Bayesian Inference

Viktor

December 3, 2019

1 Variational Inference

Variational inference [4, 1] is an alternative to MCMC in the Bayes world. In essence, the posterior distributions are approximated by optimizing over a family of testfunctions on a scalar bound, the KL diverence. In the special case where we are interested in the point estimate of of model parameters, this gets more structure and leads [13] to the EM algorithm [5, 8](which also gives an estimate for the hidden variables). [6].

For many practical applications the mean field approximation over the variational densities is applied which, leads to coordinate ascent mean-field variational inference (CAVI) [3]. CAVI can be seen as "message passing" algorithm and thus connects variational inference with graphical models [16, 15, 9, 12] (implementation https://dotnet.github.io/infer/).

From a numerical point of view there have been stochastic generalizations in order to deal with large datasets. In order to avoid computing the functional derivations explicitly for a given model analytically an "automatic differentiation variational inference" (ADVI) method has been proposed [10].

The problem. Assume that we have random variables X of observed data and Z of latent random variables. Inference of the model $p(x, z) = p(x|z)p(z)$ amounts to finding the conditional probabilities $p(z|x)$. This quantity is also known posterior. Formally this is trivial via Bayes theorem $p(z|x) = p(x, y)/p(x)$. However computing $p(x) = \int p(x, z)dz$ may be computationally intractable (for example, it may lead to very high dimensional numeric integrations for correlated latent variables). Cynically, $p(x)$ is just a normalization factor. In physics it is called partition function and in statistics its called evidence. In fact, computing the evidence is one of the most challenging parts of Bayesian methods.

The optimization problem. Variational inference tries to approximately determine $p(z|x)$ by introducing a family of approximate probability densities $q(z|x) \in \mathcal{Q}$ and then finding that member that approximates the posterior optimally with respect to the KL divergence,

$$q^*(z) = \underset{q(z) \in \mathcal{Q}}{\operatorname{argmin}} \, D(q(z) \parallel p(z|x)), \tag{1}$$

where the shorthand notation $q(z) \equiv q(z|x)$ was introduced. This transforms an integration problem to an optimization problem. However, the KL divergence of this form is of no help to the original problem (computing the marginal $p(x)$) as we still need to compute the term $p(z|x)$ which, requires the term $p(x)$. So we need to find another objective function, the evidence lower bound (ELBO),

$$\mathcal{F}(q) := \langle \log p(x, z) \rangle_{q(z)} - \langle \log q(z) \rangle_{q(z)} \tag{2}$$

Theorem 1. *Let there be given a family of probability densities $\mathcal{Q}$. Then $q^*(z)$ is a minimizer over $\mathcal{Q}$ of $D(q(z) \parallel p(z|x))$ iff it is a maximizer of the ELBO $\mathcal{F}(q)$.*

Proof. This follows from

$$\begin{aligned} D(q(z) \parallel p(z|x)) &= \langle \log q(z) \rangle_{q(z)} - \langle \log p(z|x) \rangle_{q(z)} \\ &= -\mathcal{F}(q) + \langle \log p(x) \rangle_{q(z)} \\ &= -\mathcal{F}(q) + p(x), \end{aligned}$$

and that $p(x)$ is constant with respect to $q(z)$, $\delta p(x)/\delta q(z) = 0$ □

So, first we need to specify a family of variational densities $\mathcal{Q}$ and then optimize Equation 2 over $\mathcal{Q}$. Not that both, the KL and ELBO objective functions implicitly depend on x and so does $q^*(z)$. This just reflects the dependence on the training data.

Discussion. Turning to the interpretation of variational inference rewrite

$$\begin{aligned} \mathcal{F}(q) &= \langle \log p(x|z) \rangle_{q(z)} + \langle \log p(z) \rangle_{q(z)} - \langle \log q(z) \rangle_{q(z)} \\ &= \langle \log p(x|z) \rangle_{q(z)} - D(q(z) \parallel p(z)). \end{aligned} \tag{3}$$

The density that maximizes the ELBO thus compromises between i) putting most probability weight on z where the posterior $p(x|z)$ is large and ii) beeing close to the prior $p(z)$. Furthermore if we assume that the N observed data points are iid of the form $x = x_1, \ldots, x_N$ the first expression reads

$\langle \log p(x|z) \rangle_{q(z)} = \sum_{i=1}^{N} \langle \log p(x_i|z) \rangle_{q(z)}$. Thus, the first term becomes more important for large number of observed data points. But this is just in line with the usual Bayesian behavior.

Furthermore the ELBO bounds the (log) evidence $p(x)$,

$$\log p(x) = \mathcal{F}(q) + D(q(z) \parallel p(z|x)) \geq \mathcal{F}(q) \tag{4}$$

because $D(\cdot) \geq 0$. This property lends ELBO its name and has been used for model selection under the assumption that the ELBO is a good approximation to the marginal likelihood. However, this approach is not rigorous. So we will just use the ELBO as objective function for the optimization problem (which is rigorous).

The KL-divergence enjoys the property $D(q \parallel p) = 0 \Leftrightarrow q = p$. Assume a maximizer q^* of the ELBO, which is also a minimizer of the KL-divergence. Now, if the KL-divergence $D(q^*(z) \parallel p(z|l)) = 0$ then q^* is the true posterior. However this need not be the case as the variational family typically suffers from severe restrictions due computational resources. In this case it is not clear how "good" the minimizer of the KL-divergence approximates the posterior [12, 11].

1.1 Mean-field approximation

Definition 1 (Mean field approximation)**.** *A family of pdfs $q_Z(z) \in \mathcal{Q}$ is called mean field approximation if it is of the form $q_Z(z) = \prod_\ell q_{Z_\ell}(z_\ell)$ with $q_{Z_\ell}(z_\ell) \in \mathcal{Q}_\ell$ and $Q = \otimes_\ell \mathcal{Q}_\ell$.*

Note that the definition does not assume how the latent variables are factorized. The case where the latent variables are fully factorized is sometimes called *naive mean field approximation* while the case where some latent variables are collected into groups is sometimes called *generalized mean field approximation.* In the following we use the shorthand notation $q_\ell(z_\ell) \equiv q_{Z_\ell}(z_\ell)$

Theorem 2. *Given a mean field approximation of a variational family, $\prod_\ell q_\ell(z_\ell) \in \otimes_\ell \mathcal{Q}_\ell$ with $\mathcal{Q}_\ell = \{q_\ell(z_\ell) \,|\, q_\ell(z_\ell) \text{ is pdf}\}$. Then the maximizer of the ELBO is given by*

$$q^*(z) = \prod_\ell q_\ell^*(z_\ell), \tag{5}$$

$$q^*(z_\ell) = \frac{1}{N_\ell} \exp \langle \log p(x, z_\ell, z_{-\ell}) \rangle_{q_{z_{-\ell}}}, \tag{6}$$

where $z = \{z_\ell, z_{-\ell}\}$, $z_{-\ell} = \{z_1, \ldots, \cancel{z_\ell} \ldots\}$ and $q_{z_{-\ell}} = \prod_{\neq \ell} q_i(z_i)$

Proof. Taking the functional derivative of the ELBO under the constraint that all $q_\ell(z_\ell)$ are pdfs, i.e., $\int q_\ell(z_\ell)\, dz_\ell = 1$ and $q_\ell(z_\ell) \geq 0\ \forall z_\ell$ via Lagrangian multipliers gives,

$$\begin{aligned}
&\frac{\delta}{\delta q_k(z'_k)} \left\{ \mathcal{F}(\prod_{\ell=1}^{L} q^*_\ell(z_\ell)) - \sum_{\ell=1}^{L} \lambda_\ell \left(\int q_\ell(z_\ell)\, dz_\ell - 1 \right) \right\} \\
=&\frac{\delta}{\delta q_k(z'_k)} \left\{ \int dz_1 \cdots \int dz_1\, p(x, z_1, \dots, z_L) \prod_{\ell=1}^{L} q_\ell(z_\ell) - \sum_{\ell=1}^{L} \int dz_\ell\, q_\ell(z_\ell) \log q_\ell(z_\ell) \right\} \\
&- \lambda_k \int \delta(z_k - z'_k)\, dz_k \\
=&\int \cdots \int dz_1 \dots \cancel{dz}_k \dots dz_L\, p(x, z_1, \dots, z'_k, \dots z_L) \prod_{\ell \neq k} q_\ell(z_\ell) - \log q_k(z'_k) - 1 - \lambda_k.
\end{aligned}$$

Solving for $q_k(z'_k)$ and using the notation from above gives

$$q_k(z'_k) = \frac{\exp\langle \log p(x, z'_k, z_{-k}) \rangle_{q_{z_{-k}}}}{\exp(1 + \lambda_k)} \tag{7}$$

The Lagrangian multiplier and thus the is obtained by normalization, $N_k = \exp(1 + \lambda_k) = \int dz'_k\ \exp\langle \log p(x, z'_k, z_{-k}) \rangle_{q_{z_{-k}}}$. Together with the functional from this implies that q_k is indeed a pdf. □

Note that (6) can be equivalently written in terms of complete conditionals $p(z_\ell | x, z_{-\ell})$,

$$q^*(z_\ell) = \frac{1}{N'_\ell} \exp \langle \log p(z_\ell | x, z_{-\ell}) \rangle_{q_{z_{-\ell}}}, \tag{8}$$

because $\langle \log p(x, z_\ell, z_{-\ell}) \rangle_{q_{z_{-\ell}}} = \langle \log p(z_\ell | x, z_{-\ell}) \rangle_{q_{z_{-\ell}}} - \langle \log p(x, z_{-\ell}) \rangle_{q_{z_{-\ell}}}$ and the second term does not depend on $q_\ell(z_\ell)$ and thus may be absorbed into the normalization factor.

- Theorem 1 and 2 establish the best approximation to the KL-divergence of the true posterior within a factorized family of variational densities. Recall $q(z) \equiv q(z|x)$ and $z = \{z_1, \dots, z_\ell, \dots\}$

- Integrating out definition 1 shows that the approximation to the marginal posterior corresponds to the variational component, which is also complete conditional.

$$q(z_\ell | x) \equiv q(z_\ell) = q_\ell(z_\ell) \tag{9}$$

$$q(z_\ell | z_{-\ell}, x) = q_\ell(z_\ell) \tag{10}$$

Algorithm 1: Coordinate ascent variational inference

```
Input: A model $p(x, z)$ and a dataset $x$
Result: Maximizer of the ELBO
Initialize: variational densities $q_\ell(z_\ell),\ \ell = 1 \dots m$
$n = 0$
while While the ELBO is not converged do
    for $\ell = 1 \dots m$ do
        $q^{(n+1)}(z_\ell) \propto \exp \langle \log p(x, z_\ell, z_{-\ell}) \rangle_{q^{(n+1)}_{z_{<\ell}} \cdot q^{(n)}_{z_{>\ell}}}$
    end
    Compute ELBO $\mathcal{F}(q^{(n+1)})$ from Eq. (2)
    $n = n + 1$
end
return $q^{(n)}(z) = \prod_{\ell=1}^{m} q_\ell^{(n)}(z_\ell)$
```

- From Theorem 2, $q_\ell^*(z_\ell)$ depends on all other approximate pdfs of random variables. In practice this yields an iterative scheme of computation, the *coordinate ascent mean-field variational inference algorithm* (CAVI)[4, 3]. This fix-point procedure is summarized in Algorithm 1. The short-hand notation that $q^{(n+1)}_{z_{<\ell}} \cdot q^{(n)}_{z_{>\ell}} \equiv q_1^{(n+1)} \dots q_{\ell-1}^{(n+1)} q_\ell^{(n)} \dots q_\ell^{(n)}$ was used to indicate that the updated variational densities are used as soon as they are available [3].

- Especially from Eq (8) the connection between CAVI and the Gibbs sampler in the MCMC approach [6, 7] can be seen. Both use the complete conditional $p(z_\ell | x, z_{-\ell})$ in their update steps.

- In general this formulation is not very useful for practical computations. Thus typically further structure is incorporated by using the exponential family as variational family, which leads to much simpler CAVI update equations and allows variational inference to scale to massive data [4, 2, 14].

Variational Bayesian Expectation-Maximization (VB-EM). So far there was no interpretation on the latent variables. Consider a model with latent variables $z = \{z, \theta\}$, where θ are model parameters and z are hidden (e.g., local) variables. The Variational Bayesian Exptectation-Maximisation (VB-EM) [1, 2] is just a consequence of Theorem 2 together with the mean-field ansatz $q(z,\theta) = q_z(z)q_\theta(\theta)$,

$$q_z(z) \propto \exp \langle \log p(z,x,\theta) \rangle_{q_\theta}, \tag{11}$$
$$q_\theta(\theta) \propto \exp \langle \log p(x,z,\theta) \rangle_{q_z}, \tag{12}$$

which can be equivalently written as

$$q_z(z) \propto \exp \langle \log p(z,x|\theta) \rangle_{q_\theta}, \tag{13}$$
$$q_\theta(\theta) \propto p(\theta) \exp \langle \log p(x,z|\theta) \rangle_{q_z}. \tag{14}$$

This follow $p(x,z,\theta) = p(x,z|\theta)p(\theta)$. The residual contribution $\exp \langle \log p(\theta) \rangle_{q_\theta}$ may be absorbed in the normalization factor in (13) and $\exp \langle \log p(\theta) \rangle_{q_z} = p(\theta)$ in (14).

1.2 Parametric variational distributions

1.3 Point estimates

Very often θ is the only quantity of interest, either 1) as a marginal posterior pdf, $\int p(\theta,z|x)\,dz$, or 2) by a point-estimate of it (MAP) $\mathrm{argmax}_\theta \int p(\theta,z|x)\,dz$, or 3) as a maximum likelihood estimate (MLE) $\mathrm{argmax}_\theta \int p(x,z|\theta)\,dz$. Clearly, the first case can be approximately obtained by VB-EM or equivalently, the mean-field ansatz $q(z,\theta) = q(z)q(\theta)$ and using the machinery of Theorem 2.

Expectation-Maximisation The aim of the expectation maximisation (EM) algorithm is provide a maximum likelihood estimate (MLE) of the model parameters θ under hidden random variables Z, $\mathrm{argmax}_\theta\, p(x|\theta) = \mathrm{argmax}_\theta \int p(x,z|\theta)\,dz$.

Theorem 3. *Let there be given a variational family $q \in \mathcal{Q}$. Then*

$$\mathcal{F}(q,\theta) := \langle \log p(x,z|\theta) \rangle_{q(z)} - \langle \log q(z) \rangle_{q(z)} \tag{15}$$

is a lower bound to the log-likelihood $\log p(x|\theta) \geq \mathcal{F}(q,\theta)\ \forall q$ *with equality iff* $q(z) = p(z|x,\theta) \in \mathcal{Q}$.

Proof. We have $\log p(x|\theta) = \mathcal{F}(q,\theta) + D(q(z) \parallel p(z|x,\theta)) \geq \mathcal{F}(q,\theta)$. Equality follows from the property of the KL-divergence $D(q \parallel p) = 0 \Leftrightarrow q = p$ □

Note that the variational densities only depend on the observed data $q(z) \equiv q(z|x)$ (but not on parameter θ). In contrast to the fully Baysian variational methods above we need the bounding property of theorem 3 in order to lend the maximizer of $\mathcal{F}(q,\theta)$ an interpretation as beeing the MLE.

```
Algorithm 2: Classical EM Algorithm ($n$-th update)
    E-step: $q^{(n)}(z) = p(z|x, \theta^{(n-1)})$
    M-step: $\theta^{(n)}(z) = \text{argmax}_\theta \langle \log p(x, z|\theta) \rangle_{q^{(n)}}$
```

For a given variational family $\mathcal{Q}$ we can now formulate the *constraint* (with respect to $\mathcal{Q}$) EM Algorithm [1] by optimizing functional (15):

$$q^*(z) = \text{argmax}_q \, \mathcal{F}(q, \theta) \quad \text{(E-step)} \tag{16}$$

$$\theta^* = \text{argmax}_\theta \, \mathcal{F}(q, \theta) \quad \text{(M-step)}, \tag{17}$$

where $\text{argmax}_\theta \, \mathcal{F}(q, \theta) = \text{argmax}_\theta \langle \log p(x, z|\theta) \rangle_q$ because $q(z)$ does not depend on θ.

If the varitional family is sufficiently large such that $= p(z|x, \theta) \in \mathcal{Q}$ (in other words if the conditional posterior $p(z|x, \theta)$ is computable) the the equality condition of theorem 3 can be used to arrive at the *unconstraint*, or classical EM algorithm. [5, 13]

$$q^*(z) = p(z|x, \theta) \quad \text{(E-step)} \tag{18}$$

$$\theta^* = \text{argmax}_\theta \langle \log p(x, z|\theta) \rangle_q \quad \text{(M-step)}. \tag{19}$$

The M-step follows by taking the functional derivative with respect to $q(z)$ under the constraint $\int q(z) \, dz = 1$. While the constraint EM converges to lower bound of the log-likelihood the uncontraint converges to a local maximum of the log-likelihood. Both variants lead to an iterative algorithm (Algorithm 2).

Theorem 3 and the (un)constraint EM formulas are only statements on what happens at the maximizer but not *how* the maximizer is computationally approached. This allows to speed up algorithm 2 for large data sets. Consider a case with global model parameters θ and local hidden random variables $Z = \{Z1, \dots Z_N\}$ together with N observations $X = \{X_1 \dots X_N\}$ such that $p(x, z|\theta) = \prod p(x_i, z_i|\theta)$. Then choose $q(z) = \prod q_i(z_i)$. If the goal is to learn θ there is no need to compute the E-step for all datapoints at each iteration. It is enough to perform the E-step for only one datapoint component as this increases $\mathcal{F}(q, \theta)$ for fixed θ (algorithm 3). Eventually this leads to "sparse" algorithms which have been shown to converge faster **??**.

References

[1] Matthew J. Beal. Variational algorithms for approximate bayesian inference. Technical report, 2003.

```
Algorithm 3: Sparse EM Algorithm ($n$-th update)
  Choose some data point $i$ to be updated
  E-step: $q_k^{(n)}(z) = \begin{cases} p(z_i|x_i, \theta^{(n-1)}) & k = i \\ q_k^{(n)}(z) & k \neq i \end{cases}$
          $q^{(n)}(z) = \prod_k q_k^{(n)}(z)$
  M-step: $\theta^{(n)}(z) = \mathrm{argmax}_\theta \langle \log p(x, z|\theta) \rangle_{q^{(n)}}$
```

[2] J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, M. West (eds, Matthew J. Beal, and Zoubin Ghahramani. The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures, 2003.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[4] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians, 2016.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

[6] Alan E. Gelfand and Adrian F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

[7] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, November 1984.

[8] Maya R. Gupta and Yihua Chen. Theory and use of the em algorithm. *Found. Trends Signal Process.*, 4(3):223–296, March 2011.

[9] David A. Knowles and Tom Minka. Non-conjugate variational message passing for multinomial and binary regression. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1701–1709. Curran Associates, Inc., 2011.

[10] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic differentiation variational inference, 2016.

[11] M.A.R. Leisink and H.J. Kappen. A tighter bound for graphical models, 2001.

[12] Tom Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, January 2005.

[13] Radford M. Neal and Geoffrey E. Hinton. Learning in graphical models. chapter A View of the EM Algorithm That Justifies Incremental, Sparse, and Other Variants, pages 355–368. MIT Press, Cambridge, MA, USA, 1999.

[14] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[15] Matthew Wand, John Ormerod, Simone Padoan, and Rudolf Fuhrwirth. Mean field variational bayes for elaborate distributions. *Bayesian Analysis*, 6, 12 2011.

[16] John Winn and Christopher M. Bishop. Variational message passing. *J. Mach. Learn. Res.*, 6:661–694, December 2005.