Field Theoretical Analysis of On-Line Learning of Probability Distributions

Toshiaki Aida

Department of Physics, Tokyo Institute of Technology, Oh-okayama, Meguro-ku, Tokyo, Japan
and Tokyo Metropolitan College of Aeronautical Engineering, Minami-senjyu, Arakawa-ku, Tokyo, Japan
(Received 24 May 1999)

On-line learning of probability distributions is analyzed from the field theoretical point of view. We can obtain an optimal on-line learning algorithm, since a renormalization group enables us to control the number of degrees of freedom of a system according to the number of examples. We do not learn parameters of a model, but probability distributions themselves. Therefore, the algorithm requires no *a priori* knowledge of a model.

PACS numbers: 02.50.Fz, 11.10.Hi, 84.35.+i, 89.70.+c

The methods of statistical mechanics have provided the framework to analyze learning and generalization by neural networks and have achieved considerable success. This implies the fundamental connection between learning and statistical mechanical theories.

For example, Bialek *et al.* [1] made clear the batch learning problem of probability distributions from the field theoretical point of view. They optimally controlled the number of degrees of freedom of a learning system through the scaling, which was naturally introduced by the quantum field theory. It is crucial to control the number of degrees of freedom in learning problems. Learning a lot of features requires us to introduce a large number of degrees of freedom, while too many degrees of freedom increase ambiguities when we are given a small number of examples. This problem has been considered in various ways [2].

Can we expect that the scaling leads to an optimal algorithm also in on-line learning problems? On-line learning has the advantages of less storage of data and less peak computation than batch learning. Some elaborated algorithms achieve the same optimal learning rate as the corresponding batch ones [3]. In this Letter, we will show that the scaling analysis derives an optimal on-line learning algorithm, when we identify the number of examples with a scale parameter.

We observe a sequence of D-dimensional sample points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, which are drawn independently from an unknown probability distribution $Q^*(\mathbf{x})$. How can we infer the distribution, which is specified by an infinite number of parameters in general? A finite number of sample points do not enable us to determine $Q^*(\mathbf{x})$ uniquely but allow us only to make a probabilistic description. We start from the probability of a distribution $Q(\mathbf{x})$, which is given via Bayes' rule [1].

$$P[Q \mid \mathbf{x}_1, \ldots, \mathbf{x}_N] = \frac{P[\mathbf{x}_1, \ldots, \mathbf{x}_N \mid Q]P[Q]}{P(\mathbf{x}_1, \ldots, \mathbf{x}_N)}, = \frac{Q(\mathbf{x}_1)\cdots Q(\mathbf{x}_N)P[Q]}{\int \mathcal{D}Q\, Q(\mathbf{x}_1)\cdots Q(\mathbf{x}_N)P[Q]}. \quad (1)$$

Here, $P[Q]$ denotes a "prior distribution," which is a probability density in the space of probability distributions, and limits the function Q to possible classes.

Let us consider the properties of the prior distribution. Introducing a field valuable $\phi(\mathbf{x})$ which takes a value $-\infty < \phi < \infty$, we rewrite the distribution $Q(\mathbf{x})$ as $Q(\mathbf{x}) = e^{-\phi(\mathbf{x})}/l^D$. We note that l is not a parameter but a unit length scale, which is different from the case of Ref. [1]. It only plays a role to ensure the consistency of dimensions and is fixed to be one in numerical studies. To make our learning problem well posed, it is known that the prior distribution must not give ultraviolet divergences. Therefore, we choose such a prior distribution as imposes that the function $\phi(\mathbf{x})$ has to be smooth enough ($2\alpha > D$):

$$P[Q] = \frac{1}{Z_0}\exp\left[-\frac{l^{2\alpha-D}}{2}\int d^D x\,(\partial_x^\alpha \phi)^2 - \frac{1}{l^D}\int d^D x\, F(\phi)\right] \times \delta\left[\frac{1}{l^D}\int d^D x\, e^{-\phi} - 1\right]. \quad (2)$$

Hereafter, we will use the notation $\partial_x^\alpha \equiv \partial_{x_{i_1}}\cdots\partial_{x_{i_\alpha}}$, and mean by the repeated indices $i_1, \ldots, i_\alpha$ the sums from 1 to the dimension D. We note that we have introduced an unknown function $F[\phi(\mathbf{x})]$, which will be determined later so that we may treat fluctuation effects in a renormalizable way. The delta function gives the constraint to normalize the probability distribution, and the factor Z_0 is a normalization constant.

Putting the prior (2) into Bayes' rule (1), we obtain the partition function as $Z = \frac{1}{Z_0}\int \frac{d\lambda}{2\pi}\int \mathcal{D}\phi \times \exp[-\frac{1}{g}S(\phi, \lambda)]$, where

$$S = \frac{l^{2\alpha-D}}{2}\int d^D x\,(\partial_x^\alpha \phi)^2 + \frac{1}{l^D}\int d^D x F(\phi) + i\lambda\left[\frac{1}{l^D}\int d^D x\, e^{-\phi} - 1\right] + N\int d^D x\, P_N \phi\,. \quad (3)$$

The constant g counts fluctuation effects and is set to be one in numerical analysis. $P_N(\mathbf{x})$ is the distribution

 0031-9007/99/83(17)/3554(4)$15.00

of sample data, defined as $P_N(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\delta^D(\mathbf{x} - \mathbf{x}_i)$. The action (3) explicitly shows that we have no scale parameter other than the number of example data N. Therefore, we can conclude that the number N sets the length scale to observe an unknown distribution and determines the behavior of a fluctuating system.

In order to proceed further, we will consider the asymptotic situation $N \to \infty$. We expand the action (3) around classical solutions and perform the functional integration of $\phi(\mathbf{x})$. The classical solutions $\hat{\phi}(\mathrm{x})$ and $\hat{\lambda}$ are defined by the following two equations:

$$(-1)^{\alpha} l^{2\alpha}\Delta^{\alpha}\hat{\phi} + F' - i\hat{\lambda}e^{-\hat{\phi}} = -Nl^D P_N\,, \quad (4)$$

$$\frac{1}{l^D}\int d^D x\, e^{-\hat{\phi}(\mathbf{x})} = 1\,. \quad (5)$$

If we apply the normalization condition (5) to the integration of the equation (4), we can find that $i\hat{\lambda} = N + \int d^D x\, F'[\hat{\phi}(\mathbf{x})]/l^D$, which determines the mass term in the action (3).

It is straightforward to evaluate the functional integration of $\phi(\mathbf{x})$. We will calculate the leading corrections to the classical solutions. Then, it is enough to restrict ourselves to the quadratic order estimation. Furthermore, we will consider only the leading order effect in the $N \to \infty$ limit. The integration up to quadratic terms is given by the ratio of functional determinants. We can evaluate it in a standard way [4]. As a result, the effective action is found to be

$$S(\hat{\phi}, \hat{\lambda}) + gR\int d^D x\left(\frac{N\hat{Q}}{l^{2\alpha - D}}\right)^{\frac{D}{2\alpha}}. \quad (6)$$

Here, the constant R is defined as $R = 1/(4\pi)^{D/2-1}4D\Gamma(\frac{D}{2})\sin\frac{D}{2\alpha}\pi$.

Now that we have obtained a local correction term, we can determine the unknown function $F(\hat{\phi})$. As is easily seen, a counterterm can be produced only through the following definition of the function $F(\hat{\phi})$:

$$F(\hat{\phi}) = k_0 e^{-\frac{D}{2\alpha}\hat{\phi}}, \quad (7)$$

$$k_0 = k - gRN^{\frac{D}{2\alpha}}. \quad (8)$$

Here, we have introduced a parameter k, which has the observed value at a scale N. Since a bare parameter k_0 does not depend on the scale N, we obtain the renormalization group equation which describes the scaling behavior of the parameter k.

$$\frac{dk}{dN} = g\frac{DR}{2\alpha}\frac{1}{N^{1-D/2\alpha}}\,. \quad (9)$$

As a result, we can find the scaling form of the parameter to be $k \sim gRN^{D/2\alpha}$, which does not depend on the initial value of k.

The scaling behavior is understood from the following point of view. In learning problems, the number of example data N is considered to play a role of setting our observation scale. When we have received a small number of data points, we observe a signal source at a low resolution. The more examples we have received, at the higher resolution we can investigate the source. The signal source with a true probability distribution is a bare quantity which exists independently of our observation at a length scale. The renormalization group equation (9) means that we should scale the parameter k in order to respect the invariance of the true probability distribution under the change of the scale N. Therefore, such a scaling of the parameter k is expected to lead to the optimal performance.

On the other hand, we have the relation $k_0 = k$ if we do not consider the fluctuation effects in a Bayesian setting. This is just the case of maximum likelihood estimation. Again, the invariance of k_0 under the change of N requires that we should not vary the parameter k with the number of examples N. It will be seen later that the scaling of the parameter k plays a crucial role to achieve the optimal learning rate.

An on-line learning algorithm is how to change our hypothesis about an unknown distribution after receiving a new example. It is given by the recursion relation of the expectation value $\langle\phi(\mathbf{x})\rangle$ of the field $\phi(\mathbf{x})$. In our approximation level, the expectation value is just the classical solution $\hat{\phi}(\mathbf{x})$. Therefore, an on-line learning algorithm is directly obtained from the variation of Eq. (4).

$$\Delta\langle\phi_N(\mathbf{x})\rangle \sim \frac{1}{g}\int d^D x'\, G_N(\mathbf{x}, \mathbf{x}') \times\left[-\delta^D(\mathbf{x}' - \mathbf{x}_{N+1}) + \frac{\Delta(i\hat{\lambda})_N}{l^D}e^{-\langle\phi_N(\mathbf{x}')\rangle} + \frac{D}{2\alpha l^D}\frac{dk_N}{dN}e^{-\frac{D}{2\alpha}\langle\phi_N(\mathbf{x}')\rangle}\right]. \quad (10)$$

Hereafter, we will explicitly attach the number of examples N to denote the scale. The learning rate $G_N(\mathbf{x}, \mathbf{x}')$ is a Green's function and defines a local bin size $\xi_N(\mathbf{x})$.

$$G_N(\mathbf{x}, \mathbf{x}') \sim \frac{g}{l^{2\alpha - D}}\frac{(-1)^{\alpha-1}}{(2\pi)^{\frac{D}{2}}\alpha r^{\frac{D}{2}-1}} \times\sum_{n=0}^{\alpha-1}(\gamma_n\xi_N^{-1})^{\frac{D}{2}-2\alpha+1}K_{\frac{D}{2}-1}(\gamma_n\xi_N^{-1}r)\,, \quad (11)$$

$$\xi_N(\mathbf{x}) = l\left[(i\hat{\lambda})_N e^{-\langle\phi_N(\mathbf{x})\rangle} + \left(\frac{D}{2\alpha}\right)^2 k_N e^{-\frac{D}{2\alpha}\langle\phi_N(\mathbf{x})\rangle}\right]^{-\frac{1}{2\alpha}}. \quad (12)$$

Here, r is the distance between the points $\mathbf{x}$ and $\mathbf{x}'$, and the function $K_{\frac{D}{2}-1}$ is a modified Bessel function of the second kind. γ_n is defined to be $\gamma_n \equiv e^{i(2n+1)\frac{\pi}{2\alpha}}e^{-i\frac{\pi}{2}}$. The Green's function is real and regular for any $r \geq 0$.

The first term on the right-hand side of (10) gives the effect of a new datum $\mathbf{x}_{N+1}$, which is necessarily local.

The minus sign of the term ensures the increase of the distribution $\langle Q_N \rangle \simeq e^{-\langle \phi_N \rangle}/l^D$ at the point. The second and third terms show the influence of the changes of the parameters $(i\hat{\lambda})_N$ and k_N, respectively. The Green's function smoothes them in the local length scale $\xi_N(\mathbf{x})$. We can see from Eq. (12) that the length scale should be small only where we find a lot of example data and when the number of examples N is large. Thus, the bin size $\xi_N(\mathbf{x})$ penalizes that we prepare too many degrees of freedom in order to reconstruct a probability distribution from given data.

We note that we have not yet obtained the explicit form of the change $\Delta(i\hat{\lambda})_N$. We report having found that $\Delta(i\hat{\lambda})_N \sim 1 - D^2 k_N/4\alpha^2 N e^{-(1-\frac{D}{2\alpha})\langle \phi_N(\mathbf{x}_{N+1})\rangle}$, which is obtained from the variation of $(i\hat{\lambda})_N = N - Dk_N/2\alpha \cdot \int d^D x\, (\hat{Q}_N/l^{2\alpha-D})^{D/2\alpha}$.

In the rest, we will discuss the average performance of the algorithm. Since we consider the asymptotic situation, we may expand the algorithm around the true probability distribution $Q^* \equiv e^{-\phi^*}/l^D$, which is an ultraviolet fixed point. Then, we define an error $\epsilon_N(\mathbf{x})$ as the difference $\epsilon_N(\mathbf{x}) \equiv \langle \phi_N(\mathbf{x}) \rangle - \phi^*(\mathbf{x})$ and will evaluate the dynamics of the error.

We have expanded the Green's function in the convolution of Eq. (10) around the fixed point and found that it decays as $1/N$, which is known to be optimal in the on-line learning by neural networks [5]. Putting $\langle \phi_N \rangle = \phi^* + \epsilon_N$ into the expansion, we can obtain the evolution equation of the error ϵ_N. However, it requires care to be solved, since it has the N dependence which is not extracted explicitly. The equation depends on a new datum $\mathbf{x}_{N+1}$ through the terms $\delta^D(\mathbf{x} - \mathbf{x}_{N+1})$ and $\Delta(i\hat{\lambda})_N$. The data dependence is averaged over many steps of evolutions, which is found from the following relations [1]:

$$\sum_{i=1}^{N} \delta^D(\mathbf{x} - \mathbf{x}_{i+1}) = NQ^*(\mathbf{x}) + \sqrt{N}\,\rho(\mathbf{x}), \quad (13)$$

$$\langle\!\langle \rho(\mathbf{x})\rho(\mathbf{x}') \rangle\!\rangle = Q^*(\mathbf{x})\delta^D(\mathbf{x} - \mathbf{x}'). \quad (14)$$

Here, the function $\rho(\mathbf{x})$ is a fluctuating density. Equation (13) shows that we have to separate the data-dependent terms into a leading order and a subleading one. Therefore, we have to adopt the average $\int d^D x_{N+1}\, Q^*(\mathbf{x}_{N+1})\Delta(i\hat{\lambda})_N$ as the leading order of $\Delta(i\hat{\lambda})_N$, and $\Delta(i\hat{\lambda})_N - \int d^D x_{N+1}\, Q^*(\mathbf{x}_{N+1})\Delta(i\hat{\lambda})_N$ as the subleading one. Then, the leading order term $\epsilon_N^0(\mathbf{x})$ of the large N expansion of the error $\epsilon_N(\mathbf{x}) = \epsilon_N^0(\mathbf{x}) + \epsilon_N^1(\mathbf{x}) + \ldots$ obeys the following equation:

$$N\epsilon_{N+1}^0(\mathbf{x}) - (N-1)\epsilon_N^0(\mathbf{x}) = 1 - (Q^*)^{-1}\delta^D(\mathbf{x} - \mathbf{x}_{N+1}) + g\,\frac{D^2 R}{4\alpha^2 N^{1-\frac{D}{2\alpha}}}\left[\frac{1}{(l^D Q^*)^{1-\frac{D}{2\alpha}}} - \int d^D x\, \frac{Q^*}{(l^D Q^*)^{1-\frac{D}{2\alpha}}}\right]. \quad (15)$$

Summed up with N changed 1 to N, Eq. (15) gives the following result:

$$\epsilon_N^0(\mathbf{x}) = -\frac{1}{\sqrt{N}}\, Q^*(\mathbf{x})^{-1}\rho(\mathbf{x}) + g\,\frac{DR}{2\alpha N^{1-\frac{D}{2\alpha}}} \times \left[\frac{1}{(l^D Q^*)^{1-\frac{D}{2\alpha}}} - \int d^D x\, \frac{Q^*}{(l^D Q^*)^{1-\frac{D}{2\alpha}}}\right]. \quad (16)$$

The subleading term $\epsilon_N^1(\mathbf{x})$ is found to be $O(\log N/N)[\alpha = D]$ or $O(N^{\frac{D}{2\alpha}-\frac{3}{2}})[\alpha \neq D]$. However, we can prove that it does vanish if and only if $\alpha = D$. Then, Eq. (16) is exact up to subleading order and leads to the optimal error decay (17). This means that we can infer distributions most effectively, when we choose their candidates among C^{2D}-class functions in D dimensions.

From Eq. (16), we can show that the average of the squared error $\epsilon_N(\mathbf{x})^2$ gives a universal result. This is consistent with the definition of the reparametrization invariant distance between two distributions $p(\mathbf{x}, \xi)$ and $p(\mathbf{x}, \xi + d\xi)$, which are specified by parameters ξ^i [6]: $ds^2 = g_{ij} d\xi^i d\xi^j = E[\frac{\partial}{\partial \xi^i}\log p \frac{\partial}{\partial \xi^j}\log p] d\xi^i d\xi^j = E[(d\log p)^2]$.

We have found the quadratic error as

$$\left\langle\!\!\left\langle \int d^D x\, Q^* \epsilon_N^2 \right\rangle\!\!\right\rangle = \frac{V_{\mathbf{p}} V_{\mathbf{x}}}{(2\pi)^D}\,\frac{1}{N}. \quad (17)$$

Here, the coefficients $V_{\mathbf{p}}$ and $V_{\mathbf{x}}$ denote the volumes of momentum space and of coordinate one, respectively. The constant $V_{\mathbf{p}} V_{\mathbf{x}}$ gives the number of all the possible configurations of the system. Since it is independent of how to describe the system, we have reproduced the universal asymptotic behavior, which is well known in neural network models [7]. Thus, we have obtained the $1/N$ decay of the quadratic error up to subleading order, which is considered to be optimal as in the case of neural networks [5]. We note that the optimal behavior (17) can be obtained, if and only if we develop the parameter k_N according to the renormalization group equation (9).

Figure 1 shows the result of the numerical simulation of the algorithm (10), which is applied to the learning of a two-dimensional Gaussian distribution. It is obvious that the quadratic error is asymptotically linear to the inverse of the number of examples $1/N$. This ensures the analytical result (17).

Finally, we point out the relation to previous works and make concluding remarks. Bialek *et al.* [1] gave an excellent field theoretical formulation of the batch learning of probability distributions. They controlled the number of degrees of freedom through the scaling of an infrared cutoff parameter l. Especially, they obtained the optimal scaling of a bin size $\xi_N \propto N^{-1/(2\alpha+D)}$, which includes the well-known result $\xi_N \propto N^{-1/3}$ in one-dimensional space $D = \alpha = 1$ [8]. On the other hand,

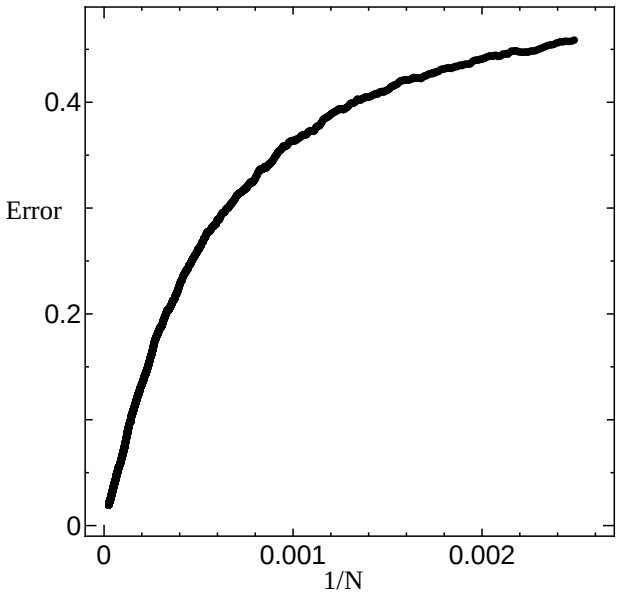

FIG. 1. The quadratic error (17) in learning the two-dimensional Gaussian distribution $Q^*(\mathbf{x}) = e^{-(x^2+y^2)/2}/2\pi[-1/2 \le x, y \le 1/2]$. We have performed the algorithm (10) with the area divided into 40×40 pieces.

we have analyzed the on-line learning of probability distributions. In our case, a renormalization group enables us to control a bin size (12) and leads to the optimal change of the hypothesis (10) through the scaling of the parameter k_N. We can find, from $(i\hat{\lambda})_N \sim N$ and $\alpha = D$, the optimal scaling of the bin size $\xi_N \propto N^{-1/2D}$ in the on-line learning scheme.

Thus, a renormalization group gives an on-line learning algorithm in a natural way, which we have proved to be optimal. Our discussion is so general that we may adopt it as a principle to derive optimal on-line learning algorithms. Furthermore, we could expect that the optimal error decay can also be achieved up to a desired order, if we extend the field theoretical analysis to the order.

The author thanks H. Nishimori and Y. Kitazawa for their valuable comments and discussions. He also thanks J. Inoue for several useful comments. This work was partially supported by the Grant-in-Aid for Scientific Research No. 10750056 from the Ministry of Education, Science and Culture, Japan.

[1] W. Bialek, C. G. Callan, and S. P. Strong, Phys. Rev. Lett. **77**, 4693 (1996).
[2] See, for example, N. Murata, S. Yoshizawa, and S. Amari, IEEE Trans. Neural Netw. **5**, 865 (1994).
[3] N. Barkai, H. S. Seung, and H. Sompolinsky, Phys. Rev. Lett. **75**, 1415 (1995); J. W. Kim and H. Sompolinsky, Phys. Rev. Lett. **76**, 3021 (1996); M. Biehl, P. Riegler, and M. Stechert, Phys. Rev. E **52**, 4624 (1995).
[4] S. Coleman, *Aspects of Symmetry* (Cambridge University Press, Cambridge, England, 1975), p. 340.
[5] M. Opper, Phys. Rev. Lett. **77**, 4671 (1996).
[6] S. Amari, *Differential-Geometrical Methods in Statistics* (Springer, Berlin, 1985).
[7] H. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. A **45**, 6056 (1992); S. Amari and N. Murata, Neural Comput. **5**, 140 (1993); M. Opper and D. Haussler, Phys. Rev. Lett. **75**, 3772 (1995).
[8] See, for example, A. Barron and T. Cover, IEEE Trans. Inf. Theory **37**, 1034 (1991).