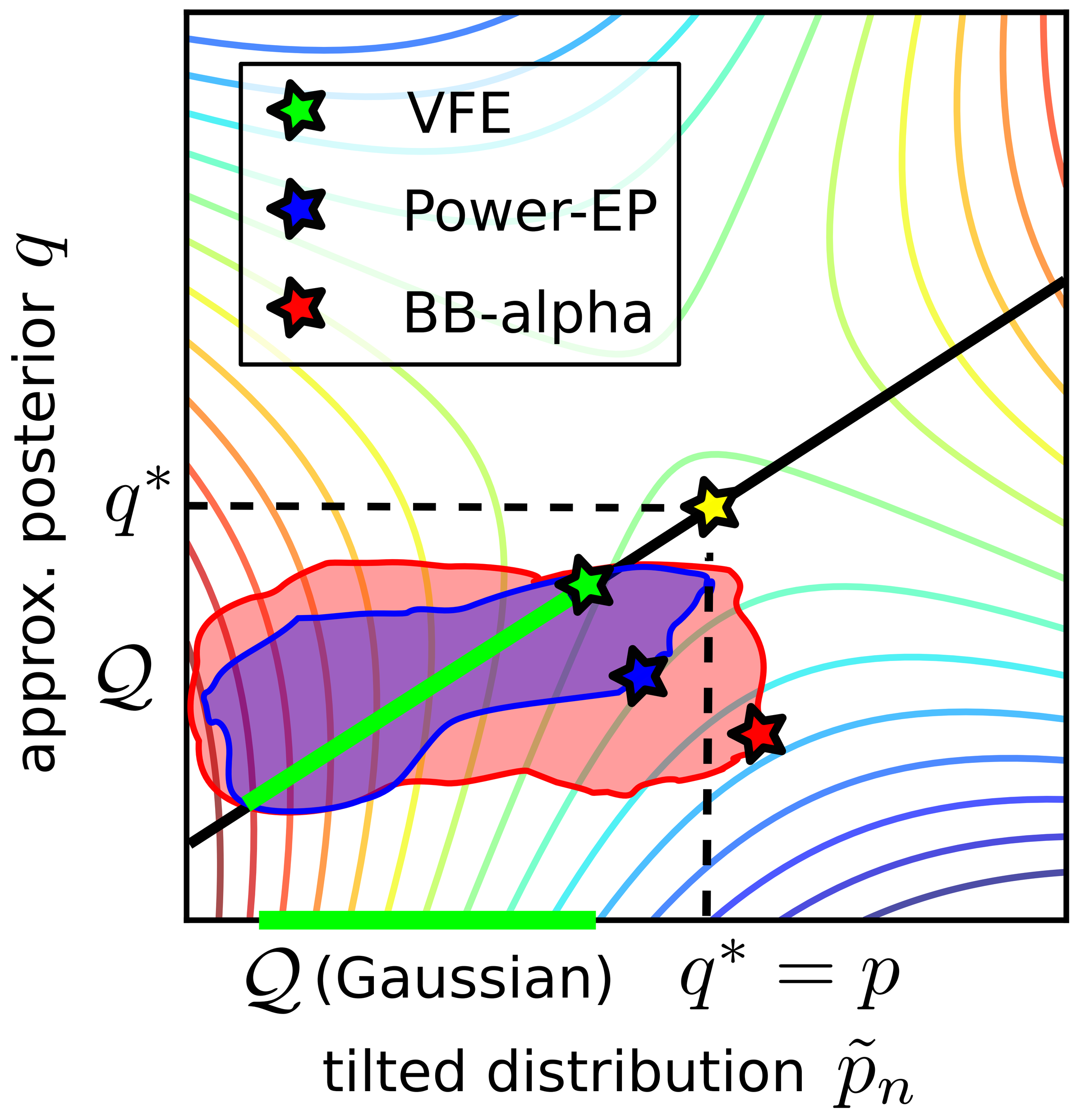
A Unifying Approximate Inference Framework from Variational Free Energy Relaxation

Yingzhen Li and Richard E. Turner, University of Cambridge

Big picture: Approximate inference as constraint relaxation



- Approximate posterior is obtained by solving a constrainted minimisation problem with the following energy (with $\sum_n \frac{1}{\alpha_n} \neq 1$):

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

- **VFE constraints:** $\tilde{p}_n = q, \forall n$; ($\mathcal{F}(q, \{\tilde{p}_n\})$ simplified to $\mathcal{F}_{\text{VFE}}(q)$)
- **Power-EP constraints:** $\mathbb{E}_q[\boldsymbol{\phi}(\boldsymbol{\theta})] = \mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}(\boldsymbol{\theta})], \forall n$; (moment matching)
- **(*NEW*) BB-α constraints:** $N\mathbb{E}_q[\boldsymbol{\phi}(\boldsymbol{\theta})] = \sum_n \mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}(\boldsymbol{\theta})]$; (moment averaging)
- **(*NEW*)** Mixing distributed EP and BB-α;
- **(*NEW*)** Extenstions to latent variable models.

From VFE to Power EP

- Target distribution $p(\boldsymbol{\theta}) \propto p_0(\boldsymbol{\theta}) \prod_n f_n(\boldsymbol{\theta})$, e,g, $f_n(\boldsymbol{\theta}) = p(\boldsymbol{x}_n|\boldsymbol{\theta})$;
- VI minimises the **variational free energy** (VFE):

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

- An equivalent optimisation problem, **subject to** $\tilde{p}_n = q, \forall n$:

$$\min_{q, \{\tilde{p}_n\}} \left(1 - \sum_n \frac{1}{\alpha_n}\right) \mathrm{KL}[q||p_0] - \sum_n \frac{1}{\alpha_n} \mathbb{E}_{\tilde{p}_n}\left[\log \frac{p_0(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})^{\alpha_n}}{\tilde{p}_n(\boldsymbol{\theta})}\right].$$

- **Constraint relaxation:** from $q = \tilde{p}_n, \forall n$ to **moment matching**:

$$\boxed{\mathbb{E}_q[\boldsymbol{\phi}(\boldsymbol{\theta})] = \mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}(\boldsymbol{\theta})], \forall n.}$$

- Introduce a new variable by using the KL duality:

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

- Write $\boldsymbol{\lambda}_{-n}$ as the Lagrange multiplier for moment matching constraints, and solve the Lagrangian:

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

$$\left(\sum_n \frac{1}{\alpha_n} - 1\right) \lambda_q(\boldsymbol{\theta}) = \sum_n \frac{1}{\alpha_n} \boldsymbol{\lambda}_{-n}^T \boldsymbol{\phi}(\boldsymbol{\theta}) + \text{const}.$$

- Defining $\lambda_q(\boldsymbol{\theta}) = \boldsymbol{\lambda}_q^T \boldsymbol{\phi}(\boldsymbol{\theta}) + \text{const}$ and substituting in the fixed point solutions, we arrive the **power-EP (dual) energy**:

$$\min_{\boldsymbol{\lambda}_q} \max_{\{\boldsymbol{\lambda}_{-n}\}} \left(\sum_n \frac{1}{\alpha_n} - 1\right) \log Z_q - \sum_n \frac{1}{\alpha_n} \log Z_n,$$

$$\text{subject to } \left(\textstyle\sum_n \frac{1}{\alpha_n} - 1\right) \boldsymbol{\lambda}_q = \textstyle\sum_n \frac{1}{\alpha_n} \boldsymbol{\lambda}_{-n}.$$

- Approximation: $q(\boldsymbol{\theta}) = \frac{1}{Z_q} p_0(\boldsymbol{\theta}) \exp\left[\boldsymbol{\lambda}_q^T \boldsymbol{\phi}(\boldsymbol{\theta})\right]$.
- **Local factor parameterisation returns power EP:** define $\boldsymbol{\lambda}_n = (\boldsymbol{\lambda}_q - \boldsymbol{\lambda}_{-n})/\alpha_n$, then rewrite $\boldsymbol{\lambda}_q = \sum_n \boldsymbol{\lambda}_n$ and $\boldsymbol{\lambda}_{-n} = \boldsymbol{\lambda}_q - \alpha_n \boldsymbol{\lambda}_n$. Now $f_n(\boldsymbol{\theta}) \approx \exp\left[\boldsymbol{\lambda}_n^T \boldsymbol{\phi}(\boldsymbol{\theta})\right]$.

BB-α & Distributed Algorithms

Deriving black-box alpha:

- Power-EP proposes N sets of constraints (main reason for memory overhead);
- Idea: reduce to **weighted moment averaging**: $\mathbb{E}_q[\phi] = \sum_n w_n \mathbb{E}_{\tilde{p}_n}[\phi]$, $\sum_n w_n = 1$;
- Choose $\alpha_n = \alpha$, $w_n = 1/N$ and solve the Lagrangian again, we arrive at the **BB-α (dual) energy**:

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

Distributed Power-EP algorithms:

- In this case factor indices are divided into subsets $N_1, N_2, ..., N_K$
- Rewrite $p(\boldsymbol{\theta}) \propto p_0(\boldsymbol{\theta}) \prod_k F_k(\boldsymbol{\theta})$, $F_k(\boldsymbol{\theta}) = \prod_{n_k \in N_k} f_{n_k}(\boldsymbol{\theta})$,
- ...and repeat the same procedure!
- Alternatively, add extra constraints $\tilde{p}_i = \tilde{p}_j \forall i, j \in N_k$.

Mixing BB-α and distributed methods:

- Distributed BB-α: let $\mathbb{E}_q[\boldsymbol{\phi}(\boldsymbol{\theta})] = \frac{1}{N} \sum_k \mathbb{E}_{\tilde{p}_k}[\boldsymbol{\phi}(\boldsymbol{\theta})]$.
- Nesting BB-α in distributed EP: let $\mathbb{E}_q[\boldsymbol{\phi}(\boldsymbol{\theta})] = \frac{1}{|N_k|} \sum_{n_k \in N_k} \mathbb{E}_{\tilde{p}_{n_k}}[\boldsymbol{\phi}(\boldsymbol{\theta})]$.

Extension: Latent Variable Models

- Assume factorised approximation $q(\boldsymbol{\theta}, \boldsymbol{z}_n) = q(\boldsymbol{\theta}) \prod_n q(\boldsymbol{z}_n)$:

$$\mathcal{F}_{\text{VFE}}(q) = \mathbb{E}_q\left[\log \frac{q(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta})} + \sum_n \log \frac{q(\boldsymbol{z}_n)}{p_0(\boldsymbol{z}_n)} - \sum_{n=1}^N \log f_n(\boldsymbol{\theta}, \boldsymbol{z}_n)\right].$$

- Decouple q to $\tilde{p}_n$ similarly, and
- ...VI considers constraints $\tilde{p}_n(\boldsymbol{\theta}, \boldsymbol{z}_n) = q(\boldsymbol{\theta}) q(\boldsymbol{z}_n), \forall n$;
- Different constraint relaxations return different algorithms!
  - Full EP: $\mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}(\boldsymbol{\theta}), \boldsymbol{\psi}(\boldsymbol{z}_n)] = \mathbb{E}_q[\boldsymbol{\phi}(\boldsymbol{\theta}), \boldsymbol{\psi}(\boldsymbol{z}_n)]$;
  - Nesting EP in VI: $\mathbb{E}_{\tilde{p}_n}[\boldsymbol{\psi}(\boldsymbol{z}_n)] = \mathbb{E}_q[\boldsymbol{\psi}(\boldsymbol{z}_n)]$, $q(\boldsymbol{\theta}) = \tilde{p}_n(\boldsymbol{\theta})$;
  - Nesting VI in EP: $\mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}(\boldsymbol{\theta})] = \mathbb{E}_q[\boldsymbol{\phi}(\boldsymbol{\theta})]$, $\tilde{p}_n(\boldsymbol{z}_n) = q(\boldsymbol{z}_n)$, $\tilde{p}_n(\boldsymbol{\theta}, \boldsymbol{z}_n) = \tilde{p}_n(\boldsymbol{\theta}) \tilde{p}_n(\boldsymbol{z}_n)$;
  - "Tilted" VMP: $\mathbb{E}_{\tilde{p}_n}[\boldsymbol{\phi}(\boldsymbol{\theta}), \boldsymbol{\psi}(\boldsymbol{z}_n)] = \mathbb{E}_q[\boldsymbol{\phi}(\boldsymbol{\theta}), \boldsymbol{\psi}(\boldsymbol{z}_n)]$, $\tilde{p}_n(\boldsymbol{\theta}, \boldsymbol{z}_n) = \tilde{p}_n(\boldsymbol{\theta}) \tilde{p}_n(\boldsymbol{z}_n)$.