

Supplementary Information for “The Impact of Architecture and Cost Function on Dissipative Quantum Neural Networks”

Tobias C. Sutter¹, Christopher Popp², and Beatrix C. Hiesmayr³

University of Vienna, Faculty of Physics, Währingerstrasse 17, 1090 Vienna.

¹Tobias.Christoph.Sutter@univie.ac.at

²Christopher.Popp@univie.ac.at

³Beatrix.Hiesmayr@univie.ac.at

A The Composite Parametrization of Isometries

In this section, we derive the composite parametrization for isometries. As the name suggests, it is obtained from the composite parametrization of the unitary group $\mathcal{U}(d)$ introduced in Ref. [1, 2].

Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces with dimensions d_1 and d_2 , respectively. Let $\mathcal{L}(\mathcal{H})$ be the set of linear operators mapping $\mathcal{H}$ into itself, and $\text{Iso}(\mathcal{H}_1, \mathcal{H}_2)$ the set of isometries from $\mathcal{H}_1$ to $\mathcal{H}_2$ with $d_1 \leq d_2$. Furthermore, let $\{|i\rangle_\ell\}_{i=0}^{d_\ell-1}$ be the computational basis of $\mathcal{H}_\ell$.

Any isometry $V \in \text{Iso}(\mathcal{H}_1, \mathcal{H}_2)$ with $d_1 \leq d_2$ can be written as

$$V = U \mathbb{1}_{d_2 \times d_1}, \quad (\text{A1})$$

where $U \in \mathcal{U}(d_2)$, and $\mathbb{1}_{d_2 \times d_1} \in \text{Iso}(\mathcal{H}_1, \mathcal{H}_2)$ denotes the $d_2 \times d_1$ matrix consisting of the first d_1 columns of $\mathbb{1}_{d_2 \times d_2}$. It can be written as

$$\mathbb{1}_{d_2 \times d_1} = \sum_{i=0}^{d_1-1} |i\rangle_2 \langle i|_1. \quad (\text{A2})$$

Note that not all basis vectors $|i\rangle_2$ of $\mathcal{H}_2$ need to appear in this sum. By (A1), the action $VAV^\dagger$ of V on $A \in \mathcal{L}(\mathcal{H}_1)$ can be interpreted as first canonically embedding A into $\mathcal{L}(\mathcal{H}_2)$ and subsequently applying the unitary U .

We can use the composite parametrization of unitary matrices [1, 2] to write U as

$$U = \left[\prod_{m=0}^{d_2-2} \left(\prod_{n=m+1}^{d_2-1} \Lambda_{m,n} \right) \right] \left[\prod_{l=0}^{d_2-1} e^{iP_l \lambda_{ll}} \right], \quad (\text{A3})$$

with

$$P_n := |n\rangle_2 \langle n|_2, \quad (\text{A4})$$

$$Y_{m,n} := -i|m\rangle_2 \langle n|_2 + i|n\rangle_2 \langle m|_2, \quad 0 \leq m < n \leq d_2 - 1, \quad (\text{A5})$$

$$\Lambda_{m,n} := e^{iP_n \lambda_{n,m}} e^{iY_{m,n} \lambda_{m,n}}, \quad (\text{A6})$$

and $\lambda_{m,n} \in [0, 2\pi]$ for $m \geq n$ and $\lambda_{m,n} \in [0, \pi/2]$ for $m < n$.

Using (A1) and (A3) we compute

$$\prod_{l=0}^{d_2-1} e^{iP_l \lambda_{ll}} \mathbb{1}_{d_2 \times d_1} = \left(\sum_{l=0}^{d_2-1} e^{i\lambda_{l,l}} |l\rangle_2 \langle l|_2 \right) \left(\sum_{k=0}^{d_1-1} |k\rangle_2 \langle k|_1 \right) \quad (\text{A7})$$

$$= \sum_{k=0}^{d_1-1} e^{i\lambda_{k,k}} |k\rangle_2 \langle k|_1, \quad (\text{A8})$$

and observe that only the first d_1 phases $\lambda_{k,k}$ are relevant for the isometry. This $d_2 \times d_1$ -dimensional matrix is explicitly given by

$$\prod_{l=0}^{d_2-1} e^{iP_l \lambda_{ll}} \mathbb{1}_{d_2 \times d_1} = \begin{pmatrix} e^{i\lambda_{0,0}} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & \dots & e^{i\lambda_{d_1-1,d_1-1}} \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{pmatrix}, \quad (\text{A9})$$

Next, we calculate

$$\Lambda_{m,n} = \mathbb{1}_{d_2 \times d_2} + (c_{m,n} - 1)P_m + (e_{n,m}c_{m,n} - 1)P_n - e_{n,m}s_{m,n}|n\rangle_2 \langle m|_2 + s_{m,n}|m\rangle_2 \langle n|_2, \quad (\text{A10})$$

where we abbreviated $\sin \lambda_{m,n} = s_{m,n}$, $\cos \lambda_{m,n} = c_{m,n}$, and $e^{i\lambda_{n,m}} = e_{n,m}$. In matrix notation, this amounts to

$$\Lambda_{m,n} = \begin{pmatrix} \mathbb{1}_{m \times m} & 0 & 0 & 0 & 0 \\ 0 & c_{m,n} & 0 & s_{m,n} & 0 \\ 0 & 0 & \mathbb{1}_{(n-m-1) \times (n-m-1)} & 0 & 0 \\ 0 & -e_{n,m} s_{m,n} & 0 & e_{n,m} c_{m,n} & 0 \\ 0 & 0 & 0 & 0 & \mathbb{1}_{(d_2-n-1) \times (d_2-n-1)} \end{pmatrix}. \quad (\text{A11})$$

By inspection, we see that $\Lambda_{m,n}$ acts trivially from the left on a matrix of the form (A9) if $m \geq d_1$. Thus, we can write the isometry (A1) as

$$V = \left[\prod_{m=0}^{d_2-2} \left(\prod_{n=m+1}^{d_2-1} \Lambda_{m,n} \right) \right] \left[\prod_{l=0}^{d_2-1} e^{iP_l \lambda_{ll}} \right] \mathbb{1}_{d_2 \times d_1} \quad (\text{A12})$$

$$= \left[\prod_{m=0}^{d_1-1} \left(\prod_{n=m+1}^{d_2-1} \Lambda_{m,n} \right) \right] \left[\prod_{l=0}^{d_1-1} e^{iP_l \lambda_{ll}} \right] \mathbb{1}_{d_2 \times d_1}. \quad (\text{A13})$$

Regarding the parameter count in (A13), we note that the first term on the right-hand side gives $\sum_{m=0}^{d_1-1} 2(d_2-m-1) = 2d_1d_2 - d_1^2 - d_1$ because each $\Lambda_{m,n}$ introduces two degrees of freedom. The second term on the right-hand side gives an additional d_1 free parameters. Thus, we find that the isometry $V \in \text{Iso}(\mathcal{H}_1, \mathcal{H}_2)$ in (A13) has $2d_1d_2 - d_1^2$ free real parameters. A simple argument shows that this is indeed the number of free real parameters of a general isometry in $\text{Iso}(\mathcal{H}_1, \mathcal{H}_2)$. Consequently, we cannot eliminate more parameters from (A13).

The parameters of the isometry (A13) can be conveniently collected in the matrix

$$(\lambda_{m,n})_{m,n} = \begin{pmatrix} \lambda_{0,0} & \cdots & \lambda_{0,d_1-1} & \cdots & \cdots & \lambda_{0,d_2-1} \\ \vdots & \ddots & \vdots & & & \vdots \\ \lambda_{d_1-1,0} & \cdots & \lambda_{d_1-1,d_1-1} & \cdots & \cdots & \lambda_{d_1-1,d_2-1} \\ \vdots & & \vdots & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ \lambda_{d_2-1,0} & \cdots & \lambda_{d_2-1,d_1-1} & 0 & \cdots & 0 \end{pmatrix}, \quad (\text{A14})$$

The diagonal entries $\lambda_{n,n}$ correspond to global phases in the respective subspaces. The entries $\lambda_{m,n}$ in the upper triangular part represent rotations in the subspaces spanned by $|n\rangle_2$ and $|m\rangle_2$, and $\lambda_{n,m}$ in the lower triangular part correspond to relative phases in these subspaces. Except for the diagonal entries $\lambda_{k,k}$, these are the same parameters needed to parameterize a general mixed state $\rho_2 \in \mathcal{D}(\mathcal{H}_2)$ of rank $d_1 \leq d_2$ (in addition to $d_1 - 1$ required mixing probabilities; cf. [1]).

A.1 Composite Parametrization for the Stinespring Isometry of a Quantum Channel

A general CPTP map $\mathcal{E} : \mathcal{D}(\mathcal{H}_1) \rightarrow \mathcal{D}(\mathcal{H}_2)$ can be written in its Stinespring representation as

$$\mathcal{E}(\rho) = \text{Tr}_{1,A}(V\rho V^\dagger), \quad (\text{A15})$$

where $V \in \text{Iso}(\mathcal{H}_1, \mathcal{H}_2 \otimes \mathcal{H}_A \otimes \mathcal{H}_1)$, and the Hilbert space $\mathcal{H}_A$ with $\dim(\mathcal{H}_A) = d_A$ corresponds to an ancilla system. We can choose $d_A = d_2$ because d_1d_2 is the maximal Kraus rank of $\mathcal{E}$ [3]. In this case, the isometry in (A13) with $d_2d_Ad_1 = d_1d_2^2$ has $d_1^2(2d_2^2 - 1)$ degrees of freedom. However, due to the unitary freedom on the space $\mathcal{H}_A \otimes \mathcal{H}_1$ (which has no physical relevance), a general CPTP map can be reduced to $d_1^2(d_2^2 - 1)$ degrees of freedom. Unfortunately, we cannot straightforwardly get rid of the $d_1^2d_2^2$ redundant parameters in (A14) because these degrees of freedom do not coincide one-to-one with the parameters $\lambda_{i,j}$.

Nonetheless, using (A13), we can write a Stinespring isometry for $\mathcal{E}$ as

$$V = U_C \mathbb{1}_{d_1d_2^2 \times d_1} = U_C (|0\rangle_2 \otimes |0\rangle_A \otimes \mathbb{1}_1), \quad (\text{A16})$$

with

$$U_C = \left[\prod_{m=0}^{d_1-1} \left(\prod_{n=m+1}^{d_1 d_2^2 - 1} \Lambda_{m,n} \right) \right] \left[\prod_{l=0}^{d_1-1} e^{i P_l \lambda_{ll}} \right]. \quad (\text{A17})$$

We thus obtain for any $\rho_1 \in \mathcal{D}(\mathcal{H}_1)$:

$$\mathcal{E}(\rho_1) = \text{Tr}_{1,A}[V \rho_1 V^\dagger] \quad (\text{A18})$$

$$= \sum_{k=0}^{d_2-1} \sum_{l=0}^{d_1-1} \langle k|_A \otimes \langle l|_1 V \rho_1 V^\dagger |k\rangle_A \otimes |l\rangle_1 \quad (\text{A19})$$

$$= \sum_{k=0}^{d_2-1} \sum_{l=0}^{d_1-1} \langle k|_A \otimes \langle l|_1 U_C |0\rangle_2 \otimes |0\rangle_A \rho_1 \langle 0|_2 \otimes \langle 0|_A U_C^\dagger |k\rangle_A \otimes |l\rangle_1 \quad (\text{A20})$$

$$= \sum_{k=0}^{d_2-1} \sum_{l=0}^{d_1-1} G_{k,l} \rho_1 G_{k,l}^\dagger, \quad (\text{A21})$$

with the Kraus operators $G_{k,l} = \langle k|_A \otimes \langle l|_1 V = \langle k|_A \otimes \langle l|_1 U_C |0\rangle_2 \otimes |0\rangle_A$.

Note that this representation of a quantum channel $\mathcal{E}$ requires us to tensor the systems $\mathcal{H}_A$ and $\mathcal{H}_2$ from the left onto $\mathcal{H}_1$ to obtain $|0\rangle_2 \otimes |0\rangle_A \otimes \mathbb{1}_1$ in (A16). In this case, the parameters in the composite parametrized isometry are reduced to the correct number $d_1^2(2d_2^2 - 1)$. If we tensor the systems $\mathcal{H}_A$ and $\mathcal{H}_2$ from the right onto $\mathcal{H}_1$, we would get $\mathbb{1}_1 \otimes |0\rangle_A \otimes |0\rangle_2$ in (A16). Consequently, the $d_1^2 d_2^4$ parameters of the full unitary U_C get reduced only by less than $d_1^2(d_2^2 - 1)^2$, and (A17) is no longer valid. The reason is that the reduction of parameters shown above cannot be carried out. This also becomes apparent when numerically optimizing DQNNs using the composite parametrization because the gradient for redundant degrees of freedom in the unitary formalism vanishes. If tensored in the “wrong” order, only some diagonal elements of the parameter matrix (A14) are irrelevant for the quantum channel and have vanishing derivative (cf. [4]). If done correctly, only the parameters $\lambda_{i,j}$ in (A14) are relevant for the optimization, i.e., have non-vanishing derivative in general. Consequently, the derivative of the cost function only needs to be calculated for those, leading to better computational performance.

B Derivatives of different cost functions

This section presents the derivatives of the different cost functions required for optimizing a DQNN by gradient descent/ascent. In Sec. 4, the resulting gradient matrix is used to update the isometry parameters, e.g., with the ADAM optimizer [5].

For example, consider an extended DQNN consisting of 5 qudits and two perceptrons U_1 and U_2 in the unitary formulation. It comprises one input, one hidden, one output, and two ancilla layers. Using the notation of Sec. 2.2, we have $U_1 \equiv U_1^{(1,2,3)}$ and $U_2 \equiv U_1^{(3,4,5)}$. Let us denote the set of variational parameters in U_m as $\mathcal{S}_m = \{\lambda_{x,y}^{(m)}\}_{x,y}$. Due to the equivalence of the isometry and the unitary picture, this is already a reduced set of parameters (cf. (A14)). The network channel is given by

$$\rho_{\text{out}} = \mathcal{E}_{\text{net}}(\rho_{\text{in}}) = \text{Tr}_{1,2,3,4} \{ U (|\mathbf{0}\rangle\langle\mathbf{0}| \otimes \rho_{\text{in}}) U^\dagger \}, \quad (\text{B1})$$

where $U = U_2 U_1$, and $|\mathbf{0}\rangle\langle\mathbf{0}| \equiv |\mathbf{0}\rangle\langle\mathbf{0}|_{5,4,3,2}$ is the initial state of the ancilla, hidden, and output layers.

Gradient optimization utilizes either the numerical or the analytical derivative of the total cost function (11). In the former case, we approximate the cost function gradient by

$$\frac{\partial C_{\text{tot}}}{\partial \lambda_{x,y}^{(m)}} = \frac{1}{N_t} \sum_{i=1}^{N_t} \frac{\partial}{\partial \lambda_{x,y}^{(m)}} C(\rho_{\text{tar}}^{(i)}, \rho_{\text{out}}^{(i)}) \quad (\text{B2})$$

$$\approx \frac{1}{N_t} \sum_{i=1}^{N_t} \frac{C(\rho_{\text{tar}}^{(i)}, \rho_{\text{out}}^{(i)}(\lambda_{x,y}^{(m)} + \varepsilon)) - C(\rho_{\text{tar}}^{(i)}, \rho_{\text{out}}^{(i)}(\lambda_{x,y}^{(m)} - \varepsilon))}{2\varepsilon} \quad (\text{B3})$$

for each $m \in \{1, 2\}$ and relevant x, y (cf. (A14)). This method requires choosing a suitable small $\varepsilon > 0$ and is generally only an approximation of the true gradient. Thus, we resort to it only when we cannot compute the cost function's analytical derivative. This is the case for the quantum Chernoff bound (25) and the quantum relative entropy (26).

For the analytic approach, we compute (B2) analytically for each $m \in \{1, 2\}$ and relevant x, y . To unclutter the notation in the following, we focus on one input-target pair of the training set (i.e., one term in the sum (B2)) and drop the superscript (i) . When evaluating $\partial C(\rho_{\text{tar}}, \rho_{\text{out}})/\partial \lambda_{x,y}^{(m)}$, we necessarily encounter $\partial \rho_{\text{out}}/\partial \lambda_{x,y}^{(m)}$ due to the chain rule. Thus, before moving on, we first take care of this. As shown in [4], we can calculate for our example DQNN

$$\frac{\partial \rho_{\text{out}}}{\partial \lambda_{x,y}^{(1)}} = \text{Tr}_{1,2,3,4} \left\{ U_2 U_1 i [\tilde{Y}_{x,y}^{(1)}, \tilde{\rho}] U_1^\dagger U_2^\dagger \right\}, \quad (\text{B4})$$

$$\frac{\partial \rho_{\text{out}}}{\partial \lambda_{x,y}^{(2)}} = \text{Tr}_{1,2,3,4} \left\{ U_2 i [\tilde{Y}_{x,y}^{(2)}, U_1 \tilde{\rho} U_1^\dagger] U_2^\dagger \right\}, \quad (\text{B5})$$

where $[\cdot, \cdot]$ denotes the commutator and $\tilde{\rho} = |\mathbf{0}\rangle\langle\mathbf{0}| \otimes \rho_{\text{in}}$. Furthermore

$$\tilde{Y}_{x,y}^{(m)} = \begin{cases} (U_{x,y}^{(m)})^\dagger Y_{x,y}^{(m)} U_{x,y}^{(m)} & , x < y \\ P_x^{(m)} & , x = y \\ (U_{x,y}^{(m)})^\dagger P_x^{(m)} U_{x,y}^{(m)} & , x > y \end{cases} \quad (\text{B6})$$

with

$$U_{x,y}^{(m)} = \begin{cases} \left[\prod_{k=y+1}^{d-1} \Lambda_{x,k}^{(m)} \right] \left[\prod_{k=x+1}^{d-2} \prod_{l=k+1}^{d-1} \Lambda_{k,l}^{(m)} \right] \prod_{l=x}^{d-1} \exp(i P_l^{(m)} \lambda_{l,l}^{(m)}) & \text{for } x < y, \\ \left[\prod_{k=x}^{d-1} \Lambda_{y,k}^{(m)} \right] \left[\prod_{k=y+1}^{d-2} \prod_{l=k+1}^{d-1} \Lambda_{k,l}^{(m)} \right] \prod_{l=y}^{d-1} \exp(i P_l^{(m)} \lambda_{l,l}^{(m)}) & \text{for } x > y, \end{cases} \quad (\text{B7})$$

and where $P_x^{(m)}$, $Y_{x,y}^{(m)}$ and $\Lambda_{k,l}^{(m)}$ as in (A4), (A5) and (A6), respectively. Generalizing the calculation for (B4) and (B5) to networks with more (unitary) perceptrons U_m is straightforward.

The remainder of this section deals with evaluating $\partial C(\rho_{\text{tar}}, \rho_{\text{out}})/\partial \lambda_{x,y}^{(m)}$ for the cost functions presented in Sec. 3. It is important to keep in mind that only ρ_{out} depends on the variational parameters $\lambda_{x,y}^{(m)}$. Hence, $\partial \rho_{\text{out}}/\partial \lambda_{x,y}^{(m)} \neq 0$ in general, but $\partial \rho_{\text{tar}}/\partial \lambda_{x,y}^{(m)} \equiv 0$. Throughout, we furthermore assume $\rho_{\text{tar}} \neq \rho_{\text{out}}$ as this can become problematic for the gradient of certain cost functions. This condition can be implemented in the training algorithm: Before computing the gradient for any optimization iteration, check if $\rho_{\text{out}} = \rho_{\text{tar}}$. If yes, the network is already optimal for this training state, and we exclude this instance from the present iteration of the gradient calculation.

B.1 Derivative of D_{HS}

A straightforward calculation using (18) yields

$$\frac{\partial D_{\text{HS}}(\rho_{\text{tar}}, \rho_{\text{out}})}{\partial \lambda_{x,y}^{(m)}} = \frac{\partial}{\partial \lambda_{x,y}^{(m)}} \sqrt{\text{Tr}((\rho_{\text{out}} - \rho_{\text{tar}})^2)} \quad (\text{B8})$$

$$= \frac{1}{2 D_{\text{HS}}(\rho_{\text{tar}}, \rho_{\text{out}})} \frac{\partial}{\partial \lambda_{x,y}^{(m)}} \text{Tr}((\rho_{\text{out}} - \rho_{\text{tar}})^2) \quad (\text{B9})$$

$$= \frac{1}{D_{\text{HS}}(\rho_{\text{tar}}, \rho_{\text{out}})} \text{Tr} \left((\rho_{\text{out}} - \rho_{\text{tar}}) \frac{\partial \rho_{\text{out}}}{\partial \lambda_{x,y}^{(m)}} \right). \quad (\text{B10})$$

B.2 Derivative of D_{Tr}

The derivative of the trace distance cost function (19) is

$$\frac{\partial D_{\text{Tr}}(\rho_{\text{tar}}, \rho_{\text{out}})}{\partial \lambda_{x,y}^{(m)}} = \frac{1}{2} \frac{\partial}{\partial \lambda_{x,y}^{(m)}} \text{Tr} \left(\sqrt{(\rho_{\text{out}} - \rho_{\text{tar}})^2} \right) \quad (\text{B11})$$

$$= \frac{1}{4} \text{Tr} \left(\left(\sqrt{(\rho_{\text{out}} - \rho_{\text{tar}})^2} \right)^{-1} \frac{\partial}{\partial \lambda_{x,y}^{(m)}} (\rho_{\text{out}} - \rho_{\text{tar}})^2 \right) \quad (\text{B12})$$

$$= \frac{1}{2} \text{Tr} \left(\left(\sqrt{(\rho_{\text{out}} - \rho_{\text{tar}})^2} \right)^{-1} (\rho_{\text{out}} - \rho_{\text{tar}}) \frac{\partial \rho_{\text{out}}}{\partial \lambda_{x,y}^{(m)}} \right). \quad (\text{B13})$$

If $A = \sqrt{(\rho_{\text{out}} - \rho_{\text{tar}})^2}$ is singular, we use the Moore-Penrose inverse [6] instead of A^{-1} . In more detail, in (B12) we used

$$\text{Tr} \left(\frac{\partial}{\partial \lambda} \sqrt{A} \right) = \text{Tr} \left(\frac{1}{2} \frac{\partial A}{\partial \lambda} (A^{1/2})^{-1} \right) \quad (\text{B14})$$

for $\lambda = \lambda_{x,y}^{(m)}$ and $A = (\rho_{\text{out}} - \rho_{\text{tar}})^2$. To evaluate this, we can utilize the power series expansion for the matrix square root, given by

$$A^{1/2} = \sum_{n=0}^{\infty} (-1)^n \binom{1/2}{n} (\mathbb{1} - A)^n, \quad (\text{B15})$$

which is convergent if the spectrum of A satisfies $\text{spec}(A) \subseteq \mathcal{D}(1, 1) \subset \mathbb{C}$, where $\mathcal{D}(1, 1)$ denotes a disk with radius 1 and centered at 1 in $\mathbb{C}$. If $\text{spec}(A) \subseteq (0, 1]$, i.e., A is non-singular, the inverse of (B15) is

$$A^{-1/2} = \sum_{n=0}^{\infty} (-1)^n \binom{-1/2}{n} (\mathbb{1} - A)^n. \quad (\text{B16})$$

For singular A we can use the Moore-Penrose pseudoinverse to define (B16). We can then calculate

$$\text{Tr} \left(\frac{\partial}{\partial \lambda} \sqrt{A} \right) = \text{Tr} \left(\frac{\partial}{\partial \lambda} \sum_{n=0}^{\infty} (-1)^n \binom{1/2}{n} (\mathbb{1} - A)^n \right) \quad (\text{B17})$$

$$= \text{Tr} \left(\sum_{n=0}^{\infty} (-1)^n \binom{1/2}{n} \sum_{k=0}^{n-1} (\mathbb{1} - A)^k \frac{\partial (\mathbb{1} - A)}{\partial \lambda} (\mathbb{1} - A)^{n-k-1} \right) \quad (\text{B18})$$

$$= \text{Tr} \left(\frac{\partial A}{\partial \lambda} \sum_{n=1}^{\infty} (-1)^{n+1} \binom{1/2}{n} n (\mathbb{1} - A)^{n-1} \right) \quad (\text{B19})$$

$$= \text{Tr} \left(\frac{1}{2} \frac{\partial A}{\partial \lambda} A^{-1/2} \right), \quad (\text{B20})$$

where in the third equality we used that the trace is cyclic, and a simple index shift in combination with the identity $\binom{1/2}{n+1} (n+1) = \frac{1}{2} \binom{-1/2}{n}$ yields the last line.

B.3 Derivative of F_1 and D_1

The derivative of the fidelity cost function (21) takes the form

$$\frac{\partial F_1(\rho_{\text{tar}}, \rho_{\text{out}})}{\partial \lambda_{x,y}^{(m)}} = \frac{\partial}{\partial \lambda_{x,y}^{(m)}} \text{Tr} \left(\sqrt{\sqrt{\rho_{\text{tar}}} \rho_{\text{out}} \sqrt{\rho_{\text{tar}}}} \right)^2 \quad (\text{B21})$$

$$= 2 \sqrt{F_1(\rho_{\text{tar}}, \rho_{\text{out}})} \text{Tr} \left(\frac{\partial}{\partial \lambda_{x,y}^{(m)}} \sqrt{\sqrt{\rho_{\text{tar}}} \rho_{\text{out}} \sqrt{\rho_{\text{tar}}}} \right) \quad (\text{B22})$$

$$= \sqrt{F_1(\rho_{\text{tar}}, \rho_{\text{out}})} \text{Tr} \left(\frac{\partial \rho_{\text{out}}}{\partial \lambda_{x,y}^{(m)}} \sqrt{\rho_{\text{tar}}} \left(\sqrt{\sqrt{\rho_{\text{tar}}} \rho_{\text{out}} \sqrt{\rho_{\text{tar}}}} \right)^{-1} \sqrt{\rho_{\text{tar}}} \right), \quad (\text{B23})$$

where in the last line we used (B14) with $A = \sqrt{\rho_{\text{tar}}} \rho_{\text{out}} \sqrt{\rho_{\text{tar}}}$. This is valid as one can show that $\text{spec}(\sqrt{\rho_{\text{tar}}} \rho_{\text{out}} \sqrt{\rho_{\text{tar}}}) \subset [0, 1]$.

For the Bures distance cost function (22), we obtain

$$\frac{\partial D_1(\rho_{\text{tar}}, \rho_{\text{out}})}{\partial \lambda_{x,y}^{(m)}} = \frac{\partial}{\partial \lambda_{x,y}^{(m)}} \sqrt{2 \left(1 - \sqrt{F_1(\rho_{\text{tar}}, \rho_{\text{out}})} \right)} \quad (\text{B24})$$

$$= \frac{1}{D_1(\rho_{\text{tar}}, \rho_{\text{out}})} \frac{\partial}{\partial \lambda_{x,y}^{(m)}} \left(1 - \sqrt{F_1(\rho_{\text{tar}}, \rho_{\text{out}})} \right) \quad (\text{B25})$$

$$= \frac{-1}{2D_1(\rho_{\text{tar}}, \rho_{\text{out}}) \sqrt{F_1(\rho_{\text{tar}}, \rho_{\text{out}})}} \frac{\partial F_1(\rho_{\text{tar}}, \rho_{\text{out}})}{\partial \lambda_{x,y}^{(m)}} \quad (\text{B26})$$

$$= \frac{-1}{2D_1(\rho_{\text{tar}}, \rho_{\text{out}})} \text{Tr} \left(\frac{\partial \rho_{\text{out}}}{\partial \lambda_{x,y}^{(m)}} \sqrt{\rho_{\text{tar}}} \left(\sqrt{\sqrt{\rho_{\text{tar}}} \rho_{\text{out}} \sqrt{\rho_{\text{tar}}}} \right)^{-1} \sqrt{\rho_{\text{tar}}} \right), \quad (\text{B27})$$

where we used (B23) in the last line.

B.4 Derivative of F_2 and D_2

To take the derivative of the fidelity F_2 , we write it as

$$F_2(\rho_{\text{tar}}, \rho_{\text{out}}) = \frac{\mathcal{A}}{\mathcal{B}} \quad (\text{B28})$$

where

$$\mathcal{A} = \text{Tr}(\rho_{\text{out}} \rho_{\text{tar}}), \quad (\text{B29})$$

$$\mathcal{B} = \max \{ \text{Tr}(\rho_{\text{out}}^2), \text{Tr}(\rho_{\text{tar}}^2) \}. \quad (\text{B30})$$

Thus, we have

$$\frac{\partial F_2(\rho_{\text{tar}}, \rho_{\text{out}})}{\partial \lambda_{x,y}^{(m)}} = \frac{\frac{\partial \mathcal{A}}{\partial \lambda_{x,y}^{(m)}} \mathcal{B} - \mathcal{A} \frac{\partial \mathcal{B}}{\partial \lambda_{x,y}^{(m)}}}{\mathcal{B}^2}. \quad (\text{B31})$$

The derivative of $\mathcal{A}$ is easily evaluated to be

$$\frac{\partial \mathcal{A}}{\partial \lambda_{x,y}^{(m)}} = \text{Tr} \left(\frac{\partial \rho_{\text{out}}}{\partial \lambda_{x,y}^{(m)}} \rho_{\text{tar}} \right). \quad (\text{B32})$$

By writing

$$\max \{ \text{Tr}(\rho_{\text{out}}^2), \text{Tr}(\rho_{\text{tar}}^2) \} = \frac{1}{2} (\text{Tr}(\rho_{\text{out}}^2) + \text{Tr}(\rho_{\text{tar}}^2) + |\text{Tr}(\rho_{\text{out}}^2) - \text{Tr}(\rho_{\text{tar}}^2)|) \quad (\text{B33})$$

$$= \frac{1}{2} \left(\text{Tr}(\rho_{\text{out}}^2 + \rho_{\text{tar}}^2) + \sqrt{\text{Tr}(\rho_{\text{out}}^2 - \rho_{\text{tar}}^2)^2} \right), \quad (\text{B34})$$

we see that $\mathcal{B}$ is not differentiable at $\rho_{\text{out}} = \rho_{\text{tar}}$. For all other cases, we can compute

$$\frac{\partial \mathcal{B}}{\partial \lambda_{x,y}^{(m)}} = \text{Tr} \left(\frac{\partial \rho_{\text{out}}}{\partial \lambda_{x,y}^{(m)}} \rho_{\text{out}} \right) \left(1 + \frac{\text{Tr}(\rho_{\text{out}}^2 - \rho_{\text{tar}}^2)}{\sqrt{\text{Tr}(\rho_{\text{out}}^2 - \rho_{\text{tar}}^2)^2}} \right) \quad (\text{B35})$$

$$= \text{Tr} \left(\frac{\partial \rho_{\text{out}}}{\partial \lambda_{x,y}^{(m)}} \rho_{\text{out}} \right) (1 + \text{sign}(\text{Tr}(\rho_{\text{out}}^2 - \rho_{\text{tar}}^2))) . \quad (\text{B36})$$

Inserting (B32) and (B36) into (B31) finally yields after some simplifications

$$\frac{\partial F_2(\rho_{\text{tar}}, \rho_{\text{out}})}{\partial \lambda_{x,y}^{(m)}} = \frac{\text{Tr} \left(\frac{\partial \rho_{\text{out}}}{\partial \lambda_{x,y}^{(m)}} (\rho_{\text{tar}} - \rho_{\text{out}} F_2(\rho_{\text{tar}}, \rho_{\text{out}}) (1 + \text{sign}(\text{Tr}(\rho_{\text{out}}^2 - \rho_{\text{tar}}^2))) \right)}{\max \{ \text{Tr}(\rho_{\text{out}}^2), \text{Tr}(\rho_{\text{tar}}^2) \}}. \quad (\text{B37})$$

For the D_2 distance (24), the derivative evaluates to

$$\frac{\partial D_2(\rho_{\text{tar}}, \rho_{\text{out}})}{\partial \lambda_{x,y}^{(m)}} = \frac{-1}{D_2(\rho_{\text{tar}}, \rho_{\text{out}})} \frac{\partial F_2(\rho_{\text{tar}}, \rho_{\text{out}})}{\partial \lambda_{x,y}^{(m)}} \quad (\text{B38})$$

$$= \frac{-\text{Tr} \left(\frac{\partial \rho_{\text{out}}}{\partial \lambda_{x,y}^{(m)}} (\rho_{\text{tar}} - \rho_{\text{out}} F_2(\rho_{\text{tar}}, \rho_{\text{out}}) (1 + \text{sign}(\text{Tr}(\rho_{\text{out}}^2 - \rho_{\text{tar}}^2)))) \right)}{D_2(\rho_{\text{tar}}, \rho_{\text{out}}) \max \{ \text{Tr}(\rho_{\text{out}}^2), \text{Tr}(\rho_{\text{tar}}^2) \}}. \quad (\text{B39})$$

References

1. Spengler, C., Huber, M. & Hiesmayr, B. C. A composite parameterization of unitary groups, density matrices and subspaces. en. *Journal of Physics A: Mathematical and Theoretical* **43**, 385306. ISSN: 1751-8113, 1751-8121. <https://iopscience.iop.org/article/10.1088/1751-8113/43/38/385306> (2024) (Sept. 2010).
2. Spengler, C., Huber, M. & Hiesmayr, B. C. Composite parameterization and Haar measure for all unitary and special unitary groups. en. *Journal of Mathematical Physics* **53**, 013501. ISSN: 0022-2488, 1089-7658. <https://pubs.aip.org/jmp/article/53/1/013501/232964/Composite-parameterization-and-Haar-measure-for> (2023) (Jan. 2012).
3. Watrous, J. *The Theory of Quantum Information* 1st ed. en. ISBN: 978-1-107-18056-7. <https://www.cambridge.org/core/product/identifier/9781316848142/type/book> (2024) (Cambridge University Press, Apr. 2018).
4. Hiesmayr, B. C. A quantum information theoretic view on a deep quantum neural network. *AIP Conference Proceedings* **3061**, 020001. ISSN: 0094-243X. <https://doi.org/10.1063/5.0203600> (2025) (Mar. 2024).
5. Kingma, D. P. & Ba, J. *Adam: A Method for Stochastic Optimization* en. arXiv:1412.6980 [cs]. Jan. 2017. <http://arxiv.org/abs/1412.6980> (2025).
6. Penrose, R. A generalized inverse for matrices. en. *Mathematical Proceedings of the Cambridge Philosophical Society* **51**, 406–413. ISSN: 0305-0041, 1469-8064. https://www.cambridge.org/core/product/identifier/S0305004100030401/type/journal_article (2025) (July 1955).