Supplementary materials of HAFVF

Anonymous

June 5, 2018

1 Proof of Proposition 1

Proof. To apply the NCVMP algorithm [Knowles and Minka(2011)], we first need to compute the inverse posterior covariance of the sufficient statistics of the beta distribution:

$$C(\mathbf{T}(w)\,|\,\boldsymbol{\phi}) = \left[\begin{array}{cc} \psi_1(\phi^\alpha) - \psi_1(\phi^\alpha+\phi^\beta) & -\psi_1(\phi^\alpha+\phi^\beta) \\ -\psi_1(\phi^\alpha+\phi^\beta) & \psi_1(\phi^\beta) - \psi_1(\phi^\alpha+\phi^\beta) \end{array}\right]$$

Next, we need to take the derivative of the expected log-joint probability wrt. $\boldsymbol{\phi}$. Noting that $\mathbb{E}_{q(w)}[w] = \frac{\phi^\alpha}{\phi^\alpha+\phi^\beta}$ and that $\mathbb{E}_{q(w)}[w]$ intervenes twice in the expression of the ELBO (in $\mathbb{E}_q\Big[\log p(\mathbf{z}\,|\,\widehat{\boldsymbol{\vartheta}})\Big]$) and in $C(\mathbf{T}(\mathbf{z})\,|\,\widehat{\boldsymbol{\vartheta}})$), we can use the chain rule and write:

$$\begin{aligned} C(\mathbf{T}(w)\,|\,\boldsymbol{\phi})^{-1}\nabla_{\boldsymbol{\phi}}\Big\{\mathbb{E}_{q(\mathbf{z})}\Big[\log p(\mathbf{z}\,|\,\widehat{\boldsymbol{\vartheta}})\Big] + \mathbb{E}_{q(w)}[\log p(\boldsymbol{\phi}\,|\,\widehat{\boldsymbol{\varphi}})] - \frac{1}{2}\,\mathbb{V}\mathrm{ar}_{q(w)}[w]C(\mathbf{T}(\mathbf{z})\,|\,\widehat{\boldsymbol{\vartheta}})\Big\} = \\ C(\mathbf{T}(w)\,|\,\boldsymbol{\phi})^{-1}\Big(\nabla_{\boldsymbol{\phi}}\mathbb{E}_{q(w)}[w]\,\delta_{\mathrm{L}} + \left[\begin{array}{c} \varphi^\alpha\left(\psi_1(\phi^\alpha) - \psi_1(\phi^\alpha+\phi^\beta)\right) - \varphi^\beta\psi_1(\phi^\alpha+\phi^\beta) \\ \varphi^\beta\left(\psi_1(\phi^\beta) - \psi_1(\phi^\alpha+\phi^\beta)\right) - \varphi^\alpha\psi_1(\phi^\alpha+\phi^\beta) \end{array}\right] + \\ \frac{1}{2}\Big(\nabla_{\boldsymbol{\phi}}\left\{\mathbb{V}\mathrm{ar}_{q(w)}[w]\right\}C(\mathbf{T}(\mathbf{z})\,|\,\widehat{\boldsymbol{\vartheta}}) + \mathbb{V}\mathrm{ar}_{q(w)}[w]\nabla_{\boldsymbol{\phi}_t}\left\{\mathbb{E}_{q_t(w)}[w]\right\}\frac{d}{d\widehat{w}}C(\mathbf{T}(\mathbf{z})\,|\,\widehat{\boldsymbol{\vartheta}})\Big)\Big) \end{aligned}$$

Expanding this final expression gives back the expression in proposition 1. □

Corollary 1. *For a given value of δ_{L} and δ_{C}, $\phi_t^\alpha > \phi_t^\beta$ implies that ϕ_t^α is more affected (positively or negatively) than ψ_t^β by δ_{L} and δ_{C}, and conversely.*

Proof. This directly follows from the fact that

$$\phi_t^\alpha > \phi_t^\beta \Leftrightarrow C\phi_t^\alpha + B\phi_t^\beta > C\phi_t^\beta + A\phi_t^\alpha$$

. □

Note also that, for a given value of ϕ_1^α, $K(\phi^\beta,\phi^\alpha) \to 0$ if $\phi_t^\beta \to 0$, and conversely for $K(\phi^\alpha,\phi^\beta)$.

2 Proof of Lemma 1

Proof. Let $g_{\widehat{\boldsymbol{\vartheta}}} := \nabla_{\widehat{\boldsymbol{\vartheta}}} p(\mathbf{z} \,|\, \widehat{\boldsymbol{\vartheta}})$ be the score function of the prior distribution and $\widehat{w} := \mathbb{E}_{q_t(w)}[w]$. We want to solve the following equality wrt δ_{L}:

$$\nabla_{\boldsymbol{\phi}_t} \left\{ \mathbb{E}_{q_t(\mathbf{z})} \left[g_{\widehat{\boldsymbol{\vartheta}}} \right] \right\} \approx \nabla_{\boldsymbol{\phi}_t} \widehat{w}\, \delta_{\mathrm{L}}$$

Which has the following solution:

$$\begin{aligned}
\nabla_{\boldsymbol{\phi}_t} \widehat{w}\, \delta_{\mathrm{L}} &= \nabla_{\boldsymbol{\phi}_t} \{\widehat{w}\} \frac{d}{d\widehat{w}} \mathbb{E}_{q_t(\mathbf{z})} \left[g_{\widehat{\boldsymbol{\vartheta}}} \right] \\
&= \nabla_{\boldsymbol{\phi}_t} \{\widehat{w}\} \times \\
&\mathbb{E}_{q_t(\mathbf{z})} \left[\mathbf{z}^T (\boldsymbol{\theta}^{\xi}_{t-1} - \boldsymbol{\theta}^{\xi}_{0}) - A(\mathbf{z}) \left(\boldsymbol{\theta}^{\eta}_{t-1} - \boldsymbol{\theta}^{\eta}_{0} \right) - \frac{d\widehat{\boldsymbol{\vartheta}}}{d\widehat{w}} \nabla_{\widehat{\boldsymbol{\vartheta}}} B \left(\widehat{\boldsymbol{\vartheta}} \right) \right] \\
&= \nabla_{\boldsymbol{\phi}_t} \left\{ \widehat{w} \right\} \times \\
&\Big(\mathbb{E}_{q_t(\mathbf{z})}[\mathbf{z}]^T \left(\boldsymbol{\theta}^{\xi}_{t-1} - \boldsymbol{\theta}^{\xi}_{0} \right) - \mathbb{E}_{q_t(\mathbf{z})}[A(\mathbf{z})] \left(\boldsymbol{\theta}^{\eta}_{t-1} - \boldsymbol{\theta}^{\eta}_{0} \right) + \\
&- \frac{d\widehat{\boldsymbol{\vartheta}}}{d\widehat{w}} \nabla_{\widehat{\boldsymbol{\vartheta}}} B \left(\widehat{\boldsymbol{\vartheta}} \right) \Big)
\end{aligned} \tag{1}$$

As $p(\mathbf{z})$ is assumed to be from the exponential family, the derivative of the log-partition function $B(\cdot)$ wrt the natural parameter $\boldsymbol{\theta}(w)$ is equal to the expected value of the sufficient statistics:

$$\frac{d\widehat{\boldsymbol{\vartheta}}}{d\widehat{w}} \nabla_{\widehat{\boldsymbol{\vartheta}}} B \left(\widehat{\boldsymbol{\vartheta}} \right) = \begin{bmatrix} \boldsymbol{\theta}^{\xi}_{t-1} - \boldsymbol{\theta}^{\xi}_{0} \\ \boldsymbol{\theta}^{\eta}_{t-1} - \boldsymbol{\theta}^{\eta}_{0} \end{bmatrix}^T \begin{bmatrix} \mathbb{E}_{p(\mathbf{z})}[\mathbf{z}] \\ \mathbb{E}_{p(\mathbf{z})}[-A(\mathbf{z})] \end{bmatrix}$$

which can be plugged into Equation (1) to retrieve the expression of Lemma 1. □

We see that the value of δ_{L} depends on two elements: the first being whether the sign of $\mathbb{E}_{q(\mathbf{z})}[\mathbf{z}] - \mathbb{E}_{p(\mathbf{z})}[\mathbf{z}]$ matches the sign of $\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}_0$, and if the difference of the expected log-partitions under the posterior q and prior p is negative (because $\boldsymbol{\theta}^{\eta}_{t-1}$ is always greater than $\boldsymbol{\theta}^{\eta}_{0}$). The first summand can therefore be understood as a measure of how much the new observations $T(x_t)$, which conditions q_t matches the observed difference between the previous posterior q_{t-1} and the initial prior p_0: if it does (i.e. both differences are negative / positive) then there is evidence that w should increase (because δ_{L} will grow, see below). The second summand somehow measures whether the new posterior q_t will on average decrease the entropy of the model $p(x \,|\, \mathbf{z})$, which is linearly determined by $A(\mathbf{z})$.

3 AdaFVF implementation

Following [Kingma and Ba(2015)], we propose a scheme similar to the Adam optimizer where the mean gradient and preconditioner are optimized according to the HAFVF. We will consider the problem of inferring the vectors m and v^2, i.e. the vectors of first and second moments of the gradient. We will use a slightly different notation than **??**: the decays β_1 and β_2 will be replaced by w_1 and w_2 respectively, for the sake of coherence.

Let us consider that the partial derivative at the iteration t follows a normal distribution with mean and covariance m_t, s_t. The conjugate prior of this distribution is a Normal Inverse Gamma distribution with parameters $\boldsymbol{\theta} := \{\mu_0, \kappa_0, \alpha_0, \beta_0\}$. One could already apply the HAFVF to this model, with the restriction that $w_1 = w_2$. To keep the constraint $w_1 < w_2$, we assume a fully factorized posterior, and factorize the joint probability defined in the paper (equation 3) to:

$$\begin{aligned}
&p(d_t, m_t, s_t^2, w_1, w_2, b | \, \mathbf{x}_{1:t-1}) \approx p(d_t \,|\, m_t, s_t^2) \times \\
&\frac{(q_{t-1}(m_t \,|\, m_{t-1}, s^2/\kappa_{t-1})^{w_1} p(s_t^2 \,|\, \alpha_{t-1}, \beta_{t-1}))^{w_2}}{Z(w, \boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_0)} \times \\
&\frac{\Big(p(m_t \,|\, m_0, s^2/\kappa_0)^{\frac{1-w_1 w_2}{1-w_2}} p(s_t^2 \,|\, \alpha_0, \beta_0)\Big)^{1-w_2}}{Z(w, \boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_0)} \times \\
&\frac{q(w_1 \,|\, \boldsymbol{\phi}_{1t-1})^b p(w_1 \,|\, \boldsymbol{\phi}_{10})^{b-1}}{Z(b, \boldsymbol{\phi}_{1t-1}, \boldsymbol{\phi}_{10})} \frac{q(w_2 \,|\, \boldsymbol{\phi}_{2t-1})^b p(w_2 \,|\, \boldsymbol{\phi}_{20})^{b-1}}{Z(b, \boldsymbol{\phi}_{2t-1}, \boldsymbol{\phi}_{20})} \times \\
&p(b \,|\, \boldsymbol{\beta}_{t-1})
\end{aligned} \tag{2}$$

where $\boldsymbol{\theta}$ is defined as the prior or approximate posterior parameters at the trial 0 or $t-1$, respectively. This new formulation ensured that the decay of m was equal to $w_1 * w_2$.

For a set of N partial derivatives, the natural implementation of the joint probability presented here before for multiple, univariate gaussians would be

$$p(\mathbf{d}, \mathbf{m}, \mathbf{s}, w, b) = \prod_{i=1}^{N} p(d^i \,|\, m^i, s^i) p(m^i, s^i \,|\, \boldsymbol{\vartheta}^i) p(w_1, w_2 \,|\, \boldsymbol{\varphi}) p(b \,|\, \boldsymbol{\beta})$$

but this model is hard to fit in practice, because the posterior over w is highly sensitive to the dimensionality of the data at the level below (see Proposition 1). In order to deal with this, we modified the above equation by taking the N$^{\text{th}}$ radical of the joint probability $p(d^i, m^i, s^i)$. The normalized log-joint probability then reads:

$$\log \widetilde{p}(\mathbf{d}, \mathbf{m}, \mathbf{s}, w, b) = \frac{\sum_{i=1}^{N} \log p(d^i \,|\, m^i, s^i) + \log p(m^i, s^i \,|\, \boldsymbol{\vartheta}^i)}{N} + \log p(w_1, w_2 \,|\, \boldsymbol{\varphi}) + \log p(b \,|\, \boldsymbol{\beta}).$$

An important consideration to make is that we fitted the HAFVF to each of the sets of weights and biases independently: this ensured that the decays were adapted to the scale of each of these gradients independently. For instance, the extreme layers of a neural network usually have a wider distribution of partial derivatives than the intermediate layers: the HAFVF can account for this and adapt accordingly. Also, unlike other approaches, our

Algorithm 1 AdaFVF algorithm. $\mathcal{L}^j(q(m, s^2, w_t, b_t))$ stands for the ELBO value, $\widehat{\mathbf{d}}$ is the vector of expected first moment of the partial derivatives, and $\widehat{\mathbf{d}^2}$ the vector of expected second moments. η is the step size.

```
**Input:** noisy function $f(\mathbf{z})$ with parameters $\mathbf{z} = \{z^j \in \mathbb{R}^{D^j}\}$ for $j = 1 : J$, hyperparameters $\{\boldsymbol{\theta}_0, \boldsymbol{\phi}_0, \boldsymbol{\beta}_0\}$, learning rate $\eta$
Initialize randomly $\mathbf{z}_1$.
for $t = 1$ to $T$ do
  get $\mathbf{d}_t = \nabla_{\mathbf{z}_t} f(\mathbf{z}_t)$
  for $j = 1$ to $J$ do
    set $i := 0$
    set $\mathcal{L}^j(q(m_t, s_t, w_t, b_t)) := -\infty$
    while $i < 100$ and $\delta\mathcal{L}^j(q(m_t, s_t, w_t, b_t)) > 0.001$ do
      $i$ += 1
      update:
      $\boldsymbol{\theta}^j_t$ s.t. $q(m_t, s^2_t | \boldsymbol{\theta}^j_t) = \frac{\exp \mathbb{E}_{q(w_{1t}, w_{2t}, b_t))}\left[\log \widetilde{p}(d, m, s^2, w_1, w_2, b\right]}{Z}$    {see CVMP}
      $\{\boldsymbol{\phi}^j_t, \boldsymbol{\beta}^j_t\} = \arg\max_{\boldsymbol{\phi}^j_t, \boldsymbol{\beta}^j_t} \mathcal{L}^j(q(m, s^2, w_t, b_t))$    {see NCVMP}
      compute $\mathcal{L}^j(q(m, s^2, w_t, b_t))$
    end while
    update ${z^j}_{t+1} = {z^j}_t - \eta * \widehat{\mathbf{d}}/\widehat{\mathbf{d}^2}$
  end for
end for
```

method does not deal with degenerate samples by ignoring the step, but by decreasing their relative importance: in this way, the algorithm can discriminate which layer should be ignored (or better, reset) and which should not.

The full algorithm is given in Section 3.

Considering the fact that the function f is supposedly computationally expensive, the computational cost of this approach comparatively not much higher than the one of other SGD algorithms: update of the variational posterior over m, s is virtually identical to the update achieved by Adam, and most of the cost comes from the (possibly grouped) computation of the expected log-joint probability and its gradient. Expensive function in the ELBO include mostly the logarithm of the rate parameter of the prior $\log \beta(w)$ and approximate posterior $\log \beta$ which appear in the Gamma log-likelihood and in the expectation of the log-variance, respectively[1]. This evaluation is the step with the highest computational burden.

A final point to emphasize it that AdaFVF requires a careful choice of prior distribution over m, s^2, w_1, w_2 and b. As stated in the main text, we used high and confident prior over w_1 and w_2 to ensure that these parameters did not increase too much (so that they kept following the current distribution of gradients) and also to ensure that, after an degenerate gradient was observed, these parameters quickly came back to a value close the prior initially chosen. For m, s^2 the value of $\boldsymbol{\theta}$ was set to $\boldsymbol{\theta}_0 = \{\mu_0 = 0, \kappa_0 = 0, \alpha_0 = 2, \beta_0 = 10^{-5}\}$. The gradient

[1] the full expression of the latter is $\mathbb{E}_q\left[\log s^2\right] = \log \beta - \psi(\alpha)$, from which only the log function need to be computed for every sample, as the effective number of observations α is supposedly identical for all the elements of z^j

over b was set to a highly informative value ($\boldsymbol{\beta} = 20$): this had the effect of allowing the posterior over w to change slowly, and helped stabilizing the algorithm. This configuration worked well across the few models we tested, regardless of the sample size or the complexity of the problem. Future theoretical and practical research should, however, be dedicated to explore in which way these choices impact the performance of the algorithm.

Although the proof of convergence of [Kingma and Welling(2013)] does not hold anymore with our formulation, we observed that this algorithm was less sensitive to noise in the gradient than Adam, and performed at least equally well for problems ranging from simple auto-encoding variational inference to complex neural network training.

References

[Kingma and Ba(2015)] Diederik P. Kingma and Jimmy Lei Ba. Adam: a Method for Stochastic Optimization. *International Conference on Learning Representations 2015*, pages 1–15, 2015.

[Kingma and Welling(2013)] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. dec 2013. URL http://arxiv.org/abs/1312.6114http://www.aanda.org/10.1051/0004-6361/201527329.

[Knowles and Minka(2011)] David Knowles and Thomas P. Minka. Non-conjugate variational message passing for multinomial and binary regression. *Nips*, pages 1–9, 2011. URL http://eprints.pascal-network.org/archive/00008459/.