Appendix

1 The duality of relaxed EP energy functions

The primary energy function of relaxed EP is

$$\min_{\boldsymbol{\eta}_i}\min_{\hat{p}_i}\max_{q}\sum_i \frac{1}{\hat{Z}_i}\int_{\mathbf{w}} \hat{p}_i(\mathbf{w})r_i(\mathbf{w})\log\frac{\hat{p}_i(\mathbf{w})}{\hat{Z}_i t_i(\mathbf{w})p(\mathbf{w})} - (n-1)\frac{1}{Z_q}\int_{\mathbf{w}} q(\mathbf{w})r_i(\mathbf{w})\log\frac{q(\mathbf{w})}{Z_q p(\mathbf{w})} + c\sum_i |\boldsymbol{\eta}_i|_1 \quad (1)$$

subject to

$$\frac{1}{\hat{Z}_i}\int_{\mathbf{w}} \phi(\mathbf{w})\hat{p}_i(\mathbf{w})r_i(\mathbf{w})d\mathbf{w} = \frac{1}{Z_q}\int_{\mathbf{w}} \phi(\mathbf{w})q(\mathbf{w})r_i(\mathbf{w})d\mathbf{w} \quad (2)$$

$$\int_{\mathbf{w}} \hat{p}_i(\mathbf{w})d\mathbf{w} = 1 \quad (3)$$

$$\int_{\mathbf{w}} q(\mathbf{w})d\mathbf{w} = 1 \quad (4)$$

$$\hat{Z}_i = \int_{\mathbf{w}} \hat{p}_i(\mathbf{w})r_i(\mathbf{w})d\mathbf{w} \quad (5)$$

$$Z_q = \int_{\mathbf{w}} q(\mathbf{w})r_i(\mathbf{w})d\mathbf{w} \quad (6)$$

$$r_i(\mathbf{w}) \propto \exp(\boldsymbol{\eta}_i^T\phi(\mathbf{w})) \quad (7)$$

where c is the constant and r_i is the relaxation factor.

Based on the KL duality bound, we obtain the dual energy function.

$$\begin{aligned}&\min_{\boldsymbol{\eta}}\min_{\nu}\max_{\lambda}(n-1)\log\int_{\mathbf{w}} p(\mathbf{w})\exp(\boldsymbol{\nu}^T\phi(\mathbf{w}) + \boldsymbol{\eta}_i^T\phi(\mathbf{w}))d\mathbf{w}\\ &-\sum_{i=1}^{n}\log\int_{\mathbf{w}} t_i(\mathbf{w})p(\mathbf{w})\exp(\boldsymbol{\lambda}_i^T\phi(\mathbf{w}) + \boldsymbol{\eta}_i^T\phi(\mathbf{w}))d\mathbf{w} + c\sum_i|\boldsymbol{\eta}_i|_1 \quad (8)\\ &(n-1)\boldsymbol{\nu} = \sum_i \boldsymbol{\lambda}_i \quad (9)\end{aligned}$$

Setting the gradient of the above function to zero gives us the fixed-point updates described in the Section 3 of the main text. The fixed-point updates, however, do not guarantee convergence. But because of the relaxed KL minimization, REP always converges in our experiments (while EP can diverge when given many outliers).

Now we prove the duality of the relaxed EP energy function. Applying the KL duality to the first term in (1)produces

$$\begin{aligned}&\frac{1}{\hat{Z}_i}\int_{\mathbf{w}} \hat{p}_i(\mathbf{w})r_i(\mathbf{w})\log\frac{\hat{p}_i(\mathbf{w})}{\hat{Z}_i t_i(\mathbf{w})p(\mathbf{w})} \quad (10)\\ =\;& \frac{1}{\hat{Z}_i}\int_{\mathbf{w}} \hat{p}_i(\mathbf{w})r_i(\mathbf{w})\log\frac{\hat{p}_i(\mathbf{w})r_i(\mathbf{w})}{\hat{Z}_i t_i(\mathbf{w})p(\mathbf{w})r_i(\mathbf{w})}\\ =\;& \max_{\lambda}\frac{1}{\hat{Z}_i}\int_{\mathbf{w}} \hat{p}_i(\mathbf{w})r_i(\mathbf{w})\boldsymbol{\lambda}_i(\mathbf{w})d\mathbf{w} - \log\int_{\mathbf{w}} t_i(\mathbf{w})p(\mathbf{w})r_i(\mathbf{w})\exp(\boldsymbol{\lambda}_i(\mathbf{w}))d\mathbf{w}\end{aligned}$$

This is because the maximum of the right side of (10) is achieved when (taking derivative to $\boldsymbol{\lambda}_i(\mathbf{w})$)

$$\frac{1}{\hat{Z}_i}\hat{p}_i(\mathbf{w})r_i(\mathbf{w}) - \frac{t_i(\mathbf{w})p(\mathbf{w})r_i(\mathbf{w})\exp(\boldsymbol{\lambda}_i(\mathbf{w}))}{\int_{\mathbf{w}} t_i(\mathbf{w})p(\mathbf{w})r_i(\mathbf{w})\exp(\boldsymbol{\lambda}_i(\mathbf{w}))d\mathbf{w}} = 0 \quad (11)$$

which means

$$\exp(\boldsymbol{\lambda}_i(\mathbf{w})) = \frac{\hat{p}_i(\mathbf{w})r_i(\mathbf{w})\int_{\mathbf{w}} t_i(\mathbf{w})p(\mathbf{w})r_i(\mathbf{w})\exp(\boldsymbol{\lambda}_i(\mathbf{w}))d\mathbf{w}}{\hat{Z}_i t_i(\mathbf{w})p(\mathbf{w})r_i(\mathbf{w})} \tag{12}$$

Inserting $\exp(\boldsymbol{\lambda}_i(\mathbf{w}))$ in (10) proves the KL duality for (10).
And from the stationary condition, we can assume w.l.o.g. that

$$\boldsymbol{\lambda}_i(\mathbf{w}) = \boldsymbol{\lambda}_i^T \phi(\mathbf{w}) \tag{13}$$

$$\begin{aligned}
&\frac{1}{\hat{Z}_i}\int_{\mathbf{w}} \hat{p}_i(\mathbf{w})r_i(\mathbf{w})\log\frac{\hat{p}_i(\mathbf{w})}{t_i(\mathbf{w})p(\mathbf{w})} &(14)\\
=&\max_{\lambda}\frac{1}{\hat{Z}_i}\int_{\mathbf{w}} \hat{p}_i(\mathbf{w})r_i(\mathbf{w})\boldsymbol{\lambda}_i^T\phi(\mathbf{w})d\mathbf{w} - \log\int_{\mathbf{w}} t_i(\mathbf{w})p(\mathbf{w})r_i(\mathbf{w})\exp(\boldsymbol{\lambda}_i^T\phi(\mathbf{w}))d\mathbf{w}
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&-\frac{1}{Z_q}\int_{\mathbf{w}} q(\mathbf{w})r_i(\mathbf{w})\log\frac{q(\mathbf{w})}{Z_q p(\mathbf{w})} &(15)\\
=&-\frac{1}{Z_q}\int_{\mathbf{w}} q(\mathbf{w})r_i(\mathbf{w})\log\frac{q(\mathbf{w})r_i(\mathbf{w})}{Z_q p(\mathbf{w})r_i(\mathbf{w})}\\
=&\min_{\boldsymbol{\nu}} -\frac{1}{Z_q}\int_{\mathbf{w}} \boldsymbol{\nu}(\mathbf{w})q(\mathbf{w})r_i(\mathbf{w})d\mathbf{w} + \log\int_{\mathbf{w}} p(\mathbf{w})r_i(\mathbf{w})\exp(\boldsymbol{\nu}(\mathbf{w}))d\mathbf{w}\\
=&\min_{\boldsymbol{\nu}} -\frac{1}{Z_q}\int_{\mathbf{w}} \boldsymbol{\nu}^T\phi(\mathbf{w})q(\mathbf{w})r_i(\mathbf{w})d\mathbf{w} + \log\int_{\mathbf{w}} p(\mathbf{w})r_i(\mathbf{w})\exp(\boldsymbol{\nu}^T\phi(\mathbf{w}))d\mathbf{w}
\end{aligned}$$

With the constraint ($(n-1)\boldsymbol{\nu} = \sum_i \boldsymbol{\lambda}_i$) and (2), we obtain the dual energy function:

$$\begin{aligned}
&\min_{\boldsymbol{\eta}}\min_{\boldsymbol{\nu}}\max_{\boldsymbol{\lambda}}(n-1)\log\int_{\mathbf{w}} p(\mathbf{w})r_i(\mathbf{w})\exp(\boldsymbol{\nu}^T\phi(\mathbf{w}))d\mathbf{w}\\
&-\sum_{i=1}^{n}\log\int_{\mathbf{w}} t_i(\mathbf{w})p(\mathbf{w})r_i(\mathbf{w})\exp(\boldsymbol{\lambda}_i^T\phi(\mathbf{w}))d\mathbf{w} + c\sum_i|\boldsymbol{\eta}_i|_1 &(16)\\
&(n-1)\boldsymbol{\nu} = \sum_i \boldsymbol{\lambda}_i &(17)
\end{aligned}$$

2 Choices of relaxation factors and the energy function

As discussed in Section 4.2 of the main text, we have multiple choices for the form of the relaxation factors. To save the computational cost, we can parameterize $\boldsymbol{\eta}_i$ (the natural parameter of the relaxation message) in a constrained form. For example, we can make $\boldsymbol{\eta}_i$ to be a scaled version of the natural parameters of the old message $\tilde{t}_i^{old}$, shown as follows.

$$\boldsymbol{\eta}_i = \eta\boldsymbol{\tau}_i^{old} \tag{18}$$

With the restricted form (18) for the relaxation factor, the REP algorithm still finds a stationary point of the energy function (1).

To see this, we first substitute $\boldsymbol{\lambda}_i = (n-1)\boldsymbol{\nu} - \sum_{k\neq i}\boldsymbol{\lambda}_i$ into (1). Then we compute the derivatives with respect to $\boldsymbol{\nu}$ and η_i, respectively. While the derivative with respect to $\boldsymbol{\nu}$ remains the same as in the case without the constraint (18), the derivative with respect to η_i is different. Zeroing this derivative we obtain:

$$\frac{p(\mathbf{w})\exp(\boldsymbol{\nu}^T\phi(\mathbf{w}) + \eta_i\boldsymbol{\tau}_{\boldsymbol{i}}^{\boldsymbol{old}^T}\phi(\mathbf{w}))}{\int_{\mathbf{w}} p(\mathbf{w})\exp(\boldsymbol{\nu}^T\phi(\mathbf{w}) + \eta_i\boldsymbol{\tau}_{\boldsymbol{i}}^{\boldsymbol{old}^T}\phi(\mathbf{w}))d\mathbf{w}}\boldsymbol{\tau}_i^{old} - \frac{t_i(\mathbf{w})p(\mathbf{w})\exp(\boldsymbol{\lambda}_i^T\phi(\mathbf{w}) + \eta_i\boldsymbol{\tau}_{\boldsymbol{i}}^{\boldsymbol{old}^T}\phi(\mathbf{w}))}{\int_{\mathbf{w}} t_i(\mathbf{w})p(\mathbf{w})\exp(\boldsymbol{\lambda}_i^T\phi(\mathbf{w}) + \eta_i\boldsymbol{\tau}_{\boldsymbol{i}}^{\boldsymbol{old}^T}\phi(\mathbf{w}))d\mathbf{w}}\boldsymbol{\tau}_i^{old} = 0 \tag{19}$$

with the constraint $\sum_i|\eta_i| < c'$. This is achieved by minimizing the penalized KL divergence

$$KL_r(t_i r_i q^{\backslash i}||qr_i) + c|\eta_i|_1\,. \tag{20}$$

as used in the relaxed EP.

3 Relaxed KL for GP classification

For GP classification, we minimize the relaxed KL divergence with l_1 penalty over b_i by line search. Here we present how to compute the value of this cost function:

$$Q(b_i) = KL_r(t_i r_i q^{\backslash i}||r_i q) + c|b_i| \tag{21}$$

Following the notations in the main text (from equations (16) to (23)), we have $Q(b_i)$ as

$$\begin{aligned}&\frac{1}{\hat{Z}_i}\left\{[(1-\epsilon)\log(1-\epsilon)-\epsilon\log\epsilon]\,\psi(z)+\epsilon\log\epsilon\right\}+\frac{1}{2v_{i,b}}(F_{i,b}-\tilde{h}_i m_{i,b})-\frac{1}{2}\log\left(1+(b_i+\frac{1}{v_{i,b}})\lambda_i^{\backslash i}\right)\\&+\frac{1}{2}\log(b_i\lambda_i^{\backslash i}+1)-\frac{1}{2}b_i(m_i^2-2m_i\tilde{h}_i+F_{i,b})+\frac{1}{2}\frac{(m_i-h_i^{\backslash i})^2}{\lambda_i^{\backslash i}+b_i^{-1}}-\log\hat{Z}_i+c|b_i| \qquad (22)\end{aligned}$$

where $\hat{Z}_i = \epsilon + (1-2\epsilon)\psi(z),$ and the term $F_{i,b}$ can be computed as follows:

$$\delta_{i,b} = (\frac{1}{v_{i,b}} - \frac{1}{v_i})^{-1} \tag{23}$$

$$a_{ii}^{new} = (\frac{1}{a_{ii}} + \frac{1}{\delta})^{-1} \tag{24}$$

$$\tilde{a}_{ii}^{new} = a_{ii}^{new}(1 - \frac{a_{ii}^{new}}{a_{ii}^{new} + b_i^{-1}}) \tag{25}$$

$$F_{i,b} = \tilde{a}_{ii}^{new} + \tilde{h}_i^2 \tag{26}$$

Using the above equations, we can efficiently optimize $Q(b_i)$ over b_i via line search.

4 Power EP for GP classification

In this section, we describe how to train GP classifiers by Power EP. The updates of Power EP are the same as equations (5.64) to (5.74) in [1], except two critical modifications:

- Replace equation (5.67) in [1] by

$$\alpha_i = \frac{1}{\sqrt{\lambda_i}} \frac{[(1-\epsilon)^u - \epsilon^u]\,\mathcal{N}(z|0,1)}{\epsilon^u + [(1-\epsilon)^u - \epsilon^u]\,\psi(z)} \tag{27}$$

  where $\psi(\cdot)$ is the standard normal cumulative density function and u is the power used by Power EP.

- Moreover, after (5.70), scale v_i by u:

$$v_i \leftarrow uv_i \tag{28}$$

5 References

[1] T. P. Minka. *A family of algorithms for approximate Bayesian inference.* PhD thesis, Massachusetts Institute of Technology, 2011.