Supplementary Materials: Dropout as a Structured Shrinkage Prior

Eric Nalisnick [1] **José Miguel Hernández-Lobato** [1 2 3] **Padhraic Smyth** [4]

1. ADD as Multiplicative Noise and Equivalence to Stochastic Depth

Just as we did for ARD, we can re-formulate ADD in its expanded parametrization, revealing its equivalent MN form:

$$\boldsymbol{h}_{n,l} = f_l(\boldsymbol{h}_{n,l-1}\boldsymbol{W}_l) + \boldsymbol{h}_{n,l-1} \overset{\text{reparametrization}}{\longrightarrow} f_l(\tau_l \boldsymbol{h}_{n,l-1}\boldsymbol{W}_l) + \boldsymbol{h}_{n,l-1} \tag{1}$$

where τ_l is the ADD scale random variable. Comparing the equation above to the corresponding equation for ARD, we see that ADD is much less sampling intensive, requiring just one noise variable for each hidden layer.

Huang et al. (2016) proposed stochastic depth resnets by applying dropout to the resnet block, i.e. $\mathbf{a}_l = \lambda_l F_l(\boldsymbol{h}_{l-1}) + \boldsymbol{h}_{l-1}$ where $F(\cdot)$ denotes a whole resnet block and λ_l is a Bernoulli random variable. For simplicity, if we assume the resnet block consists of just one non-linear transformation, then applying Bernoulli ADD in MN form to the resnet architecture yields a similar expression, as seen in Equation 1. The only difference is that the Bernoulli variable is within f_l for ADD whereas it is outside the block for stochastic depth. However, if we assume f_l is the ReLU function, then the activation is scale equivariant. If the noise's support is non-negative, then it is equivalent to move the noise variable outside the activation: $\texttt{ReLU}(\tau_l \boldsymbol{h}_{l-1}\boldsymbol{W}_l) + \boldsymbol{h}_{l-1} = \tau_l \texttt{ReLU}(\boldsymbol{h}_{l-1}\boldsymbol{W}_l) + \boldsymbol{h}_{l-1}$ for $\tau \geq 0$. Thus, we can derive the previously proposed stochastic depth regularization as a special case of our ADD framework.

2. Parametrizations and their Effect on the Jensen's Gap

Given the equivalence between training under multiplicative noise and MC integration of the scale prior, we wonder: are we working with the correct parametrization? A first glance suggests that sampling noise in the hierarchical parametrization, thereby injecting it into the prior, could provide a more stable MC estimator than one obtained by perturbing the NN. We investigate this hypothesis below through analysis of the Jensen's gap of each estimator. It quickly becomes apparent that the hierarchical parametrization has undesirable properties when the hyper-prior is used as a regularizer. For all results, we assume $\mathbb{E}[\xi^2] \leq 1$, which is appropriate since we are interested in the settings in which the prior induces strong shrinkage, i.e. $\hat{\xi}^2 \approx 0$. The precise quantity we wish to analyze is $\mathcal{J}_{\text{GAP}} = \log p(\boldsymbol{y}, \{\boldsymbol{W}_l\}_{l=1}^{L+1} | \boldsymbol{X}) - \mathbb{E}_{p(\xi)} \left[\log p(\boldsymbol{y}, \{\boldsymbol{W}_l\}_{l=1}^{L+1} | \boldsymbol{X}, \{\boldsymbol{\Xi}_l\}_{l=1}^{L})\right]$ where the first term is the model joint with the scale integrated out and the second is the lower bound induced by invoking Jensen's inequality to extract the expectation.

We first consider the model in its expanded parametrization. Denote the Jensen's gap in this setting as $\mathcal{J}_{\text{EP-GAP}}$. The following proposition reveals the gap's series representation.

Proposition 2.1. *Let the NN that parametrizes the likelihood be a 2-Lipschitz function, let k denote an integer, let $\mu_\xi = \mathbb{E}[\xi]$, and let $Var[\xi]$ denote the variance of the scale (noise) distribution. The Jensen's gap of the expanded parametrization is:*

$$\begin{aligned}\mathcal{J}_{\textit{EP-GAP}} &= \frac{1}{2}\mathit{Var}[\xi] \left|\left|\nabla_{\mu_\xi} \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L})\right|\right|_2^2 - \sum_{k=2}^{\infty} \frac{(-1)^k}{k} \left[\frac{1}{2}\,\mathit{Var}[\xi]\,\boldsymbol{\Omega}\right]^k \\ &= \frac{1}{2}\mathit{Var}[\xi] \left|\left|\nabla_{\mu_\xi} \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L})\right|\right|_2^2 - \frac{1}{8}\mathit{Var}^2[\xi]\,\boldsymbol{\Omega}^2 + \mathcal{O}\left(\mathit{Var}^3[\xi]\right)\end{aligned}$$

[1] Department of Engineering, University of Cambridge, Cambridge, United Kingdom [2] Microsoft Research, Cambridge, United Kingdom [3] Alan Turing Institute [4] Department of Computer Science, University of California, Irvine, United States of America. Correspondence to: Eric Nalisnick <e.nalisnick@eng.cam.ac.uk>.

Pre-print. Work in progress.

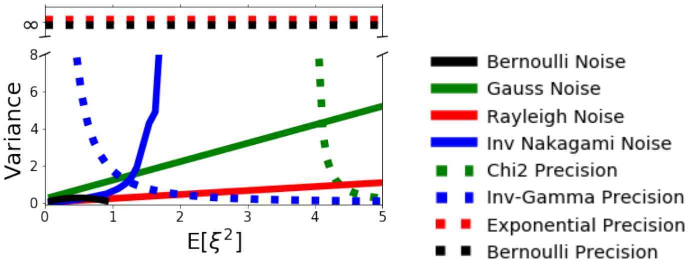

Figure 1. Noise Variances. The plot shows Var$[\xi]$ in solid lines and Var$[\xi^{-2}]$ in dashed lines for the noise distributions in Table 1 of the main text.

$$\textit{where}\;\; \boldsymbol{\Omega} = \left|\left|\nabla_{\mu_\xi} \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L})\right|\right|_2^2 + Tr\left\{\nabla^2_{\mu_\xi} \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L})\right\}$$

The proof is given in the supplementary materials. We see that the gap is a function of the variance of the scale and the norm of the log likelihood's gradient w.r.t. the scales. This latter quantity could be large. However, as $\mathbb{E}[\xi] \to 0$ so does $Var[\xi] \to 0$. We plot $\mathbb{E}[\xi^2]$ vs $Var[\xi]$ in Figure 1 (solid lines) for the noise distributions given in Table 1. Not only is $Var[\xi]$ near zero when the prior is encouraging strong shrinkage but we also see favorable scaling for all but the inverse Nakagami as $\mathbb{E}[\xi^2]$ grows.

We now move on to the hierarchical parametrization in which the random scale is left in the Gaussian prior. Denote the Jensen's gap as $\mathcal{J}_{\text{HP-GAP}}$, and we give its series representation below.

Proposition 2.2. *Let $p(\boldsymbol{W}|\boldsymbol{\Xi}^{-2})$ be a factorized, zero-mean Gaussian prior with ξ^{-2} denoting its precision. The Jensen's gap in the hierarchical parametrization is:*

$$\mathcal{J}_{\textit{HP-GAP}} = \frac{L}{2} \textit{Var}[\xi^{-2}] \; \left|\left|\nabla_{\mu_{\xi^{-2}}} \log p(\boldsymbol{W}|\mathbb{E}[\boldsymbol{\Xi}^{-2}])\right|\right|_2^2 - L \sum_{k=2}^{\infty} \frac{(-1)^k}{k} \left[\frac{1}{2}\; \textit{Var}[\xi^{-2}]\; \boldsymbol{\Psi}\right]^k$$

$$\textit{where}\;\; \boldsymbol{\Psi} = \left|\left|\nabla_{\mu_{\xi^{-2}}} \log p(\boldsymbol{W}|\mathbb{E}[\boldsymbol{\Xi}^{-2}])\right|\right|_2^2 - \frac{1}{2\mathbb{E}[\xi^{-2}]}.$$

The derivation is provided in the supplementary materials. We see that the gap depends on the variance of $1/\xi^2$ and the norm of the prior's gradient w.r.t. the precision. While the latter term is likely manageable, the former term is not. In Figure 1 we plot $\mathbb{E}[\xi^2]$ vs $Var[\xi^{-2}]$ (dashed lines) for the same scale distributions, and we see that either they asymptote as $\mathbb{E}[\xi^2] \to 0$—which occurs for Gaussian noise (green) at ~ 4 and for inverse Nakagami noise (blue) at ~ 0.5—or are infinite for the whole range—Bernoulli (black) and Rayleigh (red). Thus, the series will not converge when the shrinkage is strong, only when $\mathbb{E}[\xi^2]$ is well above one.

In summary, Proposition 2.1 shows that injecting noise into the likelihood relaxes the MAP estimate as a function of $Var[\xi]$ whereas Proposition 2.2 reveals noise in the Gaussian prior has a $Var[\xi^{-2}]$-order gap. When $\mathbb{E}[\xi^2]$ is near zero, $Var[\xi^{-2}]$ explodes for all noise distributions considered (Figure 1), resulting in an impractical objective.

Proposition 2.1: Expanded Parametrization Gap

$$\begin{aligned}
\mathcal{J}_{\text{EP-GAP}} &= \log p(\boldsymbol{y}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}|\boldsymbol{X}) - \mathbb{E}_{p(\xi)}\left[\log p(\boldsymbol{y}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}|\boldsymbol{X}, \{\boldsymbol{\Xi}_l\}_{l=1}^{L})\right] \\
&= \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}) + \log p(\{\boldsymbol{W}_l\}_{l=1}^{L+1}) - \mathbb{E}_{p(\xi)}\left[\log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\boldsymbol{\Xi}_l\}_{l=1}^{L})\right] - \log p(\{\boldsymbol{W}_l\}_{l=1}^{L+1}) \\
&= \log \mathbb{E}_{p(\xi)}\left[p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\boldsymbol{\Xi}_l\}_{l=1}^{L})\right] - \mathbb{E}_{p(\xi)}\left[\log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\boldsymbol{\Xi}_l\}_{l=1}^{L})\right].
\end{aligned}$$

Expanding $\mathbb{E}_{p(\xi)}\left[p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\boldsymbol{\Xi}_l\}_{l=1}^{L})\right]$ around $\mathbb{E}[\boldsymbol{\Xi}]$, we have

$$\mathbb{E}_{p(\xi)}\left[p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\boldsymbol{\Xi}_l\}_{l=1}^{L})\right] \approx p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L}) + \frac{1}{2} Var[\xi] Tr\left\{\nabla^2_{\mu_{xi}} p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L})\right\},$$

and dividing through by $p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L})$ inside the log yields:

$$\log \mathbb{E}_{p(\xi)}\left[p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\boldsymbol{\Xi}_l\}_{l=1}^{L})\right] \approx \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L}) + \log\left\{1 + \frac{\frac{1}{2}Var[\xi]Tr\left\{\nabla^2_{\mu_{xi}} p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L})\right\}}{p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L})}\right\}.$$

Noticing that

$$\nabla^2_{\mu_{xi}} p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L}) = p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L})\left[\left(\nabla_{\mu_{xi}} \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L})\right)^2 + \nabla^2_{\mu_{xi}} \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L})\right]$$

allows the previous expression to simplify to

$$\log \mathbb{E}_{p(\xi)}\left[p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\boldsymbol{\Xi}_l\}_{l=1}^{L})\right] \approx \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L}) + \log\left\{1 + \frac{1}{2}Var[\xi]Tr\left\{\boldsymbol{\Omega}\right\}\right\}$$

where $\boldsymbol{\Omega} = \left|\left|\nabla_{\mu_\xi} \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L})\right|\right|_2^2 + Tr\left\{\nabla^2_{\mu_\xi} \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L})\right\}$.

Next expanding the second term similarly around $\mathbb{E}[\boldsymbol{\Xi}]$, we have

$$\mathbb{E}_{p(\xi)}\left[\log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\boldsymbol{\Xi}_l\}_{l=1}^{L})\right] \approx \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L}) + \frac{1}{2}Var[\xi]Tr\left\{\nabla^2_{\mu_{xi}} \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L})\right\}.$$

Now we put both terms together while expanding the $\log\{\cdot\}$ around 1:

$$\begin{aligned}
\mathcal{J}_{\text{EP-GAP}} &\approx \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L}) + \log 1 - \sum_{k=1}^{\infty}\frac{(-1)^k}{k}\left[\frac{1}{2}Var[\xi]\boldsymbol{\Omega}\right]^k \\
&\quad - \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L}) - \frac{1}{2}Var[\xi]Tr\left\{\nabla^2_{\mu_{xi}} \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L})\right\} \\
&= -\sum_{k=1}^{\infty}\frac{(-1)^k}{k}\left[\frac{1}{2}Var[\xi]\boldsymbol{\Omega}\right]^k - \frac{1}{2}Var[\xi]Tr\left\{\nabla^2_{\mu_{xi}} \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L})\right\} \\
&= \frac{1}{2}Var[\xi]\boldsymbol{\Omega} - \frac{1}{2}Var[\xi]Tr\left\{\nabla^2_{\mu_{xi}} \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L})\right\} - \sum_{k=2}^{\infty}\frac{(-1)^k}{k}\left[\frac{1}{2}Var[\xi]\boldsymbol{\Omega}\right]^k \\
&= \frac{1}{2}Var[\xi]\left|\left|\nabla_{\mu_\xi} \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L})\right|\right|_2^2 - \sum_{k=2}^{\infty}\frac{(-1)^k}{k}\left[\frac{1}{2}Var[\xi]\boldsymbol{\Omega}\right]^k \\
&= \frac{1}{2}Var[\xi]\left|\left|\nabla_{\mu_\xi} \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}, \{\mathbb{E}[\boldsymbol{\Xi}_l]\}_{l=1}^{L})\right|\right|_2^2 - \frac{1}{8}Var^2[\xi]\boldsymbol{\Omega}^2 + \mathcal{O}(Var^3[\xi]).
\end{aligned}$$

We assume that $Var[\xi] \to 0$ (i.e. strong shrinkage) and this gives an $\mathcal{O}(Var^3[\xi])$ convergence.

Proposition 2.2: Hierarchical Parametrization Gap

$$\begin{aligned}
&\mathcal{J}_{\text{HP-GAP}} = \log p(\boldsymbol{y}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}|\boldsymbol{X}) - \mathbb{E}_{p(\xi)}\left[\log p(\boldsymbol{y}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}|\boldsymbol{X}, \{\boldsymbol{\Xi}_l\}_{l=1}^{L})\right] \\
&= \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}) + \log \mathbb{E}_{p(\xi)}\left[p(\{\boldsymbol{W}_l\}_{l=1}^{L+1}|\{\boldsymbol{\Xi}_l\}_{l=1}^{L})\right] - \log p(\boldsymbol{y}|\boldsymbol{X}, \{\boldsymbol{W}_l\}_{l=1}^{L+1}) - \mathbb{E}_{p(\xi)}\left[\log p(\{\boldsymbol{W}_l\}_{l=1}^{L+1}|\{\boldsymbol{\Xi}_l\}_{l=1}^{L})\right] \\
&= \log \mathbb{E}_{p(\xi)}\left[p(\{\boldsymbol{W}_l\}_{l=1}^{L+1}|\{\boldsymbol{\Xi}_l\}_{l=1}^{L})\right] - \mathbb{E}_{p(\xi)}\left[\log p(\{\boldsymbol{W}_l\}_{l=1}^{L+1}|\{\boldsymbol{\Xi}_l\}_{l=1}^{L})\right] \\
&= L\log \mathbb{E}_{p(\xi)}\left[p(\boldsymbol{W}|\boldsymbol{\Xi})\right] - L\mathbb{E}_{p(\xi)}\left[\log p(\boldsymbol{W}|\boldsymbol{\Xi})\right].
\end{aligned}$$

Next we reparametrize so that $\boldsymbol{\Xi}^{-2}$ denotes the Gaussian's precision. We then expand the first term around $\mathbb{E}[\boldsymbol{\Xi}^{-2}]$:

$$\mathbb{E}_{p(\xi)}\left[p(\boldsymbol{W}|\boldsymbol{\Xi}^{-2})\right] \approx p(\boldsymbol{W}|\mathbb{E}[\boldsymbol{\Xi}^{-2}]) + \frac{1}{2}Var[\xi^{-2}]Tr\left\{\nabla^2_{\mu_{\xi^{-2}}} p(\boldsymbol{W}|\mathbb{E}[\boldsymbol{\Xi}^{-2}])\right\}.$$

Dividing inside the log then allows us to write:

$$L\log\mathbb{E}_{p(\xi)}\left[p(\boldsymbol{W}|\boldsymbol{\Xi}^{-2})\right]\approx L\log p(\boldsymbol{W}|\mathbb{E}[\boldsymbol{\Xi}^{-2}])+L\log\left\{1+\frac{\frac{1}{2}Var[\xi^{-2}]Tr\left\{\nabla^2_{\mu_{\xi^{-2}}}p(\boldsymbol{W}|\mathbb{E}[\boldsymbol{\Xi}^{-2}])\right\}}{p(\boldsymbol{W}|\mathbb{E}[\boldsymbol{\Xi}^{-2}])}\right\}.$$

Noticing that

$$\nabla^2_{\mu_{\xi^{-2}}}p(\boldsymbol{W}|\mathbb{E}[\boldsymbol{\Xi}^{-2}])=p(\boldsymbol{W}|\mathbb{E}[\boldsymbol{\Xi}^{-2}])\left[\left(\nabla_{\mu_{\xi^{-2}}}\log p(\boldsymbol{W}|\mathbb{E}[\boldsymbol{\Xi}^{-2}])\right)^2+\nabla^2_{\mu_{\xi^{-2}}}\log p(\boldsymbol{W}|\mathbb{E}[\boldsymbol{\Xi}^{-2}])\right]$$

allows the previous expression to simplify to

$$L\log\mathbb{E}_{p(\xi)}\left[p(\boldsymbol{W}|\boldsymbol{\Xi}^{-2})\right]\approx L\log p(\boldsymbol{W}|\mathbb{E}[\boldsymbol{\Xi}^{-2}])+L\log\left\{1+\frac{1}{2}Var[\xi^{-2}]\boldsymbol{\Psi}\right\}$$

where

$$\boldsymbol{\Psi}=||\nabla_{\mu_{\xi^{-2}}}\log p(\boldsymbol{W}|\mathbb{E}[\boldsymbol{\Xi}^{-2}])||_2^2+Tr\left\{\nabla^2_{\mu_{\xi^{-2}}}\log p(\boldsymbol{W}|\mathbb{E}[\boldsymbol{\Xi}^{-2}])\right\}.$$

Next we can write out the second term as

$$\begin{aligned}
L\mathbb{E}_{p(\xi)}\left[\log p(\boldsymbol{W}|\boldsymbol{\Xi}^{-2})\right]&=\frac{-L}{2\sigma_0^2}\mathbb{E}[\xi^{-2}]||\boldsymbol{W}||_2^2+\frac{L}{2}\mathbb{E}[\log\xi^{-2}]-\frac{L}{2}\log 2\pi\\
&\approx\frac{-L}{2\sigma_0^2}\mathbb{E}[\xi^{-2}]||\boldsymbol{W}||_2^2+\frac{L}{2}\log\mathbb{E}[\xi^{-2}]-\frac{L}{2}\log 2\pi+\frac{L}{4}Var[\xi^{-2}]\frac{-1}{\mathbb{E}[\xi^{-2}]^2}\\
&=L\log p(\boldsymbol{W}|\mathbb{E}[\boldsymbol{\Xi}^{-2}])-Var[\xi^{-2}]\frac{L}{4\mathbb{E}[\xi^{-2}]^2}.
\end{aligned}$$

Rejoining the expressions, again expanding the log around 1, we have:

$$\begin{aligned}
\mathcal{J}_{\text{HP-GAP}}&\approx L\log p(\boldsymbol{W}|\mathbb{E}[\boldsymbol{\Xi}^{-2}])+L\log 1-L\sum_{k=1}^{\infty}\frac{(-1)^k}{k}\left[\frac{1}{2}Var[\xi^{-2}]\boldsymbol{\Psi}\right]^k-L\log p(\boldsymbol{W}|\mathbb{E}[\boldsymbol{\Xi}^{-2}])+Var[\xi^{-2}]\frac{L}{4\mathbb{E}[\xi^{-2}]^2}\\
&=-L\sum_{k=1}^{\infty}\frac{(-1)^k}{k}\left[\frac{1}{2}Var[\xi^{-2}]\boldsymbol{\Psi}\right]^k+Var[\xi^{-2}]\frac{L}{4\mathbb{E}[\xi^{-2}]^2}\\
&=\frac{L}{2}Var[\xi^{-2}]\boldsymbol{\Psi}+\frac{L}{2}Var[\xi^{-2}]\frac{1}{2\mathbb{E}[\xi^{-2}]^2}-L\sum_{k=2}^{\infty}\frac{(-1)^k}{k}\left[\frac{1}{2}Var[\xi^{-2}]\boldsymbol{\Psi}\right]^k\\
&=\frac{L}{2}Var[\xi^{-2}]||\nabla_{\mu_{\xi^{-2}}}\log p(\boldsymbol{W}|\mathbb{E}[\boldsymbol{\Xi}^{-2}])||_2^2-L\sum_{k=2}^{\infty}\frac{(-1)^k}{k}\left[\frac{1}{2}Var[\xi^{-2}]\boldsymbol{\Psi}\right]^k.
\end{aligned}$$

3. Variational EM Updates

3.1. ADD EM Updates

Below we derive the EM updates for the various hyper-priors considered. For all expressions, we assume the variational distribution on the weights factorizes, i.e. $\text{N}(\boldsymbol{W};\boldsymbol{\phi})=\prod_i\prod_j\text{N}(w_{i,j};\boldsymbol{\phi}_{i,j})$ where i denotes row and j column indices.

3.1.1. Inverse Gamma

We begin with the ELBO terms that depend on the variational variance τ:

$$\begin{aligned}
\mathcal{J}_{\text{ELBO}}(\boldsymbol{\phi},\bar{\tau}_l)&=-\text{KL}\left[\text{N}(\boldsymbol{W};\boldsymbol{\phi})||\text{N}(\boldsymbol{W}|\bar{\tau}_l)\right]+\log p(\bar{\tau}_l)\\
&=\frac{-1}{2}\sum_i\sum_j\left[\log\frac{\bar{\tau}_l}{\sigma_{i,j}^2}+\frac{\sigma_{i,j}^2}{\bar{\tau}_l}+\frac{\mu_{i,j}^2}{\bar{\tau}_l}-1\right]+\log\Gamma^{-1}(\bar{\tau}_l;\alpha,\beta)\\
&=\frac{-1}{2}\left[\sum_i\sum_j\log\frac{\bar{\tau}_l}{\sigma_{i,j}^2}+\sum_i\sum_j\frac{\sigma_{i,j}^2+\mu_{i,j}^2}{\bar{\tau}_l}-D\right]-\frac{\beta}{\bar{\tau}_l}-(\alpha+1)\log\bar{\tau}_l+\log\frac{\beta^\alpha}{\Gamma(\alpha)}.
\end{aligned}$$

Differentiating w.r.t. $\bar{\tau}_l$ and setting to zero, we have

$$\begin{aligned}
0 &= \frac{\partial}{\partial \bar{\tau}_l} \mathcal{J}_{\text{ELBO}}(\boldsymbol{\phi}, \bar{\tau}_l) \\
&= \frac{-1}{2}\left[\sum_i \sum_j \frac{1}{\bar{\tau}_l} - \frac{\sigma_{i,j}^2 + \mu_{i,j}^2}{\bar{\tau}_l^2}\right] + \frac{\beta}{\bar{\tau}_l^2} - \frac{\alpha+1}{\bar{\tau}_l} \\
&= \frac{D}{\bar{\tau}_l} - \frac{\sum_i \sum_j \sigma_{i,j}^2 + \mu_{i,j}^2}{\bar{\tau}_l^2} - \frac{2\beta}{\bar{\tau}_l^2} + \frac{2\alpha+2}{\bar{\tau}_l} \\
\frac{2\beta + \sum_i \sum_j \sigma_{i,j}^2 + \mu_{i,j}^2}{\bar{\tau}_l^2} &= \frac{D + 2\alpha + 2}{\bar{\tau}_l} \\
\bar{\tau}_l^* &= \frac{2\beta + \sum_i \sum_j \sigma_{i,j}^2 + \mu_{i,j}^2}{D + 2\alpha + 2}
\end{aligned}$$

where D are the number of dimensions (i.e. parameters) in $\boldsymbol{W}$.

3.1.2. HALF-CAUCHY

Again starting with the ELBO, we have

$$\begin{aligned}
0 &= \frac{\partial}{\partial \bar{\tau}_l} \mathcal{J}_{\text{ELBO}}(\boldsymbol{\phi}, \bar{\tau}_l) \\
&= \frac{-1}{2} \frac{\partial}{\partial \bar{\tau}_l} \sum_i \sum_j \left[\log \frac{\bar{\tau}_l}{\sigma_{i,j}^2} + \frac{\sigma_{i,j}^2}{\bar{\tau}_l} + \frac{\mu_{i,j}^2}{\bar{\tau}_l} - 1\right] + \frac{\partial}{\partial \bar{\tau}_l} \log C^+(\sqrt{\bar{\tau}_l}; 0, \eta) \\
&= \frac{-1}{2}\left[\sum_i \sum_j \frac{1}{\bar{\tau}_l} - \frac{\sigma_{i,j}^2 + \mu_{i,j}^2}{\bar{\tau}_l^2}\right] + \frac{\partial}{\partial \bar{\tau}_l} \log \frac{2}{\pi\eta(1 + \bar{\tau}_l/\eta^2)} \\
&= \frac{D}{\bar{\tau}_l} - \frac{\sum_i \sum_j \sigma_{i,j}^2 + \mu_{i,j}^2}{\bar{\tau}_l^2} + \frac{2}{\eta^2 + \bar{\tau}_l} \\
&= \frac{2\bar{\tau}_l}{\eta^2 + \bar{\tau}_l} - \frac{\sum_i \sum_j \sigma_{i,j}^2 + \mu_{i,j}^2}{\bar{\tau}_l} + D.
\end{aligned}$$

Solving[1] the above equation for $\bar{\tau}_l$ gives for a positive solution:

$$\bar{\tau}_l^* = \frac{M - \eta^2 D + \sqrt{M^2 + (2D+8)\eta^2 M + \eta^4 D^2}}{2D + 4} \quad \text{where} \quad M = \sum_i \sum_j \sigma_{i,j}^2 + \mu_{i,j}^2,$$

D is again the dimensionality, and η is the half-Cauchy's scale.

[1] We plugged the equation into Wolfram Alpha.

3.1.3. Log-Uniform

Again beginning with ELBO, we have

$$\begin{aligned}
0 &= \frac{\partial}{\partial \bar{\tau}_l} \mathcal{J}_{\text{ELBO}}(\boldsymbol{\phi}, \bar{\tau}_l) \\
&= \frac{-1}{2}\left[\sum_i \sum_j \frac{1}{\bar{\tau}_l} - \frac{\sigma_{i,j}^2 + \mu_{i,j}^2}{\bar{\tau}_l^2}\right] + \frac{\partial}{\partial \bar{\tau}_l} \log \frac{c}{\bar{\tau}_l} \\
&= \frac{D}{\bar{\tau}_l} - \frac{\sum_i \sum_j \sigma_{i,j}^2 + \mu_{i,j}^2}{\bar{\tau}_l^2} + \frac{2}{\bar{\tau}_l} \\
\frac{\sum_i \sum_j \sigma_{i,j}^2 + \mu_{i,j}^2}{\bar{\tau}_l^2} &= \frac{D+2}{\bar{\tau}_l} \\
\bar{\tau}_l^* &= \frac{\sum_i \sum_j \sigma_{i,j}^2 + \mu_{i,j}^2}{D+2}.
\end{aligned}$$

Notice that this update is the same as the inverse Gamma's as $\alpha \to 0$ and $\beta \to 0$.

3.2. ARD-ADD EM Updates

For approximate posterior inference for the joint ARD-ADD prior, we assume the posterior approximation

$$p(\boldsymbol{W}, \lambda, \tau | \boldsymbol{y}, \boldsymbol{X}) \approx q(\boldsymbol{W}; \boldsymbol{\phi})\, q(\lambda)\, q(\tau) = \text{N}(\boldsymbol{W}; \boldsymbol{\mu}_{\boldsymbol{\phi}}, \text{daig}\{\boldsymbol{\Sigma}_{\boldsymbol{\phi}}\})\, \delta[\bar{\lambda}_i]\, \delta[\bar{\tau}_l]$$

where $\boldsymbol{\phi} = \{\boldsymbol{\mu}_{\boldsymbol{\phi}}, \text{diag}\{\boldsymbol{\Sigma}_{\boldsymbol{\phi}}\}\}$, $\bar{\lambda}_i$, and $\bar{\tau}_l$ are the variational parameters. The ELBO for this approximation is

$$\begin{aligned}
&\log p(\boldsymbol{y}|\boldsymbol{X}) \geq \\
&\mathbb{E}_{q(\boldsymbol{W})}\left[\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{W})\right] - \mathbb{E}_{q(\lambda)}\mathbb{E}_{q(\tau)}\text{KL}\left[q(\boldsymbol{W}; \boldsymbol{\phi})||p(\boldsymbol{W}|\lambda, \tau)\right] - \text{KL}\left[q(\lambda)||p(\lambda)\right] - \text{KL}\left[q(\tau)||p(\tau)\right] \\
&= \mathbb{E}_{\text{N}(\boldsymbol{W})}\left[\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{W})\right] - \text{KL}\left[\text{N}(\boldsymbol{W}; \boldsymbol{\phi})||\text{N}(\boldsymbol{W}|\bar{\lambda}_i, \bar{\tau}_l)\right] - \text{KL}\left[\delta[\bar{\lambda}_i]||p(\lambda)\right] - \text{KL}\left[\delta[\bar{\tau}_l]||p(\tau)\right] \\
&= \mathbb{E}_{\text{N}(\boldsymbol{W})}\left[\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{W})\right] - \text{KL}\left[\text{N}(\boldsymbol{W}; \boldsymbol{\phi})||\text{N}(\boldsymbol{W}|\bar{\lambda}_i, \bar{\tau}_l)\right] + \log p(\bar{\lambda}_i) + \log p(\bar{\tau}_l) + \mathcal{C}
\end{aligned} \tag{2}$$

where $\mathcal{C} = \mathbb{H}[\delta[\bar{\lambda}_i]] + \mathbb{H}[\delta[\bar{\tau}_l]]$ is again a constant. Next we must find solutions for both $\bar{\lambda}_i$ and $\bar{\tau}_l$. Again we can differentiate the ELBO and set to zero:

$$\begin{aligned}
\frac{\partial}{\partial \bar{\lambda}_i}\mathcal{J}_{\text{ELBO}}(\boldsymbol{\phi}, \bar{\lambda}_i, \bar{\tau}_l) &= -\frac{\partial}{\partial \bar{\lambda}_i}\text{KL}\left[\text{N}(\boldsymbol{W}; \boldsymbol{\phi})||\text{N}(\boldsymbol{W}|\bar{\lambda}_i, \bar{\tau}_l)\right] + \frac{\partial}{\partial \bar{\lambda}_i}\log p(\bar{\lambda}_i) = 0, \\
\frac{\partial}{\partial \bar{\tau}_l}\mathcal{J}_{\text{ELBO}}(\boldsymbol{\phi}, \bar{\lambda}_i, \bar{\tau}_l) &= -\frac{\partial}{\partial \bar{\tau}_l}\text{KL}\left[\text{N}(\boldsymbol{W}; \boldsymbol{\phi})||\text{N}(\boldsymbol{W}|\bar{\lambda}_i, \bar{\tau}_l)\right] + \frac{\partial}{\partial \bar{\tau}_l}\log p(\bar{\tau}_l) = 0.
\end{aligned} \tag{3}$$

We denote the solutions as $\bar{\lambda}_i^*$ and $\bar{\tau}_l^*$, deriving them for the three hyper-priors below. Again, we update $q(\boldsymbol{W}; \boldsymbol{\phi})$ via gradient ascent:

$$\frac{\partial}{\partial \boldsymbol{\phi}}\mathcal{J}_{\text{ELBO}}(\boldsymbol{\phi}, \bar{\lambda}_i^*, \bar{\tau}_l^*) = \frac{\partial}{\partial \boldsymbol{\phi}}\mathbb{E}_{\text{N}(\boldsymbol{W};\boldsymbol{\phi})}\left[\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{W})\right] - \frac{\partial}{\partial \boldsymbol{\phi}}\text{KL}\left[\text{N}(\boldsymbol{W}; \boldsymbol{\phi})||\text{N}(\boldsymbol{W}|\bar{\lambda}_i^*, \bar{\tau}_l^*)\right]. \tag{4}$$

3.2.1. Inverse Gamma

Starting with the first line of Equation 3, we have

$$
\begin{aligned}
0 &= \frac{\partial}{\partial\bar{\lambda}_i}\mathcal{J}_{\text{ELBO}}(\boldsymbol{\phi},\bar{\lambda}_i,\bar{\tau}_l)\\
&= \frac{-1}{2}\frac{\partial}{\partial\bar{\lambda}_i}\sum_i\sum_j\left[\log\frac{\bar{\lambda}_i\bar{\tau}_l}{\sigma^2_{i,j}}+\frac{\sigma^2_{i,j}}{\bar{\lambda}_i\bar{\tau}_l}+\frac{\mu^2_{i,j}}{\bar{\lambda}_i\bar{\tau}_l}-1\right]+\frac{\partial}{\partial\bar{\lambda}_i}\log\Gamma^{-1}(\bar{\lambda}_i;\alpha,\beta)\\
&= \frac{-1}{2}\left[\sum_j\frac{1}{\bar{\lambda}_i}-\frac{\sigma^2_{i,j}+\mu^2_{i,j}}{\lambda_i^2\bar{\tau}_l}\right]+\frac{\beta}{\bar{\lambda}_i^2}-\frac{\alpha+1}{\bar{\lambda}_i}\\
&= \frac{D_i}{\bar{\lambda}_i}-\frac{\sum_j\sigma^2_{i,j}+\mu^2_{i,j}}{\bar{\lambda}_i^2\bar{\tau}_l}-\frac{2\beta}{\bar{\lambda}_i^2}+\frac{2\alpha+2}{\bar{\lambda}_i}\\
\frac{2\beta\bar{\tau}_l+\sum_j\sigma^2_{i,j}+\mu^2_{i,j}}{\bar{\lambda}_i^2\bar{\tau}_l} &= \frac{D_i+2\alpha+2}{\bar{\lambda}_i}\\
\bar{\lambda}_i^* &= \frac{2\beta\bar{\tau}_l+\sum_j\sigma^2_{i,j}+\mu^2_{i,j}}{\bar{\tau}_l(D_i+2\alpha+2)}
\end{aligned}
$$

where D_i is the number of parameters in the ith row (i.e. λ's corresponding row) of the weight matrix $\boldsymbol{W}$. Furthermore, notice that the sum over posterior parameters is across only columns (j index).

Moving on the ADD parameter, we begin with the second line of Equation 3:

$$
\begin{aligned}
0 &= \frac{\partial}{\partial\bar{\tau}_l}\mathcal{J}_{\text{ELBO}}(\boldsymbol{\phi},\bar{\lambda}_i,\bar{\tau}_l)\\
&= \frac{-1}{2}\frac{\partial}{\partial\bar{\tau}_l}\sum_i\sum_j\left[\log\frac{\bar{\lambda}_i\bar{\tau}_l}{\sigma^2_{i,j}}+\frac{\sigma^2_{i,j}}{\bar{\lambda}_i\bar{\tau}_l}+\frac{\mu^2_{i,j}}{\bar{\lambda}_i\bar{\tau}_l}-1\right]+\frac{\partial}{\partial\bar{\tau}_l}\log\Gamma^{-1}(\bar{\tau}_l;\alpha,\beta)\\
&= \frac{-1}{2}\left[\sum_i\sum_j\frac{1}{\bar{\tau}_l}-\frac{\sigma^2_{i,j}+\mu^2_{i,j}}{\bar{\lambda}_i\bar{\tau}_l^2}\right]+\frac{\beta}{\bar{\tau}_l^2}-\frac{\alpha+1}{\bar{\tau}_l}\\
&= \frac{D}{\bar{\tau}_l}-\frac{\sum_i\bar{\lambda}_i^{-1}\sum_j\sigma^2_{i,j}+\mu^2_{i,j}}{\bar{\tau}_l^2}-\frac{2\beta}{\bar{\tau}_l^2}+\frac{2\alpha+2}{\bar{\tau}_l}\\
\frac{2\beta+\sum_i\bar{\lambda}_i^{-1}\sum_j\sigma^2_{i,j}+\mu^2_{i,j}}{\bar{\tau}_l^2} &= \frac{D+2\alpha+2}{\bar{\tau}_l}\\
\bar{\tau}_l^* &= \frac{2\beta+\sum_i\bar{\lambda}_i^{-1}\sum_j\sigma^2_{i,j}+\mu^2_{i,j}}{D+2\alpha+2}.
\end{aligned}
$$

3.2.2. Half-Cauchy

Starting with the first line of Equation 3, we have

$$
\begin{aligned}
0 &= \frac{\partial}{\partial\bar{\lambda}_i}\mathcal{J}_{\text{ELBO}}(\boldsymbol{\phi},\bar{\lambda}_i,\bar{\tau}_l)\\
&= \frac{-1}{2}\frac{\partial}{\partial\bar{\lambda}_i}\sum_i\sum_j\left[\log\frac{\bar{\lambda}_i\bar{\tau}_l}{\sigma^2_{i,j}}+\frac{\sigma^2_{i,j}}{\bar{\lambda}_i\bar{\tau}_l}+\frac{\mu^2_{i,j}}{\bar{\lambda}_i\bar{\tau}_l}-1\right]+\frac{\partial}{\partial\bar{\lambda}_i}\log C^{+}(\sqrt{\bar{\lambda}_i};0,\eta)\\
&= \sum_j\frac{1}{\bar{\lambda}_i}-\frac{\sigma^2_{i,j}+\mu^2_{i,j}}{\lambda_i^2\bar{\tau}_l}+\frac{2}{\eta^2+\bar{\lambda}_i}\\
&= \frac{2\bar{\lambda}_i}{\eta^2+\bar{\lambda}_i}-\frac{\bar{\tau}_l^{-1}\sum_j\sigma^2_{i,j}+\mu^2_{i,j}}{\bar{\lambda}_i}+D_i
\end{aligned}
$$

where D_i is the number of parameters in the ith row. This final expression is of the same form we solved for ADD and therefore has the same solution with appropriately adjusted constants:

$$\bar{\lambda}_i^* = \frac{M_i - \eta^2 D_i + \sqrt{M_i^2 + (2D_i + 8)\eta^2 M_i + \eta^4 D_i^2}}{2D_i + 4} \quad \text{where} \quad M_i = \bar{\tau}_l^{-1} \sum_j \sigma_{i,j}^2 + \mu_{i,j}^2.$$

Moving on the ADD parameter, we consider the second line of Equation 3:

$$\begin{aligned}
0 &= \frac{\partial}{\partial \bar{\tau}_l} \mathcal{J}_{\text{ELBO}}(\boldsymbol{\phi}, \bar{\lambda}_i, \bar{\tau}_l) \\
&= \frac{-1}{2} \frac{\partial}{\partial \bar{\tau}_l} \sum_i \sum_j \left[\log \frac{\bar{\lambda}_i \bar{\tau}_l}{\sigma_{i,j}^2} + \frac{\sigma_{i,j}^2}{\bar{\lambda}_i \bar{\tau}_l} + \frac{\mu_{i,j}^2}{\bar{\lambda}_i \bar{\tau}_l} - 1 \right] + \frac{\partial}{\partial \bar{\tau}_l} \log C^+(\sqrt{\bar{\tau}_l}; 0, \eta) \\
&= \sum_i \sum_j \frac{1}{\bar{\tau}_l} - \frac{\sigma_{i,j}^2 + \mu_{i,j}^2}{\bar{\lambda}_i \bar{\tau}_l^2} + \frac{2}{\eta^2 + \bar{\tau}_l} \\
&= \frac{D}{\bar{\tau}_l} - \frac{\sum_i \bar{\lambda}_i^{-1} \sum_j \sigma_{i,j}^2 + \mu_{i,j}^2}{\bar{\tau}_l^2} + \frac{2}{\eta^2 + \bar{\tau}_l} \\
&= \frac{2\bar{\tau}_l}{\eta^2 + \bar{\tau}_l} - \frac{\sum_i \bar{\lambda}_i^{-1} \sum_j \sigma_{i,j}^2 + \mu_{i,j}^2}{\bar{\tau}_l} + D.
\end{aligned}$$

Again solving the same equation yields the solution:

$$\bar{\tau}_l^* = \frac{M_\lambda - \eta^2 D + \sqrt{M_\lambda^2 + (2D + 8)\eta^2 M_\lambda + \eta^4 D^2}}{2D + 4} \quad \text{where} \quad M_\lambda = \sum_i \bar{\lambda}_i^{-1} \sum_j \sigma_{i,j}^2 + \mu_{i,j}^2.$$

3.2.3. Log-Uniform

As mentioned earlier, the solution for the log-uniform distribution can be attained in the limit of the inverse Gamma. Plugging $\alpha = 0$ and $\beta = 0$ into the inverse gamma's solutions we obtain:

$$\bar{\lambda}_i^* = \frac{\sum_j \sigma_{i,j}^2 + \mu_{i,j}^2}{\bar{\tau}_l (D_i + 2)} \quad \text{and} \quad \bar{\tau}_l^* = \frac{\sum_i \bar{\lambda}_i^{-1} \sum_j \sigma_{i,j}^2 + \mu_{i,j}^2}{D + 2}.$$

4. Experimental Details

Below we provide more details on our experimental set-up. For all experiments, we used following settings, replicating the set-up of Gal & Ghahramani (2016). Optimization was done using Adam (Kingma & Ba, 2014) with a learning rate of .001 and other parameters kept at Tensorflow defaults. Training was done for 4500 epochs using batch sizes of 32 and 10 samples when calculating Monte Carlo expectations. The UCI regression data sets were divided into 20 randomized $90\% - 10\%$ train-test splits. Results are reported by training and evaluating on each split and reporting the average test metric. Test set evaluations were done by drawing 500 MC samples. The training split was standardized to zero mean and unit variance, and predictions were made by normalizing the test input based on these train set statistics and then applying the inverse transform to the prediction. RMSE and likelihood were calculated after applying the inverse transform.

Importance Sampling Experiments All results in Table 2 of the main text used a one-hidden-layer neural network with 50 hidden units and ReLU activations. We used the dropout rates and likelihood precisions reported in Table 1. These parameters were obtained from the validation set results reported by Gal & Ghahramani (2016) at: `https://github.com/yaringal/DropoutUncertaintyExps/tree/master/UCI_Datasets`.

Variational EM Experiments The ARD, ADD, and ARD-ADD results in Table 3 of the main text used neural networks with two hidden layers with one skip connection between them. ReLU activations were again used exclusively. ARD was applied to the first and last weight matrices—since ADD can't be applied without a skip connection—and thus the

	Dropout Parameter (p)	Likelihood Precision
`Boston`	.050	0.20
`Concrete`	.005	0.075
`Energy`	.005	0.75
`Kin8nm`	.010	250.00
`Power`	.005	0.10
`Wine`	.050	3.50
`Yacht`	.005	0.75

Table 1. The table above reports the dropout rate and likelihood precision used for each experiment. We use the same parameters as Gal & Ghahramani (2016).

differences in performance are due to the prior's influence on the hidden-to-hidden weights. We used *flipout* (Wen et al., 2018) to reduce the variance of the Monte Carlo expectations (10 samples). All scale priors were set to the same distribution, and we chose this distribution based on preliminary experiments on a subset of the `Protein` UCI data set (Dheeru & Karra Taniskidou, 2017). We tried the following priors: $\Gamma^{-1}(0,0)$ / log-uniform, $\Gamma^{-1}(1,1)$, $\Gamma^{-1}(3,3)$, $\Gamma^{-1}(5,5)$, $\Gamma^{-1}(1,3)$, $\Gamma^{-1}(1,5)$, $\Gamma^{-1}(1,10)$, $C^{+}(0,.5)$, $C^{+}(0,1.)$, $C^{+}(0,2.5)$, $C^{+}(0,5)$, $C^{+}(0,10.)$. $\Gamma^{-1}(3,3)$ performed best and we used it for all experiments. The fact that this one prior that was *chosen on a different data set* generalized so well speaks to the robustness of the EM algorithm.

5. Tensorflow Implementations

We have provided Python code to replicate the results presented in Tables 2 and 3 of the main text. We include the `Yacht` data set for the user's convenience; the other data sets can be downloaded from Gal & Ghahramani (2016)'s GitHub repository[2] or from the UCI repository (Dheeru & Karra Taniskidou, 2017). Below we give implementation snippets for the paper's core algorithmic contributions.

5.1. Implementation of Tail-Adaptive Importance Weights

```

log_likelihoods = # matrix of size Batch_Size x Number_of_MC_Samples

# Sort the samples
log_likelihoods_sorted, _ = tf.nn.top_k(log_likelihoods, k=n_samples, sorted=True)

# Use tf.range to provide ranks
gamma_vals = n_samples / (tf.expand_dims(tf.range(n_samples, dtype=tf.float32), 0) + 1.)

# Normalize and treat weight as a constant
normalized_weights = tf.stop_gradient(gamma_vals / tf.reduce_sum(gamma_vals, axis=1,
    keep_dims=True))

# Apply weights
final_ell_term = tf.reduce_sum(normalized_weights * log_likelihoods_sorted,
    reduction_indices=1, keep_dims=True)

# Return final expected log-likelihood term s.t. gradient will be w * d_ell
return tf.reduce_mean(final_ell_term)
```

5.2. EM Scale Updates

```
def ard_scale_update(prior_id, prior_params, mu, std):
    param_count = mu.get_shape().as_list()[1]
    moments = tf.reduce_sum(mu*mu + std*std, reduction_indices=1, keep_dims=True)

    if prior_id == "inv_gamma":
```

[2] `https://github.com/yaringal/DropoutUncertaintyExps/tree/master/UCI_Datasets`

```
        prior_opt_scale = tf.sqrt((2. * prior_params['param2'] + moments) /  (param_count + 2. * prior_params['param1'] + 2.))

    elif prior_id == "half_cauchy":
        prior_opt_scale = tf.sqrt((moments - prior_params['param1']**2 * param_count + tf.sqrt(moments**2 + (2*param_count+8) * prior_params['param1']**2 * moments + prior_params['param1']**4 * param_count**2)) / (2*param_count + 4))

    # Log-uniform implemented by just passing alpha=beta=0 to inv Gamma

    return prior_opt_scale * tf.ones_like(std)


def add_scale_update(prior_id, prior_params, mu, std):
    param_count = mu.get_shape().as_list()[0] * mu.get_shape().as_list()[1]
    moments = tf.reduce_sum(mu**2 + std**2)

    if prior_id== "inv_gamma":
        prior_opt_scale = tf.sqrt((2. * prior_params['param2'] + moments) /  (param_count + 2. * prior_params['param1'] + 2.))

    elif prior_id == "half_cauchy":
        prior_opt_scale = tf.sqrt((moments - prior_params['param1']**2 * param_count + tf.sqrt(moments**2 + (2*param_count+8.) * prior_params['param1']**2 * moments + prior_params['param1']**4 * param_count**2)) /(2*param_count + 4))

    return prior_opt_scale * tf.ones_like(std)


def ard_add_scale_update(prior_id, prior_params, old_tau, mu, std):
    param_count_cols = mu.get_shape().as_list()[1]
    param_count_rows = mu.get_shape().as_list()[0]
    moments = tf.reduce_sum(mu**2 + std**2, reduction_indices=1, keep_dims=True)

    if prior_id == "inv_gamma":
        opt_lambda = tf.sqrt( (2. * prior_params['param2'] * old_tau + moments) / (old_tau * 2. * prior_params['param1'] + old_tau * param_count_cols + old_tau * 2))
        opt_tau = tf.sqrt( (2. * prior_params['param2'] + tf.reduce_sum((1./opt_lambda) * moments)) / (param_count_cols * param_count_rows + 2 + 2 * prior_params['param1']) )

    if prior_id == "half_cauchy":
        moments_tau = (1./old_tau) * moments
        opt_lambda = tf.sqrt((moments_tau - prior_params['param1']**2 * param_count_cols + tf.sqrt(moments_tau**2 + (2*param_count_cols+8) * prior_params['param1']**2 * moments_tau + prior_params['param1']**4 * param_count_cols**2)) / (2*param_count_cols + 4))

        moments_lamb = tf.reduce_sum((1./opt_lambda) * moments)
        param_count = param_count_rows * param_count_cols
        opt_tau = tf.sqrt((moments_lamb - prior_params['param1']**2 * param_count + tf.sqrt(moments_lamb**2 + (2*param_count+8) * prior_params['param1']**2 * moments_lamb + prior_params['param1']**4 * param_count**2)) / (2*param_count + 4))

    # Log-uniform implemented by just passing alpha=beta=0 to inv Gamma

    return opt_lambda * opt_tau * tf.ones_like(std)
```

References

Dheeru, D. and Karra Taniskidou, E. UCI Machine Learning Repository, 2017. URL http://archive.ics.uci.edu/ml.

Gal, Y. and Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 1050–1059, 2016.

Huang, G., Sun, Y., Liu, Z., Sedra, D., and Weinberger, K. Q. Deep Networks with Stochastic Depth. In *European Conference on Computer Vision (ECCV)*, pp. 646–661. Springer, 2016.

Kingma, D. and Ba, J. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*, 2014.

Wen, Y., Vicol, P., Ba, J., Tran, D., and Grosse, R. Flipout: Efficient Pseudo-Independent Weight Perturbations on Mini-Batches. In *International Conference on Learning Representations*, 2018.