A Different views on fixed prior variances.

Suppose we aim to update variational posterior $q(\mu_p) = \delta(\mu_p)$ (MAP estimate) by optimizing the following ELBO lower bounding $\log p(\mathcal{D}, \mu_p)$:

$$\mathcal{L}(\mu_q, \sigma_q^2, \mu_p, \sigma_p^2) = \mathcal{L}_{data}(\mu_q, \sigma_q^2) - \frac{1}{2}\Big[\frac{(\mu_q - \mu_p)^2 + \sigma_q^2}{\gamma} - 1 + \log\frac{\gamma}{\sigma_q^2}\Big] - \frac{\mu_p^2}{2\alpha}. \tag{11}$$

we note that only the second and third term on RHS depends on μ_p. Differentiating w.r.t. μ_p gives:

$$\partial_{\mu_p}\mathcal{L}(\mu_q, \sigma_q^2, \mu_p, \sigma_p^2) = \frac{\mu_q - \mu_p}{\gamma} - \frac{\mu_p}{\alpha} = \frac{\alpha\mu_q - \alpha\mu_p - \gamma\mu_p}{\alpha\gamma}. \tag{12}$$

Setting the above equation to zero gives $\mu_p = \frac{\alpha}{\alpha+\gamma}\mu_q$. Next we calculate the coordinate ascend update:

$$\partial_{\mu_q}\mathcal{L}(\mu_q, \sigma_q^2, \mu_p, \sigma_p^2) = \partial_{\mu_q}\mathcal{L}_{data}(\mu_q, \sigma_q^2) - \frac{1}{\gamma}(\mu_q - \mu_p)|_{\mu_p = \frac{\alpha}{\alpha+\gamma}\mu_q} =$$
$$\mathcal{L}_{data}(\mu_q, \sigma_q^2) - \frac{\gamma}{\gamma(\alpha+\gamma)}\mu_q \tag{13}$$

Hence coordinate ascend with MAP learning gives the same update as applying VI to learn $q(\mu_p)$. However, MAP learning neglects the constants allowing to learn α_{reg} by optimizing ELBO.

B Derivations

We first present detailed derivation for the bound in Eq. (9).

Step I: Recall that we assume fixed prior variances and the following priors $p(\boldsymbol{W}|\boldsymbol{\mu}_p) = \mathcal{N}(\boldsymbol{W}|\boldsymbol{\mu}_p, \gamma\mathbf{1})$, $p(\boldsymbol{\mu}_p) = \mathcal{N}(\boldsymbol{\mu}_p|\mathbf{0}, \alpha\mathbf{1})$, and posteriors $q(\boldsymbol{W}|\boldsymbol{\mu}_q, \boldsymbol{\sigma}_q^2) = \mathcal{N}(\boldsymbol{W}|\boldsymbol{\mu}_q, \boldsymbol{\sigma}_q^2)$, $q(\boldsymbol{\mu}_p) = \mathcal{N}(\boldsymbol{\mu}_p|\boldsymbol{m}, \boldsymbol{\beta})$. Since both prior and posterior follow mean-field assumption, we present derivations for an individual coordinate.

Step II: To derive Φ in Eq. (6) we first calculate KL divergence between posterior and prior over μ_p:

$$D_{KL}\big(q(\mu_p)||p(\mu_p)\big) = \frac{1}{2}\Big[\frac{m^2 + \beta}{\alpha} - \log\beta + \log\alpha - 1\Big] \tag{14}$$

We have that $D_{KL}[q, p] = \mathcal{H}[q, p] - \mathcal{H}[q]$, So $\mathbb{E}_q \log p = -\mathcal{H}[q, p] = -D_{KL}[q, p] - \mathcal{H}[q]$. For $q(w|\mu_q, \sigma_q^2) = \mathcal{N}(w|\mu_q, \sigma_q^2)$, we have $\mathcal{H}[q] = \frac{1}{2}\log\sigma_q^2 + \log\sqrt{2\pi e}$, so first term in Eq. (6) is:

$$\mathbb{E}_{q(w|\mu_q, \sigma_q^2)} \log p(w|\mu_p, \sigma_p^2) =$$
$$-\frac{1}{2}\Big[\frac{\sigma_q^2 + (\mu_q - \mu_p)^2}{\gamma} + \log\gamma - \log\sigma_q^2 - 1\Big] - \frac{1}{2}\log\sigma_q^2 - \frac{1}{2}\log 2\pi e = \tag{15}$$
$$-\frac{1}{2}\Big[\frac{\sigma_q^2 + (\mu_q - \mu_p)^2}{\gamma} + \log\gamma + \log 2\pi e - 1\Big] \tag{16}$$

Hence substituting into the definition of $\Phi\big(\mu_q, \sigma_q^2, q(\mu_p, \sigma_p^2)\big)$:

$$\Phi\big(\mu_q, \sigma_q^2, q(\mu_p, \sigma_p^2)\big) = \mathbb{E}_{q(\mu_p, \sigma_p^2)}\Big[\mathbb{E}_{q(w|\mu_q, \sigma_q^2)} \log p(w|\mu_p, \sigma_p^2)\Big] - D_{KL}\big(q(\mu_p, \sigma_p^2)||p(\mu_p)p(\sigma_p^2)\big), \tag{17}$$

results in the expression:

$$\Phi = -\mathbb{E}_{q(\mu_p)}\frac{1}{2}\Big[\frac{\sigma_q^2 + (\mu_q - \mu_p)^2}{\gamma} + \log\gamma + \log 2\pi e - 1\Big] - \frac{1}{2}\Big[\frac{m^2}{\alpha} + \frac{\beta}{\alpha} + \log\alpha - \log\beta - 1\Big] =$$
$$-\frac{1}{2}\Big[\mathbb{E}_{q(\mu_p)}\frac{\sigma_q^2 + (\mu_q - \mu_p)^2}{\gamma} + \log\gamma + \log 2\pi e + \frac{m^2}{\alpha} + \frac{\beta}{\alpha} + \log\alpha - \log\beta - 2\Big]. \tag{18}$$

This expression has to be maximized w.r.t. m and β to derive q^* which is used to form Φ^*. However, using Eq. (7) gives the maximizer without needing to optimize for m:

$$\log q^*(\mu_p) \propto -\frac{(\mu_q-\mu_p)^2}{2\gamma} - \frac{\mu_p^2}{2\alpha} = -\frac{\alpha(\mu_q-\mu_p)^2}{2\alpha\gamma} - \frac{\gamma\mu_p^2}{2\alpha\gamma}$$

$$-\frac{\alpha\mu_p^2+\alpha\mu_q^2-2\alpha\mu_p\mu_q+\gamma\mu_p^2}{2\alpha\gamma} = \frac{(\alpha+\gamma)(\mu_p-\frac{\alpha}{\alpha+\gamma}\mu_q)^2+(\alpha-\frac{\alpha^2}{(\alpha+\gamma)^2})\mu_q^2}{2\alpha\gamma} = \tag{19}$$

$$\frac{(\mu_p-\frac{\alpha}{\alpha+\gamma}\mu_q)^2+(\alpha-\frac{\alpha^2}{(\alpha+\gamma)^2})\mu_q^2}{2\frac{\alpha\gamma}{\alpha+\gamma}} \tag{20}$$

This is Gaussian distribution, so $q^*(\mu_p) = \mathcal{N}(\mu_p|\frac{\alpha}{\alpha+\gamma}\mu_q, \frac{\alpha\gamma}{\alpha+\gamma})$.

Step III: We integrate $q^*(\mu_p)$ in the derived expression for Φ:

$$\mathbb{E}_{\mathcal{N}(\mu_p|\frac{\alpha}{\alpha+\gamma}\mu_q,\frac{\alpha\gamma}{\alpha+\gamma})}\frac{(\mu_q-\mu_p)^2}{\gamma} = \tag{21}$$

$$\mathbb{E}_{\mathcal{N}(\mu_p|\frac{\alpha}{\alpha+\gamma}\mu_q,\frac{\alpha\gamma}{\alpha+\gamma})}\frac{\mu_q^2+\mu_p^2-2\mu_p\mu_q}{\gamma} = \tag{22}$$

$$\frac{\mu_q^2+(\frac{\alpha}{\alpha+\gamma}\mu_q)^2+\frac{\alpha\gamma}{\alpha+\gamma}-2\frac{\alpha}{\alpha+\gamma}\mu_q^2}{\gamma} = \frac{(\mu_q-\frac{\alpha}{\alpha+\gamma}\mu_q)^2+\frac{\alpha\gamma}{\alpha+\gamma}}{\gamma} = \tag{23}$$

$$\frac{\frac{\gamma^2}{(\alpha+\gamma)^2}\mu_q^2}{\gamma} - \frac{\gamma}{(\alpha+\gamma)} = \frac{\gamma}{(\alpha+\gamma)^2}\mu_q^2 + \frac{\gamma}{(\alpha+\gamma)} \tag{24}$$

Substituting $q^*(\mu_p)$ back into Φ gives:

$$\Phi = -\frac{1}{2}\Big[\frac{\gamma}{(\alpha+\gamma)^2}\mu_q^2 + \frac{\gamma}{(\alpha+\gamma)} + \frac{\sigma_q^2}{\gamma} + \log\gamma + \log 2\pi e +$$

$$\frac{(\frac{\alpha}{\alpha+\gamma}\mu_q)^2}{\alpha} + \frac{\frac{\alpha\gamma}{\alpha+\gamma}}{\alpha} + \log\alpha - \log\frac{\alpha\gamma}{\alpha+\gamma} - 2\Big] = \tag{25}$$

$$-\frac{1}{2}\Big[\frac{\alpha+\gamma}{(\alpha+\gamma)^2}\mu_q^2 + \frac{\sigma_q^2}{\gamma} + \log\gamma + 2\frac{\gamma}{\alpha+\gamma} - \log\frac{\gamma}{\alpha+\gamma}) + \log 2\pi e - 2\Big] = \tag{26}$$

$$-\frac{1}{2}\Big[\frac{1}{\alpha+\gamma}\mu_q^2 + \frac{\sigma_q^2}{\gamma} + \log\gamma + 2\frac{\gamma}{\alpha+\gamma} - \log\frac{\gamma}{\alpha+\gamma} + \log 2\pi e - 2\Big] = \tag{27}$$

$$-\frac{1}{2}\Big[\frac{\alpha_{reg}\mu_q^2+\sigma_q^2}{\gamma} + \log\gamma + 2\frac{\gamma}{\alpha+\gamma} - \log\frac{\gamma}{\alpha+\gamma} + \log 2\pi e - 2\Big]. \tag{28}$$

Step IV: The full expression for the bound $\mathcal{L}_m^*$ becomes

$$\mathcal{L}_m^*(\mu_q,\sigma_q^2) = \mathcal{L}_{data}(\mu_q,\sigma_q^2) - \frac{1}{2}\Big[\frac{\alpha_{reg}\mu_q^2+\sigma_q^2}{\gamma} + \log\gamma + 2\alpha_{reg} - \log\alpha_{reg} - \log\sigma_q^2 - 2\Big] \tag{29}$$

We now consider the derivation of the bound $\mathcal{L}_{mv}^*$ in Eq. (10). These calculations are similar in spirit to well-known derivations of Bayesian posteriors in hierarchical models.

Step I: We employ priors $p(w) = \mathcal{N}(w|\mu_p, 1/\tau_p)$, $p(\mu_p|\tau_p) = \mathcal{N}(0, \frac{1}{t\tau_p})$ and $p(\tau_p) = \mathcal{G}(\alpha,\beta)$ and posteriors $q(w) = \mathcal{N}(w|\mu_q,\sigma_q^2)$, $q(\mu_p|\tau_p) = \mathcal{N}(\mu_p)$, $q(\tau_p) = \mathcal{G}(\tau_p)$.

Step II: We again use $\mathbb{E}_q \log p = -\mathcal{H}[q,p] = -D_{KL}[q,p] - \mathcal{H}[q]$ to derive:

$$\mathbb{E}_{q(w|\mu_q,\sigma_q^2)}\log p(w|\mu_p,\sigma_p^2) = \tag{30}$$

$$-\frac{1}{2}\Big[\frac{\sigma_q^2+(\mu_p-\mu_q)^2)}{\sigma_p^2} + \log\sigma_p^2 + 1 + \log 2\pi e\Big] = \tag{31}$$

$$-\frac{\tau_p}{2}(\sigma_q^2+(\mu_p-\mu_q)^2) + \frac{1}{2}\log\tau_p - \frac{1}{2} - \log\sqrt{2\pi e} \tag{32}$$

We first find the posterior for conditional distribution $\mu_p|\tau_p$. We have that $\log p(\mu_p|\tau_p) \propto -\frac{t\tau_p\mu_p^2}{2}$. We substitute into Eq. (7) to derive:

$$\log q^*(\mu_p|\tau_p) \propto -\frac{\tau_p}{2}(\sigma_q^2 + (\mu_p - \mu_q)^2) - \frac{1}{2}t\tau_p\mu_p^2 = \tag{33}$$

$$-\frac{\tau_p}{2}(\sigma_q^2 + \mu_p^2 + \mu_q^2 - 2\mu_q\mu_p + t\mu_p^2) \triangleq \tag{34}$$

$$-\frac{\tau_p}{2}\mu_p^2 + \tau_p\mu_q\mu_p - \frac{1}{2}t\tau_p\mu_p^2 = \tag{35}$$

$$-\frac{(1+t)\tau_p}{2}\mu_p^2 + \tau_p\mu_q\mu_p = \tag{36}$$

$$-\frac{(1+t)\tau_p}{2}(\mu_p^2 + \frac{2}{(1+t)}\mu_q\mu_p) \propto \log\mathcal{N}(\mu_p|\frac{1}{1+t}\mu_q, \frac{1}{(1+t)\tau_p}) \tag{37}$$

We now calculate the posterior for τ_p. We derive $q^*(\tau_p)$ by marginalizing μ_p from the joint density $q^*(\tau_p, \mu_p)$. We have that $\log p(\tau_p) = (\alpha - 1)\log\tau_p - \beta\tau_p$ and $\log p(\mu_p) = -t\tau_p\mu_p^2 + \frac{1}{2}\log\tau_p$. The density $q^*(\tau_p, \mu_p)$ satisfies:

$$\log q^*(\tau_p, \mu_p) \propto$$

$$-\frac{\tau_p}{2}(\sigma_q^2 + (\mu_p - \mu_q)^2) - \frac{1}{2}\log\tau_p - \frac{1}{2}t\tau_p\mu_p^2 + \frac{1}{2}\log\tau_p + (\alpha - 1)\log\tau_p - \beta\tau_p \triangleq \tag{38}$$

$$-\frac{\tau_p}{2}(\sigma_q^2 + 2\beta) + (\alpha - 1)\log\tau_p - \frac{\tau_p}{2}(t\mu_p^2 + (\mu_p - \mu_q)^2) \tag{39}$$

To marginalize μ_p we need to integrate out $\exp[-\frac{\tau_p}{2}(t\mu_p^2 + (\mu_p - \mu_q)^2)]$:

$$\int \exp(-\frac{\tau_p}{2}(t\mu_p^2 + (\mu_p - \mu_q)^2))d\mu_p = \int \exp(-\frac{\tau_p}{2}(t\mu_p^2 + \mu_p^2 + \mu_q^2 - 2\mu_p\mu_q))d\mu_p = \tag{40}$$

$$\int \exp(-\frac{\tau_p}{2}((1+t)\mu_p^2 + \mu_q^2 - 2\frac{1+t}{1+t}\mu_p\mu_q))d\mu_p = \tag{41}$$

$$\int \exp(-\frac{\tau_p(1+t)}{2}(\mu_p^2 + \frac{1}{1+t}\mu_q^2 - 2\frac{1}{1+t}\mu_p\mu_q))d\mu_p = \tag{42}$$

$$\int \exp(-\frac{\tau_p(1+t)}{2}(\mu_p^2 + \frac{1}{(1+t)^2}\mu_q^2 - 2\frac{1}{1+t}\mu_p\mu_q - \frac{1}{(1+t)^2}\mu_q^2 + \frac{1}{1+t}\mu_q^2)d\mu_p = \tag{43}$$

$$\int \exp(-\frac{\tau_p(1+t)}{2}((\mu_p - \frac{1}{1+t}\mu_q)^2 - \frac{1}{(1+t)^2}\mu_q^2 + \frac{1}{(1+t)}\mu_q^2)d\mu_p = \tag{44}$$

$$\int \exp(-\frac{\tau_p(1+t)}{2}((\mu_p - \frac{1}{1+t}\mu_q)^2) - \frac{1}{(1+t)}\mu_q^2 + \mu_q^2 d\mu_p = \tag{45}$$

$$\int \exp(-\frac{\tau_p(1+t)}{2}((\mu_p - \frac{1}{1+t}\mu_q)^2) - \frac{\tau_p}{2}\frac{t}{(1+t)}\mu_q^2)d\mu_p = \tag{46}$$

$$\exp(-\frac{t}{2(1+t)}\mu_q^2\tau_p)\frac{\sqrt{2\pi}}{\sqrt{\tau_p(t+1)}}. \tag{47}$$

So after integrating out μ_p we derive (by substituting $-\frac{t}{2(1+t)}\mu_q^2\tau_p + \log\frac{\sqrt{2\pi}}{\sqrt{\tau_p(t+1)}}$ into the joint density we obtain:

$$\log q^*(\tau_p) \propto -\frac{\tau_p}{2}(\sigma_q^2 + 2\beta) + (\alpha - 1)\log\tau_p - \frac{t}{2(1+t)}\mu_q^2\tau_p - \frac{1}{2}\log\tau_p = \tag{48}$$

$$\log q^*(\tau_p) \propto -\frac{\tau_p}{2}(\sigma_q^2 + 2\beta) + (\alpha - \frac{1}{2})\log\tau_p - \frac{t}{2(1+t)}\mu_q^2\tau_p \tag{49}$$

$$\log q^*(\tau_p) \propto -\frac{\tau_p}{2}(\sigma_q^2 + 2\beta + \frac{t}{1+t}\mu_q^2) + (\alpha - \frac{1}{2})\log\tau_p \tag{50}$$

$$\propto \log\mathcal{G}(\alpha + \frac{1}{2}, \beta + \frac{1}{2}(\sigma_q^2 + \frac{t}{(1+t)}\mu_q^2)) \tag{51}$$

Hence the optimal posteriors are $q^*(\mu_p|\tau_p) = \mathcal{N}(\frac{1}{1+t}\mu_q, \frac{1}{(1+t)\tau_p})$ and $q^*(\tau_p) = \mathcal{G}(\alpha+\frac{1}{2}, \beta+\frac{1}{2}(\sigma_q^2 + \frac{t}{(1+t)}\mu_q^2))$.

Step III: We follow by forming the expression for Φ^* and marginalizing out prior parameters. First, to form Φ^*, we need to calculate KL divergence between derived q^* and corresponding priors. We now calculate KL divergence $D_{KL}(q^*(\mu_p|\tau_p)||\mathcal{N}(\mu_p|0, \frac{1}{t\tau_p}))$:

$$
\begin{aligned}
&D_{KL}(\mathcal{N}(\mu_p|\frac{1}{1+t}\mu_q, \frac{1}{(t+1)\tau_p})||\mathcal{N}(\mu_p|0, \frac{1}{t\tau_p})) = \\
&\frac{1}{2}\Big[t\tau_p\Big(\frac{1}{(1+t)^2}\mu_q^2 + \frac{1}{(t+1)\tau_p}\Big) - \log\frac{t\tau_p}{(t+1)\tau_p} - 1\Big] = && (52)\\
&\frac{1}{2}\Big[t\tau_p\frac{1}{(1+t)^2}\mu_q^2 + \frac{t}{(t+1)} - \log\frac{t}{t+1} - 1\Big] = && (53)\\
&\frac{1}{2}\Big[\frac{t}{(1+t)^2}\tau_p\mu_q^2 + \frac{t}{(t+1)} - \log\frac{t}{t+1} - 1\Big] && (54)
\end{aligned}
$$

and recall the formula KL between Gamma distributions

$$
\begin{aligned}
&D_{KL}(\mathcal{G}(x|\alpha_p, \beta_p)||\mathcal{G}(x|\alpha_q, \beta_q)) = (\alpha_p - \alpha_q)\phi(\alpha_p) - \log\Gamma(\alpha_p) + \\
&+ \log\Gamma(\alpha_q) + \alpha_q(\log\beta_p - \log\beta_q) + \alpha_p\frac{\beta_q - \beta_p}{\beta_p}, && (55)
\end{aligned}
$$

where ϕ denotes digamma function and Γ denotes gamma function. Substituting posterior $\mathcal{G}(\sigma_p^2|\alpha + \frac{1}{2}, \beta + \frac{t}{2(t+1)}\mu_q^2 + \frac{1}{2}\sigma_q^2)$ and prior $\mathcal{G}(\sigma_p^2|\alpha, \beta)$ into the above equation gives:

$$
\begin{aligned}
&D_{KL}(\mathcal{G}(\sigma_p^2|\alpha + \frac{1}{2}, \beta + \frac{t}{2(t+1)}\mu_q^2 + \frac{1}{2}\sigma_q^2)||\mathcal{G}(\sigma_p^2|\alpha, \beta)) = \\
&= \frac{1}{2}\phi(\alpha + \frac{1}{2}) - \log\Gamma(\alpha + \frac{1}{2}) + \log\Gamma(\alpha) + \alpha\log[\beta + \frac{t}{2(t+1)}\mu_q^2 + \frac{1}{2}\sigma_q^2] \\
&- \alpha\log\beta + \frac{(\alpha + \frac{1}{2})\beta}{\beta + \frac{t}{2(t+1)}\mu_q^2 + \frac{1}{2}\sigma_q^2} - \alpha && (56)
\end{aligned}
$$

Hence, to calculate Φ^* we need to integrate the expectations in the expression:

$$
\begin{aligned}
&\Phi^* = \mathbb{E}_{q^*(\tau_p)}\mathbb{E}_{q^*(\mu_p)}\Big[-\frac{1}{2}\Big(\tau_p(\sigma_q^2 + (\mu_p - \mu_q)^2) - \log\tau_p + 1 - \log 2\pi e\Big) \\
&-D_{KL}(\mathcal{N}(\mu_p|\frac{1}{1+t}\mu_q, (t+1)\tau_p)||\mathcal{N}(\mu_p|0, t\tau_p))\Big] \\
&-D_{KL}(\mathcal{G}(\tau_p|\alpha + \frac{1}{2}, \beta + \frac{t}{2(t+1)}\mu_q^2 + \frac{1}{2}\sigma_q^2)||\mathcal{G}(\tau_p|\alpha, \beta)) && (57)
\end{aligned}
$$

We divide the above expression and first integrate out $q^*(\mu_p)$ denoting it as $\star$.

$$
\begin{aligned}
&\star = \mathbb{E}_{q^*(\mu_p)}\Big[-\frac{1}{2}\Big(\tau_p(\sigma_q^2 + (\mu_p - \mu_q)^2) - \log\tau_p + 1 - \log 2\pi e\Big) \\
&-D_{KL}(\mathcal{N}(\mu_p|\frac{1}{1+t}\mu_q, (t+1)\tau_p)||\mathcal{N}(\mu_p|0, t\tau_p))\Big] = && (58)\\
&\mathbb{E}_{q^*(\mu_p)} - \frac{1}{2}\Big(\tau_p(\sigma_q^2 + (\mu_p - \mu_q)^2) - \log\tau_p + 1 - \log 2\pi e\Big) \\
&- \frac{1}{2}\Big[\frac{t}{(1+t)^2}\tau_p\mu_q^2 + \frac{t}{(t+1)} - \log\frac{t}{t+1} - 1\Big] && (59)
\end{aligned}
$$

We have that

$$\mathbb{E}_{q^*(\mu_p)}(\mu_p - \mu_q)^2 = \mathbb{E}_{q^*(\mu_p)}\mu_p^2 + \mu_q^2 - 2\mu_q\mu_p = \tag{60}$$

$$\frac{1}{(1+t)^2}\mu_q^2 + \frac{1}{(t+1)\tau_p} + \mu_q^2 - \frac{2}{1+t}\mu_q^2 = \tag{61}$$

$$\frac{(1+t)^2 + 1 - 2(1+t)}{(1+t)^2}\mu_q^2 + \frac{1}{(t+1)\tau_p} = \tag{62}$$

$$\frac{t^2}{(1+t)^2}\mu_q^2 + \frac{1}{(t+1)\tau_p} \tag{63}$$

So by substituting into the definition of $\star$ we have:

$$\star = -\frac{1}{2}\Big(\tau_p(\sigma_q^2 + \frac{t^2}{(1+t)^2}\mu_q^2 + \frac{1}{(t+1)\tau_p}) - \log\tau_p + 1 - \log 2\pi e\Big) - \tag{64}$$

$$\frac{1}{2}\Big[\frac{t}{(1+t)^2}\tau_p\mu_q^2 + \frac{t}{(t+1)} - \log\frac{t}{t+1} - 1\Big] =$$

$$-\frac{\tau_p}{2}(\sigma_q^2 + \frac{t^2}{(1+t)^2}\mu_q^2) - \frac{1}{2(t+1)} + \frac{1}{2}\log\tau_p - \frac{1}{2} + \log\sqrt{2\pi e} -$$

$$\frac{1}{2}\Big[\frac{t}{(1+t)^2}\tau_p\mu_q^2 + \frac{t}{(t+1)} - \log\frac{t}{t+1} - 1\Big] \tag{65}$$

$$= -\frac{\tau_p}{2}(\sigma_q^2 + \frac{t}{1+t}\mu_q^2) + \log\sqrt{2\pi e} + \frac{1}{2}\log\tau_p - \frac{1}{2}\log\frac{t}{t+1}\Big] - 2 \tag{66}$$

We next need to integrate $\star$ w.r.t. $q^*(\tau_p)$ to derive Φ^*. We use the fact that for Gamma distribution $\mathcal{G}(x|\alpha,\beta)$ we have that $\mathbb{E}\log X = \psi(\alpha) - \log\beta$, where ψ denotes digamma function and $\mathbb{E}X = \frac{\alpha}{\beta}$. So we have that:

$$\mathbb{E}_{q^*(\tau_p)}\tau_p = \frac{\alpha + \frac{1}{2}}{\beta + \frac{1}{2}(\sigma_q^2 + \frac{t}{1+t}\mu_1^2))} \tag{67}$$

$$\mathbb{E}_{q^*(\tau_p)}\log\tau_p = \psi(\alpha + \frac{1}{2}) - \log\Big[\beta + \frac{1}{2}(\sigma_q^2 + \frac{t}{1+t}\mu_1^2))\Big] \tag{68}$$

$$\mathbb{E}_{q^*(\tau_p)}\star = \mathbb{E}_{q^*(\tau_p)}\Bigg[-\frac{\tau_p}{2}(\sigma_q^2 + \frac{t}{1+t}\mu_q^2) + \log\sqrt{2\pi e} + \frac{1}{2}\log\tau_p - \frac{1}{2}\log\frac{t}{t+1}\Big] - 2\Bigg] = \tag{69}$$

$$-\frac{\tau_p}{2}(\sigma_q^2 + \frac{t}{1+t}\mu_q^2) + \log\sqrt{2\pi e} - \frac{1}{2}\log\tau_p - \frac{1}{2}\log\frac{t}{t+1}\Big] = \tag{70}$$

$$-\frac{1}{2}(\frac{\alpha + \frac{1}{2}}{\beta + \frac{1}{2}(\sigma_q^2 + \frac{t}{1+t}\mu_1^2))})(\sigma_q^2 + \frac{t}{1+t}\mu_q^2) + \log\sqrt{2\pi e} +$$

$$\frac{1}{2}\psi(\alpha + \frac{1}{2}) - \frac{1}{2}\log\Big[\beta + \frac{1}{2}(\sigma_q^2 + \frac{t}{1+t}\mu_1^2))\Big] - \frac{1}{2}\log\frac{t}{t+1} \tag{71}$$

We next form Φ^*:

$$\Phi^* = \mathbb{E}_{q^*(\tau_p)}\star - D_{KL}(\mathcal{G}(\tau_p|\alpha + \frac{1}{2}, \beta + \frac{t}{2(t+1)}\mu_q^2 + \frac{1}{2}\sigma_q^2)||\mathcal{G}(\tau_p|\alpha,\beta)) = \tag{72}$$

$$-\frac{\alpha + \frac{1}{2}}{2(\beta + \frac{t}{2(t+1)}\mu_q^2 + \frac{1}{2}\sigma_q^2)}(\sigma_q^2 + \frac{t}{(1+t)}\mu_q^2) - \frac{1}{2}\log[\beta + \frac{t}{2(t+1)}\mu_q^2 + \frac{1}{2}\sigma_q^2] +$$

$$\frac{1}{2}\boldsymbol{\phi}(\alpha + \frac{1}{2}) - \Big[\frac{1}{2}\boldsymbol{\phi}(\alpha + \frac{1}{2}) - \log\Gamma(\alpha + \frac{1}{2}) + \log\Gamma(\alpha)$$

$$+\alpha\log[\beta + \frac{t}{2(t+1)}\mu_q^2 + \frac{1}{2}\sigma_q^2] - \alpha\log\beta + \frac{(\alpha + \frac{1}{2})\beta}{\beta + \frac{t}{2(t+1)}\mu_q^2 + \frac{1}{2}\sigma_q^2} - \alpha\Big] + \log\sqrt{2\pi e} = \tag{73}$$

$$= -\frac{\alpha+\frac{1}{2}}{2(\beta+\frac{t}{2(t+1)}\mu_q^2+\frac{1}{2}\sigma_q^2)}(\sigma_q^2+\frac{t}{(1+t)}\mu_q^2+2\beta)-(\alpha+\frac{1}{2})\log[\beta+\frac{t}{2(t+1)}\mu_q^2+\frac{1}{2}\sigma_q^2]+$$

$$-\Big[-\log\Gamma(\alpha+\frac{1}{2})+\log\Gamma(\alpha)]-\alpha\log\beta-\alpha\Big]+\log\sqrt{2\pi e}= \tag{74}$$

$$-(\alpha+\frac{1}{2})\log[\beta+\frac{t}{2(t+1)}\mu_q^2+\frac{1}{2}\sigma_q^2]+\log\Gamma(\alpha+\frac{1}{2})-\log\Gamma(\alpha)+\alpha\log\beta+\alpha+\log\sqrt{2\pi e}= \tag{75}$$

$$-(\alpha+\frac{1}{2})\log[\beta+\frac{t}{2(t+1)}\mu_q^2+\frac{1}{2}\sigma_q^2]+\log\frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)}+\alpha\log\beta+\alpha+\log\sqrt{2\pi e}, \tag{76}$$

where $\delta=\frac{t}{1+t}$.

Step IV: We form the bound $\mathcal{L}^*_{mv}(\mu_q,\sigma_q^2)$ as:

$$\mathcal{L}_{data}(\mu_q,\sigma_q^2)-(\alpha+\frac{1}{2})\log[\beta+\frac{t}{2(t+1)}\mu_q^2+\frac{1}{2}\sigma_q^2]+\log\frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)}+\frac{1}{2}\log\sigma_q^2+\alpha\log\beta+\alpha \tag{77}$$

C Experimental setup

Here we list all hyperparameters used in Section 4. We begin by vectorized MNIST experiment. For

experiment	batch size	lr	epochs	init logvar	α_{reg}	α	β	δ	γ	test samples
MLP MNIST	256	0.001	1000	-7	0.05	0.5	0.01	0.1	$\frac{20}{N_{in}}$	2000
LeNet	512	0.001	800	-7	0.025	0.5	0.01	0.1	$\frac{1}{N_{in}}$	500
LeNet (CIFAR10/100, SVHN)	512	0.001	800	-7	0.2	0.5	0.01	0.5	$\frac{1}{N_{in}}$	500
CNN large	512	0.001	800	-7	0.025	0.5	0.01	0.5	$\frac{1}{N_{in}}$	500
CNN large (STL10)	512	0.001	800	-7	0.2	0.5	0.01	0.3	$\frac{1}{N_{in}}$	500

Table 4: Hyperparameters used in our experiments.

LeNet and MNIST experiments, we do not use data augmentation. For larger CNN experiments, we use standard simple data augmenntation: padding in reflect mode with 4 (STL10 12) pixels and random cropping to initial size and random horizontal flip (exlcluding SVHN) with probability 0.5. For MNIST 1 epoch takes 1.2 sec, for LeNet 2.5 sec and for ResNet (CIFAR10) 1 epoch takes 12 sec.

D Additional experimental results

First we report results on UCI data sets as described in the main text but for 1 hidden layer BNN.

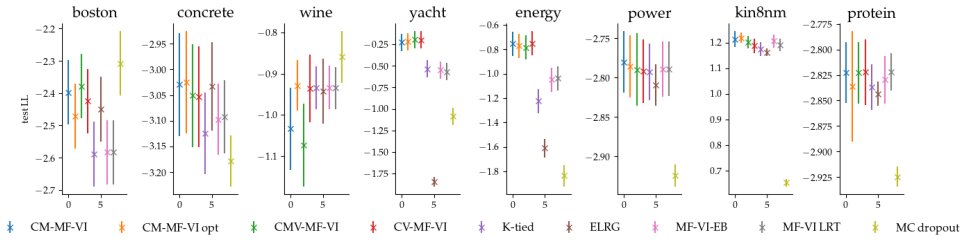

Figure 7: Comparison of collapsed bounds with other algorithms on UCI data sets (top) for 1 hidden layer BNN.

The collapsed bounds perform overall very well. They are outperformed by MC dropout on *boston* and *wine*, but MC dropout fails terribly on other data sets (we do not tune p per data set, but use one setting we of p working well for all data sets). We next report additional statistics for CNN experiments: test ECE score and test NLL on missclassified examples in the table below. We find that

the collapsed bounds als perform well in these metrics, especially in ECE score. We report the same statistics for large CNN experiment and additionally include final data terms from ELBO as they allow us to see if an algorithm is under-fitting/over-fitting. We again observe the collapsed bounds are performing well in terms of ECE score and, they have data terms which are relatively close to test NLLs reported in the main text (the gap between train NLL and test NLL is overall small and visibly smaller than for other algorithms).

test NLL miss ↓ / ECE ↓	MNIST	K-MNIST	F-MNIST	SVHN	CIFAR10
CM-MF-VI	1.63 ± 0.09	2.68 ± 0.09	2.05 ± 0.10	2.91 ± 0.06	2.27 ± 0.01
CMV-MF-VI	1.86 ± 0.09	2.85 ± 0.13	2.09 ± 0.03	2.77 ± 0.02	2.24 ± 0.05
CV-MF-VI	1.82 ± 0.07	2.75 ± 0.09	1.92 ± 0.02	2.66 ± 0.02	2.17 ± 0.02
MF-VI	1.69 ± 0.01	2.19 ± 0.01	1.84 ± 0.05	2.54 ± 0.03	2.09 ± 0.02
MC dropout	1.74 ± 0.05	2.49 ± 0.07	1.71 ± 0.02	2.14 ± 0.00	1.91 ± 0.01
MAP	2.37 ± 0.10	3.16 ± 0.10	2.04 ± 0.03	5.63 ± 0.33	2.90 ± 0.21
CM-MF-VI	0.00 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.02 ± 0.00	0.02 ± 0.00
CMV-MF-VI	0.00 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00
CV-MF-VI	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.02 ± 0.00	0.01 ± 0.00
MF-VI	0.02 ± 0.00	0.04 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.04 ± 0.01
MC dropout	0.01 ± 0.00	0.02 ± 0.00	0.03 ± 0.00	0.10 ± 0.00	0.12 ± 0.01
MAP	0.00 ± 0.00	0.03 ± 0.00	0.01 ± 0.00	0.07 ± 0.00	0.12 ± 0.02

Table 5: Image classification with LeNet CNN comparison of test ECE and test NLL on missclassified examples.

model	data set	CMV-MF-VI	CM-MF-VI	MF-VI	MC dropout	MAP
RESNET18	STL10	1.91 ± 0.02	1.91 ± 0.01	1.85 ± 0.01	3.61 ± 0.04	5.47 ± 0.12
	SVHN	2.44 ± 0.03	2.55 ± 0.01	2.29 ± 0.02	2.34 ± 0.03	7.02 ± 0.19
	CIFAR100	2.86 ± 0.01	3.25 ± 0.03	3.11 ± 0.01	3.19 ± 0.01	8.78 ± 0.07
	CIFAR10	1.94 ± 0.02	2.09 ± 0.01	1.95 ± 0.02	1.97 ± 0.01	5.92 ± 0.13
SHUFFLENET	STL10	1.89 ± 0.02	1.87 ± 0.01	1.81 ± 0.01	1.90 ± 0.00	4.10 ± 0.11
	SVHN	2.15 ± 0.02	2.21 ± 0.01	2.22 ± 0.01	1.99 ± 0.00	3.39 ± 0.05
	CIFAR100	3.14 ± 0.02	3.23 ± 0.03	3.19 ± 0.02	3.56 ± 0.01	6.95 ± 0.05
	CIFAR10	1.97 ± 0.02	1.97 ± 0.02	1.98 ± 0.02	1.90 ± 0.01	3.64 ± 0.05
ALEXNET	STL10	1.84 ± 0.00	1.86 ± 0.01	1.92 ± 0.03	2.30 ± 0.18	4.85 ± 1.00
	SVHN	2.19 ± 0.01	2.27 ± 0.03	2.10 ± 0.01	2.02 ± 0.02	6.83 ± 0.38
	CIFAR100	3.20 ± 0.01	3.28 ± 0.03	3.52 ± 0.02	3.26 ± 0.02	11.64 ± 0.21
	CIFAR10	1.91 ± 0.01	1.97 ± 0.01	1.96 ± 0.01	1.90 ± 0.01	6.74 ± 0.06
RESNET18	STL10	1.07 ± 0.01	1.12 ± 0.02	1.66 ± 0.01	0.07 ± 0.00	0.03 ± 0.00
	SVHN	0.19 ± 0.00	0.13 ± 0.00	0.34 ± 0.00	0.20 ± 0.00	0.01 ± 0.00
	CIFAR100	1.41 ± 0.01	0.72 ± 0.01	2.44 ± 0.02	0.99 ± 0.00	0.04 ± 0.00
	CIFAR10	0.43 ± 0.00	0.28 ± 0.00	0.81 ± 0.01	0.39 ± 0.00	0.03 ± 0.00
SHUFFLENET	STL10	1.02 ± 0.02	1.06 ± 0.02	1.75 ± 0.07	1.34 ± 0.01	0.24 ± 0.00
	SVHN	0.37 ± 0.00	0.36 ± 0.00	0.42 ± 0.01	1.47 ± 0.01	0.16 ± 0.00
	CIFAR100	2.03 ± 0.01	1.66 ± 0.04	2.40 ± 0.03	3.30 ± 0.01	0.33 ± 0.01
	CIFAR10	0.72 ± 0.00	0.71 ± 0.01	0.81 ± 0.01	1.41 ± 0.00	0.25 ± 0.01
ALEXNET	STL10	1.65 ± 0.05	1.60 ± 0.03	1.98 ± 0.02	0.59 ± 0.16	0.19 ± 0.11
	SVHN	0.42 ± 0.00	0.34 ± 0.01	0.70 ± 0.00	0.53 ± 0.01	0.06 ± 0.01
	CIFAR100	2.42 ± 0.03	2.19 ± 0.10	3.16 ± 0.04	1.98 ± 0.03	0.20 ± 0.01
	CIFAR10	0.82 ± 0.01	0.72 ± 0.03	1.35 ± 0.00	0.74 ± 0.03	0.06 ± 0.00
RESNET18	STL10	0.07 ± 0.01	0.07 ± 0.00	0.04 ± 0.00	0.15 ± 0.01	0.22 ± 0.00
	SVHN	0.03 ± 0.00	0.02 ± 0.00	0.04 ± 0.00	0.04 ± 0.00	0.04 ± 0.00
	CIFAR100	0.06 ± 0.00	0.02 ± 0.00	0.08 ± 0.01	0.04 ± 0.00	0.35 ± 0.00
	CIFAR10	0.04 ± 0.00	0.02 ± 0.00	0.06 ± 0.01	0.04 ± 0.00	0.11 ± 0.00
SHUFFLENET	STL10	0.07 ± 0.01	0.09 ± 0.00	0.04 ± 0.01	0.07 ± 0.01	0.25 ± 0.01
	SVHN	0.04 ± 0.01	0.04 ± 0.00	0.04 ± 0.00	0.31 ± 0.01	0.03 ± 0.00
	CIFAR100	0.04 ± 0.00	0.02 ± 0.00	0.05 ± 0.00	0.11 ± 0.00	0.39 ± 0.01
	CIFAR10	0.03 ± 0.00	0.03 ± 0.00	0.04 ± 0.00	0.13 ± 0.00	0.15 ± 0.01
ALEXNET	STL10	0.11 ± 0.02	0.09 ± 0.02	0.03 ± 0.01	0.05 ± 0.04	0.22 ± 0.03
	SVHN	0.06 ± 0.00	0.05 ± 0.01	0.11 ± 0.00	0.11 ± 0.01	0.07 ± 0.01
	CIFAR100	0.08 ± 0.01	0.08 ± 0.00	0.07 ± 0.01	0.09 ± 0.00	0.46 ± 0.00
	CIFAR10	0.07 ± 0.00	0.06 ± 0.01	0.09 ± 0.00	0.09 ± 0.00	0.19 ± 0.01

Table 6: Test NLL on missclassified examples (top), data terms (middle) and test ECE score (bottom).

E Overpruning

Here we provide more details on over-pruning in BNNs. First we consider a simple 1D regression task with 2 hidden layer BNN with a changing number of hidden units (size) with heteroscedastic Gaussian observation model. We plot contributions to predictions from 5 active units, remaining units (inactive) and predictions made by BNN (full pred). We observe that inactive units only contribute to homoscedastic predictive variance which grows with the size of the network. An increasing number of inactive weights introduce excessive noise to predictions, causing under-fitting when the size of the network grows further. These figures are in line with observations made in Figure 3.

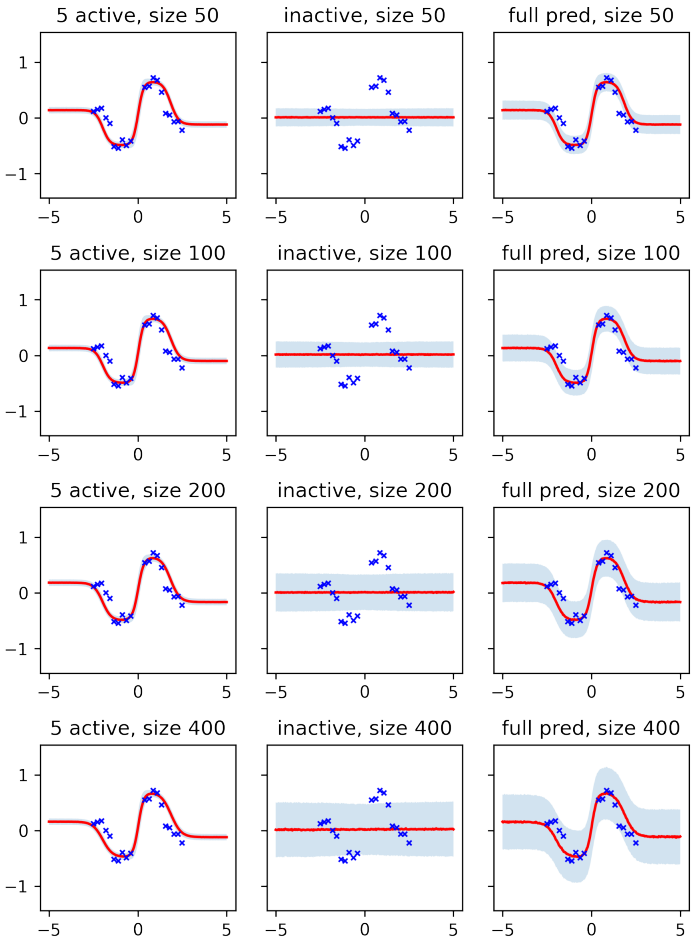

Figure 8: Overpruning for 1D regression with 2 hidden layer BNN and *tanh* activations.

Relation to heuristics. To resolve under-fitting of mean-field BNNs, previous work has re-weighted terms arising from the KL divergence between the approximate posterior to improve the performance of variational BNNs. A generic expression to learn a variational posterior assuming prior $\mathcal{N}(\mathbf{0}, \sigma_p^2\mathbf{1})$ can be expressed as:

$$\mathcal{L}_{heuristic}(\boldsymbol{\mu}_q, \boldsymbol{\sigma}_q^2) = \kappa_{data}\mathcal{L}_{data}(\boldsymbol{\mu}_q, \boldsymbol{\sigma}_q^2) - \mathbf{1}^T\left[\frac{\kappa_1}{2\sigma_p^2}\boldsymbol{\sigma}_q^2 + \frac{\kappa_2}{2\sigma_p^2}\boldsymbol{\mu}_q^2 - \frac{\kappa_3}{2}\log\boldsymbol{\sigma}_q^2 + \frac{\kappa_4}{2}\log\boldsymbol{\sigma}_p^2\right], \quad (78)$$

note that it is merely a heuristic and optimizing $\mathcal{L}_{heuristic}$ might not minimize a metric between approximate posterior $q(\boldsymbol{W}|\boldsymbol{\mu}_q, \boldsymbol{\sigma}_q^2)$ and true posterior $p(\boldsymbol{W}|\mathcal{D})$. Standard MF-VI corresponds to setting $\kappa_{data} = \kappa_1 = \kappa_2 = \kappa_3 = \kappa_4 = 1$. Posterior tempering [70] sets $\kappa_{data} = \kappa_1 = \kappa_2 = \kappa_3 = \frac{1}{T}$ and $\kappa_4 = 1$. KL down-weighting sets $\kappa_{data} = 1$, $\kappa_1 = \kappa_2 = \kappa_3 = \kappa_4 = \beta$. These approaches do not break the dependency $\kappa_2 = \kappa_1$ which we find to be inducing under-fitting. We have found one attempt to break the dependency $\kappa_2 = \kappa_1$ [45], but this was derived as a heuristic and resolving under-fitting has not been attributed to this change.