Sliced Wasserstein Meets Quantum Optics: Provable Wavefunction Tomography with Scarce Noisy Measurements

Takahiro Kajisa
Department of Informatics
The University of Electro-Communications
Tokyo, Japan
t.kajisa@uec.ac.jp

Abstract

In quantum optics, wavefunction tomography estimates a rank-one state of light, described by a single L^2 function, from measurement data. We consider homodyne detection, the most basic interferometric measurement, which samples the intensity distributions of fractional Fourier transforms of the target wavefunction at several angles. However, measurements destroy the state and the setup drifts, so such data are scarce and noisy. We therefore perform this tomography by parameterizing the wavefunction with an overparameterized neural network and minimizing the max-sliced Wasserstein-1 distance to the measured distributions by infinite-dimensional rank-1 factored gradient descent. This loss compares distributions on the line directly and reduces the analysis to classical probability theory. In this setting, we prove that every critical point at which the loss is differentiable is a global minimum. Empirically, on noisy two-mode cat states, our method keeps improving fidelity beyond the noise floor of iterative maximum-likelihood estimation.

1 Introduction

Quantum optics is a leading physical platform for quantum information processing, metrology, and communication (Braunstein & van Loock, 2005; Weedbrook et al., 2012). At the core of this platform are pure non-Gaussian states. Because they unlock the limitations of classically simulable Gaussian operations and realize universal and fault-tolerant quantum computation, they are critically important in both theoretical and experimental domains (Bartlett et al., 2002; Mari & Eisert, 2012). A prime example is the Schrödinger cat state, engineered from weak coherent states. Cat states drive both foundational physics and advanced applications, from testing CV Bell inequalities to enabling bosonic error correction and Heisenberg-limited metrology (Ourjoumtsev et al., 2006, 2007, 2009; Joo et al., 2011; Gottesman et al., 2001; Mirrahimi et al., 2014; Ofek et al., 2016). Translating these applications into practice relies entirely on our ability to precisely benchmark experimentally generated states against theoretical targets via quantum state tomography. Yet, this process faces severe fundamental bottlenecks. First, quantum states are fragile resources irreversibly destroyed by measurement, strictly limiting the information gained per shot. Second, compensating for this by extending measurement times inevitably invites experimental drift, such as optical path length fluctuations and local oscillator phase instability, breaking the assumption of identically prepared samples. As a result, tomography must operate on inherently scarce and noisy data. Hence, overcoming this data deficit to achieve high-fidelity state estimation remains a pivotal challenge for the field.

Quantum state tomography in continuous-variable optics builds upon the foundational frameworks of (Bertrand & Bertrand, 1987; Vogel & Risken, 1989). Through homodyne detection, which records interference against a local oscillator at phase θ, one samples the quadrature amplitude x from a

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

probability distribution $p_\theta(x)$. This constitutes the Radon transforms that are marginal projections of the quantum state's Wigner function $f(q, p)$ along θ (Radon, 1917; Helgason, 1999). Inverting this transform to reconstruct the Wigner function is famously ill-posed, suffering from non-uniqueness and numerical instability when subject to noise and limited measurement angles. To regularize this problem, iterative maximum likelihood estimation (iMLE) has become the de facto standard, rigorously enforcing physical constraints (positive semi-definiteness and unit trace) while guaranteeing monotonic convergence (Hradil, 1997; Lvovsky, 2004; Řeháček et al., 2007; Lvovsky & Raymer, 2009). However, iMLE reaches a critical practical bottleneck. In low-shot regimes, information deficits cause rapid performance decay. Conversely, attempting to construct a sufficiently rich dataset inevitably introduces experimental drift and noise accumulation, to which iMLE is remarkably sensitive. This misspecifies the exact likelihood function, and traps the estimation fidelity at the noise floor. Ultimately, existing techniques cannot escape the inherent experimentalist's dilemma: extending measurements to acquire more data inevitably forces the system to drift, undermining the measurement itself.

To bypass these bottlenecks, machine learning for quantum state tomography has been vigorously explored. Architectures such as restricted Boltzmann machines (Carleo & Troyer, 2017; Torlai et al., 2018; Tiunov et al., 2020), variational autoencoders (Rocchetto et al., 2018), GANs (Ahmed et al., 2021), and attention mechanisms (Cha et al., 2022) frequently outperform traditional methods in expressivity and computational efficiency. However, two fundamental flaws persist: First, the optimization lacks theoretical guarantees. Because non-convex landscapes are often riddled with spurious local minima and saddle points, conditions for global convergence remain elusive, leaving current successes largely empirical. Second, these methods fail to fully leverage the intrinsic probabilistic geometry of homodyne measurements. By relying on 2D histograms or phenomenological L^2 losses, these approaches divorce the loss function from the underlying Radon transform structure, ignoring that data inherently represent samples from directional marginal distributions. In low-shot regimes, this structural mismatch exacerbates information loss and generates artificial saddle points. Compounded by the absence of theoretical guarantees, this causes optimization failures.

Contributions. We introduce a feedforward neural network (FNN)-based tomography framework that simultaneously solves the aforementioned gaps. The theoretical schematic of this study is illustrated in Figure 1, and our specific contributions are summarized as follows:

1° **Global Convergence Guarantees.** By restricting the target to pure state as L^2 functions, we directly parameterize it via an FNN. This constitutes an infinite-dimensional rank-1 factored gradient descent, whose landscape analysis requires new techniques that have no finite-dimensional counterpart, specifically the control of the zero structure of fractional Fourier transforms. Provided the network is overparameterized and the target state possesses sufficient regularity, we rigorously prove that the resulting loss landscape is benign on the differentiable regions.

2° **Leveraging Intrinsic Data Geometry.** We adopt a loss function based on the max-sliced Wasserstein-1 (MSW1) distance. This aligns with the physics of homodyne detection, where measurements are exact slices of the state's Wigner function. Beyond its computational efficiency, this approach seamlessly treats quantum states as sets of probability distributions, ensuring high extensibility.

3° **Robust Tomography from Scarce, Noisy Data.** We numerically demonstrate our method on two-mode cat states, a critical continuous-variable quantum resource. We adopt iMLE as our baseline because it has rigorous convergence guarantees. Pushing the measurement angles, shot counts, and photon loss to their extreme limits, we show that our FNN+MSW1 framework works effectively where iMLE collapses. We also compare with a rank-1 constrained iMLE, which lacks a convergence guarantee, and find that it collapses under strong noise just as iMLE does. The rank-1 constraint alone is therefore insufficient, which highlights the importance of combining this hard constraint with a loss congruent with the data structure.

2 Preliminaries and Main Results

2.1 Wigner function

In quantum mechanics, a quantum state serves as the non-commutative analogue of a probability distribution. A quantum state is defined as a positive trace-class operator ρ on a Hilbert space $\mathcal{H}$ with

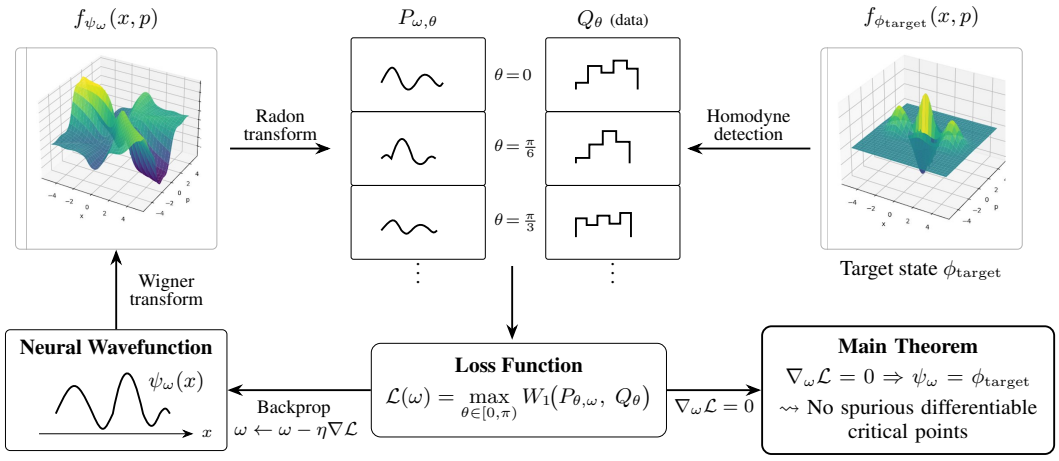


Figure 1: Flowchart of the proposed framework. As shown in the bottom left, a wavefunction is directly parameterized by a feed-forward neural network, and its marginal distributions for each slice are obtained via the Wigner-Radon transform. Simultaneously, homodyne measurements are performed on the target quantum state to acquire the empirical data distributions corresponding to each slice. These distributions are compared using the max-sliced Wasserstein-1 (MSW1) distance, and the network is optimized via gradient descent. Our main theorem guarantees that when this gradient vanishes, the system is strictly at the global minimum, thereby successfully completing the state tomography.

unit trace $\operatorname{tr}\rho = 1$. Among this set of quantum states, the rank-one operators, namely those that can be written as $\rho = \langle\psi,\cdot\rangle\psi$ using an appropriate normalized unit vector $\psi \in \mathcal{H}$ with $\|\psi\|_{L^2} = 1$, are called pure states, while all others are referred to as mixed states. Unless there is a risk of confusion, the vector $\psi \in \mathcal{H}$ is also simply referred to as a pure state. Especially, in continuous-variable quantum optics where the Hilbert space is $\mathcal{H} = L^2(\mathbb{R})$, a pure state is commonly referred to as a wavefunction, denoted by $\psi(x) \in L^2(\mathbb{R})$. Geometrically, pure states form the extreme points of the convex set of quantum states. While any mixed state can be expressed as a convex combination of pure states, a major departure from classical probability theory is that this extreme point decomposition is generally not unique. Beyond their mathematical primacy, pure states are of utmost practical importance, as the ultimate goal of continuous-variable quantum optics experiments is to generate wavefunctions exhibiting non-classical coherence and to harness them as computational resources, for instance. Motivated by both their geometric fundamentalism and practical necessity, in this study, we develop a theory of quantum optical state tomography tailored for such wavefunctions.

One equivalent representation of quantum states relies on harmonic analysis, specifically the Wigner function formalism. The Wigner function $f_\psi : \mathbb{R}^2 \to \mathbb{R}$ associated with a wavefunction ψ is defined as

$$f_\psi(x,p) := \frac{1}{\pi}\int_{-\infty}^{\infty} \psi^*(x+y)\psi(x-y)\, e^{2ipy}\, dy. \tag{1}$$

It is a well-known fact that f_ψ is real-valued, satisfies $\int_{\mathbb{R}^2} f_\psi(x,p)\, dx\, dp = 1$, and is bounded by $|f_\psi| \le 1/\pi$. However, its range is generally not restricted to non-negative values, and this negativity serves as an indicator of non-classicality. Since the Wigner function for a single degree of freedom is a function on a two-dimensional plane, specifying an angular and radial direction allows us to define the Radon transform $\mathcal{R}_\theta$ as follows:

$$P_{\psi,\theta}(x_\theta) := \mathcal{R}_\theta f_\psi(x_\theta) = \iint f_\psi(x,p)\delta(x_\theta - (x\cos\theta + p\sin\theta))dxdp.$$

From the properties of the Wigner function, it is known that $P_{\psi,\theta}$ constitutes a Borel probability measure on $\mathbb{R}$. Furthermore, its density function is given by the fractional Fourier transform as follows:

Proposition 2.1 ((Lohmann & Soffer, 1994)). *Let $\mathcal{F}_\theta$ denote the fractional Fourier transform at angle θ, and define $\psi_\theta = \mathcal{F}_\theta\psi$. Then, for the slice at each $\theta \in [0,\pi)$, the following equality holds:*

$$P_{\psi,\theta}(x_\theta) = |\psi_\theta(x_\theta)|^2. \tag{2}$$

Remark 2.2. A state ψ is called a Gaussian state when $P_{\psi,\theta}$ is a Gaussian distribution for every θ. By Hudson's theorem (Hudson, 1974), a pure state has a non-negative Wigner function if and only if it is a Gaussian state, so every pure non-Gaussian state exhibits Wigner negativity, a hallmark of non-classicality. Pure Gaussian states, moreover, are determined by their first and second moments and can be reconstructed from a few slices without the machinery developed here. Pure non-Gaussian states, such as the cat states considered below, are therefore the natural targets of our method.

2.2 Neural network assisted tomography of quantum optical wavefunctions

To tomograph an experimentally generated quantum state, quantum measurements must be conducted over a sufficient number of configurations. In quantum optics, homodyne detection is frequently employed, wherein a target quantum state ψ is interfered with a local oscillator at a relative phase θ. This is mathematically equivalent to sampling slices $P_{\psi,\theta}$ of the target state's Wigner function at angle θ. The standard tomographic procedure involves experimentally constructing a set of such slices $\{P_{\rho,\theta}\}_{\theta\in\Theta}$ for a set of measurement angles $\Theta \subset [0,\pi)$ via homodyne detection, and then solving the Radon inverse problem to identify the Wigner function f. However, the inverse problem of Radon transformation is mathematically ill-posed, so that direct computation without regularization suffers from instability against noise and data incompleteness. This background has driven the adoption of iterative maximum likelihood estimation (iMLE) and machine-learning-assisted methods, but these approaches still lack robustness or theoretical rigor. To transcend these limitations, we propose a rigorous, novel tomographic framework combining a MSW1 loss, which naturally harmonizes with the native structure of homodyne data, with an infinite-dimensional rank-1 factored gradient descent via an overparameterized neural network.

First, we introduce a learning model that approximates wavefunctions using a feed-forward neural network (FNN). Rather than general wavefunctions in $L^2(\mathbb{R})$, we deal with the Schwartz space $\psi \in \mathcal{S}(\mathbb{R})$, which encompasses important states such as a cat state. Crucially, we directly parameterize the wavefunction itself via the neural network, rather than the Wigner function. (This is justified because directly approximating the Wigner function makes it difficult to satisfy the positive semi-definiteness constraint of quantum states. Furthermore, if the target is known to be pure, parameterizing the wavefunction allows us to maximally exploit this prior knowledge.)

Definition 2.3 (The neural wavefunction)**.** Our learning model, a neural wavefunction, is given by:

$$\tilde{\psi}_\omega(x) = e^{-\beta x^2}\,\mathrm{FNN}_\omega(x), \qquad \tilde{\psi}_\omega \not\equiv 0, \qquad \psi_\omega = \frac{\tilde{\psi}_\omega}{\|\tilde{\psi}_\omega\|}. \tag{3}$$

Here, $\beta > 0$ is the width parameter of the Gaussian envelope, and $\mathrm{FNN}_\omega : \mathbb{R} \to \mathbb{R}^2$ is a fully-connected multilayer feed-forward neural network with GELU activation functions, $\mathrm{GELU}(x) = x\,\Phi(x)$ (where Φ is the CDF of the standard normal distribution). The two output components are treated as the real and imaginary parts, i.e., $\mathrm{FNN}_\omega(x) = u_\omega(x) + iv_\omega(x)$. The width β may also be trained jointly with the network as long as it stays positive, in which case it is counted among the parameters below. We do so in our experiments (Appendix B). Letting $\mathcal{I}$ be the index set of the weight and bias parameters, we denote the collection of all parameters as $\omega = (\omega_i)_{i\in\mathcal{I}}$ and the parameter space as Ω.

Remark 2.4. The normalization factor (the denominator $\|\tilde{\psi}_\omega\|$) is computed via numerical integration, automatically guaranteeing $\|\psi_\omega\|_{L^2} = 1$. Because this forcefully parameterizes wavefunctions, there is no risk of violating quantum state constraints during optimization. Furthermore, it eliminates the need for constraints on the network output, enabling unconstrained optimization with respect to ω.

Next, we define the loss function. Since quantum states can only be accessed through the results of measurements such as homodyne detection, it is natural to incorporate the native structure of such data into the loss function as much as possible. Given that homodyne measurement is the operation of sampling the Radon transform of the Wigner function, and that any Radon transform of a Wigner function yields a proper probability measure, we introduce the following max-sliced distance: for wavefunctions $\psi, \phi \in L^2(\mathbb{R})$ and their corresponding Wigner functions f_ψ, f_ϕ,

$$MSW_1(f_\psi, f_\phi) := \max_{\theta\in[0,\pi)} W_1\big(P_{\psi,\theta}, P_{\phi,\theta}\big). \tag{4}$$

Here, W_1 is the 1-dimensional Kantorovich–Rubinstein distance, i.e., $W_1(\mu,\nu) = \int_{-\infty}^{\infty} |F_\mu(x) - F_\nu(x)|\,dx$. (Strictly speaking, this is defined for elements $\mu, \nu \in \mathcal{P}_1(\mathbb{R})$, the space of Borel probability

measures with finite first moments. For brevity, we omit the discussion on integrability, but all quantum states considered in this study satisfy this condition.) Even though ψ and ϕ are quantum states, their slices $P_{\psi,\theta}$ and $P_{\phi,\theta}$ are probability distributions, which makes W_1 and MSW_1 well-defined. Moreover, MSW_1 constitutes a valid distance metric between Wigner functions. The only non-trivial property to verify is non-degeneracy ($MSW_1(f_\psi, f_\phi) = 0 \Rightarrow f_\psi = f_\phi$), but it follows immediately from the injectivity of the family of Radon transforms $\{\mathcal{R}_\theta\}_{\theta\in[0,\pi)}$. Consequently, we define the loss function for the neural wavefunction ψ_ω as follows:

Definition 2.5 (MSW_1 loss function for wavefunction tomography)**.** Let ϕ be the target wavefunction of the tomography. The loss function for the neural wavefunction ψ_ω is given by:

$$\mathcal{L}(\omega) := MSW_1(f_{\psi_\omega}, f_\phi). \tag{5}$$

Remark 2.6. In actual experiments, only a few measurement phases are available, so Θ is a small finite set, and the slice distributions are accessible only through measured data. In practice we therefore minimize

$$\mathcal{L}_\Theta(\omega) := \max_{\theta\in\Theta} W_1\big(P_{\psi_\omega,\theta},\ Q_{\theta,N}\big), \tag{6}$$

where $Q_{\theta,N}$ is the empirical distribution of N homodyne outcomes at relative phase θ, binned as in (10).

Remark 2.7 (Review of FGD)**.** Let f be a convex function over positive semi-definite matrices. Instead of minimizing f under a rank constraint $\mathrm{rank}(M) \le r$, Factored Gradient Descent (FGD) factors the matrix as $M = XX^\top$ ($X \in \mathbb{R}^{n\times r}$) and minimizes $g(X) = f(XX^\top)$ with respect to X (Burer & Monteiro, 2003). For finite-dimensional FGD, it is known that all local minima of g correspond to global minima of the constrained f with rank at most r, and that all saddle points of g are strict saddles (Bhojanapalli et al., 2016; Ge et al., 2016; Park et al., 2017). In the formulation of this paper, we similarly factor the quantum state as $\rho = \psi_\omega\psi_\omega^*$ to analyze the loss landscape $\mathcal{L}(\omega)$. In this respect, our method can be viewed as an infinite-dimensional extension of FGD.

2.3 Main result

As the main theorem of this study, we prove that the landscape of the loss function $\mathcal{L}(\omega)$ is benign on the differentiable region. First, we review the mathematical properties of the network. Due to the inherent properties of the Gaussian-enveloped GELU, the map $\omega \mapsto \psi_\omega$ is continuously differentiable as a map into the Schwartz class $\mathcal{S}(\mathbb{R})$. In particular, the output ψ_ω and its parameter derivatives $\partial\psi_\omega/\partial\omega_i$ for $i \in \mathcal{I}$ all belong to $\mathcal{S}(\mathbb{R})$. More generally, the analysis also relies on the following implicit assumptions:

Assumption 2.8 (Regularities of the neural wavefunctions)**.** For any $\omega \in \Omega$, the network ψ_ω is a non-zero real analytic function on $\mathbb{R}$. In addition, there exist $a, C_0 > 0$ such that $|\psi_\omega(x)| \le C_0\, e^{-ax^2}$ holds for all $x \in \mathbb{R}$.

When taking the parameter derivative of ψ_ω, the normalization constraint $\|\psi_\omega\| = 1$ dictates that any admissible directional derivative ϕ must satisfy $\mathrm{Re}\langle\psi_\omega, \phi\rangle = 0$. Assuming the network is sufficiently large, we introduce the following:

Assumption 2.9 (Over-parameterization)**.** For any $\omega \in \Omega$ and every tangent $\phi \in T\psi_\omega := \{\phi \in \mathcal{S}(\mathbb{R}) \mid \mathrm{Re}\langle\psi_\omega, \phi\rangle = 0\}$, there exists a direction in the parameter space $\xi \in \Omega$, given by $\xi = \sum_{i\in\mathcal{I}} a_i\omega_i$ such that $\phi = \partial\psi_\omega/\partial\xi$ holds.

Remark 2.10 (Realism of the Assumptions)**.** Since $\mathcal{S}(\mathbb{R})$ is infinite-dimensional, the parameter set Ω of a finite-sized network cannot strictly satisfy Assumption 2.9. This assumption should be understood as an idealization in the infinite-width limit, standardly employed in deep learning theory. This idealization is motivated by the Neural Tangent Kernel theory (Jacot et al., 2018), but we do not claim that it holds exactly for any finite network. As the asymptotic evaluation is not the main focus of this study, we present it in this idealized form. The empirical effectiveness of our numerical simulations will serve to substantiate these assumptions.

Under these assumptions, it can be shown that the tangent space of the distribution function for each slice is sufficiently rich to approximate the space of real-valued test functions $C^\infty_{\mathbb{R},c}(\mathbb{R})$.

Lemma 2.11. *Suppose the neural wavefunction ψ_ω satisfies Assumption 2.8 and Assumption 2.9. Then, for any $f \in C^\infty_{\mathbb{R},c}(\mathbb{R})$ and $\theta \in [0,\pi)$, there exists a sequence of direction $\xi_n \in \Omega$ in the*

parameter space such that

$$f = \lim_{n\to\infty} \frac{\partial}{\partial \xi_n} F_{P_{\psi_{\omega,\theta}}} \tag{7}$$

holds with respect to $L^1(\mathbb{R})$-norm topology. Here, $\psi_{\omega,\theta} := \mathcal{F}_\theta \psi_\omega$ and $F_{P_{\psi_{\omega,\theta}}}$ is the cumulative distribution function of the slice $P_{\psi_{\omega,\theta}}$, defined as $F_{P_{\psi_{\omega,\theta}}}(x) = \int_{-\infty}^{x} P_{\psi_{\omega,\theta}}(y)dy$.

This directly leads to the following main theorem.

Theorem 2.12. *Suppose that Assumptions 2.8 and 2.9 hold. If $\mathcal{L}$ is differentiable at ω^* and $\nabla\mathcal{L}(\omega^*) = 0$, then the model distribution exactly matches the target distribution, that is,*

$$P_{\psi_{\omega^*},\theta} = P_{\phi,\theta} \quad \forall\, \theta \in [0, \pi). \tag{8}$$

Furthermore, in this case $\psi_{\omega^*} = e^{i\alpha}\phi$ holds for some $\alpha \in \mathbb{R}$.

Remark 2.13. At a heuristic level, the hard pure-state constraint is what allows our method to look through noise. Once noise and decoherence act, the empirical slice distributions $\{Q_{\theta,N}\}_{\theta\in\Theta}$ no longer originate from any wavefunction, and an estimator that may output mixed states, such as the full-rank iMLE, fits the blurred statistics themselves and loses the interference fringes. Our method cannot do so: its output is always a pure state, ideally the one closest to the data in the geometry of the loss, and in our experiments this pure state retains the fringes of the state before the noise. The constraint alone is not sufficient under stronger noise, however, as the rank-1 constrained iMLE shows (Section 4 and Appendix B). It is its combination with the Wasserstein loss that yields high fidelity. This is a numerical observation rather than a consequence of Theorem 2.12, which assumes that the data come from a pure state.

Remark 2.14. We consider only stationary points at which $\mathcal{L}$ is differentiable and $\nabla\mathcal{L}(\omega^*) = 0$ holds. Since $\mathcal{L}$ may fail to be differentiable in general due to the maximum over θ, the analysis of local minimality at such non-differentiable points requires nonsmooth analysis, which is beyond the scope of this work. Accordingly, we make no claim about the behavior of the optimizer at such points, and we leave a guarantee that first-order methods such as Adam do not stall there to future work.

3 Numerical Experiments: Tomography of the Two-Mode Cat State

3.1 Two-mode cat states

For an amplitude $\alpha \in \mathbb{R}$, the wavefunction $\Psi(x_1, x_2)$ of a two-mode cat state is defined using the Gaussian wave packet $\psi_{\pm\alpha}(x) = \frac{1}{\pi^{1/4}} \exp\big[-\frac{1}{2}(x \mp \alpha)^2\big]$ as $\Psi(x_1, x_2) = \mathcal{N}[\psi_{+\alpha}(x_1)\psi_{+\alpha}(x_2) + \psi_{-\alpha}(x_1)\psi_{-\alpha}(x_2)]$, where $\mathcal{N} = (2 + 2\exp\big(-2\alpha^2\big))^{-1/2}$ is the normalization constant. By deriving the Wigner function for this state and applying a coordinate transformation to the center-of-mass and relative coordinates $x_\pm = (x_1 \pm x_2)/\sqrt{2},\ p_\pm = (p_1 \pm p_2)/\sqrt{2}$, we obtain the following:

Proposition 3.1. *The Wigner function f_Ψ of the two-mode cat state Ψ with parameter α can be decomposed into $f_\Psi = f_{\Psi,+} \otimes f_{\Psi,-}$, a tensor product of a Gaussian relative mode $f_{\Psi,-}(x_-, p_-) = \frac{1}{\pi} e^{-x_-^2 - p_-^2}$ and a non-Gaussian center-of-mass mode $f_{\Psi,+}$ given by*

$$f_{\Psi,+}(x_+, p_+) = \frac{\mathcal{N}^2}{\pi} \left\{ \sum_{s=\{+,-\}} e^{-(x_+ + s\sqrt{2}\alpha)^2 - p_+^2} + 2e^{-x_+^2 - p_+^2} \cos\big(2\sqrt{2}\alpha p_+\big) \right\}.$$

3.2 Experimental setup

We perform tomography of the center-of-mass mode Wigner function $f_{\Psi,+}$ in the two-mode cat state Ψ. Since the homodyne measurement of $f_{\Psi,+}$ at a relative phase θ corresponds to the configuration $\theta = \theta_1 = \theta_2$, we denote the probability distribution of that slice simply as Q_θ, emphasizing its dependence solely on θ. To account for photon loss, we introduce a parameter $\sigma > 0$ and define the noisy distribution via Gaussian convolution as

$$\tilde{Q}_\theta(x) := \big(Q_\theta * G_\sigma\big)(x), \quad G_\sigma(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\Big(-\frac{x^2}{2\sigma^2}\Big). \tag{9}$$

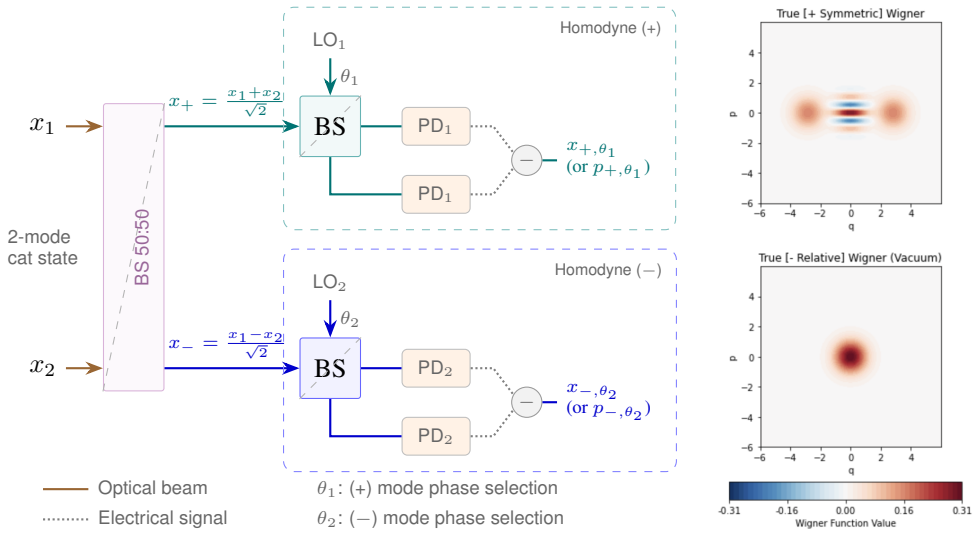


Figure 2: [Left] Schematic of the experimental circuit used in this simulation. The transformation into (+) and (−) modes is implemented by a 50:50 beam splitter (BS). Homodyne detection is achieved by interfering each of the separated beams with a reference local oscillator (LO) at a controlled phase. [Right] Wigner functions corresponding to (+) and (−) modes. The blue areas indicate the area of the negative values of Wigner functions, highlighting the non-classicality of the state. Since the (−) mode is decoupled into a Gaussian state, all the complex non-classical features are marginalized into the (+) mode. Thus, reconstructing the (+) mode is sufficient for our objective.

This is exactly the effect of a loss channel with transmissivity $\eta = 1/(1+2\sigma^2)$, which also describes the inefficiency of the homodyne detectors (Leonhardt & Paul, 1993). The loss maps each quadrature as $x \mapsto \sqrt{\eta}\,x + \sqrt{1-\eta}\,x_{\text{vac}}$ with vacuum noise of variance $1/2$, so that after rescaling by the known factor $1/\sqrt{\eta}$ the outcomes are those of the lossless state with added Gaussian noise of variance $(1-\eta)/(2\eta) = \sigma^2$. The settings $\sigma = 0.3$, 0.6, and 0.9 used below thus correspond to $\eta \approx 0.85$, 0.58, and 0.38. Unlike the rescaling, the added noise cannot be removed by calibration. It smooths out the negative interference fringes, and $\{\tilde{Q}_\theta\}$ are the slices of a mixed state. A pure-state model cannot reproduce them exactly (see Remark 2.13). Fidelities are always measured against the pure cat state before the loss.

Measurement sampling is performed from this noisy distribution. First, let Θ be a finite set of measurement phases. For each $\theta \in \Theta$, we define measurement intervals (bins) $B_{\theta,j}$ $(j = 1, \ldots, J)$ that account for the resolution of the experimental detectors (e.g., camera pixels). Let $f_{\theta,N,j}$ denote the number of times the measurement outcomes fall into the bin $B_{\theta,j}$ over N shots at phase θ, and let $\bar{x}_{\theta,N,j}$ be the mean value within each bin. Consequently, the empirical distribution $Q_{\theta,N}$ and its cumulative distribution function $F_{Q_{\theta,N}}$ for each slice can be written as:

$$F_{Q_{\theta,N}}(x) = \sum_{j:\, \bar{x}_{\theta,N,j} \le x} Q_{\theta,N}(B_{\theta,j}), \qquad Q_{\theta,N}(B_{\theta,j}) = \frac{f_{\theta,N,j}}{N}. \tag{10}$$

With these definitions, the computation of our proposed method becomes feasible.

Baseline: diluted iterative Maximum Likelihood Estimation. As a baseline, we refer to the diluted iterative maximum likelihood estimation (iMLE) method. Using the notation from the previous section, if we let $\Pi_{\theta,j}$ be the indicator function on $B_{\theta,j}$, it serves as a projection operator on $L^2(\mathbb{R})$. Here, for $t \in (0,1)$, the update rule for the estimated quantum state is given by:

$$\rho^{(k+1)} = \frac{R_t(\rho^{(k)})\rho^{(k)}R_t(\rho^{(k)})}{\operatorname{tr}\left[R_t(\rho^{(k)})\rho^{(k)}R_t(\rho^{(k)})\right]},$$

where $R_t(\rho) = I + t\big(R(\rho) - I\big)$ and $R(\rho) = \frac{1}{N|\Theta|}\sum_{\theta\in\Theta}\sum_{j=1,\ldots,J}\frac{f_{\theta,N,j}\Pi_{\theta,j}}{\operatorname{tr}(\rho\Pi_{\theta,j})}$.

Proposition 3.2 ((Řeháček et al., 2007)). *For small t, $\rho^{(k)}$ globally converges to the state ρ^* that maximizes the log-likelihood functional* $\log\mathcal{L}(\rho) = \sum_{\theta\in\Theta}\sum_{j=1,\ldots,J} f_{\theta,N,j}\log\operatorname{tr}\left(\rho\Pi_{\theta,j}\right)$.

Evaluation. The tomographic quality of the output quantum state ρ_t produced by each algorithm at time t is quantified by the overlap fidelity:

$$F(t) := 2\pi \int_{\mathbb{R}^2} f_{\Psi,+} f_{\rho_t}\, dq_+\, dp_+. \tag{11}$$

A value of $F(t) = 1$ indicates a perfect reconstruction up to a global phase. To compare convergence behavior, we plot F against the cumulative wall-clock time t of the respective algorithms.

4 Numerical experiments

We present results of numerical experiments. Implementation details are given in Appendix B. Throughout, the neural wavefunction of Definition 2.3 is a fully connected network with 8 layers of width 128 (about 10^5 parameters), trained with Adam at learning rate 3×10^{-4} for a fixed wall-clock budget of 24 s per trial. Each step evaluates all N_{angle} slices and back-propagates through the one with the largest loss. All reported settings are deliberately challenging: only $N = 1000$ shots per angle, recorded in bins of width $|B| \approx 0.37$, i.e., about three bins per period (≈ 1.11) of the interference fringes, and as few as $N_{\text{angle}} = 6$ measurement phases. For a state containing at most t photons, homodyne data are informationally complete only with $t + 1$ distinct phases (Leonhardt, 1997), and simulation studies of iMLE accordingly use 20 phases with 2×10^4 shots in total for cat states of coherent amplitude up to 2, truncated at 15 photons (Silva et al., 2018). The center-of-mass mode considered here is as large as their largest cat state: its two peaks at $x_+ = \pm\sqrt{2}\alpha$ are coherent states of amplitude $\pm\alpha = \pm 2$, and its mean photon number is about 4, so their truncation would call for at least 16 phases. With six phases the data alone therefore do not determine the state, and any reconstruction must rely on prior structure. Our method relies on purity. We show that the proposed method captures the interference fringes of the cat state.

Scarce clean data. We first examine performance as a function of the number of homodyne phase settings under the idealized photon-loss-free condition. Figure 3 compares fidelity and loss curves for $N_{\text{angle}} \in \{6, 12, 24\}$. As the number of phase settings decreases, iMLE rapidly loses fidelity, whereas the proposed method maintains high fidelity robustly. The MSW1 loss decreases steadily without stalling, consistent with Theorem 2.12.

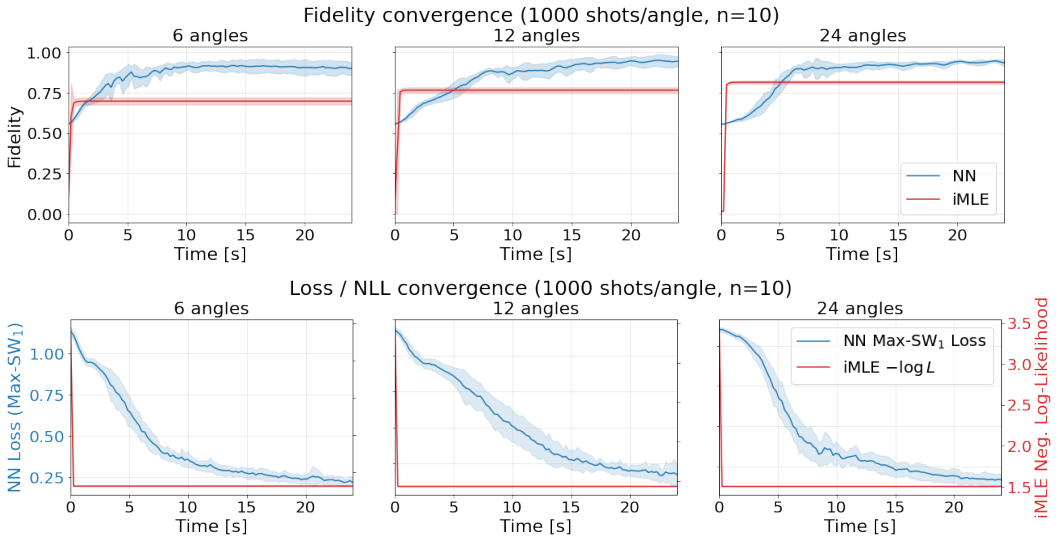


Figure 3: [Top] fidelity vs. wall-clock time for $N_{\text{angle}} = 6, 12, 24$ (left to right). [Bottom] MSW1 loss for the proposed method NN (left axis, blue) and negative log-likelihood for iMLE (right axis, red). Both losses are shown to illustrate their descent, but they are not directly comparable in magnitude. Solid lines and shaded bands denote means and standard deviations over 10 trials.

Scarce noisy data. We next fix $N_{\text{angle}} = 6$ and add photon loss, modeled as a Gaussian convolution of standard deviation σ on each slice. We test three regimes: moderate ($\sigma = 0.3$), strong ($\sigma = 0.6$),

and extreme ($\sigma = 0.9$). Figure 4 reports fidelity over time. While iMLE deteriorates sharply with increasing decoherence, the proposed method is essentially unaffected up to $\sigma = 0.6$. At $\sigma = 0.9$, the data quality is too degraded even for the proposed model to recover the fringes faithfully and a slow fidelity decay appears late in training. Reconstructed Wigner functions of a single trial at $\sigma = 0.6$ are shown in Figure 5: the proposed method retains pronounced interference fringes with negative values at the center of phase space, although its two coherent peaks are smeared along x_+, while iMLE blurs out the negative-valued fringes that are the hallmark of non-classicality.

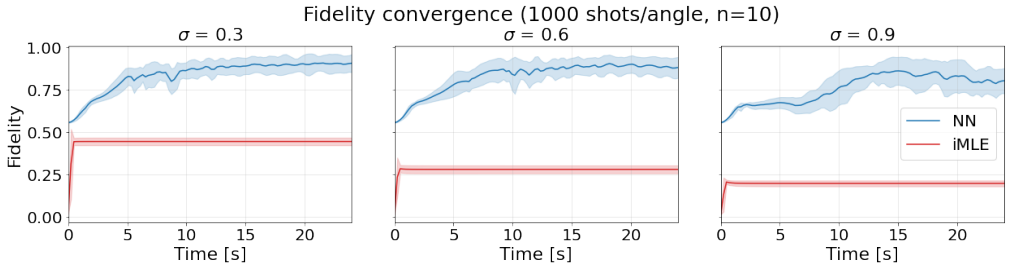


Figure 4: Performance under photon loss at $N_{\text{angle}} = 6$. Fidelity vs. wall-clock time for $\sigma \in \{0.3, 0.6, 0.9\}$ (left to right). Mean and standard-deviation bands over 10 trials.

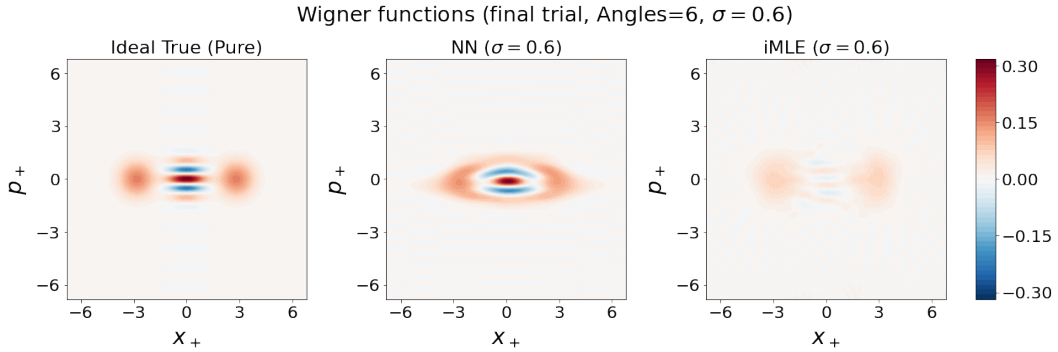


Figure 5: Reconstructed Wigner functions of a single trial at $N_{\text{angle}} = 6$ and $\sigma = 0.6$. [Left] ground-truth pure cat state. [Center] proposed method (NN). [Right] iMLE. The proposed method retains the central negative fringes characteristic of non-classicality, whereas iMLE blurs the fringes.

Scarce noisy data with rank-1 constrained iMLE. The full-rank iMLE above searches over all density matrices, whereas the proposed method enforces purity by construction. To separate the effect of this hard constraint from that of the loss function, we also run iMLE restricted to pure states. If the iteration is started from a rank-1 state $\rho^{(0)} = |\psi_0\rangle\langle\psi_0|$, every iterate remains rank-1. We draw ψ_0 uniformly at random from the unit sphere of the same truncated Fock space. Restricted to pure states, however, the likelihood is no longer maximized over a convex set, and the global convergence guarantee of Řeháček et al. (2007) does not apply. With t chosen by backtracking so that the likelihood increases monotonically, the iteration reaches only a solution of the extremal equation, and which solution it reaches depends on the initial state. Within the wall-clock budget, the iteration is rerun from fresh random initial states until the budget is exhausted, and the state with the highest training likelihood is reported.

The results are shown in Figure 6. The hard constraint alone does not recover the fringes. The rank-1 iMLE converges within a second to fidelities of about 0.75, 0.30, and 0.17 at $\sigma = 0.3$, 0.6, and 0.9, against about 0.90, 0.87, and 0.80 for the proposed method. Its spread over trials reaches a standard deviation of about 0.2 at $\sigma = 0.6$, as the pure-state likelihood becomes multimodal. With a smaller Fock truncation of dimension 40, the rank-1 iMLE exceeds the proposed method at $\sigma = 0.3$ but still falls below it under stronger noise (Remark B.1).

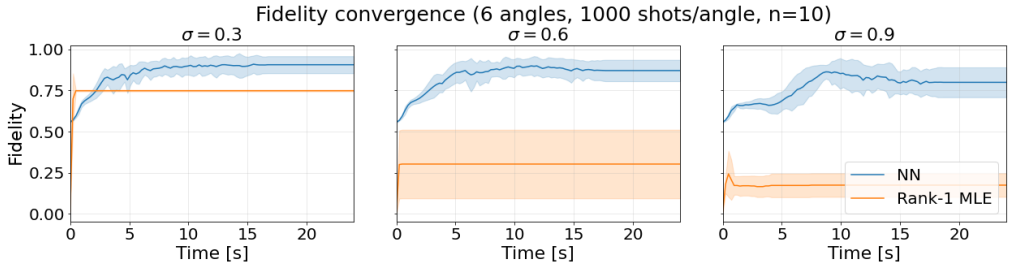


Figure 6: Proposed method (NN) vs. rank-1 constrained iMLE at $N_{\text{angle}} = 6$ with 1000 shots per angle and $\sigma \in \{0.3, 0.6, 0.9\}$ (left to right), using a Fock truncation of dimension 200 for the rank-1 iMLE. Fidelity vs. wall-clock time. Mean and standard-deviation bands over 10 trials.

Comparison with SW1 loss. A natural alternative to MSW1 replaces the maximum by an average,

$$SW_1(f_\psi, f_\phi) := \frac{1}{\pi} \int_0^\pi W_1\big(P_{\psi,\theta}, P_{\phi,\theta}\big)\, d\theta, \tag{12}$$

whose empirical version averages the slice losses over the measured angles. Empirically, SW1 reaches high fidelity faster than MSW1 (Figure 7), although at $\sigma = 0.9$ its fidelity decays later in training. Our theory, however, does not cover it. At a differentiable point of MSW1 the gradient is that of a single slice, which is what Theorem 2.12 exploits, whereas the gradient of SW1 is an average over slices, whose terms may cancel without vanishing individually. SW1 may therefore have spurious stationary points. Why such points do not appear to hinder training is left for future work.

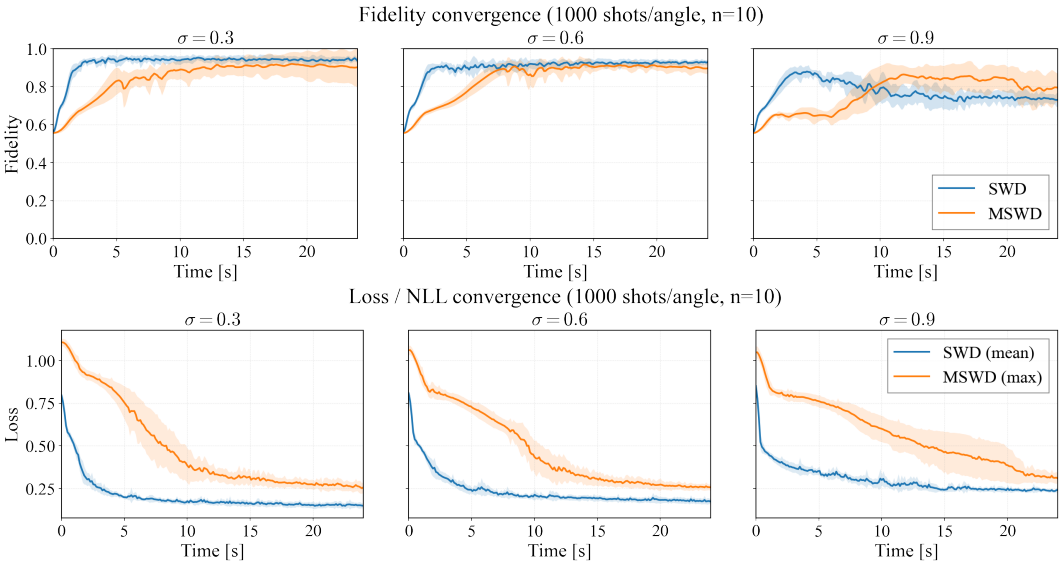


Figure 7: Fidelities and losses comparison of SW1 (blue) and MSW1 (orange) at $N_{\text{angle}} = 6$ and $\sigma = 0.3, 0.6, 0.9$. Mean trajectories over 10 trials, with standard-deviation bands.

5 Conclusion and Open Problems

We trained a neural wavefunction with the max-sliced Wasserstein-1 loss on homodyne data and proved that, under an over-parameterization assumption, every differentiable critical point of the loss is a global minimum. In simulations of a two-mode cat state with six measurement phases, the method retained fidelities of about 0.9 under photon loss up to $\sigma = 0.6$, where iMLE degraded sharply, and the rank-1 constrained iMLE shows that under strong noise the pure-state constraint alone does not suffice. A natural next step is a guarantee on the fidelity itself, for example a bound on the infidelity in terms of the attained loss.

Acknowledgments and Disclosure of Funding

This work was supported by ERATO "Super Quantum Entanglement" (Grant No. JPMJER2402) from JST.

References

S. Ahmed, C. Sánchez Muñoz, F. Nori, and A. F. Kockum. Quantum state tomography with conditional generative adversarial networks. *Phys. Rev. Lett.*, 127:140502, 2021.

S. D. Bartlett, B. C. Sanders, S. L. Braunstein, and K. Nemoto. Efficient classical simulation of continuous variable quantum information processes. *Phys. Rev. Lett.*, 88:097904, 2002.

J. Bertrand and P. Bertrand. A tomographic approach to Wigner's function. *Found. Phys.*, 17:397–405, 1987.

S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

S. L. Braunstein and P. van Loock. Quantum information with continuous variables. *Rev. Mod. Phys.*, 77: 513–577, 2005.

S. Burer and R. D. C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Math. Program.*, 95:329–357, 2003.

G. Carleo and M. Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355:602–606, 2017.

P. Cha, P. Ginsparg, F. Wu, J. Carrasquilla, P. L. McMahon, and E.-A. Kim. Attention-based quantum tomography. *Mach. Learn.: Sci. Technol.*, 3:01LT01, 2022.

R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

D. Gottesman, A. Kitaev, and J. Preskill. Encoding a qubit in an oscillator. *Phys. Rev. A*, 64:012310, 2001.

S. Helgason. *The Radon Transform*. Birkhäuser, 2nd edition, 1999.

Z. Hradil. Quantum-state estimation. *Phys. Rev. A*, 55:R1561–R1564, 1997.

R. L. Hudson. When is the Wigner quasi-probability density non-negative? *Rep. Math. Phys.*, 6(2):249–252, 1974.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

J. Joo, W. J. Munro, and T. P. Spiller. Quantum metrology with entangled coherent states. *Phys. Rev. Lett.*, 107: 083601, 2011.

U. Leonhardt. *Measuring the Quantum State of Light*. Cambridge University Press, 1997.

U. Leonhardt and H. Paul. Realistic optical homodyne measurements and quasiprobability distributions. *Phys. Rev. A*, 48:4598–4604, 1993.

A. W. Lohmann and B. H. Soffer. Relationships between the Radon–Wigner and fractional Fourier transforms. *J. Opt. Soc. Am. A*, 11(6):1798–1801, 1994.

A. I. Lvovsky. Iterative maximum-likelihood reconstruction in quantum homodyne tomography. *J. Opt. B: Quantum Semiclass. Opt.*, 6:S556–S559, 2004.

A. I. Lvovsky and M. G. Raymer. Continuous-variable optical quantum-state tomography. *Rev. Mod. Phys.*, 81: 299–332, 2009.

A. Mari and J. Eisert. Positive Wigner functions render classical simulation of quantum computation efficient. *Phys. Rev. Lett.*, 109:230503, 2012.

M. Mirrahimi, Z. Leghtas, V. V. Albert, S. Touzard, R. J. Schoelkopf, L. Jiang, and M. H. Devoret. Dynamically protected cat-qubits: a new paradigm for universal quantum computation. *New J. Phys.*, 16:045014, 2014.

N. Ofek, A. Petrenko, R. Heeres, P. Reinhold, Z. Leghtas, B. Vlastakis, Y. Liu, L. Frunzio, S. M. Girvin, L. Jiang, M. Mirrahimi, M. H. Devoret, and R. J. Schoelkopf. Extending the lifetime of a quantum bit with error correction in superconducting circuits. *Nature*, 536:441–445, 2016.

A. Ourjoumtsev, R. Tualle-Brouri, J. Laurat, and P. Grangier. Generating optical Schrödinger kittens for quantum information processing. *Science*, 312:83–86, 2006.

A. Ourjoumtsev, H. Jeong, R. Tualle-Brouri, and P. Grangier. Generation of optical 'Schrödinger cats' from photon number states. *Nature*, 448:784–786, 2007.

A. Ourjoumtsev, F. Ferreyrol, R. Tualle-Brouri, and P. Grangier. Preparation of non-local superpositions of quasi-classical light states. *Nat. Phys.*, 5:189–192, 2009.

D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi. Non-square matrix sensing without spurious local minima via the Burer–Monteiro approach. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

J. Radon. Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. *Ber. Verh. Sächs. Akad. Wiss. Leipzig*, 69:262–277, 1917.

J. Řeháček, Z. Hradil, E. Knill, and A. I. Lvovsky. Diluted maximum-likelihood algorithm for quantum tomography. *Phys. Rev. A*, 75(4):042108, 2007.

A. Rocchetto, E. Grant, S. Strelchuk, G. Carleo, and S. Severini. Learning hard quantum distributions with variational autoencoders. *npj Quantum Inf.*, 4:28, 2018.

J. L. E. Silva, S. Glancy, and H. M. Vasconcelos. Quadrature histograms in maximum-likelihood quantum state tomography. *Phys. Rev. A*, 98(2):022325, 2018.

E. S. Tiunov, V. V. Tiunova, A. E. Ulanov, A. I. Lvovsky, and A. K. Fedorov. Experimental quantum homodyne tomography via machine learning. *Optica*, 7:448–454, 2020.

G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo. Neural-network quantum state tomography. *Nat. Phys.*, 14:447–450, 2018.

K. Vogel and H. Risken. Determination of quasiprobability distributions in terms of probability distributions for the rotated quadrature phase. *Phys. Rev. A*, 40:2847–2849, 1989.

C. Weedbrook, S. Pirandola, R. García-Patrón, N. J. Cerf, T. C. Ralph, J. H. Shapiro, and S. Lloyd. Gaussian quantum information. *Rev. Mod. Phys.*, 84:621–669, 2012.

A Proof of Main Results

Proposition A.1. *Let $\psi \in \mathcal{S}(\mathbb{R})$, and assume that the zero set $Z(\psi) := \{x \in \mathbb{R} : \psi(x) = 0\}$ is discrete. Then,*

$$C^\infty_{\mathbb{R},c}(\mathbb{R}) \subset \overline{\left\{ F(x) = \int_{-\infty}^{x} \mathrm{Re}\big[\overline{\phi(y)}\psi(y)\big]\, dy \;\middle|\; \phi \in \mathcal{S}(\mathbb{R}),\ \mathrm{Re}\langle \phi, \psi\rangle = 0 \right\}}^{\|\cdot\|_{L^1}}.$$

Proof. Take an arbitrary $F \in C^\infty_{\mathbb{R},c}(\mathbb{R})$. Then $f := F' \in C^\infty_{\mathbb{R},c}(\mathbb{R})$ and it satisfies $\int f\, dx = 0$. Let $\mathrm{supp}(f) \subset [a,b]$, and set $[a,b] \cap Z(\psi) = \{x_1, \dots, x_K\}$ (which is finite by discreteness). First, by taking $\epsilon > 0$ sufficiently small, we can choose the ϵ-neighborhoods $I_k := (x_k - \epsilon, x_k + \epsilon)$ of each x_k to be mutually disjoint. Let $\chi_k \in C^\infty_{\mathbb{R},c}(\mathbb{R})$ be bump functions satisfying $0 \le \chi_k \le 1$, $\chi_k \equiv 1$ on $(x_k - \epsilon/2, x_k + \epsilon/2)$, and $\mathrm{supp}(\chi_k) \subset I_k$, and define

$$f_\epsilon(x) := f(x) \prod_{k=1}^{K} \big(1 - \chi_k(x)\big).$$

Then $f_\epsilon \in C^\infty_{\mathbb{R},c}(\mathbb{R})$ and $\mathrm{supp}(f_\epsilon) \cap Z(\psi) = \emptyset$ holds. Next, set $c_\epsilon := \int f_\epsilon\, dx$. Since $\int f\, dx = 0$, we have $c_\epsilon = -\int f(x)(1 - \prod_k (1 - \chi_k))\, dx$, and thus $|c_\epsilon| \le \|f\|_\infty 2K\epsilon$. Furthermore, taking $\eta \in C^\infty_c(\mathbb{R})$ as a function satisfying $\mathrm{supp}(\eta) \cap Z(\psi) = \emptyset$ and $\int \eta\, dx = 1$ (such η exists because $Z(\psi)$ is discrete), and setting

$$\tilde{f}_\epsilon := f_\epsilon - c_\epsilon\, \eta,$$

we obtain $\tilde{f}_\epsilon \in C^\infty_{\mathbb{R},c}(\mathbb{R})$, $\int \tilde{f}_\epsilon\, dx = 0$, and $\mathrm{supp}(\tilde{f}_\epsilon) \cap Z(\psi) = \emptyset$. Now, writing $\psi = |\psi| e^{i\alpha}$ in polar form, we define $\overline{\phi_\epsilon(x)} := \tilde{f}_\epsilon(x)/\psi(x)$. Since $\mathrm{supp}(\tilde{f}_\epsilon) \cap Z(\psi) = \emptyset$, we have $\psi(x) \neq 0$ on $\mathrm{supp}(\tilde{f}_\epsilon)$, so ϕ_ϵ is well-defined and $\phi_\epsilon \in C^\infty_c(\mathbb{R}) \subset \mathcal{S}(\mathbb{R})$. By construction, it is clear that $\mathrm{Re}\big[\overline{\phi_\epsilon(x)}\psi(x)\big] = \tilde{f}_\epsilon(x)$ for all $x \in \mathbb{R}$ holds, and from $\int \tilde{f}_\epsilon = 0$, the condition $\mathrm{Re}\langle \phi_\epsilon, \psi\rangle = 0$ is also satisfied. Therefore, setting

$$\tilde{F}_\epsilon(x) := \int_{-\infty}^{x} \tilde{f}_\epsilon(y)\, dy,$$

$\tilde{F}_\epsilon$ is an element of the set on the right-hand side, so it suffices to show that this converges to F in the L^1 norm as $\varepsilon \to 0$. Since $f - f_\epsilon = f \cdot (1 - \prod_k (1 - \chi_k))$ is supported on $\bigcup_{k=1}^K I_k$ and $\|f - f_\epsilon\|_{L^1} \le \|f\|_\infty 2K\epsilon$, we have

$$\|f - \tilde{f}_\epsilon\|_{L^1} \le \|f - f_\epsilon\|_{L^1} + |c_\epsilon|\, \|\eta\|_{L^1} \le \|f\|_\infty 2K\epsilon\, (1 + \|\eta\|_{L^1}).$$

Furthermore, by the compactness of the supports of f and $\tilde{f}_\epsilon$, there exists a positive constant M such that $\Big|\mathrm{supp}(f) \cup \mathrm{supp}(\tilde{f}_\epsilon)\Big| < M$. Therefore,

$$\|F - \tilde{F}_\epsilon\|_{L^1} = \int_{-\infty}^{\infty} \left| \int_{-\infty}^{x} (f - \tilde{f}_\epsilon)(y)\, dy \right| dx \le M\, \|f - \tilde{f}_\epsilon\|_{L^1} \le \|f\|_\infty 2MK\epsilon\, (1 + \|\eta\|_{L^1}).$$

Therefore, by taking $\varepsilon \to 0$, we see that F belongs to the L^1-closure of the set on the right-hand side. □

Proposition A.2. *Let $\psi \in S(\mathbb{R})$, $\|\psi\| = 1$ satisfy Assumption 2.8. Then for any $\theta \in [0, \pi)$, the zero set $Z(\psi_\theta) := \{x \in \mathbb{R} : \psi_\theta(x) = 0\}$ of the fractional Fourier transform $\psi_\theta := \mathcal{F}_\theta \psi$ is discrete, and each zero has finite order.*

Proof. Since $\mathcal{F}_\theta$ is a unitary operator on $L^2(\mathbb{R})$, we have $\|\psi_\theta\|_{L^2} = \|\psi\|_{L^2} = 1$, and in particular $\psi_\theta \not\equiv 0$. In the case $\theta = 0$, we have $\psi_0 = \psi$, and by the real analyticity from Assumption 2.8, $Z(\psi)$ is discrete and each zero has finite order, so the claim follows immediately. In the following, we take $\theta \in (0, \pi)$. We show that ψ_θ is the restriction to $\mathbb{R}$ of an entire function $\tilde{\psi}_\theta$ on $\mathbb{C}$. Once this is shown, since $\tilde{\psi}_\theta$ is a non-trivial entire function, its zero set $Z(\tilde{\psi}_\theta) = \{z \in \mathbb{C} : \tilde{\psi}_\theta(z) = 0\}$ is discrete in $\mathbb{C}$, and each zero has finite order. Then its restriction to $\mathbb{R}$, $Z(\psi_\theta)$, is also discrete in $\mathbb{R}$, and each zero has finite order, which gives the conclusion.

First, define $g_\theta : \mathbb{R} \to \mathbb{C}$ by $g_\theta(y) := e^{iy^2 \cot\theta/2}\,\psi(y)$. By the Gaussian-type decay from Assumption 2.8, at each point we have $|g_\theta(y)| = |\psi(y)| \le C_0\, e^{-ay^2}$. Considering the function

$$G(\xi) := \int_{\mathbb{R}} e^{-i\xi y}\, g_\theta(y)\, dy$$

with $\xi \in \mathbb{C}$ as a parameter, splitting $\xi = \xi_1 + i\xi_2$ ($\xi_1, \xi_2 \in \mathbb{R}$), the integrand of $G(\xi)$ satisfies

$$|e^{-i\xi y}\, g_\theta(y)| = |e^{-i\xi_1 y}|\,|e^{\xi_2 y}|\,|g_\theta(y)| \le C_0\, e^{\xi_2 y - ay^2} = C_0\, e^{\xi_2^2/(4a)}\, e^{-a(y-\xi_2/(2a))^2}$$

which is integrable in y, so $G(\xi)$ can be defined for each $\xi \in \mathbb{C}$. To show that $G(\xi)$ is entire, take any compact set $K \subset \mathbb{C}$ containing ξ, and set $M_K := \sup_{\xi\in K} |\mathrm{Im}\,\xi|$. From the above estimate,

$$\sup_{\xi\in K} |e^{-i\xi y}\, g_\theta(y)| \le C_0\, e^{M_K|y| - ay^2} =: h_K(y) \in L^1(\mathbb{R}).$$

Therefore, the integrand has an integrable dominating function h_K on K. Hence, for any closed curve γ in the complex plane, by Fubini's theorem and the fact that $\xi \mapsto e^{-i\xi y}$ is entire,

$$\oint_\gamma G(\xi)\, d\xi = \int_{\mathbb{R}} g_\theta(y) \left(\oint_\gamma e^{-i\xi y}\, d\xi \right) dy = 0.$$

Since this holds throughout the complex plane, by Morera's theorem, G is entire.

Now, looking at the definition of the fractional Fourier transform

$$\begin{aligned}\psi_\theta(u) &= \sqrt{\frac{1 - i\cot\theta}{2\pi}}\, e^{iu^2\cot\theta/2} \int_{\mathbb{R}} e^{iy^2\cot\theta/2 - iuy\csc\theta}\, \psi(y)\, dy \\ &= \sqrt{\frac{1 - i\cot\theta}{2\pi}}\, e^{iu^2\cot\theta/2}\, G(u\csc\theta),\end{aligned}$$

the right-hand side $u \mapsto e^{iu^2\cot\theta/2}$ is the exponential of the polynomial $iu^2\cot\theta/2$, hence entire on $\mathbb{C}$, and $u \mapsto G(u\csc\theta)$ is entire on $\mathbb{C}$ as the composition of the entire function G and the linear transformation $u \mapsto u\csc\theta$. Therefore,

$$\tilde{\psi}_\theta(u) := \sqrt{\frac{1 - i\cot\theta}{2\pi}}\, e^{iu^2\cot\theta/2}\, G(u\csc\theta)$$

is entire. Therefore, ψ_θ is the restriction to the real axis of the entire function $\tilde{\psi}_\theta$, which completes the proof. □

Proof of Lemma 2.11. Fix $\theta \in [0, \pi)$ and write $P_{\omega,\theta} := P_{\psi_\omega,\theta} = |\psi_{\omega,\theta}|^2$ for the slice in the statement. As noted before Assumption 2.8, the Gaussian-enveloped GELU network makes $\omega \mapsto \psi_\omega$ continuously differentiable as a map into $\mathcal{S}(\mathbb{R})$. Since $\mathcal{F}_\theta$ is a continuous linear automorphism of $\mathcal{S}(\mathbb{R})$, the map $\omega \mapsto \psi_{\omega,\theta} = \mathcal{F}_\theta\psi_\omega$ is continuously differentiable into $\mathcal{S}(\mathbb{R})$ as well, with $\partial\psi_{\omega,\theta}/\partial\xi = \mathcal{F}_\theta(\partial\psi_\omega/\partial\xi)$ for every direction ξ in the parameter space. By (2) and the continuity of multiplication on $\mathcal{S}(\mathbb{R})$, the map $\omega \mapsto P_{\omega,\theta}$ is continuously differentiable into $\mathcal{S}(\mathbb{R})$ with

$$\frac{\partial P_{\omega,\theta}}{\partial\xi} = 2\,\mathrm{Re}\Big[\,\overline{\psi_{\omega,\theta}}\;\mathcal{F}_\theta\Big(\frac{\partial\psi_\omega}{\partial\xi}\Big)\Big],$$

and the difference quotients $t^{-1}(P_{\omega+t\xi,\theta} - P_{\omega,\theta})$ converge to $\partial P_{\omega,\theta}/\partial\xi$ in $\mathcal{S}(\mathbb{R})$ as $t \to 0$, hence in $L^1(\mathbb{R})$. Since $\big|\int_{-\infty}^x g(y)\,dy\big| \le \|g\|_{L^1}$ for every $x \in \mathbb{R}$ and $g \in L^1(\mathbb{R})$, differentiation and integration can therefore be interchanged:

$$\frac{\partial F_{P_{\omega,\theta}}}{\partial\xi}(x) = 2\int_{-\infty}^x \mathrm{Re}\Big[\,\overline{\psi_{\omega,\theta}(y)}\;\mathcal{F}_\theta\Big(\frac{\partial\psi_\omega}{\partial\xi}\Big)(y)\Big]\,dy.$$

The normalization gives $\mathrm{Re}\langle\psi_\omega, \partial\psi_\omega/\partial\xi\rangle = 0$ for every ξ, so by Assumption 2.9 the derivative $\partial\psi_\omega/\partial\xi$ ranges over the whole tangent space $T\psi_\omega$ as ξ varies. Because $\mathcal{F}_\theta$ is a bijection of $\mathcal{S}(\mathbb{R})$ that preserves the inner product, $\mathrm{Re}\langle\psi_{\omega,\theta}, \mathcal{F}_\theta\varphi\rangle = \mathrm{Re}\langle\psi_\omega, \varphi\rangle$ for every $\varphi \in \mathcal{S}(\mathbb{R})$, so $\mathcal{F}_\theta(\partial\psi_\omega/\partial\xi)$

ranges over the whole tangent space $T\psi_{\omega,\theta} := \{\chi \in \mathcal{S}(\mathbb{R}) \mid \mathrm{Re}\langle \psi_{\omega,\theta}, \chi\rangle = 0\}$. Using $\mathrm{Re}[\overline{\psi}\chi] = \mathrm{Re}[\overline{\chi}\psi]$ and $\mathrm{Re}\langle\psi,\chi\rangle = \mathrm{Re}\langle\chi,\psi\rangle$, we conclude that, as ξ varies, $\partial F_{P_{\omega,\theta}}/\partial\xi$ ranges over the set

$$\Big\{ x \mapsto 2\int_{-\infty}^{x} \mathrm{Re}\big[\,\overline{\chi(y)}\,\psi_{\omega,\theta}(y)\big]\,dy \;\Big|\; \chi \in \mathcal{S}(\mathbb{R}),\ \mathrm{Re}\langle\chi,\psi_{\omega,\theta}\rangle = 0 \Big\}.$$

Up to the factor 2, which is immaterial because the set is a real vector space, this is the set of Proposition A.1 for $\psi = \psi_{\omega,\theta}$. Finally, $\psi_{\omega,\theta} \in \mathcal{S}(\mathbb{R})$ and, by Assumption 2.8 and Proposition A.2, its zero set is discrete, so Proposition A.1 shows that the L^1-closure of this set contains $C^\infty_{\mathbb{R},c}(\mathbb{R})$. This is the claim. □

Proof of Theorem 2.12. Let θ^* be an angle attaining the maximum in $\mathcal{L}(\omega^*)$, which exists because the slice loss $\theta \mapsto W_1(P_{\psi_{\omega^*},\theta}, P_{\phi,\theta})$ is continuous and π-periodic. As in the proof of Lemma 2.11, write $P_{\omega,\theta} := P_{\psi_\omega,\theta}$, and set

$$G_\omega := F_{P_{\omega,\theta^*}} - F_{P_{\phi,\theta^*}}, \qquad \ell(\omega) := \int_{-\infty}^{\infty} |G_\omega(x)|\,dx = W_1\big(P_{\omega,\theta^*}, P_{\phi,\theta^*}\big).$$

Fix a direction ξ in the parameter space. Since $\omega \mapsto P_{\omega,\theta^*}$ is continuously differentiable into $\mathcal{S}(\mathbb{R})$, the difference quotients $h_t := t^{-1}(P_{\omega^*+t\xi,\theta^*} - P_{\omega^*,\theta^*})$ converge in $\mathcal{S}(\mathbb{R})$ as $t \downarrow 0$, so that every seminorm of h_t stays bounded for small t. In particular, $C := \sup_{0<t\le t_0}\sup_{y\in\mathbb{R}}(1+|y|)^3|h_t(y)| < \infty$ for some $t_0 > 0$. Moreover, $\int_{\mathbb{R}} h_t\,dy = 0$ because every P_{ω,θ^*} has unit mass. Hence the difference quotients

$$\frac{G_{\omega^*+t\xi}(x) - G_{\omega^*}(x)}{t} = \int_{-\infty}^{x} h_t\,dy = -\int_{x}^{\infty} h_t\,dy$$

converge pointwise to $\partial G_{\omega^*}/\partial\xi = \partial F_{P_{\omega^*,\theta^*}}/\partial\xi$, as in the proof of Lemma 2.11. Estimating the first integral for $x \le 0$ and the second for $x \ge 0$, they are bounded by $C\int_{|x|}^{\infty}(1+s)^{-3}\,ds = \frac{C}{2}(1+|x|)^{-2}$ for all $0 < t \le t_0$. Since $\big||a|-|b|\big| \le |a-b|$, the same bound holds for the difference quotients of $|G_\omega|$. By dominated convergence, ℓ therefore has the one-sided directional derivative

$$\ell'(\omega^*;\xi) := \lim_{t\downarrow 0}\frac{\ell(\omega^*+t\xi)-\ell(\omega^*)}{t} = \int_{\{G_{\omega^*}\neq 0\}} \mathrm{sgn}\big(G_{\omega^*}\big)\,\frac{\partial G_{\omega^*}}{\partial\xi}\,dx + \int_{\{G_{\omega^*}=0\}}\Big|\frac{\partial G_{\omega^*}}{\partial\xi}\Big|\,dx.$$

Since $\mathcal{L} \ge \ell$ with equality at ω^*, the one-sided directional derivatives satisfy $\mathcal{L}'(\omega^*;\xi) \ge \ell'(\omega^*;\xi)$, and $\mathcal{L}'(\omega^*;\xi) = 0$ because $\mathcal{L}$ is differentiable at ω^* with $\nabla\mathcal{L}(\omega^*) = 0$. Adding these inequalities for ξ and $-\xi$ gives $0 \ge 2\int_{\{G_{\omega^*}=0\}}|\partial G_{\omega^*}/\partial\xi|\,dx \ge 0$. Hence the second integral in $\ell'(\omega^*;\xi)$ vanishes, so that $\ell'(\omega^*;\xi) = \int_{\{G_{\omega^*}\neq 0\}}\mathrm{sgn}(G_{\omega^*})\,\partial G_{\omega^*}/\partial\xi\,dx$ depends linearly on ξ. Since $\ell'(\omega^*;\xi) \le 0$ holds for every direction, it also holds for $-\xi$, and linearity gives $-\ell'(\omega^*;\xi) = \ell'(\omega^*;-\xi) \le 0$. Thus $\ell'(\omega^*;\xi) \ge 0$ as well, and hence $\ell'(\omega^*;\xi) = 0$ for every ξ. With $\mathrm{sgn}(0) := 0$, this reads

$$\int_{-\infty}^{\infty}\mathrm{sgn}\big(F_{P_{\omega^*,\theta^*}}(x) - F_{P_{\phi,\theta^*}}(x)\big)\,\frac{\partial F_{P_{\omega^*,\theta^*}}(x)}{\partial\xi}\,dx = 0 \qquad \text{for every direction } \xi.$$

At this point, by Lemma 2.11, the L^1-closure of the set of functions $\partial F_{P_{\omega^*,\theta^*}}/\partial\xi$ contains the real test functions $C^\infty_{\mathbb{R},c}(\mathbb{R})$, and $\mathrm{sgn}\big(F_{P_{\omega^*,\theta^*}}(x) - F_{P_{\phi,\theta^*}}(x)\big)$ is a bounded measurable function. Therefore, by a limiting argument and the fundamental principle of the calculus of variations, $\mathrm{sgn}\big(F_{P_{\omega^*,\theta^*}}(x) - F_{P_{\phi,\theta^*}}(x)\big) = 0$, that is, $F_{P_{\omega^*,\theta^*}}(x) = F_{P_{\phi,\theta^*}}(x)$ holds for almost every x. By the right-continuity of the cumulative distribution function, taking a monotonically decreasing convergent sequence (x_n) for any x such that $F_{P_{\omega^*,\theta^*}}(x_n) = F_{P_{\phi,\theta^*}}(x_n)$ holds, we have

$$F_{P_{\omega^*,\theta^*}}(x) = \lim_{n\to\infty} F_{P_{\omega^*,\theta^*}}(x_n) = \lim_{n\to\infty} F_{P_{\phi,\theta^*}}(x_n) = F_{P_{\phi,\theta^*}}(x).$$

From the above, $P_{\omega^*,\theta^*} = P_{\phi,\theta^*}$ holds, and the slice loss $W_1\big(P_{\omega^*,\theta^*}, P_{\phi,\theta^*}\big)$ at θ^* becomes zero. Since θ^* attains the maximum in $\mathcal{L}(\omega^*)$, the slice losses at all other angles vanish as well, so $\mathcal{L}(\omega^*) = 0$.

Finally, by the injectivity of the Radon transform, the Wigner functions in this case coincide completely, but when reconstructing a wavefunction from the Wigner function, an indeterminacy of a constant phase remains. Hence the coincidence holds up to $U(1)$ difference. □

B Implementation Details

Target state and phase-space domain. We target the center-of-mass mode of a two-mode cat state with $\alpha = 2$, so that the relevant displacement is $\alpha_+ = \sqrt{2}\,\alpha \approx 2.83$. The phase-space variables (x_+, p_+) are restricted to the domain $[-L, L]^2$ with $L = \alpha_+ + 4 \approx 6.83$, giving sufficient margin around the two coherent peaks of the cat state.

Measurement protocol. For each phase $\theta \in \Theta$, we generate $N = 1000$ shots from the noisy marginal $\tilde{Q}_\theta$. The measurement angles Θ are equispaced in $[0, \pi)$, with $N_{\text{angle}} \in \{6, 12, 24\}$ depending on the experiment. Detector binning is implemented on a uniform grid with bin width $|B_{\theta,j}| \approx 0.37$. The empirical CDF $F_{Q_{\theta,N}}$ is constructed from the per-bin counts and bin-mean values as described in Equation 10.

Neural-network architecture. The wave-function ansatz $\psi_\omega : \mathbb{R} \to \mathbb{C}$ is parameterized as a multilayer perceptron (MLP) with input dimension 1, output dimension 2 (real and imaginary parts), depth $D = 8$, and width $W = 128$ in each hidden layer. We use GELU activations, and the entire network output is multiplied by a Gaussian envelope $\exp\bigl(-\beta x^2\bigr)$ with β a learnable scalar parameter initialized at $\beta_0 = 0.05$. The total parameter count is approximately 1.0×10^5. At every forward pass, the output is normalized on a fine 1D grid of 1000 points over $[-10, 10]$ so that $\|\psi_\omega\|_{L^2} = 1$ holds throughout training.

Loss-computation grids. For loss evaluation, the Wigner-function reconstruction integral (Eq. 1) is discretized on grids matched to the detector resolution: an x_+-grid of 41 points and a p_+-grid of 19 points over $[-L, L]$, and a y-integration grid of 35 points over $[-(L-2), L-2]$. Higher-resolution grids (201×201) are used only for final Wigner-function visualization.

Optimizer and training schedule. We use Adam with learning rate 3×10^{-4} and default momenta $(\beta_1, \beta_2) = (0.9, 0.999)$. Each trial is run for a fixed wall-clock budget of 24 seconds of pure optimization time. Within each step, the proposed MSW1 method computes the maximum loss across all N_{angle} angles (full-batch max), while the SW1 baseline samples one angle per step (minibatch mean) for fair comparison of per-step computational cost. No learning-rate schedule, weight decay, or gradient clipping is applied.

Baseline: diluted iterative Maximum Likelihood Estimation. We implement the diluted iMLE of (Řeháček et al., 2007) in a truncated Fock basis of dimension $N_{\text{Fock}} = 200$. The dilution parameter is set to $t = 0.5$, and the initial state is the maximally mixed state $\rho^{(0)} = I_{N_{\text{Fock}}}/N_{\text{Fock}}$. After each update, we explicitly Hermitize the iterate $\rho^{(k+1)}$ to suppress numerical drift from positivity. iMLE is run on the same hardware and within the same 24-second wall-clock budget as the proposed method, under the same per-step accounting (excluding fidelity/likelihood evaluation overhead).

Remark B.1 (Choice of N_{Fock})**.** The center-of-mass mode of the target is a cat state of coherent amplitude $\alpha = 2$ with mean photon number about 4. Its population above 17 photons is below 10^{-6}, and the added noise at $\sigma \leq 0.9$ extends this threshold only to about 34 photons. A truncation at $N_{\text{Fock}} = 200$ therefore represents every state in our experiments, including the noisy ones, so the failure modes of the full-rank iMLE cannot be attributed to insufficient truncation. For the rank-1 iMLE, however, a large truncation is not harmless, presumably because it leaves the likelihood more room to overfit the binned data. With a Fock truncation of dimension 40, which still covers all states in our experiments, the rank-1 iMLE reaches fidelities of about 0.98, 0.59, and 0.35 at $\sigma = 0.3$, 0.6, and 0.9: it then exceeds the proposed method under mild noise, while under stronger noise it remains below it with a large spread over trials.

Statistical protocol. All experiments are repeated over 10 independent trials with seeds $\{1000, 1001, \ldots, 1009\}$, controlling both data sampling and neural-network initialization. Reported values are means and standard deviations across the 10 trials.

Hardware. All experiments are run on a single Apple M4 Pro using PyTorch's MPS backend for the proposed method, while the iMLE baseline runs on the same machine via NumPy on the CPU, which is the natural backend for iMLE since its inner loop is dominated by dense complex

matrix products for which CPU BLAS is highly competitive at $N_{\text{Fock}} = 200$. We verified that this choice does not disadvantage iMLE: in the regime of interest, iMLE typically completes hundreds of iterations within the 24-second budget, well past the point where its fidelity has plateaued, so the bottleneck is statistical (under-sampling of the marginals), not computational.

Algorithm 1: Neural Quantum State Tomography with the MSW_1 loss

Data: Measurement data $\mathcal{D} = \big\{(\theta, \{x_\theta^{(m)}, \mu_\theta^{(m)}\}_{m=1}^{M_\theta})\big\}_{\theta\in\Theta}$, phase-space cutoff L, integration cutoff $y_{\max}$, and discretization steps dx, dp, dy.
Result: Trained wave-function network $\psi_\omega(x) = u_\omega(x) + i\, v_\omega(x)$.
Initialize neural-network parameters ω;
while *training has not converged* **do**
  `/* 1. Wave-function normalization on a wide auxiliary grid */`
  $\mathcal{N} \leftarrow \Big(\int_{-X_{\text{norm}}}^{X_{\text{norm}}} |\psi_\omega(x')|^2\, dx'\Big)^{1/2}$ ($X_{\text{norm}} \gg L$, taken as 10 in our experiments);
  $\tilde{\psi}_\omega(x) \leftarrow \psi_\omega(x)/\mathcal{N}$;
  foreach $\theta \in \Theta$ **do**
    `/* 2. Compute the rotated Wigner function on the evaluation grid */`
    foreach *grid point* $(x_i, p_j) \in [-L, L] \times [-L, L]$ **do**
      $x_\theta \leftarrow x_i \cos\theta + p_j \sin\theta, \quad p_\theta \leftarrow x_i \sin\theta - p_j \cos\theta$;
      $W_{\omega,\theta}(x_i, p_j) \leftarrow \frac{1}{\pi}\int_{-y_{\max}}^{y_{\max}} \text{Re}\Big[\overline{\tilde{\psi}_\omega(x_\theta + y)}\, \tilde{\psi}_\omega(x_\theta - y)\, e^{2ip_\theta y}\Big]\, dy$;
    `/* 3. Marginalize along` p_j `to obtain the 1D model distribution */`
    $P_{\omega,\theta}(x_i) \leftarrow \int_{-L}^{L} W_{\omega,\theta}(x_i, p_j)\, dp_j$;
    $P_{\omega,\theta}(x_i) \leftarrow \text{Normalize}\big(\max\{0,\, P_{\omega,\theta}(x_i)\}\big)$;
    `/* 4. Per-slice 1D Wasserstein distance via CDFs */`
    Predicted CDF: $F_{P_{\omega,\theta}}(x) \leftarrow \int_{-\infty}^{x} P_{\omega,\theta}(x')\, dx'$;
    Empirical CDF: $F_{Q_{\theta,N}}(x) \leftarrow \sum_m \mu_\theta^{(m)}\, \chi\big[x_\theta^{(m)} \le x\big]$;
    $\mathcal{L}_\theta \leftarrow \int \big|F_{P_{\omega,\theta}}(x) - F_{Q_{\theta,N}}(x)\big|\, dx$;
  `/* 5. Aggregate slice losses via the max operator (`MSW_1`) */`
  $\mathcal{L}(\omega) \leftarrow \max_{\theta\in\Theta} \mathcal{L}_\theta$;
  Update ω with one Adam step on $\mathcal{L}(\omega)$;
return ψ_ω;

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Theoretical claims are stated with full assumptions in Section 2.3, and the experimental scope is described in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Discussed in Remarks 2.10, 2.13, and 2.14, and in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All assumptions are stated in Section 2.3, and complete proofs are given in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Full implementation details are provided in Appendix B and Algorithm 1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code is included in the supplemental material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All hyperparameters and settings are specified in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All results report means and standard deviations over 10 trials (Section 4, Appendix B).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Hardware and runtime details are given in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work involves no human subjects, no sensitive data, and no dual-use concerns.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is foundational research in quantum state estimation with no direct negative societal implications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.