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,

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

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 it cannot be a proper probability distribution. □

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,

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

where $\mathrm{c}_1 := -\frac{1}{2} \log(2\pi \mathrm{e} \sigma^2) - \log \mathrm{C}$, $\mathrm{c}_2 := \mathrm{c}_1 + \frac{1}{2} \log(\sigma^2)$, and we used the fact that $\chi^2(\lambda, \nu)$ is equivalent to a Poisson mixture of centralised chi-squared distributions. 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,

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

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 $f_n \in \mathrm{L}^1, \forall n \in \mathbb{N}$. 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,

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

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,

$$
\underset{\mathrm{Q}(w)}{\mathbb{E}}[\log q(w)] - \underset{\mathrm{Q}(w)}{\mathbb{E}}[\log p(w)]
$$

[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. Copyright 2018 by the author(s).

$$= \mathrm{c}_2 + \frac{1}{2}\left(\log 2 + \mathrm{e}^{-\frac{\mu^2}{2\sigma^2}} \sum_{k=0}^{\infty} \frac{(\frac{\mu^2}{2\sigma^2})^k}{k!} \psi(1/2+k) \right)$$

$$= \mathrm{c}_3 - \frac{1}{2} \left. \frac{\partial \mathrm{M}(a; 1/2; -\mu^2/(2\sigma^2))}{\partial a} \right|_{a=0} ,$$

where $\mathrm{M}(a;b;z)$ denotes the Kummer's function of the first kind, and $\mathrm{c}_3 := \mathrm{c}_2 - \frac{3}{2}\log 2 - \frac{1}{2}\gamma$. It is easy to check that Equation (3) holds for all $u = 0$ assuming $0^0 = 1$.

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 $u = 0$ is obtained 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 $[0, \infty)$. □

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 $[0, \infty)$. □

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^{d}, $\mathrm{d} \in \mathbb{N}$, be the d-dimensional Lebesgue measure.[1] P, Q will be probability measures, P with continuous density p w.r.t. the Lebesgue measure on $\mathbb{R}^{\mathrm{D}}$, 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 dimension of S, i.e. zero in

[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 use to refer to any measure defined on the Borel σ-algebra.

the countable, and $\dim(S)$ in the linear manifold case (dim being the Hamel dimension). The restriction $\mathrm{Q}|_S$ of Q to $(S, \mathcal{B}_S)$, $\mathcal{B}_S$ the trace σ-algebra, will be denoted by $\widetilde{\mathrm{Q}}$.

Assume $\widetilde{\mathrm{Q}}$ has a density q w.r.t. the counting measure on $\mathbb{Q}^{\mathrm{D}}$ if S is at most countable,[2] or w.r.t. Lebesgue measure on S in the linear manifold case. In the (at most) countable case, further assume that $\operatorname{diam}(S) < \infty$ if S is infinite (trivially true if S is finite). If S is a linear manifold, 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. We will also assume that $\log q \in \mathrm{L}^1(\widetilde{\mathrm{Q}})$. Finally, the axiom of choice is assumed throughout.

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

Proof of Theorem 4. Combine Propositions 8 and 21. □

Proof of Theorem 5. Use Proposition 9. □

Notice that the statements of Propositions 8, 9 and 21 differ slightly from those of Theorems 4 and 5 by denoting the limit as $\mathbb{E}_{\widetilde{\mathrm{Q}}} \log \frac{q}{p|_S}$ instead of $\mathbb{E}_{\mathrm{Q}} \log \frac{q}{p}$. The former is more precise in the sense that q is the density of $\widetilde{\mathrm{Q}}$ w.r.t. m_S on $(S, \mathcal{B}_S)$, and thus is not measurable w.r.t. Q, making the integral ill-defined. After swapping Q for $\widetilde{\mathrm{Q}}$, the interchange of p for $p|_S$ is necessary for similar reasons. We omitted this detail from the main text so as to meet the page limit, and to lighten the technicality of the discussion.

B.1. Convolutional approach

Before approaching the proof of Propositions 8 and 9, we note that Lemma 11 allows us to assume that $S = \mathbb{R}^{\mathrm{K}_S} \times \{0\}^{\mathrm{D}-\mathrm{K}_S}$ if S is a linear manifold w.l.o.g.

The following definitions will be useful: let Z and $\mathcal{E}$ be independent random variables respectively distributed according to the distributions $\mathrm{P}_{\mathcal{E}} := \mathcal{N}(0, I_{\mathrm{D}})$ and Q. Define the shorthands $\mathcal{E}^{(n)} := \mathcal{E}/\sqrt{n}$ and $Z^{(n)} := Z + \mathcal{E}^{(n)}$. We further define the random variables $\widetilde{\mathcal{E}} := \mathcal{E}^{(n)}_{1:\,\mathrm{K}_S} \times \{0\}^{\mathrm{D}-\mathrm{K}_S}$, where the subscript denotes the first K_S elements of the vector ($\widetilde{\mathcal{E}}^{(n)} = 0$ if $\mathrm{K}_S = 0$), $\widetilde{\mathcal{E}}^{(n)} := \widetilde{\mathcal{E}}/\sqrt{n}$, and $\widetilde{Z}^{(n)} := Z + \widetilde{\mathcal{E}}^{(n)}$. The corresponding distributions will be denoted as follows: $\mathrm{Q}^{(n)} = \mathrm{Law}(Z^{(n)})$, $\widetilde{\mathrm{Q}}^{(n)} := \mathrm{Law}(\widetilde{Z}^{(n)})$, $\mathrm{P}^{(n)}_{\mathcal{E}} := \mathrm{Law}(\mathcal{E}^{(n)})$, $\mathrm{P}_{\widetilde{\mathcal{E}}} := \mathrm{Law}(\widetilde{\mathcal{E}})$, and $\mathrm{P}^{(n)}_{\widetilde{\mathcal{E}}} := \mathrm{Law}(\widetilde{\mathcal{E}}^{(n)})$.

Notice that $(Z, Z^{(n)}, \widetilde{Z}^{(n)})$ and $(\mathcal{E}^{(n)}, \widetilde{\mathcal{E}}^{(n)})$ are deterministically coupled collections of random variables. 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.

The convolution of two Borel measures μ, ν on $\mathbb{R}^d$, $d \in \mathbb{N}$, will be denoted by $\mu \star \nu$ where for any measurable set B, $(\mu \star \nu)(B) = \int \mu(B - x)\nu(\mathrm{d}x)$. Observe $\mathrm{Q}^{(n)} = \mathrm{Q} \star \mathcal{N}_{\mathbb{R}^{\mathrm{D}}}(0, n^{-1}I)$, and $\widetilde{\mathrm{Q}}^{(n)} = \widetilde{\mathrm{Q}} \star \mathcal{N}_S(0, n^{-1}I)$ with $\mathcal{N}_S(0, n^{-1}I) = \mathrm{P}^{(n)}_{\widetilde{\mathcal{E}}}$ being the Gaussian probability measure on $(S, \mathcal{B}_S)$ (assuming $\mathcal{N}_S(\mu, \Sigma) = \delta_\mu$, the Dirac's delta distribution, if S at most countable). As a corollary of (Dudley, 2002, Proposition 9.1.6), we have,

$$q^{(n)}(x) = \int \phi^{\lambda^{\mathrm{D}}}_{x,n^{-1}I}\, q \,\mathrm{d}m_S \qquad , x \in \mathbb{R}^{\mathrm{D}}\,, \tag{10}$$

where $\phi^{\lambda^{\mathrm{D}}}_{\mu,\Sigma}$ is the density function w.r.t. λ^{D} of $\mathcal{N}(\mu, \Sigma)$ (we will omit the superscript unless confusion may arise). By an analogous argument, we obtain,

$$\widetilde{q}^{(n)}(x) = \int \phi^{m_S}_{x,n^{-1}I}\, q \,\mathrm{d}m_S \qquad , x \in S\,, \tag{11}$$

where $\phi^{m_S}_{\mu,\Sigma}(z) = \delta_{\mathrm{Kr}}\left(z - \mu\right)$ if S is at most countable (δ_{Kr} is the Kronecker's delta function), as m_S is the counting measure and $\mathcal{N}_S(\mu, \Sigma) = \delta_\mu$ (see above), and the usual density function of degenerate Gaussian if m_S is the Lebesgue measure on S. Notice that it would have been more precise to replace $\phi^{\lambda^{\mathrm{D}}}_{x,n^{-1}I}$ in Equation (10) with $\phi^{\lambda^{\mathrm{D}}}_{x,n^{-1}I}|_S$ (c.f. Lemma 22); we omit the restriction in situations where its necessity is clear from the context.

Proposition 8. *Let S be at most countable and all the relevant aforementioned assumptions hold. We consider two cases: $\log p \in \mathrm{L}^1(\mathrm{Q})$ and $\log p \notin \mathrm{L}^1(\mathrm{Q})$. If $\log p \in \mathrm{L}^1(\mathrm{Q})$, assume that the random variables $\{\log p(Z^{(n)})\}_{n \in \mathbb{N}}$ are uniformly integrable.*[3]

Then,

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

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

Proof of Proposition 8. First, assume that $\log p \in \mathrm{L}^1(\mathrm{Q})$. Because $\log q \in \mathrm{L}^1(\widetilde{\mathrm{Q}})$ by assumption, we have $\log \frac{q}{p|_S} \in$

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

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

$\mathrm{L}^1(\widetilde{\mathrm{Q}})$ by Lemma 22 and (Dudley, 2002, Theorem 4.1.10). We can thus focus on convergence of the cross-entropy and negative entropy terms individually. By Lemma 12, the cross-entropy term converges. The negative entropy term converges by Lemma 13.

It remains to investigate the case $\log p \notin \mathrm{L}^1(\mathrm{Q})$. Because Lemma 13 still holds, we can invoke Lemma 20 which establishes that both the sequence $(\mathrm{KL}\left(\mathrm{Q}^{(n)} \,\|\, \mathrm{P}\right) - s^{(n)})$ and the integral $\mathbb{E}_{\widetilde{\mathrm{Q}}} \log \frac{q}{p|_S}$ do not converge as desired. □

Proposition 9. *Let S be a linear manifold and all the relevant aforementioned assumptions hold. We consider two cases: $\log p \in \mathrm{L}^1(\mathrm{Q})$ and $\log p \notin \mathrm{L}^1(\mathrm{Q})$. If $\log p \in \mathrm{L}^1(\mathrm{Q})$, assume that the random variables $\{\log p(Z^{(n)})\}_{n\in\mathbb{N}}$ are uniformly integrable,*[4] *and that $\mathbb{E}\|Z\|_2^2 < \infty$.*

Then,

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

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

Proof of Proposition 9. First, assume that $\log p \in \mathrm{L}^1(\mathrm{Q})$. Because $\log q \in \mathrm{L}^1(\widetilde{\mathrm{Q}})$ by assumption, we have $\log \frac{q}{p|_S} \in \mathrm{L}^1(\widetilde{\mathrm{Q}})$ by Lemma 22 and (Dudley, 2002, Theorem 4.1.10). We can thus focus on convergence of the cross-entropy and negative entropy terms individually.

By Lemma 12, the cross-entropy term converges. Turning to the negative entropy term, by Lemma 14, we need to prove,

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

Lemma 15 gives $\log \widetilde{q}^{(n)}(\widetilde{Z}^{(n)}) \to \log q(Z)$ a.s. Lemma 19 then yields the convergence in mean. Therefore,

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

It remains to investigate the case $\log p \notin \mathrm{L}^1(\mathrm{Q})$. Because Lemmas 14 and 15 and thus also Lemma 19 still hold, we can invoke Lemma 20 which establishes that both the sequence $(\mathrm{KL}\left(\mathrm{Q}^{(n)} \,\|\, \mathrm{P}\right) - s^{(n)})$ and the integral $\mathbb{E}_{\widetilde{\mathrm{Q}}} \log \frac{q}{p|_S}$ do not converge as desired. □

Proposition 10. *For $f \in C(\mathbb{R}^{\mathrm{D}})$, a collection of random variables $\{f(Z^{(n)})\}_{n\in\mathbb{N}}$ 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 $c_1, \ldots, c_p \in \mathbb{R}$, and $\mathbb{E}\,\|Z\|_2^p < \infty$.*[5]

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,

$$\lim_{k\to\infty} \limsup_{n\to\infty} \mathbb{E}\, \mathbb{I}_{|\xi_n|>k} |\xi_n| = 0 \,. \tag{12}$$

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

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

with $h_p(Z)$ being a linear combination of terms $\|Z^{(n)}\|_2^k$ for $k \in 0, 1, \ldots, p$. 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 $\mathbb{E}\, \mathbb{I}_{U^{\mathrm{C}}} |f(Z^{(n)})| \leq \sup_{U^{\mathrm{C}}} |f|$ which is finite by continuity of $|f|$ and compactness of U^{C} (Heine–Borel theorem), the random variables $\{f(Z^{(n)})\}_{n\in\mathbb{N}}$ are integrable.

By Equation (12), it is sufficient if $\forall \varepsilon > 0$, $\exists k \in \mathbb{R}$ s.t.,

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

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 w.l.o.g. 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 w.l.o.g. 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 r.h.s. summands can be upper bounded,

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

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

[5] 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.

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 $A_\alpha^{(n)} := \{x \in \mathbb{R}^{\mathrm{D}} \colon \|x\|_2 \leq \alpha\sqrt{n}\} \in \mathcal{B}$, $\beta^{(n)} := \mathrm{P}_{\mathcal{E}}(A_\alpha^{(n)})$ 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 $\mathbb{I}_{A_\alpha^{(n)}}(\mathcal{E})\,\mathbb{I}_{K_\delta^{\mathrm{C}}}\left(Z + \mathcal{E}/\sqrt{n}\right) \leq \mathbb{I}_{A_\alpha^{(n)}}(\mathcal{E})\,\mathbb{I}_{K_{\delta,\alpha}^{\mathrm{C}}}(Z)$ 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 r.h.s. 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,

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

for all $k \geq \kappa$ which concludes the proof. □

AUXILIARY LEMMAS

Lemma 11. *Assume S is a linear manifold, i.e. $S = \{x \in \mathbb{R}^{\mathrm{D}} \colon x = t(z), z \in S_0\}$, where $S_0 = \mathbb{R}^{\mathrm{K}_S} \times \{0\}^{\mathrm{D}-\mathrm{K}_S}$, and $t \colon x \mapsto b + Ax$ with $b \in \mathbb{R}^{\mathrm{D}}$ and $A \in \mathbb{R}^{\mathrm{D}\times\mathrm{D}}$ orthonormal. Then,*

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

with $(s^{(n)}) \subset \mathbb{R}$, if and only if,

$$\begin{aligned}
&\lim_{n\to\infty}\{\mathrm{KL}\,((t_{\#}^{-1}\mathrm{Q}) \star \mathrm{P}_{\mathcal{E}}^{(n)}\,\|\,t_{\#}^{-1}\mathrm{P}) - s^{(n)}\} \\
&\qquad = \mathop{\mathbb{E}}_{t_{\#}^{-1}\widetilde{\mathrm{Q}}} \log\frac{q \circ t}{p|_S \circ t}\,.
\end{aligned}$$

Furthermore,

$$q \circ t = \frac{\mathrm{d}t_{\#}^{-1}\widetilde{\mathrm{Q}}}{\mathrm{d}\lambda_{S_0}}\,,$$

where $\lambda_{S_0} = t_{\#}^{-1}m_S$ is the Lebesgue measure on S_0 with the corresponding trace σ-algebra $\mathcal{B}_{S_0}$. If q is continuous and bounded, then also $q \circ t$ is continuous bounded w.r.t. the corresponding trace topology.

Proof of Lemma 11. From definition, $t^{-1}(x) = A^{\mathrm{T}}(x - b)$ which is clearly a homeomorphism from $\mathbb{R}^{\mathrm{D}}$ onto itself. Because we are working with Borel σ-algebras, we can use Lemma 7.5 in (Gray, 2011) to establish,

$$\mathrm{KL}\,(\mathrm{Q}^{(n)}\,\|\,\mathrm{P}) = \mathrm{KL}\,(t_{\#}^{-1}\mathrm{Q}^{(n)}\,\|\,t_{\#}^{-1}\mathrm{P})\,.$$

By definition, $t_{\#}^{-1}\mathrm{Q}^{(n)} = \mathrm{Law}(t^{-1}(Z + \mathcal{E}^{(n)}))$; substituting $t^{-1}(Z + \mathcal{E}^{(n)}) = A^{\mathrm{T}}(Z + \mathcal{E}^{(n)} - b) = A^{\mathrm{T}}(Z - b) + A^{\mathrm{T}}\mathcal{E}^{(n)}$. Thus by properties of the multivariate normal distribution and orthonormality of A, $t_{\#}^{-1}\mathrm{Q}^{(n)} = \mathrm{Law}(A^{\mathrm{T}}(Z - b) + A^{\mathrm{T}}\mathcal{E}^{(n)}) = \mathrm{Law}(A^{\mathrm{T}}(Z - b) + \mathcal{E}^{(n)}) = (t_{\#}^{-1}\mathrm{Q}) \star \mathrm{P}_{\mathcal{E}}^{(n)}$, and therefore for all $n \in \mathbb{N}$,

$$\mathrm{KL}\,(\mathrm{Q}^{(n)}\,\|\,\mathrm{P}) = \mathrm{KL}\,((t_{\#}^{-1}\mathrm{Q}) \star \mathrm{P}_{\mathcal{E}}^{(n)}\,\|\,t_{\#}^{-1}\mathrm{P})\,.$$

Hence the two sequences of KL divergences are the same. By the substitution formula (see, for example, (Kallenberg, 2006, Lemma 1.22)),

$$\mathop{\mathbb{E}}_{t_{\#}^{-1}\widetilde{\mathrm{Q}}} \log\frac{q \circ t}{p|_S \circ t} = \mathop{\mathbb{E}}_{\widetilde{\mathrm{Q}}} \log\frac{q}{p|_S}\,,$$

which finishes the first part of the proof.

Because t is continuous, $q \circ t$ is continuous and bounded if the same holds for q. Finally, for any $B \in \mathcal{B}_{S_0}$,

$$\begin{aligned}
t_{\#}^{-1}\widetilde{\mathrm{Q}}(B) = \widetilde{\mathrm{Q}}(t(B)) &= \int_{t(B)} q\,\mathrm{d}m_S \\
&= \int_S \mathbb{I}_B\left(t^{-1}(x)\right) q(x)\,t_{\#}\lambda_{S_0}(\mathrm{d}x) \\
&= \int_S \mathbb{I}_B(x)\,q \circ t(x)\lambda_{S_0}(\mathrm{d}x) = \int_B q \circ t\,\mathrm{d}\lambda_{S_0}\,,
\end{aligned}$$

which shows that $q \circ t = \frac{\mathrm{d}t_{\#}^{-1}\widetilde{\mathrm{Q}}}{\mathrm{d}\lambda_{S_0}}$ as desired. □

Lemma 12. *If $\{\log p(Z^{(n)})\}$ is uniformly integrable, then $\mathbb{E}_{\mathrm{Q}^{(n)}} \log p \to \mathbb{E}_{\widetilde{\mathrm{Q}}} \log p|_S$ 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 $n \to \infty$. By Lemma 22, $\mathbb{E}_{\mathrm{Q}} \log p = \mathbb{E}_{\widetilde{\mathrm{Q}}} \log p|_S$, concluding the proof. □

Lemma 13. *If S is at most countable, then,*

$$\lim_{n\to\infty}\{\mathop{\mathbb{E}}_{\mathrm{Q}^{(n)}} \log q^{(n)} + \tfrac{\mathrm{D}}{2}\log(2\pi\mathrm{e}n^{-1})\} = \mathop{\mathbb{E}}_{\widetilde{\mathrm{Q}}} \log q\,.$$

Proof of Lemma 13. The density of $\widetilde{\mathrm{Q}}$ w.r.t. the counting measure on $\mathbb{Q}^{\mathrm{D}}$ can be written using the Kronecker's delta function δ_{Kr} as $q(x) = \sum_{i \in \mathbb{N}} \rho_i \delta_{\mathrm{Kr}}(x - m_i)$, where $\rho_i \geq 0$, $\sum_{i \in \mathbb{N}} \rho_i = 1$, and $m_i \in \mathbb{Q}^{\mathrm{D}}, \forall i \in \mathbb{N}$. Recall that by Equation (10), the density of $\mathrm{Q}^{(n)}$ w.r.t. λ^{D} is,

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

We can use the properties of multivariate normal distributions and the Tonelli–Fubini's theorem to establish,

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

which can be viewed as an integral over the product space $\mathbb{N} \times \mathbb{R}^{\mathrm{D}}$ (remember S is at most countable) w.r.t. the product measure of the distribution with density $i \mapsto \rho_i$ and the Gaussian $\mathcal{N}_{\mathbb{R}^{\mathrm{D}}}(0, I)$. For any $i \in \mathbb{N}$ 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 $n \to \infty$. Furthermore, because the sum inside the logarithm is upper bounded by one, we have $|f^{(n)}(i, \xi)| = -f^{(n)}(i, \xi)$, $\forall n \in \mathbb{N}$, and since $-\log x \downarrow \infty$ as $x \downarrow 0$, we obtain $|f^{(n)}(i, \xi)| \leq -f^{(*)}(i, \xi)$ which is the negative logarithm of the i^{th} summand in $\exp[f^{(n)}(i, \xi)]$ for all $n \in \mathbb{N}$. Observing,

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

we can invoke the dominated convergence theorem to establish (using the identity $-\frac{\mathrm{D}}{2} = -\frac{\mathrm{D}}{2} \log \mathrm{e}$),

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

as $n \to \infty$, concluding the proof. □

Lemma 14. *For S a linear manifold and every $n \in \mathbb{N}$, $\mathbb{E} \log q^{(n)}(Z^{(n)})$ is equal to,*

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

Proof of Lemma 14. As stated at the beginning of this section, we can w.l.o.g. assume $S = \mathbb{R}^{\mathrm{K}_S} \times \{0\}^{\mathrm{D} - \mathrm{K}_S}$. Then,

$$\begin{aligned} &\log q^{(n)}(x) \\ &= \log\left[\int_{\mathbb{R}^{\mathrm{D}}} (2\pi n^{-1})^{-\frac{\mathrm{D}}{2}} \mathrm{e}^{-\frac{\|x - z\|_2^2}{2n^{-1}}} \mathrm{Q}(\mathrm{d}z)\right] \\ &= -\frac{\mathrm{D} - \mathrm{K}_S}{2} \log(2\pi n^{-1}) - \frac{n}{2} \left\|x_{(\mathrm{K}_S + 1)\colon \mathrm{D}}\right\|_2^2 \\ &\quad + \log \underbrace{\left[\int_S \phi^{m_S}_{x_{1\colon \mathrm{K}_S} \times \{0\}^{\mathrm{D} - \mathrm{K}_S}, n^{-1} I} \mathrm{d}\widetilde{\mathrm{Q}}\right]}_{= \widetilde{q}^{(n)}(x_{1\colon \mathrm{K}_S} \times \{0\}^{\mathrm{D} - \mathrm{K}_S})} \,, \end{aligned}$$

$\forall x \in \mathbb{R}^{\mathrm{D}}$, where we used Lemma 22 for the last equality. Using the definition $Z^{(n)} = Z + \mathcal{E}/\sqrt{n}$,

$$\begin{aligned} &\mathbb{E} \log q^{(n)}(Z^{(n)}) \\ &= \int\!\!\int \phi^{\lambda^{\mathrm{D}}}_{0, I}(\xi) \log q^{(n)}(z + \xi/\sqrt{n}) \, \lambda^{\mathrm{D}}(\mathrm{d}\xi) \mathrm{Q}(\mathrm{d}z) \\ &= -\frac{\mathrm{D} - \mathrm{K}_S}{2} \log(2\pi n^{-1}) - \frac{n}{2} \mathbb{E} \left\|\mathcal{E}_{(\mathrm{K}_S + 1)\colon \mathrm{D}}/\sqrt{n}\right\|_2^2 \\ &\quad + \int\!\!\int \phi^{m_S}_{0, I}(\xi) \log \widetilde{q}^{(n)}(z + \xi/\sqrt{n}) \, m_S(\mathrm{d}\xi) \widetilde{\mathrm{Q}}(\mathrm{d}z) \\ &= -\frac{\mathrm{D} - \mathrm{K}_S}{2} \log(2\pi n^{-1}) - \frac{\mathrm{D} - \mathrm{K}_S}{2} \\ &\quad + \int\!\!\int \phi^{m_S}_{0, I}(\xi) \log \widetilde{q}^{(n)}(z + \xi/\sqrt{n}) \, m_S(\mathrm{d}\xi) \widetilde{\mathrm{Q}}(\mathrm{d}z) \\ &= -\frac{\mathrm{D} - \mathrm{K}_S}{2} \log(2\pi \mathrm{e} n^{-1}) + \mathbb{E} \log \widetilde{q}^{(n)}(\widetilde{Z}^{(n)}) \,, \end{aligned}$$

where the first equality is by the Tonelli–Fubini's theorem, the second by Lemma 22 and the standard marginalisation properties of the Gaussian distribution (the $\log \widetilde{q}^{(n)}$ term inside the integral only depends on $\mathcal{E}^{(n)}_{1\colon \mathrm{K}_S}$), the third by the relation of independent Gaussian variables and the χ^2 distribution, and the last again by the Tonelli-Fubini's theorem and the identity $-\frac{\mathrm{D} - \mathrm{K}_S}{2} = -\frac{\mathrm{D} - \mathrm{K}_S}{2} \log \mathrm{e}$. □

Lemma 15. *If S is a linear manifold, then,*

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

Proof. Clearly $\widetilde{Z}^{(n)} = Z + \widetilde{\mathcal{E}}/\sqrt{n} \to Z$ a.s. Hence for fixed values $Z = z$ and $\widetilde{\mathcal{E}} = \xi$,

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

by the triangle inequality. The second term on the r.h.s. 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,

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

where $\mathcal{N}_S(\mu, \Sigma)$ is the Gaussian distribution on S with the corresponding moments. 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 w.l.o.g. assume $t \leq \eta$ (take $t = \eta$ if not). For $A := \{x \in S \colon \|x\|_2 < t\}$,

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

where the latter term on the r.h.s. 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,

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

where the right hand side converges monotonically to ε as desired. Therefore $\log \widetilde{q}^{(n)}(\widetilde{Z}^{(n)}) \to \log q(Z)$ a.s. □

Lemma 16. *For S 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 16. Boundedness is a simple consequence of Equation (10) and the Hölder's inequality,

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

similarly for $\widetilde{q}^{(n)}$ using Equation (11).

The proofs of continuity are analogous, 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 $f_z(x) := \exp(-\frac{n}{2} \|x - z\|_2^2)$.

We can upper bound,

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

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,

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

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,

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

$\forall (z, \xi) \in \mathbb{R}^{\mathrm{D}} \times \mathbb{R}^{\mathrm{D}}$, 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,

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

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

Lemma 17. *For S is a linear manifold, $\widetilde{q}^{(n)}$ converges pointwise to q as $n \to \infty$.*

Proof of Lemma 17. W.l.o.g. assume $S = \mathbb{R}^{\mathrm{K}_S} \times \{0\}^{\mathrm{D} - \mathrm{K}_S}$ (c.f. Lemma 11). For arbitrary $x \in S$,

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

where $\mathcal{N}_S(\mu, \Sigma)$ is the Gaussian measure on S with the corresponding moments. Because q is continuous by assumption, for every $\varepsilon > 0$, $\exists \delta > 0$ s.t. $\|(x - \xi / \sqrt{n}) - x\|_2 = \|\xi / \sqrt{n}\|_2 < \delta \implies |q(x - \xi / \sqrt{n}) - q(x)| < \varepsilon$. For any $\alpha > 0$, we can use Chebyshev's inequality to determine $\mathrm{N} \in \mathbb{N}$ s.t. $\forall n \geq \mathrm{N}$, $\mathbb{P}(\|\xi / \sqrt{n}\|_2 \geq \delta) \leq \alpha$. Define $B \subset S$ to be the ball centred at zero with radius δ. Observe,

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

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

Lemma 18. *Assume $w_1, \dots, w_k \in \mathbb{R}$ are arbitrary constants, and ε_i, $i = 1, \dots, k$, are i.i.d. standard normal variables. Define the vector $w = (w_i)_{i=1}^k$. Then for $p \geq 0$,*

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

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. □

Lemma 19. *If S is a linear manifold, $\mathbb{E}\,\|Z\|_2^2 < \infty$, and $\log \widetilde{q}^{(n)}(\widetilde{Z}^{(n)}) \to \log q(Z)$ a.s., then as $n \to \infty$,*

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

Proof of Lemma 19. 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 $\delta > 0$. W.l.o.g. 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, and thus for any compact $\bar{B}_\delta$ we can find an interval $[s_\delta - r_\delta, s_\delta + r_\delta]$ satisfying the above condition for ν and all $\nu^{(n)}$.

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 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[6] function $h_{\delta,\eta}\,\mathrm{Id}$, for some fixed $\eta > 0$, where,

$$h_{\delta,\eta}(x) = \begin{cases} 1 & \text{, if } x \in \bar{B}_\delta \\ 0 & \text{, if } x \in F_{r,\eta} \\ \frac{r_\delta + \eta - |x - s_\delta|}{\eta} & \text{, else.} \end{cases}\,,$$

with $F_{\delta,\eta}$ defined as complement of $(s_\delta - r_\delta - \eta, s_\delta + r_\delta + \eta)$.

Using the triangle inequality,

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

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

Starting with the first term on the r.h.s., 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. $\mathbb{E}_\nu\, \mathbb{I}_{\bar{B}_\delta^{\mathrm{C}}}|\mathrm{Id}| \leq \varepsilon$. Denote this number by δ_1.

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 16, each $\widetilde{q}^{(n)}$ is bounded by C_q, thus we w.l.o.g. 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 r.h.s., $\forall n \in \mathbb{N}$. Then,

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

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 $\widetilde{\mathrm{Q}}^{(n)} \ll \mathcal{N}_S(0, I), \forall n \in \mathbb{N}$, and thus,

$$\mathop{\mathbb{E}}_{\widetilde{\mathrm{Q}}^{(n)}} \mathbb{I}_{(A_\delta^{(n)})^{\mathrm{C}}} \log \frac{\widetilde{q}^{(n)}}{\phi_{0,I}^{m_S}}$$

in the above derivation is well-defined. $\widetilde{\mathrm{Q}}^{(n)} \ll \mathcal{N}_S(0, I)$ implies that $\widetilde{\mathrm{Q}}^{(n)}((A_\delta^{(n)})^{\mathrm{C}}) > 0$ if $\mathcal{N}(0, I_S)((A_\delta^{(n)})^{\mathrm{C}}) > 0$ meaning we can upper bound the first term on the r.h.s. by,

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

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

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 18, 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 $\delta \to 0$ do so independently of n by uniform tightness of $\{\widetilde{\mathrm{Q}}^{(n)}\}_{n\in\mathbb{N}}$ and the construction of $A_\delta^{(n)}$. We can thus find constants $\mathrm{N}_1 \in \mathbb{N}$ and $\delta_2 > 0$ which will make $\mathbb{E}_{\nu^{(n)}}\,\mathbb{I}_{\bar{B}_\delta^{\mathrm{C}}}|\mathrm{Id}|$, $n \geq \mathrm{N}$, arbitrarily small.

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. Setting $\delta = \delta_1 \wedge \delta_2$, we can thus find $\mathrm{N}_2 \in \mathbb{N}$ that will make $\left|\mathbb{E}_\nu(h_{\delta,\eta}\mathrm{Id}) - \mathbb{E}_{\nu^{(n)}}(h_{\delta,\eta}\mathrm{Id})\right|$ arbitrarily small.

To establish that $|\mathbb{E}_\nu(\mathrm{Id}) - \mathbb{E}_{\nu^{(n)}}(\mathrm{Id})|$ can be made arbitrarily small, simply take $\mathrm{N} = \mathrm{N}_1 \vee \mathrm{N}_2$. Hence $\mathbb{E}\log\widetilde{q}^{(n)}(\widetilde{Z}^{(n)}) \to \mathbb{E}\log q(Z)$ as $n \to \infty$. □

Lemma 20. *If* $\log p \notin \mathrm{L}^1(\mathrm{Q})$, *and,*

$$\lim_{n\to\infty}\{\mathop{\mathbb{E}}_{\mathrm{Q}^{(n)}}\log q^{(n)} - s^{(n)}\} = \mathop{\mathbb{E}}_{\widetilde{\mathrm{Q}}}\log q ,$$

then $\mathbb{E}_{\widetilde{\mathrm{Q}}}\log\frac{q}{p|_S}$ *and* $(\mathbb{E}_{\mathrm{Q}^{(n)}}\log\frac{q}{p|_S} - s^{(n)})$ *diverge.*

Proof of Lemma 20. By Lemma 22, $\log p|_S \notin \mathrm{L}^1(\widetilde{\mathrm{Q}})$. Because $\log q \in \mathrm{L}^1(\widetilde{\mathrm{Q}})$ by assumption, we have $\log\frac{q}{p|_S} \notin \mathrm{L}^1(\widetilde{\mathrm{Q}})$ which yields the first part of the claim.

We now turn to the second part, i.e. to the sequence $(\mathbb{E}_{\mathrm{Q}^{(n)}}\log\frac{q}{p|_S} - s^{(n)})$.

First, we prove that $\mathbb{E}_{\mathrm{Q}^{(n)}}\log p$ cannot converge. Since p is bounded, we can w.l.o.g. assume $\log p \leq 0$. If $\log p \notin \mathrm{L}^1(\mathrm{Q}^{(n)})$ infinitely often, $\mathbb{E}\log p(Z^{(n)})$ does not converge. Otherwise $\log p \in \mathrm{L}^1(\mathrm{Q}^{(n)}), \forall n \geq \mathrm{N}$, for some $\mathrm{N} \in \mathbb{N}$. Notice that $Z^{(n)} \to Z$ a.s., and by continuity of $\log p$, also $\log p(Z^{(n)}) \to \log p(Z)$ a.s. Because the zero function is trivially integrable and $\log p \leq 0$, we can use the reverse Fatou's lemma to establish,

$$\limsup_{n\to\infty}\mathbb{E}\log p(Z^{(n)}) \leq \mathbb{E}\log p(Z) = -\infty ,$$

where we have used that $\log p \leq 0$ and $\log p \notin \mathrm{L}^1(\mathrm{Q})$ in the last equality. Again $\mathbb{E}\log p(Z^{(n)})$ does not converge.

Now we need to prove $(\mathbb{E}_{\mathrm{Q}^{(n)}}\log\frac{q^{(n)}}{p} - s^{(n)})$ does not converge. Assume the sequence converges to some $\kappa \in \mathbb{R}$. By assumption $(\mathbb{E}_{\mathrm{Q}^{(n)}}\log q^{(n)} - s^{(n)})$ converges. Thus by (Dudley, 2002, Theorem 4.1.10), $\mathbb{E}_{\mathrm{Q}^{(n)}}\log p$ must also converge which is a contradiction of the divergence established above. Therefore $(\mathbb{E}_{\mathrm{Q}^{(n)}}\log\frac{q^{(n)}}{p} - s^{(n)})$ cannot converge, proving the second part of the claim. □

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; such function 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 21. *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* S *be at most countable and all the relevant aforementioned assumptions hold. We will consider two cases:* $\log p \in \mathrm{L}^1(\mathrm{Q})$ *and* $\log p \notin \mathrm{L}^1(\mathrm{Q})$.

Then,

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

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

Proof of Proposition 21. By assumption, $\operatorname{diam}(S) < \infty$ and thus we can find a compact set $K \subset \mathbb{R}^{\mathrm{D}}$ s.t. $S \subset K$. W.l.o.g. 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 $\log p \in \mathrm{L}^1(\mathrm{Q})$. By Lemma 22, $\log p|_S \in \mathrm{L}^1(\widetilde{\mathrm{Q}})$. Because we assumed that $\log q \in \mathrm{L}^1(\widetilde{\mathrm{Q}})$,

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

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 $q^{(n)}(k_n(x)) \to \widetilde{\mathrm{Q}}(\{x\})$, as for any $x' \in S \setminus \{x\}$, $\widetilde{\mathrm{Q}}(\{x'\}) > 0$ 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 $q^{(n)}(k_n(x)) \downarrow \widetilde{\mathrm{Q}}(\{x\})$ by being a monotonically decreasing sequence with the least upper bound equal exactly to $\widetilde{\mathrm{Q}}(\{x\})$. Note that by assumption $\widetilde{\mathrm{Q}}(\{x\}) = q(x)$ where q is the density $\widetilde{\mathrm{Q}}$ of w.r.t. the counting measure on $\mathbb{Q}^{\mathrm{D}}$, and thus $q^{(n)}(k_n(x)) \downarrow q(x)$.

The following insight will help us:

$$\sum_{a \in A^{(n)}} q^{(n)}(a) h(a) = \int q(x) h(k_n(x)) m_S(\mathrm{d}x) , \qquad (14)$$

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 r.h.s. 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,

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

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 (14)). 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:

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

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

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

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

It remains to investigate the case $\log p \notin \mathrm{L}^1(\mathrm{Q})$. Notice that our proof of convergence of $(\mathbb{E}_{\mathrm{Q}^{(n)}} \log p^{(n)} + \mathrm{D} \log \Delta_n)$ to $\mathbb{E}_{\widetilde{\mathrm{Q}}} \log p|_S$ is independent of $\log p \in \mathrm{L}^1(\mathrm{Q})$ and is facilitated using the dominated convergence theorem. The dominated convergence theorem states that the pointwise limit itself must be integrable, and thus the case $\log p \notin \mathrm{L}^1(\mathrm{Q})$ is never realised under our assumptions by Lemma 22. □

B.3. Shared auxiliary lemmas

Lemma 22. *For* Q *a probability measure on* $(\mathbb{R}^{\mathrm{D}}, \mathcal{B})$*,* $\widetilde{\mathrm{Q}}$ *its restriction to* $(S, \mathcal{B}_S)$*, and a Borel measurable function* $f \colon \mathbb{R}^{\mathrm{D}} \to \mathbb{R}$*, the following holds,*

$$\mathop{\mathbb{E}}_{\mathrm{Q}} f = \mathop{\mathbb{E}}_{\widetilde{\mathrm{Q}}} f|_S ,$$

with $f|_S$ *being the restriction of* f *to* S*.*

Proof of Lemma 22. Because $\mathbb{E}_{\mathrm{Q}} f = \int f \,\mathrm{dQ} = \int f^+ \,\mathrm{dQ} - \int f^- \,\mathrm{dQ}$, with $f^+ = f \vee 0$ and $f^- = -(f \wedge 0)$, by definition of the Lebesgue integral, we can w.l.o.g. assume $f \geq 0$ so that $\int f \,\mathrm{dQ} = \sup\{\int g \,\mathrm{dQ} \colon 0 \leq g \leq f, g \text{ simple}\}$. For any simple g, using $\operatorname{supp}(\mathrm{Q}) = S$,

$$\begin{aligned}\int_{\mathbb{R}^{\mathrm{D}}} g \,\mathrm{dQ} &= \int_{\mathbb{R}^{\mathrm{D}}} \sum_{j=1}^{n} a_j \mathbb{I}_{B_j} \,\mathrm{dQ} = \sum_{j=1}^{n} a_j \mathrm{Q}(B_j) \\ &= \sum_{j=1}^{n} a_j \mathrm{Q}(B_j \cap S) = \sum_{j=1}^{n} a_j \widetilde{\mathrm{Q}}(B_j \cap S) \\ &= \int_S \sum_{j=1}^{n} a_j \mathbb{I}_{B_j} \,\mathrm{d}\widetilde{\mathrm{Q}} = \int_S g|_S \,\mathrm{d}\widetilde{\mathrm{Q}} \,,\end{aligned}$$

with $\{a_j\}_{j=1}^{n} \subset \mathbb{R}, n \in \mathbb{N}$. Taking the supremum on both sides establishes $\int_{\mathbb{R}^{\mathrm{D}}} f \,\mathrm{dQ} = \int_S f|_S \,\mathrm{d}\widetilde{\mathrm{Q}}$. □

C. Proofs for Section 5

Proof of Proposition 6. Let us first check the assumptions of Proposition 9. Clearly, the respective densities are continuous and bounded. Furthermore, the entropy of Q is equal to $\frac{1}{2} \log \det_*(2\pi \mathrm{e} \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^{\mathrm{T}})$ where $\det_*$ is the pseudo-determinant, and thus $\log q \in \mathrm{L}^1(\widetilde{\mathrm{Q}})$. It is also clear that $\mathbb{E}_{Z\sim \mathrm{Q}} \|Z\|_2^2 < \infty$ by the relation of the squared norm of Gaussian random variables and the χ^2 distribution. We will use Proposition 10 to ensure that the collection of random variables $\{\log p(Z^{(n)})\}$, $Z^{(n)} \sim \mathrm{Q}^{(n)}$, is uniformly integrable. Observe that for all $z \in \mathbb{R}^{\mathrm{D}}$,

$$|\log p(z)| \leq \mathrm{c} + \frac{1}{2} |z^{\mathrm{T}} \Sigma^{-1} z| \leq \mathrm{c} + \frac{\|z\|_2^2}{2\gamma_0} \,,$$

where $\mathrm{c} \in \mathbb{R}_+$ is a constant, and γ_0 is the lowest eigenvalue of Σ which is higher than zero because Σ is a (strictly) positive definite matrix by assumption. As we have already established $\mathbb{E}_{Z\sim \mathrm{Q}} \|Z\|_2^2 < \infty$, Proposition 10 holds and thus Proposition 9 can be applied.

For fixed $\boldsymbol{A}$, the Q distribution has support over the subspace $S = \{\boldsymbol{x} \in \mathbb{R}^{\mathrm{D}} \colon \boldsymbol{x} = \boldsymbol{A}\boldsymbol{z}, \boldsymbol{z} \in \mathbb{R}^{\mathrm{K}}\}$. 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 substitution which reduces QKL to,

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

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 $-\frac{1}{2}\log|\boldsymbol{V}| = -\frac{1}{2}\sum_{k=1}^{\mathrm{K}} \log \boldsymbol{V}_{kk}$ 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 $\boldsymbol{a}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_k = 1/\gamma_k$ which will be lowest for the highest eigenvalues γ_k of $\boldsymbol{\Sigma}$. Differentiating the objective w.r.t. $\boldsymbol{V}_{kk}$ after substituting the optimal $\boldsymbol{A}$ yields,

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

Setting to zero, we see that $\boldsymbol{V}_{kk} = \gamma_k$, 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,

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

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,

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

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 $\boldsymbol{V}_{kk} = \gamma_k - \tau^{(n)}$ where γ_k is the eigenvalue for the k^{th} substituted eigenvector. By substituting into $\mathcal{L}$,

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

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 $|\gamma_k - (\gamma_k - \tau^{(n)})| \to 0$ as $n \to \infty$. The optimal solution thus converges to the PCA/QKL in Frobenius/Euclidean distance. □

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.