Lecture 15

Performer, Variational Autoencoders

Rethinking Attention with Performers

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, Adrian Weller

- “We introduce the first Transformer architectures, *Performers,* capable of **provably** accurate and practical estimations of regular (softmax) full rank attention, but of only linear space and time complexity and **not relying on any priors** such as sparsity or low-rankness. Performers use the *Fast Attention Via positive Orthogonal Random features* (FAVOR+) mechanism”

Generalized definition

- Define three different vectors corresponding to each word.
  - Input $x \in \mathbb{R}^d$ $x = [\underline{x_1} - -x_n]_{d \times n}$
  - Key
  - Query
  - Value

Generalized definition

- Define three different vectors corresponding to each word.
  - Input $x \in \mathbb{R}^d$ $X = [\underline{x_1} - - x_n]_{d \times n}$
  - Key $k \in \mathbb{R}^p$ $K = \underbrace{W_k^T}_{p \times d} X_{d \times n}$ $K = [\underline{x_1} - - k_n]_{p \times n}$
  - Query
  - Value

Generalized definition

- Define three different vectors corresponding to each word.
  - Input $x \in \mathbb{R}^d$ $X = [\underline{x_1} - -x_n]_{d\times n}$
  - Key $k \in \mathbb{R}^p$ $K = \underbrace{W_k^T}_{p\times d} X_{d\times n}$ $K = [\underline{x_1} - -k_n]_{p\times n}$
  - Query $q \in \mathbb{R}^p$ $Q = \underbrace{W_q^T}_{p\times d} X_{d\times n}$ $Q = [q_1 - -q_n]_{p\times n}$
  - Value

Generalized definition

- Define three different vectors corresponding to each word.
  - Input $x \in \mathbb{R}^d$ $X = [\underline{x_1} - -x_n]_{d\times n}$
  - Key $k \in \mathbb{R}^p$ $K = \underbrace{W_k^T}_{p\times d} X_{d\times n}$ $K = [\underline{x_1} - -k_n]_{p\times n}$
  - Query $q \in \mathbb{R}^p$ $Q = \underbrace{W_q^T}_{p\times d} X_{d\times n}$ $Q = [q_1 - -q_n]_{p\times n}$
  - Value $v \in \mathbb{R}^m$ $V = \underbrace{W_v^T}_{m\times d} X_{d\times n}$ $V = [\underline{v_1} - -v_n]_{m\times n}$

Generalized definition

- Define three different vectors corresponding to each word.
  - Input $x \in \mathbb{R}^d$ $X = [\underline{x_1} - -x_n]_{d\times n}$
  - Key $k \in \mathbb{R}^p$ $K = \underbrace{W_k^T}_{p\times d} X_{d\times n}$ $K = [\underline{x_1} - -k_n]_{p\times n}$
  - Query $q \in \mathbb{R}^p$ $Q = \underbrace{W_q^T}_{p\times d} X_{d\times n}$ $Q = [q_1 - -q_n]_{p\times n}$
  - Value $v \in \mathbb{R}^m$ $V = \underbrace{W_v^T}_{m\times d} X_{d\times n}$ $V = [\underline{v_1} - -v_n]_{m\times n}$

$$\underset{m\times n}{\mathcal{L}} = \underset{m\times n}{V}\ softmax\ \underbrace{\left(\frac{Q^T K}{\sqrt{P}}\right)}_{n\times n}$$

Generalized definition

- Define three different vectors corresponding to each word.
  - Input $x \in \mathbb{R}^d$ $X = [\underline{x_1} - -x_n]_{d\times n}$
  - Key $k \in \mathbb{R}^p$ $K = \underbrace{W_k^T}_{p\times d} X_{d\times n}$ $K = [\underline{x_1} - -k_n]_{p\times n}$
  - Query $q \in \mathbb{R}^p$ $Q = \underbrace{W_q^T}_{p\times d} X_{d\times n}$ $Q = [q_1 - -q_n]_{p\times n}$
  - Value $v \in \mathbb{R}^m$ $V = \underbrace{W_v^T}_{m\times d} X_{d\times n}$ $V = [\underline{v_1} - -v_n]_{m\times n}$

$$\underset{m\times n}{\mathcal{L}} = \underset{m\times n}{V}\ softmax\,(\underbrace{\frac{Q^T K}{\sqrt{P}}}_{n\times n})$$

Q^T: $n \times p$; K: $p \times n$

$O(n^2 m)$

Generalized definition

- Define three different vectors corresponding to each word.
  - Input $x \in \mathbb{R}^d$ $X = [\underline{x_1} - -x_n]_{d\times n}$
  - Key $k \in \mathbb{R}^p$ $K = \underbrace{W_k^T}_{p\times d} X_{d\times n}$ $K = [\underline{x_1} - -k_n]_{p\times n}$
  - Query $q \in \mathbb{R}^p$ $Q = \underbrace{W_q^T}_{p\times d} X_{d\times n}$ $Q = [q_1 - -q_n]_{p\times n}$
  - Value $v \in \mathbb{R}^m$ $V = \underbrace{W_v^T}_{m\times d} X_{d\times n}$ $V = [\underline{v_1} - -v_n]_{m\times n}$

$$\mathcal{L} = V\, softmax\left(\underbrace{\frac{Q^T K}{\sqrt{P}}}_{n\times n}\right)$$

$m \times n$ $m \times n$ $n \times p$ $p \times n$

$X^T \underbrace{W_q W_k} X$

$O(n^2 m)$

$K(x, y) = \phi(x)^T \phi(y)$

$\phi \quad x \rightarrow \phi(x)$

$K(x, y) = \phi(x)^T \phi(y)$

$\phi \quad x \rightarrow \phi(x)$

- most kernels can be approximated by random features

$K(x,y) = \phi(x)^T \phi(y)$

$\phi \quad x \rightarrow \phi(x)$

- most kernels can be approximated by random features
- Random features has this form:

$$K(x,y) = \phi(x)^T \phi(y)$$

$$\phi \quad x \rightarrow \phi(x)$$

- most kernels can be approximated by random features
- Random features has this form:

$$\phi(x) = \frac{h(x)}{\sqrt{r}} \left(f_1\left(\underline{\omega_1^T} x\right), f_1\left(\underline{\omega_2^T} x\right) \dots f_1\left(\underline{\omega_r^T} x\right) \dots f_l\left(\underline{\omega_1^T} x\right) \dots f_l\left(\underline{\omega_r^T} x\right)\right)$$

$K(x, y) = \phi(x)^T \phi(y)$

$\phi \quad x \rightarrow \phi(x)$

- most kernels can be approximated by random features
- Random features has this form:

$$\phi(x) = \underbrace{\frac{h(x)}{\sqrt{r}} \left(f_1\left(\underline{\omega_1^T} x\right), f_1\left(\underline{\omega_2^T} x\right) \ldots f_1\left(\underline{\omega_r^T} x\right) \ldots f_l\left(\underline{\omega_1^T} x\right) \ldots f_l\left(\underline{\omega_r^T} x\right)\right)}_{l \times r \text{ elements in vector } \phi}$$

$\phi(y)$

$K(x, y) = \phi^T(x)\phi(y)$

For example:

$h(x) = 1 \quad f_1 = \sin \quad f_2 = \cos \quad l = 2 \quad \underline{\omega} \sim N(O, I_l)$

For example:

$h(x) = 1 \quad f_1 = \sin \quad f_2 = \cos \quad l = 2 \quad \underline{\omega} \sim N(O, I_l)$

$$\phi(\underline{x}) = \frac{1}{\sqrt{r}}(\sin\left(\underline{\omega_1^T x}\right) \sin\left(\underline{\omega_2^T x}\right) \dots \sin\left(\underline{\omega_1^T x}\right) \cos\left(\underline{\omega_1^T x}\right) \dots \cos\left(\underline{\omega_r^T x}\right))$$

For example:

$h(x) = 1 \quad f_1 = \sin \quad f_2 = \cos \quad l = 2 \qquad \underline{\omega} \sim N(O, I_l)$

$$\phi(\underline{x}) = \frac{1}{\sqrt{r}}\left(\sin\left(\underline{\omega_1^T} x\right) \sin\left(\underline{\omega_2^T} x\right) \dots \sin\left(\underline{\omega_1^T} x\right) \cos\left(\underline{\omega_1^T} x\right) \dots \cos\left(\underline{\omega_r^T} x\right)\right)$$

$$K(x, y) = \phi^T \phi(y) = e^{\frac{-|x-y|^2}{\gamma}}$$

Gaussian

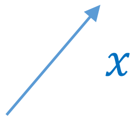
For example:

$$h(x) = 1 \quad f_1 = \sin \quad f_2 = \cos \quad l = 2 \qquad \underline{\omega} \sim N(O, I_l)$$

$$\phi(\underline{x}) = \frac{1}{\sqrt{r}}\left(\sin\left(\underline{\omega_1^T} x\right) \sin\left(\underline{\omega_2^T} x\right) \ldots \sin\left(\underline{\omega_1^T} x\right) \cos\left(\underline{\omega_1^T} x\right) \ldots \cos\left(\underline{\omega_r^T} x\right)\right)$$

$$K(x, y) = \phi^T \phi(y) = e^{\frac{-|x-y|^2}{\gamma}}$$

Gaussian



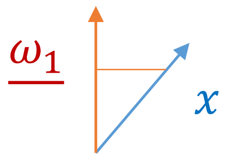
For example:

$$h(x) = 1 \quad f_1 = \sin \quad f_2 = \cos \quad l = 2 \qquad \underline{\omega} \sim N(O, I_l)$$

$$\phi(\underline{x}) = \frac{1}{\sqrt{r}}\left(\sin\left(\underline{\omega_1^T} x\right) \sin\left(\underline{\omega_2^T} x\right) \dots \sin\left(\underline{\omega_1^T} x\right) \cos\left(\underline{\omega_1^T} x\right) \dots \cos\left(\underline{\omega_r^T} x\right)\right)$$

$$K(x, y) = \phi^T \phi(y) = e^{\frac{-|x-y|^2}{\gamma}}$$

Gaussian



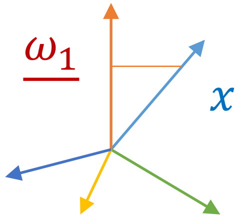
For example:

$$h(x) = 1 \quad f_1 = \sin \quad f_2 = \cos \quad l = 2 \qquad \underline{\omega} \sim N(O, I_l)$$

$$\phi(\underline{x}) = \frac{1}{\sqrt{r}}(\sin\left(\underline{\omega_1^T} x\right) \sin\left(\underline{\omega_2^T} x\right) \ldots \sin\left(\underline{\omega_1^T} x\right) \cos\left(\underline{\omega_1^T} x\right) \ldots \cos\left(\underline{\omega_r^T} x\right))$$

$$K(x, y) = \phi^T \phi(y) = e^{\frac{-|x-y|^2}{\gamma}}$$

Gaussian



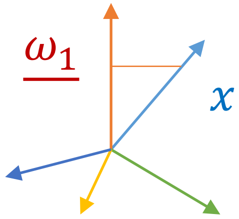
For example:

$h(x) = 1 \quad f_1 = \sin \quad f_2 = \cos \quad l = 2 \qquad \underline{\omega} \sim N(O, I_l)$

$$\phi(\underline{x}) = \frac{1}{\sqrt{r}}\left(\sin\left(\underline{\omega_1^T x}\right)\sin\left(\underline{\omega_2^T x}\right)\ldots\sin\left(\underline{\omega_1^T x}\right)\cos\left(\underline{\omega_1^T x}\right)\ldots\cos\left(\underline{\omega_r^T x}\right)\right)$$

$$K(x, y) = \phi^T \phi(y) = e^{\frac{-|x-y|^2}{\gamma}}$$

Gaussian



$$softmax\left(\frac{Q^T K}{\sqrt{P}}\right)$$

$$\sigma(\underline{s})_i = a_i = \frac{e^{s_i}}{\sum_j e^{s^j}}$$

$$\sigma(\underline{s})_i = a_i = \frac{e^{s_i}}{\sum_j e^{s_j}}$$

$$A = \exp(\frac{Q^T K}{\sqrt{P}})$$

$$\underbrace{\begin{bmatrix} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{bmatrix}_{n\times n}}_{A}$$

$$\sigma(\underline{s})_i = a_i = \frac{e^{s_i}}{\sum_j e^{s_j}}$$

$$A = \exp\left(\frac{Q^T K}{\sqrt{P}}\right)$$

$$\underbrace{\begin{bmatrix} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{bmatrix}_{n\times n}}_{A}$$

$$\underbrace{q^T_{l\times p} K_{p\times n}}_{l\times n}$$

$$\sigma(\underline{s})_i = a_i = \frac{e^{s_i}}{\sum_j e^{s^j}}$$

$$A = \exp(\frac{Q^T K}{\sqrt{P}})$$

$$\underbrace{\begin{bmatrix} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{bmatrix}_{n\times n}}_{A} \qquad \begin{matrix} 1 \\ \vdots \\ 1 \end{matrix}_{n\times 1} \qquad \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}_{n\times 1}$$

$$\underbrace{q^T_{l\times p} K_{p\times n}}_{l\times n}$$

$$\sigma(\underline{s})_i = a_i = \frac{e^{s_i}}{\sum_j e^{s^j}}$$

$$A = \exp(\frac{Q^T K}{\sqrt{P}})$$

$$\underbrace{\begin{bmatrix} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{bmatrix}_{n\times n}}_{A} \qquad \begin{matrix} 1 \\ \vdots \\ 1 \end{matrix}_{n\times 1} \qquad \begin{matrix} x \\ \vdots \\ x \end{matrix}_{n\times 1}$$

$$\underbrace{q^T_{l\times p} K_{p\times n}}_{l\times n}$$

$$diag\,\underbrace{(A\underline{1})}_{D} = \begin{bmatrix} x & 0 & 0 & 0 & - \\ 0 & x & 0 & 0 & - \\ 0 & 0 & x & 0 & - \\ \cdots & \cdots & \cdots & \cdots & \end{bmatrix}$$

$$\sigma(\underline{s})_i = a_i = \frac{e^{s_i}}{\sum_j e^{s^j}}$$

$$A = \exp(\frac{Q^T K}{\sqrt{P}})$$

$$\underbrace{\begin{bmatrix} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{bmatrix}_{n\times n}}_{A} \quad \begin{matrix} 1 \\ \vdots \\ 1 \end{matrix}\; n\times 1 \quad \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}\; n\times 1$$

$$\underbrace{q^T_{l\times p} K_{p\times n}}_{l\times n}$$

$$diag\,\underbrace{(A\underline{1})}_{D} = \begin{bmatrix} x & 0 & 0 & 0 & - \\ 0 & x & 0 & 0 & - \\ 0 & 0 & x & 0 & - \\ \cdots & \cdots & \cdots & \cdots & \end{bmatrix}$$

$$softmax\left(\frac{Q^T K}{\sqrt{P}}\right) = AD^{-1}$$

$$\sigma(\underline{s})_i = a_i = \frac{e^{s_i}}{\sum_j e^{s^j}}$$

$$A = \exp(\frac{Q^T K}{\sqrt{P}})$$

$$\underbrace{\begin{bmatrix} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{bmatrix}_{n\times n}}_{A} \quad \begin{matrix} 1 \\ \vdots \\ 1 \end{matrix}_{n\times 1} \quad \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}_{n\times 1}$$

$$\underbrace{q^T_{l\times p} K_{p\times n}}_{l\times n}$$

$$diag\,\underbrace{(A\underline{1})}_{D} = \begin{bmatrix} x & 0 & 0 & 0 & - \\ 0 & x & 0 & 0 & - \\ 0 & 0 & x & 0 & - \\ \cdots & \cdots & \cdots & \cdots & \end{bmatrix}$$

$$\begin{bmatrix} \frac{1}{x} & 0 & 0 & 0 & - \\ 0 & \frac{1}{x} & 0 & 0 & - \\ 0 & 0 & \frac{1}{x} & 0 & - \\ \cdots & \cdots & \cdots & \cdots & \end{bmatrix}$$

$$softmax\left(\frac{Q^T K}{\sqrt{P}}\right) = AD^{-1}$$

$$e^{-|x-y|^2} = e^{-(x-y)^T(x-y)} = e^{-[x^Tx+y^Ty-2x^Ty]} = e^{-x^Tx} \cdot e^{-y^Ty} \cdot e^{2x^Ty}$$

$$\underbrace{e^{-|x-y|^2}} = e^{\frac{-(x-y)^T(x-y)}{2}} = e^{\frac{-[x^T x+y^T y-2x^T y]}{2}} = e^{-x^T x} \cdot e^{-y^T y} \cdot \underbrace{e^{2x^T y}}$$

$$\underbrace{e^{-|x-y|^2}} = e^{\frac{-(x-y)^T(x-y)}{2}} = e^{\frac{-[x^Tx+y^Ty-2x^Ty]}{2}} = \underbrace{e^{\frac{-x^Tx}{2}} \cdot e^{\frac{-y^Ty}{2}} \cdot \underbrace{e^{\frac{\cancel{2}x^Ty}{\cancel{2}}}}}$$

$$\underbrace{e^{-|x-y|^2}} = e^{\frac{-(x-y)^T(x-y)}{2}} = e^{\frac{-[x^Tx+y^Ty-2x^Ty]}{2}} = \underbrace{e^{\frac{-x^Tx}{2}} \cdot e^{\frac{-y^Ty}{2}} \cdot \underbrace{e^{\frac{\cancel{2}x^Ty}{\cancel{2}}}}}$$

$$K_{gauss} \cdot e^{\frac{x^Tx}{2}} \cdot e^{\frac{y^Ty}{2}}$$

$$\underbrace{e^{-|x-y|^2}} = e^{\frac{-(x-y)^T(x-y)}{2}} = e^{\frac{-[x^Tx+y^Ty-2x^Ty]}{2}} = \underbrace{e^{\frac{-x^Tx}{2}} \cdot e^{\frac{-y^Ty}{2}} \cdot \underbrace{e^{\frac{\not{2}x^Ty}{\not{2}}}}}$$

$$\underbrace{K_{gauss} \cdot e^{\frac{x^Tx}{2}} \cdot e^{\frac{y^Ty}{2}}}_{K_{SM} \ \leftarrow e^{x^Ty}}$$

$$\underbrace{e^{-|x-y|^2}} = e^{\frac{-(x-y)^T(x-y)}{2}} = e^{\frac{-[x^Tx+y^Ty-2x^Ty]}{2}} = \underbrace{e^{\frac{-x^Tx}{2}} \cdot e^{\frac{-y^Ty}{2}} \cdot \underbrace{e^{\frac{\cancel{2}x^Ty}{\cancel{2}}}}}$$

$$\underbrace{K_{gauss} \cdot e^{\frac{x^Tx}{2}} \cdot e^{\frac{y^Ty}{2}}}_{K_{SM} \;\; \leftarrow e^{x^Ty}}$$

$$h(x) = \frac{x^Tx}{2}$$

$$e^{-|x-y|^2} = e^{\frac{-(x-y)^T(x-y)}{2}} = e^{\frac{-[x^T x + y^T y - 2x^T y]}{2}} = e^{\frac{-x^T x}{2}} \cdot e^{\frac{-y^T y}{2}} \cdot e^{\frac{2x^T y}{2}}$$

$$\underbrace{K_{gauss} \cdot e^{\frac{x^T x}{2}} \cdot e^{\frac{y^T y}{2}}}_{K_{SM} \;\leftarrow e^{x^T y}}$$

$$h(x) = \frac{x^T x}{2}$$

$$\phi(x) = \frac{h(x)}{\sqrt{r}} \left(f_1\left(\omega_1^T x\right), f_1\left(\omega_2^T x\right) \dots f_1\left(\omega_r^T x\right) \dots f_l\left(\omega_1^T x\right) \dots f_l\left(\omega_r^T x\right)\right)$$

$$\underbrace{e^{-|x-y|^2}} = e^{\frac{-(x-y)^T(x-y)}{2}} = e^{\frac{-[x^Tx+y^Ty-2x^Ty]}{2}} = \underbrace{e^{\frac{-x^Tx}{2}} \cdot e^{\frac{-y^Ty}{2}} \cdot \underbrace{e^{\frac{2x^Ty}{2}}}}$$

$$\underbrace{K_{gauss} \cdot e^{\frac{x^Tx}{2}} \cdot e^{\frac{y^Ty}{2}}}_{K_{SM} \;\leftarrow e^{x^Ty}}$$

$$h(x) = \frac{x^Tx}{2}$$

$$\phi(x) = \frac{h(x)}{\sqrt{r}}\left(f_1\left(\underline{\omega_1^T}x\right), f_1\left(\underline{\omega_2^T}x\right) \dots f_1\left(\underline{\omega_r^T}x\right) \dots f_l\left(\underline{\omega_1^T}x\right) \dots f_l\left(\underline{\omega_r^T}x\right)\right)$$

$$\phi(\underline{x}) = \frac{x^Tx}{2\sqrt{r}}\left(\sin\left(\underline{\omega_1^T}x\right) \sin\left(\underline{\omega_2^T}x\right) \dots \sin\left(\underline{\omega_1^T}x\right) \cos\left(\underline{\omega_1^T}x\right) \dots \cos\left(\underline{\omega_r^T}x\right)\right)$$

$$V\ softmax\left(\frac{Q^T K}{\sqrt{P}}\right)$$

$$\underbrace{V\ softmax\left(\frac{Q^T K}{\sqrt{P}}\right)}_{V\ Q'^T K'}$$

$$\underbrace{V\ softmax\left(\frac{Q^T K}{\sqrt{P}}\right)}_{V\ Q'^T K'}$$

$Q^T \rightarrow n \times p$

$K \rightarrow p \times n$

$$\underbrace{V\, softmax\left(\frac{Q^T K}{\sqrt{P}}\right)}_{V\, Q'^T K'}$$

$Q^T \rightarrow n \times p$

$K \rightarrow p \times n$

$O(mn^2)$

$$\underbrace{V\ softmax\left(\frac{Q^T K}{\sqrt{P}}\right)}_{V\ (Q'^T K')}$$

$n \times p$

$p \times n$

$O(mn^2)$

$$\underbrace{m \times n \quad n \times r'}_{m \times r'}$$

$r' \times n$

$r' \times n$

$O(mr'n^2)$

$$V \, softmax\left(\frac{Q^T K}{\sqrt{P}}\right)$$

$n \times p$

$p \times n$

$V\,(Q'^T K')$

$O(mn^2)$

$\underbrace{m \times n \quad n \times r'}_{m \times r'}$

$r' \times n$

$r' \times n$

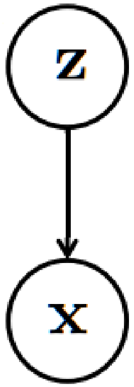
$O(mr'n^2)$

$$\begin{bmatrix} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{bmatrix}$$

Variational Auto encoder (VEA)

Variational Inference

- Problem Definition
  - Observable Data: $x = \{x_1, x_2, ..., x_n\}$
  - Hidden Variable: $z = \{z_1, z_2, ..., z_n\}$

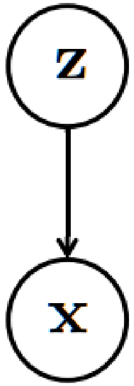


Variational Inference

- Problem Definition

  - Observable Data: $x = \{x_1, x_2, ..., x_n\}$

  - Hidden Variable: $z = \{z_1, z_2, ..., z_n\}$



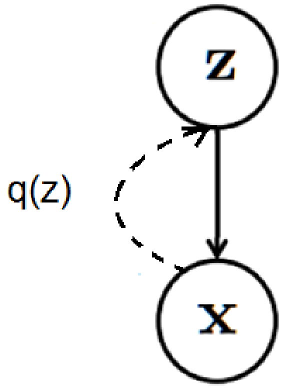
$$p(z|x) = \frac{p(z,x)}{p(x)} = \frac{p(x|z)p(z)}{\int p(x|z)p(z)dz}$$

Variational Inference

- Solutions
    - Monte Carlo Sampling
        - Metropolis Hasting
        - Gibbs Sampling
    - Variational Inference

Variational Inference

- Approximate $p(z|x)$ by $q(z)$



- Minimize the KL Divergence:

$$D_{KL}\left[q(z)||p(z|x)\right] = -\int q(z) log \frac{p(z|x)}{q(z)} dz$$

Variational Lower Bound

$$D_{KL}\left[q(z)||p(z|x)\right] = -\int q(z)log\frac{p(z|x)}{q(z)}dz$$

Variational Lower Bound

$$D_{KL}\left[q(z)||p(z|x)\right] = -\int q(z)log\frac{p(z|x)}{q(z)}dz$$

$$= -\int q(z)log\frac{p(z,x)}{q(z)p(x)}dz$$

Variational Lower Bound

$$D_{KL}\left[q(z)||p(z|x)\right] = -\int q(z)log\frac{p(z|x)}{q(z)}dz$$

$$= -\int q(z)log\frac{p(z,x)}{q(z)p(x)}dz$$

$$= -\int q(z)log\frac{p(z,x)}{q(z)}dz + \int q(z)log\big(p(x)\big)dz$$

Variational Lower Bound

$$D_{KL}\left[q(z)||p(z|x)\right] = -\int q(z)log\frac{p(z|x)}{q(z)}dz$$

$$= -\int q(z)log\frac{p(z,x)}{q(z)p(x)}dz$$

$$= -\int q(z)log\frac{p(z,x)}{q(z)}dz + \int q(z)log\left(p(x)\right)dz$$

$$= -\int q(z)\Big(log\left(p(z,x)\right) - log\left(q(z)\right)\Big)dz + log\left(p(x)\right)$$

Variational Lower Bound

$$D_{KL}\left[q(z)||p(z|x)\right] = -\int q(z) log \frac{p(z|x)}{q(z)} dz$$

$$= -\int q(z) log \frac{p(z,x)}{q(z)p(x)} dz$$

$$= -\int q(z) log \frac{p(z,x)}{q(z)} dz + \int q(z) log\big(p(x)\big) dz$$

$$= -\int q(z)\Big(log\big(p(z,x)\big) - log\big(q(z)\big)\Big) dz + log\big(p(x)\big)$$

$$= -\underbrace{\left(E_{q(z)}\left[log\big(p(z,x)\big)\right] - E_{q(z)}\left[log\big(q(z)\big)\right]\right)}_{\text{Evidence Lower Bound (ELBO)}} + log\big(p(x)\big)$$

Variational Lower Bound

$$D_{KL}\left[q(z)||p(z|x)\right] = -\underbrace{\left(E_{q(z)}\left[log\left(p(z,x)\right)\right] - E_{q(z)}\left[log\left(q(z)\right)\right]\right)}_{\text{Evidence Lower Bound (ELBO)}} + log\left(p(x)\right)$$

Variational Lower Bound

$$D_{KL}\Big[q(z)||p(z|x)\Big] = -\underbrace{\left(E_{q(z)}\Big[log\big(p(z,x)\big)\Big] - E_{q(z)}\Big[log\big(q(z)\big)\Big]\right)}_{\text{Evidence Lower Bound (ELBO)}} + log\big(p(x)\big)$$

$$D_{KL}\Big[q(z)||p(z|x)\Big] = -L\Big[q(z)\Big] + log\Big(p(x)\Big)$$

Variational Lower Bound

$$D_{KL}\Big[q(z)||p(z|x)\Big] = -\underbrace{\left(E_{q(z)}\Big[log\big(p(z,x)\big)\Big] - E_{q(z)}\Big[log\big(q(z)\big)\Big]\right)}_{\text{Evidence Lower Bound (ELBO)}} + log\big(p(x)\big)$$

$$D_{KL}\Big[q(z)||p(z|x)\Big] = -L\Big[q(z)\Big] + log\Big(p(x)\Big)$$

$$log\Big(p(x)\Big) = D_{KL}\Big[q(z)||p(z|x)\Big] + L\Big[q(z)\Big]$$

Variational Lower Bound

$$D_{KL}\Big[q(z)||p(z|x)\Big] = -\underbrace{\left(E_{q(z)}\Big[log\big(p(z,x)\big)\Big] - E_{q(z)}\Big[log\big(q(z)\big)\Big]\right)}_{\text{Evidence Lower Bound (ELBO)}} + log\big(p(x)\big)$$

$$D_{KL}\Big[q(z)||p(z|x)\Big] = -L\Big[q(z)\Big] + log\Big(p(x)\Big)$$

$$log\Big(p(x)\Big) = D_{KL}\Big[q(z)||p(z|x)\Big] + L\Big[q(z)\Big]$$

Minimizing $D_{KL}\Big[q(z)||p(z|x)\Big]$

is equal to Maximizing $L\Big[q(z)\Big]$