

Fluctuation-dissipation-type theorem in stochastic linear learning


Manhyung Han,^{1,*} Jeonghyeok Park^{2,*}, Taewoong Lee,^{3,§} and Jung Hoon Han^{4,||}

¹*Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, Korea*

²*Jesus College, Cambridge University, Jesus Lane, Cambridge CB5 8BL, United Kingdom*

³*Harvard College, Harvard University, Cambridge, Massachusetts 02138, USA*

⁴*Department of Physics, Sungkyunkwan University, Suwon 16419, Korea*

 (Received 8 June 2021; revised 5 August 2021; accepted 9 September 2021; published 20 September 2021)

The fluctuation-dissipation theorem (FDT) is a simple yet powerful consequence of the first-order differential equation governing the dynamics of systems subject simultaneously to dissipative and stochastic forces. The linear learning dynamics, in which the input vector maps to the output vector by a linear matrix whose elements are the subject of learning, has a stochastic version closely mimicking the Langevin dynamics when a full-batch gradient descent scheme is replaced by that of a stochastic gradient descent. We derive a generalized FDT for the stochastic linear learning dynamics and verify its validity among the well-known machine learning data sets such as MNIST, CIFAR-10, and EMNIST.

DOI: [10.1103/PhysRevE.104.034126](https://doi.org/10.1103/PhysRevE.104.034126)

I. INTRODUCTION

It is not an uncommon perception among the practitioners of machine learning and of theoretical many-body physics that some ideas of physics, most notably those of equilibrium and nonequilibrium statistical physics, might have significance in the fundamental understanding of the machine learning dynamics. Such a sentiment and progress along the direction has continued for some time and is still in active pursuit, mostly by the researchers in the machine learning community [1–9]. The belief in the statistical-physics foundation of the machine learning will be strengthened obviously by more examples of ideas originating from statistical physics and then manifesting themselves in the machine learning. Here, we establish one such connection, relating a fundamental theorem in near-equilibrium statistical physics [10–14] to the theory of learning dynamics [1–3,6,7,9], in particular where the learning process is *linear* and described by a stochastic equation similar to what governs the Ornstein-Uhlenbeck processes [15,16]. The theorem in question is the fluctuation-dissipation theorem (FDT).

The FDT in a strict sense refers to specific relations that hold between correlation functions and response functions of physical systems under equilibrium [16]. Here, we use the term in a more relaxed sense, referring to mathematical identities among the observable quantities under the stationary state condition. The difference between the equilibrium and the stationary state is revealed by the existence of an anti-symmetric matrix \mathbf{Q} [10–14], which will be defined shortly.

The FDT is illustrated most simply in the Langevin dynamics of a single-particle subject simultaneously to dissipative and stochastic forces

$$\dot{x} = -\gamma x + f(t), \quad (1.1)$$

where, in the context of Newtonian motion, x represents the velocity of a particle in one dimension, $-\gamma x$ is the resistive force, and $f(t)$ is the random force coming from the environment. On integrating the first-order differential equation we obtain the formally exact solution $x(t) = e^{-\gamma t} [x(0) + \int_0^t dt' e^{\gamma t'} f(t')]$ which, in the long-time limit ($t \rightarrow \infty$), yields the average

$$\langle x^2 \rangle = 2D e^{-2\gamma t} \int_0^t dt' e^{2\gamma t'} = D/\gamma, \quad (1.2)$$

assuming the white-noise correlation $\langle f(t)f(t') \rangle = 2D\delta(t-t')$. The competing tendencies of the dissipation (γ) and fluctuation (D) find balance through the identity.

Multidimensional generalization of the Langevin dynamics finds expression in

$$\dot{\mathbf{x}} = -\mathbf{\Gamma}\mathbf{x} + \mathbf{f}(t), \quad (1.3)$$

with n -dimensional variables $\mathbf{x} = (x_1, \dots, x_n)$, the $n \times n$ dissipation matrix $\mathbf{\Gamma}$, and the n -dimensional stochastic force vector \mathbf{f} obeying the zero mean $\langle \mathbf{f} \rangle = \mathbf{0}$ and the variance $\langle \mathbf{f}(t)\mathbf{f}^T(t') \rangle = 2\mathbf{D}\delta(t-t')$, in terms of the $n \times n$ diffusion matrix \mathbf{D} . From the exact solution $\mathbf{x}(t) = e^{-\mathbf{\Gamma}t} [\mathbf{x}(0) + \int_0^t e^{\mathbf{\Gamma}t'} \mathbf{f}(t') dt']$ we derive the long-time correlation average

$$\begin{aligned} \mathbf{\Sigma}(t) &= \langle \mathbf{x}(t)\mathbf{x}^T(t) \rangle \\ &= 2 \int_0^t dt' e^{\mathbf{\Gamma}(t-t')} \mathbf{D} e^{\mathbf{\Gamma}^T(t-t')}, \end{aligned} \quad (1.4)$$

and the following identity for $\mathbf{\Sigma} = \mathbf{\Sigma}(t \rightarrow \infty)$:

$$\mathbf{\Gamma}\mathbf{\Sigma} + \mathbf{\Sigma}\mathbf{\Gamma}^T = 2\mathbf{D}. \quad (1.5)$$

*These authors contributed equally to this work.

†hanmanhyung@gmail.com

‡jp868@cam.ac.uk

§taewoonglee@college.harvard.edu

||hanjemme@gmail.com

This identity relates the diffusion matrix \mathbf{D} with the dissipation matrix $\mathbf{\Gamma}$ through the correlation matrix $\mathbf{\Sigma}$ in the stationary state, for the Ornstein-Uhlenbeck processes with constant $\mathbf{\Gamma}$ and \mathbf{D} [10,11]. Extensions and applications of the theorem both in physical systems and machine learning have since appeared [2,3,14]. Owing to the identity, one can write the matrix $\mathbf{\Gamma\Sigma}$ as the sum of the symmetric (\mathbf{D}) and antisymmetric (\mathbf{Q}) matrix:

$$\mathbf{\Gamma\Sigma} = \mathbf{D} + \mathbf{Q}. \quad (1.6)$$

It was pointed out in Ref. [13] that $\mathbf{Q} = \mathbf{0}$ implies the detailed balance, otherwise one should allow the possibility $\mathbf{Q} \neq \mathbf{0}$ in the decomposition, Eq. (1.6).

In Sec. II, we derive an analogous mathematical identity for the stochastic linear learning dynamics. This is then verified, in Sec. III, through numerical experiments on several well-known machine learning data sets. Implications of our work are discussed in Sec. IV.

II. FDT IN LEARNING DYNAMICS

In the learning dynamics one is confronted with a collection of input vectors \mathbf{x}_α (e.g., pixels in a jpg file reformatted as a one-dimensional vector) and output vectors \mathbf{y}_α (e.g., classification of the picture as an image of a cat or a dog), where $1 \leq \alpha \leq N$ runs over the entire data set called the *batch*. In the linear learning dynamics one is interested in finding the matrix \mathbf{W} that minimizes the error

$$\begin{aligned} E &= \frac{1}{2N} \sum_{\alpha=1}^N (\mathbf{y}_\alpha - \mathbf{W}\mathbf{x}_\alpha)^T (\mathbf{y}_\alpha - \mathbf{W}\mathbf{x}_\alpha) \\ &\rightarrow \frac{1}{2} \text{Tr}[\mathbf{\Sigma}_{xx} \mathbf{W}^T \mathbf{W} - \mathbf{W}^T \mathbf{\Sigma}_{yx} - \mathbf{\Sigma}_{yx}^T \mathbf{W}]. \end{aligned} \quad (2.1)$$

There is a constant term which is dropped in going to the second line. The two correlation functions appearing in the second line are

$$\mathbf{\Sigma}_{xx} = \frac{1}{N} \sum_{\alpha=1}^N \mathbf{x}_\alpha \mathbf{x}_\alpha^T, \quad \mathbf{\Sigma}_{yx} = \frac{1}{N} \sum_{\alpha=1}^N \mathbf{y}_\alpha \mathbf{x}_\alpha^T. \quad (2.2)$$

The gradient descent (GD) method of finding the optimal \mathbf{W} results in the first-order differential equation for \mathbf{W} [6,7]:

$$\frac{d\mathbf{W}}{dt} = -\frac{\delta E}{\delta \mathbf{W}} = -\mathbf{W}\mathbf{\Sigma}_{xx} + \mathbf{\Sigma}_{yx}. \quad (2.3)$$

The full solution is given by $\mathbf{W}(t) = \mathbf{W}(0)e^{-\mathbf{\Sigma}_{xx}t} + \mathbf{W}_0(1 - e^{-\mathbf{\Sigma}_{xx}t})$, where $\mathbf{W}_0 = \mathbf{\Sigma}_{yx}\mathbf{\Sigma}_{xx}^{-1}$ offers the equilibrium solution.

An interesting connection to the Langevin dynamics and FDT arises when we treat $\mathbf{\Sigma}_{xx}$ and $\mathbf{\Sigma}_{yx}$ in the dynamics of Eq. (2.3) as a minibatch (not a full-batch) average. At each stage of \mathbf{W} evolution one picks a different, randomly chosen minibatch to compute the average $\mathbf{\Sigma}_{xx}(t) = N_m^{-1} \sum_{\alpha \in B(t)} \mathbf{x}_\alpha \mathbf{x}_\alpha^T$ and $\mathbf{\Sigma}_{yx}(t) = N_m^{-1} \sum_{\alpha \in B(t)} \mathbf{y}_\alpha \mathbf{x}_\alpha^T$, where N_m is the minibatch size and $B(t)$ is the particular minibatch chosen at the time t . The \mathbf{W} dynamics according to the stochastic gradient descent (SGD) scheme becomes

$$\frac{d\mathbf{W}}{dt} = -\mathbf{W}\mathbf{\Sigma}_{xx}(t) + \mathbf{\Sigma}_{yx}(t). \quad (2.4)$$

Phrased in the language of Langevin dynamics, both the dissipative [$\mathbf{\Sigma}_{xx}(t)$] and the stochastic [$\mathbf{\Sigma}_{yx}(t)$] forces are time dependent, whereas the conventional Langevin dynamics has constant dissipative force matrix and the (time-dependent) stochastic force.

The analysis of the SGD equation above is facilitated by making the change of variables as the sum of the stationary, time-independent piece and the time-dependent, fluctuating piece. Here, time t refers to the artificial time in the evolution of the learning matrix $\mathbf{W}(t)$. First, the learning matrix itself is separated as the sum of two pieces,

$$\mathbf{W}(t) = \mathbf{W}_0 + \overline{\mathbf{W}}(t), \quad (2.5)$$

with the overline representing the fluctuating part of $\mathbf{W}(t)$ and $\mathbf{W}_0 = \mathbf{\Sigma}_{yx}\mathbf{\Sigma}_{xx}^{-1}$ is the stationary solution. A similar decomposition takes place for other variables in the equation,

$$\begin{aligned} \mathbf{\Sigma}_{xx}(t) &= \mathbf{\Sigma}_{xx} + \overline{\mathbf{\Sigma}}_{xx}(t), \\ \mathbf{\Sigma}_{yx}(t) &= \mathbf{\Sigma}_{yx} + \overline{\mathbf{\Sigma}}_{yx}(t). \end{aligned} \quad (2.6)$$

By definition, the average of the fluctuation parts of the correlation functions has zero mean $\langle \overline{\mathbf{\Sigma}}_{xx} \rangle = 0 = \langle \overline{\mathbf{\Sigma}}_{yx} \rangle$ when computed after reaching convergence to equilibrium, that is, after a sufficiently long time. Instead of Eq. (2.4), we get to work with the equivalent equation

$$\frac{d\overline{\mathbf{W}}(t)}{dt} = -\overline{\mathbf{W}}(t)[\mathbf{\Sigma}_{xx} + \overline{\mathbf{\Sigma}}_{xx}(t)] + \overline{\mathbf{\Sigma}}_{yx}(t) - \mathbf{W}_0\overline{\mathbf{\Sigma}}_{xx}(t). \quad (2.7)$$

This equation describes the convergence of the learning matrix $\overline{\mathbf{W}}(t)$ in the SGD scheme.

In fact, it is possible to write down the exact solution to the stochastic Eq. (2.7) in the form of a Wiener integral, as shown in Appendix A. Not surprisingly, however, the exact solution is not amenable to further analysis unless some approximation scheme is involved. The way forward in analyzing the stochastic learning dynamics is to replace $\mathbf{\Sigma}_{xx} + \overline{\mathbf{\Sigma}}_{xx}(t)$ by its time-independent part $\mathbf{\Sigma}_{xx}$, i.e., the full-batch correlation matrix, in Eq. (2.7). Then one can find an exact solution for $\overline{\mathbf{W}}(t)$ in the form

$$\overline{\mathbf{W}}(t) = \left[\overline{\mathbf{W}}(0) + \int_0^t \overline{\mathbf{\Sigma}}'_{yx}(t') e^{\mathbf{\Sigma}_{xx}t'} \right] e^{-\mathbf{\Sigma}_{xx}t}, \quad (2.8)$$

where $\overline{\mathbf{\Sigma}}'_{yx}(t) = \overline{\mathbf{\Sigma}}_{yx}(t) - \mathbf{W}_0\overline{\mathbf{\Sigma}}_{xx}(t)$. To complete the analysis, we need also to examine the influence of the term we ignored, i.e., $\overline{\mathbf{\Sigma}}_{xx}(t)$, and this is done in Appendix A. The result, as expected, is that as long as $\overline{\mathbf{\Sigma}}_{xx}(t)$ is sufficiently small compared to $\mathbf{\Sigma}_{xx}$, the correction to the solution we obtained in Eq. (2.8) is perturbatively small in $\overline{\mathbf{\Sigma}}_{xx}(t)$.

We can write down the long-time correlation matrix for $\overline{\mathbf{W}}(t)$ as

$$\begin{aligned} \mathbf{\Sigma}_{\overline{\mathbf{W}}} &= \langle [\overline{\mathbf{W}}(t)]^T \overline{\mathbf{W}}(t) \rangle \\ &= \int_0^t dt' \int_0^t dt'' e^{\mathbf{\Sigma}_{xx}(t'-t)} \langle [\overline{\mathbf{\Sigma}}'_{yx}(t')]^T \overline{\mathbf{\Sigma}}'_{yx}(t'') \rangle e^{\mathbf{\Sigma}_{xx}(t''-t)} \\ &= \int_0^t dt' e^{\mathbf{\Sigma}_{xx}(t'-t)} 2\mathbf{D} e^{\mathbf{\Sigma}_{xx}(t'-t)}. \end{aligned} \quad (2.9)$$

The part of the $\overline{\mathbf{W}}(t)$ solution in Eq. (2.8) that depends on the initial condition $\overline{\mathbf{W}}(0)$ can be dropped, due to its exponential decay at long time t . Furthermore, we assumed a zero-range temporal correlation $\langle [\overline{\Sigma}'_{yx}(t')]^T \overline{\Sigma}'_{yx}(t'') \rangle = 2\mathbf{D}\delta(t' - t'')$ in arriving at the final form in Eq. (2.9). Whether this sort of assumption is justifiable remains to be checked, and the test performed in the next section finds the answer to be in the affirmative.

Going back to Eq. (2.9) for now, we can easily deduce from it the following identity,

$$\Sigma_{xx}\Sigma_{WW} + \Sigma_{WW}\Sigma_{xx} = 2\mathbf{D}, \quad (2.10)$$

for $\Sigma_{WW} \equiv \Sigma_{WW}(t \rightarrow \infty)$. One notes an unmistakable similarity between this relation and the one derived in Eq. (1.5) from Langevin dynamics. Since the former relation is typically known as the FDT in the Langevin dynamics, we label the mathematically analogous relation derived in Eq. (2.10) as the FDT-type relation in the stochastic linear learning dynamics. Granted, there is no physical dynamics underlying our SGD equation. In deriving the formula (2.10), however, the existence of a physical dynamics is irrelevant as long as there is a good mathematical analogy between the two situations. Although Σ_{xx} and $\Sigma'_{yx}(t)$ in Eq. (2.7) are not the true dissipative and stochastic forces as in the Langevin dynamics, we do not hesitate to label Eq. (2.10) as the FDT-type relation on the basis of the formal, mathematical analogy. Whenever possible, exploiting an exact relation such as Eq. (2.10) is helpful in the analysis of a given problem.

Technically, an even more refined form of the FDT-type relation than the one shown in Eq. (2.10) can be derived. The interested reader is referred to Appendix B. Meanwhile, we work with the more tractable version as shown in Eq. (2.10) when we try to check the validity of the FDT-type relation numerically in the next section.

III. NUMERICAL EXPERIMENTS

For sufficiently small time $t = h$ we can solve the stochastic Eq. (2.4) approximately,

$$\begin{aligned} \mathbf{W}(h) &\approx \left[\mathbf{W}(0) + \int_0^h \Sigma_{yx}(t') e^{\Sigma_{xx}(0)t'} dt' \right] e^{-\Sigma_{xx}(0)h} \\ &\approx \mathbf{W}(0)[1 - \Sigma_{xx}(0)h] + \int_0^h \Sigma_{yx}(t') dt'. \end{aligned} \quad (3.1)$$

We can further divide up the interval $t \in [0, h]$ into M equal segments, each of width $\varepsilon \equiv h/M$, and write $\Sigma_{xx}(0) \rightarrow M^{-1} \sum_{i=1}^M \Sigma_{xx}(i \cdot \Delta)$, $\int_0^h \Sigma_{yx}(t') dt' \rightarrow \varepsilon \sum_{i=1}^M \Sigma_{yx}(i \cdot \Delta)$. In the end, Eq. (3.1) turns into a recursive formula

$$\mathbf{W}^{(n+1)} = \mathbf{W}^{(n)}[1 - \varepsilon \Sigma_{xx}^{(n)}] + \varepsilon \Sigma_{yx}^{(n)}, \quad (3.2)$$

where $\Sigma_{xx}^{(n)}$ and $\Sigma_{yx}^{(n)}$ are averages over the minibatch of size MN_m . From now on, we will simply refer to MN_m as the minibatch size N_m . At sufficiently large n , $\mathbf{W}^{(n)}$ converges to \mathbf{W}_0 , with small fluctuations whose properties will be the subject of investigation.

In particular, we are interested in whether the small fluctuations near the equilibrium $\mathbf{W} = \mathbf{W}_0$ obey the relation derived in Eq. (2.10). To test the claim, we employ three

representative data sets that researchers in machine learning frequently employ: MNIST, CIFAR-10, and EMNIST Letters (abbreviated as EMNIST from here on) [17]. As is well known, MNIST and CIFAR-10 consist of ten different objectives or output vectors \mathbf{y}^α , represented by one-hot vectors $(1, 0, \dots, 0)$ through $(0, \dots, 0, 1)$. Twenty-six alphabets are represented by as many output vectors in the case of EMNIST. The pixel sizes are 28×28 for both MNIST and EMNIST, and 32×32 for CIFAR-10. There are 60 000 (50 000 and 140 000) training data samples in MNIST (CIFAR-10 and EMNIST), which we use to form the full batch.

The minibatch update scheme is implemented by randomly choosing $N_m = 5, 500$, and 100 data from the batch in the case of MNIST, CIFAR-10, and EMNIST, respectively, and using them to generate the n th correlation functions $\Sigma_{xx}^{(n)}$ and $\Sigma_{yx}^{(n)}$. Once a minibatch selection is complete, we return the data back to the full batch for the next round of minibatch selection. This could induce some correlations between different minibatches as the same data may appear repeatedly over different minibatches. We will analyze the correlations between different minibatches later on, and find the correlation effect at $n \neq n'$ to be rather small. The learning rate of $\varepsilon = 2 \times 10^{-7}$, 2×10^{-8} , 5×10^{-7} (MNIST, CIFAR-10, EMNIST) was used for the update. The update $\mathbf{W}^{(n)} \rightarrow \mathbf{W}^{(n+1)}$ takes place according to the simple update scheme in Eq. (3.2). The minibatch size and the learning rate are chosen such that a good convergence to equilibrium takes place in the iteration. We have not systematically investigated how the efficiency of the convergence depends on either N_m or ε , as that is not the main objective of this research, but only confirmed that different choices do not affect the final quality of the data as long as the convergence is achieved.

In order to derive the relation (2.10), a recourse was made to an approximation in which the mini-batch $\Sigma_{xx}(t)$ was replaced by the full-batch Σ_{xx} —see Eq. (2.7) and the ensuing discussion. It was also mentioned that this approximation does not invalidate the theorem (2.10) as long as the difference between the full-batch Σ_{xx} and the minibatch $\Sigma_{xx}(t)$ remains sufficiently small. On the other hand, the actual numerical integration of the SGD equation through Eqs. (3.1) and (3.2) is done with time-dependent, minibatch $\Sigma_{xx}^{(n)}$ rather than the time-independent, full-batch Σ_{xx} . Despite the difference, as we now discuss, the relation (2.10) is obeyed quite nicely in our simulation that uses the minibatch $\Sigma_{xx}^{(n)}$ in the updates. Figure 1 shows the Euclidean norm ratio $\|\overline{\Sigma}_{xx}^{(n)}\|/\|\Sigma_{xx}\|$ after the equilibrium has been reached. To our surprise, the ratio is not necessary very small as our analytical derivation of Eq. (2.10) forced us to assume. Nevertheless, we were able to achieve good convergence $\mathbf{W}^{(n)} \rightarrow \mathbf{W}_0$ numerically, and the formula (2.10) is nicely reproduced in the simulation.

The convergence of $\mathbf{W}^{(n)}$ at large n to the equilibrium value $\mathbf{W}_0 = \Sigma_{xy}\Sigma_{xx}^{-1}$ was checked by measuring the inner product of the $\mathbf{W}^{(n)}$ and \mathbf{W}_0 divided by their norms: $\cos[\theta^{(n)}] = \mathbf{W}^{(n)} \cdot \mathbf{W}_0 / \|\mathbf{W}^{(n)}\| \|\mathbf{W}_0\|$. The inner product of two matrices is defined by taking a product of the matrix elements sharing the same (ij) index and making a sum over all (ij) 's. The norm is the square root of the inner product of a matrix with itself. Once the steady state is reached, e.g., $\cos[\theta^{(n)}] \gtrsim 0.999$, we start calculating the \mathbf{W} -correlation matrix Σ_{WW} and the \mathbf{D} .

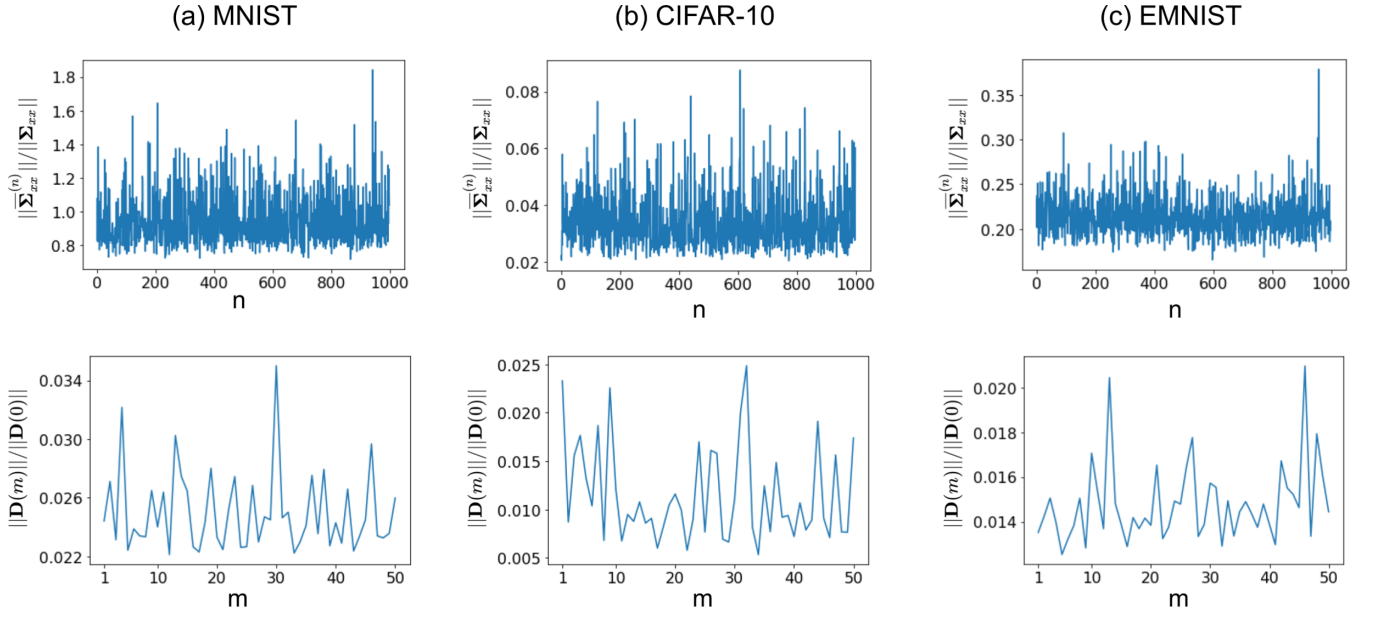


FIG. 1. (a)–(c) Euclidean norm ratio $\|\bar{\Sigma}_{xx}^{(n)}\|/\|\Sigma_{xx}\|$ over several minibatches labeled by n after the equilibrium is reached. (b) The Euclidean norm of the unequal time correlation matrix $\|\mathbf{D}(m)\|/\|\mathbf{D}(0)\|$ [Eq. (3.4)]. The minibatch size used in each figure is $N_m = 5, 500$, and 100 for MNIST, CIFAR-10, and EMNIST, respectively.

For Σ_{WW} we use

$$\Sigma_{WW} \simeq \frac{1}{n_2 - n_1} \sum_{n=n_1+1}^{n_2} [\overline{\mathbf{W}}^{(n)}]^T \overline{\mathbf{W}}^{(n)} \quad (3.3)$$

for some large n_1, n_2 well within the equilibrium region. The sampling data $n_2 - n_1 \sim 10^4$ were sufficient to guarantee good averaging and a clear image for Σ_{WW} , ready for subsequent analysis. The overline indicates that the difference between the n th learning matrix $\mathbf{W}^{(n)}$ and \mathbf{W}_0 must be used in obtaining Σ_{WW} .

Deducing the diffusion matrix \mathbf{D} is a bit more challenging. First of all, it is obtained as the correlator of $\bar{\Sigma}_{yx}^{(n)'} = \bar{\Sigma}_{yx}^{(n)} - \mathbf{W}_0 \bar{\Sigma}_{xx}^{(n)}$. In general, an unequal time correlator of these quantities may be defined as

$$\mathbf{D}(m) \equiv \frac{1}{n_2 - n_1} \sum_{n=n_1+1}^{n_2} [\bar{\Sigma}_{yx}^{(n+m)'}]^T \bar{\Sigma}_{yx}^{(n)'}. \quad (3.4)$$

We have performed an analysis of $\mathbf{D}(m)$ for several m and shown the result in Fig. 1. For $m \neq 0$, the size of the Euclidean norm $\|\mathbf{D}(m)\|$ becomes no more than 4% of the value at $m = 0$, suggesting the uncorrelated nature of the matrices at different “times.” To obtain the diffusion matrix \mathbf{D} , we choose $m = 0$ and take $\mathbf{D} = \mathbf{D}(0)$. Now that we have both Σ_{WW} and \mathbf{D} from the numerical data, we can compare their values and look for proof of proportionality between them. Both \mathbf{D} and $\Sigma_{xx} \Sigma_{WW} + \Sigma_{WW} \Sigma_{xx}$ are displayed graphically in Fig. 2. It turns out the correlators exhibit a highly periodic structure with period a coming from the $a \times a$ pixel size of each data set. (The original $a = 28$ dimension of the MNIST and EMNIST was chopped at the boundary to $a = 24$. Otherwise it was difficult to get the full-batch inverse Σ_{xx}^{-1}).

Due to the highly periodic structure of the real-space images of $\Sigma_{xx} \Sigma_{WW} + \Sigma_{WW} \Sigma_{xx}$ and \mathbf{D} , only a handful of Fourier

peaks at $\mathbf{k} = (k_x, k_y)$ given by multiples of $2\pi/a$ were significant. Figure 2 shows the Fourier components along $\mathbf{k} = (k_x, 0)$ normalized by the value at $\mathbf{k} = (0, 0)$. The near-perfect match in the Fourier analysis of both $\Sigma_{xx} \Sigma_{WW} + \Sigma_{WW} \Sigma_{xx}$ and \mathbf{D} is not *a priori* obvious, and must be attributed to the FDT theorem at work in the stochastic linear learning dynamics. The accuracy of the FDT theorem seems somewhat reduced in the case of CIFAR-10 in comparison to the other two data sets, as one can see from comparing the images in the third row in Fig. 2. We believe this is due to the conversion of the original color image to the black-and-white image before processing the CIFAR-10 data, or perhaps the complexity of the CIFAR-10 images compared to the other two cases, or a combination of both. A further investigation of the origin of the reduced accuracy of the FDT theorem for the CIFAR-10 data set is the subject of future investigation.

IV. DISCUSSION

Our work addresses a FDT-type relation in the stochastic linear learning dynamics. The relation derived in Eq. (2.10) is found to hold quite well for a number of machine learning data sets. The analogy to the Langevin dynamics naturally gives rise to an interpretation of the input covariance matrix Σ_{xx} as the effective friction, and the input-output variance Σ_{yx} as the effective stochastic force in the learning dynamics. Although it is more or less obvious from the context, we want to emphasize once again that the analogy to the celebrated fluctuating-dissipation theorem in statistical physics [15,16] is a purely formal and mathematical one. The mathematical similarity of the Langevin Eq. (1.3) and the stochastic gradient Eq. (2.4) is what makes the derivation of the FDT-type relation (2.10) possible.

We have made several attempts to go beyond the simple stochastic linear learning scheme. For one, we tried placing a

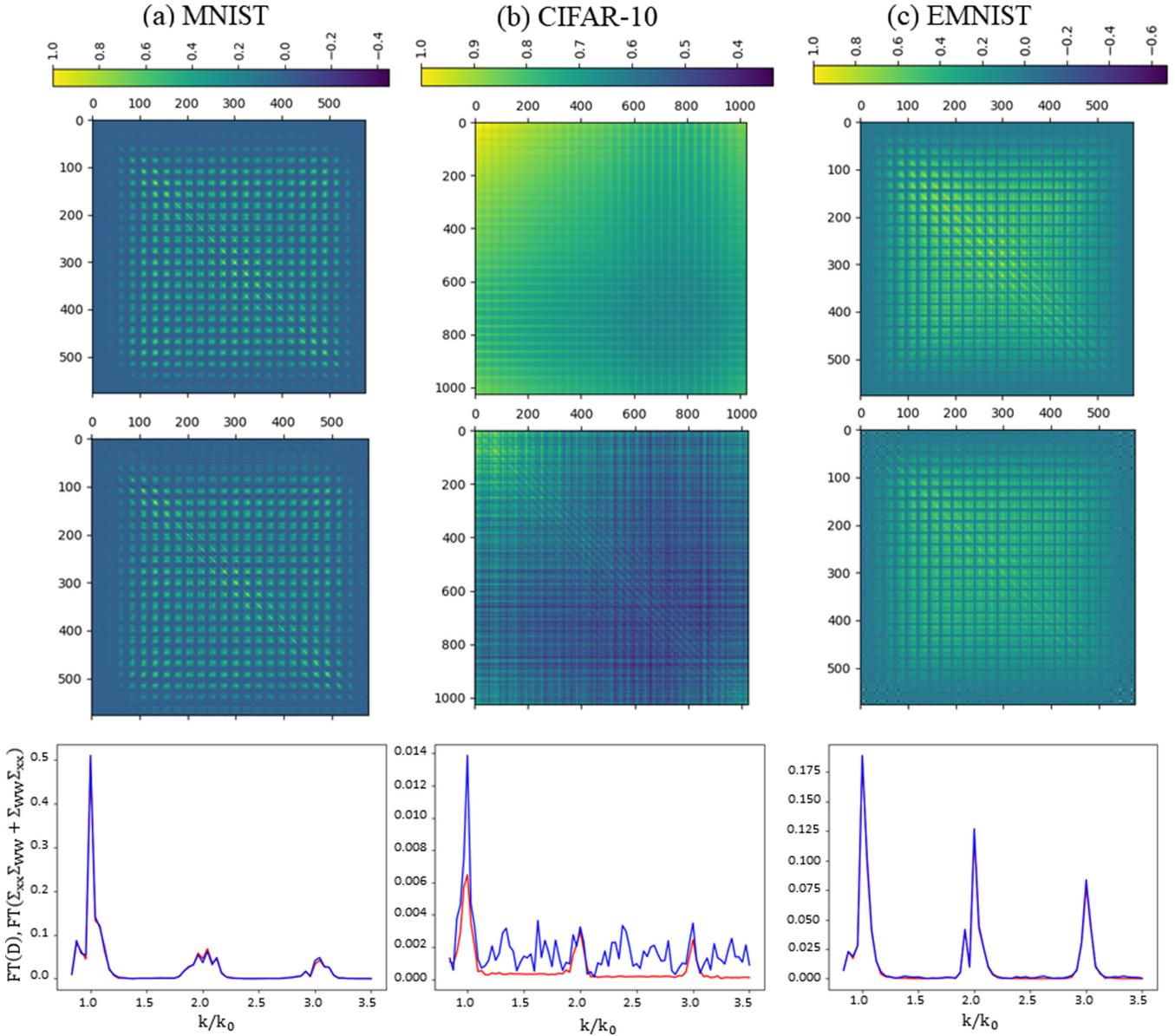


FIG. 2. Fluctuation analysis for (a) MNIST, (b) CIFAR-10, and (c) EMNIST data sets. Top: Plots of \mathbf{D} obtained from each data set. Middle: Plots of $\Sigma_{xx}\Sigma_{WW} + \Sigma_{WW}\Sigma_{xx}$. Bottom: Normalized Fourier components for \mathbf{D} (red) and $\Sigma_{xx}\Sigma_{WW} + \Sigma_{WW}\Sigma_{xx}$ (blue) plotted along $\mathbf{k} = (k_x, 0)$ with $k_0 = 2\pi/a$.

convolutional neural network (CNN) layer before the neural network layer \mathbf{W} . As shown in Appendix C, this formulation naturally leads to FDT in terms of the CNN-filtered input data sets $\mathbf{X}^\alpha = \mathbf{C} \otimes \mathbf{x}^\alpha$, where \otimes represents the CNN operation. The FDT holds with respect to the renormalized data sets \mathbf{X}^α . In another attempt, we tried introducing non-linearity explicitly by using an alternative error function $E = (2N)^{-1} \sum_{\alpha=1}^N \sum_{i=1}^n (y_i^\alpha - z_i^\alpha)^2$ with the sigmoid function $z_i^\alpha = [e^{-\sum_{j=1}^n W_{ij} x_j^\alpha} + 1]^{-1}$ parametrized by the learning matrix \mathbf{W} . Such a formulation leads to the dynamics $d\mathbf{W}/dt$ that is, unfortunately, highly nonlinear and defies further analytical treatment.

The FDT-type relation in the stochastic learning was noticed some years earlier by Yaida [9]. His derivation of the so-called FDT relation avoids any use of an explicit error function and relies solely on the stationary property of

observables after the learning process has saturated. It is a powerful formulation in the sense that the relations apply to an arbitrary learning architecture with nonlinearities. On the other hand, by avoiding the stochastic differential equation formulation, the connection that his relations have with the FDT in statistical physics becomes somewhat vague. More seriously, when our error function is used to work out his formulas, the outcome does not match our FDT formula derived in Eq. (2.10). This leads us to suspect that there may be multiple FDT-type theorems governing the stationary states of learning, with both our formula and his addressing different facets.

We have investigated whether, writing $\Sigma_{xx}\Sigma_{WW}$ in Eq. (2.10) as the sum $\Sigma_{xx}\Sigma_{WW} = \mathbf{D} + \mathbf{Q}$, there will be a significant contribution of the antisymmetric matrix \mathbf{Q} . A crude measure of the significance of \mathbf{Q} relative to \mathbf{D} is the maximum

value of the matrix elements in \mathbf{Q} divided by that of \mathbf{D} . The results are 0.12, 0.096, and 0.045 for MNIST, CIFAR-10, and EMNIST, respectively, suggesting that the antisymmetric components are probably very small and insignificant.

ACKNOWLEDGMENTS

The Python code used in the numerical experiment is available online [18]. J.H.H. acknowledges fruitful discussion with and input on the manuscript from Ping Ao, J. H. Jo, S. B. Lim, J. D. Noh, Vinit Singh, and Hayong Yun.

APPENDIX A: FULL WIENER INTEGRAL

The fluctuation-dissipation theorem (2.10) for stochastic linear learning dynamics was obtained assuming time-independent Σ_{xx} and time-dependent $\Sigma_{yx}(t)$. It means a full-batch Σ_{xx} and a minibatch $\Sigma_{yx}(t)$ are assumed in the derivation. On the other hand, the numerical integration of Eq. (2.4) or (2.7) was done in our experiment using both minibatch $\Sigma_{xx}(t)$ and $\Sigma_{yx}(t)$. In spite of the difference, the FDT seems to hold quite well numerically. We reexamine the

full stochastic equation for linear learning dynamics written in Eq. (2.7),

$$\begin{aligned} \frac{d\mathbf{W}}{dt} &= -\mathbf{W}(\Sigma_{xx} + \Sigma_{xx}(t)) + \Sigma_{yx}(t) - \mathbf{W}_0 \Sigma_{xx}(t) \\ &= -\mathbf{W}(\mathbf{A} + \mathbf{a}(t)) + \mathbf{b}(t). \end{aligned} \quad (\text{A1})$$

Notations have been simplified in the second line. The full solution to this can be found using the Wiener path integral formulation, familiarly known in physics as the Feynman path integral.

First, one makes a decomposition $\mathbf{W} \rightarrow \mathbf{W}e^{-\mathbf{A}t}$, and derives the equation in terms of the new \mathbf{W} :

$$\begin{aligned} \frac{d\mathbf{W}}{dt} &= -\mathbf{W}e^{-\mathbf{A}t}\mathbf{a}(t)e^{\mathbf{A}t} + \mathbf{b}(t)e^{\mathbf{A}t} \\ &= -\mathbf{W}\mathbf{a}_I(t) + \mathbf{b}_I(t). \end{aligned} \quad (\text{A2})$$

The subscript I is meant to indicate the ‘‘interaction picture’’ representation of the learning dynamics following similar jargon in quantum mechanics. With both $\mathbf{a}_I(t)$ and $\mathbf{b}_I(t)$ being time dependent, one can find the solution in the path integral form,

$$\begin{aligned} \mathbf{W}(t) &= \left(\mathbf{W}(0) + \int_0^t dt' \mathbf{b}_I(t') P_{t' \rightarrow 0} \exp \left[\int_0^{t'} dt'' \mathbf{a}_I(t'') \right] \right) P_{0 \rightarrow t} \exp \left[- \int_0^t \mathbf{a}_I(t') dt' \right] \\ &= \mathbf{W}(0) P_{0 \rightarrow t} \exp \left[- \int_0^t \mathbf{a}_I(t') dt' \right] + \int_0^t dt' \mathbf{b}_I(t') P_{t' \rightarrow t} \exp \left[- \int_{t'}^t \mathbf{a}_I(t'') dt'' \right]. \end{aligned} \quad (\text{A3})$$

The symbol $P_{0 \rightarrow t}$ means that the operator (matrix) defined at time $t' = 0$ is to be written at the far left, and the one at time $t' = t$ at the far right. The symbol $P_{t' \rightarrow t}$ means that the $t'' = t'$ operator appears on the far left, and the $t'' = t$ operator at the far right. The usual composition rule of path integrals gives the second expression of the second line.

The first term $\sim \mathbf{W}(0)$ can be ignored because the long-time result should not depend on the initial condition. Furthermore, we are interested in terms that are only first order in the fluctuation. Under these assumptions we can write the result in Eq. (A3) approximately,

$$\mathbf{W}(t) - \mathbf{W}_0 \approx \int_0^t dt' \mathbf{b}(t') e^{\mathbf{A}(t'-t)}. \quad (\text{A4})$$

The full definition of the \mathbf{W} matrix is restored in the above. Note that this is exactly the same expression obtained earlier in Eq. (2.8), without the initial $\mathbf{W}(0)$. Hence the FDT derived earlier is valid to the leading order in the fluctuation.

APPENDIX B: REFINEMENT OF THE FDT THEOREM

It is possible to define a more general kind of diffusion matrix than the one presented in Eq. (2.9),

$$\begin{aligned} \langle [\Sigma_{yx}(t) - \Sigma_{yx}]_{i\alpha} [\Sigma_{yx}(t') - \Sigma_{yx}]_{j\beta} \rangle \\ = 2D_{i\alpha, j\beta} \delta(t - t'), \end{aligned} \quad (\text{B1})$$

that does not involve the summation over the output indices i, j . A similar generalization for the W -correlation matrix

gives

$$\begin{aligned} \langle (W - W_0)_{ij} (W - W_0)_{kl} \rangle &= \Sigma_{ij,kl} \\ &= \int_0^t dt' 2D_{i\alpha, k\beta} [e^{\mathbf{A}(t'-t)}]_{\alpha j} [e^{\mathbf{A}(t'-t)}]_{\beta l}, \end{aligned} \quad (\text{B2})$$

where the result from Eq. (A4) is used to reach the second line.

If we fix the two output indices i and k in the above relation, then one can rewrite it in the following fashion:

$$\begin{aligned} (\Sigma_{i,k})_{jl} &= \int_0^t dt' [e^{\mathbf{A}^T(t'-t)}]_{j\alpha} (2D_{i,k})_{\alpha\beta} [e^{\mathbf{A}(t'-t)}]_{\beta l} \\ &= \int_0^t dt' [e^{\mathbf{A}^T(t'-t)} 2D_{i,k} e^{\mathbf{A}(t'-t)}]_{jl}. \end{aligned} \quad (\text{B3})$$

In other words, for a given pair of output indices (i, k) , we have a matrix relation

$$\Sigma_{i,k} = \int_0^t dt' [e^{\mathbf{A}^T(t'-t)} 2\mathbf{D}_{i,k} e^{\mathbf{A}(t'-t)}], \quad (\text{B4})$$

subject to the same kind of identity as before:

$$\begin{aligned} \mathbf{A}^T \Sigma_{i,k} + \Sigma_{i,k} \mathbf{A} &= \int_0^t dt' \frac{d}{dt'} [e^{\mathbf{A}^T(t'-t)} 2\mathbf{D}_{i,k} e^{\mathbf{A}(t'-t)}] \\ &= 2\mathbf{D}_{i,k}. \end{aligned} \quad (\text{B5})$$

In conclusion, the FDT holds irrespective of the choice of output indices (i, k) .

APPENDIX C: STOCHASTIC LINEAR LEARNING WITH CNN LAYER

Adding a CNN layer before the \mathbf{W} layer and optimizing the error function with respect to both \mathbf{W} and the CNN filter matrix turns out to be within the mathematically tractable scope. The CNN layer transforms the input vector \mathbf{x}^α into a modified input vector $\mathbf{X}^\alpha = \mathbf{C} \otimes \mathbf{x}^\alpha$ according to the recipe

$$X_j^\alpha = \sum_{i_x, i_y=1}^c C_i X_{i+j-1}^\alpha. \quad (\text{C1})$$

We switch to a two-dimensional vector notation of the indices, $\mathbf{i} = (i_x, i_y)$, $\mathbf{j} = (j_x, j_y)$, and write $\mathbf{1} = (1, 1)$. The convolution operator \mathbf{C} is a $c \times c$ dimensional matrix. The error function is the same as before, Eq. (2.1) with \mathbf{X}^α taking the place of \mathbf{x}^α :

$$E = \frac{1}{2N} \sum_{\alpha=1}^N (\mathbf{y}_\alpha - \mathbf{W}\mathbf{X}_\alpha)^T (\mathbf{y}_\alpha - \mathbf{W}\mathbf{X}_\alpha). \quad (\text{C2})$$

Minimization of the error must take place with respect to both \mathbf{W} and \mathbf{C} .

Taking the derivative of the error E with respect to an element of the convolution matrix C_i can be done by using the chain rule,

$$\begin{aligned} \frac{\partial E}{\partial C_i} &= \sum_{\alpha, j} \frac{\partial E}{\partial X_j^\alpha} \frac{\partial X_j^\alpha}{\partial C_i} \\ &= \sum_{\alpha, j} \frac{\partial E}{\partial X_j^\alpha} X_{j+i-1}^\alpha. \end{aligned} \quad (\text{C3})$$

Supplemented by $\delta E / \delta \mathbf{X}^\alpha = N^{-1} \mathbf{W}^T (\mathbf{W}\mathbf{X}^\alpha - \mathbf{y}^\alpha)$, we arrive at

$$\frac{dC_i}{dt} = \frac{1}{N} \sum_{\alpha, j} (\mathbf{W}^T \mathbf{y}^\alpha - \mathbf{W}^T \mathbf{W}\mathbf{X}^\alpha)_{j, k} X_{j+i-1}^\alpha. \quad (\text{C4})$$

The first term on the right-hand side (rhs) becomes

$$\frac{1}{N} \sum_{\alpha, j, k} W_{k, j} y_k^\alpha X_{j+i-1}^\alpha = \sum_{k, j} W_{k, j} \Sigma_{k, j+i-1}^{yx}. \quad (\text{C5})$$

This is a summation over the output index k , and a convolution with respect to the input indices. The surviving

index is i , which covers the elements of the filter matrix \mathbf{C} . We have $\mathbf{1} \leq j \leq L - c + 1$, $\mathbf{1} \leq i \leq c$, and $\mathbf{1} \leq j + i - 1 \leq L$, which keeps track of the range of indices in a correct manner. For the second term on the rhs we get

$$\begin{aligned} &\frac{1}{N} \sum_{\alpha, j} (\mathbf{W}^T \mathbf{W}\mathbf{X}^\alpha)_{j, k} X_{j+i-1}^\alpha \\ &= \frac{1}{N} \sum_{\alpha, j, k} (\mathbf{W}^T \mathbf{W})_{j, k} X_k^\alpha X_{j+i-1}^\alpha \\ &= \frac{1}{N} \sum_{\alpha, j, k, l} (\mathbf{W}^T \mathbf{W})_{j, k} C_l X_{l+k-1}^\alpha X_{j+i-1}^\alpha \\ &= \sum_l C_l \left[\sum_{j, k} (\mathbf{W}^T \mathbf{W})_{j, k} (\Sigma^{xx})_{l+k-1, j+i-1} \right]. \end{aligned} \quad (\text{C6})$$

We can define two new quantities,

$$\begin{aligned} P_{i, l} &\equiv \sum_{j, k} (\mathbf{W}^T \mathbf{W})_{j, k} (\Sigma^{xx})_{l+k-1, j+i-1} = P_{l, i}, \\ Q_i &\equiv \sum_{j, k} W_{k, j} \Sigma_{k, j+i-1}^{yx}, \end{aligned} \quad (\text{C7})$$

to simplify the equation

$$\frac{dC_i}{dt} = Q_i - \sum_l P_{i, l} C_l. \quad (\text{C8})$$

The \mathbf{W} matrix appears in various places in the definition of \mathbf{P} and \mathbf{Q} , and can be obtained from

$$\frac{d\mathbf{W}}{dt} = -\mathbf{W}\Sigma_{XX} + \Sigma_{yX}, \quad (\text{C9})$$

where

$$\begin{aligned} \Sigma_{i, j}^{XX} &= \sum_{k, l} C_k C_l \Sigma_{k+i-1, l+j-1}^{xx}, \\ \Sigma_{i, j}^{yX} &= \sum_k C_k \Sigma_{i, j+k-1}^{yx}. \end{aligned} \quad (\text{C10})$$

The two Eqs. (C8) and (C9) can be solved simultaneously by GSD. At equilibrium we have $\mathbf{C} = \mathbf{P}^{-1} \mathbf{Q}$ but this formula is a bit misleading as the filter \mathbf{C} enters implicitly in both \mathbf{P} and \mathbf{Q} as well.

[1] M. Welling and Y. W. Teh, in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11 (Omnipress, Madison, WI, 2011), pp. 681–688.
[2] Y.-A. Ma, T. Chen, and E. B. Fox, in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15 (MIT Press, Cambridge, MA, 2015), pp. 2917–2925.
[3] S. Mandt, M. D. Hoffman, and D. M. Blei, in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16 (JMLR.org, New York, 2016), pp. 354–363.

[4] P. Chaudhari and S. Soatto, in *2018 Information Theory and Applications Workshop (ITA)* (IEEE, New York, 2018), pp. 1–10.
[5] P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, and R. Zecchina, *J. Stat. Mech.: Theory Exp.* (2019) 124018.
[6] A. M. Saxe, J. L. McClelland, and S. Ganguli, Exact solutions to the nonlinear dynamics of learning in deep linear neural networks, [arXiv:1312.6120](https://arxiv.org/abs/1312.6120).
[7] A. M. Saxe, J. L. McClelland, and S. Ganguli, *Proc. Natl. Acad. Sci. U.S.A.* **116**, 11537 (2019).

- [8] Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli, *Annu. Rev. Condens. Matter Phys.* **11**, 501 (2020).
- [9] S. Yaida, Fluctuation-dissipation relations for stochastic gradient descent, [arXiv:1810.00004](https://arxiv.org/abs/1810.00004).
- [10] P. Ao, *J. Phys. A: Math. Gen.* **37**, L25 (2004).
- [11] C. Kwon, P. Ao, and D. J. Thouless, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 13029 (2005).
- [12] L. Yin and P. Ao, *J. Phys. A: Math. Gen.* **39**, 8593 (2006).
- [13] C. Kwon, J. D. Noh, and H. Park, *Phys. Rev. E* **83**, 061145 (2011).
- [14] C. Kwon and P. Ao, *Phys. Rev. E* **84**, 061106 (2011).
- [15] R. Kubo, M. Toda, and N. Hashitsume, *Statistical Physics II* (Springer, Berlin, 1991).
- [16] H. Risken and T. Frank, *The Fokker-Planck Equation* (Springer, Berlin, 1996).
- [17] J. T. G. Cohen, S. Afshar, and A. van Schaik, EMNIST: an extension of MNIST to handwritten letters, [arXiv:1702.05373](https://arxiv.org/abs/1702.05373).
- [18] <https://github.com/lemonseed117/FDT-Stochastic.git>