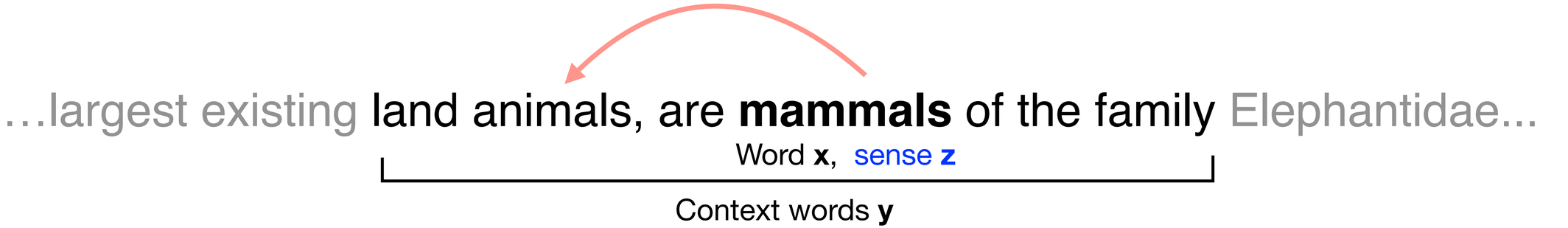
Mixture model of word's contexts

$$\pi_w|\alpha \sim \text{Dirichlet}(\alpha,\ldots,\alpha)$$

$$z_i|\boldsymbol{\pi}, x_i \sim \text{Categorical}(\pi_{x_i})$$

$$y_{ij}|x_i,z_i \sim \text{Softmax}(\text{in}_{x_i,z_i}^T \text{out}_{y_{ij}})$$

Solution: latent-variable model



Latent-variable Skip-gram

$$p(v|w,z) = \frac{\exp(\text{in}_{w,z}^T \text{out}_v)}{\sum_{v'} \exp(\text{in}_{w,z}^T \text{out}_{v'})}$$

$$p(\mathbf{y}|x,z) = \prod_j p(y_j|x,z)$$

Mixture model of word's contexts

$$\pi_w|\alpha \sim \text{Dirichlet}(\alpha,\ldots,\alpha)$$
$$z_i|\boldsymbol{\pi}, x_i \sim \text{Categorical}(\pi_{x_i})$$
$$y_{ij}|x_i,z_i \sim \text{Softmax}(\text{in}_{x_i,z_i}^T \text{out}_{y_{ij}})$$

Training via variational EM

Training via variational EM

- **Observed variables**: $\mathcal{D} = \{(x_i, \mathbf{y}_i)\}_{i=1}^{N}$ (words and contexts)

Training via variational EM

- **Observed variables**: $\mathcal{D} = \{(x_i, \mathbf{y}_i)\}_{i=1}^{N}$ (words and contexts)
- **Hidden variables**: $Z = \{z_i\}_{i=1}^{N}$ (meanings in each context, local),
  $\Pi = \{\pi_w\}_{w=1}^{V}$ (meaning probabilities, global)

Training via variational EM

- **Observed variables**: $\mathcal{D} = \{(x_i, \mathbf{y}_i)\}_{i=1}^{N}$ (words and contexts)
- **Hidden variables**: $Z = \{z_i\}_{i=1}^{N}$ (meanings in each context, local),

  $\Pi = \{\pi_w\}_{w=1}^{V}$ (meaning probabilities, global)
- **Parameters**: $\theta = \{\mathrm{in}_{w,k}\}_{w=1,k=1}^{V,K} \cup \{\mathrm{out}_w\}_{w=1}^{V}$ (word embeddings)

Training via variational EM

- **Observed variables**: $\mathcal{D} = \{(x_i, \mathbf{y}_i)\}_{i=1}^N$ (words and contexts)
- **Hidden variables**: $Z = \{z_i\}_{i=1}^N$ (meanings in each context, local),
  $\Pi = \{\pi_w\}_{w=1}^V$ (meaning probabilities, global)
- **Parameters**: $\theta = \{\text{in}_{w,k}\}_{w=1,k=1}^{V,K} \cup \{\text{out}_w\}_{w=1}^V$ (word embeddings)

$$\log p(\mathcal{D}|\theta) = \log \int p(\Pi|\alpha) \int p(Z|\Pi) \prod_{i=1}^N p(\mathbf{y}_i|x_i, z_i, \theta) dZ d\Pi \to \max_\theta$$

Intractable to compute or optimize directly

Training via variational EM

Training via variational EM

- Consider a fully factored **posterior approximation**

$$q(Z, \Pi) = \prod_{w=1}^{V} q(\pi_w) \prod_{i=1}^{N} q(z_i) \approx p(Z, \Pi | \mathcal{D}, \theta)$$

Training via variational EM

- Consider a fully factored **posterior approximation**

$$q(Z, \Pi) = \prod_{w=1}^{V} q(\pi_w) \prod_{i=1}^{N} q(z_i) \approx p(Z, \Pi | \mathcal{D}, \theta)$$

- Introduce a **variational lower bound**:

$$\begin{aligned} \mathcal{L}(q, \theta) &= \mathbb{E}_{q(Z,\Pi)} \left[\log p(\Pi, Z, Y | X, \theta) - \log q(Z, \Pi) \right] \\ &\leq \log p(\mathcal{D} | \theta) \end{aligned}$$

Training via variational EM

- Consider a fully factored **posterior approximation**

$$q(Z, \Pi) = \prod_{w=1}^{V} q(\pi_w) \prod_{i=1}^{N} q(z_i) \approx p(Z, \Pi | \mathcal{D}, \theta)$$

- Introduce a **variational lower bound**:

$$\begin{aligned} \mathcal{L}(q, \theta) &= \mathbb{E}_{q(Z,\Pi)} \left[\log p(\Pi, Z, Y | X, \theta) - \log q(Z, \Pi) \right] \\ &\leq \log p(\mathcal{D} | \theta) \end{aligned}$$

- Employ **stochastic variational inference** for optimization

E-step. Getting rid of local parameters

E-step. Getting rid of local parameters

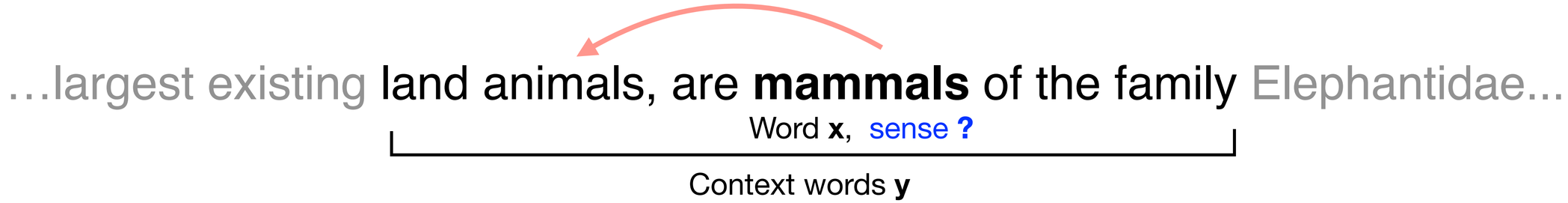
- Assume the current approximation $q(\Pi)$

E-step. Getting rid of local parameters

- Assume the current approximation $q(\Pi)$
- Solve $q^*(Z) = \prod_{i=1}^{N} q^*(z_i) = \arg\max_{q(Z)} \mathcal{L}(q(\Pi), q(Z))$

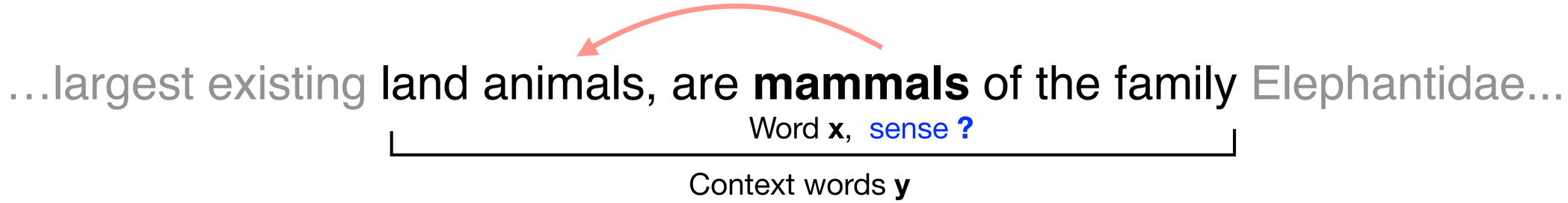
E-step. Getting rid of local parameters

- Assume the current approximation $q(\Pi)$
- Solve $q^*(Z) = \prod_{i=1}^{N} q^*(z_i) = \arg\max_{q(Z)} \mathcal{L}(q(\Pi), q(Z))$



E-step. Getting rid of local parameters

- Assume the current approximation $q(\Pi)$
- Solve $q^*(Z) = \prod_{i=1}^{N} q^*(z_i) = \arg\max_{q(Z)} \mathcal{L}(q(\Pi), q(Z))$

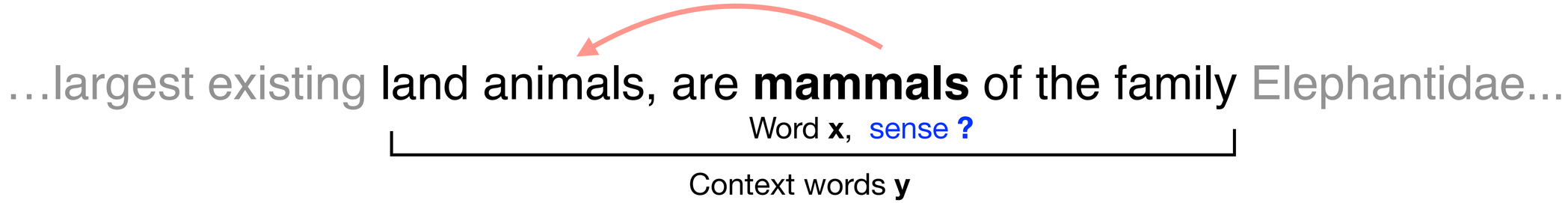


- Perform **sense disambiguation**

$$q(z_i = k)^* \propto \exp(\mathbb{E}_{q(\pi_{x_i})} \log p(z_i = k|\pi_{x_i}) + \log p(\mathbf{y}_i|x_i, z = k, \theta))$$

E-step. Getting rid of local parameters

- Assume the current approximation $q(\Pi)$
- Solve $q^*(Z) = \prod_{i=1}^{N} q^*(z_i) = \arg\max_{q(Z)} \mathcal{L}(q(\Pi), q(Z))$



- Perform **sense disambiguation**

$$q(z_i = k)^* \propto \exp(\mathbb{E}_{q(\pi_{x_i})} \log p(z_i = k|\pi_{x_i}) + \log p(\mathbf{y}_i|x_i, z = k, \theta))$$

- Consider new, **tighter** bound $\mathcal{L}^*(q(\Pi), \theta) = \mathcal{L}(q(\Pi), q^*(Z), \theta) \geq \mathcal{L}(q(\Pi), q(Z), \theta)$

M-step. Stochastic update of global parameters

M-step. Stochastic update of global parameters

- Due to conjugacy $q(\pi_w) = \mathrm{Dirichlet}(\gamma_{w,1}, \ldots, \gamma_{w,K})$

M-step. Stochastic update of global parameters

- Due to conjugacy $q(\pi_w) = \mathrm{Dirichlet}(\gamma_{w,1}, \ldots, \gamma_{w,K})$
- The VLB now has a finite-dimensional numeric parametrization

$$\mathcal{L}^*(\boldsymbol{\gamma}, \theta) = \mathbb{E}_q \left[\log p(\Pi|\boldsymbol{\alpha}) - \log q(\Pi|\boldsymbol{\gamma}) + \sum_{i=1}^{N} \sum_{k=1}^{K} q^*(z_i = k) \log p(\mathbf{y}_i|x_i, k, \theta) \right] + \underbrace{\mathcal{H}(q^*(Z))}_{\text{const}}$$

M-step. Stochastic update of global parameters

- Due to conjugacy $q(\pi_w) = \mathrm{Dirichlet}(\gamma_{w,1}, \ldots, \gamma_{w,K})$
- The VLB now has a finite-dimensional numeric parametrization

$$\mathcal{L}^*(\boldsymbol{\gamma}, \theta) = \mathbb{E}_q \left[\log p(\Pi|\boldsymbol{\alpha}) - \log q(\Pi|\boldsymbol{\gamma}) + \sum_{i=1}^{N} \sum_{k=1}^{K} q^*(z_i = k) \log p(\mathbf{y}_i | x_i, k, \theta) \right] + \underbrace{\mathcal{H}(q^*(Z))}_{\text{const}}$$

Weighted skip-gram prediction

M-step. Stochastic update of global parameters

- Due to conjugacy $q(\pi_w) = \mathrm{Dirichlet}(\gamma_{w,1}, \ldots, \gamma_{w,K})$
- The VLB now has a finite-dimensional numeric parametrization

$$\mathcal{L}^*(\boldsymbol{\gamma}, \theta) = \mathbb{E}_q \left[\log p(\Pi|\boldsymbol{\alpha}) - \log q(\Pi|\boldsymbol{\gamma}) + \sum_{i=1}^{N} \sum_{k=1}^{K} q^*(z_i = k) \log p(\mathbf{y}_i | x_i, k, \theta) \right] + \underbrace{\mathcal{H}(q^*(Z))}_{\text{const}}$$

Weighted skip-gram prediction

M-step. Stochastic update of global parameters

- Due to conjugacy $q(\pi_w) = \mathrm{Dirichlet}(\gamma_{w,1}, \ldots, \gamma_{w,K})$
- The VLB now has a finite-dimensional numeric parametrization

$$\mathcal{L}^*(\boldsymbol{\gamma}, \theta) = \mathbb{E}_q \left[\log p(\Pi|\boldsymbol{\alpha}) - \log q(\Pi|\boldsymbol{\gamma}) + \sum_{i=1}^{N} \sum_{k=1}^{K} q^*(z_i = k) \log p(\mathbf{y}_i|x_i, k, \theta) \right] + \underbrace{\mathcal{H}(q^*(Z))}_{\text{const}}$$

-KL between Dirichlet

Weighted skip-gram prediction

M-step. Stochastic update of global parameters

- Due to conjugacy $q(\pi_w) = \mathrm{Dirichlet}(\gamma_{w,1}, \ldots, \gamma_{w,K})$
- The VLB now has a finite-dimensional numeric parametrization

$$\mathcal{L}^*(\boldsymbol{\gamma}, \theta) = \mathbb{E}_q \left[\log p(\Pi|\boldsymbol{\alpha}) - \log q(\Pi|\boldsymbol{\gamma}) + \sum_{i=1}^{N} \sum_{k=1}^{K} q^*(z_i = k) \log p(\mathbf{y}_i|x_i, k, \theta) \right] + \underbrace{\mathcal{H}(q^*(Z))}_{\text{const}}$$

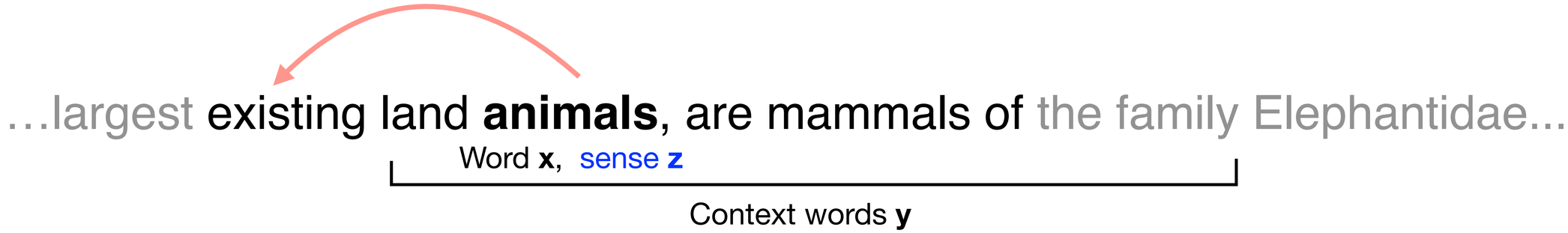
-KL between Dirichlet

Weighted skip-gram prediction

- We can use almost the same Skip-gram training algorithm!

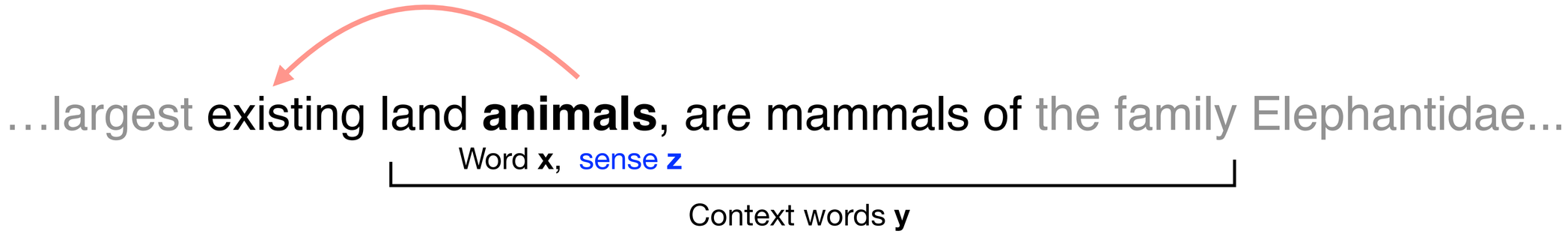
SVI for Multi-meaning Skip-gram

SVI for Multi-meaning Skip-gram



1. Sample a context $(x, \mathbf{y}) \sim \mathcal{D}$

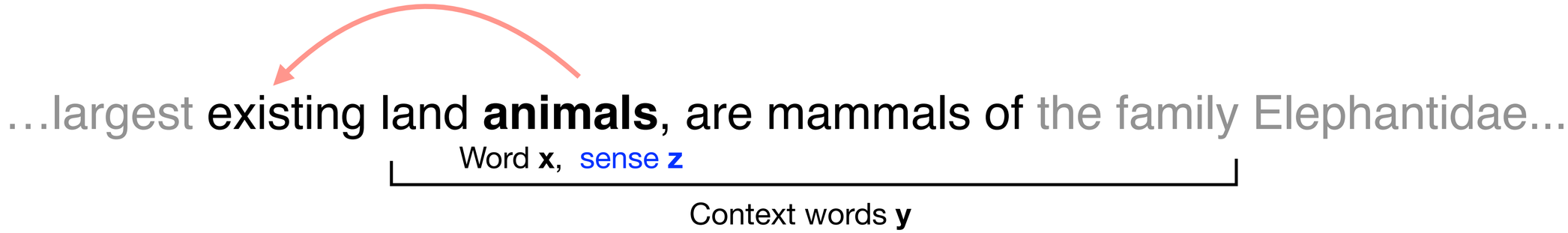
SVI for Multi-meaning Skip-gram



1. Sample a context $(x, \mathbf{y}) \sim \mathcal{D}$

2. Disambiguate the word meaning $q(z) \propto \exp(\mathbb{E}_{q(\pi_x)} \log p(z|\pi) + \log p(\mathbf{y}|x, z, \theta))$

SVI for Multi-meaning Skip-gram

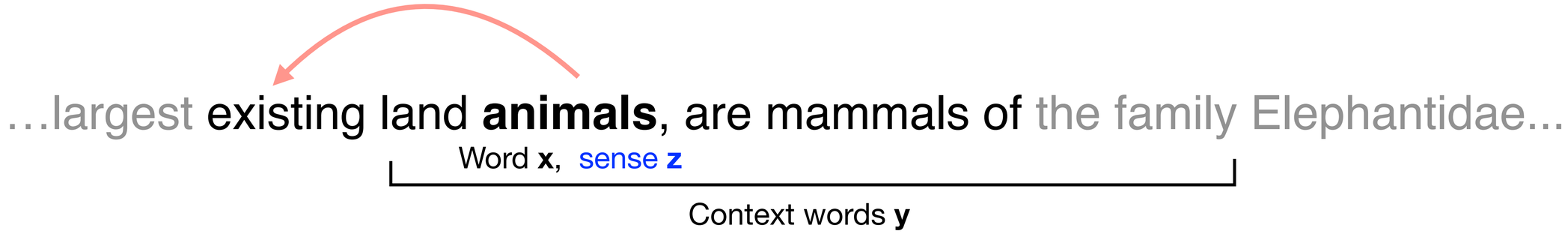


1. Sample a context $(x, \mathbf{y}) \sim \mathcal{D}$

2. Disambiguate the word meaning $q(z) \propto \exp(\mathbb{E}_{q(\pi_x)} \log p(z|\pi) + \log p(\mathbf{y}|x, z, \theta))$

3. Update word embeddings $\theta \leftarrow \theta + \eta \nabla \sum_{k=1}^{K} q(z = k) \log p(\mathbf{y}|x, z = k, \theta)$

SVI for Multi-meaning Skip-gram



1. Sample a context $(x, \mathbf{y}) \sim \mathcal{D}$
2. Disambiguate the word meaning $q(z) \propto \exp(\mathbb{E}_{q(\pi_x)} \log p(z|\pi) + \log p(\mathbf{y}|x, z, \theta))$
3. Update word embeddings $\theta \leftarrow \theta + \eta \nabla \sum_{k=1}^{K} q(z = k) \log p(\mathbf{y}|x, z = k, \theta)$
4. Update parameters of $q(\pi_x|\gamma_x) \quad \gamma_{x,k} \leftarrow \gamma_{x,k} + \eta \underbrace{(q(z = k)n_x - \gamma_{x,k})}_{\text{natural gradient}}$

How to choose the number of meanings?

How to choose the number of meanings?

- Fix the number K for each word

How to choose the number of meanings?

- Fix the number K for each word
- Use a heuristic, e.g. based on the word frequency

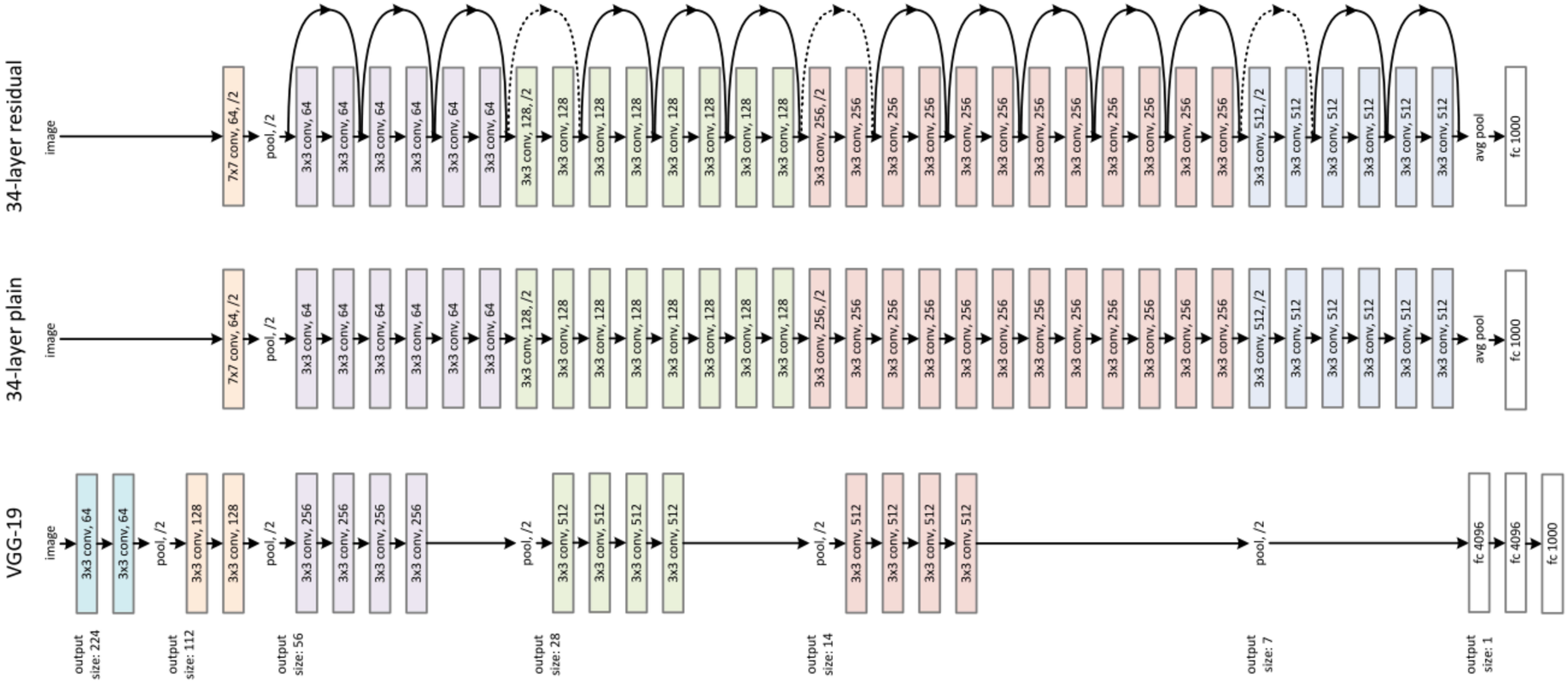
How to choose the number of meanings?

- Fix the number K for each word
- Use a heuristic, e.g. based on the word frequency
- Use an external analysis tool, e.g. part of speech tagger

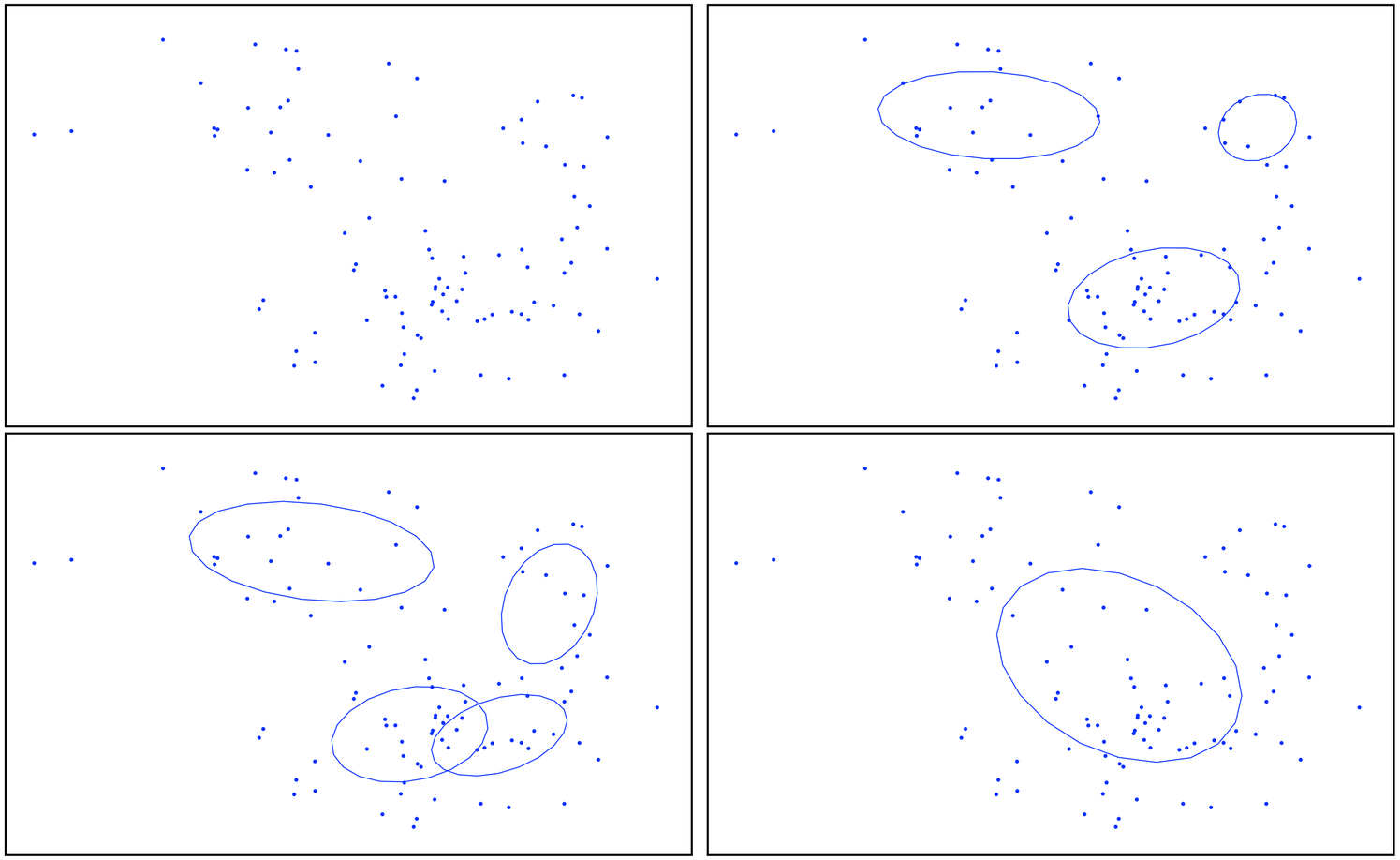
How to choose the number of meanings?

- Fix the number K for each word
- Use a heuristic, e.g. based on the word frequency
- Use an external analysis tool, e.g. part of speech tagger
- Figure out something really smart

How many layers?



How many clusters?



How many clusters?

Traditional model selection

How many clusters?

Traditional model selection

- Define a finite mixture model

$$\pi|\alpha \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$$
$$\phi_k \sim H, k = 1, \ldots, K$$
$$\mu_k, \Sigma_k \sim p_H(\mu, \Sigma)$$
$$z_i|\pi \sim \mathrm{Categortical}(\pi), i = 1, \ldots, N$$
$$x_i|z_i, \boldsymbol{\phi} \sim p(x|\phi_i)$$

How many clusters?

Traditional model selection

- Define a finite mixture model

$$\pi|\alpha \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$$
$$\phi_k \sim H, k = 1, \ldots, K$$
$$\mu_k, \Sigma_k \sim p_H(\mu, \Sigma)$$
$$z_i|\pi \sim \mathrm{Categortical}(\pi), i = 1, \ldots, N$$
$$x_i|z_i, \boldsymbol{\phi} \sim p(x|\phi_i)$$

- For each K measure $\log p(X_{\mathrm{val}}|X_{\mathrm{train}})$

How many clusters?

Traditional model selection

- Define a finite mixture model

$$\pi|\alpha \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$$
$$\phi_k \sim H, k = 1, \ldots, K$$
$$\mu_k, \Sigma_k \sim p_H(\mu, \Sigma)$$
$$z_i|\pi \sim \mathrm{Categortical}(\pi), i = 1, \ldots, N$$
$$x_i|z_i, \boldsymbol{\phi} \sim p(x|\phi_i)$$

- For each K measure $\log p(X_{\mathrm{val}}|X_{\mathrm{train}})$

- Choose K with maximum likelihood

How many clusters?

Traditional model selection

- Define a finite mixture model

$$\pi|\alpha \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$$
$$\phi_k \sim H, k = 1, \ldots, K$$
$$\mu_k, \Sigma_k \sim p_H(\mu, \Sigma)$$
$$z_i|\pi \sim \mathrm{Categortical}(\pi), i = 1, \ldots, N$$
$$x_i|z_i, \boldsymbol{\phi} \sim p(x|\phi_i)$$

- For each K measure $\log p(X_{\mathrm{val}}|X_{\mathrm{train}})$

- Choose K with maximum likelihood

Model selection via Dirichlet process

- Define an infinite model density

$$G = \sum_{k=1}^{\infty} \pi_k \delta(\phi - \phi_k)$$

$$p(x_i|G) = \sum_{k=1}^{\infty} \pi_k p(x_i|\phi_k)$$

How many clusters?

Traditional model selection

- Define a finite mixture model

$$\pi|\alpha \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$$
$$\phi_k \sim H, k = 1, \ldots, K$$
$$\mu_k, \Sigma_k \sim p_H(\mu, \Sigma)$$
$$z_i|\pi \sim \mathrm{Categortical}(\pi), i = 1, \ldots, N$$
$$x_i|z_i, \boldsymbol{\phi} \sim p(x|\phi_i)$$

- For each K measure $\log p(X_{\mathrm{val}}|X_{\mathrm{train}})$

- Choose K with maximum likelihood

Model selection via Dirichlet process

- Define an infinite model density

$$G = \sum_{k=1}^{\infty} \pi_k \delta(\phi - \phi_k)$$

$$p(x_i|G) = \sum_{k=1}^{\infty} \pi_k p(x_i|\phi_k)$$

- Choose a nonparametric prior

$$G|H, \alpha \sim \mathrm{DP}(H, \alpha)$$

How many clusters?

Traditional model selection

- Define a finite mixture model

$$\pi|\alpha \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$$
$$\phi_k \sim H, k = 1, \ldots, K$$
$$\mu_k, \Sigma_k \sim p_H(\mu, \Sigma)$$
$$z_i|\pi \sim \mathrm{Categortical}(\pi), i = 1, \ldots, N$$
$$x_i|z_i, \boldsymbol{\phi} \sim p(x|\phi_i)$$

- For each K measure $\log p(X_{\mathrm{val}}|X_{\mathrm{train}})$
- Choose K with maximum likelihood

Model selection via Dirichlet process

- Define an infinite model density

$$G = \sum_{k=1}^{\infty} \pi_k \delta(\phi - \phi_k)$$

$$p(x_i|G) = \sum_{k=1}^{\infty} \pi_k p(x_i|\phi_k)$$

- Choose a nonparametric prior

$$G|H, \alpha \sim \mathrm{DP}(H, \alpha)$$

- Select the model one-shot

$$p(x|X) = \int \underbrace{p(G|X)}_{\text{model}} p(x|G) dG$$

$$p(G|X) \propto \underbrace{p(G)p(X|G)}_{\text{model selection}}$$

What is a good nonparametric prior?

What is a good nonparametric prior?

- Allows potentially infinite number of clusters

What is a good nonparametric prior?

- Allows potentially infinite number of clusters
- For a finite number of data points **n**, the number of clusters is **K** << **n**

What is a good nonparametric prior?

- Allows potentially infinite number of clusters
- For a finite number of data points **n**, the number of clusters is **K** << **n**
- Model complexity (number of clusters) grows with more data available

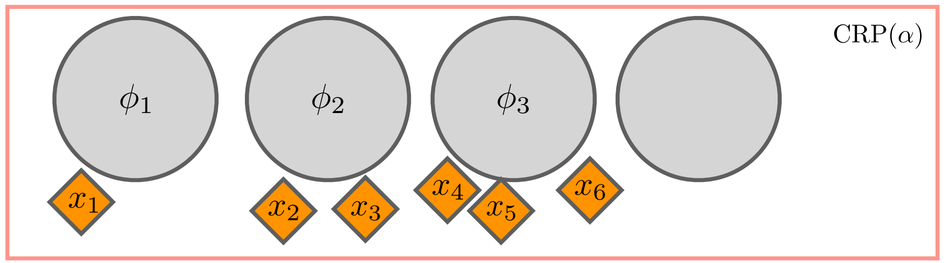
What is a good nonparametric prior?

- Allows potentially infinite number of clusters
- For a finite number of data points **n**, the number of clusters is **K** << **n**
- Model complexity (number of clusters) grows with more data available

$$\mathrm{DP}(H, \alpha)$$

What is a good nonparametric prior?

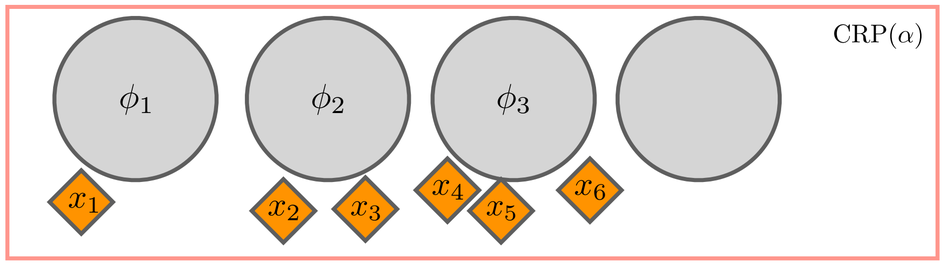
- Allows potentially infinite number of clusters
- For a finite number of data points **n**, the number of clusters is **K** << **n**
- Model complexity (number of clusters) grows with more data available



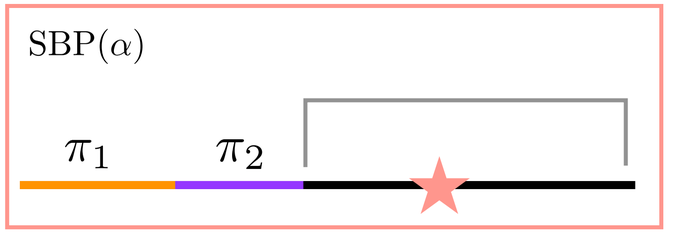
$\Leftrightarrow \mathrm{DP}(H, \alpha)$

What is a good nonparametric prior?

- Allows potentially infinite number of clusters
- For a finite number of data points **n**, the number of clusters is **K** << **n**
- Model complexity (number of clusters) grows with more data available

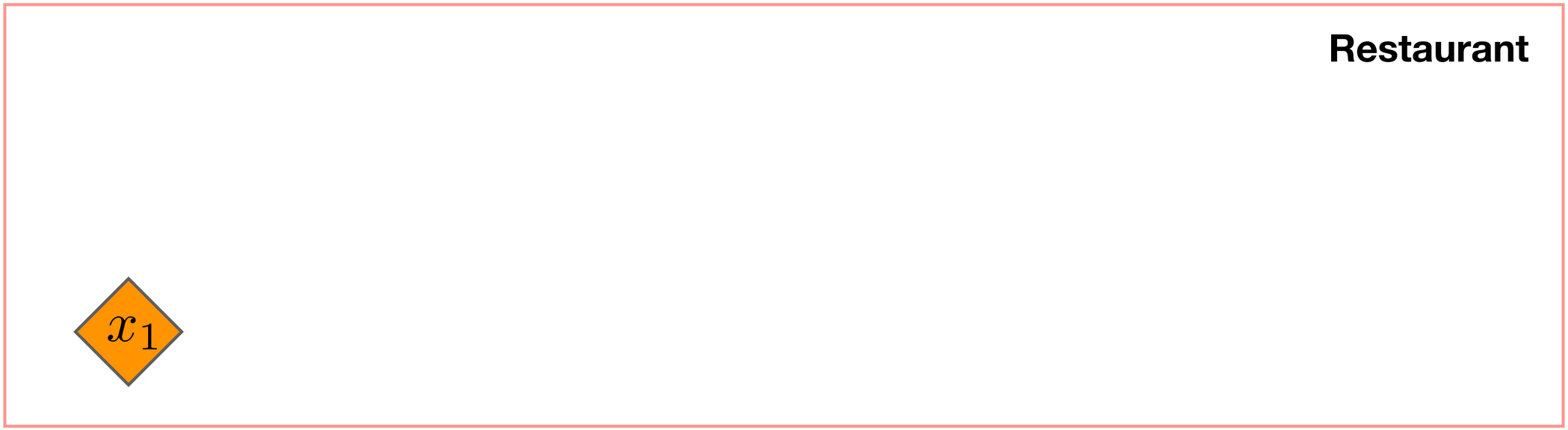


$\Leftrightarrow \mathrm{DP}(H, \alpha) \Leftrightarrow$

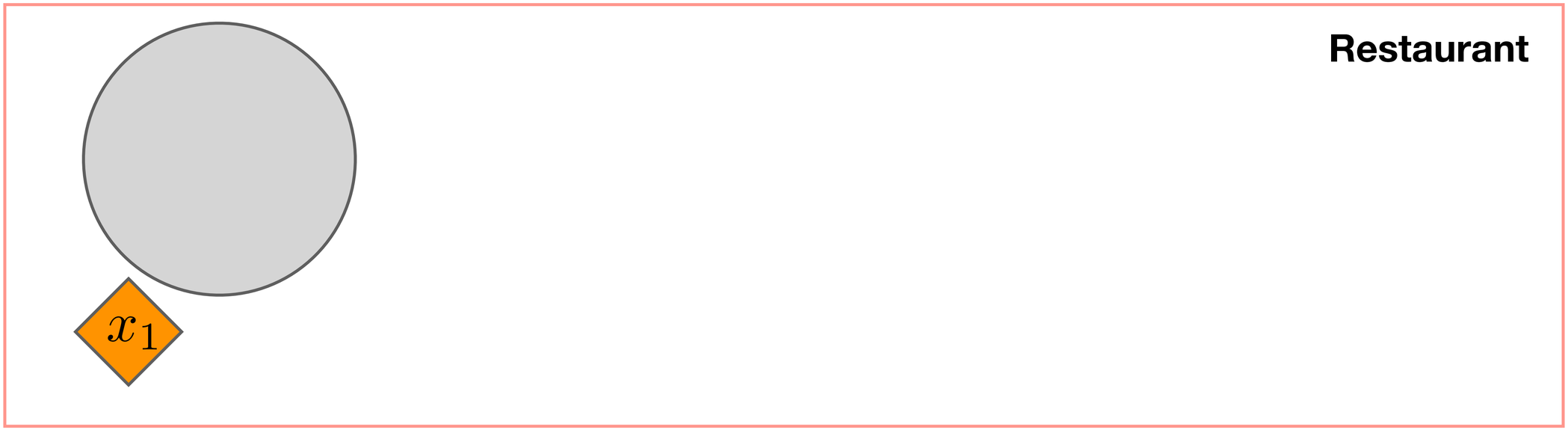


Chinese Restaurant Process

Chinese Restaurant Process

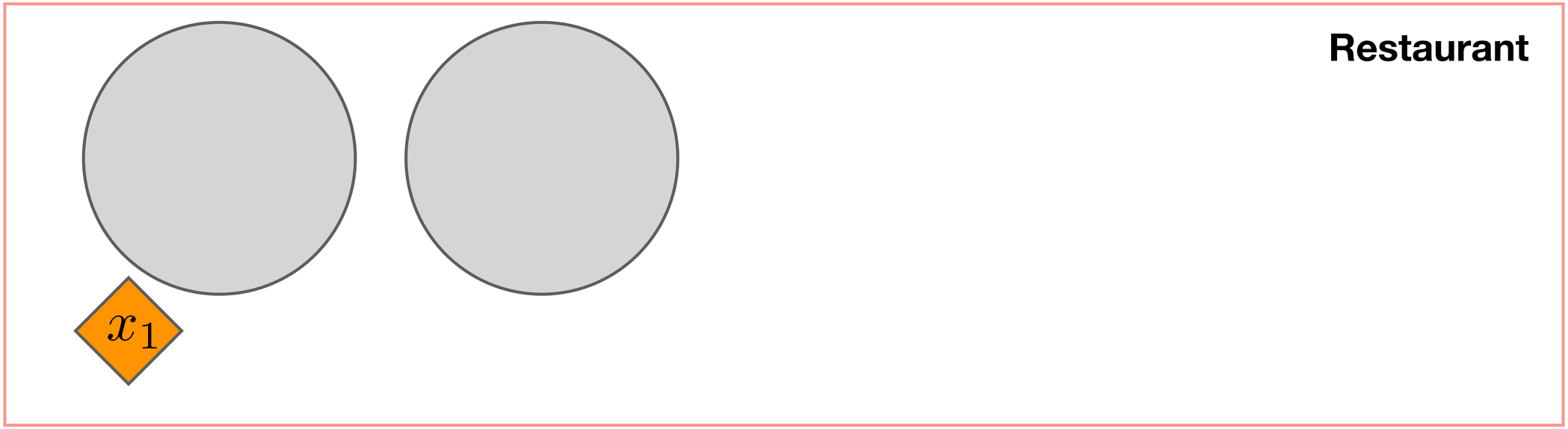


Chinese Restaurant Process



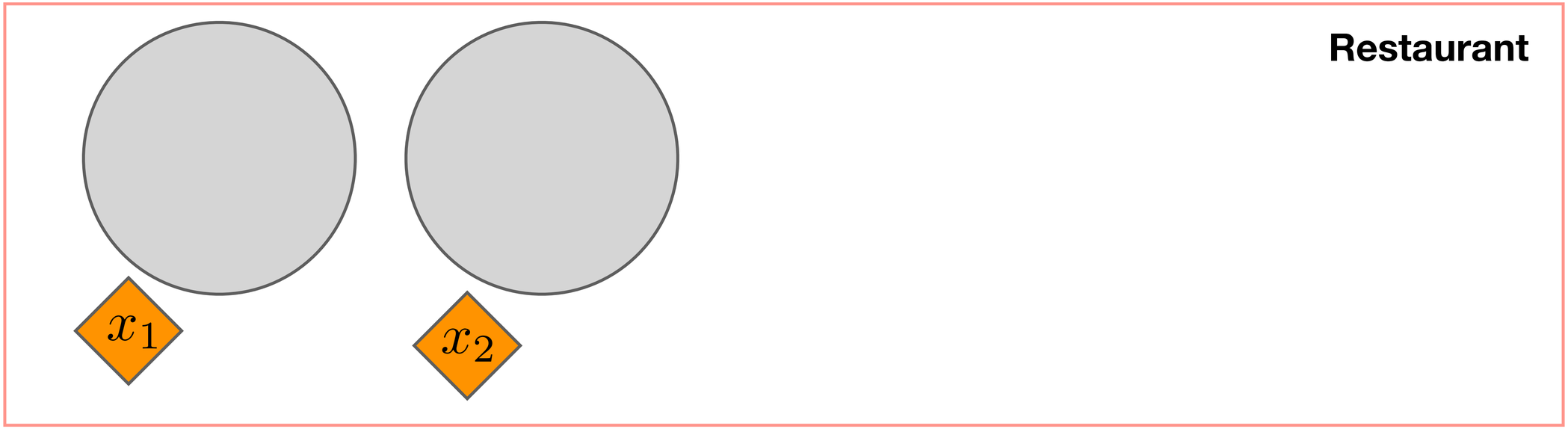
$z_1 = 1$

Chinese Restaurant Process



$z_1 = 1$

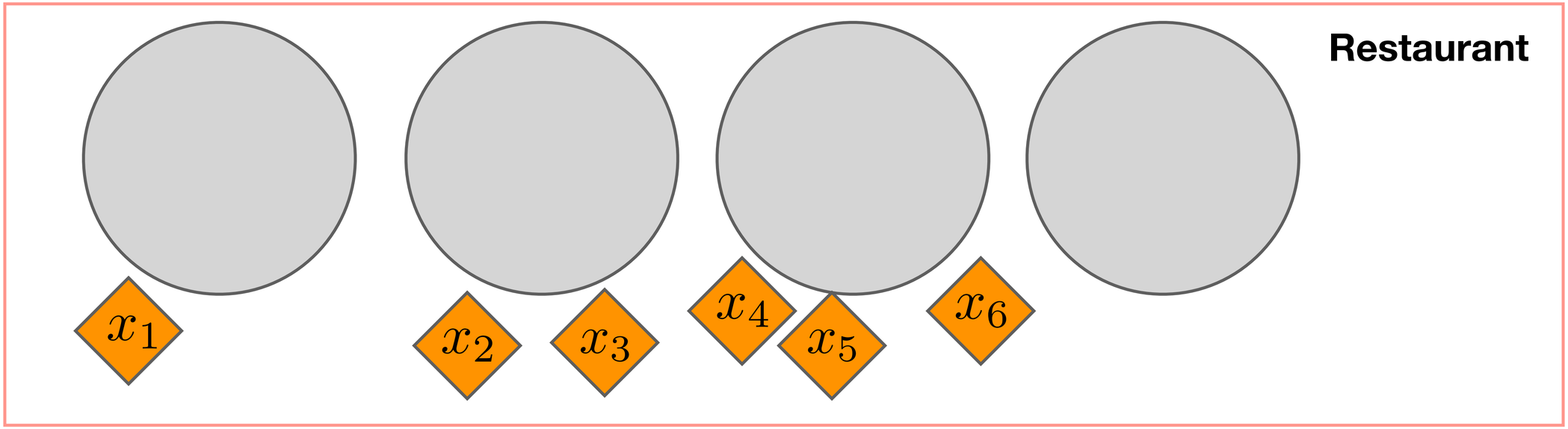
Chinese Restaurant Process



$z_1 = 1$

$z_2 = 2$

Chinese Restaurant Process

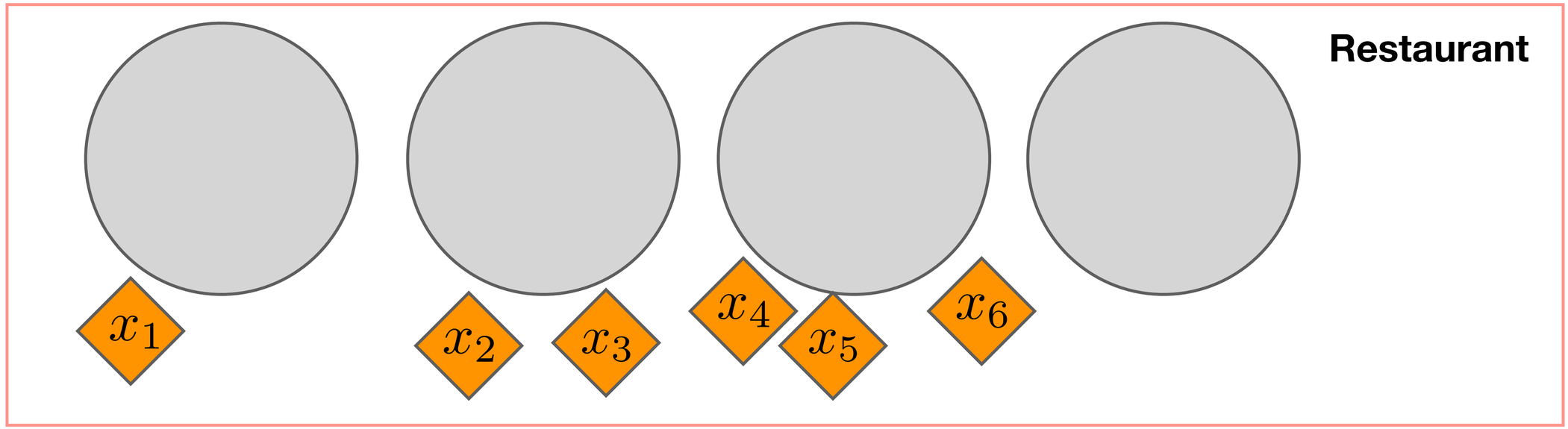


$z_1 = 1$

$z_2 = 2$

...

Chinese Restaurant Process



$z_1 = 1$

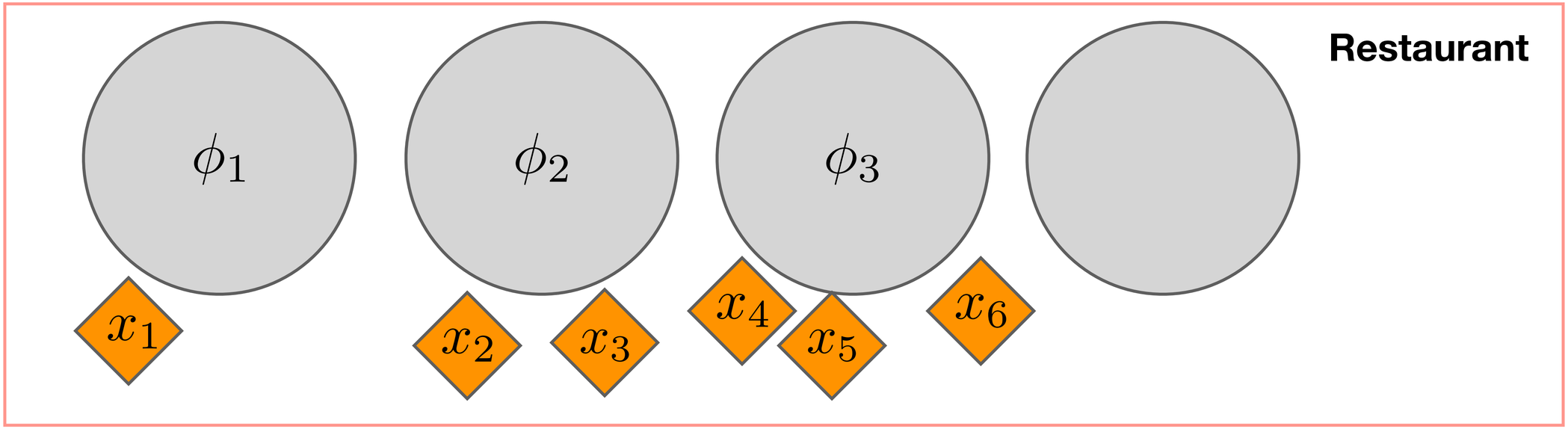
$z_2 = 2$

...

$$p(z_i = k|z_{<i}) \propto \begin{cases} n_k, & k \leq K, \\ \alpha, & k = K+1 \end{cases}$$

$$p(Z) = \prod_{i=1}^{N} p(z_i|z_{<i})$$

Chinese Restaurant Process



$z_1 = 1$

$z_2 = 2$

...

$$p(z_i = k | z_{<i}) \propto \begin{cases} n_k, & k \leq K, \\ \alpha, & k = K+1 \end{cases}$$

$$p(Z) = \prod_{i=1}^{N} p(z_i | z_{<i})$$

$$\phi_k \sim H, \quad (\phi_k \sim p(\phi|H))$$

$$x_i | Z, \Phi \sim p(x_i | \phi_k)$$

How many meanings?

How many meanings?

Finite mixture of word's contexts

$$\pi_w|\alpha \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$$
$$z_i|\boldsymbol{\pi}, x_i \sim \text{Categorical}(\pi_{x_i})$$
$$y_{ij}|x_i, z_i \sim \text{Softmax}(\text{in}_{x_i,z_i}^T \text{out}_{y_{ij}})$$

How many meanings?

Finite mixture of word's contexts

$$\pi_w|\alpha \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$$

$$z_i|\boldsymbol{\pi}, x_i \sim \text{Categorical}(\pi_{x_i})$$

$$y_{ij}|x_i, z_i \sim \text{Softmax}(\text{in}_{x_i,z_i}^T \text{out}_{y_{ij}})$$

How many meanings?

Finite mixture of word's contexts

$$\pi_w|\alpha \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$$
$$z_i|\boldsymbol{\pi}, x_i \sim \text{Categorical}(\pi_{x_i})$$
$$y_{ij}|x_i, z_i \sim \text{Softmax}(\text{in}_{x_i,z_i}^T \text{out}_{y_{ij}})$$

Infinite mixture of word's contexts

$$z_i|z_{<i} \sim \text{CRP}(\alpha)$$
$$y_{ij}|x_i, z_i \sim \text{Softmax}(\text{in}_{x_i,z_i}^T \text{out}_{y_{ij}})$$

How many meanings?

Finite mixture of word's contexts

$$\pi_w|\alpha \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$$
$$z_i|\boldsymbol{\pi}, x_i \sim \text{Categorical}(\pi_{x_i})$$
$$y_{ij}|x_i, z_i \sim \text{Softmax}(\text{in}_{x_i,z_i}^T \text{out}_{y_{ij}})$$

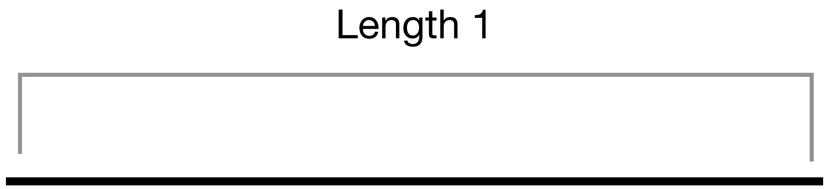
Infinite mixture of word's contexts

$$z_i|z_{<i} \sim \text{CRP}(\alpha)$$
$$y_{ij}|x_i, z_i \sim \text{Softmax}(\text{in}_{x_i,z_i}^T \text{out}_{y_{ij}})$$

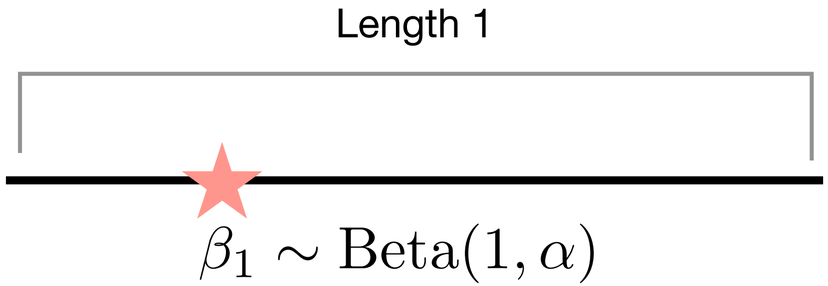
Problem: data is not iid anymore, inconvenient for SVI

Stick-breaking process

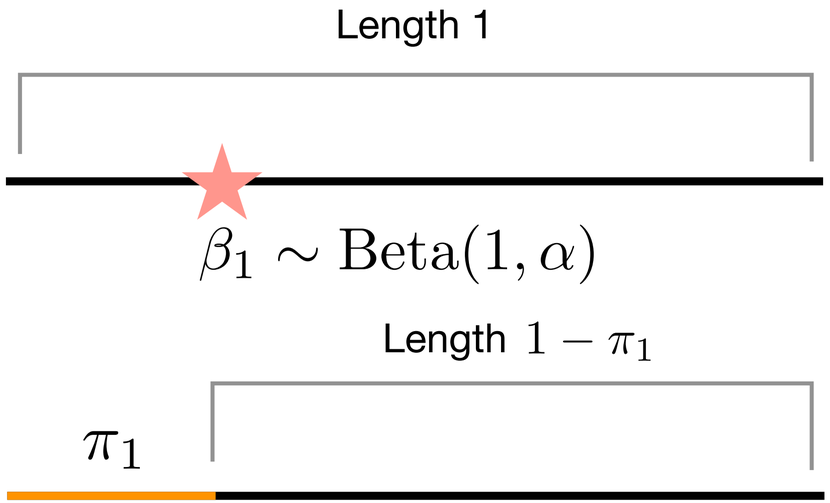
Stick-breaking process



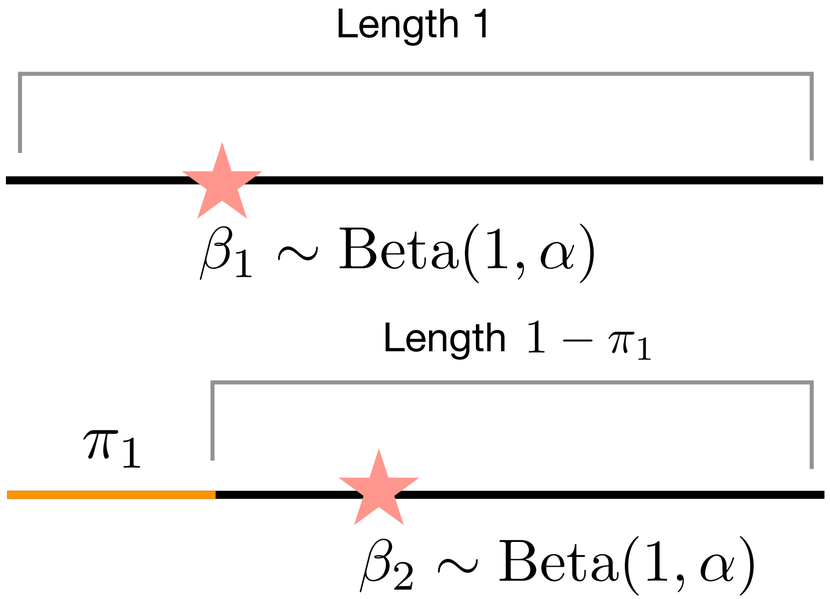
Stick-breaking process



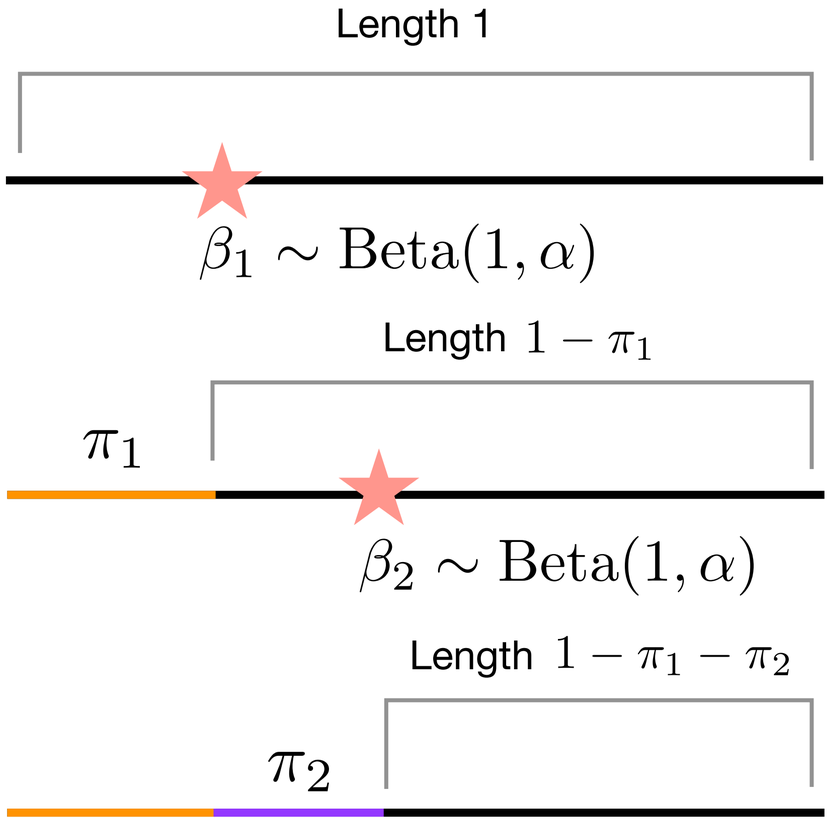
Stick-breaking process



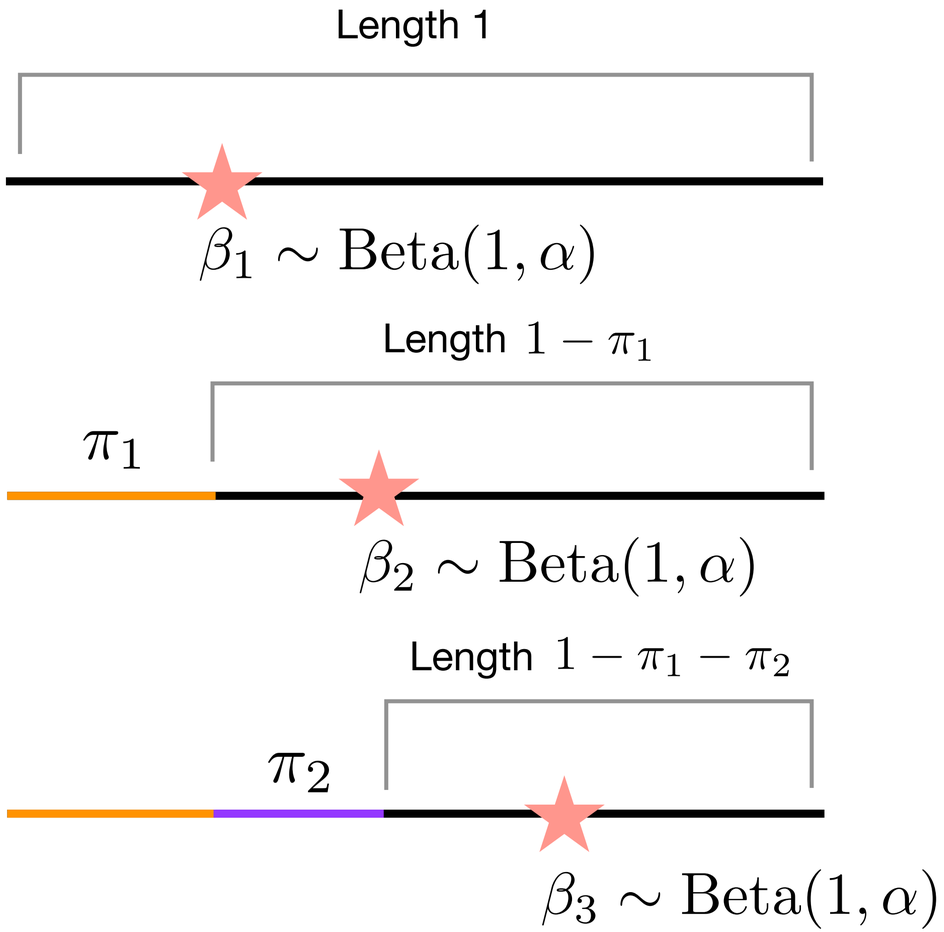
Stick-breaking process



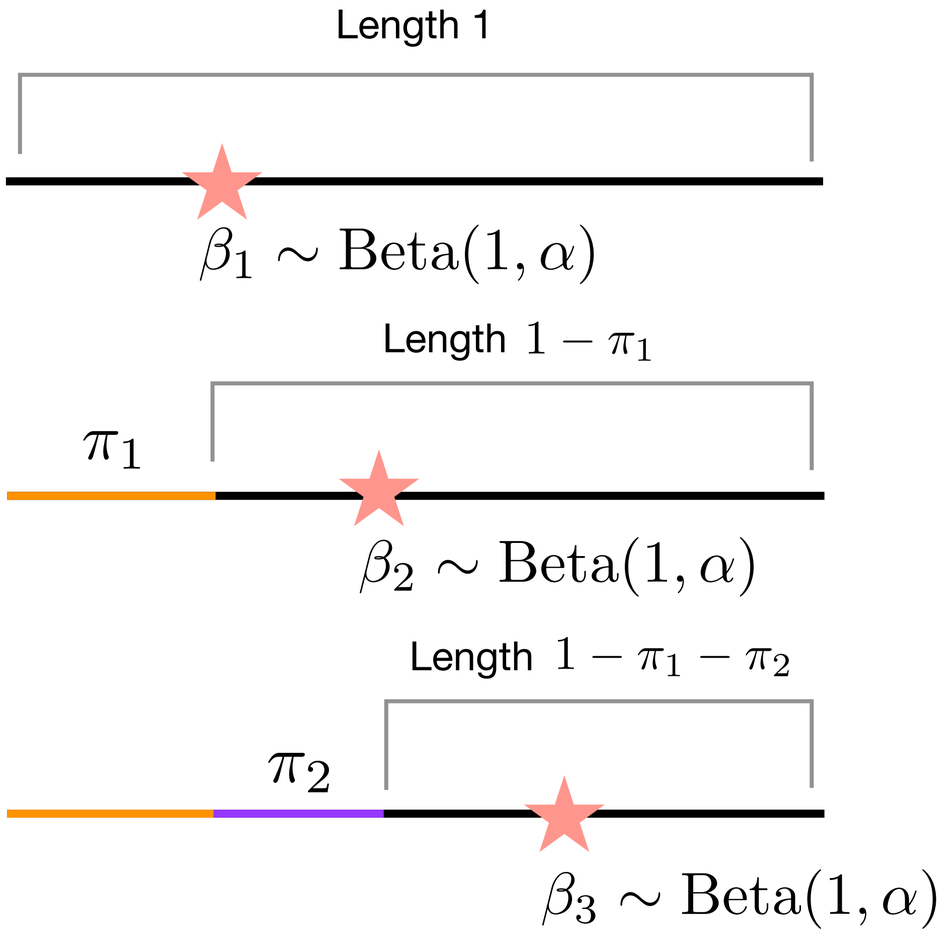
Stick-breaking process



Stick-breaking process

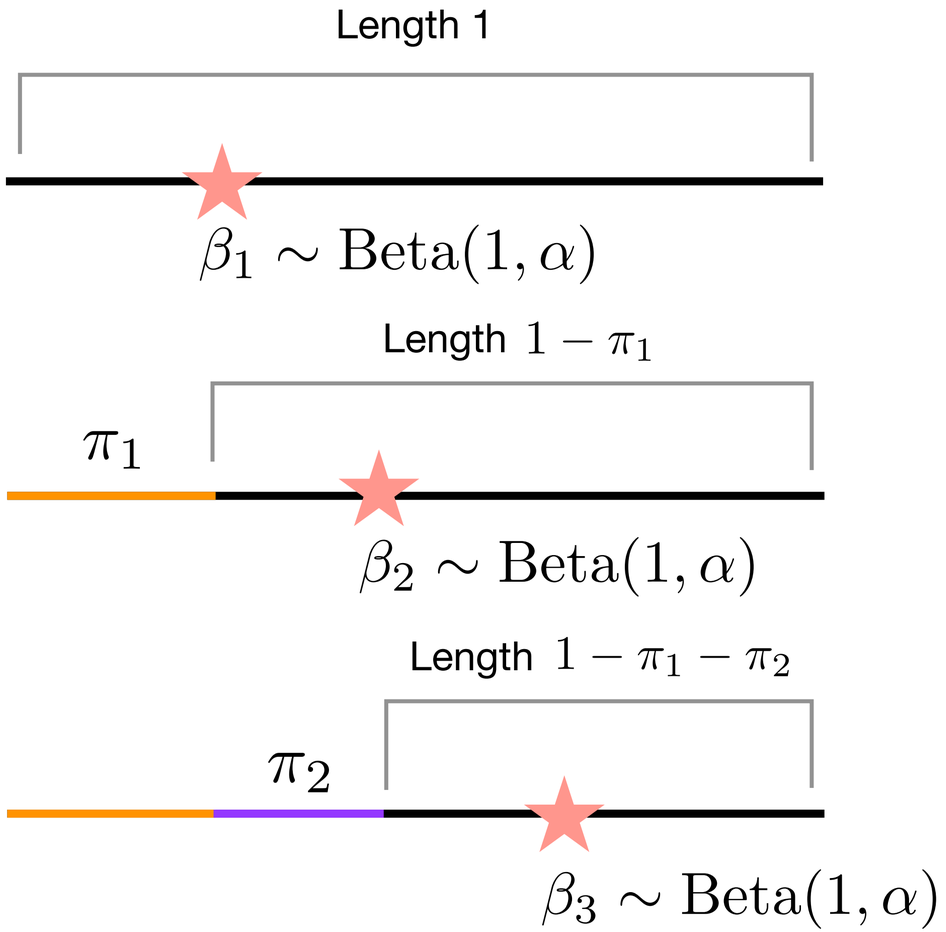


Stick-breaking process



$$
\begin{aligned}
\beta_k &\sim \text{Beta}(1, \alpha) \\
\phi_k &\sim H, \quad k = 1, 2, \ldots \\
\pi_k &= \beta_k \prod_{t<k} (1 - \beta_t) = \beta_k (1 - \sum_{t<k} \pi_t) \\
z_i | \pi &\sim \text{Categorical}(\pi) \\
x_i | z_i &\sim p(x_i | \phi_k)
\end{aligned}
$$

Stick-breaking process



$$\beta_k \sim \text{Beta}(1, \alpha)$$
$$\phi_k \sim H, \quad k = 1, 2, \ldots$$
$$\pi_k = \beta_k \prod_{t<k} (1 - \beta_t) = \beta_k (1 - \sum_{t<k} \pi_t)$$
$$z_i | \pi \sim \text{Categorical}(\pi)$$
$$x_i | z_i \sim p(x_i | \phi_k)$$

$$G = \sum_{k=1}^{\infty} \pi_k \delta(\phi - \phi_k) \sim \text{DP}(H, \alpha) \qquad z \sim \text{CRP}(\alpha)$$

Adaptive Skip-gram

- Prior over infinitely-many meanings
- Automatic model selection (more data - more meanings)
- Control of the meaning granularity

Infinite mixture of word's contexts

$$\beta_k|\alpha \sim \text{Beta}(1, \alpha), \quad k = 1, 2, \ldots$$

$$\pi_k = \beta_k \prod_{t<k} (1 - \beta_t)$$

$$z_i|\boldsymbol{\beta}, x_i \sim \text{Categorical}(\pi_{x_i})$$

$$y_{ij}|x_i, z_i \sim \text{Softmax}(\text{in}_{x_i,z_i}^T \text{out}_{y_{ij}})$$

Variational inference for SBP

Variational inference for SBP

- Finite-dimensional variational approximation

$$q(\boldsymbol{\beta}, \boldsymbol{\phi}, Z) = \prod_{k=1}^{\infty} q(\beta_k) q(\phi_k) \prod_{i=1}^{N} q(z_i) \approx p(\boldsymbol{\beta}, \boldsymbol{\phi}, Z | X, \alpha)$$

Variational inference for SBP

- Finite-dimensional variational approximation

$$q(\boldsymbol{\beta}, \boldsymbol{\phi}, Z) = \prod_{k=1}^{\infty} q(\beta_k) q(\phi_k) \prod_{i=1}^{N} q(z_i) \approx p(\boldsymbol{\beta}, \boldsymbol{\phi}, Z | X, \alpha)$$

$$\pi_k = \beta_k \prod_{t<k} (1 - \beta_t)$$

$$q(\beta_T) = \delta(\beta_T - 1), \quad \underbrace{q(\beta_k) = \text{Beta}(1, \alpha), q(\phi_k) = p_H(\phi_k)}_{\text{prior}}, k > T$$

Variational inference for SBP

- Finite-dimensional variational approximation

$$q(\boldsymbol{\beta}, \boldsymbol{\phi}, Z) = \prod_{k=1}^{\infty} q(\beta_k) q(\phi_k) \prod_{i=1}^{N} q(z_i) \approx p(\boldsymbol{\beta}, \boldsymbol{\phi}, Z | X, \alpha)$$

$$\pi_k = \beta_k \prod_{t<k} (1 - \beta_t)$$

$$q(\beta_T) = \delta(\beta_T - 1), \quad \underbrace{q(\beta_k) = \text{Beta}(1, \alpha), q(\phi_k) = p_H(\phi_k)}_{\text{prior}}, k > T$$

- Due to prior conjugacy $q(\beta_k) = \text{Beta}(\beta_k | a_k, b_k)$

Variational inference for SBP

- Finite-dimensional variational approximation

$$q(\boldsymbol{\beta}, \boldsymbol{\phi}, Z) = \prod_{k=1}^{\infty} q(\beta_k) q(\phi_k) \prod_{i=1}^{N} q(z_i) \approx p(\boldsymbol{\beta}, \boldsymbol{\phi}, Z | X, \alpha)$$

$$\pi_k = \beta_k \prod_{t<k} (1 - \beta_t)$$

$$q(\beta_T) = \delta(\beta_T - 1), \quad \underbrace{q(\beta_k) = \text{Beta}(1, \alpha), q(\phi_k) = p_H(\phi_k)}_{\text{prior}}, k > T$$

- Due to prior conjugacy $q(\beta_k) = \text{Beta}(\beta_k | a_k, b_k)$

$$\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}_q \left[\log p(\mathbf{x} | \boldsymbol{\beta}, \boldsymbol{\pi}, \mathbf{z}) - \text{KL}(q(\boldsymbol{\beta}, \boldsymbol{\pi}, \mathbf{z}) || p(\boldsymbol{\beta}, \boldsymbol{\pi}, \mathbf{z}) \right] \\
&= \mathbb{E}_q [\log p(\mathbf{x} | \boldsymbol{\beta}, \boldsymbol{\pi}, \mathbf{z}) - \sum_{k=1}^{T} \text{KL}(q(\beta_k) q(\phi_k) || p(\beta_k) p(\phi_k)) \\
&\quad - \sum_{k=T+1}^{\infty} \text{KL}(p(\beta_k) p(\phi_k) || p(\beta_k) p(\phi_k)) - \text{KL}(q(\mathbf{z}) || p(\mathbf{z} | \boldsymbol{\beta}))]
\end{aligned}$$

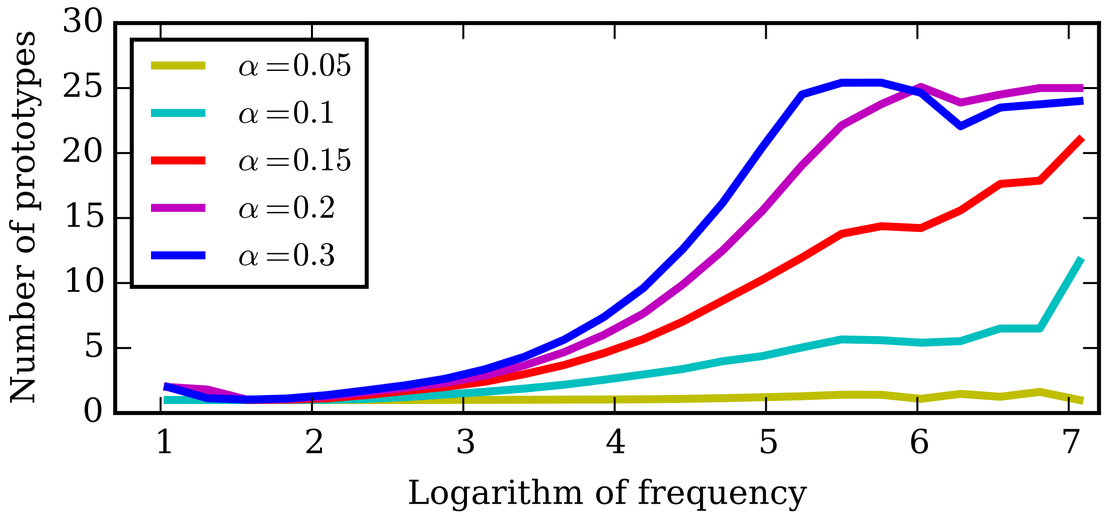
Sense discovery

Sense discovery

ALPHA	$p(z)$	nearest neighbours
		“light”
Skip-Gram	1.00	far-red, emitting
0.075	0.28	armoured, amx-13, kilcrease
	0.72	bright, sunlight, luminous
0.1	0.09	tvärbanan, hudson-bergen
	0.17	dark, bright, green
	0.09	4th, dragoons, 2nd
	0.26	radiation, ultraviolet
	0.28	darkness, shining, shadows
	0.11	self-propelled, armored
		“core”
Skip-Gram	1.00	cores, components, i7
0.075	0.3	competencies, curriculum
	0.34	cpu, cores, i7, powerxcell
	0.36	nucleus backbone
0.1	0.21	reactor, hydrogen-rich
	0.13	intel, processors
	0.27	curricular, competencies
	0.15	downtown, cores, center
	0.24	nucleus, rag-tag, roster

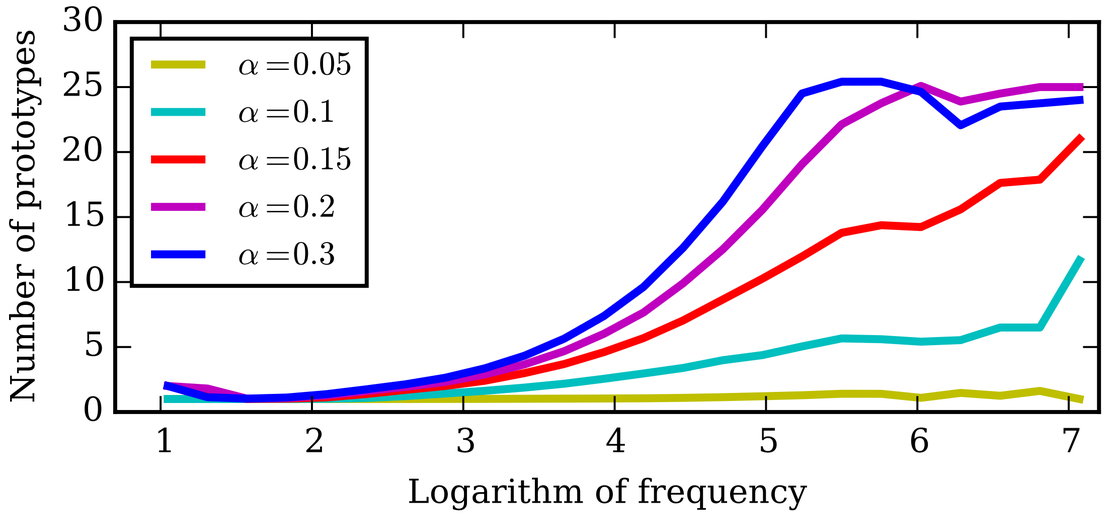
Sense discovery

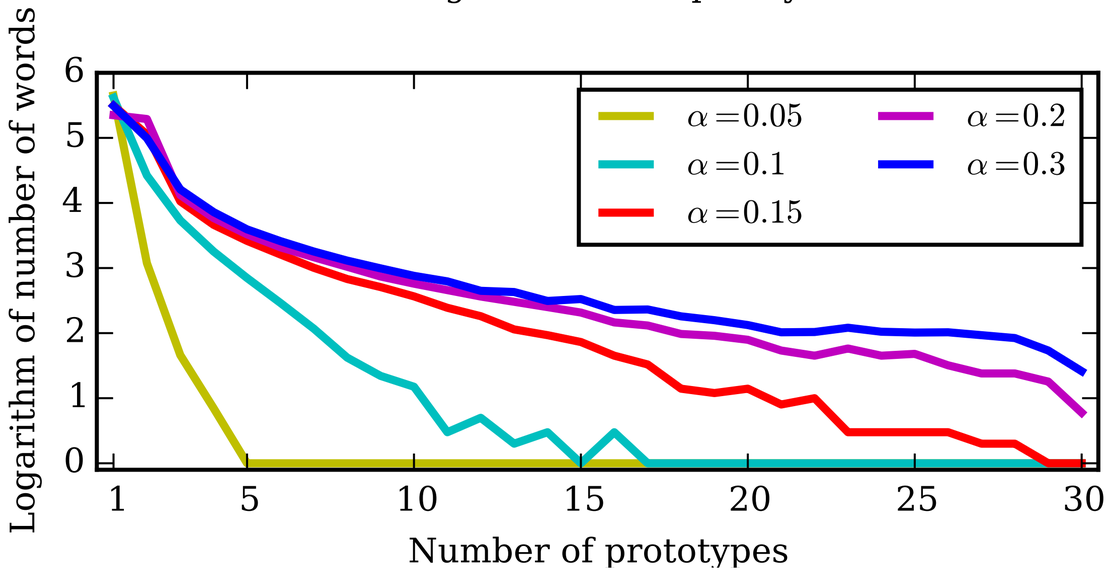
ALPHA	$p(z)$	nearest neighbours
		"light"
Skip-Gram	1.00	far-red, emitting
0.075	0.28	armoured, amx-13, kilcrease
	0.72	bright, sunlight, luminous
0.1	0.09	tvärbanan, hudson-bergen
	0.17	dark, bright, green
	0.09	4th, dragoons, 2nd
	0.26	radiation, ultraviolet
	0.28	darkness, shining, shadows
	0.11	self-propelled, armored
		"core"
Skip-Gram	1.00	cores, components, i7
0.075	0.3	competencies, curriculum
	0.34	cpu, cores, i7, powerxcell
	0.36	nucleus backbone
0.1	0.21	reactor, hydrogen-rich
	0.13	intel, processors
	0.27	curricular, competencies
	0.15	downtown, cores, center
	0.24	nucleus, rag-tag, roster



Sense discovery

ALPHA	$p(z)$	nearest neighbours
		"light"
Skip-Gram	1.00	far-red, emitting
0.075	0.28	armoured, amx-13, kilcrease
	0.72	bright, sunlight, luminous
0.1	0.09	tvärbanan, hudson-bergen
	0.17	dark, bright, green
	0.09	4th, dragoons, 2nd
	0.26	radiation, ultraviolet
	0.28	darkness, shining, shadows
	0.11	self-propelled, armored
		"core"
Skip-Gram	1.00	cores, components, i7
0.075	0.3	competencies, curriculum
	0.34	cpu, cores, i7, powerxcell
	0.36	nucleus backbone
0.1	0.21	reactor, hydrogen-rich
	0.13	intel, processors
	0.27	curricular, competencies
	0.15	downtown, cores, center
	0.24	nucleus, rag-tag, roster





Model selection

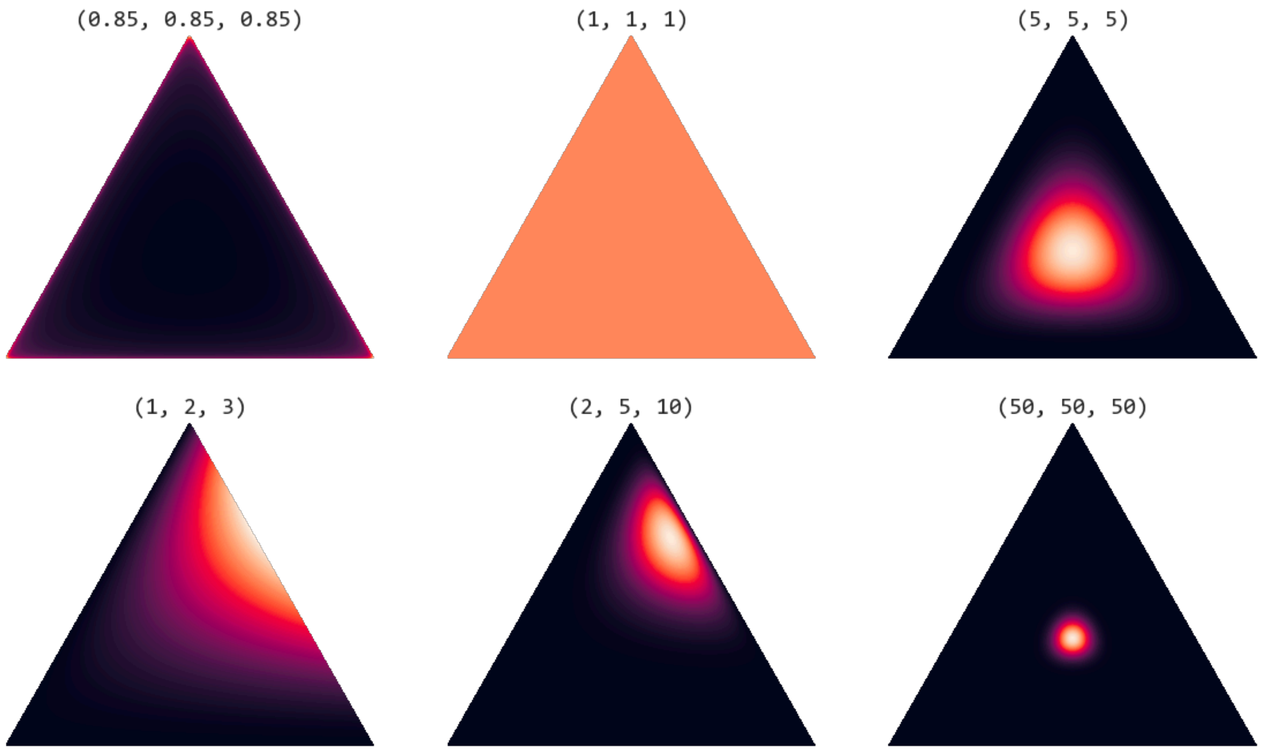
MODEL	LOG-LIKELIHOOD	ARI
Skip-Gram.300D	-7.403	-
Skip-Gram.600D	-7.387	-
AdaGram.300D $\alpha = 0.05$	-7.399	0.007
AdaGram.300D $\alpha = 0.1$	-7.385	0.226
AdaGram.300D $\alpha = 0.15$	-7.382	**0.268**
AdaGram.300D $\alpha = 0.2$	-7.378	0.254
AdaGram.300D $\alpha = 0.25$	**-7.375**	0.250
AdaGram.300D $\alpha = 0.5$	-7.387	0.230

Breaking Sticks and Ambiguities with Adaptive Skip-gram

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin and Dmitry Vetrov. In *AISTATS* 2016.

https://github.com/sbos/AdaGram.jl
https://bitbucket.org/sbos/adagram_deepbayes2019

Dirichlet distribution



Distribution over (discrete distribution)

$$\mathrm{Dir}(\pi|\alpha) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\pi_k^{\alpha_k-1}$$

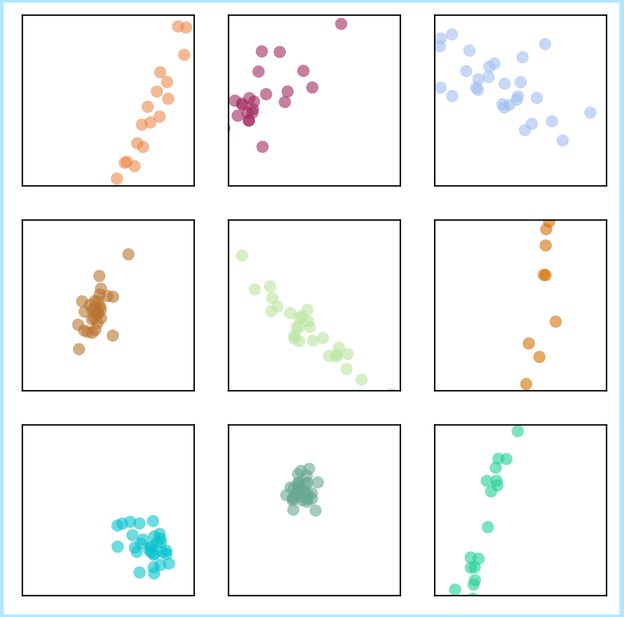
$$\pi_1,\ldots,\pi_K \sim \mathrm{Dir}(\alpha_1,\ldots,\alpha_K)$$

$$\sum_{k=1}^{K}\pi_k = 1, \quad \pi_k \geq 0$$

Dirichlet process

Base measure (prior)

$$H = \mathrm{NW}(\mu_0, \lambda, W, \nu)$$

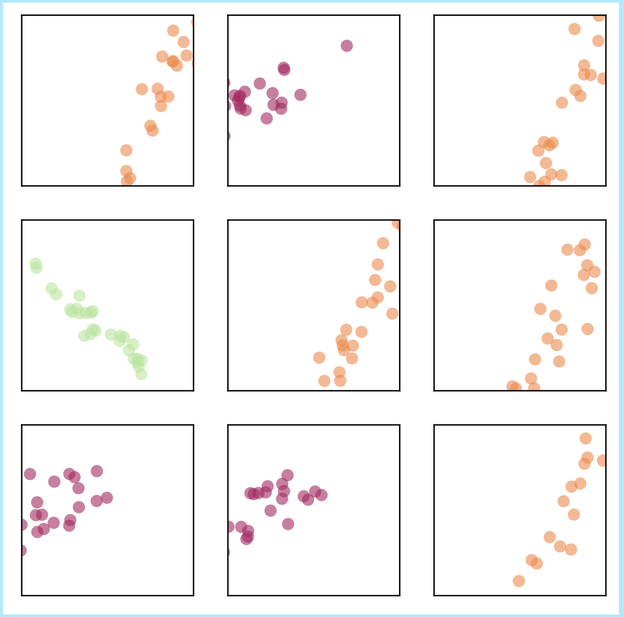


$$\mu, \Sigma \sim H$$

$$\mathrm{DP}(H, \alpha)$$

DP sample

$$G \sim \mathrm{DP}(H, \alpha)$$



$$\mu, \Sigma \sim G$$

Gaussian vs Dirichlet

Gaussian process (function estimation)

Prior

$$f \sim \mathrm{GP}(\mu, \Sigma)$$

Gaussian vs Dirichlet

Gaussian process (function estimation)

Prior

$$f \sim \mathrm{GP}(\mu, \Sigma)$$

Posterior over functions

$$p(f|X,Y) = \frac{p(f)p(Y|f,X)}{p(X,Y)}$$

Gaussian vs Dirichlet

Gaussian process (function estimation)

Prior

$$f \sim \mathrm{GP}(\mu, \Sigma)$$

Posterior over functions

$$p(f|X,Y) = \frac{p(f)p(Y|f,X)}{p(X,Y)}$$

Posterior predictive

$$p(y|x,X,Y) = \int p(f|X,Y)p(y|f,x)df$$

Gaussian vs Dirichlet

Gaussian process (function estimation)

Prior

$$f \sim \mathrm{GP}(\mu, \Sigma)$$

Posterior over functions

$$p(f|X,Y) = \frac{p(f)p(Y|f,X)}{p(X,Y)}$$

Posterior predictive

$$p(y|x,X,Y) = \int p(f|X,Y)p(y|f,x)df$$

Dirichlet process (density estimation)

Prior

$$G \sim \mathrm{DP}(H, \alpha)$$

Gaussian vs Dirichlet

Gaussian process (function estimation)

Prior

$$f \sim \mathrm{GP}(\mu, \Sigma)$$

Posterior over functions

$$p(f|X,Y) = \frac{p(f)p(Y|f,X)}{p(X,Y)}$$

Posterior predictive

$$p(y|x,X,Y) = \int p(f|X,Y)p(y|f,x)df$$

Dirichlet process (density estimation)

Prior

$$G \sim \mathrm{DP}(H, \alpha)$$

Posterior over densities

$$p(G|X) = \frac{p(G)p(X|G)}{p(X)}$$

Gaussian vs Dirichlet

Gaussian process (function estimation)

Prior

$$f \sim \mathrm{GP}(\mu, \Sigma)$$

Posterior over functions

$$p(f|X,Y) = \frac{p(f)p(Y|f,X)}{p(X,Y)}$$

Posterior predictive

$$p(y|x,X,Y) = \int p(f|X,Y)p(y|f,x)df$$

Dirichlet process (density estimation)

Prior

$$G \sim \mathrm{DP}(H, \alpha)$$

Posterior over densities

$$p(G|X) = \frac{p(G)p(X|G)}{p(X)}$$

Posterior predictive

$$p(x|X) = \int p(G|X)p(x|G)dG$$

Finite projections

Gaussian process (function estimation)

Projection is Gaussian

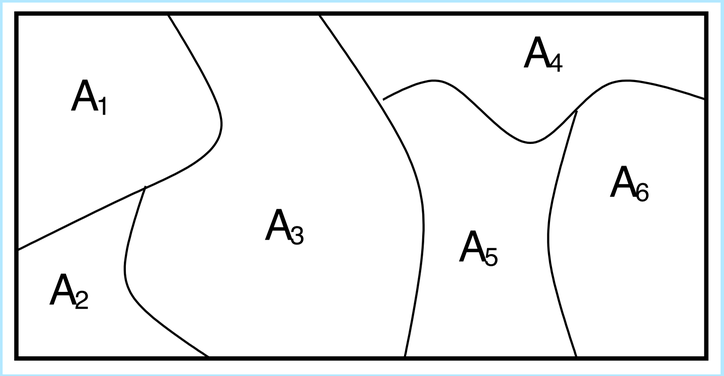
$$Y = f(X) \sim \mathcal{N}(\mu(X), \Sigma(X))$$

Dirichlet process (density estimation)

Projection is Dirichlet

$$(G(A_1), \ldots, G(A_n)) \sim \mathrm{Dir}(\alpha H(A_1), \ldots, \alpha H(A_n))$$

$$A_1 \cup \ldots \cup A_n = \mathcal{X}, \quad G \sim \mathrm{DP}(H, \alpha)$$



References

- **Skip-gram**: Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

- **Dirichlet process**

  - Ferguson, Thomas S. "A Bayesian analysis of some nonparametric problems." The annals of statistics (1973): 209-230.

  - Sethuraman, Jayaram. "A constructive definition of Dirichlet priors." Statistica sinica (1994): 639-650.

  - Teh, Yee Whye. "Dirichlet processes: Tutorial and practical course." *Gatsby Computational Neuroscience Unit, University College London* (2007).

References

- **Variational inference for Dirichlet process**

  - Blei, David M., and Michael I. Jordan. "Variational inference for Dirichlet process mixtures." Bayesian analysis 1.1 (2006): 121-143.

Image credits

- Slide 2: https://blog.evjang.com/2016/08/variational-bayes.html
- Slide 3: SHUTTERSTOCK / KYSLYNSKAHAL, https://artsyvoiceprint.com/products/tattoo-sound-wave-custom-art-black-white-digital-file, https://en.wikipedia.org/wiki/Support-vector_machine, https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
- Slide 33: https://towardsdatascience.com/dirichlet-distribution-a82ab942a879
- Slide 36: Yee Whye The, August 2007, MLSS