Variational Inference

Probabilistic Graphical Models

Jerónimo Hernández-González

Approximate inference
Alternatives

- Sampling:
  - Forward sampling
  - MCMC
  - Gibbs sampling
- Optimization:
  - Loopy Belief Propagation
  - Expectation Propagation
  - Variational approaches

Variational inference

Variational Inference: when?

- The objective is inference; to answer queries as
$$P(\boldsymbol{Y}|\boldsymbol{E} = \boldsymbol{e}) \qquad \text{for } \boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{E}, \boldsymbol{H})$$
- The joint $P(\boldsymbol{X})$ is too complex to use exact inference
- Sampling approaches: expensive? Known how to sample?

Variational Inference: what?

1. Take a family of distributions $\mathcal{Q}$ over $\boldsymbol{Y}$ where inference is **simple**

   What is it a family of distributions?

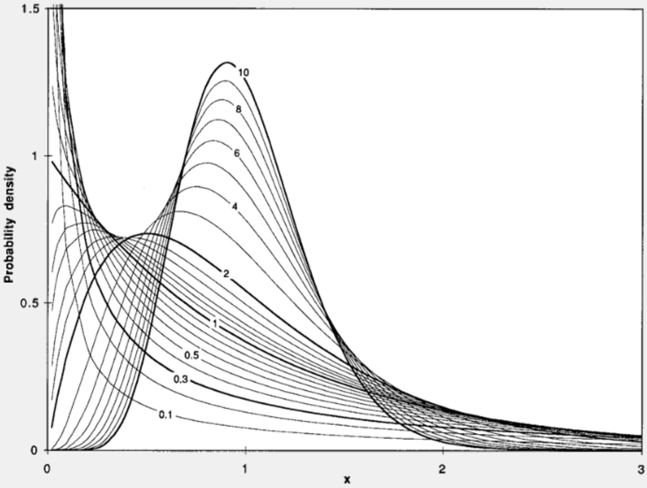
2. Find the distribution $Q(\boldsymbol{Y}) \in \mathcal{Q}$ which is **closest** to $P(\boldsymbol{Y}|\boldsymbol{E} = \boldsymbol{e})$

   Projection of $P(\boldsymbol{Y}|\boldsymbol{E} = \boldsymbol{e})$

3. Use $Q(\boldsymbol{Y})$ instead of $P(\boldsymbol{Y}|\boldsymbol{E} = \boldsymbol{e})$ to approximate the answer

Family of distributions

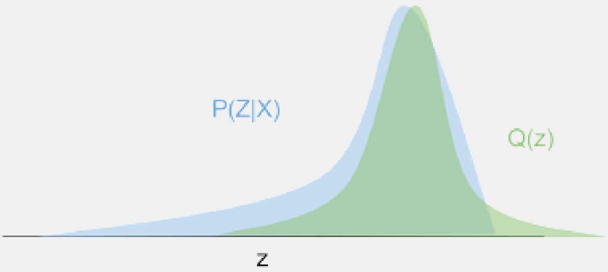
Family of Gamma distributions



Background

Family of distributions

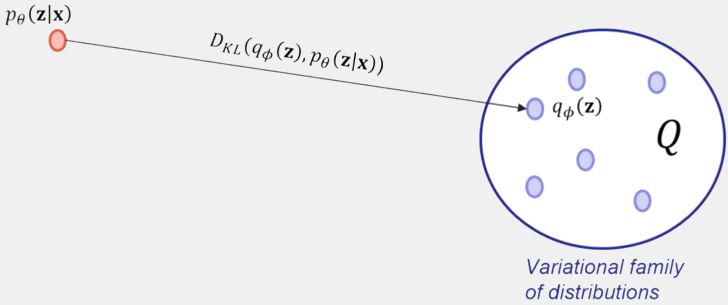
The objective is to just perform inference on an easy, parametric distribution $Q(\boldsymbol{Y};\phi)$,
We use a specific set of parameters Φ so that Q is as close as possible to $P(\boldsymbol{Y}|\boldsymbol{E}=\boldsymbol{e})$



Background

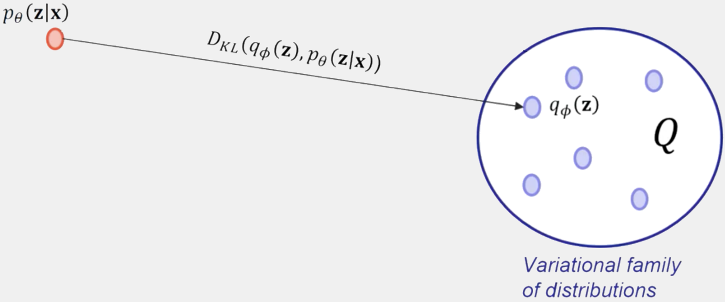
Projection

Given a point p_θ (specific distribution) and a set $\mathcal{Q}$ (family of distributions), find the point $q_\phi \in \mathcal{Q}$ that is closest to p_θ



Projection (II): An optimization problem

Which is the $q_\phi \in \mathcal{Q}$ (a distribution in the family) that minimizes the distance to a given point p_θ (specific distribution)?



Background

Distance: **Kullback-Liebler divergence** or Relative entropy

$$
\begin{aligned}
KL(P||Q) &= \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)} \\
&= \mathbb{E}_P [\log P(x)] - \mathbb{E}_P [\log Q(x)]
\end{aligned}
$$

Properties:

- Positivity. $KL(P||Q) \geq 0$, and $KL(P||Q) = 0 \rightarrow P = Q$
- No symmetry: $KL(P||Q) \neq KL(Q||P)$
- No triangle inequality
- If $P(X, Y)$ and $Q(X, Y)$ satisfy that $X \perp\!\!\!\perp Y$ then

  $$KL(P(X, Y)||Q(X, Y)) = KL(P(X)||Q(X)) + KL(P(Y)||Q(Y))$$

Variational inference (revisited)

Variational Inference: what?

1. Take a family of distributions $\mathcal{Q}$ over $\boldsymbol{Y}$ where inference is **simple**

   What is it a family of distributions?

2. Find the distribution $Q(\boldsymbol{Y}) \in \mathcal{Q}$ which is **closest** to $P(\boldsymbol{Y}|\boldsymbol{E} = \boldsymbol{e})$
   That is, find Q^* as the one that minimizes the distance of P to $\mathcal{Q}$:

$$Q^* = \arg \min_{Q \in \mathcal{Q}} KL(Q||P).$$

3. Use $Q(\boldsymbol{Y})$ instead of $P(\boldsymbol{Y}|\boldsymbol{E} = \boldsymbol{e})$ to approximate the answer

Variational inference

The Evidence Lower BOund (ELBO)

$$Q^*(\boldsymbol{Y}) = \arg \min_{Q(\boldsymbol{Y})\in\mathcal{Q}} KL(Q(\boldsymbol{Y})||P(\boldsymbol{Y}|\boldsymbol{e}))$$

$$\begin{aligned} KL(Q(\boldsymbol{Y})||P(\boldsymbol{Y}|\boldsymbol{e})) &= \mathbb{E}[\log Q(\boldsymbol{Y})] - \mathbb{E}[\log P(\boldsymbol{Y}|\boldsymbol{e})] \\ &= \mathbb{E}[\log Q(\boldsymbol{Y})] - \mathbb{E}[\log P(\boldsymbol{Y},\boldsymbol{e})] + \log P(\boldsymbol{e}) \\ &= \log P(\boldsymbol{e}) - ELBO(Q) \end{aligned}$$

$$\begin{aligned} ELBO(Q) &= \mathbb{E}[\log P(\boldsymbol{Y},\boldsymbol{e})] - \mathbb{E}[\log Q(\boldsymbol{Y})] \\ &= \mathbb{E}[\log P(\boldsymbol{e})] + \mathbb{E}[\log P(\boldsymbol{Y}|\boldsymbol{e})] - \mathbb{E}[\log Q(\boldsymbol{Y})] \end{aligned}$$

Instead of minimizing $KL(Q(\boldsymbol{Y})||P(\boldsymbol{Y}|\boldsymbol{e}))$,

we look for the Q (i.e., its parameters Φ) that maximizes the evidence lower bound, $ELBO(Q;\Phi)$

Mean-field algorithm

The simplest variational approximation

- Use the family $\mathcal{Q}$ of distributions that factorize over the variables independently:

$$Q(\boldsymbol{X}) = \prod_{i=1}^{v} Q(X_i)$$

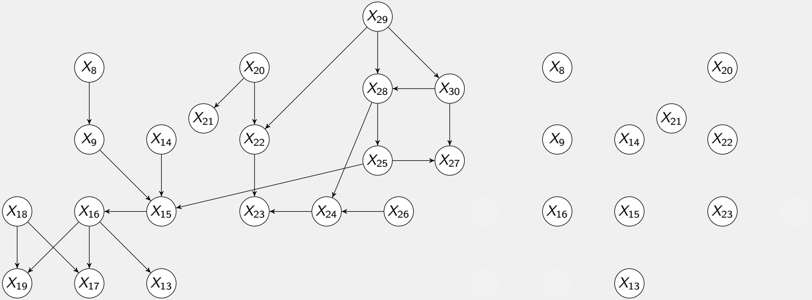
- Find the distribution $Q(\boldsymbol{Y}) \in \mathcal{Q}$ **closest** to $P(\boldsymbol{Y}|\boldsymbol{E} = \boldsymbol{e})$
  That is, find Q^* as the I-projection of P onto $\mathcal{Q}$:

$$Q^* = \arg \min_{Q \in \mathcal{Q}} KL(Q||P).$$

Mean-field algorithm

Family of fully factorized distributions

$$P(X_8, X_9, X_{14}, X_{13}, X_{15}, X_{16}, X_{20}, X_{21}, X_{22}, X_{23} | X_{17}, X_{29}, X_{28}, X_{27})$$



$$Q(\boldsymbol{Y}) = Q_8(X_8) \times Q_9(X_9) \times Q_{14}(X_{14}) \times Q_{13}(X_{13}) \times Q_{15}(X_{15}) \times Q_{16}(X_{16}) \times Q_{20}(X_{20}) \times Q_{21}(X_{21}) \times Q_{22}(X_{22}) \times Q_{23}(X_{23})$$

Mean-field algorithm

A coordinate-descent like algorithm

How to do the update of each $Q(Y_i)$?

The update in coordinate descent:

$$x_i \leftarrow \arg\min_{x_i'} f(x_i', x_{-i})$$

translates to:

$$\begin{aligned} Q(Y_i) &= \arg\min_{Q'(Y_i)} KL\Big(Q'(Y_i)Q(\boldsymbol{Y}_{-i})||P\Big) \\ &= \arg\min_{Q'(Y_i)} KL\Big(Q'(Y_i)\prod_{j\neq i} Q(Y_j)||P\Big) \end{aligned}$$

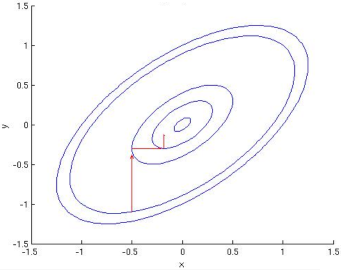
Background

Coordinate-descent

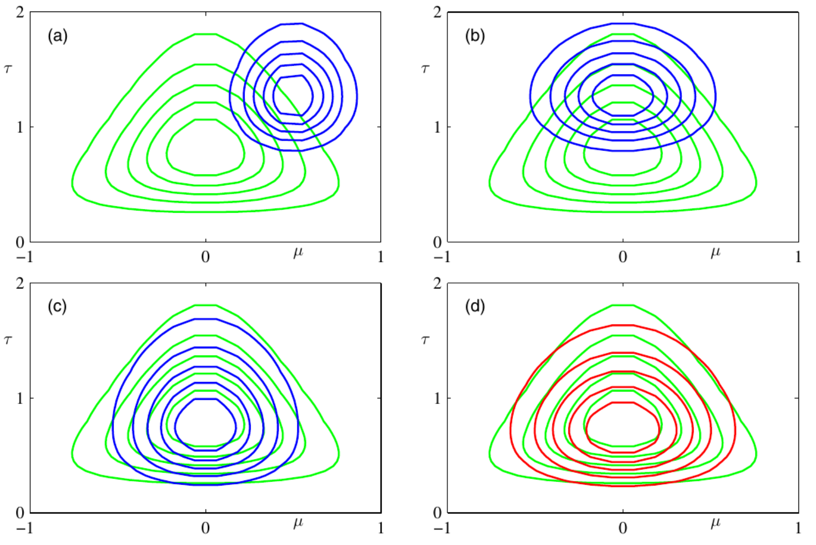
At each step, move one coordinate to the **minimum** while keeping the other coordinates fixed

f is a function from $\mathbb{R}^N \rightarrow \mathbb{R}$

function CoordinateDescent(f, x^0)
  $x \leftarrow x^0$
  while Not Converged **do**
    for $n \in \{1, \ldots, N\}$ **do** ▷ Sequentially!
      $x_i \leftarrow \arg\min_{x_i'} f(x_i', x_{-i})$
    end for
  end while
end function



Coordinate-descent



Mean-field algorithm

A coordinate-descent like algorithm (II)

How to do the update of each $Q(Y_i)$?

For MF, the objective is:

$$\arg\min_{Q'(Y_i)} KL\Big(Q'(Y_i)Q(\boldsymbol{Y}_{-i})||P\Big)$$

It can be seen that the ELBO decomposes

$$\log P(\boldsymbol{e}) + \sum_i \mathbb{E}[\log P(y_i|\boldsymbol{y}_{1:i-1}, \boldsymbol{e})] - \mathbb{E}_i[\log Q(y_i)]$$

and we can use that for each variable y_i:

$$\mathbb{E}[\log P(y_i|\boldsymbol{y}_{-i}, \boldsymbol{e})] - \mathbb{E}_i[\log Q(y_i)] + const.$$

To obtain, after some maths:

$$\hat{Q}_i(y_i) \propto \exp\left(\mathbb{E}_{Q_{-i}}\left[\log P(y_i, \boldsymbol{Y}_{-i}, \boldsymbol{e})\right]\right)$$

Mean-field algorithm

A coordinate-descent like algorithm (IV)

```
function GeneralMeanFieldApproximation(P)
    for all Q_i do
        Init(Q_i)
    end for
    while Not Converged do
        for all Q_i do                                          ▷ Sequentially!
            Q̃_i ← exp(E_{Q_{-i}}[log P(Y_i, Y_{-i}, X)])
            Q_i ← Normalize(Q̃_i)
        end for
    end while
end function
```

$$\tilde{Q}_i \leftarrow \exp\left(\mathbb{E}_{Q_{-i}}\left[\log P(Y_i, \boldsymbol{Y}_{-i}, \boldsymbol{X})\right]\right)$$

$$Q_i \leftarrow \text{Normalize}(\tilde{Q}_i)$$

Mean-field algorithm

$Q(Y_i)$ update when P factorizes

If P factorizes s.t. $P(\boldsymbol{X}) = \frac{1}{Z} \prod_{\phi \in \Phi} \phi(\boldsymbol{X}_\phi)$

Only the factors ϕ that contain the variable Y_i can be considered:

$$\hat{Q}_i(y_i) \propto \exp\left(\sum_{\substack{\phi \in \Phi : \\ Y_i \in d(\phi)}} \mathbb{E}_{Q_{d(\phi)\setminus y_i}} \left[\log \phi(y_i, X_{d(\phi)\setminus y_i}) \right] \right)$$

where $d(\phi)$ is the scope of the factor ϕ, i.e., the variables in ϕ.

Summary

- **Main idea**: Find in a family of probability distributions the best member to supplant a probability distribution of interest
- That best member is used instead of the probability distribution of interest to answer the queries

  Inference is approximated as these two might be (slightly) different
- We might want a family the members of which are simple to use in practice
- Inference as an optimization problem, usually done by coordinate ascent
- It requires mathematical derivations for each application of VI

  Really? Black-box VI

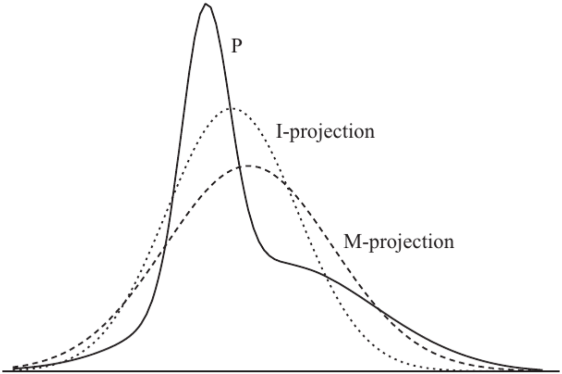
Reverse or forward KL

Let P be a distribution and $\mathcal{Q}$ be a family of distributions.

- Forwards KL:
$$Q^* = \arg\min_{Q \in \mathcal{Q}} KL(P||Q).$$
a.k.a. M-projection or moment projection
Infinite if $Q(x) = 0$ and $P(x) > 0$. Thus, if $P(x) > 0$, look for Q s.t. $Q(x) > 0$
- Reverse KL:
$$Q^* = \arg\min_{Q \in \mathcal{Q}} KL(Q||P).$$
a.k.a. I-projection or information projection
Infinite if $P(x) = 0$ and $Q(x) > 0$. Thus, if $P(x) = 0$, look for Q s.t. $Q(x) = 0$

Variational Inference

Understanding I-projections and M-projections



Forwards KL leads to M-projection

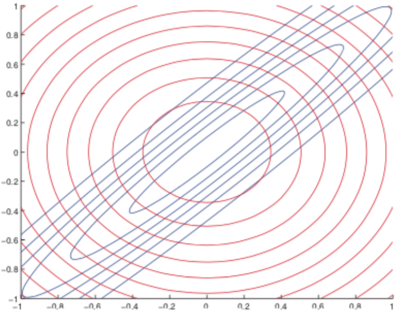
$$KL(P||Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}$$

Reverse KL leads to I-projection

$$KL(Q||P) = \sum_{x \in X} Q(x) \log \frac{Q(x)}{P(x)}$$

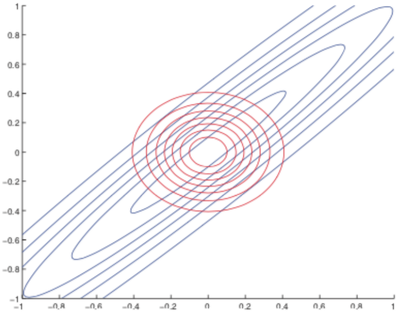
Variational Inference

Understanding I-projections and M-projections



Forwards KL leads to M-projection

$$KL(P||Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}$$

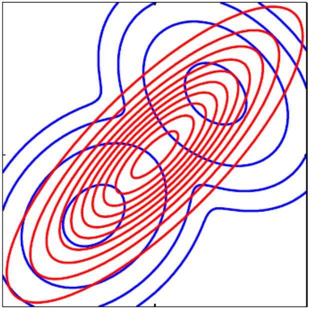


Reverse KL leads to I-projection

$$KL(Q||P) = \sum_{x \in X} Q(x) \log \frac{Q(x)}{P(x)}$$

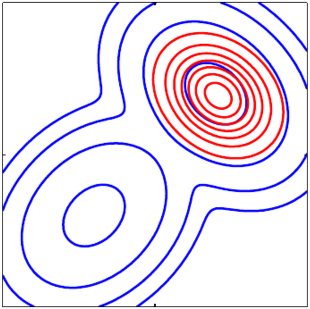
Variational Inference

Understanding I-projections and M-projections



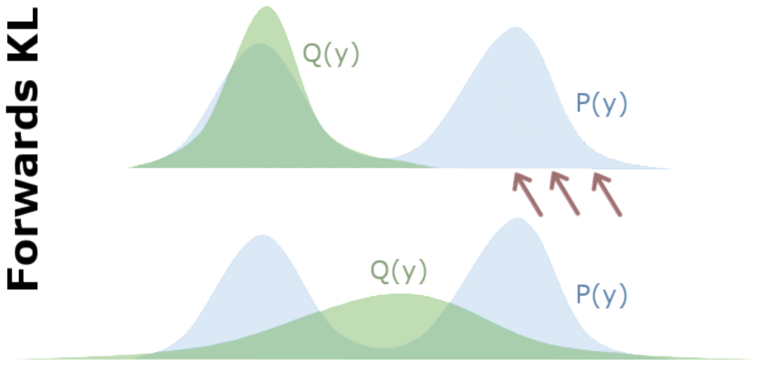
Forwards KL leads to M-projection

$$KL(P||Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}$$

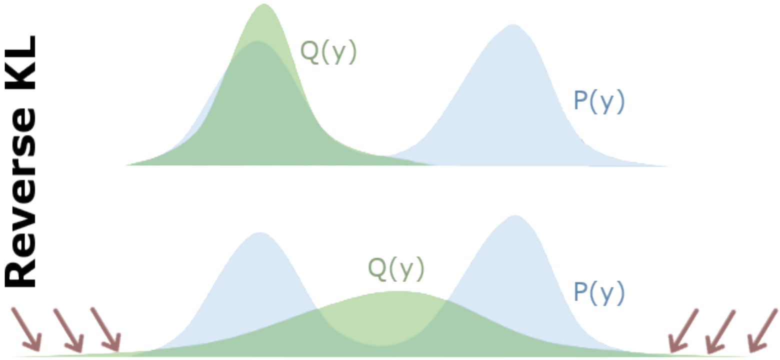


Reverse KL leads to I-projection

$$KL(Q||P) = \sum_{x \in X} Q(x) \log \frac{Q(x)}{P(x)}$$



M-projection: $\arg\min_Q KL(P||Q) = \arg\min_Q \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}$



I-projection: $\arg \min_Q KL(Q||P) = \arg \min_Q \sum_{x \in X} Q(x) \log \frac{Q(x)}{P(x)}$

Variational inference

To keep deepening

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). *Variational Inference: A Review for Statisticians.* Journal of the American Statistical Association, 112(518), 859–877

Fox, C. W., & Roberts, S. J. (2011). *A tutorial on variational Bayesian inference.* Artificial Intelligence Review, 38(2), 85–95

Variational Inference

Probabilistic Graphical Models

Jerónimo Hernández-González