Appendix A for the paper (Basile and Tamburini, 2017): Calculations of the Probability Function and its Gradient in a Quantum Language Model

1 Starting Formulae

In this section we recall the relevant definitions used in the main paper (Basile and Tamburini, 2017). We work in the Hilbert space $\mathcal{H}_{ancilla} \otimes \mathcal{H}_{system} = \mathbb{C}^D \otimes \mathbb{C}^N = \mathbb{C}^{DN}$. We use an orthonormal basis $\{|w\rangle\}$ for $\mathcal{H}_{system}$ where each vector is labelled by a word $w \in \{1, ..., N\}$. On $\mathcal{H}_{system}$, we define orthogonal projectors Π_w on each basis vector (strictly speaking the span thereof). Their matrix components read $(\Pi_w)_{ij} = \delta_{iw}\delta_{jw},\ i, j = 1, ..., N$. Notice that in this expression we canonically identify each word w with its corresponding number.

Using the projectors Π_w on $\mathcal{H}_{system}$ we define projector on the full Hilbert space by

$$\Pi_w^{(2)} = \mathbf{I}_D \otimes \Pi_w. \tag{1}$$

The state of the model is encoded in a density matrix ρ, a positive semidefinite matrix of unit trace, $\mathrm{Tr}(\rho) = 1$. The dynamics are described by unitary evolution operators.

1.1 Unitary Evolution Operators

As explained in the main paper, we used a set $\{V(w)\}_{w\in\{1,...,N\}}$ of unitary operators, one for each word, which are constructed using a smaller set of p unitary operators $\{U_1, ..., U_p\}$, denoted in the following with an array $\mathbf{U} = (U_1, ..., U_p)$, and an embedding $\alpha : \{1, ..., N\} \to \mathbb{R}^p$.

The embedding is used to construct real vectors $\alpha(w) = (\alpha_1(w), ..., \alpha_p(w))$, and the components of these vectors are used to build the set of evolution operators in the following fashion:

$$w \mapsto V(w) = \prod_{i=1}^{p} U_i^{\alpha_i(w)}. \tag{2}$$

1.2 Probability Function

Equipped with this structure, we recall the final formula for the probability function $P_n(\mathbf{U}) \equiv P(\mathbf{w}|\mathbf{U})$ for the occurence of a sequence $\mathbf{w} = (w_1, ..., w_n)$ of words

$$P_n(\mathbf{U}) = \mathrm{Tr}(\Pi_{w_n}^{(2)} ... V^\dagger(w_2)\Pi_{w_2}^{(2)} V^\dagger(w_1)\Pi_{w_1}^{(2)} \rho_0 \Pi_{w_1}^{(2)} V(w_1)\Pi_{w_2}^{(2)} V(w_2) ... \Pi_{w_n}^{(2)}). \tag{3}$$

The $p(DN)^2$ real parameters of the model are encoded in the array $\mathbf{U}$, while the initial density matrix ρ_0 of the system is specified by a maximum-likelihood estimation on initial words in the training corpus.

2 Calculations

In this section we provide details on the calculations we used to considerably simplify the computation of $P_n(\mathbf{U})$, as well as an exact expression for its unconstrained gradient, namely the gradient with respect to the parameters without constraining any matrix U_j to be unitary. For the reader's convenience, we first state the final results.

2.1 Simplified Formula for the Probability

We obtained a formula for $P_n(\mathbf{U})$ in terms of products of $D \times D$ matrices

$$P_n(\mathbf{U}) = \mathrm{Tr}(T^\dagger R T), \tag{4}$$

where $T = T^{(2)}T^{(3)}...T^{(n)}$ and each matrix in the product is given by the entries

$$T^{(k)}_{i,j} = [V(w_{k-1})]_{Ni+w_{k-1},Nj+w_k},$$

where $i, j = 0, ..., D-1$. We used indices starting from 0 for convenience. The matrix R is given by the entries

$$R_{i,j} = (\rho_0)_{Ni+w_1,Nj+w_1}.$$

2.2 The Gradient

Finally, the gradient is defined as follows: expanding $\mathbf{U} \mapsto \mathbf{U} + t\mathbf{Z}$ for small t one gets an expression of the form

$$P_n(\mathbf{U} + t\mathbf{Z}) = P_n(\mathbf{U}) + t\mathrm{Re}(\mathbf{Z}^\dagger \mathbf{G}) + O(t^2), \tag{5}$$

where $\mathbf{Z}^\dagger\mathbf{G} \equiv \sum_{j=1}^p Z_j^\dagger G_j$ and the gradient $\mathbf{G}$ is given in components by a complicated formula. In order to state the formula we first define a few ingredients.

In addition to the matrices R and $T^{(k)}$, we need $D \times ND$ matrices $Q_k,\ k = 1, ..., n$ defined by the entries

$$(Q_k)_{jA} \equiv \delta_{Nj+w_k,A},$$

with indicies $j = 0, ..., D-1$ and $A = 1, ..., DN$. We construct the following truncations of the evolution operators, labelled *lesser* and *greater* products respectively.

$$V^{<j}(w) \equiv \prod_{i=1}^{j-1} U_i^{\alpha_i(w)},$$

$$V^{>j}(w) \equiv \prod_{i=j+1}^{n} U_i^{\alpha_i(w)}.$$

Then, we introduce a spectral decomposition for each matrix U_j as $U_j = S_j D_j S_j^\dagger$, whose existence is guaranteed by the spectral theorem being unitary matrices also normal as well. The diagonal matrices $D_j = \mathrm{diag}(u_1, ..., u_{ND})$ contain the eigenvalues of U_j. Finally, we define the following matrices $C_j(\alpha)$, constructed in terms of these eigenvalues.

$$[C_j(\alpha)]_{AB} = \frac{\overline{u_A}^{\alpha} - \overline{u_B}^{\alpha}}{\overline{u_A} - \overline{u_B}} \text{ if } u_A \neq u_B$$

$$[C_j(\alpha)]_{AB} = \alpha \overline{u_A}^{\alpha-1} \text{ if } u_A = u_B$$

where the overline denotes complex conjugation. We can now state the formula for the gradient components G_j

$$G_j = 2S_j \sum_{k=2}^{n} \Big\{ \Big[S_j^{\dagger} \Big(V^{<j}(w_{k-1})^{\dagger} Q_{k-1}^{T} \Big(\prod_{l=2}^{k-1} T^{(l)} \Big)^{\dagger} R T \Big(\prod_{l=k+1}^{n} T^{(l)} \Big)^{\dagger} Q_k V^{>j}(w_{k-1})^{\dagger} \Big) S_j \Big] \cdot C_j(\alpha_j(w_{k-1})) \Big\} S_j^{\dagger} \tag{6}$$

where $\cdot$ denotes *entrywise* matrix multiplication.

3 Details of the calculations

In this section we proceed to show details of the calculations that lead to equations (4) and (6).

3.1 Calculations for the Probability

Starting from equation (3), we make use of the fact that projectors are idempotent, that is $(\Pi_w^{(2)})^2 = \Pi_w^{(2)}$. Also, in order not to overload the notation, we define $\Pi_k \equiv \Pi_{w_k}^{(2)}$ and $V_k \equiv V(w_2)$. Doubling each projector inside the trace, except the first and last, equation (3) becomes

$$\begin{aligned} P_n(\mathbf{U}) &= \mathrm{Tr}\Big((\Pi_n V_{n-1}^{\dagger} \Pi_{n-1})(\Pi_{n-1} V_{n-2}^{\dagger} \Pi_{n-2}) ... (\Pi_2 V_1^{\dagger} \Pi_1)(\Pi_1 \rho_0 \Pi_1)(\Pi_1 V_1 \Pi_2) ... (\Pi_{n-1} V_{n-1} \Pi_n) \Big) \\ &= \mathrm{Tr}\Big((\Pi_{n-1} V_{n-1} \Pi_n)^{\dagger} (\Pi_{n-2} V_{n-2} \Pi_{n-1})^{\dagger} ... (\Pi_1 V_1 \Pi_2)^{\dagger} (\Pi_1 \rho_0 \Pi_1)(\Pi_1 V_1 \Pi_2) ... (\Pi_{n-1} V_{n-1} \Pi_n) \Big) \\ &= \mathrm{Tr}\Big(((\Pi_1 V_1 \Pi_2) ... (\Pi_{n-1} V_{n-1} \Pi_n))^{\dagger} (\Pi_1 \rho_0 \Pi_1)(\Pi_1 V_1 \Pi_2) ... (\Pi_{n-1} V_{n-1} \Pi_n) \Big). \end{aligned} \tag{7}$$

We are thus led to look at the matrices $(\Pi_{k-1} V_{k-1} \Pi_k)$. In components, using indices $A, B, C, D = 1, ..., DN$ they are given by

$$(\Pi_{k-1} V_{k-1} \Pi_k)_{AB} = (\Pi_{k-1})_{AC} (V_{k-1})_{CD} (\Pi_k)_{DB},$$

using the convention of summing over repeated indices for convenience. In order to work with projectors defines as tensor products we employ composite indices $A \to (i, a)$, where $i = 0, ..., D-1$ and $a = 1, ..., N$. Specifically the map is given by $A = Ni + a$. Doing the same for each index we end up with

$$\begin{aligned} (\Pi_{k-1})_{AC} (V_{k-1})_{CD} (\Pi_k)_{DB} &= \delta_{ik} \delta_{a,w_{k-1}} \delta_{c,w_{k-1}} (V_{k-1})_{Nk+c,Nl+d} \delta_{lj} \delta_{d,w_k} \delta_{b,w_k} \\ &= (V_{k-1})_{Ni+w_{k-1},Nj+w_k} \delta_{a,w_{k-1}} \delta_{b,w_k} \\ &\equiv T_{ij}^{(k)} \delta_{a,w_{k-1}} \delta_{b,w_k} \\ &\equiv T_{ij}^{(k)} \theta_{ab}^{(k-1,k)}. \end{aligned} \tag{8}$$

It is easy to check that the matrices $\theta^{(k-1,k)}$ satisfy $\theta^{(k-1,k)}\theta^{(k,k+1)} = \theta^{(k-1,k+1)}$ and $\theta^{(k,k)} = \Pi_{w_k}$, the $N \times N$ projector *on the system space only*. The last remaining factor in equation (7) can be computed in the same fashion

$$\begin{aligned}(\Pi_1\rho_0\Pi_1)_{AB} &= (P_1)_{AC}(\rho_0)_{CD}(\Pi_1)_{DB} \\ &= \delta_{ik}\delta_{a,w_1}\delta_{c,w_1}(\rho_0)_{Nk+c,Nl+d}\delta l j \delta d, w_1\delta b, w_1 \\ &(\rho_0)_{Ni+w_1,Nj+w_1}\delta a, w_1\delta b, w_1 \\ &\equiv R_{ij}(\Pi_{w_1})_{ab}.\end{aligned} \tag{9}$$

We see that the trace in equation (7) factorises into the product of a trace on the D-dimensional ancilla space and a trace on the N-dimensional system space. This trace gives 1, leaving us with equation (4). Substituting equations (8) and (9) into the trace we get

$$\begin{aligned}P_n(\mathbf{U}) &= \mathrm{Tr}\Big(((\Pi_1 V_1\Pi_2)...(\Pi_{n-1}V_{n-1}\Pi_n))^\dagger(\Pi_1\rho_0\Pi_1)(\Pi_1 V_1\Pi_2)...(\Pi_{n-1}V_{n-1}\Pi_n)\Big) \\ &= \mathrm{Tr}_D(T^\dagger RT)\mathrm{Tr}_N((\theta^{1,2})\theta^{2,3}...\theta^{n-1,n})^\dagger\Pi_{w_1}(\theta^{1,2})\theta^{2,3}...\theta^{n-1,n}) \\ &= \mathrm{Tr}_D(T^\dagger RT)\mathrm{Tr}_N(\theta^{n,1}\theta^{1,1}\theta^{1,n}) \\ &= \mathrm{Tr}_D(T^\dagger RT)\mathrm{Tr}_N(\theta^{n,n}) \\ &= \mathrm{Tr}_D(T^\dagger RT),\end{aligned} \tag{10}$$

thus recovering equation (4).

3.2 Calculations for the Gradient

The calculation of the gradient is more involved. It can be broken down in various steps, owing to the complicated functional dependence of the probability on the parameters: a trace of products of submatrices of products of (arbitrary real!) powers of the U_j matrices. In the following, to compute the generic component G_j of the gradient, we perform the Taylor expansion (5) deforming only $U_j \mapsto U_j + tZ$, leaving all the other matrices invariant. This means that, in the many products that appear in the formulae, one is forced to employ what essentially is Leibniz's rule, in the schematic form

$$\prod_{j=1}^{n}(A_j + tB_j) = \prod_{j=1}^{n}A_j + t\sum_{k=1}^{n}\left(\prod_{j<k}A_j\right)B_k\left(\prod_{j>k}A_j\right) + O(t^2). \tag{11}$$

This is where much of the structure in equation (6) comes from. We will use this formula a lot in the computation.

3.2.1 Setup

The 'lowest layer' of dependence is also the most complicated to Taylor expand for small deformations of U_j: real powers of the form $U_j^{\alpha_j(w)}$, which appear in each evolution operator in products, are not trivial to differentiate or Taylor expand when U_j is a matrix. Indeed, the expansion involved a double Taylor series and various rearrangings, as well as the spectral decompositions for the U_j. This step is where the entrywise product in equation (6) and the matrices $C_j(\alpha)$

show up. For now, let us first simply denote the result of the expansion of these matrix powers as

$$U_j^{\alpha_j(w)} = U_j^{\alpha_j(w)} + tB_j(w) + O(t^2), \tag{12}$$

so that we can use the matrices $B_j(w)$ to compute the gradient. The last step will be the computation of $B_j(w)$.

3.2.2 Derivation of the formula

We now move on to expanding the formula for $P_n(\mathbf{U})$ for small deformations $U_j \mapsto U_j + tZ$, leaving all the other matrices $U_i,\ i \neq j$ invariant. We use the matrices $B_j(w)$ defined in (12) before moving on to their computation for clarity. Given $B_j(w)$, the remaining steps are in fact nothing but expanding products of matrices and then taking the trace. In order to put the final result in the canonical form (5) we will also need to make use of properties of the trace.

The first ingredient to expand are the matrices $T^{(k)}$ which make up the product T. Recall that each matrix $T^(k)$ is a submatrix of V_{k-1}, which is expanded to

$$\begin{aligned} V_{k-1} = \prod_{i=1}^{p} U_i^{\alpha_i(w_{k-1})} &\mapsto \prod_{i=1}^{p} U_i^{\alpha_i(w_{k-1})} + t\left(\prod_{i<j} U_i^{\alpha_i(w_{k-1})}\right) B_j(w_{k-1}) \left(\prod_{i>j} U_i^{\alpha_i(w_{k-1})}\right) + O(t^2) \\ &= V_{k-1} + tV^{<j}(w_{k-1})B_j(w_{k-1})V^{>j}(w_{k-1}) + O(t^2). \end{aligned} \tag{13}$$

Denoting with square brackets $[V^{<j}(w_{k-1})B_j(w_{k-1})V^{>j}(w_{k-1})]$ the relevant submatrix which defines the variation of $T^{(k)}$, we can proceed to compute the variation of the product $T \mapsto T + t\delta T + O(t^2)$ with the formula (11)

$$T = T^(2)...T^(n) \mapsto T + t\sum_{k=2}^{n}\left(\prod_{l=2}^{k-1} T^{(l)}\right)[V^{<j}(w_{k-1})B_j(w_{k-1})V^{>j}(w_{k-1})]\left(\prod_{l=k+1}^{n} T^{(l)}\right) + O(t^2). \tag{14}$$

In order to work with submatrices it is useful to use the matrices Q_k defined in Section 2. It is in fact easy to check the formula $[M]_{Ni+w_k,Nj+w_l} = [Q_k M Q_l^T]_{ij}$ for any $N \times N$ matrix M). In this fashion we can rewrite the variation, which will be substituted in the following formulae, as

$$\delta T = \sum_{k=2}^{n}\left(\prod_{l=2}^{k-1} T^{(l)}\right) Q_{k-1} V^{<j}(w_{k-1})B_j(w_{k-1})V^{>j}(w_{k-1}) Q_{k-1}^T \left(\prod_{l=k+1}^{n} T^{(l)}\right)$$

Finally, the trace in equation (4) can be expanded in the following fashion:

$$\begin{aligned} \mathrm{Tr}(T^\dagger RT) &\mapsto \mathrm{Tr}(T^\dagger RT) + t\mathrm{Tr}((\delta T)^\dagger)RT) + t\mathrm{Tr}(T^\dagger)R\delta T) + O(t^2) \\ &= \mathrm{Tr}(T^\dagger RT) + t\mathrm{Re}\big(2\mathrm{Tr}((\delta T)^\dagger RT)\big) + O(t^2), \end{aligned} \tag{15}$$

where we used the property $\mathrm{Tr}(A^\dagger B) = \overline{(\mathrm{Tr}(B^\dagger A))}$ in order to put the variation in a form akin to that of equation (5). To do this in a complete fashion we will also make use of the ciclicity $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$, but first we move on to the computation of $B_j(w)$.

3.2.3 Variation of the Matrix Power

What is left is to compute the variation of the real matrix power, which in this section we denote as U^α to avoid notation overload. The key point is that, having to work with matrices, we have to use the definition $U^\alpha = \exp(\alpha \ln U)$, where to define the logarithm we will have to use the analytic continuation of the Taylor series

$$\ln U \equiv \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} (U - \mathbf{I})^n.$$

Convergence of this series in the operator norm puts conditions on U which are not satisfied in general, even at zeroth order in Z: U has eigenvalues of the form $e^{i\theta}$ which lie on the unit circle in the complex plane. Nevertheless we can proceed by working with suitable U, because the end result is free of logarithms and can be thus analytically continued to unitary U. Let us now proceed in the Taylor expansion. We have

$$(U + tZ)^\alpha = \exp(\alpha \ln(U + tZ)),$$

and the logarithm, using (11) and the spectral expansion $U = SDS^\dagger$ with unitary S, is expanded to

$$\begin{aligned} \ln(U + tZ) &= \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} (U - \mathbf{I} + tZ)^n \\ &= \ln U + t \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \sum_{k=1}^{n} (U - \mathbf{I})^{k-1} Z (U - \mathbf{I})^{n-k} + O(t^2) \\ &= \ln U + tS \left[\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \left(\sum_{k=1}^{n} (D - \mathbf{I})^{k-1} (S^\dagger Z S) (D - \mathbf{I})^{n-k} \right) \right] S^\dagger + O(t^2). \end{aligned} \tag{16}$$

We now face the task of evaluating the sum in the round brackets and then the series in the square brackets. The finite sum can be evaluated using index notation, and this is where the entrywise multiplication comes in. Using indices $A, B = 1, ..., DN$ as in our conventions, the AB entry of the matrix sum in round brackets is given by

$$\begin{aligned} \left(\sum_{k=1}^{n} (D - \mathbf{I})^{k-1} (S^\dagger Z S) (D - \mathbf{I})^{n-k} \right)_{AB} &= \sum_{k=1}^{n} (u_A - 1)^{k-1} (S^\dagger Z S)_{AB} (u_B - 1)^{n-k} \\ &= \left(\sum_{k=1}^{n} (u_A - 1)^{k-1} (u_B - 1)^{n-k} \right) (S^\dagger Z S)_{AB}, \end{aligned} \tag{17}$$

which is then expressed as the entrywise product of $(S^\dagger Z S)$ and the matrix in the round brackets. For different eigenvalues $u_A \neq u_B$ we get

$$\begin{aligned}\sum_{k=1}^{n}(u_A-1)^{k-1}(u_B-1)^{n-k} &= \left(\sum_{k=1}^{n}\left(\frac{u_A-1}{u_B-1}\right)^{k-1}\right)(u_B-1)^{n-1}\\ &= \frac{\left(\frac{u_A-1}{u_B-1}\right)^{n}-1}{\left(\frac{u_A-1}{u_B-1}\right)-1}(u_B-1)^{n-1}\\ &= \frac{(u_A-1)^n-(u_B-1)^n}{u_A-u_B},\end{aligned} \tag{18}$$

while for equal eigenvalues $u_A = u_B$ we get what is expected by taking the limit $u_A \to u_B$

$$\sum_{k=1}^{n}(u_A-1)^{k-1}(u_A-1)^{n-k} = n(u_A-1)^{n-1}. \tag{19}$$

This property is preserved when we then sum the series by uniform convergence of the Taylor series (which we then have to analytically continue at the end), as one can verify explicitly, to get the expression in the round brackets. Leaving out the entrywise multiplication with $(S^\dagger Z S)$, the entries of the matrix series are thus

$$\begin{aligned}\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}\left(\sum_{k=1}^{n}(u_A-1)^{k-1}(u_B-1)^{n-k}\right) &= \frac{\ln u_A-\ln u_B}{u_A-u_B} \text{ if } u_A \neq u_B\\ &= \frac{1}{u_A} \text{ if } u_A = u_B\\ &\equiv L_{AB},\end{aligned} \tag{20}$$

and from this it follows that the matrix in square brackets in (16) is the entrywise multiplication $L\cdot(S^\dagger Z S)$. We can now substitute this result in the exponential. Using (11) yet again in the Taylor series of the exponential (which has an infinite radius of convergence), we get

$$\begin{aligned}\exp(\alpha\ln(U+tZ)) &= \exp\Big(\alpha\ln U+\alpha t S\big(L\cdot(S^\dagger Z S)\big)S^\dagger+O(t^2)\Big)\\ &= \sum_{n=0}^{\infty}\frac{\alpha^n}{n!}\Big(\ln U+tS\big(L\cdot(S^\dagger Z S)\big)S^\dagger\Big)^n+O(t^2)\\ &= U^\alpha+t\sum_{n=0}^{\infty}\frac{\alpha^n}{n!}\sum_{k=1}^{n}(\ln U)^{k-1}\Big[S\big(L\cdot(S^\dagger Z S)\big)S^\dagger\Big](\ln U)^{n-k}+O(t^2).\end{aligned} \tag{21}$$

We can perform the same kind of steps we used before to evaluate the series, ending up with entrywise multiplication. Using the spectral expansion for the logarithm, $(\ln U)^k = S(\ln D)^k S^\dagger$, we get

$$\sum_{n=0}^{\infty}\frac{\alpha^n}{n!}\sum_{k=1}^{n}(\ln U)^{k-1}\Big[S\big(L\cdot(S^\dagger Z S)\big)S^\dagger\Big](\ln U)^{n-k} = S\left[\sum_{n=0}^{\infty}\frac{\alpha^n}{n!}\sum_{k=1}^{n}(\ln D)^{k-1}\Big[\big(L\cdot(S^\dagger Z S)\big)\Big](\ln D)^{n-k}\right]S^\dagger \tag{22}$$

and the sum with the diagonal matrix D, in entries, is evaluated to

$$\begin{aligned}\left[\sum_{k=1}^{n}(\ln D)^{k-1}\Big[S\big(L\cdot(S^\dagger ZS)\big)S^\dagger\Big](\ln D)^{n-k}\right]_{AB} &= \left(\sum_{k=1}^{n}(\ln u_A)^{k-1}(\ln u_B)^{n-k}\right)\big(L\cdot(S^\dagger ZS)\big)_{AB}\\ &= \left(\sum_{k=1}^{n}\Big(\frac{\ln u_A}{\ln u_B}\Big)^{k-1}\right)(\ln u_B)^{n-1}\big(L\cdot(S^\dagger ZS)\big)_{AB}\\ &= \left(\frac{(\ln u_A)^n-(\ln u_B)^n}{\ln u_A-\ln u_B}\right)\big(L\cdot(S^\dagger ZS)\big)_{AB} \text{ if } u_A\neq u_B\\ &= n(\ln u_A)^{n-1}\big(L\cdot(S^\dagger ZS)\big)_{AB} \text{ if } u_A=u_B.\end{aligned} \tag{23}$$

Inserting (23) in the series, we finally get

$$U^\alpha \mapsto U^\alpha + tSMS^\dagger + O(t^2), \tag{24}$$

where the matrix M is defined, in entries by,

$$\begin{aligned}M_{AB} &= \left(\frac{u_A^\alpha-u_B^\alpha}{\ln u_A-\ln u_B}\right)L_{AB}(S^\dagger ZS)_{AB} \text{ if } u_A\neq u_B\\ &= \alpha u_A^\alpha L_{AB}(S^\dagger ZS)_{AB} \text{ if } u_A=u_B\\ &= C(\alpha)^\dagger_{AB}(S^\dagger ZS)_{AB} = [C(\alpha)^\dagger\cdot(S^\dagger ZS)]_{AB}\end{aligned} \tag{25}$$

where we used the definition of the matrix $C(\alpha)$, omitting the j index which we restore in tue full formula to denote that we put $U=U_j$ in the above calculations. In order to put the expansion as in (5), we need the adjoint of an entrywise product. It is easy to verify that $(A\cdot B)^\dagger) = A^\dagger\cdot B^\dagger$. This ends the calculation of the variation of the matrix power, which yields the matrices $B_j(w)$ in the form

$$B_j(w) = S_j[C_j(\alpha(w))^\dagger\cdot(S_j^\dagger ZS_j)]S_j^\dagger.$$

3.2.4 Putting everything together

We are finally ready to compute the gradient components G_j. Equation (15) gives us the (first order) variation of $P_n(\mathbf{U})$ as

$$P_n(\mathbf{U}) \mapsto P_n(\mathbf{U}) + t\mathrm{Re}(2\mathrm{Tr}\big((\delta T)^\dagger RT\big) + O(t^2) \tag{26}$$

and we computed

$$\delta T = \sum_{k=2}^{n}\left(\prod_{l=2}^{k-1}T^{(l)}\right)Q_{k-1}V^{<j}(w_{k-1})B_j(w_{k-1})V^{>j}(w_{k-1})Q_{k-1}^T\left(\prod_{l=k+1}^{n}T^{(l)}\right)$$

as well as the matrices $B_j(w) = S_j[C_j(\alpha(w))^\dagger\cdot(S_j^\dagger ZS_j)]S_j^\dagger$. Thus we need to compute the adjoint

$$(\delta T)^\dagger = \sum_{k=2}^{n} \left(\prod_{l=k+1}^{n} T^{(l)} \right)^\dagger Q_{k-1} V^{>j}(w_{k-1})^\dagger B_j(w_{k-1})^\dagger V^{<j}(w_{k-1})^\dagger Q_{k-1}^T \left(\prod_{l=2}^{k-1} T^{(l)} \right)^\dagger$$

and the adjoint of $B_j(w)$, which is given by $B_j(w)^\dagger = S_j[C_j(\alpha(w)) \cdot (S_j^\dagger Z^\dagger S_j)]S_j^\dagger$. The full expression for the variation $(\delta T)^\dagger$ that we shall insert into (26) then reads

$$(\delta T)^\dagger = \sum_{k=2}^{n} \left(\prod_{l=k+1}^{n} T^{(l)} \right)^\dagger Q_{k-1} V^{>j}(w_{k-1})^\dagger S_j[C_j(\alpha(w_{k-1})) \cdot (S_j^\dagger Z^\dagger S_j)] S_j^\dagger V^{<j}(w_{k-1})^\dagger Q_{k-1}^T \left(\prod_{l=2}^{k-1} T^{(l)} \right)^\dagger. \tag{27}$$

The form of equation (27) translates into a variation for the probability of the schematic form

$$\mathrm{Re} \sum_{k=2}^{n} 2\mathrm{Tr}\Big(\mathcal{A}_k \Big[\mathcal{C}_k \cdot (S^\dagger Z^\dagger S) \Big] \mathcal{B}_k \Big),$$

and has to be put into a form where $Z^\dagger$ is to the left of everything else inside the trace. In order to do this we focus on the trace inside the sum. Using index notation, and the fact that the matrices $C_j(\alpha(w_{k-1}))$ (which is represented by $\mathcal{C}_k$ in the schematic expression above) are symmetric, we get

$$\begin{aligned}
\mathrm{Tr}\Big(\mathcal{A}_k \Big[\mathcal{C}_k \cdot (S^\dagger Z^\dagger S) \Big] \mathcal{B}_k \Big) &= \mathrm{Tr}\Big((\mathcal{B}_k \mathcal{A}_k) \Big[\mathcal{C}_k \cdot (S^\dagger Z^\dagger S) \Big] \Big) \\
&= (\mathcal{B}_k \mathcal{A}_k)_{AB} \Big[\mathcal{C}_k \cdot (S^\dagger Z^\dagger S) \Big]_{AB} \\
&= (\mathcal{B}_k \mathcal{A}_k)_{AB} (\mathcal{C}_k)_{BA} (S^\dagger)_{BC} (Z^\dagger)_{CD} (S)_{DA} \\
&= (Z^\dagger)_{CD} (S)_{DA} (\mathcal{B}_k \mathcal{A}_k)_{AB} (\mathcal{C}_k)_{AB} (S^\dagger)_{BC} \\
&= (Z^\dagger)_{CD} (S)_{DA} [(\mathcal{B}_k \mathcal{A}_k) \cdot \mathcal{C}_k]_{BA} (S^\dagger)_{BC} \\
&= \mathrm{Tr}\Big(Z^\dagger S [(\mathcal{B}_k \mathcal{A}_k) \cdot \mathcal{C}_k] S^\dagger \Big).
\end{aligned} \tag{28}$$

All that remains now is to substitute the matrices $\mathcal{A}_k, \mathcal{B}_k$ into the variation, which corresponds to the gradient $G = 2\sum_{k=2}^{n} S[(\mathcal{B}_k \mathcal{A}_k) \cdot \mathcal{C}_k] S^\dagger$ in schematic notation. In particular, these matrices can be read off from (26) and (27) as

$$\mathcal{A}_k = \left(\prod_{l=k+1}^{n} T^{(l)} \right)^\dagger Q_{k-1} V^{>j}(w_{k-1})^\dagger S_j$$

$$\mathcal{B}_k = S_j^\dagger V^{<j}(w_{k-1})^\dagger Q_{k-1}^T \left(\prod_{l=2}^{k-1} T^{(l)} \right)^\dagger RT$$

and substituting in the above formula finally leads to

$$G_j = 2S_j \sum_{k=2}^{n} \Big\{ \Big[S_j^\dagger \Big(V^{<j}(w_{k-1})^\dagger Q_{k-1}^T \Big(\prod_{l=2}^{k-1} T^{(l)} \Big)^\dagger RT \Big(\prod_{l=k+1}^{n} T^{(l)} \Big)^\dagger Q_k V^{>j}(w_{k-1})^\dagger \Big) S_j \Big] \cdot C_j(\alpha_j(w_{k-1})) \Big\} S_j^\dagger \tag{29}$$

which is the equation (6).

References

Ivano Basile and Fabio Tamburini. 2017. Towards quantum language models. In *Proc. of Empirical Methods in Natural Language Processing - EMNLP 2017*.