LDA with Amortized Inference

Nanbo Sun

Abstract

This report describes how to frame Latent Dirichlet Allocation (LDA) as a Variational Auto-Encoder (VAE) and use the Amortized Variational Inference (AVI) to optimize it. 1. We will introduce the LDA and use Mean Field Variational Inference (MFVI) to optimize it. 2. We collapse the topics in LDA model because we can not do backpropagation through a categorical variable. 3. We will introduce Gaussian VAE with AVI. 4. We frame the collapsed LDA as VAE and do AVI.

1 Prerequisite

Before reading this report, you **must** read the following papers carefully.

1. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
2. Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.
3. Figurnov, M., Mohamed, S., & Mnih, A. (2018). Implicit Reparameterization Gradients. arXiv preprint arXiv:1805.08498.

2 LDA with MFVI

2.1 Generative Process

For document d with K topics and V unique words,

1. draw a topic mixture $\theta \sim \text{Dir}(\alpha)$, where θ is a K-vector, and $\sum_k \theta_k = 1$;
2. for each of the N_m "word counts," independently
   (a) draw a topic $z_n \sim \text{Mult}(\theta)$;
   (b) draw a word $w_n \sim p\left(w_n \mid z_n, \beta\right)$, where β is a $K \times V$ matrix, each row of which defines a multinomial distribution over all the voxels. β_{ij} is the probability that word j appears given topic i.

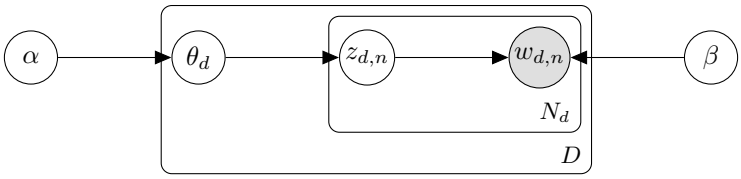

Figure 1: LDA model.

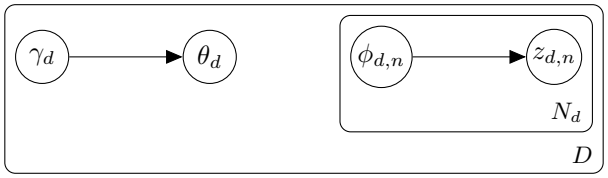

Figure 2: Variational distribution used to approximate the posterior in LDA.

2.2 Constructing the Lower Bound

From Figure 2, the variational distribution used to approximate the true posterior is factorizable as

$$q(\theta, \mathbf{z} \mid \gamma, \phi) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n).$$

The lower bound $\mathcal{L}(\gamma, \phi \mid \alpha, \beta)$ of the single-document[1] log likelihood $\log p(\mathbf{w} \mid \alpha, \beta)$ is computed using Jensen's inequality as follows

$$\begin{aligned}
\log p(\mathbf{w} \mid \alpha, \beta) &= \log \int \sum_{\mathbf{z}} p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) \, \mathrm{d}\theta \\
&= \log \int \sum_{\mathbf{z}} \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) \, q(\theta, \mathbf{z} \mid \gamma, \phi)}{q(\theta, \mathbf{z} \mid \gamma, \phi)} \mathrm{d}\theta \\
&= \log \int \sum_{\mathbf{z}} q(\theta, \mathbf{z} \mid \gamma, \phi) \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{q(\theta, \mathbf{z} \mid \gamma, \phi)} \mathrm{d}\theta \\
&= \log E_q \left\{ \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{q(\theta, \mathbf{z} \mid \gamma, \phi)} \right\} \\
&\geq E_q \{\log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)\} - E_q \{\log q(\theta, \mathbf{z} \mid \gamma, \phi)\} \\
&\triangleq \mathcal{L}(\gamma, \phi \mid \alpha, \beta).
\end{aligned} \tag{1}$$

The difference between the log likelihood and its lower bound can be proven to be in fact the KL divergence between the variational distribution and the true posterior.

$$\begin{aligned}
&\log p(\mathbf{w} \mid \alpha, \beta) - \mathcal{L}(\gamma, \phi \mid \alpha, \beta) \\
&= E_q \{\log p(\mathbf{w} \mid \alpha, \beta)\} - E_q \{\log p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)\} + E_q \{\log q(\theta, \mathbf{z} \mid \gamma, \phi)\} \\
&= E_q \left\{ \log \frac{p(\mathbf{w} \mid \alpha, \beta) \, q(\theta, \mathbf{z} \mid \gamma, \phi)}{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)} \right\}
\end{aligned}$$

[1] This also explains why the document subscript is dropped for simplicity hereafter.

$$= E_q \left\{ \log \frac{q\left(\theta, \mathbf{z} \mid \gamma, \phi\right)}{p\left(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta\right)} \right\}$$
$$= \mathrm{D}_{\mathrm{KL}}(q\left(\theta, \mathbf{z} \mid \gamma, \phi\right) \parallel p\left(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta\right)).$$

Therefore, maximizing the lower bound is equivalent to minimizing the KL divergence. That is, the variational distribution automatically approaches to the real posterior as we maximize the lower bound.

2.3 Expanding the Lower Bound

To maximize the lower bound, we first need to spell out the lower bound $\mathcal{L}\left(\gamma, \phi \mid \alpha, \beta\right)$ in terms of the model parameters (α, β) and the variational parameters (γ, ϕ). Continuing from (1), we have

$$\begin{aligned}
\mathcal{L}\left(\gamma, \phi \mid \alpha, \beta\right) &= E_q\left\{\log p\left(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta\right)\right\} - E_q\left\{\log q\left(\theta, \mathbf{z} \mid \gamma, \phi\right)\right\} \\
&= E_q\left\{\log \frac{p\left(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta\right)}{q\left(\theta, \mathbf{z} \mid \gamma, \phi\right)}\right\} \\
&= E_q\left\{\log \frac{p\left(\theta \mid \alpha\right) p\left(\mathbf{z} \mid \theta\right) p\left(\mathbf{w} \mid \mathbf{z}, \beta\right)}{q\left(\theta \mid \gamma\right) q\left(\mathbf{z} \mid \phi\right)}\right\} \qquad (2) \\
&= E_q\left\{\log p\left(\theta \mid \alpha\right)\right\} + E_q\left\{\log p\left(\mathbf{z} \mid \theta\right)\right\} + E_q\left\{\log p\left(\mathbf{w} \mid \mathbf{z}, \beta\right)\right\} \\
&\quad - E_q\left\{\log q\left(\theta \mid \gamma\right)\right\} - E_q\left\{\log q\left(\mathbf{z} \mid \phi\right)\right\}.
\end{aligned}$$

We now further expand each of the five terms in (2).

The first term is

$$\begin{aligned}
E_q\left\{\log p\left(\theta \mid \alpha\right)\right\} &= E_q\left\{\log \frac{\Gamma\left(\sum_{i=1}^K \alpha_i\right)}{\prod_{k=1}^K \Gamma\left(\alpha_k\right)} \prod_{k=1}^K \theta_k^{\alpha_k - 1}\right\} \\
&= E_q\left\{\log \Gamma\left(\sum_{i=1}^K \alpha_i\right) + \sum_{k=1}^K (\alpha_k - 1) \log \theta_k - \sum_{k=1}^K \log \Gamma\left(\alpha_k\right)\right\} \\
&= \sum_{k=1}^K (\alpha_k - 1) E_q\left\{\log \theta_k\right\} + \log \Gamma\left(\sum_{i=1}^K \alpha_i\right) - \sum_{k=1}^K \log \Gamma\left(\alpha_k\right) \\
&= \sum_{k=1}^K (\alpha_k - 1)\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^K \gamma_i\right)\right) + \log \Gamma\left(\sum_{i=1}^K \alpha_i\right) - \sum_{k=1}^K \log \Gamma\left(\alpha_k\right),
\end{aligned}$$

where $\Psi(\cdot)$ is the digamma function, the first derivative of the log Gamma function. The final line is due to the following property of the Dirichlet distribution as a member of the exponential family. If $\theta \sim Dir(\alpha)$, then $E_{p(\theta|\alpha)}\left\{\log \theta_i\right\} = \Psi(\alpha_i) - \Psi(\sum_{i=1}^K \alpha_i)$.

The second term is

$$\begin{aligned}
E_q\left\{\log p\left(\mathbf{z} \mid \theta\right)\right\} &= E_q\left\{\log \prod_{n=1}^N p\left(z_n \mid \theta\right)\right\} \\
&= E_q\left\{\log \prod_{n=1}^N \prod_{k=1}^K \theta_k^{\mathbb{1}_z(n,k)}\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{n=1}^{N} \sum_{k=1}^{K} E_q \left\{ \mathbb{1}_z(n,k) \log \theta_k \right\} \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} E_q \left\{ \mathbb{1}_z(n,k) \right\} E_q \left\{ \log \theta_k \right\} \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \phi_{n,k} \left(\Psi(\gamma_k) - \Psi \left(\sum_{i=1}^{K} \gamma_i \right) \right),
\end{aligned}
$$

where $\phi_{n,k}$ is the probability of the nth word being produced by topic k, and $\mathbb{1}(\cdot)$ is the indicator function.

We expand **the third term** as

$$
\begin{aligned}
E_q \left\{ \log p \left(\mathbf{w} \mid \mathbf{z}, \beta \right) \right\} &= E_q \left\{ \log \prod_{n=1}^{N} p \left(w_n \mid z_n, \beta \right) \right\} \\
&= E_q \left\{ \log \prod_{n=1}^{N} \prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{k,v}^{\mathbb{1}_z(n,k) \mathbb{1}_w(n,v)} \right\} \\
&= E_q \left\{ \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{v=1}^{V} \mathbb{1}_z(n,k) \mathbb{1}_w(n,v) \log \beta_{k,v} \right\} \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{v=1}^{V} E_q \left\{ \mathbb{1}_z(n,k) \right\} \mathbb{1}_w(n,v) \log \beta_{k,v} \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{v=1}^{V} \phi_{n,k} \mathbb{1}_w(n,v) \log \beta_{k,v}.
\end{aligned}
$$

Very similar to the first term, **the fourth term** is expanded as

$$
E_q \left\{ \log q \left(\theta \mid \gamma \right) \right\} = \sum_{k=1}^{K} (\gamma_k - 1) \left(\Psi(\gamma_k) - \Psi \left(\sum_{i=1}^{K} \gamma_i \right) \right) + \log \Gamma \left(\sum_{k=1}^{K} \gamma_k \right) - \sum_{k=1}^{K} \log \Gamma \left(\gamma_k \right).
$$

Finally, **the fifth term** is expanded as

$$
\begin{aligned}
E_q \left\{ \log q \left(\mathbf{z} \mid \phi \right) \right\} &= E_q \left\{ \log \prod_{n=1}^{N} q \left(z_n \mid \phi_n \right) \right\} \\
&= E_q \left\{ \log \prod_{n=1}^{N} \prod_{k=1}^{K} \phi_{n,k}^{\mathbb{1}_z(n,k)} \right\} \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} E_q \left\{ \mathbb{1}_z(n,k) \right\} \log \phi_{n,k} \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \phi_{n,k} \log \phi_{n,k}.
\end{aligned}
$$

Therefore, the fully expanded lower bound is

$$
\mathcal{L} \left(\gamma, \phi \mid \alpha, \beta \right) = \sum_{k=1}^{K} (\alpha_k - 1) \left(\Psi(\gamma_k) - \Psi \left(\sum_{i=1}^{K} \gamma_i \right) \right) + \log \Gamma \left(\sum_{i=1}^{K} \alpha_i \right) - \sum_{k=1}^{K} \log \Gamma \left(\alpha_k \right)
$$

$$
\begin{aligned}
&+ \sum_{n=1}^{N}\sum_{k=1}^{K} \phi_{n,k}\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right) \\
&+ \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{v=1}^{V} \phi_{n,k}\mathbb{1}_w(n,v)\log\beta_{k,v} \\
&- \sum_{k=1}^{K}(\gamma_k - 1)\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right) - \log\Gamma\left(\sum_{k=1}^{K}\gamma_k\right) + \sum_{k=1}^{K}\log\Gamma\left(\gamma_k\right) \\
&- \sum_{n=1}^{N}\sum_{k=1}^{K}\phi_{n,k}\log\phi_{n,k}.
\end{aligned} \tag{3}
$$

2.4 Maximizing the Lower Bound

In this section, we maximize the lower bound w.r.t. the variational parameters ϕ and γ. Recall that as the maximization runs, the KL divergence between the variational distribution and the true posterior drops (E-step of the variational EM algorithm).

2.4.1 Variational Multinomial

We first maximize Equation (3) w.r.t. $\phi_{n,k}$. Since $\sum_{k=1}^{K}\phi_{n,k} = 1$, this is a constrained optimization problem that can be solved by the Lagrange multiplier method. The Lagrangian w.r.t. $\phi_{n,k}$ is

$$
\mathcal{L}_{[\phi_{n,k}]} = \phi_{n,k}\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right) + \phi_{n,k}\log\beta_{k,v} - \phi_{n,k}\log\phi_{n,k} + \lambda_n\left(\sum_{i=1}^{K}\phi_{n,i} - 1\right),
$$

where λ_n is the Lagrange multiplier. Taking the derivative, we get

$$
\frac{\partial}{\partial\phi_{n,k}}\mathcal{L}_{[\phi_{n,k}]} = \Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right) + \log\beta_{k,v} - \log\phi_{n,k} - 1 + \lambda_n.
$$

Setting this derivative to zero yields

$$
\begin{aligned}
\phi_{n,k} &= \beta_{k,v}\exp\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right) + \lambda_n - 1\right) \\
&\propto \beta_{k,v}\exp\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right).
\end{aligned}
$$

2.4.2 Variational Dirichlet

Now we maximize Equation (3) w.r.t. γ_k, the kth component of the Dirichlet parameter. Only the terms containing γ_k are retained.

$$
\mathcal{L}_{[\gamma]} = \sum_{k=1}^{K}(\alpha_k - 1)\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right)
$$

$$+\sum_{n=1}^{N}\phi_{n,k}\left(\Psi(\gamma_k)-\Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right)$$
$$-\sum_{k=1}^{K}(\gamma_k-1)\left(\Psi(\gamma_k)-\Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right)-\log\Gamma\left(\sum_{i=1}^{K}\gamma_i\right)+\sum_{k=1}^{K}\log\Gamma\left(\gamma_k\right)$$

Taking the derivative w.r.t. γ_k, we have

$$\begin{aligned}
\frac{\partial}{\partial\gamma_k}\mathcal{L}_{[\gamma]} &= \left(\Psi'(\gamma_k)-\Psi'\left(\sum_{i=1}^{K}\gamma_i\right)\right)(\alpha_k-1)\\
&+\left(\Psi'(\gamma_k)-\Psi'\left(\sum_{i=1}^{K}\gamma_i\right)\right)\sum_{n=1}^{N}\phi_{n,k}\\
&-\left(\Psi'(\gamma_k)-\Psi'\left(\sum_{i=1}^{K}\gamma_i\right)\right)(\gamma_k-1)-\left(\Psi(\gamma_k)-\Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right)\\
&-\frac{\Psi\left(\sum_{i=1}^{K}\gamma_i\right)}{\Gamma\left(\sum_{i=1}^{K}\gamma_i\right)}+\frac{\Psi(\gamma_k)}{\Gamma\left(\gamma_k\right)}\\
&=\left(\Psi'(\gamma_k)-\Psi'\left(\sum_{i=1}^{K}\gamma_i\right)\right)\left(\alpha_k+\sum_{n=1}^{N}\phi_{n,k}-\gamma_k\right)-\Psi(\gamma_k)+\Psi\left(\sum_{i=1}^{K}\gamma_i\right)\\
&-\Psi\left(\sum_{i=1}^{K}\gamma_i\right)+\Psi(\gamma_k)\\
&=\left(\Psi'(\gamma_k)-\Psi'\left(\sum_{i=1}^{K}\gamma_i\right)\right)\left(\alpha_k+\sum_{n=1}^{N}\phi_{n,k}-\gamma_k\right).
\end{aligned}$$

Setting it to zero, we have

$$\gamma_k=\alpha_k+\sum_{n=1}^{N}\phi_{n,k}.$$

2.5 Estimating Model Parameters

The previous section is the E-step of the variational EM algorithm, where we tighten the lower bound w.r.t. the variational parameters; this section is the M-step, where we maximize the lower bound w.r.t. the model parameters α and β. Now we add back the document subscript to consider the whole corpus.

By the assumed exchangeability of the documents, the overall log likelihood of the corpus is just the sum of all the documents' log likelihoods, and the overall lower bound is just the sum of the individual lower bounds. Again, only the terms involving β are left in the overall lower bound. Adding the Lagrange multipliers, we obtain

$$\mathcal{L}_{[\beta]}=\sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{k=1}^{K}\sum_{v=1}^{V}\phi_{d,n,k}\mathbb{1}_w(d,n,v)\log\beta_{k,v}+\sum_{k=1}^{K}\lambda_k\left(\sum_{v=1}^{V}\beta_{k,v}-1\right).$$

Taking the derivative w.r.t. $\beta_{k,v}$ and setting it to zero, we have

$$\begin{aligned}\frac{\partial}{\partial \beta_{k,v}}\mathcal{L}_{[\beta]} &= \sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{d,n,k}\mathbb{1}_w(d,n,v)\frac{1}{\beta_{k,v}} + \lambda_k = 0 \\ &\Rightarrow \beta_{k,v} = -\frac{1}{\lambda_k}\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{d,n,k}\mathbb{1}_w(d,n,v) \\ &\Rightarrow \beta_{k,v} \propto \sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{d,n,k}\mathbb{1}_w(d,n,v).\end{aligned}$$

Similarly, for α, we have

$$\begin{aligned}\mathcal{L}_{[\alpha]} &= \sum_{d=1}^{D}\left(\sum_{k=1}^{K}(\alpha_k - 1)\left(\Psi(\gamma_{d,k}) - \Psi\left(\sum_{i=1}^{K}\gamma_{d,i}\right)\right) + \log\Gamma\left(\sum_{i=1}^{K}\alpha_i\right) - \sum_{k=1}^{K}\log\Gamma\left(\alpha_k\right)\right) \\ \frac{\partial}{\partial \alpha_k}\mathcal{L}_{[\alpha]} &= \sum_{d=1}^{D}\left(\Psi(\gamma_{d,k}) - \Psi\left(\sum_{i=1}^{K}\gamma_{d,i}\right) + \Psi\left(\sum_{i=1}^{K}\alpha_i\right) - \Psi(\alpha_k)\right) \\ &= \sum_{d=1}^{D}\left(\Psi(\gamma_{d,k}) - \Psi\left(\sum_{i=1}^{K}\gamma_{d,i}\right)\right) + D\left(\Psi\left(\sum_{i=1}^{K}\alpha_i\right) - \Psi(\alpha_k)\right).\end{aligned}$$

Since the derivative also depends on other $\alpha_{k'\neq k}$, we compute the Hessian

$$\frac{\partial^2}{\partial \alpha_k \partial \alpha_{k'}}\mathcal{L}_{[\alpha]} = D\Psi'\left(\sum_{i=1}^{K}\alpha_i\right) - D\delta(k-k')\Psi(\alpha_k),$$

and notice that its form allows for the linear-time Newton-Raphson algorithm.

3 Collapsed LDA without topics

3.1 Generative Process

For document d with K topics and V unique words,

1. draw a topic mixture $\theta \sim \mathrm{Dir}(\alpha)$, where θ is a K-vector, and $\sum_k \theta_k = 1$;
2. for each of the N_m "word counts," independently
   (a) draw a word $w_n \sim p\left(w_n \mid \theta, \beta\right)$, where β is a $K \times V$ matrix, each row of which defines a multinomial distribution over all the voxels. β_{ij} is the probability that word j appears given topic i.

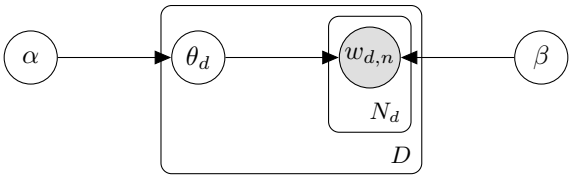

Figure 3: Collapsed LDA model.

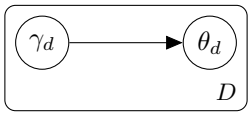

Figure 4: Variational distribution used to approximate the posterior in collapsed LDA.

3.2 Comparing LDA vs collapsed LDA

Table 1: Comparing LDA vs collapsed LDA.

	LDA	collapsed LDA
prior	$p(\theta) = \text{Dir}(\alpha)$	$p(\theta) = \text{Dir}(\alpha)$
likelihood	$p(w\|z = k, \beta) = \text{Cat}(\beta_k)$	$p(w\|\theta, \beta) = \text{Cat}(\theta\beta)$
posterior	$p(z, \theta\|w)$	$p(\theta\|w)$
approximate posterior	$q(z, \theta\|\phi, \gamma) = q(z\|\phi)q(\theta\|\gamma) = \text{Cat}(\phi)\text{Dir}(\gamma)$	$q(\theta\|\gamma) = \text{Dir}(\gamma)$

Proof of likelihood for collapsed LDA

$$\begin{aligned} p(w_{dn}|\theta_d, \beta) &= \sum_{z_{dn}} p(z_{dn}, w_{dn}|\theta_d, \beta) \\ &= \sum_{z_{dn}} p(z_{dn}|\theta)p(w_{dn}|z_{dn}, \beta) \\ &= \sum_k \theta_{dk}\beta_{kw} \end{aligned}$$

4 Gaussian VAE with AVI

4.1 Cost Function

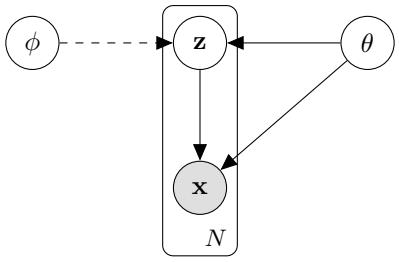

Figure 5: SVI

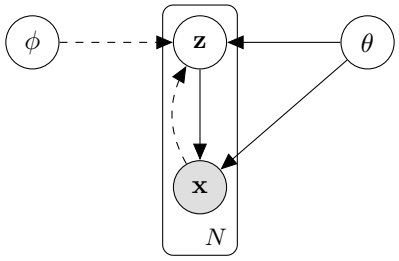

Figure 6: VAE

circles	random variables
shaded circles	observed random variables
unshaded circles	hidden random variables
N	number of samples
$\boldsymbol{\theta}$	generative model parameters
$\boldsymbol{\phi}$	variational approximation parameters

$$
\begin{aligned}
\mathscr{L}(\theta, \phi; \boldsymbol{x}^{(i)}) &= E_{q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})}\left[-\log q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)}) + \log p_\theta(\boldsymbol{z}, \boldsymbol{x}^{(i)})\right] \\
&= \log p_\theta(\boldsymbol{x}^{(i)}) - D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)}) \| p_\theta(\boldsymbol{z}|\boldsymbol{x}^{(i)})) && \text{(lower bound)} \\
&= -D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)}) \| p_\theta(\boldsymbol{z}, \boldsymbol{x}^{(i)})) && \text{(joint-constrastive)} \\
&= E_{q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})}\left[\log p_\theta(\boldsymbol{x}^{(i)}|\boldsymbol{z})\right] - D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)}) \| p_\theta(\boldsymbol{z})) && \text{(prior-contrastive)} \\
&= \text{Reconstruction - Regularization}
\end{aligned}
$$

$\mathscr{L}(\theta, \phi; \boldsymbol{x}^{(i)})$ is the (variational) lower bound on the marginal log likelihood of the data point $\boldsymbol{x}^{(i)}$. The sum $\sum_{i=1}^{N} \mathscr{L}(\theta, \phi; \boldsymbol{x}^{(i)})$ is the evidence lower bound objective (**ELBO**).

4.2 Optimization

We want to maximize $\mathscr{L}(\theta, \phi; \boldsymbol{x}^{(i)})$ in order to maximize the marginal likelihood $\log p_\theta(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)})$ of the data.

Therefore, we need to differentiate and optimize $\mathscr{L}(\theta, \phi; \boldsymbol{x}^{(i)})$ with respect to θ and ϕ.

- Differentiate with respect to θ

$$\begin{aligned}
\nabla_\theta \mathscr{L}(\theta, \phi; \boldsymbol{x}^{(i)}) &= \nabla_\theta E_{q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})}\left[-\log q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)}) + \log p_\theta(\boldsymbol{z}, \boldsymbol{x}^{(i)})\right] \\
&= E_{q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})}\left[-\nabla_\theta \log q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)}) + \nabla_\theta \log p_\theta(\boldsymbol{z}, \boldsymbol{x}^{(i)})\right] \text{ (differentiate inside expectation)} \\
&= E_{q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})}\left[-0 + \frac{\nabla_\theta p_\theta(\boldsymbol{z}, \boldsymbol{x}^{(i)})}{p_\theta(\boldsymbol{z}, \boldsymbol{x}^{(i)})}\right]
\end{aligned}$$

- Differentiate with respect to ϕ

$$\begin{aligned}
\nabla_\phi \mathscr{L}(\theta, \phi; \boldsymbol{x}^{(i)}) &= \nabla_\phi E_{q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})}\left[-\log q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)}) + \log p_\theta(\boldsymbol{z}, \boldsymbol{x}^{(i)})\right] \\
&\neq E_{q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})}\left[-\nabla_\phi \log q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)}) + \nabla_\phi \log p_\theta(\boldsymbol{z}, \boldsymbol{x}^{(i)})\right] \quad \text{(can't move inside expectation)}
\end{aligned}$$

4.2.1 Score Function Estimators

Idea

Differentiate the density $q_\phi(\boldsymbol{z})$

$$\begin{aligned}
\nabla_\phi E_{q_\phi(\boldsymbol{z})}\left[f(\boldsymbol{z})\right] &= \nabla_\phi \int_z f(\boldsymbol{z}) q_\phi(\boldsymbol{z}) dz \\
&= \int_z f(\boldsymbol{z}) \left[\nabla_\phi q_\phi(\boldsymbol{z})\right] dz && \text{(differentiate the density } q_\phi(\boldsymbol{z})\text{)} \\
&= \int_z f(\boldsymbol{z}) \left[q_\phi(\boldsymbol{z}) \nabla_\phi \log q_\phi(\boldsymbol{z})\right] dz && \text{(REINFORCE trick)} \\
&= \int_z \left[f(\boldsymbol{z}) \nabla_\phi \log q_\phi(\boldsymbol{z})\right] q_\phi(\boldsymbol{z}) dz \\
&= E_{q_\phi(\boldsymbol{z})}\left[f(\boldsymbol{z}) \nabla_\phi \log q_\phi(\boldsymbol{z})\right] && \text{(score function estimator)} \\
&\approx \frac{1}{L} \sum_{l=1}^{L} f(\boldsymbol{z}^{(l)}) \nabla_\phi \log q_\phi(\boldsymbol{z}^{(l)}) && \text{(Monte Carlo approximation)}
\end{aligned}$$

This estimator has been used in LDA with MFVI or SVI.

4.2.2 Pathwise Gradient Estimators

4.2.2.1 Explicit Reparametrization Gradients

Idea

- Differentiate the function $f(\boldsymbol{z})$
- Apply a standardization function $S_\phi(\boldsymbol{z}) = \epsilon \sim q(\epsilon)$ to remove dependence of $q_\phi(\boldsymbol{z})$ on ϕ
- $S_\phi(\boldsymbol{z})$ should be continuously differentiable w.r.t. parameters ϕ and argument $\boldsymbol{z}$ and invertible $\boldsymbol{z} = S_\phi^{-1}(\epsilon)$

$$
\begin{aligned}
\nabla_\phi E_{q_\phi(\boldsymbol{z})}\left[f(\boldsymbol{z})\right] &= \nabla_\phi \int_z f(\boldsymbol{z}) q_\phi(\boldsymbol{z}) dz \\
&= \nabla_\phi \int_\epsilon f(S_\phi^{-1}(\epsilon)) q_\phi(S_\phi^{-1}(\epsilon)) \frac{dS_\phi^{-1}(\epsilon)}{d\epsilon} d\epsilon && \text{(integration by subsititution)} \\
&= \nabla_\phi \int_\epsilon f(S_\phi^{-1}(\epsilon)) q(\epsilon) d\epsilon && (q(\epsilon) = q_\phi(S_\phi^{-1}(\epsilon)) \frac{dS_\phi^{-1}(\epsilon)}{d\epsilon}) \\
&= E_{q(\epsilon)}\left[\nabla_\phi f(S_\phi^{-1}(\epsilon))\right] \\
&= E_{q(\epsilon)}\left[\nabla_{\boldsymbol{z}} f(S_\phi^{-1}(\epsilon)) \nabla_\phi S_\phi^{-1}(\epsilon)\right] && \text{(chain rule)}
\end{aligned}
$$

4.2.2.2 Implicit Reparametrization Gradients

Idea

$$
\begin{aligned}
S_\phi(\boldsymbol{z}) &= \epsilon \\
\nabla_{\boldsymbol{z}} S_\phi(\boldsymbol{z}) \nabla_\phi \boldsymbol{z} + \nabla_\phi S_\phi(\boldsymbol{z}) &= 0 && \text{(take the gradient } \nabla_\phi \text{ and chain rule)} \\
\nabla_\phi \boldsymbol{z} &= -(\nabla_{\boldsymbol{z}} S_\phi(\boldsymbol{z}))^{-1} \nabla_\phi S_\phi(\boldsymbol{z})
\end{aligned}
$$

$$
\begin{aligned}
\nabla_\phi E_{q_\phi(\boldsymbol{z})}\left[f(\boldsymbol{z})\right] &= \nabla_\phi \int_z f(\boldsymbol{z}) q_\phi(\boldsymbol{z}) dz \\
&= \nabla_\phi \int_\epsilon f(S_\phi^{-1}(\epsilon)) q_\phi(S_\phi^{-1}(\epsilon)) \frac{dS_\phi^{-1}(\epsilon)}{d\epsilon} d\epsilon && \text{(integration by subsititution)} \\
&= \nabla_\phi \int_\epsilon f(S_\phi^{-1}(\epsilon)) q(\epsilon) d\epsilon && (q(\epsilon) = q_\phi(S_\phi^{-1}(\epsilon)) \frac{dS_\phi^{-1}(\epsilon)}{d\epsilon}) \\
&= E_{q(\epsilon)}\left[\nabla_\phi f(S_\phi^{-1}(\epsilon))\right] \\
&= E_{q(\epsilon)}\left[\nabla_{\boldsymbol{z}} f(S_\phi^{-1}(\epsilon)) \nabla_\phi S_\phi^{-1}(\epsilon)\right] && \text{(chain rule)} \\
&= E_{q(\epsilon)}\left[\nabla_{\boldsymbol{z}} f(S_\phi^{-1}(\epsilon)) \nabla_\phi \boldsymbol{z}\right] \\
&= E_{q(\epsilon)}\left[-\nabla_{\boldsymbol{z}} f(S_\phi^{-1}(\epsilon)) (\nabla_{\boldsymbol{z}} S_\phi(\boldsymbol{z}))^{-1} \nabla_\phi S_\phi(\boldsymbol{z})\right] \\
&= \int_\epsilon \left[-\nabla_{\boldsymbol{z}} f(z) (\nabla_{\boldsymbol{z}} S_\phi(\boldsymbol{z}))^{-1} \nabla_\phi S_\phi(\boldsymbol{z})\right] q(\epsilon) d\epsilon \\
&= \int_\epsilon \left[-\nabla_{\boldsymbol{z}} f(z) (\nabla_{\boldsymbol{z}} S_\phi(\boldsymbol{z}))^{-1} \nabla_\phi S_\phi(\boldsymbol{z})\right] q_\phi(\boldsymbol{z}) d\boldsymbol{z} \\
&= E_{q_\phi(\boldsymbol{z})}\left[-\nabla_{\boldsymbol{z}} f(z) (\nabla_{\boldsymbol{z}} S_\phi(\boldsymbol{z}))^{-1} \nabla_\phi S_\phi(\boldsymbol{z})\right]
\end{aligned}
$$

4.2.2.3 Comparision between explicit and implicit

Table 1: Comparison of the two reparameterization types. While they provide the same result, the implicit version is easier to implement for distributions such as Gamma because it does not require inverting the standardization function $\mathcal{S}_{\boldsymbol{\phi}}(\boldsymbol{z})$.

	Explicit reparameterization	Implicit reparameterization (proposed)
Forward pass	Sample $\boldsymbol{\varepsilon} \sim q(\boldsymbol{\varepsilon})$ Set $\boldsymbol{z} \leftarrow \mathcal{S}_{\boldsymbol{\phi}}^{-1}(\boldsymbol{\varepsilon})$	Sample $\boldsymbol{z} \sim q(\boldsymbol{z}\vert\boldsymbol{\phi})$
Backward pass	Set $\nabla_{\boldsymbol{\phi}}\boldsymbol{z} \leftarrow \nabla_{\boldsymbol{\phi}}\mathcal{S}_{\boldsymbol{\phi}}^{-1}(\boldsymbol{\varepsilon})$ Set $\nabla_{\boldsymbol{\phi}} f(\boldsymbol{z}) \leftarrow \nabla_{\boldsymbol{z}} f(\boldsymbol{z})\nabla_{\boldsymbol{\phi}}\boldsymbol{z}$	Set $\nabla_{\boldsymbol{\phi}}\boldsymbol{z} \leftarrow -(\nabla_{\boldsymbol{z}}\mathcal{S}_{\boldsymbol{\phi}}(\boldsymbol{z}))^{-1}\nabla_{\boldsymbol{\phi}}\mathcal{S}_{\boldsymbol{\phi}}(\boldsymbol{z})$ Set $\nabla_{\boldsymbol{\phi}} f(\boldsymbol{z}) \leftarrow \nabla_{\boldsymbol{z}} f(\boldsymbol{z})\nabla_{\boldsymbol{\phi}}\boldsymbol{z}$

4.3 Model

Table 2: Comparing LDA vs collapsed LDA vs Gaussian VAE.

	LDA	collapsed LDA	VAE
prior	$p(\theta) = \text{Dir}(\alpha)$	$p(\theta) = \text{Dir}(\alpha)$	$p_\theta(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
likelihood	$p(w\vert z = k, \beta) = \text{Cat}(\beta_k)$	$p(w\vert\theta, \beta) = \text{Cat}(\theta\beta)$	$p_\theta(\boldsymbol{x}\vert\boldsymbol{z}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{I})$
posterior	$p(z, \theta\vert w)$	$p(\theta\vert w)$	$p_\theta(\mathbf{z}\vert\mathbf{x})$
approximate posterior	$q(z, \theta\vert\phi, \gamma) = q(z\vert\phi)q(\theta\vert\gamma) = \text{Cat}(\phi)\text{Dir}(\gamma)$	$q(\theta\vert\gamma) = \text{Dir}(\gamma)$	$q_\phi(\mathbf{z}\vert\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{I})$

$$
\begin{aligned}
\mathscr{L}(\theta, \phi; \boldsymbol{x}^{(i)}) &= E_{q_\phi(\boldsymbol{z}\vert\boldsymbol{x}^{(i)})}\left[\log p_\theta(\boldsymbol{x}^{(i)}\vert\boldsymbol{z})\right] - D_{KL}(q_\phi(\boldsymbol{z}\vert\boldsymbol{x}^{(i)})\|p_\theta(\boldsymbol{z})) \\
\mathscr{L}^B(\theta, \phi; \boldsymbol{x}^{(i)}) &\approx \frac{1}{L}\sum_{l=1}^{L}\log p_\theta(\boldsymbol{x}^{(i)}\vert\boldsymbol{z}^{(i,l)}) - D_{KL}(q_\phi(\boldsymbol{z}\vert\boldsymbol{x}^{(i)})\|p_\theta(\boldsymbol{z})) \\
&= \frac{1}{L}\sum_{l=1}^{L}\log\mathcal{N}(\boldsymbol{x}^{(i)}; \mu, \sigma^2\boldsymbol{I}) - D_{KL}(q_\phi(\boldsymbol{z}\vert\boldsymbol{x}^{(i)})\|p_\theta(\boldsymbol{z})) \\
&= -\frac{1}{L}\sum_{l=1}^{L}\frac{|\boldsymbol{x}^{(i)} - \mu|^2}{\sigma^2} - D_{KL}(q_\phi(\boldsymbol{z}\vert\boldsymbol{x}^{(i)})\|p_\theta(\boldsymbol{z})) \quad \text{(discarding some constants)} \\
&= -\frac{1}{L}\sum_{l=1}^{L}\frac{|\boldsymbol{x}^{(i)} - \mu|^2}{\sigma^2} + \frac{1}{2}\sum_{j=1}^{J}(1 + \log\sigma_j^2 - \mu_j^2 - \sigma_j^2) \\
&= -MSE + Regularization
\end{aligned}
$$

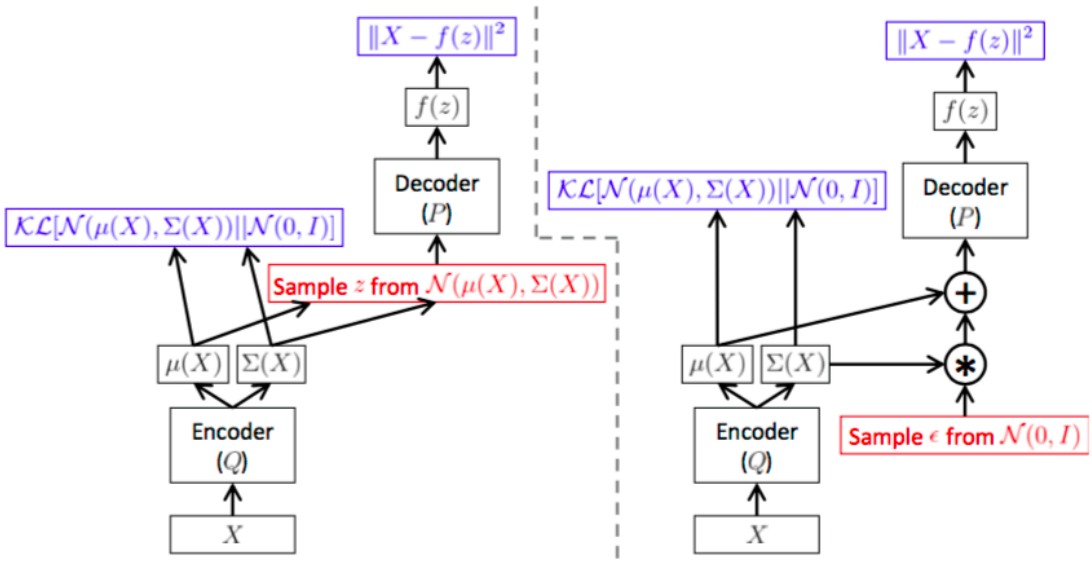

Figure 7: Gaussian VAE

4.3.1 Encoder

$$\left[\boldsymbol{\mu}^{(i)}, \boldsymbol{\sigma}^{(i)}\right] = MLP(\boldsymbol{x}^{(i)})$$

- Explicit reparameterization

$$\boldsymbol{\epsilon}^{(l)} \sim p(\boldsymbol{\epsilon}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$
$$\boldsymbol{z}^{(i,l)} = S_{\phi}^{-1}(\boldsymbol{\epsilon}^{(l)}, \boldsymbol{x}^{(i)}) = \boldsymbol{\mu}^{(i)} + \boldsymbol{\sigma}^{(i)} \odot \boldsymbol{\epsilon}^{(l)}$$

- Implicit reparameterization

$$\boldsymbol{z}^{(i,l)} \sim q_{\phi}(\boldsymbol{z}|\boldsymbol{x}^{(i)}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}^{(i)}, \boldsymbol{\sigma}^{2(i)}\boldsymbol{I})$$

4.3.2 Decoder

- θ without reparameterization

$$\boldsymbol{x}^{(i)} \sim p_{\theta}(\boldsymbol{x}^{(i)}|\boldsymbol{z}^{(i,l)}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{I})$$

- θ with reparametrization

$$[\boldsymbol{\mu}, \boldsymbol{\sigma}] = MLP(\boldsymbol{x}^{(i)})$$
$$\boldsymbol{x}^{(i)} \sim p_{\theta}(\boldsymbol{x}^{(i)}|\boldsymbol{z}^{(i,l)}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{I}))$$

5 Collapsed LDA as VAE with AVI

5.1 Model

Table 3: Comparing LDA vs collapsed LDA vs Gaussian VAE.

	LDA	collapsed LDA	VAE
prior	$p(\theta) = \text{Dir}(\alpha)$	$p(\theta) = \text{Dir}(\alpha)$	$p_\theta(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
likelihood	$p(w\|z = k, \beta) = \text{Cat}(\beta_k)$	$p(w\|\theta, \beta) = \text{Cat}(\theta\beta)$	$p_\theta(\boldsymbol{x}\|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{I})$
posterior	$p(z, \theta\|w)$	$p(\theta\|w)$	$p_\theta(\mathbf{z}\|\mathbf{x})$
approximate posterior	$q(z, \theta\|\phi, \gamma) = q(z\|\phi)q(\theta\|\gamma) = \text{Cat}(\phi)\text{Dir}(\gamma)$	$q(\theta\|\gamma) = \text{Dir}(\gamma)$	$q_\phi(\mathbf{z}\|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{I})$

$$\mathscr{L}(\theta, \phi; \boldsymbol{x}^{(i)}) = E_{q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})}\left[\log p_\theta(\boldsymbol{x}^{(i)}|\boldsymbol{z})\right] - D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})\|p_\theta(\boldsymbol{z}))$$

$$\begin{aligned}\mathscr{L}^B(\theta, \phi; \boldsymbol{x}^{(i)}) &\approx \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(\boldsymbol{x}^{(i)}|\boldsymbol{z}^{(i,l)}) - D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})\|p_\theta(\boldsymbol{z})) \\ &= \frac{1}{L}\sum_{l=1}^{L} \log Cat(\boldsymbol{w}^{(i)}; \theta\beta) - D_{KL}(Dir(\theta; \gamma)\|Dir(\theta; \alpha))\end{aligned}$$

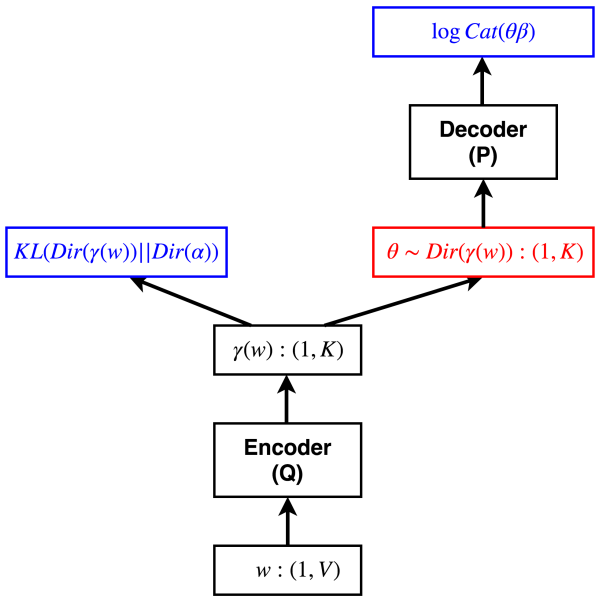

Figure 8: LDA VAE

5.1.1 Encoder

- Implicit reparameterization

$$\boldsymbol{\gamma}^{(i)} = MLP(\boldsymbol{w}^{(i)})$$
$$\boldsymbol{\theta}^{(i,l)} \sim q_\phi(\boldsymbol{\theta}|\boldsymbol{w}^{(i)}) = \text{Dir}(\boldsymbol{\theta}; \boldsymbol{\gamma}^{(i)})$$

5.1.2 Decoder

- θ without reparameterization

$$\boldsymbol{w}^{(i)} \sim p_\theta(\boldsymbol{w}^{(i)}|\boldsymbol{\theta}^{(i,l)}) = \text{Cat}(\boldsymbol{w}; \boldsymbol{\theta}\boldsymbol{\beta})$$