A Unifying Approximate Inference Framework from Variational Free Energy Relaxation

Yingzhen Li
University of Cambridge
Cambridge, CB2 1PZ, UK
yl494@cam.ac.uk

Richard E. Turner
University of Cambridge
Cambridge, CB2 1PZ, UK
ret26@cam.ac.uk

Abstract

This short paper extends the free-energy derivations of variational inference, loopy belief propagation and expectation propagation (EP) to a wider range of approximate inference methods including power EP, distributed EP, and black-box alpha-divergence minimisation. The framework provides a very flexible framework for the design of variational algorithms that can mix versions of any of the aforementioned algorithms inside a single coherent algorithm. The framework is general, extending to latent variable models, for example.

1 Relaxations for Variational Free Energy

Consider approximating an intractable posterior $p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z} p_0(\boldsymbol{\theta}) \prod_{n=1}^N p(\boldsymbol{x}_n|\boldsymbol{\theta})$. Such posteriors arise in many contexts e.g. density estimation and regression/classification (for supervised learning the likelihood is replaced by $p(\boldsymbol{y}_n|\boldsymbol{x}_n, \boldsymbol{\theta})$). In the following we use the short-hand $f_n(\boldsymbol{\theta}) = p(\boldsymbol{x}_n|\boldsymbol{\theta})$ for the likelihood functions. Variational inference (VI) [1, 2] employs KL-divergence minimisation to obtain a tractable approximation $q(\boldsymbol{\theta})$ by

$$\min_q \mathrm{KL}[q||p] \Leftrightarrow \min_q \mathcal{F}_{\mathrm{VFE}}(q) = \mathbb{E}_q \left[\log q(\boldsymbol{\theta}) - \log p_0(\boldsymbol{\theta}) - \sum_{n=1}^N \log f_n(\boldsymbol{\theta}) \right]. \tag{1}$$

The RHS expression is called the *variational free energy* (VFE) [1, 2], defined on the set of valid probability densities $\{q : \int q(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1\}$. Minimising this free energy is equivalent to KL-divergence minimisation as Z does not depend on $\boldsymbol{\theta}$ and q. Crucially, it is a *global* optimisation objective wrt. a single q distribution, hence we term it a *global* approximation method. On the other hand, expectation propagation (EP) [3] is a *local* approximation method that can outperform VFE on a variety of tasks, e.g. Gaussian Process classification [4] (although this issue is less clear cut for more modern applications of VFE/EP). VI and EP are the most popular and foundational algorithms for distributional approximate Bayesian inference.

A number of papers have considered the connections between VI and EP, e.g. see [5, 6]. We first revisit these relationships again by a derivation that transforms the VFE into the *Bethe free energy* [7, 8] and then to EP. This is done in sections 1.1 and 1.2 with a standard approach that includes, 1. argument decoupling, 2. constraint relaxation, and 3. dual form representation. In later sections we extend this procedure to power EP and other modern variational models, and also to models that contain latent variables. We further demonstrate the flexibility of this framework by showing how it allows new variants to be derived that mix different VI and EP-like algorithms. A sketch of this procedure is provided in the main text and the full derivations can be found in appendix.

Submitted to the Approximate Inference workshop, NIPS 2016, Barcelona, Spain.

1.1 From VFE to Bethe Free Energy

First we make use of the additivity of logarithm to transform VFE optimisation into an equivalent optimisation problem:

$$\min_{q,\{\tilde{p}_n\}} (1-N)\mathrm{KL}[q||p_0] - \sum_n \mathbb{E}_{\tilde{p}_n}\left[\log \frac{p_0(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})}{\tilde{p}_n(\boldsymbol{\theta})}\right] \quad \text{subject to } \tilde{p}_n = q, \forall n. \tag{2}$$

Full details can be found in appendix, eqns. (13) and (14). In summary we have rearranged terms in the VFE and then **decoupled** the global q distributions into a set of local distributions $\tilde{p}_n$, which are then tied together by introducing equality constraints. The first approximation can now be made by **relaxing** the equality constraints to moment matching constraints $\mathbb{E}_{\tilde{p}_n}[\boldsymbol{\theta}^k] = \mathbb{E}_q[\boldsymbol{\theta}^k]$ for $k \in \mathbb{N}$, and a further crude relaxation suggests moment matching just for the first M moments[1] $\mathbb{E}_{\tilde{p}_n}[\boldsymbol{\theta}^m] = \mathbb{E}_q[\boldsymbol{\theta}^m], m = 1, 2, ..., M$. In the following we use a vectorial function $\boldsymbol{\phi}(\boldsymbol{\theta})$ to summarise these constraints as $\mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}] = \mathbb{E}_q[\boldsymbol{\phi}]$, where as an example $\boldsymbol{\phi}(\boldsymbol{\theta}) = [\boldsymbol{\theta}, \boldsymbol{\theta}\boldsymbol{\theta}^T]$ for Gaussian EP. In general $\boldsymbol{\phi}$ can contain any polynomial terms or other basis functions. This relaxation returns the following constrained optimisation problem:

$$\begin{aligned} &\min_{q,\{\tilde{p}_n\}} \mathcal{F}_{\text{Bethe}}(\{\tilde{p}_n\}, q) \quad \text{subject to } \mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}] = \mathbb{E}_q[\boldsymbol{\phi}], \forall n, \\ &\mathcal{F}_{\text{Bethe}}(\{\tilde{p}_n\}, q) = (1-N)\mathrm{KL}[q||p_0] - \sum_n \mathbb{E}_{\tilde{p}_n}\left[\log \frac{p_0(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})}{\tilde{p}_n(\boldsymbol{\theta})}\right]. \end{aligned} \tag{3}$$

$\mathcal{F}_{\text{Bethe}}(\{\tilde{p}_n\}, q)$ is the *Bethe free energy* [7, 8] that is usually presented in the context of probabilistic graphical models. Below we show how to derive its dual form, which is usually discussed in EP literature [9, 10, 11].

Remark. Minka [9] discussed (3) as a minimax problem ($\min_{\{\tilde{p}_n\}} \max_q$) instead, which we believe is incorrect. First (3) relaxes the constraints in (2), meaning both should have the same optimisation direction. Then since (2) just decouples VFE (1) with equality constraints, it will be a pure minimisation with the same stationary points. For graphical models the Bethe free energy optimisation problem is formulated as a pure minimisation problem like above (3), e.g. in [6, 12] but also interestingly in pages 3-4 of [9]. On the other hand, Minka's derivation of the dual energy takes a different approach and does not require the minimax assumption of the primal problem.

1.2 From Bethe to EP: a Dual Form Representation

We provide a derivation in a similar way as [12], starting from a note on the KL duality[2]

$$-\mathrm{KL}[q||p_0] = \min_{\lambda_q(\boldsymbol{\theta})} -\mathbb{E}_q[\lambda_q(\boldsymbol{\theta})] + \log \mathbb{E}_{p_0}\left[\exp[\lambda_q(\boldsymbol{\theta})]\right]. \tag{4}$$

The RHS upper-bound to the KL term is in the same spirit as the variational lower-bound for $\log Z$, in that both employ convex duality. Equality is achieved by $q(\boldsymbol{\theta}) \propto p_0(\boldsymbol{\theta}) \exp[\lambda_q(\boldsymbol{\theta})]$. Substitution into (3) then yields a transformed energy that we denoted as $\mathcal{F}_{\text{Bethe}}(\{\tilde{p}_n\}, q, \lambda_q(\boldsymbol{\theta}))$ (see (16) in appendix). Denote $\boldsymbol{\lambda}_{-n}$ as the Lagrange multiplier for moment matching and ν, ν_n for the normalisation constraints of q and $\tilde{p}_n$, respectively. This returns the following Lagrangian

$$\min_{q,\{\tilde{p}_n\},\lambda_q(\boldsymbol{\theta})} \max_{\{\boldsymbol{\lambda}_{-n},\nu_n,\nu\}} \mathcal{F}_{\text{Bethe}}(\{\tilde{p}_n\}, q, \lambda_q(\boldsymbol{\theta})) + \sum_n \boldsymbol{\lambda}_{-n}^T(\mathbb{E}_q[\boldsymbol{\phi}] - \mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}]) + ... \tag{5}$$

where we have omitted the terms associated with ν_n and ν for notational ease (see (17) in appendix). Solving the fixed points for $\tilde{p}_n$ and ν_n returns $\tilde{p}_n(\boldsymbol{\theta}) = \frac{1}{Z_n} p_0(\boldsymbol{\theta}) f_n(\boldsymbol{\theta}) \exp\left[\boldsymbol{\lambda}_{-n}^T \boldsymbol{\phi}(\boldsymbol{\theta})\right]$, where the normalising constant is $Z_n = \int p_0(\boldsymbol{\theta}) f_n(\boldsymbol{\theta}) \exp\left[\boldsymbol{\lambda}_{-n}^T \boldsymbol{\phi}(\boldsymbol{\theta})\right] d\boldsymbol{\theta}$. Also it is straight-forward to evaluate the fixed point condition for q: $(N-1)\lambda_q(\boldsymbol{\theta}) = \sum_n \boldsymbol{\lambda}_{-n}^T \boldsymbol{\phi}(\boldsymbol{\theta}) + \nu$. We explicitly specify $\lambda_q(\boldsymbol{\theta}) = \boldsymbol{\lambda}_q^T \boldsymbol{\phi}(\boldsymbol{\theta}) + \frac{1}{N-1}\nu$ w.l.o.g., so now the stationary condition changes to $(N-1)\boldsymbol{\lambda}_q = \sum_n \boldsymbol{\lambda}_{-n}$. Also as shown in appendix the constant ν can be dropped. Importantly, substituting $\tilde{p}_n$ back to (5) and enforcing the fixed point condition for q yields the *EP energy* [9]:

$$\min_{\boldsymbol{\lambda}_q} \max_{\{\boldsymbol{\lambda}_{-n}\}} (N-1)\log \mathbb{E}_{p_0}\left[\exp[\boldsymbol{\lambda}_q^T \boldsymbol{\phi}(\boldsymbol{\theta})]\right] - \sum_n \log Z_n \quad \text{subject to } (N-1)\boldsymbol{\lambda}_q = \sum_n \boldsymbol{\lambda}_{-n}. \tag{6}$$

[1] The zeroth moment matching constraint is replaced by the constraint that $\tilde{p}_n$ integrates to 1.

[2] We include this step in order to connect to the EP energy with optimisation arguments all in the dual space.

Notice now the optimisation problem over q is dropped since (6) does not depend on it. To obtain the approximate posterior back, we make use of the tightness of the KL duality, and define $q(\boldsymbol{\theta}) = \frac{1}{Z_q} p_0(\boldsymbol{\theta}) \exp\left[\boldsymbol{\lambda}_q^T \boldsymbol{\phi}(\boldsymbol{\theta})\right]$ with $\log Z_q = \log \mathbb{E}_{p_0}\left[\exp[\boldsymbol{\lambda}_q^T \boldsymbol{\phi}(\boldsymbol{\theta})]\right]$. The expectation consistent approximate inference (EC) algorithm [10] is a special case with $p_0(\boldsymbol{\theta}) \propto 1$ and $N = 2$.

Remark. In the EP literature $\tilde{p}_n(\boldsymbol{\theta})$ is named the *tilted distribution*, and $\boldsymbol{\lambda}_{-n}$ is the natural parameter of the *cavity distribution* for factor f_n. In practice (6) requires a double-loop algorithm to guarantee convergence [13]. Instead EP [3] parametrises the (natural parameters of) local approximating factors by $\boldsymbol{\lambda}_n = \boldsymbol{\lambda}_q - \boldsymbol{\lambda}_{-n}$, with the goal of $\exp[\boldsymbol{\lambda}_n^T \boldsymbol{\phi}(\boldsymbol{\theta})]$ capturing the effect of $f_n(\boldsymbol{\theta})$ on the exact posterior. EP runs a fixed point iteration algorithm to find the stationary points for $\{\boldsymbol{\lambda}_n\}_{n=1}^N$. The constraint is enforced by calculating $\boldsymbol{\lambda}_q = \sum_n \boldsymbol{\lambda}_n$ and $\boldsymbol{\lambda}_{-n} = \sum_{m \neq n} \boldsymbol{\lambda}_m$ after each update.

1.3 From VFE to Power EP

We now extend the above approach to power EP [14] which is a new contribution, although fairly straightforward. This procedure includes one modification to the Bethe free energy. Assume for each factor f_n a power value $\alpha_n \neq 0$ is associated, with $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_N)$ and $\sum_n \frac{1}{\alpha_n} \neq 1$. Then the Bethe free energy with moment matching constraints is modified to (see (21) in appendix):

$$\mathcal{F}_{\boldsymbol{\alpha}}(q, \{\tilde{p}_n\}) = \left(1 - \sum_n \frac{1}{\alpha_n}\right) \mathrm{KL}[q||p_0] - \sum_{n=1}^N \frac{1}{\alpha_n} \mathbb{E}_{\tilde{p}_n}\left[\log \frac{p_0(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})^{\alpha_n}}{\tilde{p}_n(\boldsymbol{\theta})}\right]. \tag{7}$$

Calculations following section 1.2 also reveal the change of the fixed point condition for q to $(\sum_n \frac{1}{\alpha_n} - 1)\boldsymbol{\lambda}_q = \sum_n \frac{1}{\alpha_n} \boldsymbol{\lambda}_{-n}$. Define q as an exponential family distribution with natural parameter $\boldsymbol{\lambda}_q$ as before, and $\boldsymbol{\lambda}_n = (\boldsymbol{\lambda}_q - \boldsymbol{\lambda}_{-n})/\alpha_n$. We arrive at the power EP objective:

$$\left(\sum_n \frac{1}{\alpha_n} - 1\right) \log Z_q - \sum_n \frac{1}{\alpha_n} \log \int p_0(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})^{\alpha_n} \exp\left[(\boldsymbol{\lambda}_q - \alpha_n \boldsymbol{\lambda}_n)^T \boldsymbol{\phi}(\boldsymbol{\theta})\right] d\boldsymbol{\theta}. \tag{8}$$

The iterative process also enforces $\boldsymbol{\lambda}_q = \sum_n \boldsymbol{\lambda}_n$. It is shown in [5] that (8) becomes an lower-bound of $-\log Z$ when $\alpha_n > 0$ and $\sum_n \frac{1}{\alpha_n} < 1$. On the other hand, taking $\alpha_n \to 0, \forall n$ recovers $\mathcal{F}_{\text{VFE}}$ but now the q distribution is restricted to have an exponential family form.

2 Further Constraint Relaxations by Weighted Averaging

In EP we need N Lagrange multipliers $\{\boldsymbol{\lambda}_{-n}\}$ because of the individual moment matching constraints $\mathbb{E}_q[\boldsymbol{\phi}] = \mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}]$. This translates into a large memory burden for large datasets and "big" models. To solve this issue, we start from the re-formulated energy function (7), and then replace the N equality constraints $q = \tilde{p}_n$ by a single one that we call *weighted averaged moment matching*: $\mathbb{E}_q[\boldsymbol{\phi}] = \sum_n w_n \mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}]$, with $\boldsymbol{w} = (w_1, ..., w_n)$ denote the weighing vector that sum to 1. The motivation here is to reduce the number of Lagrange multipliers (thus saving memory) but to ensure that q still resembles the tilted distributions. Empirical evaluations have shown that this relaxation method returns state-of-the-art performance for some problems [15] even though it can be a very poor approximation to retaining N equality constraints in (2).

We proceed to solve the corresponding constrained optimisation problem. We also apply the KL duality (4) to (7) (but for simplicity now we directly use $\boldsymbol{\lambda}_q^T \boldsymbol{\phi}(\theta)$), and denote the transformed energy (in a similar way as before) as $\mathcal{F}_{\boldsymbol{\alpha}}(\{\tilde{p}_n\}, q, \boldsymbol{\lambda}_q)$. Now we have the following Lagrangian

$$\min_{q, \{\tilde{p}_n\}, \boldsymbol{\lambda}_q} \max_{\lambda, \{\nu_n\}, \nu} \mathcal{F}_{\boldsymbol{\alpha}}(\{\tilde{p}_n\}, q, \boldsymbol{\lambda}_q) + \boldsymbol{\lambda}^T (\mathbb{E}_q[\boldsymbol{\phi}] - \sum_n w_n \mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}]) + ... \tag{9}$$

This returns $\tilde{p}_n(\boldsymbol{\theta}) = \frac{1}{Z_n} p_0(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})^{\alpha_n} \exp\left[\alpha_n w_n \boldsymbol{\lambda}^T \boldsymbol{\phi}(\boldsymbol{\theta})\right]$, but the fixed point condition for q becomes $\left(\sum_n \frac{1}{\alpha_n} - 1\right) \boldsymbol{\lambda}_q = \boldsymbol{\lambda}$, indicating $\boldsymbol{\lambda}$ as a function of $\boldsymbol{\lambda}_q$. Thus we can directly obtain a single-loop algorithm for the following minimisation

$$\min_{\boldsymbol{\lambda}_q} \left(\sum_n \frac{1}{\alpha_n} - 1\right) \log Z_q - \sum_n \frac{1}{\alpha_n} \log \int p_0(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})^{\alpha_n} \exp\left[\beta_n \boldsymbol{\lambda}_q \boldsymbol{\phi}(\boldsymbol{\theta})\right] d\boldsymbol{\theta} \tag{10}$$

with $\beta_n = \left(\sum_m \frac{1}{\alpha_m} - 1\right) \alpha_n w_n$. Black-box alpha (BB-$\alpha$) [15] is recovered by a special case of the above via defining $\alpha_n = \alpha, \forall n$ and $w_n = (\frac{1}{\alpha_n} - \frac{1}{N})/(\sum_m \frac{1}{\alpha_m} - 1) = 1/N$, which leads to $\beta_n = 1 - \alpha/N$. In this case the weighing vector $\boldsymbol{w}$ is implicitly defined by the choice of α_n.

Remark. The original derivation of BB-α in [15] was ad hoc starting from the power EP energy (8) and then tying the local parameters $\boldsymbol{\lambda}_n$. The derivation here provides a rigorous justification from a constrained primal energy optimisation perspective.

3 Distributed Variational Methods

Variational methods have been shown to be very efficient for distributed computing, e.g. see [16, 17, 18] for EP and [19] for VFE. In the new set-up this is equivalent to employing different decoupling strategies. One way to see this is to first group the factors into K subset-level factors $F_k = \prod f_{n_k}$ with the corresponding index sets $N_k = \{n_k\}$, where usually $N_i \cap N_j = \emptyset,\ i \neq j$ and $\cup_k N_k = \{1, ..., N\}$. Now we perform EP at the group level. Precisely, we rewrite the variational free energy (1) and also assume for simplicity $\alpha_k \neq 0$ and $\alpha_{n_k} = \alpha_k$ for $n_k \in N_k$:

$$\mathcal{F}_{\text{VFE}}(q) = \left(1 - \sum_k \frac{1}{\alpha_k}\right) \text{KL}[q||p_0] - \sum_k \frac{1}{\alpha_k} \mathbb{E}_q \left[\log \frac{p_0(\boldsymbol{\theta}) F_k(\boldsymbol{\theta})^{\alpha_k}}{q(\boldsymbol{\theta})}\right]. \tag{11}$$

One can easily recover the distributed EP objective by repeating the decoupling and Lagrangian computation procedures as in section 1.1 and 1.2, with the moment-matching constraints applied to $\tilde{p}_k$ for subset-level factor F_k rather than individual factor f_n. An equivalent derivation starts from power EP in section 1.3 but with an extra set of constraints restricting $\tilde{p}_i = \tilde{p}_j, \forall i, j \in N_k$. Simple calculations reveal that in this case $1/\alpha_k$ in (11) equals to the sum of $1/\alpha_{n_k}$ for all $n_k \in N_k$.

Monte Carlo (MC) methods have been employed to compute these moments since now we incorporate multiple factors (data points) in the second integral. The EP/MC mixed approach combines the advantages from both world: it provides more accurate approximations than full EP since it is less "local", while it remains faster than full MC methods (as the tilted distribution contains less complex factors) and is straight-forward to parallelise. In the extreme case with $K = 1$ and $\alpha \neq 0, 1$, the above derivation recovers the *variational Rényi bound* [20], with the q distribution restricted to exponential families. Rényi divergences [21] have been adopted in [20] to allow extensions to all valid q distributions, and a connection to BB-α was also derived.

We further present a mixed approach that nests BB-α in distributed EP, and again for simplicity we assume $\alpha_{n_k} = \alpha_k$ for $n_k \in N_k$. We still decouple all the q distributions associated with factor $\tilde{f}_n$ to $\tilde{p}_n$ as in (7), but then relax the equality constraints to $\frac{1}{|N_k|} \sum_{n_k \in N_k} \mathbb{E}_{\tilde{p}_{n_k}}[\boldsymbol{\phi}] = \mathbb{E}_q[\boldsymbol{\phi}],\ \forall k$. Solving the Lagrangian and defining $\boldsymbol{\lambda}_q = \sum_k \boldsymbol{\lambda}_k$ returns the following dual energy:

$$(\sum_k \frac{|N_k|}{\alpha_k} - 1) \log Z_q - \sum_k \frac{1}{\alpha_k} \sum_{n_k \in N_k} \log \int p_0(\boldsymbol{\theta}) f_{n_k}(\boldsymbol{\theta})^{\alpha_k} \exp\left[(\boldsymbol{\lambda}_q - \frac{\alpha_k}{|N_k|} \boldsymbol{\lambda}_k)^T \boldsymbol{\phi}(\boldsymbol{\theta})\right] d\boldsymbol{\theta},$$

This means the local parameters $\boldsymbol{\lambda}_k$ are updated in a BB-α fashion, while the final approximation q is constructed in an EP way. Potentially this approach can both reduce the approximation bias of BB-α (since more than 1 local parameter is in use) as well as being computationally faster than distributed EP (as often now the moments for the tilted distribution become tractable).

4 Discussion

We have presented a principled way to construct a number of approximate inference algorithms, by manipulating the variational free energy though likelihood grouping, argument decoupling, and constraint relaxation steps. The algorithms discussed so far are not directly applicable to latent variable models, but they can be extended to do so as described appendix B. Gaussian EP can be viewed as a continuous version of *belief propagation* (BP) [22, 23], and we have also shown an equivalent Bethe free energy definition (3) for posterior approximation. Theoretical properties of BP and Bethe free energy have been (and continue to be) extensively studied (e.g. see [6]), however much less existing work has analysed the convergence and energy approximation accuracy of EP. Future work will focus on these theoretical issues connecting to the rich literature of variational inference, message passing and constrained optimisation.

References

[1] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[2] M. J. Beal, *Variational algorithms for approximate Bayesian inference*. PhD thesis, University College London, 2003.

[3] T. Minka, "Expectation propagation for approximate Bayesian inference," in *Uncertainty in Artificial Intelligence*, vol. 17, pp. 362–369, 2001.

[4] M. Kuss and C. E. Rasmussen, "Assessing approximate inference for binary gaussian process classification," *Journal of Machine Learning Research*, vol. 6, no. Oct, pp. 1679–1704, 2005.

[5] T. Minka, "Divergence measures and message passing," tech. rep., Technical report, Microsoft Research, 2005.

[6] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1-2, pp. 1–305, 2008.

[7] Bethe, "Statistical theory of superlattices," *Proc. R. Soc. Lond. A*, vol. 150, no. 871, pp. 552–575, 1935.

[8] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Bethe free energy, kikuchi approximations, and belief propagation algorithms," in *Neural Information Processing Systems*, 2001.

[9] T. Minka, "The ep energy function and minimization schemes," tech. rep., Technical report, 2001.

[10] M. Opper and O. Winther, "Expectation consistent approximate inference," *The Journal of Machine Learning Research*, vol. 6, pp. 2177–2204, 2005.

[11] M. Seeger, "Expectation propagation for exponential families," tech. rep., 2005.

[12] T. Heskes, "Stable fixed points of loopy belief propagation are local minima of the bethe free energy," in *Advances in neural information processing systems*, pp. 343–350, 2002.

[13] T. Heskes and O. Zoeter, "Expectation propagation for approximate inference in dynamic bayesian networks," in *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pp. 216–223, Morgan Kaufmann Publishers Inc., 2002.

[14] T. Minka, "Power ep," Tech. Rep. MSR-TR-2004-149, Microsoft Research, Cambridge, 2004.

[15] J. M. Hernández-Lobato, Y. Li, M. Rowland, D. Hernández-Lobato, T. Bui, and R. E. Turner, "Black-box α-divergence minimization," in *International Conference of Machine Learning*, 2016.

[16] A. Gelman, A. Vehtari, P. Jylänki, C. Robert, N. Chopin, and J. P. Cunningham, "Expectation propagation as a way of life," *arXiv:1412.4869*, 2014.

[17] M. Xu, B. Lakshminarayanan, Y. W. Teh, J. Zhu, and B. Zhang, "Distributed bayesian posterior sampling via moment sharing," in *Neural Information Processing Systems*, 2014.

[18] Y. W. Teh, L. Hasenclever, T. Lienart, S. Vollmer, S. Webb, B. Lakshminarayanan, and C. Blundell, "Distributed bayesian learning with stochastic natural-gradient expectation propagation and the posterior server," *arXiv preprint arXiv:1512.09327*, 2015.

[19] T. Broderick, N. Boyd, A. Wibisono, A. C. Wilson, and M. I. Jordan, "Streaming variational bayes," in *Neural Information Processing Systems* (C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, eds.), pp. 1727–1735, 2013.

[20] Y. Li and R. E. Turner, "Rényi divergence variational inference," in *Neural Information Processing Systems*, 2016.

[21] A. Rényi, "On measures of entropy and information," *Fourth Berkeley symposium on mathematical statistics and probability*, vol. 1, 1961.

[22] J. Pearl, "Reverend bayes on inference engines: A distributed hierarchical approach," in *The Second National Conference on Artificial Intelligence (AAAI-82)*, 1982.

[23] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2282–2312, 2005.

[24] T. Salimans and D. A. Knowles, "Fixed-form variational posterior approximation through stochastic linear regression," *Bayesian Analysis*, vol. 8, no. 4, pp. 837–882, 2013.

[25] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations*, 2014.

[26] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International Conference of Machine Learning*, 2014.

[27] J. Winn and C. M. Bishop, "Variational message passing," *Journal of Machine Learning Research*, vol. 6, no. Apr, pp. 661–694, 2005.

A Derivations for Free-Energy Relaxations

In this section a full detailed derivation of sections 1.1 and 1.2 in the main text is provided.

To start with, the KL-divergence is defined as

$$\mathrm{KL}[q(\boldsymbol{\theta})||p_0(\boldsymbol{\theta})] = \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta})} d\boldsymbol{\theta}. \tag{12}$$

A.1 From VFE to Bethe Free Energy: Derivations

First we make use of the additivity of logarithm to rewrite the VFE as

$$\begin{aligned}
\mathcal{F}_{\text{VFE}}(q) &= \mathbb{E}_q \left[\log q(\boldsymbol{\theta}) - \log p_0(\boldsymbol{\theta}) - \sum_n \log f_n(\boldsymbol{\theta}) \right] \\
&= \mathbb{E}_q \left[\log \frac{q(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta})} - N \log \frac{q(\boldsymbol{\theta}) p_0(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta}) q(\boldsymbol{\theta})} - \sum_n \log f_n(\boldsymbol{\theta}) \right] \\
&= (1-N)\mathrm{KL}[q||p_0] - \sum_n \mathbb{E}_q \left[\log \frac{p_0(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right].
\end{aligned} \tag{13}$$

Note here the energy is represented by a collection of "local VFE" (the second term in the last line). An equivalent optimisation problem **decouples** those local variational problems, but to obtain the same solution a set of equality constraints is imposed:

$$\min_{q, \{\tilde{p}_n\}} (1-N)\mathrm{KL}[q||p_0] - \sum_n \mathbb{E}_{\tilde{p}_n} \left[\log \frac{p_0(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})}{\tilde{p}_n(\boldsymbol{\theta})} \right] \quad \text{subject to } \tilde{p}_n = q, \forall n. \tag{14}$$

The equality constraints are equivalent to constraining the moment generating function of $\tilde{p}_n$, denoted as $M_{\tilde{p}_n}(\boldsymbol{t}) = \mathbb{E}_{\tilde{p}_n}[e^{\boldsymbol{t}^T \boldsymbol{\theta}}]$, to be equal to $M_q(\boldsymbol{t})$[3]. As stated in the main text, these can be **relaxed** to matching all the moments $\mathbb{E}_{\tilde{p}_n}[\boldsymbol{\theta}^m] = \mathbb{E}_q[\boldsymbol{\theta}^m]$ for $m \in \mathbb{N}$, and under some smoothness assumption for q, matching all the moments is equivalent to equality constraints. A further crude relaxation suggests moment matching just for a vectorial function $\boldsymbol{\phi}(\boldsymbol{\theta})$, i.e. $\mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}] = \mathbb{E}_q[\boldsymbol{\phi}]$. After relaxation we arrive at the *Bethe free energy* presented in the main text.

A.2 From Bethe to EP: a Dual Form Representation: Derivations

We first provide a note on the derivation of KL duality. Since $\mathrm{KL}[q||p_0]$ is convex in q, a dual form representation of KL goes to

$$\mathrm{KL}[q||p_0] = \max_{\lambda_q(\boldsymbol{\theta})} \langle \lambda_q(\boldsymbol{\theta}), q \rangle - A(q), \tag{15}$$

where $\lambda_q(\boldsymbol{\theta})$ is some arbitrary functional, $A(q)$ represent the dual of KL satistfying $\mathrm{KL}[q||p_0] = A^*(q)$. In the space of distributions the inner product $\langle \lambda_q(\boldsymbol{\theta}), q \rangle$ is defined as the integral $\int q(\boldsymbol{\theta}) \lambda_q(\boldsymbol{\theta}) d\boldsymbol{\theta} = \mathbb{E}_q[\lambda_q(\boldsymbol{\theta})]$. Some calculations also reveal that $A^*(q) = \log \mathbb{E}_{p_0} [\exp[\lambda_q(\boldsymbol{\theta})]]$. Combining all these results returns the KL duality stated in the main text. The equality is achieved when $q(\boldsymbol{\theta}) \propto p_0(\boldsymbol{\theta}) \exp[\lambda_q(\boldsymbol{\theta})]$.

Substitution of the KL duality into (3) yields the transformed energy

$$\mathcal{F}_{\text{Bethe}}(\{\tilde{p}_n\}, q, \lambda_q(\boldsymbol{\theta})) = (1-N)\mathbb{E}_q[\lambda_q(\boldsymbol{\theta})] + (N-1) \log \mathbb{E}_{p_0} [\exp[\lambda_q(\boldsymbol{\theta})]] - \sum_n \mathbb{E}_{\tilde{p}_n} \left[\log \frac{p_0(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})}{\tilde{p}_n(\boldsymbol{\theta})} \right]. \tag{16}$$

Denote $\boldsymbol{\lambda}_{-n}$ as the Lagrange multiplier for moment matching and ν, ν_n for the normalisation constraints of q and $\tilde{p}_n$, respectively. This returns the following Lagrangian

$$\begin{aligned}
\min_{q, \{\tilde{p}_n\}, \lambda_q(\boldsymbol{\theta})} \max_{\{\boldsymbol{\lambda}_{-n}, \nu_n, \nu\}} & \mathcal{F}_{\text{Bethe}}(\{\tilde{p}_n\}, q, \lambda_q(\boldsymbol{\theta})) + \sum_n \boldsymbol{\lambda}_{-n}^T (\mathbb{E}_q[\boldsymbol{\phi}] - \mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}]) \\
& + \sum_n \nu_n \left(\int \tilde{p}_n(\boldsymbol{\theta}) d\boldsymbol{\theta} - 1 \right) + \nu \left(\int q(\boldsymbol{\theta}) d\boldsymbol{\theta} - 1 \right).
\end{aligned} \tag{17}$$

We first compute the fixed points for $\tilde{p}_n$ and ν_n, which returns

$$\tilde{p}_n(\boldsymbol{\theta}) = \frac{1}{Z_n} p_0(\boldsymbol{\theta}) f_n(\boldsymbol{\theta}) \exp \left[\boldsymbol{\lambda}_{-n}^T \boldsymbol{\phi}(\boldsymbol{\theta}) \right], \quad \log Z_n = 1 + \nu_n.$$

[3] Technically speaking $q = \tilde{p}_n$ up to zero measure.

Also Z_n serves as the normalising constant of $\tilde{p}_n$, meaning

$$Z_n = \int p_0(\boldsymbol{\theta}) f_n(\boldsymbol{\theta}) \exp\left[\boldsymbol{\lambda}_{-n}^T \boldsymbol{\phi}(\boldsymbol{\theta})\right] d\boldsymbol{\theta}.$$

Then we evaluate the fixed point condition for q, returning

$$(N-1)\lambda_q(\boldsymbol{\theta}) = \sum_n \boldsymbol{\lambda}_{-n}^T \boldsymbol{\phi}(\boldsymbol{\theta}) + \nu.$$

Substituting $\tilde{p}_n$, ν_n back into (17) and enforcing the above condition, the Lagrangian changes to

$$\min_{q,\lambda_q(\boldsymbol{\theta})} \max_{\{\boldsymbol{\lambda}_{-n},\nu\}} (N-1)\log \mathbb{E}_{p_0}\left[\exp[\lambda_q(\boldsymbol{\theta})]\right] - \sum_n \log Z_n - \nu \quad \text{subject to} (N-1)\lambda_q(\boldsymbol{\theta}) = \sum_n \boldsymbol{\lambda}_{-n}^T \boldsymbol{\phi}(\boldsymbol{\theta}) + \nu. \tag{18}$$

Now the above function does not depend on q, so the optimisation problem of q can be eliminated. We further assume $\lambda_q(\boldsymbol{\theta}) = \boldsymbol{\lambda}_q^T \boldsymbol{\phi}(\boldsymbol{\theta}) + \frac{1}{N-1}\nu$ w.l.o.g. so that the constraint changes to $(N-1)\boldsymbol{\lambda}_q = \sum_n \boldsymbol{\lambda}_{-n}$. Using this new definition, $\log \mathbb{E}_{p_0}\left[\exp[\lambda_q(\boldsymbol{\theta})]\right] = \log \mathbb{E}_{p_0}\left[\exp[\boldsymbol{\lambda}_q^T \boldsymbol{\phi}(\boldsymbol{\theta})]\right] + \frac{1}{N-1}\nu$, which means the Lagrangian now is independent with ν as well, so that we can drop the optimisation problem of it. With all these set-ups we arrive at the *EP energy*

$$\min_{\boldsymbol{\lambda}_q} \max_{\{\boldsymbol{\lambda}_{-n}\}} (N-1)\log \mathbb{E}_{p_0}\left[\exp[\boldsymbol{\lambda}_q^T \boldsymbol{\phi}(\boldsymbol{\theta})]\right] - \sum_n \log Z_n \quad \text{subject to } (N-1)\boldsymbol{\lambda}_q = \sum_n \boldsymbol{\lambda}_{-n}. \tag{19}$$

Finally to obtain the approximate posterior from this dual energy optimisation, we make use of the tightness of the KL duality, and define

$$q(\boldsymbol{\theta}) = \frac{1}{Z_q} p_0(\boldsymbol{\theta}) \exp\left[\boldsymbol{\lambda}_q^T \boldsymbol{\phi}(\boldsymbol{\theta})\right], \quad \log Z_q = \log \mathbb{E}_{p_0}\left[\exp[\boldsymbol{\lambda}_q^T \boldsymbol{\phi}(\boldsymbol{\theta})]\right]. \tag{20}$$

This means by constraint relaxations, we implicitely defined the approximate posterior q to have an exponential family form, with the sufficient statistic $\boldsymbol{\phi}(\boldsymbol{\theta})$ defined by the selection of moments.

A.3 From VFE to Power EP: Derivations

The modified energy function also comes from rearranging terms in VFE:

$$\begin{aligned}
\mathcal{F}_{\text{VFE}}(q) &= \mathbb{E}_q\left[\log q(\boldsymbol{\theta}) - \log p_0(\boldsymbol{\theta}) - \sum_n \log f_n(\boldsymbol{\theta})\right] \\
&= \mathbb{E}_q\left[\log \frac{q(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta})} - \left(\sum_n \frac{1}{\alpha_n}\right) \log \frac{q(\boldsymbol{\theta}) p_0(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta}) q(\boldsymbol{\theta})} - \sum_n \log \left(f_n(\boldsymbol{\theta})^{\alpha_n}\right)^{1/\alpha_n}\right] \\
&= \left(1 - \sum_n \frac{1}{\alpha_n}\right) \text{KL}[q||p_0] - \sum_n \frac{1}{\alpha_n} \mathbb{E}_q\left[\log \frac{p_0(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})^{\alpha_n}}{q(\boldsymbol{\theta})}\right].
\end{aligned} \tag{21}$$

Decoupling to $\tilde{p}_n$ and relaxing the equality constraints to moment maching returns the energy function $\mathcal{F}_{\boldsymbol{\alpha}}(q, \{\tilde{p}_n\})$ (and thus power) EP in the main text. With the KL duality and all the Lagrange multipliers defined accordingly, the Lagrangian becomes

$$\begin{aligned}
\min_{q,\{\tilde{p}_n\},\lambda_q(\boldsymbol{\theta})} \max_{\{\boldsymbol{\lambda}_{-n},\nu_n,\nu\}} & \mathcal{F}_{\boldsymbol{\alpha}}(\{\tilde{p}_n\}, q, \lambda_q(\boldsymbol{\theta})) + \sum_n \frac{1}{\alpha_n} \boldsymbol{\lambda}_{-n}^T (\mathbb{E}_q[\boldsymbol{\phi}] - \mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}]) \\
& + \sum_n \nu_n \left(\int \tilde{p}_n(\boldsymbol{\theta}) d\boldsymbol{\theta} - 1\right) + \nu \left(\int q(\boldsymbol{\theta}) d\boldsymbol{\theta} - 1\right),
\end{aligned} \tag{22}$$

and here we scale the multipliers $\boldsymbol{\lambda}_{-n}$ with $1/\alpha_n$ just to follow the conventions of power EP.

B Mixing Variational Methods for Latent Variable Models

The algorithms we discussed in the main text are not directly applicable to latent variable models, and in this appendix we provide several EP-like recipes for them. As we shall see again, different decoupling and constraint relaxation strategy returns algorithms that have different global and local behaviour.

Assume now the exact posterior becomes $p(\boldsymbol{\theta}, \{\boldsymbol{z}_n\}|\{\boldsymbol{x}_n\}) \propto p_0(\boldsymbol{\theta}) \prod_n p_0(\boldsymbol{z}_n) p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{\theta})$. We note that the algorithms discussed below also applies to models that has *intermediate-level variables*, i.e. those latent variables that are attached to a subset of data. The goal is to approximate the exact posterior of both $\boldsymbol{\theta}$ and $\boldsymbol{z}_n$, and we assume factorised approximation $q(\boldsymbol{\theta}, \{\boldsymbol{z}_n\}) = q(\boldsymbol{\theta}) \prod_n q(\boldsymbol{z}_n|\boldsymbol{x}_n)$. Note that in VI/VB literature the local variational approximation is often denoted as $q(\boldsymbol{z}_n)$. However as at optimum $q(\boldsymbol{z}_n)$ depends on $\boldsymbol{x}_n$, here we explicitly write down this dependence as $q(\boldsymbol{z}_n|\boldsymbol{x}_n)$. This notation is also convenient for *amortised inference* [24, 25, 26] which uses a recognition model (i.e. sharing parameters between $q(\boldsymbol{z}_n)$) to parameterise the local variational distribution.

B.1 Full Power EP/BB-α Treatment

We repeat the term rearranging and argument decoupling procedure as in (7). This returns:

$$\mathcal{F}_{\boldsymbol{\alpha}}(q, \{\tilde{p}_n\}) = \left(1 - \sum_n \frac{1}{\alpha_n}\right) \mathrm{KL}[q(\boldsymbol{\theta})||p_0(\boldsymbol{\theta})] + \sum_n \left(1 - \frac{1}{\alpha_n}\right) \mathrm{KL}[q(\boldsymbol{z}_n|\boldsymbol{x}_n)||p_0(\boldsymbol{z}_n)] \\ - \sum_n \frac{1}{\alpha_n} \mathbb{E}_{\tilde{p}_n}\left[\log \frac{p_0(\boldsymbol{\theta}) p_0(\boldsymbol{z}_n) p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{\theta})^{\alpha_n}}{\tilde{p}_n(\boldsymbol{\theta}, \boldsymbol{z}_n)}\right], \tag{23}$$

subject to $\tilde{p}_n(\boldsymbol{\theta}, \boldsymbol{z}_n) = q(\boldsymbol{\theta}) q(\boldsymbol{z}_n|\boldsymbol{x}_n), \forall n$. The next step is to relax the equality constraint to moment-matching constraints denoted as $\mathbb{E}_{\tilde{p}_n}[\boldsymbol{\Phi}(\boldsymbol{\theta}, \boldsymbol{z}_n)] = \mathbb{E}_q[\boldsymbol{\Phi}(\boldsymbol{\theta}, \boldsymbol{z}_n)]$. The choice of the sufficient statistic $\boldsymbol{\Phi}$ also plays an important role here, and for simplicity we omit the dependence to the observations $\boldsymbol{x}_n$, and assume a *factorised sufficient statistic*, i.e. $\boldsymbol{\Phi}(\boldsymbol{\theta}, \boldsymbol{z}_n) = [\boldsymbol{\phi}(\boldsymbol{\theta}), \boldsymbol{\psi}(\boldsymbol{z}_n)]$. We leave the general case to future work.

Now we proceed to solve the fixed points using similar methods presented in the main text. We still use the KL duality (4) for $q(\boldsymbol{\theta})$, and write that for the latent variables as

$$-\mathrm{KL}[q(\boldsymbol{z}_n|\boldsymbol{x}_n)||p_0(\boldsymbol{z}_n)] = \min_{\eta(\boldsymbol{z}_n, \boldsymbol{x}_n)} -\mathbb{E}_q[\eta(\boldsymbol{z}_n, \boldsymbol{x}_n)] + \log \mathbb{E}_{p_0}\left[\exp[\eta(\boldsymbol{z}_n, \boldsymbol{x}_n)]\right]. \tag{24}$$

To simplify computations we assume $\eta(\boldsymbol{z}_n, \boldsymbol{x}_n) = \boldsymbol{\eta}(\boldsymbol{x}_n)^T \boldsymbol{\psi}(\boldsymbol{z}_n)$ w.l.o.g. Substitution into (23) returns the modified form $\mathcal{F}_{\boldsymbol{\alpha}}(q, \{\tilde{p}_n\}, \boldsymbol{\lambda}_q, \{\boldsymbol{\eta}(\boldsymbol{x}_n)\})$ which is defined in a similar way as (16). Associating Lagrange multiplier $\boldsymbol{\lambda}_{-n}$ and $\boldsymbol{\eta}_n$ for the moment matching constraints of $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ respectively (and also those for normalisation), we have the following Lagrangian

$$\mathcal{F}_{\boldsymbol{\alpha}}(q, \{\tilde{p}_n\}, \boldsymbol{\lambda}_q, \{\boldsymbol{\eta}(\boldsymbol{x}_n)\}) + \sum_n \frac{1}{\alpha_n}\left[\boldsymbol{\lambda}_{-n}^T (\mathbb{E}_q[\boldsymbol{\phi}] - \mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}]) + \boldsymbol{\eta}_n^T (\mathbb{E}_q[\boldsymbol{\psi}] - \mathbb{E}_{\tilde{p}_n}[\boldsymbol{\psi}])\right] + ... \tag{25}$$

where we have omitted the multipliers for the normalisation constraints. Finding the fixed point wrt. $\tilde{p}_n$ returns the following tilted distribution:

$$\tilde{p}_n(\boldsymbol{\theta}, \boldsymbol{z}_n) = \frac{1}{Z_n} p_0(\boldsymbol{\theta}) p_0(\boldsymbol{z}_n) p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{\theta})^{\alpha_n} \exp\left[\boldsymbol{\lambda}_{-n}^T \boldsymbol{\phi}(\boldsymbol{\theta}) + \boldsymbol{\eta}_n^T \boldsymbol{\psi}(\boldsymbol{z}_n)\right]. \tag{26}$$

Also zeroing the gradient wrt. q yields the fixed point conditions (up to a constant) $\left(\sum_n \frac{1}{\alpha_n} - 1\right) \boldsymbol{\lambda}_q = \sum_n \frac{1}{\alpha_n} \boldsymbol{\lambda}_{-n}$ just like in power EP and $(1 - \alpha_n)\boldsymbol{\eta}(\boldsymbol{x}_n) = \boldsymbol{\eta}_n$. The second one is similar to the BB-α case so we directly assume it holds and drop the optimisation problem of $\boldsymbol{\eta}_n$. Furthermore, substituting $\tilde{p}_n$ back into (25) and zeroing the gradient wrt. q, we arrive at the following power EP energy

$$(\sum_n \frac{1}{\alpha_n} - 1) \log \int p_0(\boldsymbol{\theta}) \exp\left[\boldsymbol{\lambda}_q^T \boldsymbol{\phi}(\boldsymbol{\theta})\right] d\boldsymbol{\theta} + \sum_n (\frac{1}{\alpha_n} - 1) \log \int p_0(\boldsymbol{z}_n) \exp\left[\boldsymbol{\eta}(\boldsymbol{x}_n)^T \boldsymbol{\psi}(\boldsymbol{z}_n)\right] d\boldsymbol{z}_n \\ - \sum_n \frac{1}{\alpha_n} \log \int p_0(\boldsymbol{\theta}) p_0(\boldsymbol{z}_n) p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{\theta})^{\alpha_n} \exp\left[\boldsymbol{\lambda}_{-n}^T \boldsymbol{\phi}(\boldsymbol{\theta}) + (1 - \alpha_n)\boldsymbol{\eta}(\boldsymbol{x}_n)^T \boldsymbol{\psi}(\boldsymbol{z}_n)\right] d\boldsymbol{\theta} d\boldsymbol{z}_n \\ \text{subject to } (\sum_n \frac{1}{\alpha_n} - 1)\boldsymbol{\lambda}_q = \sum_n \frac{1}{\alpha_n} \boldsymbol{\lambda}_{-n}. \tag{27}$$

To make the KL duality tight we define the approximation q obtained from the dual energy optimisation as

$$q(\boldsymbol{\theta}) = \frac{1}{Z_q} p_0(\boldsymbol{\theta}) \exp\left[\boldsymbol{\lambda}_q^T \boldsymbol{\phi}(\boldsymbol{\theta})\right], \quad q(\boldsymbol{z}_n|\boldsymbol{x}_n) = \frac{1}{Z_q(\boldsymbol{x}_n)} p_0(\boldsymbol{z}_n) \exp\left[\boldsymbol{\eta}(\boldsymbol{x}_n)^T \boldsymbol{\psi}(\boldsymbol{z}_n)\right].$$

We also present a much cleaner version of the energy function by substituting the q distributions into the power EP energy (with a further definition $\boldsymbol{\lambda}_n = \frac{1}{\alpha_n}(\boldsymbol{\lambda}_q - \boldsymbol{\lambda}_{-n})$) and rearranging terms:

$$-\log \int p_0(\boldsymbol{\theta}) \exp\left[\boldsymbol{\lambda}_q^T \boldsymbol{\phi}(\boldsymbol{\theta})\right] d\boldsymbol{\theta} - \sum_n \frac{1}{\alpha_n} \log \mathbb{E}_{q(\boldsymbol{\theta}) q(\boldsymbol{z}_n|\boldsymbol{x}_n)}\left[\left(\frac{p_0(\boldsymbol{z}_n) p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{\theta})}{q(\boldsymbol{z}_n|\boldsymbol{x}_n) \exp\left[\boldsymbol{\lambda}_n^T \boldsymbol{\phi}(\boldsymbol{\theta})\right]}\right)^{\alpha_n}\right] \\ \text{subject to } \boldsymbol{\lambda}_q = \sum_n \boldsymbol{\lambda}_n. \tag{28}$$

This term rearranging proce The BB-α variant can be derived similarly by further constraint relaxation. The detailed derivation is omitted here, but in summary we keep the constraint for $\boldsymbol{\psi}$ but modify the other constraint by $\mathbb{E}_{q(\boldsymbol{\theta})}[\boldsymbol{\phi}] = \frac{1}{N} \sum_n \mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}]$. One can show that the dual energy becomes (again after rearranging terms)

$$-\sum_n \frac{1}{\alpha} \log \mathbb{E}_{q(\boldsymbol{\theta}) q(\boldsymbol{z}_n|\boldsymbol{x}_n)}\left[\left(\frac{p_0(\boldsymbol{\theta})^{\frac{1}{N}} p_0(\boldsymbol{z}_n) p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{\theta})}{q(\boldsymbol{\theta})^{\frac{1}{N}} q(\boldsymbol{z}_n|\boldsymbol{x}_n)}\right)^{\alpha}\right], \tag{29}$$

with $q(\boldsymbol{\theta})$ and $q(\boldsymbol{z}_n|\boldsymbol{x}_n)$ defined as exponential family distributions. Extensions to general q distributions can be justified using Rényi divergences minimisation methods [20].

B.2 Nesting Power-EP/BB-α in VI

This section discusses how to use power EP/BB-α as an inner loop computation for variational inference. It starts from (23), but only applies constraint relaxations to the moments for the latent variables. In other words, now the constraints are $\mathbb{E}_q[\boldsymbol{\psi}(\boldsymbol{z}_n)] = \mathbb{E}_{\tilde{p}_n}[\boldsymbol{\psi}(\boldsymbol{z}_n)]$ and $q(\boldsymbol{\theta}) = \tilde{p}_n(\boldsymbol{\theta})$. For computational convenience we assume $\tilde{p}_n(\boldsymbol{\theta}, \boldsymbol{z}_n) = q(\boldsymbol{\theta})\tilde{p}_n(\boldsymbol{z}_n|\boldsymbol{\theta})$ w.l.o.g. so that there is only one set of constraints $\mathbb{E}_{q(\boldsymbol{z}_n|\boldsymbol{x}_n)}[\boldsymbol{\psi}(\boldsymbol{z}_n)] = \mathbb{E}_{q(\boldsymbol{\theta})\tilde{p}_n(\boldsymbol{z}_n|\boldsymbol{\theta})}[\boldsymbol{\psi}(\boldsymbol{z}_n)]$ (besides the normalisation ones). Denote $\boldsymbol{\eta}_n$ as the associated Lagrange multipliers for the moment maching constraints, then solving the corresponding Lagrangian yields $(1-\alpha_n)\boldsymbol{\eta}(\boldsymbol{x}_n) = \boldsymbol{\eta}_n$ again, and

$$\tilde{p}_n(\boldsymbol{z}_n|\boldsymbol{\theta}) = \frac{1}{Z_n(\boldsymbol{x}_n, \boldsymbol{\theta})} p_0(\boldsymbol{z}_n) p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{\theta})^{\alpha_n} \exp\left[(1-\alpha_n)\boldsymbol{\eta}(\boldsymbol{x}_n)^T \boldsymbol{\psi}(\boldsymbol{z}_n)\right]. \tag{30}$$

Note that now the normalising constant $Z_n(\boldsymbol{x}_n, \boldsymbol{\theta})$ is also a function of $\boldsymbol{\theta}$. We still keep the free-form optimisation for $q(\boldsymbol{\theta})$. Substitution of $\tilde{p}_n(\boldsymbol{z}_n|\boldsymbol{\theta})$ into the Lagrangian returns a "mixed form" of energy (again after rearranging terms)

$$\min_q \mathrm{KL}[q(\boldsymbol{\theta})||p_0(\boldsymbol{\theta})] - \sum_n \mathbb{E}_{q(\boldsymbol{\theta})}\left[\frac{1}{\alpha_n}\log \mathbb{E}_{q(\boldsymbol{z}_n|\boldsymbol{x}_n)}\left[\left(\frac{p_0(\boldsymbol{z}_n)p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{\theta})}{q(\boldsymbol{z}_n|\boldsymbol{x}_n)}\right)^{\alpha_n}\right]\right], \tag{31}$$

where $q(\boldsymbol{z}_n|\boldsymbol{x}_n)$ is restricted to have an exponential family form $q(\boldsymbol{z}_n|\boldsymbol{x}_n) \propto p_0(\boldsymbol{z}_n)\exp\left[\boldsymbol{\eta}(\boldsymbol{x}_n)^T\boldsymbol{\psi}(\boldsymbol{z}_n)\right]$.

Remark. A naive change of the constraints to $\mathbb{E}_q[\boldsymbol{\phi}(\boldsymbol{\theta})] = \mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}(\boldsymbol{\theta})]$ and $q(\boldsymbol{z}_n|\boldsymbol{x}_n) = \tilde{p}_n(\boldsymbol{z}_n)$ does *not* lead to a nested approach of VI in Power-EP/BB-α. We omit the details here, but a try-out for the BB-α variant returns the following energy

$$\min_q \sum_n \mathrm{KL}[q(\boldsymbol{z}_n|\boldsymbol{x}_n)||p_0(\boldsymbol{z}_n)] - \sum_n \frac{1}{\alpha}\mathbb{E}_{q(\boldsymbol{z}_n|\boldsymbol{x}_n)}\log \mathbb{E}_{q(\boldsymbol{\theta})}\left[\left(\frac{p_0(\boldsymbol{\theta})^{1/N}p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{\theta})}{q(\boldsymbol{\theta})^{1/N}}\right)^{\alpha}\right], \tag{32}$$

with $q(\boldsymbol{\theta})$ also restricted as an exponential family distribution.

B.3 Nesting VI in Power-EP/BB-α

Recall in the beginning we mentioned that the decoupling process plays a crucial role on the dual form of the energy. Now we decouple the variational distributions in a different way:

$$\mathcal{F}_{\boldsymbol{\alpha}}(q, \{\tilde{p}_n\}) = \left(1 - \sum_n \frac{1}{\alpha_n}\right)\mathrm{KL}[q(\boldsymbol{\theta})||p_0(\boldsymbol{\theta})] - \sum_n \frac{1}{\alpha_n}\mathbb{E}_{\tilde{p}_n(\boldsymbol{\theta})q(\boldsymbol{z}_n|\boldsymbol{x}_n)}\left[\log \frac{p_0(\boldsymbol{\theta})p_0(\boldsymbol{z}_n)^{\alpha_n}p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{\theta})^{\alpha_n}}{\tilde{p}_n(\boldsymbol{\theta})q(\boldsymbol{z}_n|\boldsymbol{x}_n)^{\alpha_n}}\right], \tag{33}$$

In this case we only relax the constraints for latent variable posterior approximations to $\tilde{p}_n(\boldsymbol{\theta})$ to moment matching: $\mathbb{E}_q[\boldsymbol{\phi}] = \mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}]$. We also reuse the derivations in section 1.3 of the main text by noticing now

$$\log f_n(\boldsymbol{\theta}) = \mathbb{E}_{q(\boldsymbol{z}_n|\boldsymbol{x}_n)}\left[\log \frac{p_0(\boldsymbol{z}_n)p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{\theta})}{q(\boldsymbol{z}_n|\boldsymbol{x}_n)}\right], \tag{34}$$

and thus omit the detailed expression of the energy functions. Readers are referred to (8) in the main text. This algorithm (with the power EP variant) returns the same stationary points as VI if the model is conjugate, i.e. $\log p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{\theta})$ is linear in $\boldsymbol{\phi}(\boldsymbol{\theta})$.

B.4 Variational Message Passing between Tilted Distributions

Applying similar derivation as in B.3 to the decoupling 23 returns a slightly different algorithm that applies variational message passing (VMP) [27] to the tilted distributions. Here we also directly enforce the factorisation assumption, i.e. $\tilde{p}(\boldsymbol{\theta}, \boldsymbol{z}_n) = \tilde{p}_n(\boldsymbol{\theta})\tilde{p}_n(\boldsymbol{z}_n)$ subject to $\tilde{p}_n(\boldsymbol{\theta}) = q(\boldsymbol{\theta})$ and $\tilde{p}_n(\boldsymbol{z}_n) = q(\boldsymbol{z}_n|\boldsymbol{x}_n)$. This returns the same fixed point as with the joint tilted distribution version if we optimise the free energy under equality constraints. However the stationary points differs from those derived above when relaxing the constraint to moment matching. This is because, as $\tilde{p}$ factorises, the fixed point solution of the Lagrangian should contain "messages" sent between $\tilde{p}_n(\boldsymbol{\theta})$ and $\tilde{p}(\boldsymbol{z}_n)$. To be precise, we solve the fixed point of the Lagrangian (25) (but with factorised $\tilde{p}$). The fixed point conditions for q remain the same, but those for $\tilde{p}$ become

$$\tilde{p}_n(\boldsymbol{\theta}) = \frac{1}{Z_n}p_0(\boldsymbol{\theta})\exp\left[\boldsymbol{\lambda}_{-n}^T\boldsymbol{\phi}(\boldsymbol{\theta}) + \alpha_n\mathbb{E}_{\tilde{p}(\boldsymbol{z}_n)}[\log p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{\theta})]\right], \tag{35}$$

$$\tilde{p}_n(\boldsymbol{z}_n) = \frac{1}{Z_n(\boldsymbol{x}_n)}p_0(\boldsymbol{z}_n)\exp\left[\boldsymbol{\eta}_n^T\boldsymbol{\psi}(\boldsymbol{z}_n) + \alpha_n\mathbb{E}_{\tilde{p}_n(\boldsymbol{\theta})}[\log p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{\theta})]\right]. \tag{36}$$

Substituting these new fixed point equations back to the Lagrangian returns a different dual energy function

$$\begin{aligned}(\sum_n \frac{1}{\alpha_n} - 1)\log Z_q + \sum_n (\frac{1}{\alpha_n} - 1)\log Z_q(\boldsymbol{x}_n) - \sum_n \frac{1}{\alpha_n}\left(\log Z_n + \log Z_n(\boldsymbol{x}_n)\right) \\ + \sum_n \mathbb{E}_{\tilde{p}_n(\boldsymbol{\theta})\tilde{p}_n(\boldsymbol{z}_n)}\left[\log p(\boldsymbol{x}_n|\boldsymbol{z}_n,\boldsymbol{\theta})\right] \\ \text{subject to } (\sum_n \frac{1}{\alpha_n} - 1)\boldsymbol{\lambda}_q = \sum_n \frac{1}{\alpha_n}\boldsymbol{\lambda}_{-n}.\end{aligned} \tag{37}$$

The BB-α version can also be derived similarly which we omit here.

The proposed algorithm can be extended to $\alpha_n = 1$, which returns the results in section B.3 and subject to the discussions there. Otherwise the algorithm behaves differently since 1) the local messages are computed using the tilted distributions and 2) for non-conjugate models the tilted distributions contain more complex structure compared to q. In below we provide a fixed point iteration procedure to find the stationary distributions.

1 Select a datapoint $\boldsymbol{x}_n$;

2 Compute cavity distribution $q_{-n}(\boldsymbol{\theta})$ (either using power EP or BB-α);

3 Compute cavity distribution $q_{-1}(\boldsymbol{z}_n)$ when $\alpha_n \neq 1$, otherwise $q_{-1}(\boldsymbol{z}_n) = p_0(\boldsymbol{z}_n)$;

4 Run VMP/VI on this single datapoint $\boldsymbol{x}_n$ with prior terms changed to $q_{-n}(\boldsymbol{\theta})$ and $q_{-1}(\boldsymbol{z}_n)$. This procedure calculates $\tilde{p}_n(\boldsymbol{\theta})$ and $\tilde{p}(\boldsymbol{z}_n)$;

5 Compute the moment matching update $q_{new}(\boldsymbol{\theta}) \rightarrow \text{proj}[\tilde{p}_n(\boldsymbol{\theta})]$;

6 Compute the final update for q with q_{new} (either using power EP or BB-α).

C A Note on the Decoupling Procedure

In the derivations so far, each likelihood term $p(\boldsymbol{x}_n|\boldsymbol{\theta})$ (or $p(\boldsymbol{z}_n|\boldsymbol{x}_n,\boldsymbol{\theta})$) is associated to only one tilted distribution $\tilde{p}_n$. This is not necessary, and in general $\log p(\boldsymbol{x}_n|\boldsymbol{\theta})$ can be split in any arbitrary way. In other words, the decoupling procedure depends on the way a factor graph is defined, which in general can be specified in any arbitrary way (as long as the function represented by that factor graph remains unchanged). For example, in distributed variational methods we assumed the subsets are disjoint to each other, i.e. $N_i \cap N_j = \emptyset$. A relaxed form for this assumption would associate a vector $\boldsymbol{\mu}_k = (\mu_{k1}, \mu_{k2}, ..., \mu_{kN})$ to each of the K groups (recall N is the total number of datapoints), where the subset-level factor F_k is constructed as $\log F_k = \sum_n \mu_{kn} \log f_n$. The set of these vectors $\{\boldsymbol{\mu}_k\}$ should satisfy $\sum_k \mu_{kn} = 1, \forall n$.