Iterative Configuration of Programmable Unitary Converter Based on Few-Layer Redundant Multiplane Light Conversion

Yoshitaka Taguchi,[1,*] Yunzhuo Wang,[2] Ryota Tanomura,[1] Takuo Tanemura,[1] and Yasuyuki Ozeki[1]

[1] *Department of Electrical Engineering and Information Systems, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan*

[2] *Preferred Networks Inc. Otemachi Bldg., 1-6-1 Otemachi, Chiyoda-ku, Tokyo 100-0004 Japan*

(Received 25 January 2023; revised 3 April 2023; accepted 6 April 2023; published 1 May 2023)

Programmable unitary photonic devices are emerging as promising tools to implement unitary transformation for quantum information processing, machine learning, and optical communication. These devices typically use a rectangular mesh of Mach-Zehnder interferometers, which has a clear mathematical structure and can be configured deterministically. However, this mesh architecture is sensitive to fabrication errors, and the correction techniques are still under investigation. In contrast, the multiplane light-conversion (MPLC) architecture is more robust against fabrication errors, but a deterministic method for configuring the converter has not yet been developed due to its complex mathematical structure. In this work, we propose a fast iterative configuration method for MPLC, following the mathematical review of the matrix distance and proposal of an alternative norm. We show through numerical simulations that adding a few redundant layers significantly improves the convergence of the MPLC architecture, making it a practical and attractive option. We also consider the effects of finite resolution and cross-talk in phase shifters in our simulations. In addition, we propose a phase-insensitive distance suited for applications using only intensity detections. Our method demonstrates orders of magnitude better accuracy and a 20-fold speedup compared to previous approaches.

DOI: 10.1103/PhysRevApplied.19.054002

I. INTRODUCTION

Programmable unitary transformations implemented on integrated photonic platforms are becoming a powerful tool for a variety of applications, including quantum photonics [1–7], machine learning [8–14], and optical communication [15–19]. Accurate realization of a given unitary transformation is critical, as the fidelity of computational results and the error of optical communication can be significantly affected by the precision of the realized transformation. A common approach to synthesizing unitary transformations is to use a mesh of Mach-Zehnder interferometers (MZIs) known as the Clements architecture [20], which consists of phase shifters and beam splitters (BSs). This architecture is attractive because its mathematical structure is decomposable, allowing the required phase shift in each MZI to be explicitly determined from the given unitary transformation. However, physical implementation artifacts such as deviation in the splitting ratio of BSs can result in errors in the synthesized transformation. These errors can become significant as the number of optical modes increases [21]. Several design proposals have been made in an effort to reduce or eliminate this error, with the goal of achieving a precise, customizable, and fabrication-error-tolerant unitary transformation that can be applied to scalable and reliable applications.

To address the challenge of implementation artifacts in the Clements architecture, several approaches have been proposed. One approach is local error correction, which involves fixing each MZI and can be applied to any MZI-based architecture, but requires prior knowledge of passive and active components [22]. Another approach is the measurement of components with on-chip power monitors, which allows for the calibration of each MZI but also increases the size of the chip and the complexity of wiring [23,24]. Self-configuration and 3-MZI approaches utilize an additional BS to achieve partially perfect linear operation and employ a feedback loop to adjust each phase shift using only output signals [25,26]. While this method allows for infinite scalability, it also increases the size of the circuit and may have issues with stability [27]. It is worth noting that these approaches primarily consider the artifacts of passive BSs in the circuit, and do not sufficiently consider the artifacts of phase shifters, such as cross-talk.

*ytaguchi@ginjo.t.u-tokyo.ac.jp

Another architecture employs a series connection of phase-shifter arrays and unitary transformations to achieve a highly robust universal synthesis of unitary matrices that is resistant to fabrication errors. This architecture, also known as the multiplane light-conversion (MPLC) architecture [28–31], is particularly robust because each unitary transformation can be selected from a wide range of possible unitaries [32–35]. The unitary transformation can be almost any well-known N-mode mixer, which can significantly increase the flexibility and tolerance to fabrication errors. However, configuring the phase shifters in this architecture is challenging, and no explicit configuration method has been known due to its complex mathematical structure. The optimization of this architecture must deal with the many local minima present in its high-dimensional parameter space [34]. As a result, previous reports have relied on heuristic global searches, such as basin hopping and simulated annealing, to configure the phase shifters [34–36]. However, these methods are time consuming and suffer from exponentially increasing search times as the parameter space dimension increases. To address this issue, a machine-learning-based configuration algorithm has been proposed [37]. While this algorithm may offer a solution, it requires an accurate initial estimation of the structure and may result in decreased matrix fidelity if the initial estimation contains errors.

In this research, we present an alternative, fast, and iteratively configurable MPLC architecture that does not require prior knowledge and relies only on output signals. This approach involves adding a few redundant layers to the existing MPLC architecture and using derivative-based optimization with gradient approximation. This additional layer redundancy significantly improves the optimization performance of the MPLC architecture, in contrast to the similar approach used for the Clements architecture [21, 38], which adds a large number of redundant layers. When compared to numerical optimization of the Clements architecture without redundancy [38], our proposed method achieves a 5 orders of magnitude better accuracy with 1/20 fewer iterations for $N = 128$ modes of transformation. Additionally, our proposed method is able to achieve 5 orders of magnitude better accuracy and is 23 times faster in configuration compared to the previous report that used a heuristic algorithm to optimize the MPLC architecture [39].

This paper is structured as follows. Before discussing the main results, we begin by discussing useful general properties of unitary matrix optimization and introducing another distance in Sec. II. One key property we cover is that unitary matrix optimization essentially has no local minima. Additionally, we propose another distance, a phase-insensitive variant of the Frobenius norm, which is invariant under phase shifts at the output modes. Previously, the standard Frobenius norm has even been used in phase-insensitive applications. In Sec. III, we investigate the optimization properties of the MPLC architecture with a few redundant layers of parametrization. While the parametrization of a unitary matrix can cause optimization to fall into local minima, we demonstrate through numerical simulations that these can be effectively avoided by adding a few redundant layers. Our results show that this architecture can be efficiently optimized using well-known local minimization algorithms, such as the gradient-descent algorithm, while the Clements architecture cannot. We also study the statistical properties of convergence. In Sec. IV, we examine practical scenarios, such as when the gradient of the system is not available, only intensity detection is used, and cross-talk between phase shifters exists. We evaluate the impact of gradient approximation and cross-talk on the proposed method, and show that it still performs well, albeit with a reduction in achieved matrix accuracy after optimization or an increase in the number of iteration until convergence. The phase-insensitive distance exhibits similar optimization properties. In Sec. V, we conclude the paper.

II. MATRIX DISTANCE USING THE FROBENIUS NORM

This section presents some general mathematical properties of the Frobenius norm and proposes another distance. We begin by defining the concept of unimodality for functions on the unitary group $U(N)$ and show that the matrix distance using the Frobenius norm exhibits this unimodality. We then clarify the range and expected value of the norm. Additionally, we introduce the phase-insensitive matrix distance for applications that only use intensity detection.

A. Unimodality on $U(N)$

Here, the concept of unimodality for a function on $U(N)$ is introduced. Unimodality is typically defined for probability distributions [40]. For a multivariable function $f : \mathbb{R}^N \to \mathbb{R}$, unimodality is defined through the level set $L(f, \alpha) = \{\mathbf{x} \,|\, f(\mathbf{x}) \leq \alpha, \mathbf{x} \in \mathbb{R}^N\}$ and the convexity of $L(f, \alpha)$ [41,42]. In this paper, we extend this definition to functions on $U(N)$ by considering the path connectedness of $L(f, \alpha)$, as $U(N)$ is not a convex set.

Definition: A function $f : U(N) \to \mathbb{R}$ is called *unimodal* if the level set $L(f, \alpha) = \{X | f(X) \leq \alpha, X \in U(N)\}$ is path connected for any $\alpha \in \mathbb{R}$.

In other words, any local minimum of a *unimodal* function on $U(N)$ is also a global minimum.

B. Unimodality of the Frobenius norm

We prove that the unitary matrix distance using the Frobenius norm is unimodal. The distance between two unitary matrices $d(X, U)$ is defined as $\|X - U\|_F$, where

$\|A\|_F = \sqrt{\mathrm{Tr}\left[A^\dagger A\right]}$ is the Frobenius norm. It is worth noting that the matrix distance using mean square error (MSE) $\sum_{i,j} |X_{ij} - U_{ij}|^2$ is equivalent to $d(X,U)^2$. Given a unitary matrix $U \in U(N)$, we show that the function $f_U : U(N) \to \mathbb{R}$, defined as $f_U(X) = d(X,U)$ is unimodal. First, from the definition of the Frobenius norm, $f_U(X)^2$ is simplified as

$$\begin{aligned} f_U(X)^2 &= \mathrm{Tr}\left[(X-U)^\dagger (X-U)\right] \\ &= 2N - 2\mathrm{Re}\left[\mathrm{Tr}\left[U^\dagger X\right]\right]. \end{aligned} \tag{1}$$

We write the eigenvalues of $U^\dagger X$ as $\lambda_k (1 \le k \le N)$. Since both U and X are unitary, all the eigenvalues λ_k satisfy $|\lambda_k| = 1$. Therefore, the eigenvalues can be written as $\lambda_k = e^{i\theta_k}$, where $-\pi \le \theta_k \le \pi$. Using these eigenvalues, Eq. (1) can be simplified further as

$$f_U(X)^2 = 2N - 2\sum_{k=1}^{N} \cos\theta_k. \tag{2}$$

Equation 2 implies that $f_U(X)^2$ is *unimodal* because $\cos\theta_k$ is unimodal over the range $-\pi \le \theta \le \pi$ and their sum is also unimodal. An algebraic proof of this unimodality is provided in the Appendix.

C. Range and normalization

We derive the range of $f_U(X)^2$ from Eq. (2) and propose a proper normalization for the distance. The maximum of $f_U(X)^2$ is $4N$ if and only if $\theta_k = \pm\pi$ for all k, and the minimum is 0 if and only if $\theta_k = 0$ for all k. In previous studies, $f_U(X)^2$ has been normalized by N [25], $2N$ [38], or N^2 [35]. Here, we propose a normalization by $4N$, which yields $0 \le f_U(X)^2/4N \le 1$. This is a good normalization of the norm with a range from 0 to 1 that is independent of N.

D. Expected value

To calculate the expected value $\mathbb{E}\left[f_U(X)^2/4N\right]$, the distribution of θ_k is considered. If X is sampled from the Haar measure, then $U^\dagger X$ is also Haar random due to the invariance of the Haar measure. As a result, the eigenvalues of $U^\dagger X$ are uniformly distributed on the unit circle $|c| = 1$, and we have $\theta_k \sim U(-\pi,\pi)$. Because

$$\mathbb{E}\left[\cos\theta_k\right] = \int_{-\pi}^{\pi} \frac{1}{2\pi} \cos\theta \, d\theta = 0, \tag{3}$$

the expected value of the second term in the Eq. (2) is 0. We now conclude that $\mathbb{E}\left[f_U(X)^2/4N\right] = 2N/4N = 0.5$. This fact is observed numerically in the initial value of the convergence plots in Sec. III.

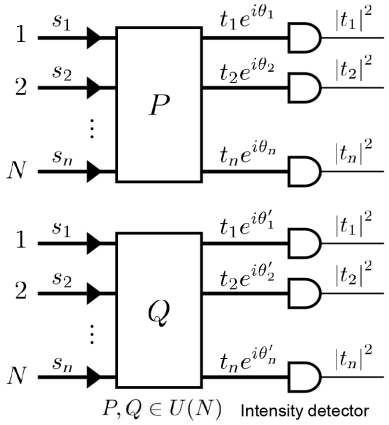

FIG. 1. A complex vector $(s_1, s_2, \ldots, s_n)^\top$ is input into two unitary conversion devices, whose transfer matrices are represented as P and Q. The outputs from these devices are identical, with the exception of the phase degrees of freedom at the outputs, when evaluated using a phase-insensitive distance.

E. Phase-insensitive distance

Here, we introduce a phase-insensitive variant of the matrix distance using the Frobenius norm. This variant is suitable for applications that detect only the intensity of the output modes, as the distance should not be affected by the output phases from the unitary converter. Applications that benefit from this phase-insensitive distance include machine learning and quantum photonics, where photodiodes or photon-number counters are placed at the output ports. Figure 1 shows a scenario where a complex vector $(s_1, s_2, \ldots, s_n)^\top$ is input into two unitary conversion devices. The transfer matrix for these devices is represented by P and Q, and their complex outputs are in polar form as $te^{i\theta}$. The only difference in the output vectors from these two devices is in their phase, with $\theta_i \ne \theta'_i$. In applications that detect only the intensity of output modes, these two matrices P and Q are treated the same and a suitable matrix distance is introduced for this purpose. In the following discussion, the matrix U represents the given target unitary matrix, and the matrix X represents the actual conversion achieved by the unitary converter device.

To investigate the effect of output phases from the unitary converter, we represent the unitary matrices $U^\dagger$ and X as

$$\begin{aligned} U^\dagger &= \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_n \end{bmatrix} \\ X &= \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix}, \end{aligned} \tag{4}$$

where $\mathbf{u}_i$ are column vectors of $U^\dagger$ and $\mathbf{x}_i^\top$ are row vectors of X. Since $U^\dagger$ and X are unitary matrices, the norms of $\mathbf{u}_i$

and $\mathbf{x}_i$ are all equal to 1. In this context, the output phases of the unitary converter correspond to the phases of $\mathbf{x}_i$. The original Frobenius norm changes its value under the global phase change of $\mathbf{x}_i$, which is defined as replacing $\mathbf{x}_i$ with $e^{i\theta_i}\mathbf{x}_i$.

We analyze the dependence of the Frobenius norm on the global phase of each $\mathbf{x}_i$. The Frobenius norm is expanded using the column vectors $\mathbf{u}_i$ as

$$\begin{aligned}\|X-U\|_F^2 &= \|XU^\dagger - I\|_F^2 \\ &= \|X\mathbf{u}_1-\mathbf{e}_1\|^2+\|X\mathbf{u}_2-\mathbf{e}_2\|^2+\cdots+\|X\mathbf{u}_n-\mathbf{e}_n\|^2,\end{aligned} \tag{5}$$

where $\mathbf{e}_i$ is a unit column vector whose ith element is 1 and the others are 0. Expanding the term $\|X\mathbf{u}_1-\mathbf{e}_1\|^2$, we obtain

$$\begin{aligned}&\|X\mathbf{u}_1-\mathbf{e}_1\|^2 \\ &\quad= |\mathbf{x}_1^\top\mathbf{u}_1-1|^2+|\mathbf{x}_2^\top\mathbf{u}_2|^2+\cdots+|\mathbf{x}_n^\top\mathbf{u}_n|^2.\end{aligned} \tag{6}$$

Except for the first term $|\mathbf{x}_1^\top\mathbf{u}_1-1|^2$, the other terms $|\mathbf{x}_i^\top\mathbf{u}_i|^2 (i\geq 2)$ are invariant under the global phase change in $\mathbf{x}_i$. Only the first term depends on the global phase of $\mathbf{x}_1$. Therefore, the term $\|X\mathbf{u}_1-\mathbf{e}_1\|^2$ is invariant under the global phase change in the row vectors $\mathbf{x}_2,\mathbf{x}_3,\ldots\mathbf{x}_n$. Similarly, $\|X\mathbf{u}_i-\mathbf{e}_i\|^2$ is independent of the global phase of $\mathbf{x}_j$ where $j\neq i$. When the global phase of $\mathbf{x}_i$ is changed, the term $|\mathbf{x}_i^\top\mathbf{u}_i-1|^2$ takes its minimum value if and only if $\mathbf{x}_i^\top\mathbf{u}_i$ is a positive real number, because

$$|\mathbf{x}_i^\top\mathbf{u}_i-1|^2 = |\mathbf{x}_i^\top\mathbf{u}_i|^2-2\mathrm{Re}\left[\mathbf{x}_i^\top\mathbf{u}_i\right]+1 \tag{7}$$

and only the second term $2\mathrm{Re}\left[\mathbf{x}_i^\top\mathbf{u}_i\right]$ is dependent on the global phase of $\mathbf{x}_i$. This motivates the idea of replacing all instances of $|\mathbf{x}_i^\top\mathbf{u}_i-1|^2$ with $\left(|\mathbf{x}_i^\top\mathbf{u}_i|-1\right)^2$ in the Frobenius norm $\|X-U\|_F^2$. As discussed in Eqs. (6) and (7), the minimum value of the Frobenius norm under this substitution is the same as the minimum value of the original Frobenius norm, due to the phase-dependence property.

Based on the aforementioned consideration, we propose a phase-insensitive matrix distance. Given unitary matrix U, we define the distance function $h_U : U(N)\to\mathbb{R}$ using only terms whose form is $|\mathbf{x}_i^\top\mathbf{u}_j|$, which can be obtained using intensity measurements at the outputs from the unitary converter. We define positive real numbers $a_{ij}=|\mathbf{x}_i^\top\mathbf{u}_j|^2\geq 0$. All the a_{ij} can be obtained by multiplying column vector $\mathbf{u}_i$ with matrix X through the unitary converter, because

$$X\mathbf{u}_j = \begin{bmatrix}\mathbf{x}_1^\top\mathbf{u}_j\\ \mathbf{x}_2^\top\mathbf{u}_j\\ \vdots\\ \mathbf{x}_n^\top\mathbf{u}_j\end{bmatrix} \tag{8}$$

and $|\mathbf{x}_i^\top\mathbf{u}_j|^2$ can be obtained through the intensity measurement of $\mathbf{x}_i^\top\mathbf{u}_j$. By expanding all the terms in Eq. (5) with Eq. (6) and applying the substitution discussed with respect to Eq. (7), we define $h_U(X)$ as

$$\begin{aligned}&h_U(X) = \\ &(\sqrt{a_{11}}-1)^2+a_{12} &&+\cdots &&+a_{1n}\\ &+a_{21} &&+(\sqrt{a_{22}}-1)^2+\cdots &&+a_{2n}\\ &\vdots &&\vdots &&+(\sqrt{a_{ii}}-1)^2\cdots\\ &+a_{n1} &&+a_{n2} &&+\cdots &&+(\sqrt{a_{nn}}-1)^2\\ &=\sum_{ij}\left(\delta_{ij}-\left|\left[XU^\dagger\right]_{ij}\right|\right)^2.\end{aligned} \tag{9}$$

This function is independent of the global phase of $\mathbf{x}_i$, that is, the phases of each output from the unitary converter, and has a minimum value identical to the original Frobenius norm distance function $f_U(X)$.

The function $h_U(X)$ is also unimodal. This can be proven by reductio ad absurdum. Suppose $h_U(X)$ is not a unimodal function and has multiple disconnected local minima. Then, for all matrices X' such that $h_U(X')$ is a local minimum, there must exist at least one set of phases $(\theta_1,\theta_2,\ldots\theta_n)$ that defines a matrix

$$Y = \begin{bmatrix}e^{i\theta_1}\mathbf{x}_1'^\top\\ e^{i\theta_2}\mathbf{x}_2'^\top\\ \vdots\\ e^{i\theta_n}\mathbf{x}_n'^\top\end{bmatrix} \tag{10}$$

such that $f_U(Y)$ is also a local minimum, where $\mathbf{x}_i'^\top$ is the row vector of X'. The existence of such a matrix Y follows from the definition of $h_U(X)$. However, this would mean that $f_U(X)$ also has multiple disconnected local minima, which is a contradiction. Therefore, $h_U(X)$ must be unimodal.

III. FEW-LAYER REDUNDANT PARAMETERIZATION

The global optimization property guaranteed by unimodality discussed in the previous section is derived without any assumptions about the matrix being optimized; however, the matrix synthesized by the device is parametrized by physical parameters and unimodality may not always hold in this parameter space. We examine this issue and propose a solution to mitigate this difficulty. Let $l_U(X)\geq 0$ be a distance function from a desired unitary matrix U, and $X(\mathbf{p})$ be a unitary matrix realized by a physical converter that is parametrized by a real parameter vector $\mathbf{p}$. Each element in $\mathbf{p}$ corresponds to the amount of phase shift in the actual device. If $l_U(X)$ is unimodal,

its gradient becomes a zero vector only when $X = \pm U$. However, when optimizing an actual unitary conversion device, we need to consider the scalar optimization of $l_U(X(\mathbf{p}))$. If the Jacobian of $X(\mathbf{p})$ is full rank at any $\mathbf{p}$, meaning there exists infinitesimal parameter changes $\Delta\mathbf{p}$ for any infinitesimal matrix changes ΔX, then the function $l_U(X(\mathbf{p}))$ also has a single minimum due to the aforementioned unimodal property of $l_U(X)$. By increasing the number of layers in the unitary converter device, the degree of freedom in the parameter space increases, which may make the Jacobian of $X(\mathbf{p})$ more likely to be full rank. In this section, we demonstrate that increasing the number of layers in unitary converter devices by a few from its minimum requirement significantly improves the optimization of MPLC architecture using a gradient-based optimization algorithm.

A. Device definition and redundancy

We present the mathematical definition of the unitary converters and the few-layer redundant parameterization. Figure 2(a) shows the architecture of the MZI-based unitary converter, which is commonly referred to as Clements architecture [20]. The MZI consists of two $50:50$ BSs and two phase shifters, which can realize an arbitrary $U(2)$ transformation. In this paper, we do not consider any imperfections of the MZI. Figure 2(b) shows the structure of the MPLC architecture. Each layer consists of an N-port fixed unitary converter A_i and an array of N single-mode phase shifters. After m layers, another array of phase shifters is placed in a similar manner to the Clements architecture. The overall transformation of this device, denoted as X, is given by

$$X = L_{m+1}A_mL_m\cdots A_2L_2A_1L_1, \tag{11}$$

where A_i is the transfer matrix of a N-port unitary converter and L_i is expressed as

$$L_i = \begin{bmatrix} e^{i\theta_{i1}} & & & \\ & e^{i\theta_{i2}} & & \\ & & \ddots & \\ & & & e^{i\theta_{in}} \end{bmatrix}. \tag{12}$$

For any $i \neq j$, the matrices A_i and A_j are different. The total number of degrees of freedom in this matrix is $(m+1)(N-1)+1$. This is because each phase-shifter array has $N-1$ degree of freedom due to the loss of one degree of freedom from the global phase, and the entire device has an additional degree of freedom, the global phase. The N-port fixed unitary converter A_i can be implemented using a multiport directional coupler [33], multimode interference coupler [30], or other multiport unitary transform devices. The device should be carefully chosen to ensure that the overall transformation X is universal. The mixing

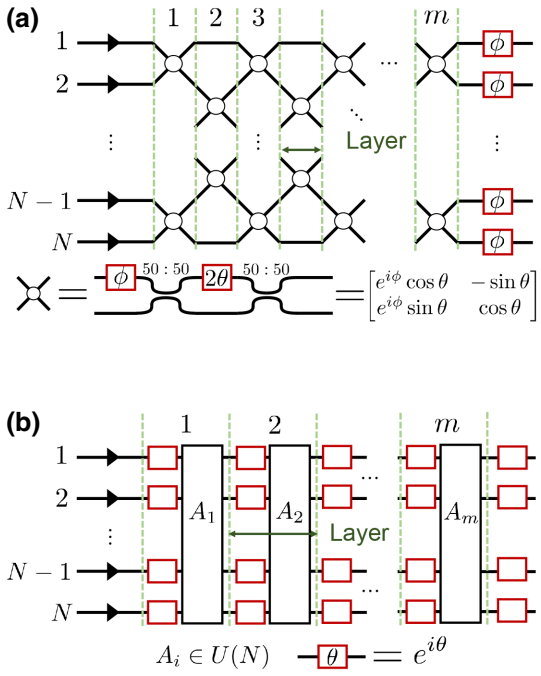

FIG. 2. Schematics of the $N \times m$ Clements architecture (a) and $N \times m$ MPLC architecture (b). The left ports are inputs, and the right ports are outputs. The number of layers in each architecture is specified by m. In the Clements architecture, each layer contains either $N/2$ or $(N-1)/2$ MZI nodes, represented by white circles in the figure. Each MZI node consists of two phase shifters, represented by the variables ϕ and θ. In the MPLC-based unitary converter (b), the architecture consists of an N-port fixed unitary converter represented by A, followed by an array of N single-mode phase shifters.

entropy of a device can be used as a measure of universality [43,44]. To realize an arbitrary $U(N)$ transformation, the total number of degrees of freedom must exceed N^2 [45]. For the Clements architecture, the number of layers m must satisfy $m \geq N$ [20]. Similarly, the number of layers m for the MPLC architecture must also satisfy $m \geq N$, which follows from $(m+1)(N-1)+1 \geq N^2$. In this context, a few-layer redundant parameterized architecture is defined as having $m = N+1, N+2$ layers for both architectures.

B. Optimization problem setting and algorithm

We formulate the matrix optimization problem as follows. We have real parameter variables expressed as a vector $\mathbf{p}$. The number of parameters depends on m and N. We define the normalized cost function $\mathcal{L}$ between two matrices as

$$\mathcal{L}(\mathbf{p}) = \frac{1}{4N}\|X(\mathbf{p}) - U\|_F^2, \tag{13}$$

where $X(\mathbf{p})$ is the unitary matrix realized physically by the parameter vector $\mathbf{p}$, U is the target matrix to be

achieved, and $\|\cdot\|_F$ is the Frobenius norm. The cost function $\mathcal{L}$ is divided by $4N$ as discussed in Sec. II C, and $0 \leq \mathcal{L} \leq 1$ is always satisfied. At the beginning of the optimization, parameters are initialized using uniform distribution ranging from 0 to 2π, and the target unitary matrix U is sampled from the Haar measure using the `stats` module of SciPy [46]. For the MPLC architecture, the matrix A_i for $1 \leq i \leq m$ is also sampled from Haar measure. After initializing the parameters and matrices, the cost function $\mathcal{L}$ is optimized using the quasi-Newton optimization method, limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) [47] implemented in `optimize` module of SciPy [46]. The derivative of the cost function $\mathcal{L}$ required for L-BFGS is calculated using automatic differentiation with the JAX framework [48]. This method starts from the initial parameters and modifies them at each step until convergence to the local minimum, where $d\mathcal{L}/d\mathbf{p} = \mathbf{0}$. In each layer, we have N real parameters to represent N phase shifts in both Clements and MPLC architectures. The optimization is run 64 times while changing the initial parameters to investigate the statistical behavior. Cases with $N = 8$ and $N = 32$ are investigated.

C. Results

Figure 3 shows the convergence plots when the number of layers is changed. The convergence plot of the cost function is recorded for 64 optimization trials. The shaded area shows the range of minimum and maximum values, the dotted line shows the 25% and 75% quantiles, and the solid line shows the median of the trials. For both Clements and MPLC architecture, the insufficient parameterized layer setting results in a large amount of errors. For MPLC architecture, the nonredundant case of $m = N$ results in a large variance of error, especially for $N = 8$. This suggests the presence of many local minima in the parameter space of the MPLC architecture, as previously reported in Ref. [34]. Although the variance of nonredundant setting $m = N$ of $N = 32$ is smaller than that of $N = 8$, the error still remains for $N = 32$, indicating the presence of inevitable local minima for this condition as well. When we increase the number of layers and add redundant degrees of freedom, the variance and error become small, as shown in the cases of $m = N + 1, N + 2$. In contrast, the Clements architecture results in a large variance of error for all conditions, even though it has the sufficient number of degrees of freedom.

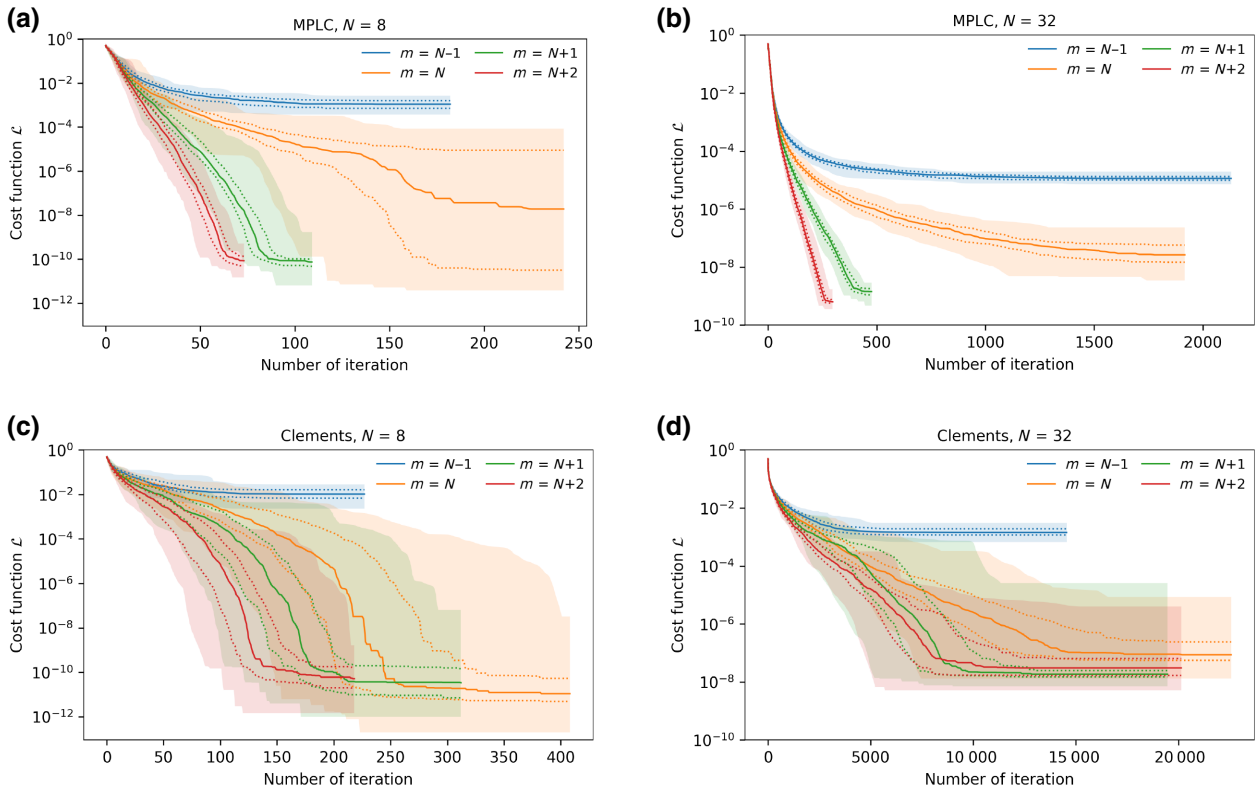

FIG. 3. Convergence plots for MPLC and Clements architectures. The vertical axis shows the value of the cost function $\mathcal{L}$ defined by Eq. (13), and the horizontal axis shows the number of iterations. The shaded area represents the minimum and maximum values, the solid line represents the median, and the dotted line represents the 25% and 75% quantiles over 64 optimization trials. MPLC architecture with (a) $N = 8$, (b) $N = 32$, Clements architecture with (c) $N = 8$, and (d) $N = 32$.

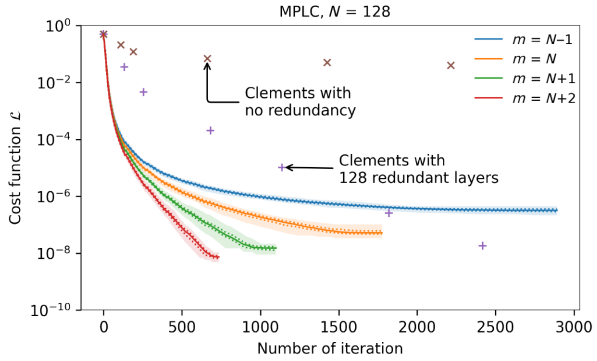

FIG. 4. Comparison of the large case of $N = 128$ with the previous study [38].

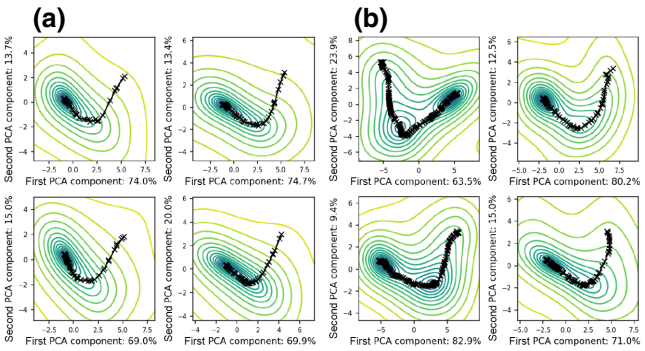

FIG. 5. The PCA projections of the optimization trajectories and contour plot of the log of the loss function for the MPLC architecture (a) and the Clements architecture (b) for $N = 8, m = N + 1$. Each of the four figures shows the trajectory of the optimization when the initial parameters and the target matrix are randomly changed.

For the cases with a large number of ports, $N = 128$, Fig. 4 shows the performance comparison with the previous study [38] of the Clements architecture with redundancy. When compared with no redundancy, the previous study converges at $\mathcal{L} = 1.4 \times 10^{-2}$ after 20 000 iterations, while MPLC architecture yields a result that is 6 orders of magnitude better than the Clements architecture with 1/20 fewer iterations. The MPLC architecture with a redundant layer $m = N + 1$ still outperforms in terms of both convergence speed and accuracy, even compared with the Clements architecture with 128 redundant layers.

We visualize the optimization trajectory and loss function in the parameter space using the method reported in Refs. [49,50] and in the Supplemental Material of Ref. [51]. The optimization trajectory is the path of the parameters in a high-dimensional space created by the optimization. We store the parameter history at each step of the optimization and apply principal component analysis (PCA) to that history. The first and second PCA components are used to project the high-dimensional path onto a two-dimensional space. The visualization is performed for $N = 8, m = N + 1$. Figure 5 shows the projected trajectories and contour plots of the log of the loss function in the projected subspace. The contour plot for the MPLC architecture is like a simple elliptic unimodal function, while that of the Clements architecture is more complex. The difference in the contour plots between these architectures suggests the reason for the convergence plot difference shown in Fig. 3.

IV. OPTIMIZATION UNDER PRACTICAL SETTINGS

We discuss three challenges that must be addressed when applying the proposed method in Sec. III to real device optimization. First, the optimization method used in this section requires the gradient of the target function. While the gradient can be taken physically [52], it requires additional external equipment, which makes the system bulky and not scalable. Second, the optimization method uses the complex amplitudes at the output for optimization. Reading the complex amplitudes in a real device requires coherent detectors at the output, which complicates the device. While some applications require coherent detection, many photonics-based optical computing platforms and quantum computing with photonic chips use intensity detection. Third, real devices have cross-talk between phase shifters, which is not considered in the optimization method. Cross-talk is especially problematic in thermo-optic phase shifters [53,54], although they are attractive due to their small footprints.

In this section, we report the results of derivative-free optimization with the original and phase-insensitive norm, using only the output signal, and examine the effect of cross-talk. We first introduce the mathematical formulation and then present the numerical results. The optimization method used is the same as in Sec. III B. These results pave the way for the design of optical unitary converters without the need for additional components, making the platform more scalable and versatile.

A. Gradient approximation of multivariate function

We use numerical gradient approximation, which uses only function values to approximate the analytical gradient, and investigate the effect of this approximation on the optimization behavior. The gradient of multivariate function $\nabla f\ (x_1, x_2, \ldots, x_n)$ is approximated by

$$\nabla f\ (x_1, x_2, \ldots, x_n) \approx \begin{bmatrix} \frac{f\ (x_1+\Delta, x_2, \ldots, x_n) - f\ (x_1, \ldots, x_n)}{\Delta} \\ \frac{f\ (x_1, x_2+\Delta, \ldots, x_n) - f\ (x_1, \ldots, x_n)}{\Delta} \\ \vdots \\ \frac{f\ (x_1, x_2, \ldots, x_n+\Delta) - f\ (x_1, \ldots, x_n)}{\Delta} \end{bmatrix}, \quad (14)$$

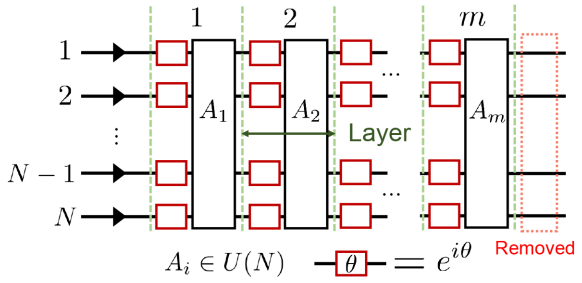

FIG. 6. Schematics of the MPLC architecture with the last phase-shifter array removed.

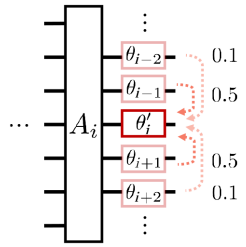

FIG. 7. Cross-talk model of the MPLC architecture.

where $\Delta \ll 1$ represents a finite difference. Calculating gradient approximation requires the same number of function evaluations as the number of parameters. We show that a derivative-based algorithm using such gradient approximation can still be effective for optimizing unitary matrices.

B. Definition of device with intensity detection

We evaluate the optimization behavior of the phase-insensitive distance introduced in Sec. II E, which uses only intensity detectors at the outputs. We expect the phase-insensitive distance to behave similarly to the phase-sensitive distance during optimization because it also has a unimodal property, as shown in Sec. II E. In order to test this, we remove the last phase shifter array from the MPLC architecture, as shown in Fig. 6. Although this removes the N degree of freedom from the architecture, we still expect the optimization behavior to be similar to that of a standard phase-sensitive norm.

C. Model of cross-talk

We model cross-talk by considering the interaction between adjacent phase shifters. The cross-talk is represented by a linear combination of phase shifts, which can be expressed as

$$\theta_i = \sum_j \alpha_{ij}\theta_j . \tag{15}$$

The coupling model and coupling coefficients α_{ij} are shown schematically in Fig. 7. The coupling in the following simulations is formulated as $\theta'_i = 0.1\theta_{i-2} + 0.5\theta_{i-1} + \theta_i + 0.5\theta_{i+1} + 0.1\theta_{i+2}$. If the coupled parameter $\mathbf{p}' = w(\mathbf{p})$ is reversible, the unitary matrix $X(\mathbf{p}')$ realized by the coupled parameter will have a full-rank Jacobian if the original $X(\mathbf{p})$ has a full-rank Jacobian. We use the gradient approximation for optimization.

D. Results

Figure 8 shows the convergence plots for gradient approximation using the standard Frobenius-norm-based distance. The approximation is calculated using a Δ value of 2^{-10}, which corresponds to phase shifts with 10-bit resolution. When a redundant layer is added, the MPLC architecture shows numerical-accuracy limited performance (for $m = N+1, N+2$) with a small variance in the error. Each iteration of the optimization requires the evaluation of the distance the same number of times as the number of parameters due to the gradient approximation. For example, when $N = 8$ and $m = N+1$, each optimization requires $8 \times (8+2) = 80$ evaluations of the distance. Using the MPLC architecture, the optimization converges at 100 iterations for this case, so the total number of evaluations is approximately 8000.

We examine the Δ dependence of the small error variance observed in Fig. 8, which arises from the gradient approximation. Figure 9 shows the optimization results for each gradient approximation accuracy using a redundant layer setting of $m = N+1$. As the finite difference Δ becomes smaller, the final error also becomes smaller. If the accuracy of the gradient approximation is not sufficient, meaning Δ is not small enough, the variance of the optimization result is very small. For example, the error is in the range $[1.5 \times 10^{-5}, 1.8 \times 10^{-5}]$ for the $\Delta = 2^{-6}$ case, $[2.3 \times 10^{-7}, 2.5 \times 10^{-7}]$ for the $\Delta = 2^{-9}$ case, and $[3.5 \times 10^{-9}, 4.5 \times 10^{-9}]$ for the $\Delta = 2^{-12}$ case. If the accuracy of the gradient approximation is sufficient

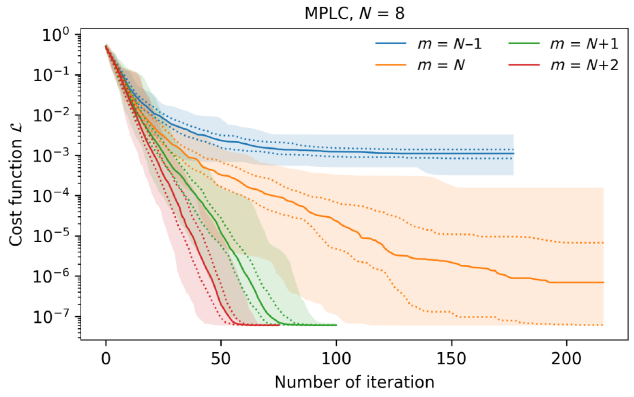

FIG. 8. Convergence plots for the MPLC architecture with gradient approximation, where $\Delta = 2^{-10}$. The 64 optimization trials are shown in the same manner as in Fig. 3.

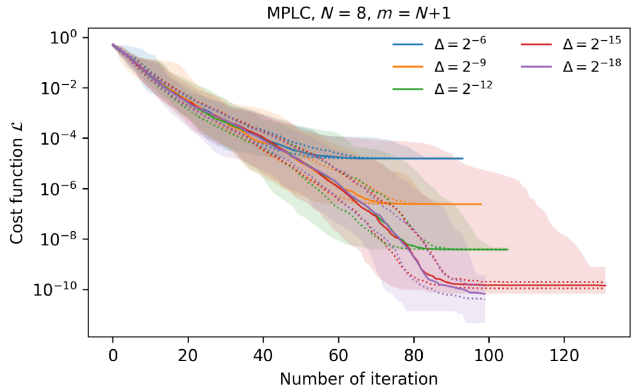

FIG. 9. Comparison of final error when changing the accuracy of gradient approximation for the MPLC architecture. The 64 optimization trials are shown in the same manner as in Fig. 3.

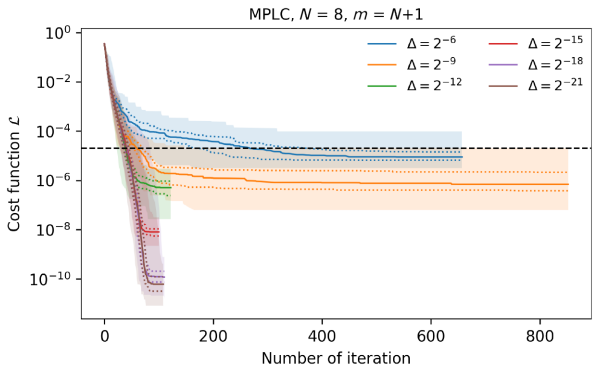

FIG. 11. Comparison of final error when changing the accuracy of gradient approximation for the MPLC architecture using phase-insenstive distance. The black dashed line represents the optimization result in a previous study [35], where $\mathcal{L} = 2.0 \times 10^{-5}$ was achieved.

($\Delta \leq 2^{-15}$), the result has a non-negligible variance similar to the one shown in Fig. 3(a). The error is in the range $\left[7.9 \times 10^{-10}, 6.8 \times 10^{-11}\right]$ for the $\Delta = 2^{-15}$ case and $\left[3.7 \times 10^{-10}, 4.6 \times 10^{-12}\right]$ for the $\Delta = 2^{-18}$ case. These results provide criteria for designing the DAC resolution of a unitary converter system.

We study the optimization property using the phase-insensitive distance. Figure 10 shows the optimization result obtained with an analytical gradient. The convergence plot is similar to the one shown in Fig. 3, in spite of the reduced degree of freedom. However, when using gradient approximation and comparing the accuracy dependence, the phase-insensitive distance shows a different result from the standard Frobenius-norm-based distance, as shown in Fig. 11. All the optimization results have a large variance, as opposed to the cases where $\Delta = 2^{-6}, 2^{-9}, 2^{-12}$ in Fig. 9. The finite difference Δ must be smaller than 2^{-18} to achieve an optimization result comparable to the one obtained using an analytical gradient. The black dashed line shows the optimization result by simulated annealing in a previous study [35]. They achieved an error of $f_{\mathrm{MSE}} = -50$ dB, which corresponds to $\mathcal{L} = 2 \times 10^{-5}$. This error can be achieved using our method with a finite difference of $\Delta \leq 2^{-9}$, and further improvement by orders of magnitude is possible with more accurate gradient approximation. When the accuracy is sufficient, our optimization method converges after about 100 iterations. To compare the speed of our method with a previous experimental report of MPLC architecture optimization using intensity detection [39], we also conduct optimization for a case with $N = 4$ and $m = N + 1$. The optimization converged after about 45 iterations with $\Delta = 2^{-18}$. As each iteration requires $4 \times (4 + 2) = 24$ evaluations, the total number of iterations required for convergence is 1080, representing a 23-fold speedup compared to the previous report (approximately 25 200 evaluations).

The effect of cross-talk when using the phase-insensitive distance and the approximated gradient is shown in Fig. 12. The gradient approximation is calculated using $\Delta = 2^{-12}$. Although cross-talk caused a larger error and increased the

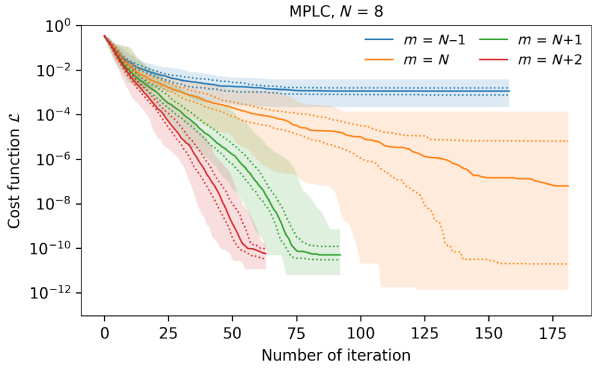

FIG. 10. Convergence plots for the MPLC architecture using phase-insensitive distance with $N = 8$. The 64 optimization trials are shown in the same manner as in Fig. 3.

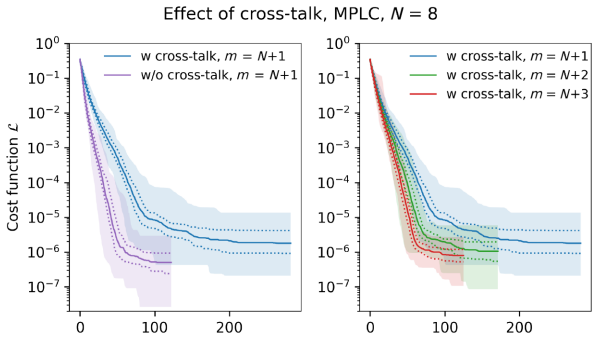

FIG. 12. Comparison of final error under cross-talk using an approximated gradient with $\Delta = 2^{-12}$ and phase-insenstive distance with $N = 8$.

number of iterations until convergence, the performance degradation can be mitigated by adding a few layers of additional redundancy. This result suggests that it may be possible to optimize the device end-to-end, including both matrix optimization and phase-shifter calibration.

V. CONCLUSION

We propose a fast and iteratively configurable MPLC architecture for realizing precise and fabrication-error-tolerant unitary transformation. Our numerical results show that adding a few redundant layers to the MPLC architecture significantly improves optimization behavior. We also examine the effect of artifacts, such as cross-talk, and find that the proposed architecture can be optimized end to end. In addition to proposing an alternative architecture and optimization method, we analyze the distance between unitary matrices using the Frobenius norm. We introduce the concept of unimodality for functions on the unitary group and prove that the matrix distance using the Frobenius norm has this property. We also calculate the expected value and range of the matrix distance. We also introduce the phase-insensitive norm, which is useful for applications that use only intensity detections. We believe that this approach will enable the scalable and robust implementation of optical unitary converters and expand the use of photonic integrated circuits in various fields.

ACKNOWLEDGMENTS

We wish to acknowledge Sho Yasui for a fruitful discussion. This work is supported by JST CREST Grant No. JPMJCR1872, Japan.

APPENDIX: UNIMODALITY OF THE FROBENIUS NORM

Here, we present an algebraic proof of the unimodality of the Frobenius norm. Let $X, U \in U(N)$.

Theorem: *Given U, if $f_U(X) = \|X - U\|_F$ has a local minimum at some X, then it is the global minimum.*

Proof : The squared Frobenius norm can be expressed as $\|X - U\|_F^2 = 2N - 2\mathrm{Re}\left[\mathrm{Tr}\left[U^\dagger X\right]\right]$, and since this is a local minimum, the term $\mathrm{Re}\left[\mathrm{Tr}\left[U^\dagger X\right]\right]$ is a local maximum. Consider $\mathrm{Re}\left[\mathrm{Tr}\left[U^\dagger X\right]\right]$ as a function on the manifold $U(N)$. We can investigate its critical points by examining the directional derivative of the function with respect to the tangent vector X' at X. The tangent vector X' can be represented as $X' = ZX$, where Z is a skew-Hermitian matrix and X is any matrix on the unitary group $U(N)$. This is because the Lie algebra $\mathfrak{u}(N)$ of the unitary group $U(N)$ is composed of skew-Hermitian matrices. (Another proof for $X' = ZX$. Consider an identity $XX^\dagger = I$. Taking the derivative of both sides, we get $X'X^\dagger + X(X^\dagger)' = O$. Then we can rewrite it as $X'X^\dagger = -X(X^\dagger)' = -(X'X^\dagger)^\dagger$. Let $Z = X'X^\dagger$. Then, we have $Z = -Z^\dagger$ which indicates Z is a skew-Hermitian matrix. Since $Z = X'X^\dagger$, we conclude that $X' = ZX$.) When X is at the critical point, then $\mathrm{Re}\left[\mathrm{Tr}\left[U^\dagger X'\right]\right] = 0$ is satisfied for any X'. Substituting $X' = ZX$, we obtain

$$\mathrm{Re}\left[\mathrm{Tr}\left[U^\dagger X'\right]\right] = \mathrm{Re}\left[\mathrm{Tr}\left[ZXU^\dagger\right]\right] = 0. \tag{A1}$$

To further expand this equation, we consider two sets of special matrices Z^1 and Z^2, whose matrix $Z^1_{ij} \in Z^1, Z^2_{ij} \in Z^2$ is indexed by $1 \le i,j \le N$ with $i \ne j$. The (k,l)th element of these matrices $[\cdot]_{kl}$ is defined as follows:

$$[Z^1_{ij}]_{kl} = \delta_{ik}\delta_{jl} - \delta_{il}\delta_{jk}, \tag{A2}$$

$$[Z^2_{ij}]_{kl} = i(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}). \tag{A3}$$

For example, each matrix set includes the following matrices:

$$Z^1_{12} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ -1 & 0 & & & \\ 0 & & & & \\ \vdots & & & & \vdots \\ 0 & & & \cdots & 0 \end{bmatrix}$$

$$Z^2_{23} = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & i & 0 & & \\ 0 & i & 0 & & & \\ 0 & 0 & & & & \\ \vdots & & & & & \vdots \\ 0 & & & & \cdots & 0 \end{bmatrix}. \tag{A4}$$

After substituting $Z^1_{ij}, Z^2_{ij}, 1 \le i,j \le N$ for Z in Eq. (A1), we obtain

$$\begin{cases} \mathrm{Re}\left[\left[XU^\dagger\right]_{ij}\right] - \mathrm{Re}\left[\left[XU^\dagger\right]_{ji}\right] = 0 \\ \mathrm{Im}\left[\left[XU^\dagger\right]_{ij}\right] + \mathrm{Im}\left[\left[XU^\dagger\right]_{ji}\right] = 0 \end{cases}, \tag{A5}$$

which leads to $XU^\dagger = (XU^\dagger)^\dagger$. Therefore,

$$(XU^\dagger)^2 = I. \tag{A6}$$

The unitary matrix $XU^\dagger$ can be diagonalized using a regular matrix V, and a diagonal matrix D whose diagonal elements $d_i \in \mathbb{C}$ satisfies $|d_i| = 1$ because $XU^\dagger$ is a unitary matrix. We can express this diagonalization as $XU^\dagger = VDV^{-1}$. Substituting $XU^\dagger$ in Eq. (A6), we obtain $D = D^\dagger$. Since $|d_i| = 1$, we have $d_i = \pm 1$. Now consider the original local maximum term $\mathrm{Re}\left[\mathrm{Tr}\left[U^\dagger X\right]\right]$. We can rewrite it as $\mathrm{Re}\left[\mathrm{Tr}\left[U^\dagger X\right]\right] = \mathrm{Re}\left[\mathrm{Tr}\left[VDV^{-1}\right]\right] = \mathrm{Re}\left[\mathrm{Tr}\left[D\right]\right] = \mathrm{Re}\left[\sum_i d_i\right]$. This value is obviously maximized when $d_i =$

$+1$ for all i. If $d_i = -1$ for some i, we can rotate this value to $d_i = +1$ along the unit circle $|c| = 1$ in the complex plane and still achieve the maximum value. Therefore, if $\mathrm{Re}\left[\mathrm{Tr}\left[U^\dagger X\right]\right]$ is a local maximum, it is also a global maximum. As a result, we now conclude that if $\|X - U\|_F \geq 0$ is a local minimum, then it must also be a global minimum. ■

[1] J. Carolan, C. Harrold, C. Sparrow, E. Martín-López, N. J. Russell, J. W. Silverstone, P. J. Shadbolt, N. Matsuda, M. Oguma, M. Itoh, G. D. Marshall, M. G. Thompson, J. C. F. Matthews, T. Hashimoto, J. L. O'Brien, and A. Laing, Universal linear optics, Science **349**, 711 (2015).

[2] J. Wang, F. Sciarrino, A. Laing, and M. G. Thompson, Integrated photonic quantum technologies, Nat. Photonics **14**, 273 (2020).

[3] A. W. Elshaari, W. Pernice, K. Srinivasan, O. Benson, and V. Zwiller, Hybrid integrated quantum photonic circuits, Nat. Photonics **14**, 285 (2020).

[4] J. Carolan, M. Mohseni, J. P. Olson, M. Prabhu, C. Chen, D. Bunandar, M. Y. Niu, N. C. Harris, F. N. C. Wong, M. Hochberg, S. Lloyd, and D. Englund, Variational quantum unsampling on a quantum photonic processor, Nat. Phys. **16**, 322 (2020).

[5] E. Pelucchi, G. Fagas, I. Aharonovich, D. Englund, E. Figueroa, Q. Gong, H. Hannes, J. Liu, C.-Y. Lu, N. Matsuda, J.-W. Pan, F. Schreck, F. Sciarrino, C. Silberhorn, J. Wang, and K. D. Jöns, The potential and global outlook of integrated photonics for quantum technologies, Nat. Rev. Phys. **4**, 194 (2022).

[6] Y. Chi, J. Huang, Z. Zhang, J. Mao, Z. Zhou, X. Chen, C. Zhai, J. Bao, T. Dai, H. Yuan, M. Zhang, D. Dai, B. Tang, Y. Yang, Z. Li, Y. Ding, L. K. Oxenløwe, M. G. Thompson, J. L. O'Brien, Y. Li, Q. Gong, and J. Wang, A programmable qudit-based quantum processor, Nat. Commun. **13**, 1166 (2022).

[7] L. S. Madsen, F. Laudenbach, M. F. Askarani, F. Rortais, T. Vincent, J. F. F. Bulmer, F. M. Miatto, L. Neuhaus, L. G. Helt, M. J. Collins, A. E. Lita, T. Gerrits, S. W. Nam, V. D. Vaidya, M. Menotti, I. Dhand, Z. Vernon, N. Quesada, and J. Lavoie, Quantum computational advantage with a programmable photonic processor, Nature **606**, 75 (2022).

[8] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, Deep learning with coherent nanophotonic circuits, Nat. Photonics **11**, 441 (2017).

[9] M. Prabhu, C. Roques-Carmes, Y. Shen, N. Harris, L. Jing, J. Carolan, R. Hamerly, T. Baehr-Jones, M. Hochberg, V. Čeperić, J. D. Joannopoulos, D. R. Englund, and M. Soljačić, Accelerating recurrent Ising machines in photonic integrated circuits, Optica **7**, 551 (2020).

[10] H. Zhang, M. Gu, X. D. Jiang, J. Thompson, H. Cai, S. Paesani, R. Santagati, A. Laing, Y. Zhang, M. H. Yung, Y. Z. Shi, F. K. Muhammad, G. Q. Lo, X. S. Luo, B. Dong, D. L. Kwong, L. C. Kwek, and A. Q. Liu, An optical neural chip for implementing complex-valued neural network, Nat. Commun. **12**, 457 (2021).

[11] S. Pai, Z. Sun, T. W. Hughes, T. Park, B. Bartlett, I. A. D. Williamson, M. Minkov, M. Milanizadeh, N. Abebe, F. Morichetti, A. Melloni, S. Fan, O. Solgaard, and D. A. B. Miller, Experimentally realized in situ backpropagation for deep learning in nanophotonic neural networks, arXiv:2205.08501.

[12] F. Ashtiani, A. J. Geers, and F. Aflatouni, An on-chip photonic deep neural network for image classification, Nature **606**, 501 (2022).

[13] S. Ohno, R. Tang, K. Toprasertpong, S. Takagi, and M. Takenaka, Si microring resonator crossbar array for on-chip inference and training of the optical neural network, ACS Photonics **9**, 2614 (2022).

[14] S. Bandyopadhyay, A. Sludds, S. Krastanov, R. Hamerly, N. Harris, D. Bunandar, M. Streshinsky, M. Hochberg, and D. Englund, Single chip photonic deep neural network with accelerated training, arXiv:2208.01623.

[15] N. K. Fontaine, C. R. Doerr, M. A. Mestre, R. R. Ryf, P. J. Winzer, L. L. Buhl, Y. Sun, X. Jiang, and R. Lingle, in *OFC/NFOEC* (IEEE, Los Angeles, CA, USA, 2012), pp. 1–3.

[16] A. Annoni, E. Guglielmi, M. Carminati, G. Ferrari, M. Sampietro, D. A. Miller, A. Melloni, and F. Morichetti, Unscrambling light—automatically undoing strong mixing between modes, Light: Sci. Appl. **6**, e17110 (2017).

[17] D. Melati, A. Alippi, A. Annoni, N. Peserico, and A. Melloni, Integrated all-optical MIMO demultiplexer for mode- and wavelength-division-multiplexed transmission, Opt. Lett. **42**, 342 (2017).

[18] K. Choutagunta, I. Roberts, D. A. B. Miller, and J. M. Kahn, Adapting Mach–Zehnder mesh equalizers in direct-detection mode-division-multiplexed links, J. Lightwave Technol. **38**, 723 (2020).

[19] R. Tanomura, R. Tang, G. Soma, S. Ishimura, T. Tanemura, and Y. Nakano, in *2022 European Conference on Optical Communication (ECOC)* (IEEE, Basel, Switzerland, 2022), pp. 1–4.

[20] W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walmsley, Optimal design for universal multiport interferometers, Optica **3**, 1460 (2016).

[21] R. Burgwal, W. R. Clements, D. H. Smith, J. C. Gates, W. S. Kolthammer, J. J. Renema, and I. A. Walmsley, Using an imperfect photonic network to implement random unitaries, Opt. Express **25**, 28236 (2017).

[22] S. Bandyopadhyay, R. Hamerly, and D. Englund, Hardware error correction for programmable photonics, Optica **8**, 1247 (2021).

[23] D. A. B. Miller, Self-configuring universal linear optical component [Invited], Photon. Res. **1**, 1 (2013).

[24] D. A. B. Miller, Setting up meshes of interferometers—reversed local light interference method, Opt. Express **25**, 29233 (2017).

[25] R. Hamerly, S. Bandyopadhyay, and D. Englund, Asymptotically fault-tolerant programmable photonics, Nat. Commun. **13**, 6831 (2022).

[26] R. Hamerly, S. Bandyopadhyay, and D. Englund, Accurate Self-Configuration of Rectangular Multiport Interferometers, Phys. Rev. Appl. **18**, 024019 (2022).

[27] R. Hamerly, S. Bandyopadhyay, and D. Englund, Stability of Self-Configuring Large Multiport Interferometers, Phys. Rev. Appl. **18**, 024018 (2022).

[28] J.-F. Morizur, L. Nicholls, P. Jian, S. Armstrong, N. Treps, B. Hage, M. Hsu, W. Bowen, J. Janousek, and H.-A. Bachor, Programmable unitary spatial mode manipulation, J. Opt. Soc. Am. A **27**, 2524 (2010).

[29] G. Labroille, B. Denolle, P. Jian, P. Genevaux, N. Treps, and J.-F. Morizur, Efficient and mode selective spatial mode multiplexer based on multi-plane light conversion, Opt. Express **22**, 15599 (2014).

[30] R. Tang, T. Tanemura, and Y. Nakano, Integrated reconfigurable unitary optical mode converter using MMI couplers, IEEE Photonics Technol. Lett. **29**, 971 (2017).

[31] R. Tang, T. Tanemura, S. Ghosh, K. Suzuki, K. Tanizawa, K. Ikeda, H. Kawashima, and Y. Nakano, Reconfigurable all-optical on-chip MIMO three-mode demultiplexing based on multi-plane light conversion, Opt. Lett. **43**, 1798 (2018).

[32] R. Tang, T. Tanemura, and Y. Nakano, in *2017 Opto-Electronics and Communications Conference (OECC) and Photonics Global Conference (PGC)* (IEEE, Singapore, 2017), pp. 1–3.

[33] R. Tanomura, R. Tang, S. Ghosh, T. Tanemura, and Y. Nakano, Robust integrated optical unitary converter using multiport directional couplers, J. Lightwave Technol. **38**, 60 (2020).

[34] M. Y. Saygin, I. V. Kondratyev, I. V. Dyakonov, S. A. Mironov, S. S. Straupe, and S. P. Kulik, Robust Architecture for Programmable Universal Unitaries, Phys. Rev. Lett. **124**, 010501 (2020).

[35] R. Tanomura, R. Tang, T. Umezaki, G. Soma, T. Tanemura, and Y. Nakano, Scalable and Robust Photonic Integrated Unitary Converter Based on Multiplane Light Conversion, Phys. Rev. Appl. **17**, 024071 (2022).

[36] R. Tanomura, R. Tang, T. Tanemura, and Y. Nakano, Integrated InP optical unitary converter with compact half-integer multimode interferometers, Opt. Express **29**, 43414 (2021).

[37] S. Kuzmin, I. Dyakonov, and S. Kulik, Architecture agnostic algorithm for reconfigurable optical interferometer programming, Opt. Express **29**, 38429 (2021).

[38] S. Pai, B. Bartlett, O. Solgaard, and D. A. B. Miller, Matrix Optimization on Universal Unitary Photonic Devices, Phys. Rev. Appl. **11**, 064044 (2019).

[39] R. Tanomura, R. Tang, T. Suganuma, K. Okawa, E. Kato, T. Tanemura, and Y. Nakano, Monolithic InP optical unitary converter based on multi-plane light conversion, Opt. Express **28**, 25392 (2020).

[40] E. W. Weisstein, Unimodal, from *mathworld* – A wolfram web resource, https://mathworld.wolfram.com/Unimodal.html.

[41] T. W. Anderson, The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities, Proceedings of the American Mathematical Society **6**, 170 (1955).

[42] G. Anescu, A heuristic fast gradient descent method for unimodal optimization, J. Adv. Math. Comput. Sci. **26**, 1 (2018).

[43] R. Tang, R. Tanomura, T. Tanemura, and Y. Nakano, Ten-port unitary optical processor on a silicon photonic chip, ACS Photonics **8**, 2074 (2021).

[44] R. Tanomura, Y. Taguchi, R. Tang, T. Tanemura, and Y. Nakano, in *Conference on Lasers and Electro-Optics/Pacific Rim (CLEOPR) 2022* (Optica Publishing Group, 2022), p. CWP13A–02.

[45] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, Experimental Realization of any Discrete Unitary Operator, Phys. Rev. Lett. **73**, 58 (1994).

[46] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, SciPy 1.0: Fundamental algorithms for scientific computing in Python, Nat. Methods **17**, 261 (2020).

[47] R. Fletcher, *Practical Methods of Optimization* (John Wiley & Sons, New York, NY, USA, 1987), 2nd ed.

[48] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, JAX: Composable transformations of Python+NumPy programs (2018).

[49] M. Gallagher and T. Downs, Visualization of learning in multilayer perceptron networks using principal component analysis, IEEE Trans. Syst. Man Cybern. B (Cybern.) **33**, 28 (2003).

[50] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, Visualizing the loss landscape of neural nets, arXiv:1712.09913.

[51] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, in *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018), Vol. 31.

[52] T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, Training of photonic neural networks through in situ backpropagation and gradient measurement, Optica **5**, 864 (2018).

[53] M. Jacques, A. Samani, E. El-Fiky, D. Patel, Z. Xing, and D. V. Plant, Optimization of thermo-optic phase-shifter design and mitigation of thermal crosstalk on the soi platform, Opt. Express **27**, 10456 (2019).

[54] B. V. Gurses, R. Fatemi, A. Khachaturian, and A. Hajimiri, Large-scale crosstalk-corrected thermo-optic phase shifter arrays in silicon photonics, IEEE J. Sel. Top. Quantum Electron. **28**, 6101009 (2022).