A beginner's guide to variational inference

Haziq Jamil

Social Statistics
London School of Economics and Political Science

1 February 2018

Social Statistics Meeting

http://socialstats.haziqj.ml

Outline

Introduction

- Consider a statistical model where we have observations $\mathbf{y} = (y_1, \dots, y_n)$ and also some latent variables $\mathbf{z} = (z_1, \dots, z_m)$.

[1] With some caveats which will be discussed.

Introduction

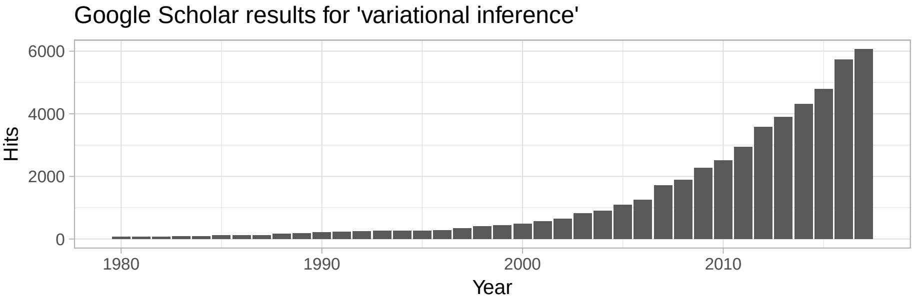
- Consider a statistical model where we have observations $\mathbf{y} = (y_1, \dots, y_n)$ and also some latent variables $\mathbf{z} = (z_1, \dots, z_m)$.
- Want to evaluate the intractable integral

$$\mathcal{I} := \int p(\mathbf{y}|\mathbf{z})p(\mathbf{z}) \, \mathrm{d}\mathbf{z}$$

  - Bayesian posterior analysis
  - Random effects models
  - Mixture models
- Variational inference approximates the "posterior" by a tractably close distribution in the KL sense.
- Advantages:
  - Computationally tractable
  - Convergence easily assessed
  - Works well in practice[1]

[1] With some caveats which will be discussed.

In the literature

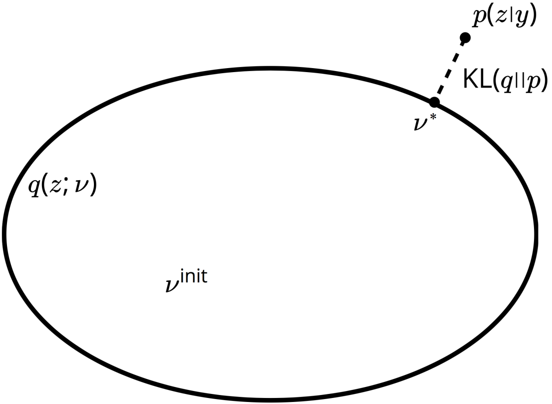


- Well known in the machine learning community.
- In social statistics:
  - E. A. Erosheva et al. (2007). "Describing disability through individual-level mixture models for multivariate binary data". *Ann. Appl. Stat*, 1.2, p. 346
  - J. Grimmer (2010). "An introduction to Bayesian inference via variational approximations". *Political Analysis* 19.1, pp. 32–47
  - Y. S. Wang et al. (2017). "A Variational EM Method for Mixed Membership Models with Multivariate Rank Data: an Analysis of Public Policy Preferences". arXiv: 1512.08731

Recommended texts

- M. J. Beal and Z. Ghahramani (2003). "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures". In: *Bayesian Statistics 7. Proceedings of the Seventh Valencia International Meeting*. Ed. by J. M. Bernardo et al. Oxford: Oxford University Press, pp. 453–464
- C. M. Bishop (2006). *Pattern Recognition and Machine Learning*. Springer
- K. P. Murphy (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press
- D. M. Blei et al. (2017). "Variational inference: A review for statisticians". *J. Am. Stat. Assoc,* to appear

Idea



- Minimise Kullbeck-Leibler divergence using calculus of variations

$$\mathrm{KL}(q\|p) = -\int \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z})} q(\mathbf{z}) \, \mathrm{d}\mathbf{z}$$

- **ISSUE**: KL$(q\|p)$ is intractable.

D. M. Blei (2017). "Variational Inference: Foundations and Innovations". URL: https://simons.berkeley.edu/talks/david-blei-2017-5-1

The Evidence Lower Bound (ELBO)

- Let $q(\mathbf{z})$ be some density function to approximate $p(\mathbf{z}|\mathbf{y})$.

The Evidence Lower Bound (ELBO)

- Let $q(\mathbf{z})$ be some density function to approximate $p(\mathbf{z}|\mathbf{y})$. Then the log-marginal density can be decomposed as follows:

$$\log p(\mathbf{y}) = \log p(\mathbf{y}, \mathbf{z}) - \log p(\mathbf{z}|\mathbf{y})$$

The Evidence Lower Bound (ELBO)

- Let $q(\mathbf{z})$ be some density function to approximate $p(\mathbf{z}|\mathbf{y})$. Then the log-marginal density can be decomposed as follows:

$$\begin{aligned}\log p(\mathbf{y}) &= \log p(\mathbf{y}, \mathbf{z}) - \log p(\mathbf{z}|\mathbf{y}) \\ &= \int \left\{ \log \frac{p(\mathbf{y}, \mathbf{z})}{q(\mathbf{z})} - \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z})} \right\} q(\mathbf{z}) \, d\mathbf{z} \\ &= \mathcal{L}(q) + \mathrm{KL}(q \| p) \\ &\geq \mathcal{L}(q)\end{aligned}$$

The Evidence Lower Bound (ELBO)

- Let $q(\mathbf{z})$ be some density function to approximate $p(\mathbf{z}|\mathbf{y})$. Then the log-marginal density can be decomposed as follows:

$$\begin{aligned}\log p(\mathbf{y}) &= \log p(\mathbf{y},\mathbf{z}) - \log p(\mathbf{z}|\mathbf{y}) \\ &= \int \left\{ \log \frac{p(\mathbf{y},\mathbf{z})}{q(\mathbf{z})} - \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z})} \right\} q(\mathbf{z}) \, d\mathbf{z} \\ &= \mathcal{L}(q) + \mathrm{KL}(q\|p) \\ &\geq \mathcal{L}(q)\end{aligned}$$

- $\mathcal{L}$ is referred to as the "lower-bound", and it serves as a surrogate function to the marginal.
- Maximising $\mathcal{L}(q)$ is equivalent to minimising $\mathrm{KL}(q\|p)$.
- **ISSUE**: $\mathcal{L}(q)$ is (generally) not convex.

Comparison to the EM algorithm

- Suppose for this part, the marginal density $p(\mathbf{y}|\theta)$ depends on parameters θ.
- In the EM algorithm, the true posterior density is used, i.e. $q(\mathbf{z}) \equiv p(\mathbf{z}|\mathbf{y},\theta)$.
- Thus,

$$\begin{aligned}
\log p(\mathbf{y}|\theta) &= \int \left\{ \log \frac{p(\mathbf{y},\mathbf{z}|\theta)}{p(\mathbf{z}|\mathbf{y},\theta)} - \cancel{\log \frac{p(\mathbf{z}|\mathbf{y},\theta)}{p(\mathbf{z}|\mathbf{y},\theta)}} \right\} p(\mathbf{z}|\mathbf{y},\theta^{(t)}) \, \mathrm{d}\mathbf{z} \\
&= \mathrm{E}_{\theta^{(t)}}[\log p(\mathbf{y},\mathbf{z}|\theta)] - \mathrm{E}_{\theta^{(t)}}[\log p(\mathbf{z}|\mathbf{y},\theta)] \\
&= Q(\theta|\theta^{(t)}) + \text{entropy}.
\end{aligned}$$

- Minimising the KL divergence corresponds to the E-step.
- For any θ,

$$\begin{aligned}
\log p(\mathbf{y}|\theta) - \log p(\mathbf{y}|\theta^{(t)}) &= Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) + \Delta\text{entropy} \\
&\geq Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}).
\end{aligned}$$

Factorised distributions (Mean-field theory)

- Maximising $\mathcal{L}$ over all possible q not feasible. Need some restrictions, but only to achieve tractability.
- Suppose we partition elements of $\mathbf{z}$ into M disjoint groups $\mathbf{z} = (\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(M)})$, and assume

$$q(\mathbf{z}) = \prod_{j=1}^{M} q_j(\mathbf{z}^{(j)}).$$

Factorised distributions (Mean-field theory)

- Maximising $\mathcal{L}$ over all possible q not feasible. Need some restrictions, but only to achieve tractability.
- Suppose we partition elements of $\mathbf{z}$ into M disjoint groups $\mathbf{z} = (\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(M)})$, and assume

$$q(\mathbf{z}) = \prod_{j=1}^{M} q_j(\mathbf{z}^{(j)}).$$

- Under this restriction, the solution to $\arg\max_q \mathcal{L}(q)$ is

$$\tilde{q}_j(\mathbf{z}^{(j)}) \propto \exp\big(\,\mathrm{E}_{-j}[\log p(\mathbf{y}, \mathbf{z})]\big) \tag{1}$$

for $j \in \{1, \dots, m\}$.

Factorised distributions (Mean-field theory)

- Maximising $\mathcal{L}$ over all possible q not feasible. Need some restrictions, but only to achieve tractability.
- Suppose we partition elements of $\mathbf{z}$ into M disjoint groups $\mathbf{z} = (\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(M)})$, and assume

$$q(\mathbf{z}) = \prod_{j=1}^{M} q_j(\mathbf{z}^{(j)}).$$

- Under this restriction, the solution to $\arg\max_q \mathcal{L}(q)$ is

$$\tilde{q}_j(\mathbf{z}^{(j)}) \propto \exp\left(\mathrm{E}_{-j}[\log p(\mathbf{y}, \mathbf{z})]\right) \tag{1}$$

  for $j \in \{1, \dots, m\}$.
- In practice, these unnormalised densities are of recognisable form (especially if conjugate priors are used).

Coordinate ascent mean-field variational inference (CAVI)

- The optimal distributions are coupled with another, i.e. each $\tilde{q}_j(\mathbf{z}^{(j)})$ depends on the optimal moments of $\mathbf{z}^{(k)}$, $k \in \{1, \dots, M : k \neq j\}$.
- One way around this to employ an iterative procedure.
- Assess convergence by monitoring the lower bound

$$\mathcal{L}(q) = \mathrm{E}_q[\log p(\mathbf{y}, \mathbf{z})] - \mathrm{E}_q[\log q(\mathbf{z})].$$

Algorithm 1 CAVI

initialise Variational factors $q_j(\mathbf{z}^{(j)})$
while $\mathcal{L}(q)$ not converged **do**
    for $j = 1, \dots, M$ **do**
        $\log q_j(\mathbf{z}^{(j)}) \leftarrow \mathrm{E}_{-j}[\log p(\mathbf{y}, \mathbf{z})] + \text{const.}$
    end for
    $\mathcal{L}(q) \leftarrow \mathrm{E}_q[\log p(\mathbf{y}, \mathbf{z})] - \mathrm{E}_q[\log q(\mathbf{z})]$
end while
return $\tilde{q}(\mathbf{z}) = \prod_{j=1}^{M} \tilde{q}_j(\mathbf{z}^{(j)})$

Exponential families

- For the mean-field variational method, suppose that each complete conditional is in the exponential family:

$$p(\mathbf{z}^{(j)}|\mathbf{z}_{-j},\mathbf{y}) = h(\mathbf{z}^{(j)}) \exp\big(\eta_j(\mathbf{z}_{-j},\mathbf{y}) \cdot \mathbf{z}^{(j)} - A(\eta_j)\big).$$

- Then, from (1),

$$\begin{aligned} \tilde{q}_j(\mathbf{z}^{(j)}) &\propto \exp\big(\operatorname{E}_{-j}[\log p(\mathbf{z}^{(j)}|\mathbf{z}_{-j},\mathbf{y})]\big) \\ &= \exp\Big(\log h(\mathbf{z}^{(j)}) + \operatorname{E}[\eta_j(\mathbf{z}_{-j},\mathbf{y})] \cdot \mathbf{z}^{(j)} - \operatorname{E}[A(\eta_j)]\Big) \\ &\propto h(\mathbf{z}^{(j)}) \exp\Big(\operatorname{E}[\eta_j(\mathbf{z}_{-j},\mathbf{y})] \cdot \mathbf{z}^{(j)}\Big) \end{aligned}$$

  is also in the same exponential family.
- C.f. Gibbs conditional densities.
- **ISSUE**: What if not in exponential family? Try importance sampling or Metropolis sampling.

Zero-forcing vs Zero-avoiding

- Back to the KL divergence:

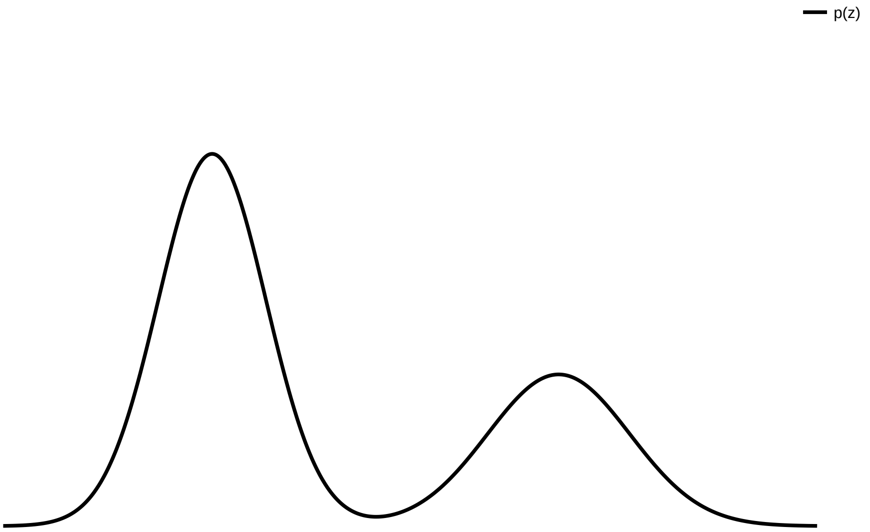
$$\mathrm{KL}(q\|p) = \int \log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{y})} q(\mathbf{z}) \, \mathrm{d}\mathbf{z}$$

- $\mathrm{KL}(q\|p)$ is large when $p(\mathbf{z}|\mathbf{y})$ is close to zero, unless $q(\mathbf{z})$ is also close to zero (*zero-forcing*).
- **ISSUE**: What about other measures of closeness? For instance,

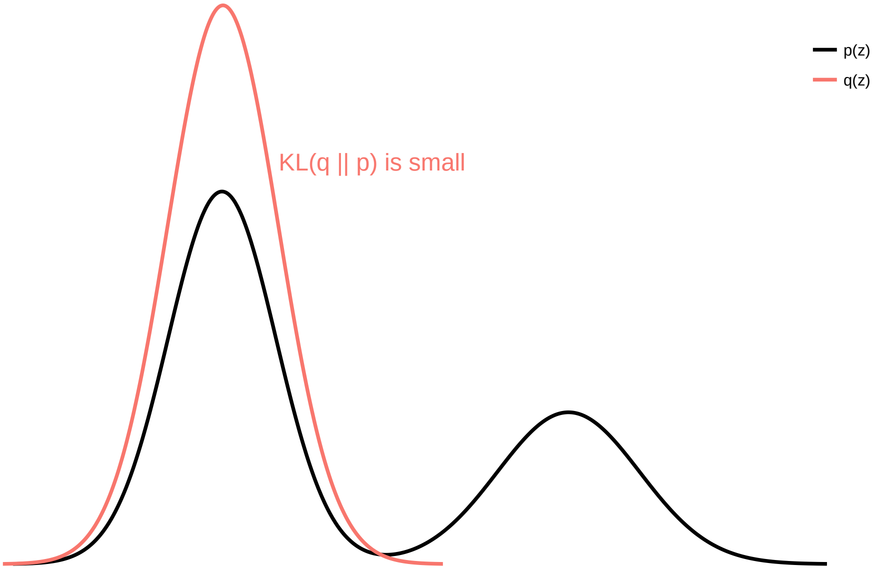
$$\mathrm{KL}(p\|q) = \int \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z}|\mathbf{y})} p(\mathbf{z}|\mathbf{y}) \, \mathrm{d}\mathbf{z}.$$

- This gives the Expectation Propagation (EP) algorithm.
- It is *zero-avoiding*, because $\mathrm{KL}(p\|q)$ is small when both $p(\mathbf{z}|\mathbf{y})$ and $q(\mathbf{z})$ are non-zero.

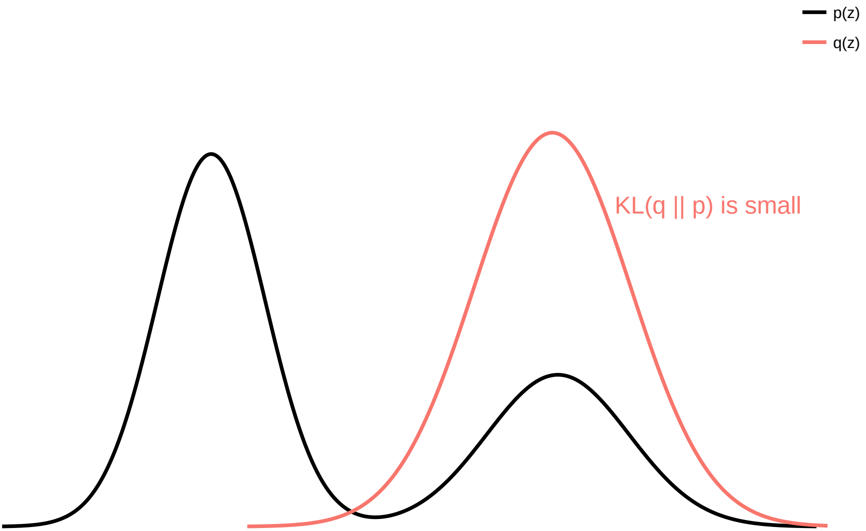
Zero-forcing vs Zero-avoiding (cont.)



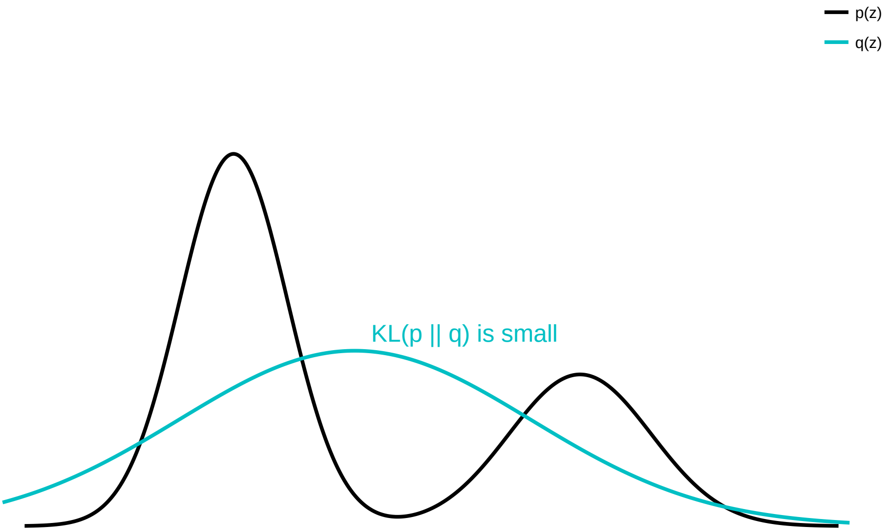
Zero-forcing vs Zero-avoiding (cont.)



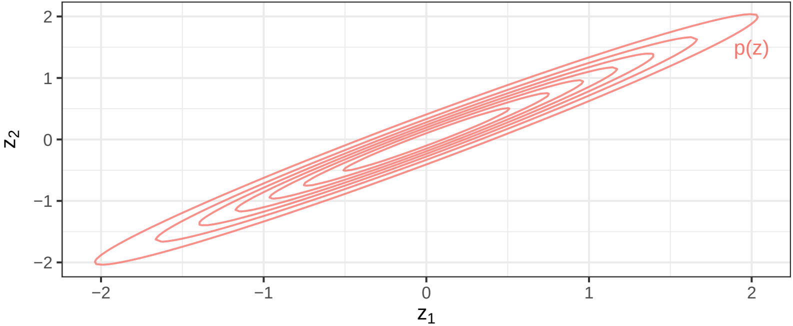
Zero-forcing vs Zero-avoiding (cont.)



Zero-forcing vs Zero-avoiding (cont.)

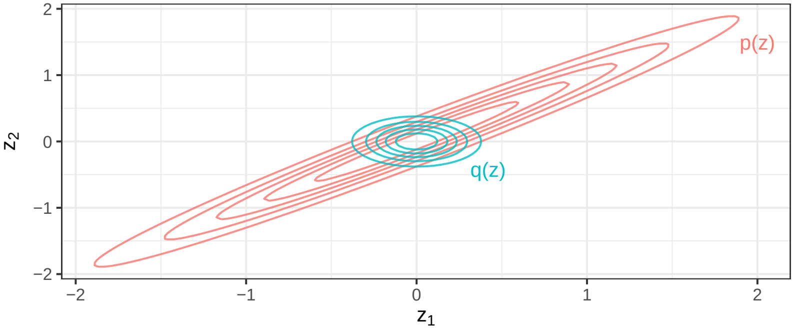


Distortion of higher order moments



- Consider $\mathbf{z} = (z_1, z_2)^\top \sim \mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Psi}^{-1})$, $\mathrm{Cov}(z_1, z_2) \neq 0$.

Distortion of higher order moments

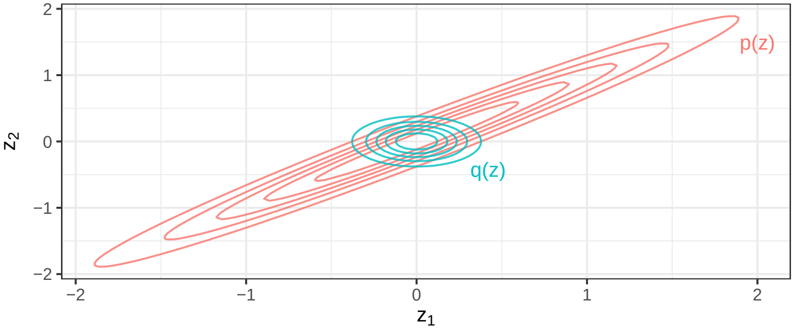


- Consider $\mathbf{z} = (z_1, z_2)^\top \sim \mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Psi}^{-1})$, $\mathrm{Cov}(z_1, z_2) \neq 0$.
- Approximating $p(\mathbf{z})$ by $q(\mathbf{z}) = q(z_1)q(z_2)$ yields

$$\tilde{q}(z_1) = \mathrm{N}(z_1 | \mu_1, \boldsymbol{\Psi}_{11}^{-1}) \quad \text{and} \quad \tilde{q}(z_2) = \mathrm{N}(z_2 | \mu_2, \boldsymbol{\Psi}_{22}^{-1})$$

and by definition, $\mathrm{Cov}(z_1, z_2) = 0$ under $\tilde{q}$.

Distortion of higher order moments



- Consider $\mathbf{z} = (z_1, z_2)^\top \sim \mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Psi}^{-1})$, $\mathrm{Cov}(z_1, z_2) \neq 0$.
- Approximating $p(\mathbf{z})$ by $q(\mathbf{z}) = q(z_1)q(z_2)$ yields

$$\tilde{q}(z_1) = \mathrm{N}(z_1 | \mu_1, \boldsymbol{\Psi}_{11}^{-1}) \quad \text{and} \quad \tilde{q}(z_2) = \mathrm{N}(z_2 | \mu_2, \boldsymbol{\Psi}_{22}^{-1})$$

  and by definition, $\mathrm{Cov}(z_1, z_2) = 0$ under $\tilde{q}$.
- This leads to underestimation of variances (widely reported in the literature—Zhao and Marriott 2013).

Non-convexity of ELBO

Means that it converges to a local optima rather than the global optima.
Show multiple restarts yield different ELBO.

Quality of approximation

- Variational inference converges to a different optimum than ML, except for certain models (Gunawardana and Byrne 2005).
- But not much can be said about the quality of approximation.
- Statistical properties not well understood—what is its statistical profile relative to the exact posterior?
- Speed though

Advanced topics

- Local variational bounds
  - Not using the mean-field assumption.
  - Instead, find a bound for the marginalising integral $\mathcal{I}$.
  - Used for Bayesian logistic regression as follows:

$$\mathcal{I} = \int \text{logit}(x^\top \beta) p(\beta) \, \text{d}\beta \geq \int f(x^\top \beta, \xi) p(\beta) \, \text{d}\beta.$$

- Stochastic variational inference
  - VI on its own doesn't offer much computational advantages.
  - Use ideas from stochastic optimisation—gradient based improvement of ELBO from subsamples of the data.
  - Scales to massive data.
- Black box variational inference
  - Beyond exponential families and model-specific derivations.

References I

Beal, M. J. and Z. Ghahramani (2003). "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures". In: *Bayesian Statistics 7*. Proceedings of the Seventh Valencia International Meeting. Ed. by J. M. Bernardo, A. P. Dawid, J. O. Berger, M. West, D. Heckerman, M. Bayarri, and A. F. Smith. Oxford: Oxford University Press, pp. 453–464.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blei, D. M. (2017). "Variational Inference: Foundations and Innovations". URL: https://simons.berkeley.edu/talks/david-blei-2017-5-1.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). "Variational inference: A review for statisticians". *Journal of the American Statistical Association*, to appear.

References II

Erosheva, E. A., S. E. Fienberg, and C. Joutard (2007). “Describing disability through individual-level mixture models for multivariate binary data”. *Annals of Applied Statistics,* 1.2, p. 346.

Grimmer, J. (2010). “An introduction to Bayesian inference via variational approximations”. *Political Analysis* 19.1, pp. 32–47.

Gunawardana, A. and W. Byrne (2005). “Convergence theorems for generalized alternating minimization procedures”. *Journal of machine learning research* 6.Dec, pp. 2049–2073.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Wang, Y. S., R. Matsueda, and E. A. Erosheva (2017). “A Variational EM Method for Mixed Membership Models with Multivariate Rank Data: an Analysis of Public Policy Preferences”. arXiv: 1512.08731.

References III

Zhao, H. and P. Marriott (2013). "Diagnostics for Variational Bayes approximations". arXiv: 1309.5117.