

Efficient learning of lattice gauge theories with fermions

Shreya Shukla^{*,*}, Yukari Yamauchi^{†,†}, Andrey Y. Lokhov^{‡,‡}, Scott Lawrence^{§,§}, and Abhijith Jayakumar^{||,||}

Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA



(Received 6 January 2026; accepted 16 June 2026; published 28 July 2026)

We introduce a learning method for recovering action parameters in lattice field theories. Our method is based on the minimization of a convex loss function constructed using the Schwinger-Dyson relations. We show that score matching, a popular learning method, is a special case of our construction of an infinite family of valid loss functions. Importantly, our general Schwinger-Dyson-based construction applies to gauge theories and models with Grassmann-valued fields used to represent dynamical fermions. In particular, we extend our method to realistic lattice field theories including quantum chromodynamics.

DOI: [10.1103/xlz6-yzxc](https://doi.org/10.1103/xlz6-yzxc)

I. INTRODUCTION

Lattice field theory provides systematic, nonperturbative access to a wide range of quantum systems, most famously including quantum chromodynamics (QCD), the theory of strong interactions. Given a family of lattice field theories (such as lattice QCD), there are two computational tasks one might perform. Most commonly, we fix a set of parameters and couplings, and compute expectation values over the probability distribution thus defined. The inverse task can also be defined: given a set of expectation values, what lattice action was used to compute them?

This inverse task, sometimes called the learning problem, is the subject of this work. The learning problem is not a mere mathematical curiosity: it is motivated by a number of applications. One of the central objects in quantum field theory is the renormalization flow. By sampling from the Euclidean lattice path integral at one set of parameters, coarse graining, and then solving the learning problem, we can directly study how the coupling constants change under renormalization [1,2]. Renormalization flow can be, and commonly is, studied by a number of methods, including perturbation theory and other analytic expansions [3], Monte Carlo studies [4], and the functional renormalization group [5]. Renormalization studies as proposed here—performed by sampling from lattice data, coarse graining,

and then learning—are fully nonperturbative, apply to any theory that can be efficiently sampled on the lattice, and (uniquely) are equally applicable to renormalization studies in which the degrees of freedom differ between the UV and IR theories.

In principle, one can solve this learning problem by repeatedly performing the forward procedure, for different coupling constants, until the expectation values are found to match. This is inefficient for two reasons. First, computing expectation values is computationally difficult; second, the space of possible couplings can be high-dimensional and therefore expensive to search through. This paper is motivated by the desire for a more efficient approach. Indeed, our approach is inspired by an observation that the learning problem can often be solved *more* efficiently than the forward problem [6].

In the case of certain statistical models, the learning problem is efficiently solved by a popular method known as *score matching* [7] that avoids the computation of the partition function. Standard score matching may be applied to lattice field theories with continuous variables, as was done in [2], but encounters computational difficulties in the case of Grassmann variables that are used to model fermions in important physical theories. Moreover, score matching requires observation of a specific set of correlators that may not be conveniently obtained, and might not be the most data-efficient estimator. This motivates a search for a more general family of estimators that would conveniently apply to relevant lattice gauge theories.

Our work begins with the observation that a set of equations known as Schwinger-Dyson relations can be used as a foundation for a learning method. The Schwinger-Dyson relations are a set of constraints on the expectation values of a statistical system, which can be efficiently determined by inspection of the action, with only a trivial amount of computation. These relations have been used before as a method of variance reduction [8–10]; in this

*Contact author: sshukla@lanl.gov

†Contact author: yyamauchi@lanl.gov

‡Contact author: lokhov@lanl.gov

§Contact author: srlawrence@lanl.gov

||Contact author: abhijithj@lanl.gov

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP³.

work we show how they may be used as the basis of a learning algorithm. In Sec. II, we show how to construct a broad family of efficient loss functions based on suitable subset of Schwinger-Dyson relations. Interestingly, we show that score matching is a special case of our construction corresponding informed by the sufficient statistics of the model. Although we leave detailed analysis of sample and computational complexity to future work, we demonstrate in Appendix A that score matching is in general not the most efficient member of this family of estimators.

While score matching is not obviously applicable to theories of Grassmann variables, Schwinger-Dyson equations do exist for such theories. In Sec. III we show that these Schwinger-Dyson relations may be used for learning of action parameters given expectation values. We demonstrate this approach on a lattice Thirring model.

Arguably the most important lattice field theories, including lattice QCD, are gauge theories. In Sec. IV we discuss the interaction of Schwinger-Dyson-based learning and score matching with gauge symmetry. We further detail how these learning methods may be used in the case of full-fledged lattice QCD. Finally in Sec. V we consider the prospects for, and limitations of, these learning methods.

II. LEARNING METHOD

In this work we are concerned with a particular type of family of probability distributions: exponential families. These are families of probability distributions $p(x)$ that have the form

$$p(x) = \frac{1}{Z} e^{-S_0(x) - \sum_i \theta_i^* Q_i(x)}, \quad (1)$$

for known functions S_0 and Q_i and potentially high-dimensional x , and Z is the normalization often referred to as the partition function. Such families are widespread in both statistical and quantum field theories. The coefficients $\{\theta_i^*\}$ (collectively denoted as θ^*) are termed the couplings and specify a member of the exponential family. In the context of field theory there is an additional condition: the functions S_0 and Q_i are polynomials in the “fields” x .

The Schwinger-Dyson relations [11–13] are an infinite tower of linear constraints on expectation values, commonly used in the context of quantum field theory. In this Section, we show how they can be used as a basis for a learning method. We also show that score matching [7], a widely used method in machine learning literature to solve the learning problem associated with the exponential family of distributions, emerges as just one choice among the infinite family of loss functions that can be constructed by considering the Schwinger-Dyson relations.

As simple illustration of the main ideas in this work, consider the following system, which may be considered a

0-dimensional ϕ^4 theory with expectation values defined as follows:

$$\langle f(x) \rangle = \int dx f(x) e^{-\frac{\alpha^*}{2} x^2 - \frac{\lambda^*}{4} x^4}, \quad (2)$$

for scalar x . The learning problem is as follows: given a few such expectation values (perhaps $\langle x^2 \rangle$ and $\langle x^4 \rangle$), determine the values of the “couplings” α^* and λ^* . For example, if we are promised $\lambda^* = 0$, this is particularly easy, as we may immediately calculate $\langle x^2 \rangle = (\alpha^*)^{-1}$.

In the case where λ^* does not vanish, expectation values of the form $\langle x^{2k} \rangle$ are not straightforwardly expressed as a function of the couplings. However, a subset of expectation values are analytically related to each other by the couplings, and we can use these to determine the unknown couplings that define the theory, as we show below.

A. Learning from Schwinger-Dyson relations

Analytical relations between expectation values are termed Schwinger-Dyson relations, and are easily obtained via integration by parts. For any distribution over $\mathbb{R}$ of the form $e^{-S(x)}$ [with polynomial $S(x)$ of even degree ≥ 2], and any polynomial $P(x)$, observe that

$$\int_{-\infty}^{+\infty} dx \frac{d}{dx} (P(x) e^{-S(x)}) = 0. \quad (3)$$

This expression generalizes to higher-dimensional spaces (like $\mathbb{R}^N$) and compact integration domains [like $SU(N)$]. It immediately follows that, for any first-derivative operator ∂ and any function $P(x)$ such that $e^{-S(x)} P(x)$ is sufficiently quickly decaying, we have the following linear constraint on the expectation values:

$$\langle P(x) \partial S(x) \rangle = \langle \partial P(x) \rangle. \quad (4)$$

In the case of our 0-dimensional ϕ^4 -like example, the general Schwinger-Dyson relation reads

$$k \langle x^{k-1} \rangle = \alpha^* \langle x^{k+1} \rangle + \lambda^* \langle x^{k+3} \rangle. \quad (5)$$

When the parameters α^*, λ^* are known, this provides a useful constraint on the expectation values. In our case, the expectation values are known but the action parameters are not. The above equation makes clear that knowing $\langle x^2 \rangle$, $\langle x^4 \rangle$, and $\langle x^6 \rangle$ is sufficient to determine α^*, λ^* by linear algebra.

The remainder of this paper will consist of variations and elaborations on this theme. We are provided (perhaps stochastic estimates of) expectation values of some theory. We know the action of the theory has a particular form, but the coupling constants in this action are not known. We wish to determine the couplings. In order to do this, we write down a sufficiently large set of Schwinger-Dyson

relations. We find the couplings that would minimize the amount of violation of these Schwinger-Dyson relations, and these couplings are our estimate of the couplings that were used to produce the original observables.

Let us formalize this procedure for the generic model of the type (1). First, for a choice of n test polynomials $P_1, \dots, P_n$, define the residual vector

$$r_j(\theta) = \langle \partial P_j \rangle - \langle P_j \partial S \rangle, \quad (6)$$

where $S = S_0 + \sum_k \theta_k Q_k$ denotes action parametrized by the optimization parameters θ . This residual vanishes when evaluated at the true parameters: $r_j(\theta^*) = 0$. It will also be convenient to define

$$M_{ji} = \langle P_j \partial Q_i \rangle \quad \text{and} \quad b_j = \langle \partial P_j \rangle - \langle P_j \partial S_0 \rangle. \quad (7)$$

In this notation the residual is defined by $r_j(\theta) = b_j - \sum_i M_{ji} \theta_i$.

With the residual defined, we can construct a convex loss function from any positive-definite quadratic form. Concretely, for any positive definite symmetric matrix A we define

$$\mathcal{L}_A(\theta) = \sum_{i,j} r_i^T(\theta) A_{ij} r_j(\theta). \quad (8)$$

Minimization of the loss function defined in (8) for a certain choice of test polynomials $\{P_i\}$ and positive definite matrix A constitutes our proposed learning method for lattice gauge theories. When provided with exact expectation values—that is, when there are no statistical errors—this loss function is minimized precisely when $r = 0$, at which point the true couplings θ^* are recovered. In the case of finite samples, the Schwinger-Dyson equations will not be satisfied exactly, but the $\hat{\theta}$ that minimizes $\mathcal{L}_A$ results in an estimator for the true parameters θ^* . The estimator properties will crucially depend on the choice of both $\{P_i\}$ and A .

Suppose that we have chosen some basis of polynomials $P_1, \dots, P_n$. Under what circumstances is the knowledge of the corresponding Schwinger-Dyson expectation values sufficient to uniquely fix the parameters θ^* ? As we discuss below, the choice $P_i = \partial Q_i$ is sufficient for estimating the model parameters (and any choice that is a superset will also suffice), but how much freedom in the choice of test polynomials $\{P_i\}$ is available? Intuitively, as long as the $\{P_i\}$ are independent and “generic” with respect to the known lattice symmetries,¹ and there are at least as many of

them as there are free parameters, the parameter estimates $\hat{\theta}$ should be uniquely determined.

We can understand geometrically the conditions under which the parameters θ^* are uniquely fixed, as follows. We begin by defining an inner product on the space of polynomials,

$$\langle f, g \rangle_p = \int dx p(x) f(x) g(x). \quad (9)$$

This inner product also defines the orthogonal space $\mathcal{V}^\perp$ of a vector space $\mathcal{V}$. With these definitions we observe that the action of the matrix M may be written in terms of this inner product,

$$(Mv)_i = \left\langle P_i, \sum_j \partial Q_j v_j \right\rangle_p. \quad (10)$$

Assuming positive-definite A , the minimum of the loss function Eq. (8) is unique if and only if $\ker M = \{0\}$. By inspection of Eq. (10), the kernel of M is precisely those vectors v that produce linear combinations $v_j \partial Q_j$, which are orthogonal to all linear combinations of the $\{P_i\}$. The kernel vanishes if and only if there is no linear combination of the $\{\partial Q_j\}$ that yields a function orthogonal to all P_i . Thus a condition both necessary and sufficient for θ^* to be uniquely determined is,

$$\text{span}(\{P_i\})^\perp \cap \text{span}(\{\partial Q_j\}) = \{0\}. \quad (11)$$

Assume that this condition does not hold, i.e., there exists $q \in \text{span}(\{P_i\})^\perp \cap \text{span}(\{\partial Q_j\})$ such that $q \neq 0$. Since $q \in \text{span}(\{\partial Q_j\})$, there exists $v^* \neq 0$ such that $q = \sum_j v_j^* \partial Q_j$, implying that $\langle P_i, \sum_j v_j^* \partial Q_j \rangle_p = 0$ for all i and $v^* \in \ker M$. For a solution θ_0 of the Schwinger-Dyson equations, we have for any $t \in \mathbb{R}$

$$M(\theta_0 + tv^*) = b + 0 = b, \quad (12)$$

and v^* manifests as a flat direction in the loss function Eq. (8), implying that we no longer have a unique minimum.

Note that the condition Eq. (11) cannot in general be checked until expectation values are known, as the definition of $\mathcal{V}^\perp$ depends on the expectation values with respect to p .

B. Relation to score matching

Minimization of the loss function defined in (8) for any sufficient set of test polynomials $\{P_i\}$ discussed above and any positive definite matrix A yields parameter estimates $\hat{\theta}$ that converge to θ^* in expectation. Here, we discuss an important special case for the choice of $\{P_i\}$ and A that recovers a widespread learning method for estimating

¹In the one-dimensional example above, note that the Schwinger-Dyson relations generated by $P_k = x^{2k}$ provide no information whatsoever about λ^* , precisely because of the $\mathbb{Z}_2$ symmetry.

probability distributions with continuous variables: score matching [7]. The loss function of score matching corresponds to a choice of test polynomials which is guaranteed to be sufficient, and in part for this reason, we will use score matching as the basis of our investigation of gauge theories in Sec. IV.

We start by providing a quick background on score matching. As before, assume that expectation values with respect to an exponential family distribution $p(x)$ of the type (1) are known, but parameters θ^* are unknown. Define an ansatz distribution $\pi(x; \theta)$ in the same family, parametrized by a vector θ . To find θ^* , we minimize the following *score-matching* loss function over the parameters θ ,

$$\mathcal{L}(\theta) = \left\langle \left\| \frac{\partial}{\partial x} \log p(x) - \frac{\partial}{\partial x} \log \pi(x; \theta) \right\|^2 \right\rangle_{p(x)}, \quad (13)$$

where $\frac{\partial}{\partial x} \log p(x)$ is often termed the *score function* of the distribution $p(x)$. In the score matching loss function, the expectation value is taken over the empirical probability distribution $p(x)$. In the multidimensional case, $\frac{\partial}{\partial x}$ denotes the vector-valued gradient, and $\|\cdot\|$ denotes the L_2 norm of the vector of partial derivatives. Notice, using the fact that p vanishes nowhere, that θ^* is the unique minimum (and in fact the unique zero) of the loss function $\mathcal{L}(\theta)$.

As written, it is not practical to minimize this loss, because evaluating $\mathcal{L}(x)$ requires direct knowledge of $p(x)$. The fact that expectation values are taken over $p(x)$ is not an obstacle—we know such expectation values by assumption—but the fact that the precise expectation value to be taken explicitly depends on $\log p(x)$, is. However, with some algebraic manipulation (see [7] for details), we can rewrite the same loss function, up to a θ -independent constant, as

$$\mathcal{L}(\theta) = \left\langle \left(\frac{\partial}{\partial x} S_\theta \right)^2 \right\rangle - 2 \left\langle \frac{\partial^2}{\partial x^2} S_\theta \right\rangle. \quad (14)$$

Here we have labeled by S_θ the logarithm of the ansatz distribution $\pi(x; \theta)$. This defines the score matching loss used in practice: this form of the loss function is readily evaluated, given samples over the empirical distribution $p(x)$.

Let us illustrate the loss (14) with our toy example of the 0-dimensional ϕ^4 theory, with the trial distribution $\pi(x; \alpha, \lambda) \propto e^{-\frac{\alpha}{2}x^2 - \frac{\lambda}{4}x^4}$. Hence, the action and its first two derivatives are given by

$$\begin{aligned} S_\theta &= \frac{\alpha}{2}x^2 + \frac{\lambda}{4}x^4, \\ dS_\theta &= \alpha x + \lambda x^3, \quad d^2S_\theta = \alpha + 3\lambda x^2. \end{aligned} \quad (15)$$

As a result we can write the score-matching loss function directly in terms of low-order expectation values, obtaining,

$$\mathcal{L}(\alpha, \lambda) = (\alpha^2 - 6\lambda)\langle x^2 \rangle + 2\alpha\lambda\langle x^4 \rangle + \lambda^2\langle x^6 \rangle - 2\alpha. \quad (16)$$

Given the expectation values $\langle x^{2k} \rangle$ for $k = 1, 2, 3$, this loss function is minimized at the true parameters α_*, λ_* .

Now let us see how score matching connects to the Schwinger-Dyson relations discussed above for the distributions of the type (1). At the minimum of the loss function $\mathcal{L}(\theta)$ defined in (14), derivatives with respect to θ must vanish,

$$\frac{\partial}{\partial \theta_i} \mathcal{L}(\theta) = 2\langle (\partial Q_i)(\partial S) \rangle - 2\langle \partial^2 Q_i \rangle = 0. \quad (17)$$

Defining $P_i = \partial Q_i$, this is precisely a Schwinger-Dyson relation of the form of Eq. (4). This correspondence tells us that the minimum of the loss function automatically satisfies a certain subset of the Schwinger-Dyson equations,² determined by the free parameters θ that were used to define the ansatz $\pi(x; \theta)$.

We can also go in the other direction, showing that the score matching loss is a special case of Eq. (8). In order to recover the score loss, take the test polynomials to be $P_j = \partial Q_j$. Then $M_{ji} = \langle \partial Q_j \partial Q_i \rangle$, and taking $A = M^{-1}$, we recover the score-matching loss,

$$\begin{aligned} \mathcal{L}^{\text{SM}}(\theta) &= \sum_{i,j} r_i^T(\theta) M_{ij}^{-1} r_j(\theta) \\ &= b^T M b - 2b^T \theta + \theta^T M \theta. \end{aligned} \quad (18)$$

To summarize, we showed that score matching loss that has an interpretation of matching of score functions in (13), is a special case of our family of estimators defined in Eq. (8). These generalized family can be constructed starting with a sufficient number of Schwinger-Dyson relations that need to be satisfied at the minimum of the loss function. With this principle in mind, in the next Section we show how the general construction is essential for dealing with Grassmann variables that are used to model fermions.

III. GRASSMANN VARIABLES

Path integrals for fermionic theories are commonly constructed as integrals over Grassmann variables.

²When performing score matching we *minimize* the score function, rather than merely seeking a saddle point. Just as the saddle-point condition corresponds to (some subset of) the Schwinger-Dyson relations, it may be seen that the minimization condition—the Hessian of $\mathcal{L}$ must be positive semidefinite—corresponds to (some subset of) positivity relations $\langle f^2 \rangle \geq 0$. This fact presumably has implications for the interplay of score-matching and sign problems, which we will not address in this work.

Functions of Grassmann variables are not maps from one space to another, but instead formal expressions, with rules for linear combination, multiplication, differentiation, and integration. For completeness, we include a brief review of the properties of Grassmann variables, and path integrals built off of them, in Appendix B. We refer the reader to a standard reference (e.g., [14]) for more details.

A lattice path integral for a fermionic system typically has the form

$$Z = \int \left(\prod_i d\bar{\eta}_i d\eta_i \right) e^{-S(\bar{\eta}, \eta)}. \quad (19)$$

Here $\bar{\eta}_i$ and η_i are independent Grassmann variables on lattice site i , which are by convention taken to transform into each other under complex conjugation. Needless to say, this path integral does not directly correspond to any statistical system in the usual sense. The action $S(\bar{\eta}, \eta)$ is not describing the logarithm of a probability distribution on any space. Instead it is to be treated as a formal expression. Nevertheless, expectation values are defined in the usual way, and we can ask to determine properties of $S(\bar{\eta}, \eta)$ given those expectation values.

A. Generalities

To guide the eye, we will again work with a concrete example before giving the general procedure. We consider a one-site model, in which the action is necessarily quadratic in the Grassmann degrees of freedom. The partition function of this model is given by

$$Z_1(\theta^*) = \int d\bar{\eta} d\eta e^{-\theta^* \bar{\eta} \eta}, \quad (20)$$

and is readily evaluated to be $Z_1(\theta^*) = \theta^*$. Expectation values in this system are defined in the usual way. A straightforward application of the rules of Grassmann arithmetic shows that the only nontrivial expectation value is $\langle \eta \bar{\eta} \rangle = (\theta^*)^{-1}$.

This explicit evaluation of the expectation value immediately translates into a learning algorithm, but we can get the same result in a more generalizable way by considering Schwinger-Dyson relations. The Schwinger-Dyson relations are derived in the same way as in the bosonic case, and for this one-site model there are two,

$$\begin{aligned} \left\langle \frac{\partial}{\partial \eta} \right\rangle &= - \left\langle \eta \frac{\partial}{\partial \eta} \theta^* \bar{\eta} \eta \right\rangle, \text{ and} \\ \left\langle \frac{\partial}{\partial \bar{\eta}} \right\rangle &= - \left\langle \bar{\eta} \frac{\partial}{\partial \bar{\eta}} \theta^* \bar{\eta} \eta \right\rangle. \end{aligned} \quad (21)$$

Thanks to the symmetry $\bar{\eta} \leftrightarrow \eta$, these two expressions are identical, and we recover $1 = \theta^* \langle \eta \bar{\eta} \rangle$ as expected above. Note that the Schwinger-Dyson equations which would be

derived by considering (for example) $\int d\bar{\eta} d\eta \frac{\partial}{\partial \eta} (\bar{\eta} e^{-S})$ trivially vanish.

At this point it is straightforward to construct a loss function by means of Eq. (8). The Schwinger-Dyson relations above define the residual vector r , and we may take A to be the identity. The following loss function is thus obtained:

$$\mathcal{L}(\theta) = (1 - \theta \langle \bar{\eta} \eta \rangle)^2. \quad (22)$$

We may now generalize to the case of arbitrarily many Grassmann fields. The general Schwinger-Dyson relation has the form

$$\langle \partial_{\eta/\bar{\eta}} P \rangle = - \langle P \partial_{\eta/\bar{\eta}} S \rangle, \quad (23)$$

for Grassmann-odd polynomials P . For a sufficiently large basis of such relations and expectation values, one can recover the parameters of the action. The general rules and conditions for identifiability mirror those mentioned in Sec. II.

Notice that score matching as defined in (13) does not apply to Grassmann theories in an unambiguous way, because a suitable norm must be defined on Grassmann variables. In the one-dimensional example above, a sensible score-matching loss function is

$$\mathcal{L}(\theta) = \left\langle \left\| \frac{\partial S(\theta_*)}{\partial \eta} - \frac{\partial S(\theta)}{\partial \eta} \right\|^2 \right\rangle. \quad (24)$$

Here the norm $\|f\|^2$ is defined as $\bar{f}f$, where Grassmann variables are transformed under complex conjugation in the conventional way as $\bar{\eta}$. Direct evaluation of this loss function reveals that it has a unique minimum at $\theta = \theta_*$ as desired. However we can already see that this choice is “brittle”: if instead we had taken the derivatives with respect to $\bar{\eta}$, then the expectation value would be negative $\mathcal{L}(\theta)$ would be unbounded below.

Once again, this loss cannot be estimated just from observations as it explicitly depends on the unknown parameter, θ^* , that we are trying to learn. As in Sec. II, we must remove the explicit dependence of $\mathcal{L}(\theta)$ on θ^* , dropping all θ -independent terms. The result is

$$\mathcal{L}(\theta) = \left\langle \frac{\partial S_\theta}{\partial \bar{\eta}} \frac{\partial S_\theta}{\partial \eta} \right\rangle + 2 \left\langle \frac{\partial}{\partial \eta} \frac{\partial}{\partial \bar{\eta}} S_\theta \right\rangle + \text{const.} \quad (25)$$

This loss function, a naïve translation of the score-matching loss to the Grassmann case, is usable only in the special case where the action S is quadratic in the Grassmann variables. Higher-order terms in the action vanish when the norm is taken (because the norm defined above obeys $\|\bar{\eta} \eta\| = 0$).

B. Lattice Thirring model

To demonstrate learning in a theory with Grassmann fields, we consider a lattice Thirring model defined with staggered fermions. This is a two-dimensional statistical theory with a single Grassmann variable χ at each site. For any polynomial $\mathcal{O}$ of the Grassmann fields, an expectation value is defined by the Grassmann path integral

$$\langle \mathcal{O} \rangle = \frac{\int d\bar{\chi} d\chi \mathcal{O}(\bar{\chi}, \chi) e^{-S_{\text{Th}}(\bar{\chi}, \chi)}}{\int d\bar{\chi} d\chi e^{-S_{\text{Th}}(\bar{\chi}, \chi)}}. \quad (26)$$

The action S_{Th} consists of the standard action of free staggered fermions [15], combined with a quartic vector interaction,

$$S_{\text{Th}} = m \sum_x \bar{\chi}_x \chi_x + \sum_{\langle x, x' \rangle} \bar{\chi}_x M_{xx'} \chi_{x'} + \frac{g^2}{2} \sum_{x, \nu} j_\nu^2(x). \quad (27)$$

Here the second sum is taken over all pairs of neighboring lattice sites on an $L \times L$ lattice with periodic boundary conditions, and $\nu = 0, 1$ denotes the Euclidean directions. The massless staggered Dirac operator is

$$M_{xx'} = \sum_{\nu=0}^1 \frac{\eta_{x, \nu}}{2} (\delta_{x+\hat{\nu}, x'} - \delta_{x-\hat{\nu}, x'}), \quad (28)$$

$$\eta_{x,0} = 1, \quad \eta_{x,1} = (-1)^{x_0},$$

while the conserved staggered vector current on the link (x, ν) is

$$j_\nu(x) = \frac{\eta_{x, \nu}}{2} (\bar{\chi}_x \chi_{x+\hat{\nu}} + \bar{\chi}_{x+\hat{\nu}} \chi_x). \quad (29)$$

In the continuum limit, this results in a theory of two flavors of (two-component) Dirac fermions. For our purposes we will not need to consider the continuum limit: we are only interested in the lattice theory on its own.

In order to numerically simulate the Thirring model (via lattice Monte Carlo simulation), we introduce a bosonic auxiliary field in such a way that the action is quadratic in the fermion fields. This allows the Grassmann variables to be integrated out analytically. Using the Hubbard-Stratonovich formalism, we have for each link (x, ν) ,

$$\exp\left(-\frac{g^2}{2} j_\nu\right) \propto \int dA_\nu \exp\left[-\frac{A_\nu^2}{2} - igA_\nu j_\nu\right]. \quad (30)$$

The resulting action, after integrating out the fermions, depends only on the auxiliary vector field A ,

$$S_{\text{Th,bose}} = -\log \det D[A, g] + \frac{1}{2} \sum_{x, \nu} A_\nu^2(x). \quad (31)$$

The staggered Dirac operator in the A background is

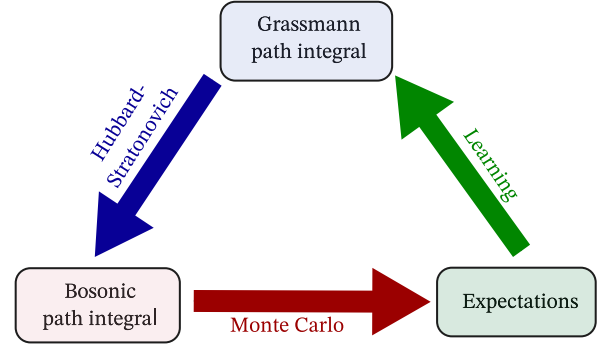


FIG. 1. Workflow for the demonstration of learning of Grassmannian actions in the case of the lattice Thirring model. We begin with an action written purely in terms of Grassmann fields. Bosonization (via Hubbard-Stratonovich) yields a theory in which the fermions can be integrated out, leaving only c-number fields. Expectation values in this theory are obtained by Monte Carlo sampling, and from those expectation values we use Schwinger-Dyson-based learning to recover directly the original (purely Grassmann) theory.

$$D_{xx'}[A, g] = m\delta_{x, x'} + \sum_{\nu=0}^1 \frac{\eta_{x, \nu}}{2} [(1 + igA_\nu(x))\delta_{x+\hat{\nu}, x'} - (1 - igA_\nu(x))\delta_{x-\hat{\nu}, x'}]. \quad (32)$$

As detailed in Appendix C, the Schwinger-Dyson relations used are those defined by taking $P = \partial_{\chi_k} Q$ in the equation,

$$\langle \partial_{\chi_k} P \rangle = -\langle P \partial_{\chi_k} S \rangle, \quad (33)$$

for $Q \in \{Q^{(1)}, Q^{(2)}\}$, where $Q_{ab}^{(1)} = \bar{\chi}_a \chi_b$ and $Q_{ap}^{(2)} = \bar{\chi}_{a+\rho} \chi_{a+\rho} \bar{\chi}_p \chi_a$.

To demonstrate Schwinger-Dyson learning of Grassmann-valued field theories, we perform the following procedure, portrayed in Fig. 1. We first choose bare couplings m and g . Using $S_{\text{Th,bose}}$, we sample field configurations A . From these configurations we compute the fermion propagator, and from there, a set of observables sufficient to infer the bare parameters m and g . Finally, we fit the Schwinger-Dyson relations obtained from S_{Th} into the loss in Eq. (8) after setting the matrix $A = \mathbb{I}$, obtaining estimated bare couplings $\tilde{m}$ and $\tilde{g}$. This style of demonstration is selected in part to show that the Schwinger-Dyson learning procedure operates directly on (formal) expectation values, rather than requiring samples from a probability distribution.

Note that the advantage of the developed procedure is that unlike in [2], we are able to directly learn the original action, which is defined on Grassmann-valued fields, although we cannot sample from this action because it does not define a probability distribution on any space. The two systems are related by an exact duality: all expectation

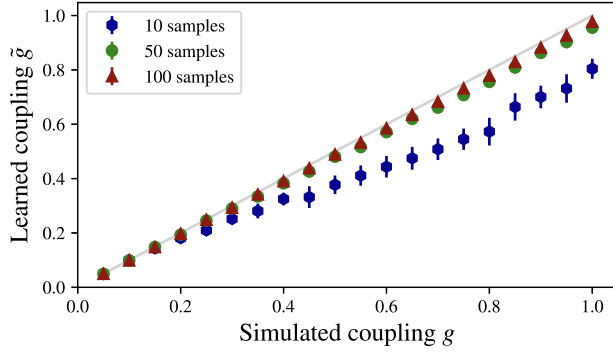


FIG. 2. Estimates of the bare coupling of the Thirring model, as the bare coupling used for simulation is changed. All calculations are performed at a bare mass of $m = 0.1$ on a 12×12 lattice.

values are the same on both sides of the Hubbard-Stratonovich transformation.

We now detail the procedure for using Schwinger-Dyson equations to learn the couplings in S_{Th} , given samples from $S_{\text{Th,bose}}$. From each of these samples we compute the Dirac operator D as defined by Eq. (32). This is inverted to obtain a propagator, again per sample, D^{-1} . This propagator allows the computation of fermion multipoint functions via Wick contraction (after which an average is taken over all configurations).

If we had exact expectation values then we would learn the couplings by enforcing Eq. (33) for some sufficiently large set of (Grassmann-odd) polynomials P . The expectation values estimated by averaging over lattice configurations come with statistical errors and will not satisfy these Schwinger-Dyson relations exactly. Therefore our strategy is to construct a convex loss function of the couplings, as in Eq. (8). This approach also allows us to make use of more Schwinger-Dyson relations than are strictly necessary (in the case of exact data) to uniquely determine the couplings. The linear system constructed from these relations is therefore overdetermined.

Consider the vector $r_{i,xy} = \langle \partial_{\chi_k} P_{i,xy} \rangle + \langle P_{i,xy} \partial_{\chi_k} S \rangle$, where $P_{1,ab} = \partial_{\chi_k} Q_{ab}^{(1)}$ and $P_{2,ap} = \partial_{\chi_k} Q_{ap}^{(2)}$. The loss is

$$\mathcal{L}(m, g^2) = \sum_{i,x,y} r_{i,xy}^T(m, g^2) \cdot r_{i,xy}(m, g^2), \quad (34)$$

and is quadratic in m and g^2 . The observables in this case are the objects $\langle \partial P \rangle$ and $\langle P \partial S \rangle$. A different choice of P will result in a different set of observables, and as long as it satisfies Eq. (11), learning is feasible.

Finally, we solve for the optimal couplings by minimizing this loss function—this amounts only to solving a linear system. The results of this learning procedure are shown in Fig. 2. We examine multiple values of the coupling g , holding the mass $m = 0.1$ and the lattice size $L = 12$ fixed. All parameters are reported in lattice units $a = 1$. For each value of bare coupling g , two fits are performed, one using

50 samples and the other using 100. Errors shown are obtained with the statistical bootstrap, and represent one standard deviation of the spread of couplings reported by the learning algorithm.

Figure 2 makes apparent two facts about the learning algorithm. First, quite a small number of samples is sufficient to obtain reasonably high precision (at the strongest couplings, a few percent) in the determination of the bare coupling. Second, the estimator defined by the learning algorithm is *biased* for finite samples (recovering the correct parameters in the limit of the large number of samples). We note that the presence of a bias is standard and explicitly appears in the analysis of nearly all learning algorithms [6].

Although a small number of samples is sufficient to obtain a reasonable estimate of the coupling constants, more samples can be used to improve the estimators. As demonstrated in Appendix D, the precision to which the couplings are observed scales with $N^{-1/2}$, where N is the number of samples available.

IV. GAUGE SYMMETRIES

In this Section we consider one final extension of the learning formalism: the case of theories with gauge symmetries. From the point of view of a Euclidean path integral (or equivalently, a statistical field theory), a gauge symmetry is a local symmetry of the action, a family of automorphisms ϕ of field space such that $S = S \circ \phi$. Gauge symmetries occur in the lattice field theories of greatest interest for the standard model, and in particular in lattice QCD.

From one perspective, gauge symmetries introduce no complications whatsoever. A theory that happens to have gauge symmetry can be thought of as just a particular point in a large family of theories that generically lack gauge symmetry, and the formalism of Sec. II applies unmodified. However, in the context of gauge theories one ordinarily applies an additional constraint, which is that no object that is not gauge invariant is defined.³ It is therefore important that, both in the case of score matching and Schwinger-Dyson-based learning, only gauge-invariant expectation values are considered.⁴ The first subsection below shows how this is accomplished, and the second subsection applies this formalism, together with the results on

³When considering gauge theories in the Hamiltonian formalism, this is accomplished by declaring the Hilbert space of the theory to be those states that are not altered by gauge transformations. From this perspective there is no such thing as “gauge symmetry,” but instead a restriction of the space of states and operators.

⁴The restriction to gauge-invariant quantities is not just important for philosophical reasons, but because there are a variety of numerical methods that give access only to gauge-invariant quantities—and this is true of *all* experimental methods.

Grassmann-valued fields of the previous Section, to detail the learning problem of lattice QCD.

A. Generalities

Yang-Mills (in four spacetime dimensions) is the quintessential example of a gauge theory, and a natural starting point for studying QCD. In this Section we consider the Wilson discretization of $SU(N)$ Yang Mills [16].

This theory is defined on a hypercubic lattice in four dimensions. Define V to be the total number of lattice sites. To each edge in the lattice (of which there are $4V$) is associated a single $SU(N)$ degree of freedom. Thus a field configuration U consists of $4V$ $SU(N)$ matrices. We consider U to be a spacetime-indexed function of the lattice: $U_\mu(x)$ lives on the link between x and $x + \hat{\mu}$.

On this space of field configurations we define the Wilson action as follows:

$$S_{\text{YM}}(U) = \frac{2}{g^2} \sum_{x, \mu < \nu} \text{Re Tr}(I - P_{\mu\nu}(x)). \quad (35)$$

Here g is a coupling constant, and the only free parameter in the action. The sum is taken over all “plaquettes”—that is, over all loops of four distinct links. The notation $P_{\mu\nu}(x)$ refers to the product of links along that loop,

$$P_{\mu\nu}(x) = U_\nu(x)^\dagger U_\mu(x + \hat{\nu})^\dagger U_\nu(x + \hat{\mu}) U_\mu(x). \quad (36)$$

With this action written, it is a small matter to see the gauge symmetry. For each site x , select a matrix $V(x)$. The matrices U then transform according to

$$U_\mu(x) \mapsto V(x + \hat{\mu}) U_\mu(x) V^\dagger(x) \quad (37)$$

which is manifestly a symmetry of S_{YM} .

Although it is not directly visible from the Yang-Mills action, one can imagine that at each site x lives a vector space $V^{(x)} \approx \mathbb{C}^N$. One should think of $U_\mu(x)$ not as an automorphism of $\mathbb{C}^N$, but instead as a linear map $V^{(x)} \rightarrow V^{(x+\hat{\mu})}$. The gauge transformations at x correspond to automorphisms of $V^{(x)}$ (acting trivially on the spaces that live at all other sites), and this motivates the transformation law of the fields U given above. This view also makes it clear that if U is the link from $x \rightarrow y$, then $U^\dagger$ serves a symmetric role as the link from $y \rightarrow x$.

Expectation values are defined in the usual way, via a path integral over the space $SU(N)^{\otimes 4V}$ of field configurations

$$\langle \mathcal{O} \rangle = \frac{\int [\prod_{\mu, x} dU_\mu(x)] \mathcal{O}(U) e^{-S_{\text{YM}}[U]}}{\int [\prod_{\mu, x} dU_\mu(x)] e^{-S_{\text{YM}}[U]}}. \quad (38)$$

To set up the learning problem, we imagine that we are given access to expectation values over an empirical

distribution which is drawn from the family of actions defined by

$$S_\beta(U) = \beta \sum_{x, \mu < \nu} \text{Re Tr}(I - P_{\mu\nu}(x)). \quad (39)$$

By assumption there exists some β^* such that the empirical distribution is proportional to $e^{-S_{\beta^*}(U)}$. We wish to learn this value of β^* by inspection of the expectation values.

To illustrate the construction, we first consider the score matching approach, which provides intuition for gauge-invariant observables needed. We then show how the parameters may also be determined from Schwinger-Dyson relations. Let us begin by defining the score function. As discussed, we impose that this score function be gauge-invariant. A natural choice is,

$$\mathcal{L}(\beta) = \left\langle \left\| \frac{\partial}{\partial U_\mu^a(x)} \log p(U) - \frac{\partial}{\partial U_\mu^a(x)} \log \pi(U; \beta) \right\|^2 \right\rangle. \quad (40)$$

As before, the derivative operator is defined as a vector indexed over spacetime (x), direction of links (μ), and color (a). Here the partial derivative is defined as follows. For any function $f(U)$, we define

$$\frac{\partial}{\partial U^a} f(U) = \lim_{\epsilon \rightarrow 0} \frac{f(e^{i\epsilon T_a} U) - f(U)}{\epsilon}. \quad (41)$$

Here the T_a are a basis of generators of $SU(N)$. When $N = 2$ these would conventionally be the Pauli matrices; for $N = 3$ they are the eight Gell-Mann matrices. In any case we adopt the normalization given by the orthonormality relation $\text{Tr } T_i T_j = 2\delta_{ij}$.

To see that Eq. (40) is gauge invariant, we must consider how these derivatives behave under gauge transformations. Consider performing a gauge transformation by $V \in SU(N)$ at site x . Then the derivative with respect to $U(x)$ transforms according to

$$\frac{\partial}{\partial U^a(x)} P_{\mu\nu} \mapsto \rho_{\text{adj}}(V)_{ab} \frac{\partial}{\partial U^b(x)} P_{\mu\nu}, \quad (42)$$

where ρ_{adj} is the adjoint representation (written in the same basis as was chosen to define T_a) of the group $SU(N)$.

The L_2 norm used to define $\mathcal{L}$ is now gauge invariant due to the fact that a representation is a homomorphism and the group is unitary: $\rho(V)^\dagger \rho(V) = \rho(V^\dagger V) = 1$. By similar reasoning we find that the diagonal of the Hessian is also gauge invariant.

We now proceed to remove the explicit β^* dependence from $\mathcal{L}$ in the usual way, neglecting terms that are independent of β and integrating by parts. Up to a constant, the score-matching loss is now

$$\mathcal{L}(\beta) = \left\langle \left| \frac{\partial}{\partial U_\mu^a(x)} S \right|^2 \right\rangle - 2 \left\langle \frac{\partial^2}{\partial U_\mu^a(x)^2} S \right\rangle, \quad (43)$$

which is manifestly gauge invariant as discussed. The minimum of this score function satisfies

$$0 = \nabla \mathcal{L} = \beta \left\langle \left| \frac{\partial}{\partial U_\mu^a(x)} R \right|^2 \right\rangle - \left\langle \frac{\partial^2}{\partial U_\mu^a(x)^2} R \right\rangle \quad (44)$$

where we have defined

$$R = \sum_{x, \mu < \nu} \text{Re Tr}(I - P_{\mu\nu}(x)). \quad (45)$$

The second derivative term is given by the quadratic Casimir of the $SU(N)$ fundamental representation,

$$C_2(N)I = \sum_i T_i^2 = 2 \left(\frac{N^2 - 1}{N} \right) I. \quad (46)$$

The other term—the norm of the first derivative—is evaluated by using the Fierz identity

$$\sum_a T_a^{ij} T_a^{kl} = 2 \left(\delta^{il} \delta_{jk} - \frac{1}{N} \delta^{ij} \delta^{kl} \right). \quad (47)$$

The second-derivative term only affects one plaquette at a time, and becomes a sum over all plaquettes. The first-derivative term is more complicated, since the same $\frac{\partial}{\partial U_\mu^a(x)}$ derivative acts nontrivially on any plaquette which includes the link from x to $x + \hat{\mu}$. The result is a sum over the 36 (ordered) pairs of plaquettes which involve that link,

$$\left| \frac{\partial}{\partial U_\mu^a(x)} R \right|^2 = \sum_{P_1, P_2} \left[\text{Re Tr}(P_1 P_2) + \text{Re Tr}(P_1^\dagger P_2) - \frac{2}{N} \text{Re Tr}(P_1) \text{Re Tr}(P_2) \right]. \quad (48)$$

Here P_1 and P_2 are products of four links, beginning and ending at site x , corresponding to plaquettes which include the link between x and $x + \hat{\mu}$.

Putting this all together, and making use of the lattice symmetries of translation and rotation invariance, we find that β^* can be determined by two expectation values,

$$\beta^* \sum_{P_1, P_2} \left\langle \left[\text{Re Tr}(P_1 P_2) + \text{Re Tr}(P_1^\dagger P_2) - \frac{2}{N} \text{Re Tr}(P_1) \text{Re Tr}(P_2) \right] \right\rangle = 2 \frac{N^2 - 1}{N} \langle \text{Re Tr } P_{\mu\nu}(x) \rangle. \quad (49)$$

This has the form of a Schwinger-Dyson relation, but as before, this is not the only Schwinger-Dyson relation we could have used to determine the value of β . The most general Schwinger-Dyson relation on the Yang-Mills lattice reads

$$\left\langle \frac{\partial P}{\partial U^a(x)} \right\rangle = \left\langle P \frac{\partial S}{\partial U^a(x)} \right\rangle, \quad (50)$$

for some choice of site x and adjoint index a . Most such relations will involve, and therefore determine, β^* . However, this equation does not in general relate two gauge-invariant quantities. Suppose for example that P is the trace of some plaquette—then neither its derivative, nor its product with the derivative of the action, is gauge invariant.

To ensure gauge invariance, it is necessary that the function P be chosen to transform in the adjoint representation of the site at which the derivative is being taken, and then sum over all $N^2 - 1$ components. That is, fixing a lattice site x , the relation

$$\sum_a \left\langle \frac{\partial \text{Tr } T_a W(x)}{\partial U^a(x)} \right\rangle = \sum_a \left\langle \text{Tr } T_a W(x) \frac{\partial S}{\partial U^a(x)} \right\rangle \quad (51)$$

is manifestly gauge invariant for any loop $W(x)$ that begins and ends at site x . The simplest example is of course to take W to be a plaquette, as discussed above.

B. Lattice quantum chromodynamics

We conclude by discussing QCD on the lattice, which requires all the technology developed in this paper so far. For concreteness we consider a lattice theory of gauge fields using the Wilson action, and Wilson-discretized fermions (see [17,18] for standard references on lattice QCD). The action of this theory is a function both of the gauge fields and the (Grassmann-valued) fermions,

$$S_{\text{QCD}}(U, \bar{\psi}, \psi) = S_{\text{YM}}(U) + \bar{\psi}^{(f)}(x) D(U) \psi^{(f)}(y), \quad (52)$$

where $D(U)$ is the lattice Dirac operator. The quark fields $\psi^{(f)}(x)$, for each flavor f and at each site x , are $4N$ component objects, implicitly indexed by color and Dirac index. For Wilson fermions we take the Dirac matrix D to be

$$D_W(U)_{xaa}^{y\beta b} = (m_f + 4r)\delta_{xy}\delta_{\alpha\beta}\delta_{ab} - \frac{1}{2}\sum_{\mu}\left[[r - \gamma_{\mu}]_{\alpha\beta}\delta_{y,x+\hat{\mu}}[U_{\mu}(x)]_{ab} + [r + \gamma_{\mu}]_{\alpha\beta}\delta_{y,x-\hat{\mu}}[U_{\mu}^{\dagger}(x - \hat{\mu})]_{ab}\right], \quad (53)$$

where r is the Wilson parameter (typically $r = 1$), removing fermion doublers at the cost of breaking the chiral symmetry explicitly. Latin indices (a, b) refer to color, and Greek (α, β) to spin components. The parameters of the action are β in S_{YM} and masses m_f of quarks, which we collectively denote as θ in the remaining discussion.

As in the case of Yang Mills, it is straightforward to construct a manifestly gauge-invariant score function via the L_2 norm,

$$\mathcal{L}(\theta) = \sum_f \left\langle \left\| \frac{\partial}{\partial \psi^{(f)}} [S_{QCD} - S_{QCD}^{(\theta)}] \right\|^2 \right\rangle + \left\langle \left\| \frac{\partial}{\partial U_{\mu}^a(x)} [S_{QCD} - S_{QCD}^{(\theta)}] \right\|^2 \right\rangle. \quad (54)$$

Here the norm, defined in Eq. (24), is taken in the N -dimension color space, spacetime, and Dirac space for the first term. The norm in the second term follows the definition in Eq. (40). Gauge invariance of the score function follows from the transformation rules of the derivatives of the action. For typical quark bilinear terms in the action,

$$B(x, \mu) = \bar{\psi}(x)U_{\mu}(x)\psi(x + \mu), \quad (55)$$

their derivatives with respect to quark and gauge fields transform in the (anti)fundamental and adjoint representation, respectively,

$$\frac{\partial}{\partial U_{\mu}^a(x)} B(x, \mu) \mapsto \rho_{\text{adj}}(V(x))_{ab} \frac{\partial}{\partial U_{\mu}^b(x)} B(x, \mu), \quad (56)$$

$$\frac{\partial}{\partial \bar{\psi}_a(x)} B(x, \mu) \mapsto \rho_{\text{fund}}(V(x))_{ab} \frac{\partial}{\partial \bar{\psi}_b(x)} B(x, \mu). \quad (57)$$

Note that we have dropped irrelevant Dirac and flavor indices on quark fields. The derivative of the action with the quark fields ψ transforms in the fundamental representation, making the first term in Eq. (54) gauge invariant. Combining the transformation rule of the plaquettes, Eq. (42), and of quark terms $Q(x, \hat{\mu})$ above, the derivatives of the action with gauge fields transform under the adjoint representation, thus the second term in the score function is also gauge invariant.

After dropping constant terms from the empirical distribution, the score function reads

$$\mathcal{L}(\theta) = \sum_f \left\langle \left| \frac{\partial}{\partial \psi^{(f)}} S_{QCD}^{(\theta)} \right|^2 + 2 \frac{\partial}{\partial \psi^{(f)}} \frac{\partial}{\partial \bar{\psi}^{(f)}} S_{QCD}^{(\theta)} \right\rangle + \left\langle \left| \frac{\partial}{\partial U_{\mu}^a(x)} S_{QCD}^{(\theta)} \right|^2 - 2 \frac{\partial^2}{\partial U_{\mu}^a(x)^2} S_{QCD}^{(\theta)} \right\rangle. \quad (58)$$

For completeness we briefly discuss the construction of gauge-invariant Schwinger-Dyson equations in lattice QCD. Schwinger-Dyson equations for lattice QCD can be constructed via the derivatives of functions of $U, \bar{\psi}, \psi$ with respect to these fields. When differentiating with respect to the gauge fields we again obtain relations of the form of Eq. (51); when differentiating with respect to quark fields we obtain,

$$\left\langle \sum_a \frac{\partial P_a}{\partial \bar{\psi}_{\alpha,a}^{(f)}(x)} \right\rangle = - \left\langle \sum_a P_a \frac{\partial S}{\partial \bar{\psi}_{\alpha,a}^{(f)}(x)} \right\rangle. \quad (59)$$

To have the both sides of the equation be manifestly gauge invariant, we want to choose as P_a a function that transforms in the antifundamental representation. One simple choice is $P_{\alpha',a} = \bar{\psi}_{\alpha',a}^{(f)}(x')$, for which the Schwinger-Dyson equation reads

$$N\delta_{\alpha\alpha'}\delta_{xx'} = - \sum_{y,\beta,a,b} \langle \bar{\psi}_{\alpha',a}^{(f)}(x) D_W(U)_{xaa}^{y\beta b} \psi_{\beta,b}^{(f)}(y) \rangle. \quad (60)$$

This Schwinger-Dyson relation allows the quark mass m_f to be determined; the gauge coupling β is not involved and cannot be determined from this equation. A Schwinger-Dyson equation that fixes β can be derived via Eq. (51).

V. DISCUSSION

We have presented a broad family of learning methods based on the lattice Schwinger-Dyson relations. Because Schwinger-Dyson relations are available for Grassmann-valued fields, our method allows learning of couplings in theories with fermions. Similarly, as (a certain subset of) lattice Schwinger-Dyson equations are gauge invariant, we obtain a learning algorithm for lattice gauge theories (including QCD) that operates exclusively on gauge-invariant data. Score matching, a venerable approach to learning exponential families of distributions, is a specific case of a broad family of learning methods based on Schwinger-Dyson relations. However, by expanding the family of learning methods considerably beyond score matching, we obtain a way to learn the action parameters with potentially lower cost, as there is more choice in the expectation values that must be measured. A concrete case

where the Schwinger-Dyson-based learning procedure is advantageous, in the sense of providing an estimator of lower variance than that of score matching, is presented in Appendix A.

Throughout this work, we have neglected the inevitable statistical errors in estimates of expectation values, and we have neglected the computational costs associated to determining those expectation values. The existence of these errors and costs is one of the motivations for our generalization to arbitrary Schwinger-Dyson relations. As shown in the Appendix, in general the Schwinger-Dyson relations represented in the score function are not the ones that yield the smallest statistical errors. We leave to future work a proper treatment of these issues, which are critical to applying these methods on larger lattices.

This work is motivated in part by the application suggested in [1,2], of studying renormalization flows on the lattice. We anticipate that the technology developed here will allow that renormalization program to be applied efficiently to (among others) theories of interacting fermions.

In this work we have assumed that the data used from learning comes from a properly mixed Monte Carlo calculation. However, prior work has shown that samples from a metastable phase—where the whole of field space has not been properly sampled—are still sufficient to accurately learn the action [19]. We anticipate that the same phenomenon applies the method described in this paper. For example, in ϕ^4 theory at strong coupling, the $\mathbb{Z}_2$ symmetry $\phi \leftrightarrow -\phi$ is spontaneously broken, and many Monte Carlo algorithms will sample only one of the two sectors. Provided such samples, the learning method outlined here will still produce the correct coefficients in the algorithm (at sufficiently large volume). This may be understood by noting that, due to the fact that the probability weight e^{-S} is small in the region between phases, the Schwinger-Dyson equations hold to good accuracy even when the path integral is taken only over one of the phases.

This paper has been entirely limited to theories defined via Euclidean lattice path integrals (whether bosonic or Grassmannian). Schwinger-Dyson equations are also available for, and can in principle be used to learn from, correlation functions computed from a real-time path integral. Whether such a method can be of use is unclear. Real-time lattice correlators are famously difficult to extract due to the sign problem [20], even once a sensible (i.e., unitary) action has been defined [21]. Similarly, we have not addressed learning from Hamiltonian data, although prior work has studied this in both the ground state [22] and at finite temperature [23].

A. Remark on integration contours

One should not make the mistake of thinking that Schwinger-Dyson relations determine all facts about a (lattice) quantum field theory. Crucial nonperturbative physics, such as instantons, are not constrained by these

equations. A simple example is accessible analytically: consider a one-site lattice model, with expectation values given by

$$\langle f \rangle := \frac{\int_{\gamma} dz f(z) e^{-z^4}}{\int_{\gamma} dz e^{-z^4}}. \quad (61)$$

Here the integrals are to be interpreted as complex contour integrals, over some contour $\gamma \subset \mathbb{C}$. There are multiple contours γ that might be chosen. For instance one can take $\gamma = \mathbb{R}$, giving the “obvious” integral along the real line. The expectation values are also convergent, for polynomial f , when we select $\gamma = i\mathbb{R}$ —here the integral is performed along the imaginary axis. The resulting expectation values are not equal; however, both sets of expectation values satisfy the Schwinger-Dyson relations.

One can see the same phenomenon by considering the Schwinger-Dyson relations in isolation, without reference to the notion of an integration contour. A complete set of Schwinger-Dyson relations for this simple model reads, for positive integers n : $\langle z^{n-1} \rangle = 4\langle z^{n+3} \rangle$. These are recursion relations, and there exists a solution for any choice of $\langle z^2 \rangle$. This choice may be seen to be degenerate with the choice of homology class (albeit over $\mathbb{R}$, instead of $\mathbb{Z}$ as usual).

Knowledge of a sufficient number of expectation values is sufficient in principle to fix the choice of contour—or equivalently, to select a single solution from vector space of solutions to the Schwinger-Dyson relations. However the methods in this paper do not provide a way to extract this information efficiently.

This question regarding what contour should be used is relevant to physical theories. The practical effect of choosing the wrong integration contour is either to neglect nonperturbative “instantonlike” contributions [24], or to render the integral entirely nonconvergent (as in the Euclidean gravitational path integral [25], where the question of the correct choice of integration contour is indeed central [26–28]). In quantum field theory, the contour ambiguity may in principle be fixed by demanding unitarity (in Euclidean quantum field theory, reflection positivity); in general it is unclear how one should select the “physical” integration contour.

ACKNOWLEDGMENTS

S. S., A. J., and A. Y. L. acknowledge support from the U.S. Department of Energy/Office of Science Advanced Scientific Computing Research Program. S. L. is supported by a Richard P. Feynman fellowship from the Laboratory Directed Research and Development (LDRD) program of Los Alamos National Laboratory (LANL). Y. Y. is supported by a Darleane C. Hoffman fellowship from the LANL LDRD program. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy, under Contract No. 89233218CNA000001.

S. S. contributed equally to this work. Y. Y., A. Y. L., S. L., and A. J. were listed in reverse alphabetical order.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available upon publication because it is not technically feasible and/or the cost of preparing, depositing, and hosting the data would be prohibitive within the terms of this research project. The data are available from the authors upon reasonable request.

APPENDIX A: ESTIMATOR VARIANCES IN ONE DIMENSION

The purpose of this appendix is to demonstrate that the coupling estimators yielded by score matching are not, in general, the most efficient estimators available, even within the family of estimators constructed from Schwinger-Dyson equations.

There are multiple meanings that can be assigned to the phrase “most efficient,” and which one is relevant depends on context. For example in lattice QCD, a premium is placed on minimizing the number of inversions of the Dirac matrix that are required. Here we show that Schwinger-Dyson learning can yield estimators of smaller variance (and therefore, requiring fewer samples for a fixed precision) than score matching.

Consider the exponential family of distributions specified by the action

$$S_\lambda(x) = \frac{1}{2}x^2 + \frac{\lambda}{4}x^4. \quad (\text{A1})$$

The relevant learning problem is to estimate λ^* given samples from S_{λ^*} , for some fixed λ^* . The variance of any estimator we provide will depend on the precise value of λ^* . To make calculations explicit, we select $\lambda^* = 0$. The first two nontrivial Schwinger-Dyson equations read

$$1 = \langle x^2 \rangle + \lambda \langle x^4 \rangle, \text{ and} \quad (\text{A2})$$

$$3\langle x^2 \rangle = \langle x^4 \rangle + \lambda \langle x^6 \rangle. \quad (\text{A3})$$

The second of these relations is the one that yields the score-matching estimator,

$$\hat{\lambda}_{\text{SM}} = \frac{3\langle x^2 \rangle - \langle x^4 \rangle}{\langle x^6 \rangle}. \quad (\text{A4})$$

However, by taking linear combinations, these two Schwinger-Dyson relations together define a one-parameter family of estimators of λ^* . Parametrizing by a real number u , these estimators are

$$\hat{\lambda}_u = \frac{u + (3 - u)\langle x^2 \rangle - \langle x^4 \rangle}{u\langle x^4 \rangle + \langle x^6 \rangle}. \quad (\text{A5})$$

This parametrization has been chosen so that $u = 0$ yields the score-matching estimator: $\hat{\lambda}_0 = \hat{\lambda}_{\text{SM}}$.

We will now compute the standard deviation $\sigma(u)$ of the estimator $\hat{\lambda}_u$. Because $\lambda^* = 0$, all expectation values involved may be computed immediately by Wick’s theorem. The covariance of the random variables x^2 and x^4 is given (in the basis $\{x^2, x^4\}$) by

$$\Sigma = \begin{pmatrix} 2 & 12 \\ 12 & 96 \end{pmatrix}. \quad (\text{A6})$$

The corresponding gradient of the estimator function $\hat{\lambda}_u(\langle x^2 \rangle, \langle x^4 \rangle)$ is

$$\nabla \hat{\lambda}_u = \begin{pmatrix} \frac{\partial \hat{\lambda}_u}{\partial \langle x^2 \rangle} \\ \frac{\partial \hat{\lambda}_u}{\partial \langle x^4 \rangle} \end{pmatrix} = \frac{1}{3u + 15} \begin{pmatrix} 3 - u \\ -1 \end{pmatrix}. \quad (\text{A7})$$

Finally we obtain the variance of the estimator $\hat{\lambda}_u$, for the case $\lambda^* = 0$, in the limit of a large number of samples N ,

$$N^{1/2} \sigma(u)^2 \approx (\nabla \hat{\lambda}_u)^T \Sigma (\nabla \hat{\lambda}_u) = \frac{2u^2 + 12u + 42}{(3u + 15)^2}. \quad (\text{A8})$$

It is apparent that $u = 0$ is not the minimum, and therefore that the information in Eq. (A2) is useful in reducing the variance of the learning process below what is obtained by score matching alone.

Of course there is no reason to limit oneself to using a *single* Schwinger-Dyson relation. As shown in the body of this paper an arbitrary number of Schwinger-Dyson relations can be used to construct a loss function, and the resulting estimator will have still-lower variance than may be obtained simply by optimizing u .

APPENDIX B: GRASSMANN VARIABLES

Let η_i for $i = 1, \dots, n$ be Grassmann variables. Any two Grassmann variables anticommute: $\eta_i \eta_j = -\eta_j \eta_i$. It follows that Grassmanns are nilpotent: $\eta_i^2 = 0$. There are 2^n distinct monomials of n Grassmann variables.

A function of Grassmann variables is defined as an arbitrary $\mathbb{C}$ -linear combination of Grassmann monomials. Thus the space of functions on n Grassmann variables is a vector space of dimension 2^n . For a function $f(\eta)$, we may define the exponential via Taylor expansion,

$$e^{f(\eta)} = \sum_k \frac{1}{k!} f(\eta)^k. \quad (\text{B1})$$

There are a finite number of terms in this expression: For $k > 2^n$, all $f(\eta)^k$ necessarily vanish.

Differentiation and integration act in precisely the same way on Grassmann variables, and are defined by,

$$\frac{d}{d\eta_i}\eta_j = \int d\eta_i \eta_j = \delta_{ij}. \quad (\text{B2})$$

Their action on more complicated expressions is defined by linearity. Note that the infinitesimals also anticommute, so that (for example) $\int d\eta_1 d\eta_2 \eta_1 \eta_2 = -1$.

APPENDIX C: SCHWINGER-DYSON RELATIONS FOR MASSIVE THIRRING MODEL

Consider the action for the massive Thirring model,

$$S_{\text{Th}} = \sum_{x,x'} \bar{\chi}_x (M_{xx'} + m\delta_{xx'}) \chi_{x'} + \frac{g^2}{2} \sum_{x,\nu} j_\nu^2(x), \quad (\text{C1})$$

where the massless staggered Dirac operator is defined in Eq. (28) and the current $j_\nu(x)$ is defined in Eq. (29). The square of the current expands to give a four-Fermi interaction: $j_\nu^2(x) = -\frac{1}{2} \bar{\chi}_{x+\nu} \chi_{x+\nu} \bar{\chi}_x \chi_x$. We will define Schwinger-Dyson relations for operators $Q_{ab}^{(1)} = \bar{\chi}_a \chi_b$ and $Q_{ap}^{(2)} = \bar{\chi}_{a+\rho} \chi_{a+\rho} \bar{\chi}_a \chi_a$. For $Q \in \{Q^{(1)}, Q^{(2)}\}$, the relations are

$$\langle \partial_{\chi_k} \partial_{\bar{\chi}_k} Q \rangle = -\langle (\partial_{\bar{\chi}_k} Q) \cdot (\partial_{\chi_k} S) \rangle. \quad (\text{C2})$$

Define $X_a \equiv \bar{\chi}_a \chi_a$, $X_{ab} \equiv \bar{\chi}_a \chi_b$, $K_{ab,c} = X_{ab} X_c$, $K_{a,bc} = X_a X_{bc}$ and $\Gamma_{abc} \equiv X_a X_b X_c$. The Schwinger-Dyson relations read,

$$\sum_i (m\delta_{ia} + M_{ia}) \langle X_{ib} \rangle - \frac{g^2}{4} \sum_\nu (\langle K_{ab,a+\nu} + K_{ab,a-\nu} \rangle) = \delta_{ab}, \quad (\text{C3})$$

$$\sum_i (m\delta_{ia} + M_{ia}) \langle K_{ia,a+\rho} \rangle + (m\delta_{i,a+\rho} + M_{i,a+\rho}) \langle K_{a,ia+\rho} \rangle - \frac{g^2}{4} \sum_\nu (\langle \Gamma_{a,a+\rho,a\pm\nu} \rangle + \langle \Gamma_{a,a+\rho,a+\rho\pm\nu} \rangle) = \langle X_{a+\rho} \rangle + \langle X_a \rangle. \quad (\text{C4})$$

For a fixed auxiliary field configuration A , the fermionic expectation values follow from Wick's theorem,

$$\langle X_{ab} \rangle \equiv \langle \bar{\chi}_a \chi_b \rangle = D[A]_{ba}^{-1} \quad \langle X_a X_b \rangle = D[A]_{aa}^{-1} D[A]_{bb}^{-1} - D[A]_{ab}^{-1} D[A]_{ba}^{-1}, \quad (\text{C5})$$

where the propagator $D[A, g]$ is defined in Eq. (32). The six-point functions $\langle \Gamma_{abc} \rangle$ are computed analogously. Full expectation values are then obtained by averaging over the ensemble of configurations. Substituting these into Eqs. (C3) and (C4) yields a linear system in the parameters m and g^2 , which can be solved to recover the bare couplings.

APPENDIX D: SCALING

This appendix is a brief supplement to the numerical results of Sec. III B, to demonstrate the scaling of the errors in our method as the number of samples is increased.

We run the same experiment, but on a smaller lattice of only 6×6 sites, and with up to 12800 samples. The results are shown in Fig. 3, where the absolute error in the estimator is plotted against the number of samples consumed. As noted in the main text, the estimator is biased, so that the (purely statistical) error estimates

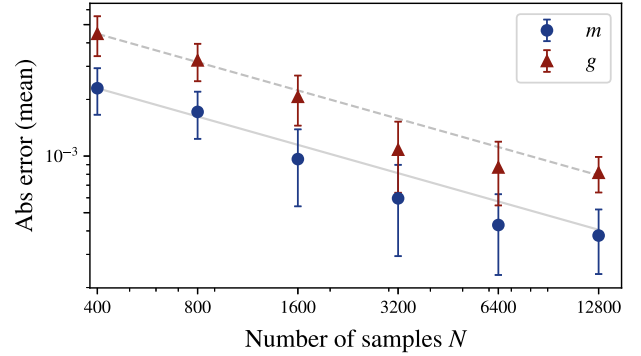


FIG. 3. Errors in parameters m and g with error bars on a 6×6 lattice. The errors scale approximately as $1/\sqrt{N}$, where N is the number of samples.

exclude the true value. The bias, as well as the statistical errors, both fall as $N^{-1/2}$ as the number of samples is increased.

- [1] P. E. Shanahan, A. Trewartha, and W. Detmold, Machine learning action parameters in lattice quantum chromodynamics, *Phys. Rev. D* **97**, 094506 (2018).
- [2] S. Shukla, A. Jayakumar, and A. Y. Lokhov, Learning of statistical field theories, [arXiv:2511.09859](#).
- [3] M. Moshe and J. Zinn-Justin, Quantum field theory in the large N limit: A review, *Phys. Rep.* **385**, 69 (2003).
- [4] R. H. Swendsen, Monte Carlo renormalization group, *Phys. Rev. Lett.* **42**, 859 (1979).
- [5] N. Dupuis, L. Canet, A. Eichhorn, W. Metzner, J. M. Pawłowski, M. Tissier, and N. Wschebor, The nonperturbative functional renormalization group and its applications, *Phys. Rep.* **910**, 1 (2021).
- [6] A. Y. Lokhov, M. Vuffray, S. Misra, and M. Chertkov, Optimal structure and parameter learning of ising models, *Sci. Adv.* **4**, e1700791 (2018).
- [7] A. Hyvärinen, Estimation of non-normalized statistical models by score matching, *J. Mach. Learn. Res.* **6**, 695 (2005).
- [8] S. Lawrence, Schwinger-Dyson control variates for lattice fermions, [arXiv:2404.10707](#).
- [9] T. Bhattacharya, S. Lawrence, and J.-S. Yoo, Control variates for lattice field theory, *Phys. Rev. D* **109**, L031505 (2024).
- [10] P. F. Bedaque and H. Oh, Leveraging neural control variates for enhanced precision in lattice field theory, *Phys. Rev. D* **109**, 094519 (2024).
- [11] F. J. Dyson, The s matrix in quantum electrodynamics, *Phys. Rev.* **75**, 1736 (1949).
- [12] J. Schwinger, On the Green's functions of quantized fields. I, *Proc. Natl. Acad. Sci. U.S.A.* **37**, 452 (1951).
- [13] J. Schwinger, On the Green's functions of quantized fields. II, *Proc. Natl. Acad. Sci. U.S.A.* **37**, 455 (1951).
- [14] M. E. Peskin and D. V. Schroeder, *An Introduction to Quantum Field Theory* (Addison-Wesley, Reading, USA, 1995).
- [15] J. B. Kogut and L. Susskind, Hamiltonian formulation of Wilson's lattice gauge theories, *Phys. Rev. D* **11**, 395 (1975).
- [16] K. G. Wilson, Confinement of quarks, *Phys. Rev. D* **10**, 2445 (1974).
- [17] C. Gattringer and C. B. Lang, *Quantum Chromodynamics on the Lattice: An Introductory Presentation*, Lecture Notes in Physics Vol. 788 (Springer, Berlin, Heidelberg, 2010).
- [18] I. Montvay and G. Münster, *Quantum Fields on a Lattice*, Cambridge Monographs on Mathematical Physics (Cambridge University Press, Cambridge, England, 1994).
- [19] A. Jayakumar, A. Y. Lokhov, S. Misra, and M. Vuffray, Discrete distributions are learnable from metastable samples, [arXiv:2410.13800](#).
- [20] S. Lawrence and Y. Yamauchi, Normalizing flows and the real-time sign problem, *Phys. Rev. D* **103**, 114509 (2021).
- [21] G. Kanwar and M. L. Wagman, Real-time lattice gauge theory actions: Unitarity, convergence, and path integral contour deformations, *Phys. Rev. D* **104**, 014513 (2021).
- [22] X.-L. Qi and D. Ranard, Determining a local Hamiltonian from a single eigenstate, *Quantum* **3**, 159 (2019).
- [23] A. Bakshi, A. Liu, A. Moitra, and E. Tang, Learning quantum Hamiltonians at any temperature in polynomial time, in *56th Annual ACM Symposium on Theory of Computing* (2023), [arXiv:2310.02243](#).
- [24] S. Lawrence, R. Weller, C. Peterson, and P. Romatschke, Instantons, analytic continuation, and PT-symmetric field theory, *Phys. Rev. D* **108**, 085013 (2023).
- [25] G. W. Gibbons, S. W. Hawking, and M. J. Perry, Path integrals and the indefiniteness of the gravitational action, *Nucl. Phys.* **B138**, 141 (1978).
- [26] M. Kontsevich and G. Segal, Wick rotation and the positivity of energy in quantum field theory, *Q. J. Math. Oxford Ser.* **72**, 673 (2021).
- [27] E. Witten, A note on complex spacetime metrics, [arXiv:2111.06514](#).
- [28] B. Banihashemi and T. Jacobson, On the lapse contour in the gravitational path integral, *Phys. Rev. D* **111**, 066014 (2025).