Affine Independent Variational Inference Supplementary Material

Edward Challis **David Barber**
Department of Computer Science
University College London, UK
{edward.challis,david.barber}@cs.ucl.ac.uk

1 Continuous partial derivatives

Potential functions $g(x) : \mathbb{R} \to \mathbb{R}$ that are piecewise smooth with a finite number of discontinuities have expectation $\left\langle g(\mathbf{w}^\mathsf{T}\mathbf{x})\right\rangle_{q_\mathbf{w}(\mathbf{w}|\mathbf{A},\mathbf{b},\boldsymbol{\theta})}$ that is smooth in $\mathbf{A}, \mathbf{b}$ provided that $q_\mathbf{v}(\mathbf{v})$ is smooth in $\mathbf{v}$. In this context we say that a function is smooth if it has continuous second order partial derivatives. Specifically, we require that $g(x)$ is piecewise smooth and so can be expressed as a sum of functions

$$g(\mathbf{w}^\mathsf{T}\mathbf{x}) = \sum_{j=1}^{J} g_j(\mathbf{w}^\mathsf{T}\mathbf{x})\mathbb{I}\left[\mathbf{w}^\mathsf{T}\mathbf{x} \in \mathcal{C}_j\right] \tag{1.1}$$

where $\mathbb{I}[x]$ is an indicator function equal to 1 when x is true and zero otherwise, $\{\mathcal{C}_j\}_{j=1}^{J}$ form a disjoint partition of $\mathbb{R}$ and for which $g_j(x)$ are smooth in x. For site projection potentials of this form the derivative *w.r.t.* A_{mn} of the expectation can be expressed as

$$\frac{\partial}{\partial A_{mn}}\left\langle \log g(\mathbf{w}^\mathsf{T}\mathbf{x})\right\rangle = \sum_{j=1}^{J}\frac{\partial}{\partial A_{mn}}\int_{\mathbf{w}^\mathsf{T}\mathbf{x}\in\mathcal{C}_j} q_\mathbf{w}(\mathbf{w}|\mathbf{A},\mathbf{b},\boldsymbol{\theta})g_j(\mathbf{w}^\mathsf{T}\mathbf{x})d\mathbf{w} \tag{1.2}$$

$$= \sum_{j=1}^{J}\int_{\mathbf{w}^\mathsf{T}\mathbf{x}\in\mathcal{C}_j} g_j(\mathbf{w}^\mathsf{T}\mathbf{x})\frac{\partial}{\partial A_{mn}}q_\mathbf{w}(\mathbf{w}|\mathbf{A},\mathbf{b},\boldsymbol{\theta})d\mathbf{w}. \tag{1.3}$$

Exchanging the order of the derivative and integral operators is possible since $g_j(\mathbf{w}^\mathsf{T}\mathbf{x})$ is smooth in $\mathbf{w}$ for $\mathbf{w}^\mathsf{T}\mathbf{x} \in \mathcal{C}_j$ by assumption and $q_\mathbf{w}(\mathbf{w}|\mathbf{A},\mathbf{b},\boldsymbol{\theta})$ is smooth in $\mathbf{w}, \mathbf{A}, \mathbf{b}$. To see that $q_\mathbf{w}(\mathbf{w}|\mathbf{A},\mathbf{b},\boldsymbol{\theta})$ is smooth in $\mathbf{A}, \mathbf{b}$ we note that it is the composition of $q_\mathbf{v}(\mathbf{v}|\boldsymbol{\theta})$ which is smooth by assumption and $\mathbf{v} = \mathbf{A}^{-1}(\mathbf{w} - \mathbf{b})$ which is also smooth. Note that the limits of each integral in equation (1.2) do not depend on A_{mn} and so do not contribute additional terms to equation (1.3). Second order derivatives carry through similarly by again passing the derivative under the integral sign in equation (1.3).

2 AI KL bound and gradient evaluation

We describe how to efficiently numerically evaluate the AI KL bound and associated gradients (with respect to the parameters $\mathbf{A} = \mathbf{LU}, \mathbf{b}$ and $\boldsymbol{\theta}$).

2.1 Entropy

The entropy's contribution to the AI KL bound can be written

$$H\left[q_\mathbf{w}(\mathbf{w})\right] = \log\left|\det\left(\mathbf{A}\right)\right| + \sum_{d=1}^{D} H\left[q_{v_d}(v_d|\theta_d)\right], \tag{2.1}$$

where $H\left[q(v_d|\theta_d)\right]$ is the univariate differential entropy of the base density $q_{v_d}(v_d|\theta_d)$. The derivative of equation (2.1) *w.r.t.* $\mathbf{A}$ is then

$$\frac{\partial}{\partial \mathbf{A}} H\left[q(\mathbf{v})\right] = \mathbf{A}^{-\mathsf{T}}. \tag{2.2}$$

The derivatives of the marginal base density's entropy, $H\left[q(v_d|\theta_d)\right]$, depend on the parametric form of the chosen base density $q_{v_d}(v_d|\theta_d)$ and the parameter θ_d only. For the results presented only two base densities were used: the skew-normal and the generalised-normal. The entropy and respective derivatives of these distributions are presented in section(4).

For the LU parameterised bound, such that $\mathbf{A} = \mathbf{LU}$ for $\mathbf{L}$ lower triangular and $\mathbf{U}$ upper triangular we have

$$\log\left|\det\left(\mathbf{A}\right)\right| = \sum_{d=1}^{D} \log L_{dd} + \log U_{dd}. \tag{2.3}$$

Thus the partial derivatives of the entropy with respect to $\mathbf{L}$ and $\mathbf{U}$ are given by

$$\frac{\partial}{\partial L_{mn}} H\left[q_{\mathbf{w}}(\mathbf{w})\right] = \delta_{mn}\frac{1}{L_{mn}},$$
$$\frac{\partial}{\partial U_{mn}} H\left[q_{\mathbf{w}}(\mathbf{w})\right] = \delta_{mn}\frac{1}{U_{mn}},$$

where δ_{mn} is the Kronecker delta.

2.2 Site projection potentials

In the main text we showed that the expectation of $g\left(\mathbf{x}^{\mathsf{T}}\mathbf{w}\right)$ with respect to $q_{\mathbf{w}}(\mathbf{w})$ can be efficiently computed by using the FFT. In this section first we review this result making clear each step of the derivation. Second, we show how the derivatives of $\left\langle g\left(\mathbf{x}^{\mathsf{T}}\mathbf{w}\right)\right\rangle$ with respect to $\mathbf{A}, \mathbf{b}, \boldsymbol{\theta}$ can also be efficiently computed.

Computing the expectation. The expectation $\left\langle g(\mathbf{x}^{\mathsf{T}}\mathbf{w})\right\rangle_{q_{\mathbf{w}}(\mathbf{w})}$ for $g : \mathbb{R} \to \mathbb{R}$ some non-linear function, $\mathbf{x} \in \mathbb{R}^D$ some fixed vector and $q_{\mathbf{w}}(\mathbf{w})$ an AI density is equivalent to the univariate expectation $\left\langle g(y)\right\rangle_{q_y(y)}$ where the density $q_y(y)$ can be obtained by using the Fourier transform. To show this first we write the expectation of $g(\mathbf{x}^{\mathsf{T}}\mathbf{w})$ *w.r.t.* $q_{\mathbf{w}}(\mathbf{w})$ as an expectation with respect to $q_{\mathbf{v}}(\mathbf{v})$,

$$\left\langle g\left(\mathbf{x}^{\mathsf{T}}\mathbf{w}\right)\right\rangle_{q_{\mathbf{w}}(\mathbf{w})} = \int g\left(\mathbf{x}^{\mathsf{T}}\mathbf{w}\right) q_{\mathbf{w}}(\mathbf{w}) d\mathbf{w} = \int g\left(\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{v} + \mathbf{x}^{\mathsf{T}}\mathbf{b}\right) q_{\mathbf{v}}(\mathbf{v}) d\mathbf{v}. \tag{2.4}$$

The last equality in equation (2.4) is obtained by making the substitution $\mathbf{w} = \mathbf{A}\mathbf{v} + \mathbf{b}$. For a cleaner notation, in what follows we let $\boldsymbol{\alpha} = \mathbf{A}^{\mathsf{T}}\mathbf{x}$ and $\beta = \mathbf{x}^{\mathsf{T}}\mathbf{b}$. We now substitute $g(\boldsymbol{\alpha}^{\mathsf{T}}\mathbf{v} + \beta) = \int \delta(y - \boldsymbol{\alpha}^{\mathsf{T}}\mathbf{v} - \beta) g(y) dy$, where $\delta(x)$ is the Dirac delta function, into equation (2.4) to give us

$$\left\langle g\left(\mathbf{w}^{\mathsf{T}}\mathbf{x}\right)\right\rangle_{q_{\mathbf{w}}(\mathbf{w})} = \int g(y) \int \prod_d q_{v_d}(v_d|\theta_d) \delta\left(y - \boldsymbol{\alpha}^{\mathsf{T}}\mathbf{v} - \beta\right) d\mathbf{v} dy = \left\langle g(y)\right\rangle_{q_y(y)}. \tag{2.5}$$

In equation (2.5) above $q_y(y)$ is the density of the random variable y defined as the linear projection of the random variables $\mathbf{v}$ such that $y = \boldsymbol{\alpha}^{\mathsf{T}}\mathbf{v} + \beta$. Thus the univariate marginal density $q_y(y)$ is defined by the integral

$$q_y(y) = \int \delta\left(y - \boldsymbol{\alpha}^{\mathsf{T}}\mathbf{v} - \beta\right) \prod_d q_{v_d}(v_d|\theta_d) d\mathbf{v}.$$

Whilst this integral is generally intractable we can make the substitution $\delta(x) = \int e^{2\pi i t x} dt$ to give us

$$q_y(y) = \int\int e^{2\pi i t\left(y - \boldsymbol{\alpha}^{\mathsf{T}}\mathbf{v} - \beta\right)} \prod_d q_{v_d}(v_d|\theta_d) d\mathbf{v} dt \tag{2.6}$$

$$= \int e^{2\pi i t y} e^{-2\pi i t \beta} \prod_d \int e^{-2\pi i t \alpha_d v_d} q_{v_d}(v_d|\theta_d) dv_d dt. \tag{2.7}$$

We now inspect each term in equation (2.7). First we consider an individual factor of the group product:

$$\int q_{v_d}(v_d)e^{-2\pi i t\alpha_d v_d}dv_d = \frac{1}{|\alpha_d|}\int q_{v_d}\left(\frac{u_d}{\alpha_d}\right)e^{-2\pi i t u_d}du_d \tag{2.8}$$

$$= \int q_{u_d}(u_d|\theta_d)\, e^{-2\pi i t u_d}du_d = \tilde{q}_{u_d}(t), \tag{2.9}$$

where the first equality comes from making the substitution $u_d = \alpha_d v_d$. This substitution defines the univariate density $q_{u_d}(u_d|\theta_d) = \frac{1}{|\alpha_d|}q_{v_d}(\frac{u_d}{\alpha_d}|\theta_d)$. Thus each factor of the group product in equation (2.7) is the Fourier transform of the density $q_{u_d}(u_d|\theta_d)$. The $e^{-2\pi i t\beta}$ factor corresponds to the Fourier transform of a delta mean shift $\delta(y-\beta)$. Putting this together equation (2.7) can be interpreted as the inverse Fourier transform of the product of the Fourier transforms of $\{q_{u_d}(u_d|\theta_d)\}_{d=1}^D$ and of the β mean shift. Algebraically this gives us an expression for the marginal $q_y(y)$ in the form

$$q_y(y) = \int e^{2\pi i t y}e^{-2\pi i t\beta}\prod_d \tilde{q}_{u_d}(t)dt. \tag{2.10}$$

This result is a reworking of the D-fold convolution theorem for probability densities. We provide the derivation here so that it may form the basis of subsequent derivations required to evaluate the AI KL bound's derivatives as univariate integrals.

Numerical evaluation. Since only in very special cases we have simple analytic forms for the univariate density $q_y(y)$ we resort to numerical methods to evaluate it. To do so we evaluate equation (2.10) replacing $\{q_{u_d}(u_d|\theta_d)\}_{d=1}^D$ with their discrete lattice approximations $\{\hat{q}_{u_d}(u_d|\theta_d)\}_{d=1}^D$. We now show that making this substitution results in $\hat{q}_y(y)$ as defined in equation(2.6) which can be efficiently computed by utilising the FFT algorithm.

First, we must define the set of lattice points used to evaluate the discrete approximate densities $\{\hat{q}_{u_d}(u_d|\theta_d)\}_{d=1}^D$. The user defines the number of lattice points $K \in \mathbb{N}$ according to their computational budget or accuracy requirements. The accuracy can be roughly assessed by computing the difference in the expectation using K and $2K$ lattice points. The lattice end points are chosen such that $[l_1, l_k] = [-\nu\sigma_y, \nu\sigma_y]$ where σ_y is the standard deviation of the random variable y given by $\sigma_y^2 = \sum_d \alpha_d^2 \mathrm{var}(v_d)$. ν is a user defined parameter, in our experiments we set $\nu = 6$ and double K until the bound value changes by less than 10^{-3}. The lattice points $[l_1, ..., l_K]$ are evenly spaced such that $\Delta = l_{k+1} - l_k$ is constant for all k.

The continuous Fourier transform of the lattice density $\hat{q}_{u_d}(u_d|\theta_d)$ takes the form

$$\tilde{\hat{q}}_{u_d}(t) := \int e^{-2\pi i t u_d}\hat{q}_{u_d}(u_d|\theta_d)du_d = \sum_{k=1}^K \pi_{dk}e^{-2\pi i t l_k}. \tag{2.11}$$

Taking the inverse Fourier transform of the product of these transforms, as $q(y)$ is defined in equation (2.10), we get

$$\hat{q}_y(y) = \int e^{2\pi i t(y-\beta)}\prod_d \sum_{k_d=1}^K \pi_{dk_d}e^{-2\pi i t l_{k_d}}dt \tag{2.12}$$

$$= \sum_{[k_1,...,k_D]}\int e^{2\pi i t\left(y-\beta-\sum_d l_{k_d}\right)}\prod_d \pi_{dk_d}dt \tag{2.13}$$

$$= \sum_{[k_1,...,k_D]}\delta\left(y-\beta-\sum_d l_{k_d}\right)\prod_d \pi_{dk_d}, \tag{2.14}$$

where the sum in equation (2.14) refers to the sum over the K^D permutations of the D dimensional cartesian product of lattice point indices $[k_1, ..., k_D]$. We note that $l_{k_d} = l_k$, the subscript is only to distinguish the different permutations of the sum.

Equation(2.14) describes a mixture of delta distributions and is the exact result from computing the convolution of the lattice approximate densities by means of the continuous Fourier transform.

Importantly, the K^D mixtures in equation (2.14) collapse to just DK distinct delta points since l_k are evenly spaced.

When $D = 2$:

$$\hat{q}_y(y) = \sum_{j=1}^{K} \sum_{k=1}^{K} \pi_{1j}\pi_{2k}\delta(y - \beta - l_j - l_k). \tag{2.15}$$

We can see from equation (2.15) above that $\hat{q}_y(y)$ is a mixture of $2K$ delta densities evenly spaced at lattice points $[2l_1, ..., 2l_K]$,

$$\hat{q}_y(y) = \sum_{n=1}^{2K} \rho_n \delta(y - \beta - l_n) \tag{2.16}$$

for suitably defined $\boldsymbol{\rho}$. For a single lattice point l_m,

$$\rho_m = \sum_{i,j:i+j=m} \pi_{1j}\pi_{2k} = \sum_{n=1}^{2K} \pi'_{1n}\pi'_{2(m-n)} = \left[\texttt{ifft}\left[\texttt{fft}[\boldsymbol{\pi}'_1] \cdot \texttt{fft}[\boldsymbol{\pi}'_2]\right]\right]_m, \tag{2.17}$$

Here $\boldsymbol{\pi}'$ refers to the zero padded vector of delta mixture weights $\boldsymbol{\pi}' = [\boldsymbol{\pi}, \mathbf{0}]$ such that $\mathbf{0}$ is a K dimensional vector of zeros. If $m-n < 1$ we extend the indices $\boldsymbol{\pi}'_{m-n} := \boldsymbol{\pi}'_{2K+m-n}$; this extension is valid and does not affect the convolution due to the zero padding of $\boldsymbol{\pi}'$. The last equality in the expression above is the statement of the discrete Fourier transform convolution theorem.

The result can be extended to higher dimensions $D > 2$ by induction using the associativity of the convolution operator and the fact that lattice point locations are invariant to convolution to give

$$\hat{q}_y(y) = \sum_{n=1}^{DK} \rho_n \delta(y - \beta - l_n) \quad \text{where} \quad \boldsymbol{\rho} = \texttt{ifft}\left[\prod_d \texttt{fft}\left[\boldsymbol{\pi}'_d\right]\right] \tag{2.18}$$

For general D, $\boldsymbol{\pi}'$ refers to the zero padded vector of delta mixture weights $\boldsymbol{\pi}' = [\boldsymbol{\pi}, \mathbf{0}]$ such that $\mathbf{0}$ is a $(D-1)K$ dimensional vector of zeros.

Computing the derivative *w.r.t.* A Taking the derivative of $\left\langle g\left(\mathbf{w}^\mathsf{T}\mathbf{x}\right)\right\rangle$ with respect to A_{mn} we obtain

$$\frac{\partial}{\partial A_{mn}}\left\langle g(\mathbf{w}^\mathsf{T}\mathbf{x})\right\rangle = x_n \int q_{\mathbf{v}}(\mathbf{v}) g'\left(\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{v} + \mathbf{b}^\mathsf{T}\mathbf{x}\right) v_m d\mathbf{v}. \tag{2.19}$$

As previously mentioned the above form is not equivalent to $x_n \left\langle v_m g'(y)\right\rangle_{q_y(y)}$. It can however be expressed as a one dimensional integral:

$$\begin{aligned}
\frac{\partial}{\partial A_{mn}}\left\langle g\left(\mathbf{w}^\mathsf{T}\mathbf{x}\right)\right\rangle &= x_n \int v_m \prod_{d=1}^{D} q_{v_d}(v_d|\theta_d) g'\left(\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{v} + \mathbf{b}^\mathsf{T}\mathbf{x}\right) d\mathbf{v} \\
&= x_n \int v_m \prod_{d=1}^{D} q_{v_d}(v_d|\theta_d) \int \delta\left(y - \boldsymbol{\alpha}^\mathsf{T}\mathbf{v} - \beta\right) g'(y) dy d\mathbf{v} \\
&= x_n \int v_m q_{v_m}(v_m|\theta_m) \prod_{d\neq m} q_{v_d}(v_d|\theta_d) \int \delta\left(y - \boldsymbol{\alpha}^\mathsf{T}\mathbf{v} - \beta\right) g'(y) dy d\mathbf{v} \\
&= x_n \int g'(y) \int v_m q_{v_m}(v_m|\theta_m) \prod_{d\neq m} q_{v_d}(v_d|\theta_d) \delta\left(y - \boldsymbol{\alpha}^\mathsf{T}\mathbf{v} - \beta\right) d\mathbf{v} dy
\end{aligned}$$

where $g'(y) = \frac{d}{dy}g(y)$, $\boldsymbol{\alpha} = \mathbf{A}^\mathsf{T}\mathbf{x}$ and $\beta = \mathbf{b}^\mathsf{T}\mathbf{x}$. To evaluate the expression above we define the univariate weighting function $d_m(y)$,

$$d_m(y) := \int v_m q_{v_m}(v_m|\theta_m) \prod_{d\neq m} q_{v_d}(v_d|\theta_d) \delta\left(y - \boldsymbol{\alpha}^\mathsf{T}\mathbf{v} - \beta\right) d\mathbf{v}. \tag{2.20}$$

Using this weighting function the gradient can simply be expressed as

$$\frac{\partial}{\partial A_{mn}} \left\langle g\left(\mathbf{w}^\mathsf{T}\mathbf{x}\right)\right\rangle = x_n \int g'(y) d_m(y) dy.$$

We evaluate $d_m(y)$ by means of computing its Fourier transform. The Fourier transform of $d_m(y)$ is given by

$$\begin{aligned}
\tilde{d}_m(t) &= \int e^{-2\pi i t y} \int v_m q_{v_m}(v_m|\theta_m) \prod_{d\neq m} q_{v_d}(v_d|\theta_d) \delta\left(y - \boldsymbol{\alpha}^\mathsf{T}\mathbf{v} - \beta\right) d\mathbf{v} dy \\
&= e^{-2\pi i t \beta} \int v_m q_{v_m}(v_m|\theta_m) e^{-2\pi i t \alpha_m v_m} \prod_{d\neq m} q_{v_d}(v_d|\theta_d) e^{-2\pi i t \alpha_d v_d} d\mathbf{v} \\
&= e^{-2\pi i t \beta} \times \tilde{e}_m(t) \times \prod_{d\neq m} \tilde{q}_{u_d}(t|\theta_d)
\end{aligned}$$

where $\tilde{e}_m(t|\theta_m)$ is the Fourier transform of the univariate expectation

$$\tilde{e}_m(t) := \int v_m q_{v_m}(v_m|\theta_m) e^{-2\pi i t \alpha_m v_m} dv_m = \int \frac{u_m}{\alpha_m} q_{u_m}\left(u_m|\theta_m\right) e^{-2\pi i t u_m} du_m.$$

Computing the derivative *w.r.t.* $\mathbf{b}$ Taking the derivative of $\left\langle g\left(\mathbf{w}^\mathsf{T}\mathbf{x}\right)\right\rangle$ with respect to b_m we get

$$\begin{aligned}
\frac{\partial}{\partial b_m} \left\langle g\left(\mathbf{w}^\mathsf{T}\mathbf{x}\right)\right\rangle &= x_m \int \prod_{d=1}^{D} q_{v_d}(v_d|\theta_d) g'\left(\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{v} + \mathbf{b}^\mathsf{T}\mathbf{x}\right) d\mathbf{v} \\
&= x_m \int q_y(y) g'(y) dy.
\end{aligned}$$

Computing the derivative *w.r.t.* $\boldsymbol{\theta}$ Taking the derivative of $\left\langle g\left(\mathbf{w}^\mathsf{T}\mathbf{x}\right)\right\rangle$ with respect to θ_m we get

$$\begin{aligned}
\frac{\partial}{\partial \theta_m} \left\langle g\left(\mathbf{w}^\mathsf{T}\mathbf{x}\right)\right\rangle &= \frac{\partial}{\partial \theta_m} \int \prod_{d=1}^{D} q_{v_d}(v_d|\theta_d) g\left(\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{v} + \mathbf{b}^\mathsf{T}\mathbf{x}\right) d\mathbf{v} \\
&= \int \left[\frac{\partial}{\partial \theta_m} \prod_{d\neq m} q_{v_d}(v_d|\theta_d) q_{v_m}(v_m|\theta_m)\right] g\left(\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{v} + \mathbf{b}^\mathsf{T}\mathbf{x}\right) d\mathbf{v} \\
&= \int g(y) \int \left[\frac{\partial}{\partial \theta_m} q_{v_m}(v_m|\theta_m)\right] \prod_{d\neq m} q_{v_d}(v_d|\theta_d) \delta\left(y - \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{v} - \mathbf{b}^\mathsf{T}\mathbf{x}\right) dy d\mathbf{v}
\end{aligned}$$

Similar to the gradient of $\left\langle \log f_n(\mathbf{w}^\mathsf{T}\mathbf{x}_n)\right\rangle$ with respect to A_{mn} we define a derivative weighting function $\tilde{p}'_d$ such that

$$\begin{aligned}
\tilde{p}'_d(t) &:= \int e^{-2\pi i t y} \int \left[\frac{\partial}{\partial \theta_m} q_{v_m}(v_m|\theta_m) \prod_{d\neq m} q_{v_d}(v_d|\theta_d)\right] \delta\left(y - \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{v} - \mathbf{b}^\mathsf{T}\mathbf{x}\right) dy d\mathbf{v} \\
&= e^{-2\pi i t \beta} \left[\prod_{d\neq m} \tilde{q}_{u_d}(t|\theta_d)\right] \int e^{-2\pi i t \alpha_m v_m} \frac{\partial}{\partial \theta_m} p(v_m|\theta_m) dv_m.
\end{aligned}$$

For $p'_d(y)$ the inverse Fourier transform of $\tilde{p}'_d(t)$ we obtain the gradient

$$\frac{\partial}{\partial \theta_m} \left\langle g\left(\mathbf{w}^\mathsf{T}\mathbf{x}\right)\right\rangle = \int p'_d(y) g(y) dy.$$

2.3 Gaussian potentials

For the Gaussian potential $\mathcal{N}\left(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ we have its log expectation under $q_\mathbf{w}(\mathbf{w})$ given by

$$2\left\langle \log \mathcal{N}\left(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\right\rangle = -D \log 2\pi - \log\det\left(\boldsymbol{\Sigma}\right) - \left\langle \mathbf{w}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{w}\right\rangle + 2\left\langle \mathbf{w}\right\rangle \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}. \quad (2.21)$$

To evaluate this expression we precompute the Cholesky decomposition of Gaussian precision matrix $\boldsymbol{\Sigma}^{-1} = \mathbf{P}^\mathsf{T}\mathbf{P}$, which scales $O\left(D^3\right)$ and only needs to be performed once. Since $\langle \mathbf{w} \rangle = \mathbf{A}\langle \mathbf{v} \rangle + \mathbf{b}$ and $\left\langle \mathbf{v}^\mathsf{T}\mathbf{B}\mathbf{v} \right\rangle = \langle \mathbf{v} \rangle^\mathsf{T} \mathbf{B} \langle \mathbf{v} \rangle + \mathrm{trace}\left(\mathbf{B}\mathrm{cov}(\mathbf{v})\right)$ we have that

$$\begin{aligned}\left\langle \mathbf{w}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{w} \right\rangle &= \left\langle \mathbf{v}^\mathsf{T}\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{A}\mathbf{v} \right\rangle + 2\left\langle \mathbf{v}^\mathsf{T}\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{b} \right\rangle + \mathbf{b}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{b} \\ &= \langle \mathbf{v} \rangle^\mathsf{T} \mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{A}\langle \mathbf{v} \rangle + \mathrm{trace}\left(\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{A}\mathrm{cov}(\mathbf{v})\right) \\ &\qquad + 2\langle \mathbf{v} \rangle^\mathsf{T} \mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{b} + \mathbf{b}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{b} \\ \langle \mathbf{w} \rangle \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} &= \langle \mathbf{v} \rangle^\mathsf{T} \mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{b}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\end{aligned}$$

where $\mathrm{cov}(\mathbf{v}) = \mathrm{diag}\left(\mathrm{var}(\mathbf{v})\right) = \mathbf{D}$ since $\mathbf{v}$ are assumed independent. All terms in the expression above, except for the trace term, can be computed as a sequence of matrix vector products. To compute the trace term we use $\mathrm{trace}\left(\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{A}\mathrm{cov}(\mathbf{v})\right) = \mathrm{vec}\left(\mathbf{PLUD}^{\frac{1}{2}}\right)^\mathsf{T} \mathrm{vec}\left(\mathbf{PLUD}^{\frac{1}{2}}\right)$, where $\mathrm{vec}\left(\mathbf{X}\right)$ constructs a column vector by concatenating the columns of the matrix $\mathbf{X}$ and $\mathbf{D}^{\frac{1}{2}}$ is the square root of the diagonal covariance matrix, which scale $O\left(D^3\right)$ for general $\boldsymbol{\Sigma}$. When $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ this reduces to $O\left(D^2\right)$.

Derivative *w.r.t.* $\mathbf{A}$ The derivatives of the above form with respect to $\mathbf{A}$ and $\mathbf{b}$ are

$$\begin{aligned}\frac{\partial}{\partial \mathbf{A}} 2\left\langle \log \mathcal{N}\left(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \right\rangle = \frac{\partial}{\partial \mathbf{A}} &- \langle \mathbf{v} \rangle^\mathsf{T} \mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{A}\langle \mathbf{v} \rangle - \mathrm{trace}\left(\mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{A}\mathrm{cov}(\mathbf{v})\right) \\ &- 2\langle \mathbf{v} \rangle^\mathsf{T} \mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{b} + 2\langle \mathbf{v} \rangle^\mathsf{T} \mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + 2\mathbf{b}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu},\end{aligned}$$

which can be expressed

$$\frac{\partial}{\partial \mathbf{A}} \left\langle \log \mathcal{N}\left(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \right\rangle = -\boldsymbol{\Sigma}^{-1}\mathbf{A}\left(\langle \mathbf{v} \rangle \langle \mathbf{v} \rangle^\mathsf{T} + \mathrm{cov}(\mathbf{v})\right) + \langle \mathbf{v} \rangle \left(\boldsymbol{\mu}\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{b}\right)^\mathsf{T}, \quad (2.22)$$

and computed using sequential matrix vector multiplies and vector outer products.

Derivative *w.r.t.* $\mathbf{b}$ The derivative with respect to $\mathbf{b}$ is given by

$$\begin{aligned}\frac{\partial}{\partial \mathbf{b}} 2\left\langle \log \mathcal{N}\left(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \right\rangle &= \frac{\partial}{\partial \mathbf{b}} - 2\langle \mathbf{v} \rangle^\mathsf{T} \mathbf{A}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{b} - \mathbf{b}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{b} + 2\mathbf{b}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ &= -2\boldsymbol{\Sigma}^{-1}\left(\mathbf{A}\langle \mathbf{v} \rangle + \mathbf{b} + \boldsymbol{\mu}\right).\end{aligned}$$

2.4 Derivatives *w.r.t.* $\mathbf{L}, \mathbf{U}$

We extend the above results to the LU decomposition of the transformation matrix such that $\mathbf{A} = \mathbf{LU}$ where $\mathbf{L}$ is lower triangular and $\mathbf{U}$ upper triangular matrices. We apply the chain rule, noting that $A_{mn} = \sum_k L_{mk}U_{kn}$ to give

$$\frac{\partial A_{mn}}{\partial L_{uv}} = \delta_{mu}U_{vn} \quad \text{and} \quad \frac{\partial A_{mn}}{\partial U_{st}} = \delta_{tn}L_{ms}$$

for δ_{ab} the Kronecker delta. Thus to compute the derivative of $F(\mathbf{A}) = F(\mathbf{LU})$ we have that

$$\begin{aligned}\frac{\partial}{\partial L_{uv}} F(\mathbf{A}) &= \sum_{mn} \frac{\partial}{\partial A_{mn}} F(\mathbf{A})\delta_{mu}U_{vn} \quad \text{when} \quad u \geq v \quad \text{and zero otherwise} \\ \frac{\partial}{\partial U_{st}} F(\mathbf{A}) &= \sum_{mn} \frac{\partial}{\partial A_{mn}} F(\mathbf{A})\delta_{tn}L_{ms} \quad \text{when} \quad t \geq s \quad \text{and zero otherwise.}\end{aligned}$$

3 Blockwise concavity

Here we present a simple reworking, and extension, of the concavity result originally provided in [1] for log-concave potentials $\{f_n\}_{n=1}^N$. Whilst the bound is jointly concave in $\mathbf{L}$ and $\mathbf{b}$ or $\mathbf{U}$ and $\mathbf{b}$ it is not jointly concave in $\mathbf{L}$ and $\mathbf{U}$ simultaneously.

The entropy of the AI bound is clearly concave in both $\mathbf{L}$ and $\mathbf{U}$ being a sum of log terms acting on individual elements of $\mathbf{L}$ and $\mathbf{U}$.

The Gaussian potential contribution to the AI bound is a negative quadratic in $\mathbf{L}$ or $\mathbf{U}$. To see this we consider the Gaussian contribution, omitting constants *w.r.t.* $\mathbf{U}$, $\mathbf{L}$ and $\mathbf{b}$ we have that

$$2\left\langle \log \mathcal{N}\left(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Sigma}\right)\right\rangle \overset{c.}{=} -\bar{\mathbf{v}}^{\mathsf{T}}\mathbf{U}^{\mathsf{T}}\mathbf{L}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{L}\mathbf{U}\bar{\mathbf{v}} - \operatorname{trace}\left(\mathbf{U}^{\mathsf{T}}\mathbf{L}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{L}\mathbf{U}\mathbf{D}\right) - 2\bar{\mathbf{v}}^{\mathsf{T}}\mathbf{U}^{\mathsf{T}}\mathbf{L}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{b} - \mathbf{b}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{b} + 2\bar{\mathbf{v}}\mathbf{U}^{\mathsf{T}}\mathbf{L}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + 2\mathbf{b}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$$

where $\bar{\mathbf{v}} = \langle \mathbf{v} \rangle$ and $\mathbf{D} = \operatorname{diag}\left(\operatorname{var}(\mathbf{v})\right)$. Keeping $\mathbf{L}$ fixed and denoting $\mathbf{X} = \mathbf{L}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{L}$ we get

$$2\left\langle \log \mathcal{N}\left(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Sigma}\right)\right\rangle \overset{c.}{=} -\bar{\mathbf{v}}^{\mathsf{T}}\mathbf{U}^{\mathsf{T}}\mathbf{X}\mathbf{U}\bar{\mathbf{v}} - \operatorname{trace}\left(\mathbf{U}^{\mathsf{T}}\mathbf{X}\mathbf{U}\mathbf{D}\right) - 2\bar{\mathbf{v}}^{\mathsf{T}}\mathbf{U}^{\mathsf{T}}\mathbf{L}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{b} - \mathbf{b}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{b} + 2\bar{\mathbf{v}}\mathbf{U}^{\mathsf{T}}\mathbf{L}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + 2\mathbf{b}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \quad (3.1)$$

which is a negative quadratic in $\mathbf{U}$ and $\mathbf{b}$ and is thus jointly concave in these parameters. A similar analysis carries through for $\mathbf{L}$ keeping $\mathbf{U}$ fixed.

Without loss of generality we can consider the concavity of a single non-linear site potential's contribution to the AI bound. We have that

$$E(\mathbf{A},\mathbf{b}) := \langle \log f_n(\mathbf{w}) \rangle = \int q_{\mathbf{v}}(\mathbf{v}) g(\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{v} + \mathbf{b}^{\mathsf{T}}\mathbf{x}) d\mathbf{v} \quad (3.2)$$

where $g(x) = \log f(x)$ for nonlinear site function $f(x) : \mathbb{R} \to \mathbb{R}^+$, $f(x)$ are assumed log-concave and so $\forall \theta \in [0,1]$

$$g\left(\theta x + (1-\theta)y\right) \geq \theta g(x) + (1-\theta)g(y). \quad (3.3)$$

Thus considering two transformation matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ we have that

$$E\left(\theta\mathbf{A}_1 + (1-\theta)\mathbf{A}_2, \theta\mathbf{b}_1 + (1-\theta)\mathbf{b}_2\right) = \left\langle g\left(\theta\left(\mathbf{x}^{\mathsf{T}}\mathbf{A}_1\mathbf{v} + \mathbf{b}_1^{\mathsf{T}}\mathbf{x}\right) + (1-\theta)\left(\mathbf{x}^{\mathsf{T}}\mathbf{A}_2\mathbf{v} + \mathbf{b}_2^{\mathsf{T}}\mathbf{x}\right)\right)\right\rangle,$$

using the concavity of g and the linearity of the expectation operator we have

$$E(\theta\mathbf{A}_1 + (1-\theta)\mathbf{A}_2, \theta\mathbf{b}_1 + (1-\theta)\mathbf{b}_2) \geq \theta\left\langle g(\mathbf{x}^{\mathsf{T}}\mathbf{A}_1\mathbf{v} + \mathbf{b}_1^{\mathsf{T}}\mathbf{x})\right\rangle + (1-\theta)\left\langle g(\mathbf{x}^{\mathsf{T}}\mathbf{A}_2\mathbf{v} + \mathbf{b}_2^{\mathsf{T}}\mathbf{x})\right\rangle$$

and thus the non-linear site functions contribute terms that are concave in $\mathbf{A}$ to the AI KL bound. Concavity in $\mathbf{L}$ follows through by letting $\mathbf{x} = \mathbf{U}\mathbf{x}$, similarly the converse holds for concavity in $\mathbf{U}$ keeping $\mathbf{L}$ fixed.

4 Base distributions

We present the entropy and gradients required to perform AI KL variational inference with the skew-normal and generalised-normal distributions.

4.1 Skew-normal

The skew-normal distribution, $\mathcal{SN}(v|\mu,\sigma,\nu)$, is parameterised,

$$\mathcal{SN}(v|\mu,\sigma,\nu) = \frac{2}{\sigma}\phi\left(\frac{v-\mu}{\sigma}\right)\Phi\left(\nu\left(\frac{v-\mu}{\sigma}\right)\right) \quad (4.1)$$

where $\phi(z) = \mathcal{N}\left(z|0,1\right)$, $\Phi(z) = \int_{-\infty}^{z}\phi(x)dx$, location parameter $\mu \in \mathbb{R}$, scale parameter $\sigma \in \mathbb{R}^+$ and skew parameter $\nu \in \mathbb{R}$. When $\nu = 0$ we recover the Gaussian density $\mathcal{N}\left(v|\mu,\sigma^2\right)$. For AI KL approximate inference $\mathbf{A},\mathbf{b}$ parameterise the covariance and the location of $q_{\mathbf{w}}(\mathbf{w})$ thus we only require to specify the skew of each base density $q_{v_d}(v_d)$ and so we fix $\mu = 0$ and $\sigma = 1$ in all experiments, letting $\theta_d = \nu$.

Derivatives To evaluate the derivative of $\mathcal{SN}(v|\mu,\sigma,\nu)$ with respect to ν we use the fact that $f'(x) = f(x)\frac{d}{dx}\log f(x)$ and present the derivatives of $\log \mathcal{SN}(v|\mu,\sigma,\nu)$ with respect to ν

$$\frac{\partial}{\partial \nu}\log \mathcal{SN}(v|\mu,\sigma,\nu) = \frac{r\phi(\nu r)}{\Phi(r)} \quad \text{where} \quad r = \frac{v-\mu}{\sigma}. \tag{4.2}$$

Moments The first two moments of the distribution are given by

$$\begin{aligned}\langle v\rangle &= \mu + \sigma\delta\sqrt{2/\pi}\\ \mathrm{var}(v) &= \sigma^2\left(1 - \frac{2\delta^2}{\pi}\right)\end{aligned}$$

where $\delta = \frac{\nu}{\sqrt{1+\nu^2}}$.

Entropy The authors are not aware of an analytic form for the skew-normal density's entropy. Therefore we used univariate rectangular quadrature to compute these terms.

4.2 Generalised-normal

The generalised-normal distribution, $\mathcal{GN}(v|\mu,\alpha,\beta)$, is given by

$$\mathcal{GN}(v|\mu,\alpha,\beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)}e^{-\left(\frac{x-\mu}{\alpha}\right)^\beta} \tag{4.3}$$

where $\Gamma(x)$ is the Gamma function $\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}dt$, location parameter $\mu \in \mathbb{R}$, scale parameter $\alpha \in \mathbb{R}^+$ and shape parameter $\beta \in \mathbb{R}^+$. In practice we constrain $\beta > 1$ to ensure differentiability of the KL bound.

Derivatives The derivative of the log density with respect to β is

$$\frac{\partial}{\partial\beta}\log\mathcal{GN}(v|\mu,\alpha,\beta) = \frac{1}{\beta} + \frac{1}{\beta^2}g\left(\frac{1}{\beta}\right) - \left(\frac{|v-\mu|}{\alpha}\right)^\beta \log\left(\frac{|v-\mu|}{\alpha}\right). \tag{4.4}$$

Moments The first two moments of the generalised-normal distribution are:

$$\begin{aligned}\langle v\rangle &= \mu,\\ \mathrm{var}(v) &= \frac{\Gamma(3/\beta)}{\Gamma(1/\beta)}.\end{aligned}$$

Entropy The generalised-normal admits an analytic form for the differential entropy

$$H\left[\mathcal{GN}(v|\mu,\alpha,\beta)\right] = \frac{1}{\beta} - \log\left[\frac{\beta}{2\alpha\Gamma(1/\beta)}\right] \tag{4.5}$$

which in turn has the gradient

$$\frac{\partial}{\partial\beta}H\left[\mathcal{GN}(v|\mu,\alpha,\beta)\right] = -\frac{1}{\beta^2} - \frac{1}{\beta} + \psi\left(\frac{1}{\beta}\right)\frac{1}{\beta^2} \tag{4.6}$$

where $\psi(x)$ is the digamma function defined as $\psi(x) = \frac{d}{dx}\log\Gamma(x)$.

References

[1] E. Challis and D. Barber. Concave Gaussian Variational Approximations for Inference in Large-Scale Bayesian Linear Models. In *International Conference on Artificial Intelligence and Statistics, AISTATS*, 2011.