Appendix: Variational Bayesian dropout: pitfalls and fixes

Jiri Hron [1] **Alexander G. de G. Matthews** [1] **Zoubin Ghahramani** [1 2]

A. Proofs for Section 3

Notation and identities used throughout this section: $\psi(x)$ for the digamma function, $\psi(x+1) = \psi(x) + 1/x$, $\psi(k+1) = \mathrm{H}_k - \gamma$ where H_k is the k^{th} harmonic number and γ is the Euler–Mascheroni's constant, $\mathrm{Ei}(x) = -\int_{-x}^{\infty} \mathrm{e}^{-t}/t \,\mathrm{d}t$ is the exponential integral function, $\sum_{k=1}^{\infty} u^k \mathrm{H}_k / k! = \mathrm{e}^u(\gamma + \log u - \mathrm{Ei}(-u))$ (Dattoli & Srivastava, 2008; Gosper, 1996), and $\sum_{k=1}^{\infty} u^k/(k!\, k) = \mathrm{Ei}(u) - \gamma - \log u$ (Harris, 1957); the last two identities hold for $u > 0$. Importantly, we define $0^0 := 1$ unless stated otherwise.

Proof of Proposition 1. Denote the likelihood value by $\epsilon > 0$. Take an arbitrary number r such that $\epsilon > r > 0$. By continuity, we can find $\delta > 0$ such that $|w - 0| < \delta$ implies that the likelihood value is greater than r; let $A \ni 0$ denote the open ball of radius δ centred at 0. Because both the prior density and the likelihood function only take non-negative values, we can apply the Tonelli–Fubini's theorem to obtain,

$$\mathrm{Z} = \int_{\mathbb{R}^{\mathrm{D}-1}} p(\boldsymbol{W}_{\neg w}) \left[\int_{\mathbb{R}} p(w) p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{W}) \,\mathrm{d}w \right] \mathrm{d}\boldsymbol{W}_{\neg w}$$
$$> \int_{\mathbb{R}^{\mathrm{D}-1}} p(\boldsymbol{W}_{\neg w}) \left[\int_A \frac{\mathrm{C}}{|w|} r \,\mathrm{d}w \right] \mathrm{d}\boldsymbol{W}_{\neg w} = \infty \,,$$

where $\boldsymbol{W}_{\neg w}$ is a shorthand for $\boldsymbol{W} \setminus w$. When $\mathrm{Z} = \infty$, the measure of $\mathbb{R}^{\mathrm{D}}$ under $\mathrm{P}(\boldsymbol{W} \mid \boldsymbol{X}, \boldsymbol{Y})$ is infinite, and thus $p(\boldsymbol{W} \mid \boldsymbol{X}, \boldsymbol{Y})$ cannot be a proper probability density. $\square$

Proof of Proposition 2. Using standard identities about Gaussian random variables, and the fact that $v := \varepsilon^2$, $\varepsilon \sim \mathcal{N}(\mu/\sigma, 1)$, follows the non-central chi-squared distribution $\chi^2(\lambda, \nu)$ with $\nu = 1$ degrees of freedom and non-centrality parameter $\lambda = (\mu/\sigma)^2$, we have,

$$\mathop{\mathbb{E}}_{\mathrm{Q}(w)}[\log \mathrm{Q}(w)] - \mathop{\mathbb{E}}_{\mathrm{Q}(w)}[\log \mathrm{P}(w)]$$
$$= \mathop{\mathbb{E}}_{\mathrm{Q}(w)}[\log \mathrm{Q}(w)] - \log \mathrm{C} + \frac{1}{2} \mathop{\mathbb{E}}_{\mathrm{Q}(w)}[\log |w|^2]$$
$$= \mathrm{c}_1 + \frac{1}{2} \mathop{\mathbb{E}}_{\varepsilon \sim \mathcal{N}(\mu/\sigma, 1)}[\log \sigma^2 \varepsilon^2]$$
$$= \mathrm{c}_1 + \frac{1}{2}\left(\log \sigma^2 + \mathop{\mathbb{E}}_{v \sim \chi^2(\mu^2/\sigma^2, 1)}[\log v] \right)$$
$$= \mathrm{c}_2 + \frac{1}{2} \int_0^\infty \sum_{k=0}^{\infty} \mathrm{e}^{-\frac{\mu^2}{2\sigma^2}} \frac{(\frac{\mu^2}{2\sigma^2})^k}{k!} \frac{v^{k-\frac{1}{2}} \mathrm{e}^{-\frac{v}{2}}}{2^{k+\frac{1}{2}} \Gamma(k + \frac{1}{2})} \log v \,\mathrm{d}v \,,$$

where $\mathrm{c}_1 := -\frac{1}{2}\log(2\pi\mathrm{e}\sigma^2) - \log \mathrm{C}$, and we used the fact that $\chi^2(\lambda, \nu)$ is equivalent to a Poisson mixture of centralised chi-squared distributions. Let us define,

$$f_n(v) := \sum_{k=0}^{n} \mathrm{e}^{-\frac{\mu^2}{2\sigma^2}} \frac{(\frac{\mu^2}{2\sigma^2})^k}{k!} \frac{v^{k-\frac{1}{2}} \mathrm{e}^{-\frac{v}{2}}}{2^{k+\frac{1}{2}} \Gamma(k + \frac{1}{2})} \log v \,,$$

and rewrite the last integral as,

$$\int_0^\infty \lim_{n\to\infty} f_n(v) \mathrm{d}v$$
$$= \int_0^1 \lim_{n\to\infty} f_n(v) \mathrm{d}v + \int_1^\infty \lim_{n\to\infty} f_n(v) \mathrm{d}v \,.$$

Observe that $f_n \geq 0, \forall n \in \mathbb{N}$ and $f_n \uparrow f_\infty$ pointwise on $v \in [1, \infty)$, and $f_n < 0, \forall n \in \mathbb{N}$ and $f_n \downarrow f_\infty$ pointwise on $v \in [0, 1)$, for f_∞ defined as the pointwise limit of f_n. Hence we can use the monotone convergence theorem as long as the $|\int f_0(v) \mathrm{d}v| < \infty$. Using the identity $\mathbb{E}_{v\sim\chi^2(0,\nu)}[\log v] = \psi(\nu/2) - \log(1/2)$, we have,

$$\int_0^\infty f_n(v) \mathrm{d}v = \log 2 + \mathrm{e}^{-\frac{\mu^2}{2\sigma^2}} \sum_{k=0}^{n} \frac{(\frac{\mu^2}{2\sigma^2})^k}{k!} \psi(1/2 + k) \,,$$

which means that $|f_n| \in \mathrm{L}^1$ for all $n \in \mathbb{N}$. Because both $\int_0^1 |f_n(v)| \mathrm{d}v$ and $\int_1^\infty |f_n(v)| \mathrm{d}v$ are upper-bounded by $\int_0^\infty |f_n(v)| \mathrm{d}v$, we can apply the monotone convergence theorem to equate,

$$\int_0^1 \lim_{n\to\infty} f_n(v) \mathrm{d}v = \int_0^1 \lim_{n\to\infty} f_n(v) \mathrm{d}v$$
$$\int_1^\infty \lim_{n\to\infty} f_n(v) \mathrm{d}v = \int_1^\infty \lim_{n\to\infty} f_n(v) \mathrm{d}v \,,$$

and thus by Theorem 4.1.10 in (Dudley, 2002) conclude $\int_0^\infty f_\infty(v) \mathrm{d}v = \lim_{n\to\infty} \int_0^\infty f_n(v) \mathrm{d}v$. Substituting back,

$$\mathop{\mathbb{E}}_{\mathrm{Q}(w)}[\log \mathrm{Q}(w)] - \mathop{\mathbb{E}}_{\mathrm{Q}(w)}[\log \mathrm{P}(w)]$$

[1]Department of Engineering, University of Cambridge, Cambridge, United Kingdom [2]Uber AI Labs, San Francisco, California, USA. Correspondence to: Jiri Hron <jh2084@cam.ac.uk>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

$$= \mathrm{c}_2 + \frac{1}{2}\left(\log 2 + \mathrm{e}^{-\frac{\mu^2}{2\sigma^2}} \sum_{k=0}^{\infty} \frac{(\frac{\mu^2}{2\sigma^2})^k}{k!} \psi(1/2+k) \right)$$
$$= \mathrm{c}_3 - \frac{1}{2} \left.\frac{\partial \mathrm{M}(a;1/2;-\mu^2/(2\sigma^2))}{\partial a}\right|_{a=0},$$

where $\mathrm{M}(a;b;z)$ denotes the Kummer's function of the first kind, $\mathrm{c}_2 \coloneqq \mathrm{c}_1 + \frac{1}{2}\log(\sigma^2)$, and $\mathrm{c}_3 \coloneqq \mathrm{c}_2 - \frac{3}{2}\log 2 - \frac{1}{2}\gamma$. It is easy to check that Equation (3) holds for all $u \geq 0$ as long as we define $0^0 = 1$, and keep $0^k = 0, \forall k > 0$.

The last equality above was obtained using Wolfram Alpha (Wolfram—Alpha, 2017b); to validate this result, we performed an extensive numerical test, and will now show that the series indeed converges for $u = \mu^2/(2\sigma^2) \in [0,\infty)$, i.e. for all plausible values of u. The comparison test gives us convergence for $u \in (0,\infty)$:

$$\sum_{k=0}^{\infty} \frac{u^k}{k!}\psi(1/2+k) < \psi(1/2) + \sum_{k=1}^{\infty}\frac{u^k}{k!}\psi(1+k)$$
$$= \psi(1/2) + \sum_{k=1}^{\infty}\frac{u^k}{k!}(\mathrm{H}_k - \gamma)$$
$$= \psi(1/2) + \mathrm{e}^u(\gamma + \log u - \mathrm{Ei}(-u)) - \gamma(\mathrm{e}^u - 1)$$
$$= \psi(1/2) - \gamma + \mathrm{e}^u(\log u - \mathrm{Ei}(-u)),$$

where we use the fact that the individual summands are non-negative for $k \geq 1$ (which is also means we need not take the absolute value explicitly). It is trivial to check that the series converges at $u = 0$, and thus we have convergence for all $u \in [0,\infty)$.

To obtain the derivative with respect to u, we use the infinite series formulation from Equation (3), and the fact that the derivative of a power series within its radius of convergence is equal to the sum of its term-by-term derivatives (see (Gowers, 2014) for a nice proof). Using that only the infinite series in Equation (3) depends on u, we obtain,

$$\nabla_u \mathrm{e}^{-u} \sum_{k=0}^{\infty}\frac{u^k}{k!}\psi(1/2+k)$$
$$= \nabla_u \left(\mathrm{e}^{-u}\psi(1/2) + \mathrm{e}^{-u}\sum_{k=1}^{\infty}\frac{u^k}{k!}\psi(1/2+k)\right)$$
$$= -\mathrm{e}^{-u}\psi(1/2) + \mathrm{e}^{-u}\sum_{k=1}^{\infty}\left(\frac{u^{k-1}}{(k-1)!}\psi(1/2+k)\right) - \mathrm{e}^{-u}\sum_{k=1}^{\infty}\left(\frac{u^k}{k!}\psi(1/2+k)\right)$$
$$= \mathrm{e}^{-u}(\psi(3/2) - \psi(1/2)) + \mathrm{e}^{-u}\sum_{k=1}^{\infty}\left(\frac{u^k}{k!}\psi(3/2+k)\right) - \mathrm{e}^{-u}\sum_{k=1}^{\infty}\left(\frac{u^k}{k!}\psi(1/2+k)\right)$$
$$= 2\mathrm{e}^{-u} + \mathrm{e}^{-u}\sum_{k=1}^{\infty}\frac{u^k}{k!}\frac{1}{1/2+k} = \mathrm{e}^{-u}\sum_{k=0}^{\infty}\frac{u^k}{k!}\frac{1}{1/2+k}$$
$$= \frac{2D_+(\sqrt{u})}{\sqrt{u}},$$

for $u > 0$ and is equal to 2 if $u = 0$; in our case, the condition $u \geq 0$ is satisfied by definition; to obtain the expression in Equation (5), notice that the above series is multiplied by $1/2$ in Equation (3). Equality of the last infinite series to $2D_+(\sqrt{u})/\sqrt{u}$, was again obtained using Wolfram Alpha (Wolfram—Alpha, 2017a); the result was numerically validated, and convergence on $u \in (0,\infty)$ can again be established using the comparison test:

$$\sum_{k=0}^{\infty}\left|\frac{u^k}{k!}\frac{1}{1/2+k}\right| = \sum_{k=0}^{\infty}\frac{u^k}{k!}\frac{1}{1/2+k} < 2 + \sum_{k=1}^{\infty}\frac{u^k}{k!}\frac{1}{k}$$
$$= 2 + \mathrm{Ei}(u) - \gamma - \log u.$$

The convergence at $u = 0$ can be checked trivially, yielding convergence for all $u \in [0,\infty)$.

$D_+(u)$ and $\sqrt{u}$ are continuous on $(0,\infty)$, and $\sqrt{u} > 0$; hence $D_+(u)/\sqrt{u}$ is continuous on $(0,\infty)$, and from definition of the Dawson integral $\lim_{u\to 0_+} D_+(\sqrt{u})/\sqrt{u} = 1$, i.e. the gradient is continuous in u on $[0,\infty)$. $\square$

Proof of Corollary 3. We use the conclusion of Proposition 2 which established differentiability for $u \in [0,\infty)$ (and thus continuity on the same interval). To show that $\mathrm{KL}\left(\mathrm{Q}(w) \| \mathrm{P}(w)\right)$ is strictly increasing for $u \in [0,\infty)$, it is sufficient to observe,

$$\nabla_u \mathrm{KL}\left(\mathrm{Q}(w)\|\,\mathrm{P}(w)\right) = \frac{1}{2}\mathrm{e}^{-u}\sum_{k=0}^{\infty}\frac{u^k}{k!}\frac{1}{1/2+k} > 0,$$

because each summand is strictly positive for $u \in [0,\infty)$ (given $0^0 = 1$). By a simple application of the mean value theorem, we conclude $\mathrm{KL}\left(\mathrm{Q}(w)\|\,\mathrm{P}(w)\right)$ is strictly increasing in u on $[0,\infty)$. $\square$

B. Proofs for Section 4

Throughout this section, let $(\mathbb{R}^{\mathrm{D}}, \|\cdot\|_2)$ be the D-dimensional Euclidean metric space, $\mathcal{T}$ the usual topology, and $\mathcal{B}$ the corresponding Borel σ-algebra. Let λ^{M} be the M-dimensional Lebesgue measure[1]. P, Q will be probability measures, P with continuous density p w.r.t. λ^{D}, and Q concentrated on some $S \in \mathcal{B}$, which is either (at most) countable or a linear manifold. Let K_S be the Hausdorff

[1]More precisely the restriction of the M-dimensional Lebesgue measure to the corresponding Borel σ-algebra. We will be using the term Lebesgue measure instead of the sometimes used term *Borel measure* which we used to refer to any measure defined on the Borel σ-algebra.

dimension of S, i.e. zero in the countable, and $\dim(S)$ in the linear manifold case (with $\dim$ denoting the Hamel dimension). Q has a density q w.r.t. the counting measure for the countable $\mathbb{Q}^{\mathrm{D}}$,[2] or w.r.t. λ^{K_S} in the linear manifold case. In the (at most) countable case, further assume that $\operatorname{diam}(S) < \infty$ if S is infinite. If S is a linear manifold, further assume that q is continuous w.r.t. the trace topology $\mathcal{T}_S$, and that both q and p are bounded; denote the bounds on densities q and p by C_q and C_p respectively. We will be using m_S as a shorthand for either of the corresponding dominating measures of q. Finally, the convolution of two Borel measures μ, ν on $\mathbb{R}^{\mathrm{D}}$ will be denoted by $\mu \star \nu$ where for any $B \in \mathcal{B}$ we have $(\mu \star \nu)(B) = \int_{\mathbb{R}^{\mathrm{D}}} \mu(B - x) \mathrm{d}\nu(x)$.

We will be using the following fact: because $(\mathbb{R}^{\mathrm{D}}, \|\cdot\|_2)$ is a complete separable metric space, every finite Borel measure is regular by Ulam's theorem (Dudley, 2002, Theorem 7.1.4), and thus tight by definition. Hence for any probability measure P on $(\mathbb{R}^{\mathrm{D}}, \mathcal{B})$ and every $\varepsilon > 0$, there exists a compact set $C \in \mathcal{B}$ s.t. $\mathrm{P}(C) > 1 - \varepsilon$. The axiom of choice is assumed throughout.

The proofs of Theorems 4 and 5 will be divided into propositions, each proven in a subsection corresponding to the limiting construction used.

Proof of Theorem 4. Combine Propositions 8 and 17. □

Proof of Theorem 5. Use Proposition 9. □

B.1. Convolutional approach

Before approaching the proof of Proposition 9, observe that we can simplify the case of S being a linear manifold by WLOG assuming that $S = \mathbb{R}^{\mathrm{K}_S} \times \{0\}^{\mathrm{D}-\mathrm{K}_S}$, i.e. the space of K_S-dimensional vectors padded out by zeros at the end. This is because we have defined q and p to be the densities w.r.t. the corresponding Lebesgue measures which are translation and rotation invariant.

The following definitions will become handy: let Z and $\mathcal{E}$ be random variables respectively distributed according to the laws Q and $\mathrm{P}_{\mathcal{E}} = \mathcal{N}(0, I_{\mathrm{D}})$. Define the shorthands $\mathcal{E}^{(n)} := \mathcal{E}/\sqrt{n}$ and $Z^{(n)} := Z + \mathcal{E}^{(n)}$. We will further define the random variables $\widetilde{\mathcal{E}}^{(n)} := \mathcal{E}^{(n)}_{1:\mathrm{K}_S}$ and $\widetilde{Z}^{(n)} := Z^{(n)}_{1:\mathrm{K}_S}$ where the $1\colon \mathrm{K}_S$ denotes reducing the corresponding vectors to their first K_S components. The relevant distributions will be denoted as follows: $\mathrm{P}^{(n)}_{\mathcal{E}} := \mathrm{Law}(\mathcal{E}^{(n)})$, $\mathrm{P}^{(n)}_{\widetilde{\mathcal{E}}} := \mathrm{Law}(\widetilde{\mathcal{E}}^{(n)})$, $\mathrm{Q}^{(n)} = \mathrm{Law}(Z^{(n)})$, and $\widetilde{\mathrm{Q}}^{(n)} := \mathrm{Law}(\widetilde{Z}^{(n)})$. Notice that $(\mathcal{E}^{(n)}, \widetilde{\mathcal{E}}^{(n)})$ and $(Z^{(n)}, \widetilde{Z}^{(n)})$ are both deterministically coupled, joint laws being the corresponding push-forwards of the $\mathrm{P}_{\mathcal{E}}$ and Q distributions. Also observe that we only convolve the approximating distribution with the Gaussian noise, and not the target P. Hence $\mathrm{P}^{(n)} = \mathrm{P}, \forall n \in \mathbb{N}$; we will thus omit the superscript here.

We will also use the following construction: let $\bar{B}_r(x) \subset \mathbb{R}^k$, $k \in \mathbb{N}$, be a closed ball centred at $x \in \mathbb{R}^{\mathrm{K}_S}$ and with radius $r > 0$. Then for some fixed $\eta > 0$, we define a continuous compactly supported[3] function $h_{r,\eta}$ where,

$$h_{r,\eta}(z) = \begin{cases} 1 & \text{, if } x \in \bar{B}_r(x) \\ 0 & \text{, if } x \in F_{r,\eta} \\ \frac{r+\eta-\|z-x\|_r}{\eta} & \text{, else.} \end{cases} , \tag{10}$$

with $F_{\delta,\eta}$ defined as complement of the ball $B_{r+\eta}(x)$.

Finally observe $\mathrm{Q}^{(n)} = \mathrm{Q} \star \mathcal{N}(0, n^{-1} I_{\mathrm{D}})$, and $\widetilde{\mathrm{Q}}^{(n)} = \mathrm{Q} \star \mathcal{N}(0, n^{-1} I_{\mathrm{K}_S})$ by the standard marginalisation properties of Gaussian distributions. As a corollary of (Dudley, 2002, Proposition 9.1.6), we have,

$$q^{(n)}(x) = \int \phi_{x, n^{-1} I_{\mathrm{D}}}(z) q(z) \, m_S(\mathrm{d}z) \qquad , x \in \mathbb{R}^{\mathrm{D}} , \tag{11}$$

$$\widetilde{q}^{(n)}(x) = \int \phi_{x, n^{-1} I_{\mathrm{K}_S}}(z) q(z) \, m_S(\mathrm{d}z) \quad , x \in S , \tag{12}$$

where $\phi_{\mu,\Sigma}$ is the density function of $\mathcal{N}(\mu, \Sigma)$. In Equation (12), it would be more precise to write $\phi_{x, n^{-1} I_{\mathrm{K}_S}}(z_{1:\mathrm{K}_S})$ by which we get rid off the trailing zeros (c.f. beginning of this section). Because $\|x - z\|_2^2 = \|x_{1:\mathrm{K}_S} - z_{1:\mathrm{K}_S}\|_2^2$ for all $x, z \in \mathbb{R}^{\mathrm{K}_S} \times \{0\}^{\mathrm{D}-\mathrm{K}_S}$, we omit the subscript to reduce clutter unless confusion may arise.

Proposition 8. *Let the relevant assumptions at the beginning of Appendix B and in Theorem 4 hold. We consider two cases:* $\log \frac{q}{p} \in \mathrm{L}^1(\mathrm{Q})$ *and* $\log \frac{q}{p} \notin \mathrm{L}^1(\mathrm{Q})$*. If* $\log \frac{q}{p} \in \mathrm{L}^1(\mathrm{Q})$*, further assume that the collection of random variables* $\{\log p(Z^{(n)})\}_{n \in \mathbb{N}}$ *is uniformly integrable.*[4]

Then,

$$\lim_{n \to \infty} \left\{ \mathrm{KL}\left(\mathrm{Q}^{(n)} \,\|\, \mathrm{P}\right) - s^{(n)} \right\} = \mathop{\mathbb{E}}_{\mathrm{Q}} \left(\log \tfrac{q}{p} \right) ,$$

with $s^{(n)} := -\frac{\mathrm{D}}{2} \log(2\pi \mathrm{e} n^{-1})$.

Proof of Proposition 8. First, assume that $\log \frac{q}{p} \in \mathrm{L}^1(\mathrm{Q})$. By Lemma 11, we can focus on convergence of the cross-entropy and negative entropy individually. By Lemma 12, the cross-entropy term converges.

Notice that the density w.r.t. the counting measure can be written using the Kronecker's delta function δ_{Kr} as $q(x) =$

[2] We use the countable measure on rationals to avoid having to deal with a dominating measure that is not σ-finite.

[3] Support is the closure of the set where the function is non-zero.

[4] A useful sufficient condition is provided in Proposition 10.

$\sum_{i\in\mathbb{N}} \rho_i \delta_{\mathrm{Kr}}\,(x - m_i)$, where $\rho_i \geq 0$, $\sum_{i\in\mathbb{N}} \rho_i = 1$, and $m_i \in \mathbb{R}^{\mathrm{D}}, \forall i \in \mathbb{N}$. Then the convolved density w.r.t. λ^{D} is,

$$q^{(n)}(x) = \sum_{i\in\mathbb{N}} \rho_i\, \phi_{m_i, n^{-1} I_{\mathrm{D}}}(x)\,.$$

Hence we can use the properties of multivariate normal distributions and the Tonelli–Fubini's theorem to write,

$$\int q^{(n)} \log q^{(n)}\, \mathrm{d}\lambda^{\mathrm{D}} = -\frac{\mathrm{D}}{2}\log(2\pi n^{-1}) + \sum_{i\in\mathbb{N}} \int \rho_i \phi_{0, I_{\mathrm{D}}}(\xi) \log\Bigg[\sum_{j\in\mathbb{N}} \rho_j \mathrm{e}^{-\frac{\left\| m_i + \xi/\sqrt{n} - m_j \right\|_2^2}{2n^{-1}}}\Bigg] \mathrm{d}\lambda^{\mathrm{D}}\,,$$

which can be viewed as an integral over the product space $S \times \mathbb{R}^{\mathrm{D}}$ w.r.t. the product measure of Q and $\mathcal{N}(0, I_{\mathrm{D}})$. For any fixed $i \in \mathbb{N}$ and $\xi \in \mathbb{R}^{\mathrm{D}}$, define,

$$f^{(n)}(i, \xi) \coloneqq \log\Bigg[\sum_{j\in\mathbb{N}} \rho_j \exp\bigg(-\frac{\left\| m_i + \xi/\sqrt{n} - m_j \right\|_2^2}{2n^{-1}}\bigg)\Bigg]\,.$$

Then $f^{(n)}(i,\xi) \to \log[\rho_i \exp(-\|\xi\|_2^2/2)] =: f^{(*)}(i,\xi)$ pointwise as $n \to \infty$. In fact, because the terms inside the logarithm are all non-negative and $\rho_i \exp(-\|\xi\|_2^2/2)$ is the i^{th} summand, we get $f^{(n)}(i,\xi) \downarrow f^{(*)}(i,\xi)$ by monotonicity of the logarithm. Because $f^{(n)}(i,\xi) \leq \log(1) = 0$, we can use the monotone convergence theorem to establish,

$$\sum_{i\in\mathbb{N}} \rho_i \mathop{\mathbb{E}}_{\mathcal{N}(0,I_{\mathrm{D}})}(f^{(n)}(i,\xi)) \downarrow \sum_{i\in\mathbb{N}} \rho_i \mathop{\mathbb{E}}_{\mathcal{N}(0,I_{\mathrm{D}})}(f^{(*)}(i,\xi))\,.$$

Solving the limit integral,

$$\sum_{i\in\mathbb{N}} \rho_i \mathop{\mathbb{E}}_{\mathcal{N}(0,I_{\mathrm{D}})}(f^{(*)}(i,\xi)) = \sum_{i\in\mathbb{N}} \rho_i \log(\rho_i) - \frac{\mathrm{D}}{2}\,,$$

we conclude (using $\log \mathrm{e} = 1$),

$$\int q^{(n)} \log q^{(n)}\, \mathrm{d}\lambda^{\mathrm{D}} + \frac{\mathrm{D}}{2}\log(2\pi \mathrm{e} n^{-1}) \to \sum_{i\in\mathbb{N}} \rho_i \log(\rho_i) = \mathop{\mathbb{E}}_{\mathrm{Q}}(\log q)\,.$$

It remains to show that if $\log \frac{q}{p} \notin \mathrm{L}^1(\mathrm{Q})$, the sequence $\{\mathrm{KL}\,(\mathrm{Q}^{(n)} \| \, \mathrm{P}) - s_{\mathrm{K}_S}^{(n)}\}_{n\in\mathbb{N}}$ also diverges.

Because we have $q(x) = \sum_i \rho_i \delta_{\mathrm{Kr}}\,(x - m_i)$, and $q^{(n)}(x) = \sum_i \rho_i \phi_{m_i, n^{-1} I_{\mathrm{D}}}(x)$, we can write,

$$\frac{\mathrm{D}}{2}\log(2\pi n^{-1}) + \log q^{(n)}(x) = \log\left[\sum_i \rho_i \mathrm{e}^{-\frac{n}{2}\|x - m_i\|_2^2}\right],$$

and thus we can define $\widetilde{q}^{(n)}(x) \coloneqq \sum_i \rho_i \mathrm{e}^{-\frac{n}{2}\|x-m_i\|_2^2}$ for the (at most) countable support case. Clearly $\widetilde{q} \to q$ pointwise. To establish continuity, notice that the $\sum_i \rho_i = 1$ requirement implies that $\forall \varepsilon > 0$, $\exists k \in \mathbb{N}$ s.t. $\sum_{i>k} \rho_i < \varepsilon/2$, and that for any $x, y \in \mathbb{R}^{\mathrm{D}}$ and $i \in \mathbb{N}$,

$$\left| \mathrm{e}^{-\frac{n}{2}\|x-m_i\|_2^2} - \mathrm{e}^{-\frac{n}{2}\|y-m_i\|_2^2} \right| < 1\,.$$

Because individual summands are continuous, for any $x \in \mathbb{R}^{\mathrm{D}}$, we can take the minimum amongst radii which guarantee that each term will not change by more than $\frac{\varepsilon}{2k}$ for any $y \in \mathbb{R}^{\mathrm{D}}$ sufficiently close. Hence $\widetilde{q}^{(n)}$ is continuous for every $n \in \mathbb{N}$.

Notice that we only need to show we only need to show,

$$\mathbb{E}\,|\log \tfrac{\widetilde{q}^{(n)}(Z^{(n)})}{p(Z^{(n)})}| \to \infty\,.$$

If $|\log \frac{\widetilde{q}^{(n)}(Z^{(n)})}{p(Z^{(n)})}|$ is not a.s. finite then we are done. In the case when a.s. finiteness holds, it must be true that $\widetilde{q}^{(n)}(\widetilde{Z}^{(n)}) > 0 \implies p(Z^{(n)}) > 0$ a.s. Thus by continuity of the logarithm, absolute value, p, $\widetilde{q}^{(n)}$ (see above), and the pointwise convergence $\widetilde{q}^{(n)} \to q$ and a.s. convergence of both $Z^{(n)}$ to Z, we have $|\log \frac{\widetilde{q}^{(n)}(Z^{(n)})}{p(Z^{(n)})}| \to |\log \frac{q(Z)}{p(Z)}|$ a.s. Hence we can use Fatou's lemma to establish,

$$\infty = \mathbb{E}\big|\log \tfrac{q(Z)}{p(Z)}\big| \leq \liminf_{n\to\infty} \mathbb{E}\big|\log \tfrac{q^{(n)}(\widetilde{Z}^{(n)})}{p(Z^{(n)})}\big|\,.$$

which means $\{\mathrm{KL}\,(\mathrm{Q}^{(n)} \| \, \mathrm{P}) - s^{(n)}\}_{n\in\mathbb{N}}$ diverges. $\square$

Proposition 9. *Let the relevant assumptions stated at the beginning of Appendix B and in Theorem 4 hold. We consider two cases:* $\log \frac{q}{p} \in \mathrm{L}^1(\mathrm{Q})$ *and* $\log \frac{q}{p} \notin \mathrm{L}^1(\mathrm{Q})$. *If* $\log \frac{q}{p} \in \mathrm{L}^1(\mathrm{Q})$, *assume that the collection of random variables* $\{\log p(Z^{(n)})\}_{n\in\mathbb{N}}$ *is uniformly integrable,*[5] *and that* $\mathbb{E}\|Z\|_2^2 < \infty$.

Then,

$$\lim_{n\to\infty} \Big\{\mathrm{KL}\,(\mathrm{Q}^{(n)} \| \, \mathrm{P}) - s_{\mathrm{K}_S}^{(n)}\Big\} = \mathop{\mathbb{E}}_{\mathrm{Q}}\big(\log \tfrac{q}{p}\big)\,,$$

with $s_{\mathrm{K}_S}^{(n)} \coloneqq -\frac{\mathrm{D}-\mathrm{K}_S}{2}\log(2\pi \mathrm{e} n^{-1})$.

Proof of Proposition 9. First, assume that $\log \frac{q}{p} \in \mathrm{L}^1(\mathrm{Q})$. By Lemma 11, we can focus on convergence of the cross-entropy and negative entropy individually. By Lemma 12, the cross-entropy term converges.

For the negative entropy term, WLOG assume $S = \mathbb{R}^{\mathrm{K}_S} \times \{0\}^{\mathrm{D}-\mathrm{K}_S}$. By Lemma 13, we need to prove that,

$$\mathbb{E}\left(\log \widetilde{q}^{(n)}(\widetilde{Z}^{(n)})\right) \to \mathbb{E}\left(\log q(Z)\right)\,.$$

First, we will establish that $\log \widetilde{q}^{(n)}(\widetilde{Z}^{(n)}) \to \log q(Z)$ a.s. By definition, $\widetilde{Z}^{(n)} = Z_{1:\,\mathrm{K}_S} + \widetilde{\mathcal{E}}/\sqrt{n}$ (the subscript/padding with zeros where appropriate will be again

[5] A useful sufficient condition is provided in Proposition 10.

omitted from now on). Clearly, $Z + \widetilde{\mathcal{E}}/\sqrt{n} \to Z$ a.s. Hence by the triangle inequality for fixed values $Z = z$ and $\widetilde{\mathcal{E}} = \xi$,

$$\begin{aligned} &\left|\log \widetilde{q}^{(n)}(z + \xi/\sqrt{n}) - \log q(z)\right| \\ &\quad\leq \left|\log \widetilde{q}^{(n)}(z + \xi/\sqrt{n}) - \log q(z + \xi/\sqrt{n})\right| \\ &\qquad + \left|\log q(z + \xi/\sqrt{n}) - \log q(z)\right|, \end{aligned} \tag{13}$$

The second term on the RHS goes to zero with $n \to \infty$ by continuity of q. Turning to the first term, we can use the continuity of the logarithm to see that we only need to show that $\forall \varepsilon > 0$, $\exists \mathrm{N} \in \mathbb{N}$ s.t. $|\widetilde{q}^{(n)}(z + \xi/\sqrt{n}) - q(z + \xi/\sqrt{n})| < \varepsilon$ for all $n \geq \mathrm{N}$. Observe,

$$\begin{aligned} &|\widetilde{q}^{(n)}(z + \tfrac{\xi}{\sqrt{n}}) - q(z + \tfrac{\xi}{\sqrt{n}})| \\ &\leq \int \left| q(z + \tfrac{\xi+u}{\sqrt{n}}) - q(z + \tfrac{\xi}{\sqrt{n}})\right| \mathcal{N}(0, I_{\mathrm{K}_S})(\mathrm{d}u)\,. \end{aligned}$$

Because q is continuous, it is uniformly continuous on compact sets. Hence we can fix $\eta > 0$ and define $F := \bar{B}_{\|\xi\|_2+\eta}(z)$, the closed ball centred at z with radius $\|\xi\|_2+\eta$, which is compact by the Heine–Borel theorem. Use uniform continuity to find $t > 0$ s.t. $\forall (x, y) \in F$ with $\|x - y\|_2 < t$ implies $|q(x) - q(y)| < \varepsilon$, and WLOG assume $t \leq \eta$ (take $t = \eta$ if not). For $A := \{x \in \mathbb{R}^{\mathrm{K}_S} : \|x\|_2 < t\}$,

$$\begin{aligned} &\int \left| q(z + \tfrac{\xi+u}{\sqrt{n}}) - q(z + \tfrac{\xi}{\sqrt{n}})\right| \mathcal{N}(0, I_{\mathrm{K}_S})(\mathrm{d}u) \\ &\leq \int \mathbb{I}_A\left(\tfrac{u}{\sqrt{n}}\right) \left| q(z + \tfrac{\xi+u}{\sqrt{n}}) - q(z + \tfrac{\xi}{\sqrt{n}})\right| \mathcal{N}(0, I_{\mathrm{K}_S})(\mathrm{d}u) \\ &\quad + \mathrm{C}_q \mathcal{N}(0, n^{-1} I_{\mathrm{K}_S})(A^{\mathrm{C}})\,, \end{aligned}$$

where the latter term on the RHS clearly vanishes as $n \to \infty$. Because $\|z + \frac{\xi+u}{\sqrt{n}} - z\|_2 \leq \|\xi\|_2 + \|\frac{u}{\sqrt{n}}\|_2 < \|\xi\|_2 + t$ and $t \leq \eta$, the first integral is clearly over a subset of F. Since $\|z + \frac{\xi+u}{\sqrt{n}} - z + \frac{\xi}{\sqrt{n}}\|_2 = \|\frac{u}{\sqrt{n}}\|_2$ which is lower than t on A by definition, the uniform continuity yields an upper bound,

$$|\widetilde{q}^{(n)}(z + \tfrac{\xi}{\sqrt{n}}) - q(z + \tfrac{\xi}{\sqrt{n}})| < \varepsilon + \mathrm{C}_q \mathcal{N}(0, n^{-1} I_{\mathrm{K}_S})(A^{\mathrm{C}})\,,$$

where the right hand side converges monotonically to ε as desired. Therefore,

$$\log \widetilde{q}^{(n)}(\widetilde{Z}^{(n)}) \to \log q(Z) \quad \text{a.s.}$$

The convergence in mean is proved next.

We define $Y := \log q(Z)$ and $\widetilde{Y}^{(n)} := \log \widetilde{q}^{(n)}(\widetilde{Z}^{(n)})$ and the corresponding probability measures $\nu := \mathrm{Law}(Y)$, $\nu^{(n)} := \mathrm{Law}(\widetilde{Y}^{(n)})$. Because a.s. convergence implies convergence in distribution, we have $\nu^{(n)} \to \nu$ weakly. Hence $\{\nu^{(n)}\}_{n\in\mathbb{N}}$ is uniformly tight by Proposition 9.3.4 in (Dudley, 2002), and so is $\{\nu^{(n)}\}_{n\in\mathbb{N}} \cup \{\nu\}$.

Therefore we can find a compact set $\bar{B}_\delta$ s.t. $\nu(\bar{B}_\delta) > 1 - \delta$ and $\nu^{(n)}(\bar{B}_\delta) > 1 - \delta, \forall n \in \mathbb{N}$ for any $\delta > 0$. WLOG we can assume that $\bar{B}_\delta$ is a closed interval as compactness is equivalent to closedness and boundedness for Euclidean spaces by the Heine–Borel theorem. Thus for any compact $\bar{B}_\delta$ we can find a closed (compact) interval $[s_\delta - r_\delta, s_\delta + r_\delta]$ which includes it.

Convergence in distribution implies that for any $f \in C_b(\mathbb{R})$, $\mathbb{E}\, f(\widetilde{Y}^{(n)}) \to \mathbb{E}\, f(Y)$ as $n \to \infty$. The identity function Id on $\mathbb{R}^{\mathrm{K}_S}$ is trivially continuous for the usual topology, but not bounded; however it is bounded on compact sets like $\bar{B}_\delta$. We thus approximate Id by a continuous compactly supported functions $h_{\delta,\eta}\,\mathrm{Id}$ where $h_{\delta,\eta}$ is constructed as in Equation (10) with $r = r_\delta$ for some $\eta > 0$.

Using the triangle inequality,

$$\begin{aligned} &\left|\mathop{\mathbb{E}}_{\nu}(\mathrm{Id}) - \mathop{\mathbb{E}}_{\nu^{(n)}}(\mathrm{Id})\right| \leq \left|\mathop{\mathbb{E}}_{\nu}(\mathrm{Id}) - \mathop{\mathbb{E}}_{\nu}(h_{\delta,\eta}\mathrm{Id})\right| \\ &+ \left|\mathop{\mathbb{E}}_{\nu}(h_{\delta,\eta}\mathrm{Id}) - \mathop{\mathbb{E}}_{\nu^{(n)}}(h_{\delta,\eta}\mathrm{Id})\right| + \left|\mathop{\mathbb{E}}_{\nu^{(n)}}(h_{\delta,\eta}\mathrm{Id}) - \mathop{\mathbb{E}}_{\nu^{(n)}}(\mathrm{Id})\right|. \end{aligned}$$

Starting with the first term on the RHS, we can upper bound,

$$\left|\mathop{\mathbb{E}}_{\nu}(\mathrm{Id}) - \mathop{\mathbb{E}}_{\nu}(h_{\delta,\eta}\mathrm{Id})\right| \leq \mathop{\mathbb{E}}_{\nu}\left|(1 - h_{\delta,\eta})\mathrm{Id}\right| \leq \mathop{\mathbb{E}}_{\nu} \mathbb{I}_{\bar{B}_\delta^{\mathrm{C}}}|\mathrm{Id}|\,,$$

and observe that $\mathbb{E}_\nu|\mathrm{Id}| \leq -\mathbb{E}_{\mathrm{Q}}(\log \bar{q}) + \left|\log \mathrm{C}_q\right|$, $\bar{q} := q/\mathrm{C}_q$, which by $\log q \in \mathrm{L}^1(\mathrm{Q})$ implies that $\mathrm{Id} \in \mathrm{L}^1(\nu)$. Because any finite number of integrable functions is uniformly integrable, we can use Theorem 10.3.5 in (Dudley, 2002) to conclude that $\forall \varepsilon > 0$, there exists $\delta > 0$ s.t. $\mathbb{E}_\nu\, \mathbb{I}_{\bar{B}_\delta^{\mathrm{C}}}|\mathrm{Id}| \leq \varepsilon$.

Turning to the last term, we can again upper bound $\left|\mathbb{E}_{\nu^{(n)}}(h_{\delta,\eta}\mathrm{Id}) - \mathbb{E}_{\nu^{(n)}}(\mathrm{Id})\right|$ with $\mathbb{E}_{\nu^{(n)}}\, \mathbb{I}_{\bar{B}_\delta^{\mathrm{C}}}|\mathrm{Id}|$, $\forall n \in \mathbb{N}$. In this case, it will be beneficial to revert to the original representation:

$$\mathop{\mathbb{E}}_{\nu^{(n)}} \mathbb{I}_{\bar{B}_\delta^{\mathrm{C}}}|\mathrm{Id}| = \mathop{\mathbb{E}}_{\widetilde{\mathrm{Q}}^{(n)}} \mathbb{I}_{(A_\delta^{(n)})^{\mathrm{C}}}|\log \widetilde{q}^{(n)}|\,,$$

with $A_\delta^{(n)} := (\log \widetilde{q}^{(n)})^{-1}(\bar{B}_\delta)$; observe that because $\nu^{(n)} = (\log \widetilde{q}^{(n)})_{\#}\widetilde{\mathrm{Q}}^{(n)}$, $\widetilde{\mathrm{Q}}^{(n)}(A_\delta^{(n)}) > 1 - \delta, \forall n \in \mathbb{N}$, by definition. By Lemma 14, each $\widetilde{q}^{(n)}$ is bounded by C_q, thus we WLOG assume that $|\log \widetilde{q}^{(n)}| = -\log \widetilde{q}^{(n)}$ as the normalisation by C_q will only add a vanishing term $\mathrm{C}_q \widetilde{\mathrm{Q}}^{(n)}((A_\delta^{(n)})^{\mathrm{C}}) \leq \mathrm{C}_q \delta$ on the RHS, $\forall n \in \mathbb{N}$. Then,

$$\begin{aligned} &\mathop{\mathbb{E}}_{\widetilde{\mathrm{Q}}^{(n)}} \mathbb{I}_{(A_\delta^{(n)})^{\mathrm{C}}}|\log \widetilde{q}^{(n)}| \\ &= -\mathop{\mathbb{E}}_{\widetilde{\mathrm{Q}}^{(n)}}(\mathbb{I}_{(A_\delta^{(n)})^{\mathrm{C}}} \log \widetilde{q}^{(n)}) \pm \mathop{\mathbb{E}}_{\widetilde{\mathrm{Q}}^{(n)}}(\mathbb{I}_{(A_\delta^{(n)})^{\mathrm{C}}} \log \phi_{0,I_{\mathrm{K}_S}}) \\ &= -\mathop{\mathbb{E}}_{\widetilde{\mathrm{Q}}^{(n)}}\left(\mathbb{I}_{(A_\delta^{(n)})^{\mathrm{C}}} \log \frac{\widetilde{q}^{(n)}}{\phi_{0,I_{\mathrm{K}_S}}}\right) \\ &\quad - \mathop{\mathbb{E}}_{\widetilde{\mathrm{Q}}^{(n)}}(\mathbb{I}_{(A_\delta^{(n)})^{\mathrm{C}}} \log \phi_{0,I_{\mathrm{K}_S}}) \end{aligned}$$

$$\leq \ -\widetilde{\mathrm{Q}}^{(n)}((A_\delta^{(n)})^{\mathrm{C}}) \log \frac{\widetilde{\mathrm{Q}}^{(n)}((A_\delta^{(n)})^{\mathrm{C}})}{\mathcal{N}(0,I_S)((A_\delta^{(n)})^{\mathrm{C}})} - \mathop{\mathbb{E}}_{\widetilde{\mathrm{Q}}^{(n)}} (\mathbb{I}_{(A_\delta^{(n)})^{\mathrm{C}}} \log \phi_{0,I_{\mathrm{K}_S}}) ,$$

where the inequality is by Equation (7) on p. 177 in (Gray, 2011), and the fact that non-degenerate Gaussian distributions on Euclidean spaces are equivalent to the corresponding Lebesgue measure (i.e. $\mathcal{N}(\mu,\Sigma) \ll \lambda^k$ and $\lambda^k \ll \mathcal{N}(\mu,\Sigma)$ for all $k \in \mathbb{N}, \mu \in \mathbb{R}^k$ and positive definite Σ) which means that $\widetilde{\mathrm{Q}}^{(n)} \ll \mathcal{N}(0, I_{\mathrm{K}_S}), \forall n \in \mathbb{N}$, and thus the $\mathrm{KL}\,(\widetilde{\mathrm{Q}}^{(n)} \| \, \mathcal{N}(0, I_{\mathrm{K}_S}))$ is well-defined. Because $\widetilde{\mathrm{Q}}^{(n)} \ll \mathcal{N}(0, I_{\mathrm{K}_S})$, $\mathcal{N}(0,I_S)((A_\delta^{(n)})^{\mathrm{C}}) > 0$ if $\widetilde{\mathrm{Q}}^{(n)}((A_\delta^{(n)})^{\mathrm{C}}) > 0$ which means we can upper bound the first term on the RHS by,

$$-\widetilde{\mathrm{Q}}^{(n)}((A_\delta^{(n)})^{\mathrm{C}}) \log \widetilde{\mathrm{Q}}^{(n)}((A_\delta^{(n)})^{\mathrm{C}}) ,$$

which vanishes as $\delta \to 0$. The second term is equal to,

$$-\widetilde{\mathrm{Q}}^{(n)}((A_\delta^{(n)})^{\mathrm{C}}) \tfrac{\mathrm{K}_S}{2} \log(2\pi) - \tfrac{1}{2}\, \mathbb{E}\, \mathbb{I}_{(A_\delta^{(n)})^{\mathrm{C}}} \left\| Z + \widetilde{\mathcal{E}}/\sqrt{n} \right\|_2^2 ,$$

where the first term again vanishes as $\delta \to 0$. Combining $\Gamma(0) = 1, \Gamma(\frac{1}{2}) = \sqrt{\pi}$ and Lemma 16, the latter term can be upper bounded by,

$$\mathbb{E}(\mathbb{I}_{(A_\delta^{(n)})^{\mathrm{C}}} \|Z\|_2^2) + \frac{\mathbb{E}\|Z\|_2}{\sqrt{2\pi n}} + \frac{\mathbb{E}\,\|\widetilde{\mathcal{E}}\|_2^2}{n} .$$

As $\mathbb{E}\,\|\widetilde{\mathcal{E}}\|_2^2 = \mathrm{K}_S$, the last term will vanish as $n \to \infty$. Because we have assumed $\mathbb{E}\|Z\|_2^2 < \infty$, Hölder's inequality yields $\mathbb{E}\|Z\|_2 < \infty$ and thus the second term will also disappear as $n \to \infty$. $\mathbb{E}\|Z\|_2^2 < \infty$ can also be used to determine that the singleton set $\{\|Z\|_2^2\}$ is uniformly integrable and thus again by Theorem 10.3.5 in (Dudley, 2002) $\mathbb{E}(\mathbb{I}_{(A_\delta^{(n)})^{\mathrm{C}}} \|Z\|_2^2) \to 0$ as $\delta \to 0$. Notice that the terms that vanish with $\delta \to 0$ will do so independently of n by uniform tightness of $\{\widetilde{\mathrm{Q}}^{(n)}\}_{n\in\mathbb{N}}$ and the construction of $A_\delta^{(n)}$.

Finally, the second term in our original upper bound, $\left|\mathbb{E}_\nu(h_{\delta,\eta}\mathrm{Id}) - \mathbb{E}_{\nu^{(n)}}(h_{\delta,\eta}\mathrm{Id})\right|$ will tend to zero as $n \to \infty$ for fixed $\delta > 0$ and $\eta > 0$ as $h_{\delta,\eta}\mathrm{Id} \in C_b(\mathbb{R})$. η is only introduced for $h_{\delta,\eta}\mathrm{Id}$ to be a continuous compactly supported function and thus can be set to an arbitrary positive number. Because we only need $\delta \to 0$ and $n \to \infty$ for a finite number of terms from above, we can take the respective minimum and maximum over these which will yield some $\delta_0 > 0$ and $\mathrm{N}_0 \in \mathbb{N}$. If we fix $\delta = \delta_0$ and take maximum between N_0 and the minimum N necessary for $\left|\mathbb{E}_\nu(h_{\delta,\eta}\mathrm{Id}) - \mathbb{E}_{\nu^{(n)}}(h_{\delta,\eta}\mathrm{Id})\right|$ to be sufficiently small, the $\left|\mathbb{E}_\nu(\mathrm{Id}) - \mathbb{E}_{\nu^{(n)}}(\mathrm{Id})\right|$ can be made arbitrarily small.

It remains to show that if $\log \frac{q}{p} \notin \mathrm{L}^1(\mathrm{Q})$, the sequence $\{\mathrm{KL}\,(\mathrm{Q}^{(n)} \| \,\mathrm{P}) - s_{\mathrm{K}_S}^{(n)}\}_{n\in\mathbb{N}}$ also diverges. Lemmas 13 to 15 only depend on boundedness of q; therefore we only need,

$$\mathbb{E}\,|\log \tfrac{\widetilde{q}^{(n)}(\widetilde{Z}^{(n)})}{p(Z^{(n)})}| \to \infty .$$

If $|\log \frac{\widetilde{q}^{(n)}(\widetilde{Z}^{(n)})}{p(Z^{(n)})}|$ is not a.s. finite then we are done. In the case when a.s. finiteness holds, it must be true that $\widetilde{q}^{(n)}(\widetilde{Z}^{(n)}) > 0 \iff p(Z^{(n)}) > 0$ a.s. and thus by continuity of the logarithm, absolute value, p, $\widetilde{q}^{(n)}$ (Lemma 14), pointwise convergence $\widetilde{q}^{(n)} \to q$ (Lemma 15), and a.s. convergence of both $Z^{(n)}$ and $\widetilde{Z}^{(n)}$ to Z, we have $|\log \frac{\widetilde{q}^{(n)}(\widetilde{Z}^{(n)})}{p(Z^{(n)})}| \to |\log \frac{q(Z)}{p(Z)}|$ a.s. Hence we can use Fatou's lemma to establish,

$$\infty = \mathbb{E}\big|\log \tfrac{q(Z)}{p(Z)}\big| \leq \liminf_{n\to\infty} \mathbb{E}\big|\log \tfrac{\widetilde{q}^{(n)}(\widetilde{Z}^{(n)})}{p(Z^{(n)})}\big| .$$

which means $\{\mathrm{KL}\,(\mathrm{Q}^{(n)} \| \,\mathrm{P}) - s_{\mathrm{K}_S}^{(n)}\}_{n\in\mathbb{N}}$ diverges. □

Proposition 10. *A collection of random variables $\{f(Z^{(n)})\}_{n\in\mathbb{N}}$, $f \in C(\mathbb{R}^{\mathrm{D}})$, is uniformly integrable if there exists some $r > 0$ s.t. $\forall x \in \mathbb{R}^{\mathrm{D}}$ with $\|x\|_2 > r$, $|f(x)| \leq h_p(x)$ where $h_p \colon \mathbb{R}^{\mathrm{D}} \to \mathbb{R}$, $x \mapsto \sum_{j=1}^p c_j \|x\|_2^j$, for some $c_1, \dots, c_p \in \mathbb{R}$, and $\mathbb{E}\,\|Z\|_2^p < \infty$.*[6]

Proof of Proposition 10. Kallenberg (2006, p. 44, Equation (5)) states that a sequence of integrable random variables $\{\xi_n\}_{n\in\mathbb{N}}$ is uniformly integrable iff,

$$\lim_{k\to\infty} \limsup_{n\to\infty} \mathbb{E}\,\mathbb{I}_{|\xi_n|>k} |\xi_n| = 0 . \tag{14}$$

Let us first ensure that random variables $\{f(Z^{(n)})\}_{n\in\mathbb{N}}$ are integrable. Defining $U \coloneqq \{x \in \mathbb{R}^{\mathrm{D}} \colon \|x\|_2 > r\}$,

$$\mathbb{E}\,\mathbb{I}_U \big|f(Z)\big| \leq \mathbb{E}\,\mathbb{I}_U h_p(Z) ,$$

with $h_p(Z)$ being a linear combination of terms $\|Z^{(n)}\|_2^k$ for $k \in 0, 1, \dots, p$. By Cauchy–Bunyakovsky–Schwarz,

$$\begin{aligned} &\mathbb{E}\,\mathbb{I}_U \|Z^{(n)}\|_2^k \leq \mathbb{E}\,\|Z + \mathcal{E}/\sqrt{n}\|_2^k \\ &\leq 2^{\frac{3k}{2}-1} \left(\mathbb{E}\,\|Z\|_2^k + 2\,\mathbb{E}\,\|Z\|_2^{\frac{k}{2}} \|\tfrac{\mathcal{E}}{\sqrt{n}}\|_2^{\frac{k}{2}} + \mathbb{E}\,\|\tfrac{\mathcal{E}}{\sqrt{n}}\|_2^k \right) . \end{aligned}$$

As $\mathbb{E}\,\|Z\|_2^t < \infty$ for all $t \in [0,p]$ by Hölder's inequality and the assumption $\mathbb{E}\,\|Z\|_2^p < \infty$, the second and third summands will go to 0 as $n \to \infty$, and the first term is finite. Because $\mathbb{E}\,\mathbb{I}_{U^{\mathrm{C}}} |f(Z^{(n)})| \leq \mathrm{C}_f \coloneqq \sup_{U^{\mathrm{C}}} |f|$ which is finite by continuity of $|f|$ and compactness of U^{C} (Heine–Borel), the random variables $\{f(Z^{(n)})\}_{n\in\mathbb{N}}$ are integrable.

By Equation (14), it is sufficient if $\forall \varepsilon > 0, \exists k \in \mathbb{R}$ s.t.,

$$\limsup_{n\to\infty} \mathbb{E}\,\mathbb{I}_{|f(Z^{(n)})|>k} |f(Z^{(n)})| < \varepsilon .$$

[6] Proposition 10 can be straightforwardly extended to polynomials in any *p-norm* $\|x\|_p = (\sum_{i=1}^{\mathrm{D}} x_i^p)^{1/p}$, $p \in [1,\infty)$ by strong equivalence of p-norms on finite Euclidean spaces.

Because any finite collection of integrable random variables is uniformly integrable, we can find $\delta > 0$ s.t. $\forall B \in \mathcal{B}$ with $\mathrm{Q}(B) \leq \delta$, $\mathbb{E}\, \mathbb{I}_B \|Z\|_2^j \leq \varepsilon / (2^{\frac{3j}{2}-1}|c_j|)$ for $j = 1, \ldots, p$. We WLOG assumed $c_j > 0, \forall j$ as otherwise we could just ignore the corresponding terms.

By tightness of Q, for every $\delta > 0$ there exists a compact set $K_{\delta,\alpha}$ s.t. $\mathrm{Q}(K_{\delta,\alpha}) > 1 - \delta$ (the purpose of α will become clear later). Because we are on a finite Euclidean space, $K_{\delta,\alpha}$ is bounded and thus we can WLOG assume $K_{\delta,\alpha} = \bar{B}_{r_\delta - \alpha}(s_\delta)$, a closed ball centred at $s_\delta \in \mathbb{R}^{\mathrm{D}}$ with radius $r_\delta - \alpha$, for some $\alpha > 0$, s.t. $r_\delta - \alpha > r$, i.e. $K_{\delta,\alpha}^{\mathrm{C}} \subset U$. Clearly $K_{\delta,\alpha} \subset K_\delta := \bar{B}_{r_\delta}(s_\delta)$ and thus $\mathrm{Q}(K_\delta) > 1 - \delta$. Define $\kappa = \sup_{K_\delta} |f|$ which is a finite constant by continuity of f and compactness of K_δ. We will now show,

$$\limsup_{n\to\infty} \mathbb{E}\, \mathbb{I}_{|f|>\kappa} |f(Z^{(n)})| < \varepsilon\,.$$

By the assumption $|f| \leq h_p$ on U, we have,

$$\begin{aligned}
&\mathbb{E}\, \mathbb{I}_{|f|>\kappa_\delta} |f(Z^{(n)})| \leq \mathbb{E}\, \mathbb{I}_{K_\delta^{\mathrm{C}}} |f(Z^{(n)})| \\
&\leq \sum_{j=1}^{p} c_j\, \mathbb{E}\, \mathbb{I}_{K_\delta^{\mathrm{C}}} \|Z^{(n)}\|_2^j = \sum_{j=1}^{p} c_j\, \mathbb{E}\, \mathbb{I}_{K_\delta^{\mathrm{C}}} \|Z + \mathcal{E}/\sqrt{n}\|_2^j\,,
\end{aligned}$$

where each of the RHS summands can be upper bounded,

$$2^{\frac{3j}{2}-1} \left(\mathbb{E}\, \mathbb{I}_{K_\delta^{\mathrm{C}}} \|Z\|_2^j + 2\, \mathbb{E}\, \|Z\|_2^{\frac{k}{2}} \|\tfrac{\mathcal{E}}{\sqrt{n}}\|_2^{\frac{j}{2}} + \mathbb{E}\, \|\tfrac{\mathcal{E}}{\sqrt{n}}\|_2^j \right).$$

As before, all but the first term will vanish as $n \to \infty$ and thus we can ignore them in evaluation of the $\limsup$. Ignoring the multiplicative constants for a moment, we turn our attention to the $\mathbb{E}\, \mathbb{I}_{K_\delta^{\mathrm{C}}}(Z^{(n)}) \|Z\|_2^j = \mathbb{E}\, \mathbb{I}_{K_\delta^{\mathrm{C}}}(Z + \mathcal{E}/\sqrt{n}) \|Z\|_2^j$ where the noise term remained inside the indicator random variable by construction of the upper bound.

Define $A_\alpha^{(n)} := \{\mathcal{E} \colon \|\mathcal{E}\|_2 \leq \alpha\sqrt{n}\} \in \mathcal{B}$, $\beta^{(n)} := \mathrm{P}_{\mathcal{E}}(A_\alpha^{(n)})$ and observe $\beta^{(n)} \uparrow 1$. Because $\|Z + \mathcal{E}/\sqrt{n}\|_2 \leq \|Z\|_2 + \|\mathcal{E}/\sqrt{n}\|_2$ by the triangle inequality, and $(Z + \mathcal{E}/\sqrt{n}) \in K_\delta^{\mathrm{C}}$ iff $\|Z + \mathcal{E}/\sqrt{n}\|_2 > r_\delta$ by definition, we have $\mathbb{I}_{A_\alpha^{(n)}}(\mathcal{E})\, \mathbb{I}_{K_\delta^{\mathrm{C}}}(Z + \mathcal{E}/\sqrt{n}) \leq \mathbb{I}_{A_\alpha^{(n)}}(\mathcal{E})\, \mathbb{I}_{K_{\delta,\alpha}^{\mathrm{C}}}(Z)$ for all $n \in \mathbb{N}$. Therefore,

$$\begin{aligned}
&\mathbb{E}[(\mathbb{I}_{A_\alpha^{(n)}}(\mathcal{E}) + \mathbb{I}_{(A_\alpha^{(n)})^{\mathrm{C}}}(\mathcal{E}))\, \mathbb{I}_{K_\delta^{\mathrm{C}}}(Z + \mathcal{E}/\sqrt{n})\, \|Z\|_2^j] \\
&\leq \mathbb{E}[\mathbb{I}_{A_\alpha^{(n)}}(\mathcal{E})\, \mathbb{I}_{K_{\delta,\alpha}^{\mathrm{C}}}(Z)\, \|Z\|_2^j] + \mathbb{E}[\mathbb{I}_{(A_\alpha^{(n)})^{\mathrm{C}}}(\mathcal{E})\, \|Z\|_2^j] \\
&= \beta^{(n)}\, \mathbb{E}[\mathbb{I}_{K_{\delta,\alpha}^{\mathrm{C}}}(Z)\, \|Z\|_2^j] + (1 - \beta^{(n)})\, \mathbb{E}\, \|Z\|_2^j\,.
\end{aligned}$$

Because $\mathbb{E}\, \|Z\|_2^j < \infty$ by Hölder's inequality and $\beta^{(n)} \uparrow 1$, the limit and thus $\limsup$ of the RHS is clearly,

$$\mathbb{E}[\mathbb{I}_{K_{\delta,\alpha}^{\mathrm{C}}}(Z)\, \|Z\|_2^j] < \frac{\varepsilon}{2^{\frac{3j}{2}-1}|c_j|}\,,$$

where the upper bound is by uniform integrability of $\|Z\|_2^j$ and the construction of $K_{\delta,\alpha}$. Substituting back,

$$\limsup_{n\to\infty} \mathbb{E}\, \mathbb{I}_{|f|>\kappa} |f(Z^{(n)})| < \varepsilon\,,$$

which concludes the proof. □

Auxiliary lemmas

Lemma 11. *If* $\log \frac{q}{p} \in \mathrm{L}^1(\mathrm{Q})$ *and* $\{\log p(Z^{(n)})\}_{n\in\mathbb{N}}$ *are uniformly integrable, then* $\log q \in \mathrm{L}^1(\mathrm{Q})$*, and,*

$$\mathop{\mathbb{E}}_{\mathrm{Q}} \left(\log \tfrac{q}{p} \right) = \mathop{\mathbb{E}}_{\mathrm{Q}}(\log q) - \mathop{\mathbb{E}}_{\mathrm{Q}}(\log p)\,.$$

Proof. By uniform integrability of $\{\log p(Z^{(n)})\}_{n\in\mathbb{N}}$ and (Dudley, 2002, Theorem 10.3.6) $\log p \in \mathrm{L}^1(\mathrm{Q})$. By Theorem 4.1.10 in (Dudley, 2002), $\log \frac{q}{p} \in \mathrm{L}^1(\mathrm{Q})$ and $\log p \in \mathrm{L}^1(\mathrm{Q})$ imply $\log q \in \mathrm{L}^1(\mathrm{Q})$, and the equality from above holds by the same theorem. □

Lemma 12. *If* $\{\log p(Z^{(n)})\}$ *is uniformly integrable, then* $\mathbb{E}_{\mathrm{Q}^{(n)}}(\log p) \to \mathbb{E}_{\mathrm{Q}}(\log p)$ *as* $n \to \infty$*.*

Proof of Lemma 12. Notice that $\|Z^{(n)} - Z\|_2 = \|\mathcal{E}/\sqrt{n}\|_2$ by definition, and therefore $Z^{(n)} \to Z$ a.s. By the continuity of p and of the logarithm function, the continuous mapping theorem yields $\log p(Z^{(n)}) \to \log p(Z)$ a.s. Since we have assumed that the collection of random variables $\{\log p(Z^{(n)})\}$ is uniformly integrable and a.s. convergence implies convergence in probability, we can use Theorem 10.3.6 in (Dudley, 2002) to deduce $\mathbb{E}_{\mathrm{Q}^{(n)}}(\log p) \to \mathbb{E}_{\mathrm{Q}}(\log p)$ as $n \to \infty$. □

Lemma 13. *For* S *is a linear manifold and every* $n \in \mathbb{N}$*,* $\mathbb{E}(\log q^{(n)}(Z^{(n)}))$ *is equal to,*

$$\mathbb{E}\left(\log \widetilde{q}^{(n)}(\widetilde{Z}^{(n)}) \right) - \frac{\mathrm{D} - \mathrm{K}_S}{2} \log(2\pi \mathrm{e} n^{-1})\,.$$

Proof of Lemma 13. As stated at the beginning of this section, we can WLOG assume $S = \mathbb{R}^{\mathrm{K}_S} \times \{0\}^{\mathrm{D}-\mathrm{K}_S}$. Then,

$$\begin{aligned}
\log q^{(n)}(x) &= \log \left[\int (2\pi n^{-1})^{-\frac{\mathrm{D}}{2}} \mathrm{e}^{-\frac{\|x-z\|_2^2}{2n^{-1}}} \mathrm{Q}(\mathrm{d}z) \right] \\
&= -\frac{\mathrm{D} - \mathrm{K}_S}{2} \log(2\pi n^{-1}) - \frac{n}{2} \left\| x_{(\mathrm{K}_S+1)\colon \mathrm{D}} \right\|_2^2 \\
&\quad + \log \left[\int \phi_{z_{1\colon \mathrm{K}_S}, n^{-1} I_{\mathrm{K}_S}}(x_{1\colon \mathrm{K}_S}) \mathrm{Q}(\mathrm{d}z) \right],
\end{aligned}$$

$\forall x \in \mathbb{R}^{\mathrm{D}}$. Using the definition $Z^{(n)} = Z + \mathcal{E}/\sqrt{n}$,

$$\begin{aligned}
&\mathbb{E}(\log q^{(n)}(Z^{(n)})) \\
&= \int\!\!\int \phi_{0, I_{\mathrm{D}}}(\varepsilon) \log q^{(n)}(z + \varepsilon/\sqrt{n})\, \lambda^{\mathrm{D}}(\mathrm{d}\varepsilon) \mathrm{Q}(\mathrm{d}z)
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{\mathrm{D}-\mathrm{K}_S}{2}\log(2\pi n^{-1}) - \frac{n}{2}\mathop{\mathbb{E}}_{\widetilde{\mathcal{E}}\sim\mathcal{N}(0,I_{\mathrm{D}-\mathrm{K}_S})}\left\|\widetilde{\mathcal{E}}/\sqrt{n}\right\|_2^2 \\
&\quad + \iint \phi_{0,I_{\mathrm{K}_S}}(\varepsilon)\log\widetilde{q}^{(n)}(z+\varepsilon/\sqrt{n})\,\lambda^{\mathrm{K}_S}(\mathrm{d}\varepsilon)\mathrm{Q}(\mathrm{d}z) \\
&= -\frac{\mathrm{D}-\mathrm{K}_S}{2}\log(2\pi n^{-1}) - \frac{\mathrm{D}-\mathrm{K}_S}{2} \\
&\quad + \iint \phi_{0,I_{\mathrm{K}_S}}(\varepsilon)\log\widetilde{q}^{(n)}(z+\varepsilon/\sqrt{n})\,\lambda^{\mathrm{K}_S}(\mathrm{d}\varepsilon)\mathrm{Q}(\mathrm{d}z) \\
&= -\frac{\mathrm{D}-\mathrm{K}_S}{2}\log(2\pi\mathrm{e}n^{-1}) + \mathbb{E}\left(\log\widetilde{q}^{(n)}(\widetilde{Z}^{(n)})\right),
\end{aligned}
$$

where the first equality is by the Tonelli–Fubini's theorem, and we used standard properties of the Gaussian distribution. □

Lemma 14. *For S is a linear manifold and every $n \in \mathbb{N}$, $q^{(n)}$ and $\widetilde{q}^{(n)}$ are both bounded by the constant C_q and continuous for $\mathcal{T}$ and $\mathcal{T}_S$ respectively.*

Proof of Lemma 14. Boundedness is a simple consequence of Equations (11) and (12) and the Hölder's inequality,

$$
\begin{aligned}
q^{(n)}(x) &= \left\|\phi_{x,n^{-1}I_{\mathrm{D}}}\, q\right\|_{\mathrm{L}^1(m_S)} \\
&\le \left\|\phi_{x,n^{-1}I_{\mathrm{D}}}\right\|_{\mathrm{L}^1(m_S)}\|q\|_{\mathrm{L}^\infty(m_S)} = \mathrm{C}_q\,;
\end{aligned}
$$

similarly for $\widetilde{q}^{(n)}$.

The proofs of continuity are analogical, therefore we will only discuss the one for q. Notice that for any $x, y \in \mathbb{R}^{\mathrm{D}}$,

$$
\left|q^{(n)}(x) - q^{(n)}(y)\right| \propto \left|\int f_z(x) - f_z(y)\mathrm{Q}(\mathrm{d}z)\right|,
$$

with $f_z(x) \coloneqq \exp(-\frac{n}{2}\|x-z\|_2^2)$.

We can upper bound,

$$
\left|\int f_z(x) - f_z(y)\mathrm{Q}(\mathrm{d}z)\right| \le \int\left|f_z(x) - f_z(y)\right|\mathrm{Q}(\mathrm{d}z),
$$

which suggests it would be sufficient to show that the collection of functions $\{f_z\}_{z\in\mathbb{R}^{\mathrm{D}}}$ is uniformly equicontinuous. A sufficient condition for uniform equicontinuity is $\{f_z\}_{z\in\mathbb{R}^{\mathrm{D}}} \subset \mathrm{Lip}(\mathbb{R}^{\mathrm{D}}, \mathrm{L})$ where $\mathrm{Lip}(\mathbb{R}^{\mathrm{D}}, \mathrm{L})$ is the set of real-valued Lipschitz continuous functions on $\mathbb{R}^{\mathrm{D}}$ with Lipschitz constant L. Because each f_z is smooth, we can use Taylor expansion to equate,

$$
f_z(x) = f_z(y) + (x-y)^{\mathrm{T}} f_z'(\xi)
$$

with $f_z'\colon \mathbb{R}^{\mathrm{D}} \to \mathbb{R}^{\mathrm{D}}$ the derivative of f_z, for some $\xi \in \mathbb{R}^{\mathrm{D}}$. Using the Cauchy–Bunyakovsky–Schwarz inequality,

$$
\left|f_z(x) - f_z(y)\right| \le \|x-y\|_2\left\|f_z'(\xi)\right\|_2,
$$

which means it is sufficient to show $\left\|f_z'(\xi)\right\|_2$ is uniformly bounded in $(z,\xi) \in \mathbb{R}^{\mathrm{D}} \times \mathbb{R}^{\mathrm{D}}$ to establish $\{f_z\}_{z\in\mathbb{R}^{\mathrm{D}}} \subset \mathrm{Lip}(\mathbb{R}^{\mathrm{D}}, \mathrm{L})$. Simple algebra shows that,

$$
\left\|f_z'(\xi)\right\|_2 = n f_z(\xi)\|\xi - z\|_2 \le \sqrt{\tfrac{n}{\mathrm{e}}},
$$

$\forall(z,\xi) \in \mathbb{R}^{\mathrm{D}} \times \mathbb{R}^{\mathrm{D}}$, with equality when $\|\xi - z\|_2 = n^{-\frac{1}{2}}$. Hence we can see that $\{f_z\}_{z\in\mathbb{R}^{\mathrm{D}}} \subset \mathrm{Lip}(\mathbb{R}^{\mathrm{D}}, \mathrm{L})$ for $\mathrm{L} = \sqrt{\frac{n}{\mathrm{e}}}$, and thus the family of functions $\{f_z\}_{z\in\mathbb{R}^{\mathrm{D}}}$ is uniformly equicontinuous.

Therefore, $\forall\varepsilon > 0, \exists\delta > 0$ s.t. $\|x-y\|_2 < \delta \implies |f_z(x) - f_z(y)| < \varepsilon$ for all $z \in \mathbb{R}^{\mathrm{D}}$. Substituting back,

$$
\left|q^{(n)}(x) - q^{(n)}(y)\right| < \left(\tfrac{n}{2\pi}\right)^{\frac{\mathrm{D}}{2}}\varepsilon,
$$

whenever $\|x-y\|_2 < \delta$, and thus $q^{(n)}$ is continuous. □

Lemma 15. *For S is a linear manifold, $\widetilde{q}^{(n)}$ converges pointwise to q as $n \to \infty$.*

Proof of Lemma 15. WLOG assume $S = \mathbb{R}^{\mathrm{K}_S} \times \{0\}^{\mathrm{D}-\mathrm{K}_S}$. For arbitrary $x \in \mathbb{R}^{\mathrm{K}_S}$,

$$
\widetilde{q}^{(n)}(x) = \int q(x - \xi/\sqrt{n})\,\mathcal{N}(0, I_{\mathrm{K}_S})(\mathrm{d}\xi),
$$

where we implicitly pad x and ξ by zeros as $q\colon S \to \mathbb{R}$. Because q is continuous by assumption, for every $\varepsilon > 0$, $\exists\delta > 0$ s.t. $\|(x - \xi/\sqrt{n}) - x\|_2 = \|\xi\|_2 < \delta \implies |q(x - \xi/\sqrt{n}) - q(x)| < \varepsilon$. For any $\alpha > 0$, we can use Chebyshev's inequality to determine $\mathrm{N} \in \mathbb{N}$ s.t. $\forall n \ge \mathrm{N}$, $\mathbb{P}(\|\xi/\sqrt{n}\|_2 \ge \delta) \le \alpha$. Define $B \subset \mathbb{R}^{\mathrm{K}_S}$ to be the ball centred at zero with radius δ. Then we can upper bound,

$$
\begin{aligned}
&\left|\widetilde{q}^{(n)}(x) - q(x)\right| \\
&\le \int\left|q(x - \xi/\sqrt{n}) - q(x)\right|\mathcal{N}(0, I_{\mathrm{K}_S})(\mathrm{d}\xi) \\
&< \varepsilon + \int_{B^{\mathrm{C}}}\left|q(x - \xi/\sqrt{n}) - q(x)\right|\mathcal{N}(0, I_{\mathrm{K}_S})(\mathrm{d}\xi) \\
&\le \varepsilon + 2\mathrm{C}_q\alpha,
\end{aligned}
$$

i.e. $\widetilde{q}^{(n)} \to q$ as $n \to \infty$ pointwise. □

Lemma 16. *Assume $w_1, \ldots, w_k \in \mathbb{R}$ are arbitrary constants, and ε_i, $i = 1, \ldots, k$, are i.i.d. standard normal variables. Define the vector $w = (w_i)_{i=1}^k$. Then for $p \ge 0$,*

$$
\mathbb{E}\left|\sum_{i=1}^k w_i\varepsilon_i\right|^p = \|w\|_2^p\frac{2^{\frac{p}{2}}\Gamma(\frac{p+1}{2})}{\Gamma(\frac{1}{2})}.
$$

Proof. Use the linearity of the dot product and Gaussianity of ε_i's to obtain,

$$
\mathbb{E}\left|\sum_{i=1}^k w_i\varepsilon_i\right|^p = \mathbb{E}\left|\|w\|_2\tilde{\varepsilon}\right|^p = \|w\|_2^p\;\mathbb{E}\;|\tilde{\varepsilon}|^p,
$$

where $\tilde{\varepsilon}$ is a standard normal random variable. The result is then obtained by realising that powers of standard normal are distributed according to Generalised Gamma variable for which the expectation is known. $\square$

B.2. Discretisation approach

We define the notion of a *discretiser*, a measurable function $k\colon \mathbb{R}^{\mathrm{D}} \to A$ where A is a finite set the members of which will be called *cells*. We will consider discretisers that divide each axis of $\mathbb{R}^{\mathrm{D}}$ into two half-intervals in the tails and many equal sized intervals in the middle; the size of these will be denoted by Δ. Thus if k divides a single axis into M cells, the total number of cells in $\mathbb{R}^{\mathrm{D}}$ will be M^{D}. We will consider sequences of discretisers $(k_n)_{n\in\mathbb{N}}$ where each k_n produces discretisation which is a refinement of the previous one, i.e. it only divides existing cells into smaller ones.

We say that a sequence of discretisers is *asymptotically exact* if for every $x \in \mathbb{R}^{\mathrm{D}}$ we have,

$$\bigcap_{n\in\mathbb{N}} \bigcap_{a\in A^{(n)}\colon k_n(x)=a} k_n^{-1}(a) = \{x\}\,,$$

i.e. any two distinct points will end up in different cells eventually. With a slight abuse of notation, we abbreviate this as $\lim_{n\to\infty} k_n(x) = \{x\}$.

We further define a function $x_n\colon A^{(n)} \to \mathbb{R}^{\mathrm{D}}$ which accepts a cell and returns an element that maps to that particular cell; the particular algorithm of picking a representative of the cell is not important, but at least one such algorithm must exist by the axiom of choice.

Finally, we denote the *quantised densities* w.r.t. the counting measure for P and Q respectively by $p^{(n)}(a) = \mathrm{P}(k_n^{-1}(a))$ and $q^{(n)}(a) = \mathrm{Q}(k_n^{-1}(a))$.

Proposition 17. *Consider an asymptotically exact sequence of discretisers $(k_n)_{n\in\mathbb{N}}$, the corresponding sequence of finite spaces $(A^{(n)})_{n\in\mathbb{N}}$, and discretisation intervals $(\Delta_n)_{n\in\mathbb{N}}$. Let the assumptions stated above and in Theorem 4 hold.*

Then,

$$\lim_{n\to\infty} \left\{ \mathrm{KL}\left(\mathrm{Q}^{(n)} \,\|\, \mathrm{P}^{(n)}\right) - s^{(n)} \right\} = \mathop{\mathbb{E}}_{\mathrm{Q}}\left(\log \tfrac{q}{p}\right),$$

with $s^{(n)} = -(\mathrm{D} - \mathrm{K}_S)\log(\Delta_n)$.

Proof of Proposition 17. By assumption, $\operatorname{diam}(S) < \infty$ and thus we can find a compact set $K \subset \mathbb{R}^{\mathrm{D}}$ s.t. $S \subset K$. WLOG define $R_+ \supset K$ to be the smallest hyper-rectangle of strictly positive Lebesgue measure s.t. it can be padded out by hypercubes with side Δ_1 (by extending the lengths of sides of R to be positive multiples of Δ_1; by the assumption that each k_n refines existing cells, and that the cells are equal sized, $k_n(R_+)$ will only produce equal sized cells for all $n \in \mathbb{N}$); R_+ exists by the Heine–Borel theorem.

The n^{th} discretised KL is defined as,

$$\mathrm{KL}\left(\mathrm{Q}^{(n)} \,\|\, \mathrm{P}^{(n)}\right) = \sum_{a\in A^{(n)}} q^{(n)}(a) \log \frac{q^{(n)}(a)}{p^{(n)}(a)}\,.$$

From now on, we will drop the input to the individual quantised densities unless confusion may arise.

We start with the case $\log \frac{q}{p} \in \mathrm{L}^1(\mathrm{Q})$. Because we have assumed that $\log p \in \mathrm{L}^1(\mathrm{Q})$ if $\log \frac{q}{p} \in \mathrm{L}^1(\mathrm{Q})$,

$$\mathop{\mathbb{E}}_{\mathrm{Q}}(\log \tfrac{q}{p}) = \mathop{\mathbb{E}}_{\mathrm{Q}}(\log q) - \mathop{\mathbb{E}}_{\mathrm{Q}}(\log p)\,,$$

by Theorem 4.1.10 in (Dudley, 2002), and thus we can focus on the negative entropy and cross-entropy terms separately.

Starting with the negative entropy term, notice that for any $x \in S$, we have $q^{(n)}(k_n(x)) \to \mathrm{Q}(\{x\})$, as for any $x' \in S \setminus \{x\}$, $\mathrm{Q}(\{x'\}) > 0$ and there exists $\mathrm{N} \in \mathbb{N}$ s.t. $\left\|x - x'\right\|_2 > \sqrt{\mathrm{D}}\Delta_n$ (the maximum distance of points in a single cell) for all $n \geq \mathrm{N}$. Thus $q^{(n)}(k_n(x)) \downarrow \mathrm{Q}(\{x\})$ by being a monotonically decreasing sequence with the least upper bound equal exactly to $\mathrm{Q}(\{x\})$. Note that by assumption $\mathrm{Q}(\{x\}) = q(x)$ where q is the density of Q w.r.t. the counting measure on S, and thus $q^{(n)}(k_n(x)) \downarrow q(x)$.

The following insight will help us:

$$\sum_{a\in A^{(n)}} q^{(n)}(a) h(a) = \int q(x) h(k_n(x)) m_S(\mathrm{d}x)\,, \tag{15}$$

for any $h\colon A^{(n)} \to \mathbb{R}$; note that the definition of $A^{(n)}$ makes $h(k_n(x))$ a simple function and thus measurable which means the RHS is well-defined. We can thus use continuity and monotonicity of the logarithm to establish $\log q^{(n)}(k_n(x)) \downarrow \log q(x)$ pointwise and the fact that $\log q^{(n)}(k_n(x)) \leq 0$ as $q^{(n)}(k_n(x)) \leq 1, \forall x$, and apply the monotone convergence theorem to establish,

$$\sum_{A^{(n)}} q^{(n)} \log q^{(n)} \downarrow \int q \log q \,\mathrm{d}m_S\,.$$

We now turn to the cross-entropy term. Because R_+ is compact, we can define,

$$\alpha_n := \max_{a\in k_n(R_+)} \left| \sup[\log p(k_n^{-1}(a))] - \inf[\log p(k_n^{-1}(a))] \right|\,,$$

and observe $\alpha_n \downarrow 0$ as $n \to \infty$ because $\log p$ is continuous, and thus uniformly continuous on R_+. Notice,

$$\begin{aligned} &\left| \sum_{a\in A^{(n)}} q^{(n)}(a) (\log[p^{(n)}(a)] - \log[p(x_n(a))\Delta_n^{\mathrm{D}}]) \right| \\ &\leq \sum_{a\in A^{(n)}} q^{(n)}(a) \left| \log[p^{(n)}(a)] - \log[p(x_n(a))\Delta_n^{\mathrm{D}}] \right| \end{aligned}$$

$$\leq \sum_{a \in A^{(n)}} q^{(n)}(a) \alpha_n \leq \alpha_n \,,$$

using that $q^{(n)} = 0$ outside of $k_n(R_+)$. Because $\alpha_n \downarrow 0$ as $n \to \infty$, we can approximate $\log[p^{(n)}(a)\Delta_n^{\mathrm{D}}]$ by $\log p(x_n(a)) + \mathrm{D} \log \Delta_n$.

Since $\lim_{n\to\infty} k_n(x) = \{x\}$ by assumption, we have $x_n(k_n(x)) \to x$ pointwise by $\left\|x - x'\right\|_2 \leq \sqrt{\mathrm{D}}\Delta_n$ for any x' s.t. $k_n(x) = k_n(x')$. By continuity of the logarithm, $\log p(x_n(k_n(x))) \to \log p(x)$ pointwise (i.e. $\log p(x_n(a))$ can be substituted for the function $h(a)$ in Equation (15)). Because R_+ is compact, we can define $\kappa := \sup_{R_+} |\log p|$ which will be finite by the continuity of $\log p$. Hence $|\log p(x_n(k_n(x)))| \leq \kappa$, and we can apply the dominated convergence theorem:

$$\sum_{a \in A^{(n)}} q^{(n)}(a) \log p(x_n(a)) \to \int q \log p \, \mathrm{d} m_S \,.$$

Putting the results in previous paragraphs together, we arrive at the following limit,

$$\sum_{A^{(n)}} q^{(n)} \log \frac{q^{(n)}}{p^{(n)}} + \mathrm{D} \log \Delta_n \to \int q \log \frac{q}{p} \, \mathrm{d} m_S \,,$$

where we are implicitly using the previously derived equality $\mathbb{E}_{\mathrm{Q}} \log \frac{q}{p} = \mathbb{E}_{\mathrm{Q}}(\log \mathrm{Q}) - \mathbb{E}_{\mathrm{Q}}(\log p)$.

To finish the proof, we must prove that the sequence $\{\mathrm{KL}\,(\mathrm{Q}^{(n)} \| \, \mathrm{P}^{(n)}) - s^{(n)}\}$ diverges if $\log \frac{q}{p} \notin \mathrm{L}^1(\mathrm{Q})$. By our above derivations, this is equivalent to proving that,

$$\int q(x) \left| \log \frac{q^{(n)}(k_n(x))}{p(x_n(k_n(x)))} \right| m_S(\mathrm{d}x) \to \infty \,.$$

If $p(x_n(k_n(x))) = 0$ and $q^{(n)}(k_n(x)) > 0$ for at least one $x \in \mathbb{Q}^{\mathrm{D}}$ for each $n \in \mathbb{N}$ then we are done. If this is not the case, notice that the results $q^{(n)}(k_n(x)) \to q(x)$ and $p(x_n(k_n(x))) \to p(x)$, both pointwise, are independent of integrability of $\log \frac{q}{p}$. By continuity of the logarithm and the absolute value function, and the assumption that $p(x_n(k_n(x))) = 0 \implies q^{(n)}(k_n(x)) = 0$ for all $x \in \mathbb{Q}^{\mathrm{D}}$ for each $n \in \mathbb{N}$,

$$q(x) \left| \log \frac{q^{(n)}(k_n(x))}{p(x_n(k_n(x)))} \right| \to q(x) \left| \log \frac{q(x)}{p(x)} \right| \,,$$

pointwise on $\mathbb{Q}^{\mathrm{D}}$. Therefore we can use Fatou's lemma to prove that also in this case,

$$\begin{aligned} \infty &= \int q(x) \left| \log \frac{q(x)}{p(x)} \right| m_S(\mathrm{d}x) \\ &\leq \liminf_{n\to\infty} \int q(x) \left| \log \frac{q^{(n)}(k_n(x))}{p(x_n(k_n(x)))} \right| m_S(\mathrm{d}x) \,, \end{aligned}$$

which concludes the proof. □

C. Proofs for Section 5

Proof of Proposition 6. For fixed $\boldsymbol{A}$, the q has support over the subspace $S = \{\boldsymbol{x} \in \mathbb{R}^{\mathrm{D}} \,|\, \boldsymbol{x} = \boldsymbol{A}\boldsymbol{z}, \boldsymbol{z} \in \mathbb{R}^{\mathrm{K}}\}$. If $\boldsymbol{z} \sim \mathcal{N}_{\mathrm{K}}(0, \boldsymbol{V})$, then $\boldsymbol{A}\boldsymbol{z} \sim \mathcal{N}_{\mathrm{D}}(0, \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^{\mathrm{T}})$. Hence we can perform change of coordinates so that QKL reduces to,

$$\int_S \phi_{0,\boldsymbol{V}}(\boldsymbol{z}) \log \frac{\phi_{0,\boldsymbol{V}}(\boldsymbol{z})}{\phi_{0,\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{A}}(\boldsymbol{z})} \lambda^{\mathrm{K}}(\mathrm{d}\boldsymbol{z})$$

where we have used the identity $(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{A})^{-1}\boldsymbol{z} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{z}$ for any $\boldsymbol{z} \in \mathbb{R}^{\mathrm{K}}$. The first term equals $-1/2 \log|\boldsymbol{V}| = -1/2 \sum_{k=1}^{\mathrm{K}} \log \boldsymbol{V}_{kk}$ up to an additive constant, and the second to $\mathrm{Tr}\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{V}\right)$ up to another additive constant. For a constant $\mathrm{C} \in \mathbb{R}$, the integral equals,

$$\mathrm{C} - \frac{1}{2}\sum_{k=1}^{\mathrm{K}} \log \boldsymbol{V}_{kk} + \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{V}\right) .$$

The second term can be rewritten as,

$$\mathrm{Tr}\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{V}\right) = \sum_{k=1}^{\mathrm{K}} \boldsymbol{V}_{kk} \boldsymbol{a}_k^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{a}_k \,,$$

where $\boldsymbol{a}_k$ is the k^{th} column of the $\boldsymbol{A}$ matrix. Because this is an additive loss term in the above QKL, and $\boldsymbol{V}_{kk} > 0$ by the construction of S, it is minimised when the $\boldsymbol{a}_k$ vectors are aligned with the top K eigenvectors of $\boldsymbol{\Sigma}$ because then $\boldsymbol{a}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_k = 1/\lambda_k$ which will be lowest for the highest eigenvalues λ_k of $\boldsymbol{\Sigma}$. Differentiating the objective w.r.t. $\boldsymbol{V}_{kk}$ after substituting the optimal $\boldsymbol{A}$ yields,

$$-\frac{1}{2}\frac{1}{\boldsymbol{V}_{kk}} + \frac{1}{2}\frac{1}{\lambda_k} \,.$$

Setting to zero, we see that $\boldsymbol{V}_{kk} = \lambda_k$, i.e. matching the eigenvalues of $\boldsymbol{\Sigma}$ is the optimal solution. □

Proof of Proposition 6. The n^{th} KL is up to an additive constant equal to,

$$\mathcal{L} := \mathrm{Tr}\left((\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^{\mathrm{T}} + \tau^{(n)}\boldsymbol{I})\Sigma^{-1}\right) - \log\left|\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^{\mathrm{T}} + \tau^{(n)}\boldsymbol{I}\right| .$$

Using some matrix calculus identities from (Petersen et al., 2008), the derivatives w.r.t. the individual parameters are,

$$\begin{aligned} \nabla_{\boldsymbol{A}}\mathcal{L} &= \Sigma^{-1}\boldsymbol{A} - (\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^{\mathrm{T}} + \tau^{(n)}\boldsymbol{I})^{-1}\boldsymbol{A} \,, \\ \nabla_{\mathrm{diag}(\boldsymbol{V})}\mathcal{L} &= \mathrm{diag}[\boldsymbol{A}^{\mathrm{T}}(\Sigma^{-1} - (\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^{\mathrm{T}} + \tau^{(n)}\boldsymbol{I})^{-1})\boldsymbol{A}] \,. \end{aligned}$$

Defining a new diagonal matrix $\widehat{\boldsymbol{V}}_{kk}^{(n)} = \boldsymbol{V}_{kk} + \tau^{(n)}$, and using the orthogonality of $\boldsymbol{A}$'s columns, we have,

$$\begin{aligned} \nabla_{\boldsymbol{A}}\mathcal{L} &= \Sigma^{-1}\boldsymbol{A} - \boldsymbol{A}(\widehat{\boldsymbol{V}}^{(n)})^{-1} \,, \\ \nabla_{\mathrm{diag}(\boldsymbol{V})}\mathcal{L} &= \mathrm{diag}[\boldsymbol{A}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{A} - (\widehat{\boldsymbol{V}}^{(n)})^{-1}] \,. \end{aligned}$$

Setting the first formula above to zero leads to an eigenvector problem, hence we know that the columns of $\boldsymbol{A}$ must be eigenvectors of $\boldsymbol{\Sigma}$. Setting the second formula to zero yields,

$$\boldsymbol{V}_{kk} = (\boldsymbol{a}_k^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{a}_k)^{-1} - \tau^{(n)} .$$

which after substitution of $\boldsymbol{a}_k$ by an eigenvector leads to $\boldsymbol{V}_{kk} = \lambda_k - \tau^{(n)}$ where λ_k is the eigenvalue for the k^{th} substituted eigenvector. By substituting into $\mathcal{L}$,

$$\mathrm{C} + \sum_{k=1}^{\mathrm{K}} \frac{\lambda_k}{\lambda_k} - \log(\lambda_k - \tau^{(n)}) \, ,$$

where C is a constant, we see that to the objective is minimised when the eigenvectors corresponding to the highest eigenvalues are selected. Hence the solution for $\boldsymbol{A}$ is the same as for PCA for all $n \in \mathbb{N}$, and $|\lambda_k - (\lambda_k - \tau^{(n)})| \to 0$ as $n \to \infty$. The optimal solution thus converges to the PCA/QKL in Frobenius/Euclidean distance. $\square$

References

Dattoli, G. and Srivastava, H. A Note on Harmonic Numbers, Umbral Calculus and Generating Functions. *Applied Mathematics Letters*, 21(7):686–693, 2008.

Dudley, R. M. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2002. ISBN 9780521007542.

Gosper, R. W. Harmonic Summation and Exponential GFS, 1996.

Gowers, T. Differentiating Power Series. https://gowers.wordpress.com/2014/02/22/differentiating-power-series/, February 2014.

Gray, R. M. *Entropy and Information Theory*. Springer Science & Business Media, 2011.

Harris, F. E. Tables of the Exponential Integral Ei(x). *Mathematical Tables and Other Aids to Computation*, 11(57): 9–16, 1957.

Kallenberg, O. *Foundations of Modern Probability*. Springer Science & Business Media, 2006.

Petersen, K. B. et al. The matrix cookbook. 2008.

Wolfram—Alpha. https://goo.gl/sZoiuC, November 2017a.

Wolfram—Alpha. https://goo.gl/A5bxLh, November 2017b.