PHYSICAL REVIEW
LETTERS

VOLUME 77 2 DECEMBER 1996 NUMBER 23

Field Theories for Learning Probability Distributions

William Bialek,[1] Curtis G. Callan,[2] and Steven P. Strong[1]
[1]*NEC Research Institute, 4 Independence Way, Princeton, New Jersey 08540*
[2]*Department of Physics, Princeton University, Princeton, New Jersey 08544*
(Received 25 July 1996)

Imagine being shown N samples of random variables drawn independently from the same distribution. What can you say about the distribution? In general, of course, the answer is nothing, unless you have some prior notions about what to expect. From a Bayesian point of view one needs an *a priori* distribution on the space of possible probability distributions, which defines a scalar field theory. In one dimension, free field theory with a normalization constraint provides a tractable formulation of the problem, and we discuss generalizations to higher dimensions. [S0031-9007(96)01804-2]

PACS numbers: 02.50.Fz, 02.50.Ey, 03.70.+k, 87.10.+e

As we watch the successive flips of a coin (or the meanderings of stock prices), we ask ourselves if what we see is consistent with the conventional probabilistic model of a fair coin. More quantitatively, we might try to fit the data with a definite model that, as we vary parameters, includes the fair coin and a range of possible biases. The estimation of these underlying parameters is the classical problem of statistical inference or "inverse probability," and has its origins in the foundations of probability theory itself [1]. But when we observe continuous variables, the relevant probability distributions are functions, not finite lists of numbers as in the classical examples of flipping coins or rolling dice. In what sense can we infer these functions from a finite set of examples? In particular, how do we avoid the solipsistic inference in which each data point we have observed is interpreted as the location of a narrow peak in the underlying distribution?

Let the variable of interest be x with probability distribution $Q(x)$; we start with the one dimensional case. We are given a set of points $x_1, x_2, \ldots, x_N$ that is drawn independently from $Q(x)$, and are asked to estimate $Q(x)$ itself. One approach is to assume that all possible $Q(x)$ are drawn from a space parametrized by a finite set of coordinates, implicitly excluding distributions that have many sharp features. In this case, it is clear that the number of examples N can eventually overwhelm the number of parameters K [2]. Although the finite dimensional case is often of practical interest, one would like a formulation faithful to the original problem of estimating a function rather than a limited number of parameters.

No finite number of examples will determine uniquely the whole function $Q(x)$, so we require a probabilistic description. Using Bayes' rule, we can write the probability of the function $Q(x)$ given the data as

$$P[Q(x) \mid x_1, x_2, \ldots, x_N] = \frac{P[x_1, x_2, \ldots, x_N \mid Q(x)]P[Q(x)]}{P(x_1, x_2, \ldots, x_N)} \quad (1)$$

$$= \frac{Q(x_1)Q(x_2)\cdots Q(x_N)P[Q(x)]}{\int [dQ(x)]Q(x_1)Q(x_2)\cdots Q(x_N)P[Q(x)]}, \quad (2)$$

where we make use of the fact that each x_i is chosen independently from the distribution $Q(x)$, and $P[Q(x)]$ summarizes our *a priori* hypotheses about the form of $Q(x)$. If asked for an explicit estimate of $Q(x)$, one might try to optimize the estimate so that the mean-square deviation from the correct answer is, at each point x, as small as possible. This optimal least-square estimator $Q_{\rm est}(x; \{x_i\})$ is the average of $Q(x)$ in the conditional distribution of Eq. (2), which can be written as

$$Q_{\rm est}(x; \{x_i\}) = \frac{\langle Q(x)Q(x_1)Q(x_2)\cdots Q(x_N)\rangle^{(0)}}{\langle Q(x_1)Q(x_2)\cdots Q(x_N)\rangle^{(0)}}, \quad (3)$$

where by $\langle\cdots\rangle^{(0)}$ we mean expectation values with respect to the *a priori* distribution $P[Q(x)]$. The prior distribution

$P[Q(x)]$ is a scalar field theory, and the n-point functions of this theory are precisely the objects that determine our inferences from the data.

The restriction of the distribution $Q(x)$ to a finite dimensional space represents, in the field theoretic language, a sharp ultraviolet cutoff scheme. Several authors have considered the problem of choosing among distributions with different numbers of parameters, which corresponds to assuming that the true theory, $P[Q(x)]$, has a hard ultraviolet cutoff whose unknown location is to be set by this choice. As in field theory itself, one would like to have a theory in which the observable quantities—like your best estimate $Q_{\rm est}(x;\{x_i\})$—are *independent* of the cutoff as the cutoff goes to infinity. Our Bayesian approach will provide this.

The prior distribution, $P[Q(x)]$, should capture our prejudice that the distribution $Q(x)$ is smooth, so $P[Q(x)]$ must penalize large gradients, as in conventional field theories. To have a field variable $\phi(x)$ that takes on a full range of real values $(-\infty < \phi < \infty)$, we write

$$Q(x) = \frac{1}{\ell_0}\,\exp[-\phi(x)], \tag{4}$$

where ℓ_0 is an arbitrary length scale. Then we take ϕ to be a free scalar field with a constraint to enforce normalization of $Q(x)$. Thus $\phi(x)$ is chosen from a probability distribution

$$P_\ell[\phi(x)] = \frac{1}{Z}\exp\left[-\frac{\ell}{2}\int dx(\partial_x\phi)^2\right] \times \delta\left[1 - \frac{1}{\ell_0}\int dx\, e^{-\phi(x)}\right], \tag{5}$$

where Z is a normalization constant and ℓ is a length scale that defines the hypothesized smoothness of the distribution $Q(x)$. We write $P_\ell[\phi(x)]$ to remind us that we have chosen a particular value for ℓ, and we will later consider averaging over a distribution of ℓ's, $P(\ell)$. The objects of interest are the correlation functions

$$\langle Q(x_1)Q(x_2)\cdots Q(x_N)\rangle^{(0)} = \int D\phi\, P_\ell[\phi(x)]\prod_{i=1}^{N}\frac{1}{\ell_0}\exp[-\phi(x_i)] \tag{6}$$

$$= \frac{1}{\ell_0^N}\,\frac{1}{Z}\int\frac{d\lambda}{2\pi}\int D\phi\,\exp[-S(\phi;\lambda)], \tag{7}$$

where, by introducing the Fourier representation of the delta function, we define the action

$$S(\phi;\lambda) = \frac{\ell}{2}\int dx(\partial_x\phi)^2 + i\frac{\lambda}{\ell_0}\int dx\, e^{-\phi(x)} + \sum_{i=1}^{N}\phi(x_i) - i\lambda\,. \tag{8}$$

We evaluate the functional integral in Eq. (7) in a semiclassical approximation, which becomes accurate as N becomes large. Keeping only the configuration that extremizes the action—the pure classical approximation, with no fluctuations—is equivalent to maximum likelihood estimation, which chooses the distribution, $Q(x)$, that maximizes $P[Q(x)\,|\,\{x_i\}]$. In our case, integration over fluctuations will play a crucial role in setting the proper value of the scale ℓ.

The classical equations of motion for ϕ and λ are

$$\frac{\delta S(\phi;\lambda)}{\delta\phi(x)} = \frac{\partial S(\phi;\lambda)}{\partial\lambda} = 0\,, \tag{9}$$

which imply

$$\ell\partial_x^2\phi_{\rm cl}(x) + i\frac{\lambda_{\rm cl}}{\ell_0}\,e^{-\phi_{\rm cl}(x)} = \sum_{i=1}^{N}\delta(x - x_i)\,, \tag{10}$$

$$\frac{1}{\ell_0}\int dx\, e^{-\phi_{\rm cl}(x)} = 1\,. \tag{11}$$

Integrating Eq. (10) and comparing with Eq. (11), we find that $i\lambda_{\rm cl} = N$, provided that $\partial\phi(x)$ vanishes as $|x| \to \infty$ [3]; as is often the case, the steepest descent approximation to the integral requires us to deform the integral into the complex λ plane. If the points $\{x_i\}$ are actually chosen from a distribution $P(x)$, then, as $N \to \infty$, we hope that $\phi_{\rm cl}(x)$ will converge to $-\ln[\ell_0 P(x)]$. This would guarantee that our average over all possible distributions $Q(x)$ is dominated by configurations $Q_{\rm cl}(x)$ that approximate the true distribution. So we write $\phi_{\rm cl}(x) = -\ln[\ell_0 P(x)] + \psi(x)$ and expand Eq. (10) to first order in $\psi(x)$. In addition we notice that the sum of delta functions can be written as

$$\sum_{i=1}^{N}\delta(x - x_i) = NP(x) + \sqrt{N}\,\rho(x)\,, \tag{12}$$

where $\rho(x)$ is a fluctuating density such that

$$\langle\rho(x)\rho(x')\rangle = P(x)\delta(x - x')\,. \tag{13}$$

The (hopefully) small field $\psi(x)$ obeys the equation

$$[\ell\partial_x^2 - NP(x)]\psi(x) = \sqrt{N}\,\rho(x) + \ell\partial_x^2\ln P(x)\,, \tag{14}$$

which we can solve by WKB methods because of the large factor N:

$$\psi(x) = \int dx'\,K(x,x')\,[\sqrt{N}\,\rho(x') + \ell\partial_x^2\ln P(x')]\,, \tag{15}$$

$$K(x,x') \sim \frac{1}{2\sqrt{N}}\,[\ell^2 P(x)P(x')]^{-1/4} \times \exp\left[-\int_{\min(x,x')}^{\max(x,x')} dy\sqrt{\frac{NP(y)}{\ell}}\right]. \tag{16}$$

Thus the "errors" $\psi(x)$ in our estimate of the distribution involve an average of the fluctuating density over a region of (local) size $\xi \sim [\ell/NP(x)]^{1/2}$. The average systematic error and the mean-square random error are easily computed in the limit $N \to \infty$ because this length scale becomes small. We find

$$\langle \psi(x) \rangle = \frac{\ell}{NP(x)}\, \partial_x^2 \ln P(x) + \cdots , \tag{17}$$

$$\langle [\delta\psi(x)]^2 \rangle = \frac{1}{4}\, \frac{1}{\sqrt{NP(x)\ell}} + \cdots , \tag{18}$$

justifying our claim that the classical solution converges to the correct distribution.

The complete semiclassical result is

$$\langle Q(x_1)Q(x_2)\cdots Q(x_N)\rangle^{(0)} \approx \frac{1}{\ell^N}\, R \exp[-S(\phi_{\rm cl}; \lambda = -iN)], \tag{19}$$

where R is the ratio of determinants,

$$R = \left[\frac{\det[-\ell\partial_x^2 + NQ_{\rm cl}(x)]}{\det(-\ell\partial_x^2)} \right]^{-1/2}. \tag{20}$$

This has to be computed a bit carefully—there is no restoring force for fluctuations λ, but these can be removed by fixing the spatially uniform component of $\phi(x)$, which enforces normalization of $Q(x)$ [4]. Then the computation of the determinants is standard [5], and we find

$$R = \exp\left[-\frac{1}{2} \left(\frac{N}{\ell} \right)^{1/2} \int dx \sqrt{Q_{\rm cl}(x)} \right], \tag{21}$$

where as before we use the limit $N \to \infty$ to simplify the result [6]. It is interesting to note that R can also be written as $\exp[-(1/2)\int dx\, \xi^{-1}]$, so the fluctuation contribution to the effective action counts the number of independent "bins" (of size $\sim\xi$) that describe $Q(x)$.

Putting the factors together, we find that

$$\langle Q(x_1)Q(x_2)\cdots Q(x_N)\rangle^{(0)} \approx \prod_{i=1}^{N} P(x_i) \exp[-F(x_1, x_2, \ldots, x_N)], \tag{22}$$

where the correction term F is given by

$$F(\{x_i\}) = \frac{1}{2} \left(\frac{N}{\ell} \right)^{1/2} \int dx\, P^{1/2}(x) e^{-\psi(x)/2} + \frac{\ell}{2} \int dx (\partial_x \ln P - \partial_x \psi)^2 + \sum_{i=1}^{N} \psi(x_i) . \tag{23}$$

The crucial point, which can be verified from the explicit solution in Eq. (16), is that $F(\{x_i\})$ is finite, even when multiple points x_i approach each other. Hence our estimate of the probability distribution from Eq. (3) is finite even when we ask about $Q(x)$ at the points where we have been given examples: we are in one dimension where ultraviolet divergences are not a problem.

Although our theory is finite in the ultraviolet, we do have an arbitrary length scale ℓ. This means that we define, *a priori*, a scale on which variations of the probability $Q(x)$ are viewed as "too fast." One would rather let all scales in our estimate of the distribution $Q(x)$ emerge from the data points themselves. We can restore scale invariance (perhaps scale indifference is a better term here) by viewing ℓ itself as a parameter that needs to be determined. Thus, as a last step in evaluating the functional integral, we should integrate over ℓ, weighted by some prior distribution, $P(\ell)$, for values of this parameter. The hope is that this integral will be dominated by some scale, ℓ_*, that is determined primarily by the structure of $Q(x)$ itself, at least in the large N limit. As long as our *a priori* knowledge about ℓ can be summarized by a reasonably smooth distribution, then, at large N, ℓ_* must be the minimum of F, since this is the only place where ℓ appears with coefficients that grow as powers of N. To see how this works we compute the average value of F and minimize with respect to ℓ. Up to constant factors, this amounts to balancing the ℓ dependence of the kinetic energy against that of the fluctuation determinant. The result is

$$\ell_* \propto N^{1/3} \left[\frac{\int dx\, P^{1/2}}{\int dx (\partial_x \ln P)^2} \right]^{2/3}. \tag{24}$$

Strictly, one should use a particular value of F and not its average, but fluctuations are of lower order in N and do not change the qualitative result $\ell_* \propto N^{1/3}$.

If the fluctuation effects were ignored, as in maximum likelihood estimation, ℓ would be driven to zero and we would be overly sensitive to the details of the data points. This parallels the discussion of "Occam factors" in the finite dimensional case, where the phase space factors from integration over the parameters $\{g_\mu\}$ serve to discriminate against models with larger numbers of parameters [2]. It is not clear from the discussion of finite dimensional models, however, whether these factors are sufficiently powerful to reject models with an infinite number of parameters. Here we see that, even in an infinite dimensional setting, the fluctuation terms are sufficient to control the estimation problem and select a model with finite, N-dependent, complexity.

Because we are trying to estimate a function, rather than a finite number of parameters, we must allow ourselves to give a more and more detailed description of the function $Q(x)$ as we see more examples; this is quantified by the scale ξ_* on which the estimated distribution is forced to be smooth. With the selection of the optimal ℓ from Eq. (24), we see that $\xi_* \propto (\ell_*/N)^{1/2} \propto N^{-1/3}$. The classical solution converges to the correct answer

with a systematic error, from Eq. (17), that vanishes as $\langle\psi\rangle \propto N^{-2/3}$, while the random errors have a variance [Eq. (18)] that vanishes with the same power of N. We can understand this result by noting that in a region of size ξ_* there are, on average, $N_{\rm ex} \sim NP(x)\xi_*$ examples, which scales as $N_{\rm ex} \propto N^{2/3}$; the random errors then have a standard deviation $\delta\psi_{\rm rms} \sim 1/\sqrt{N_{\rm ex}}$ [7].

What happens in higher dimensions? If we keep the free field theory then we will have problems with ultraviolet divergences in the various correlation functions of the field $\phi(x)$. Because $Q(x) = (1/\ell)\exp[-\phi(x)]$, ultraviolet divergences in ϕ mean that we cannot define a normalizable distribution for the value of Q at a single point in the continuum limit. In terms of information theory [8], if functions $Q(x)$ are drawn from a distribution functional with ultraviolet divergences, then even specifying the function $Q(x)$ to finite precision requires an infinite amount of information.

As an alternative, we can consider higher derivative actions in higher dimensions. All the calculations are analogous to those summarized above, so here we list only the results. If we write, in D dimensions, $Q(x) = (1/\ell_0^D)\exp[-\phi(x)]$, and choose a prior distribution

$$P[\phi(x)] = \frac{1}{Z}\exp\left[-\frac{\ell^{2\alpha-D}}{2}\int d^Dx(\partial_x^\alpha\phi)^2\right] \times \delta\left[1 - \frac{1}{\ell_0^D}\int d^Dx\, e^{-\phi(x)}\right]; \quad (25)$$

then to insure finiteness in the ultraviolet we must have $2\alpha > D$. The saddle point equations lead to a distribution that smooths the examples on a scale $\xi \sim (\ell^{2\alpha-D}/NQ)^{1/2\alpha}$, and the fluctuation determinant makes a contribution to the action $\propto \int d^Dx[NQ(x)/\ell^{2\alpha-D}]^{D/2\alpha}$. Again we find the optimal value of ℓ as a compromise between this term and the kinetic energy, resulting in $\ell_* \propto N^{D/(4\alpha^2-D^2)}$. Then the optimal value of ξ becomes $\xi_* \propto N^{-1/(2\alpha+D)}$, so that the estimated distribution is smooth in volumes of dimension ξ_*^D that contain $N_{\rm ex} \sim NQ\xi^D \sim N^{2\alpha/(2\alpha+D)}$ examples. Then the statistical errors in the estimate will behave as

$$\delta\psi_{\rm rms} \propto \delta Q/Q \sim N_{\rm ex}^{-1/2} \sim N^{-\mu}, \quad (26)$$

with the "error exponent" $\mu = \alpha/(2\alpha + D)$. Note that since $2\alpha > D$, the exponent $1/4 < \mu < 1/2$. The most rapid convergence, $\mu = 1/2$, occurs if $Q(x)$ is drawn from a family of arbitrarily smooth ($\alpha \to \infty$) distributions, so we can choose fixed, small bins in which to accumulate the samples, leading to the naive $1/\sqrt{N}$ counting statistics. If we assume that our prior distribution functional is local, then α must be an integer and we can have $\mu \to 1/4$ only as $D \to \infty$, so that the slowest possible convergence occurs in infinite dimension.

The fact that higher dimensional functions are more difficult to learn is often called the "curse of dimensionality." We see that this is not just a quantitative problem—unless we hypothesize that higher dimensional functions are drawn from ensembles with proportionately higher order notions of smoothness, one would require an infinite amount of information to specify the function at finite precision. Once we adopt these more stringent smoothness hypotheses, however, the worst that happens is a reduction in the error exponent μ by a factor of 2.

Is there a more general motivation for the choice of action in Eq. (25)? First, this action gives the maximum entropy distribution consistent with a fixed value of $\int d^Dx(\partial_x^\alpha\phi)^2$, and by integrating over ℓ we integrate over these fixed values. Thus our action is equivalent to the rather generic assumption that probability distributions are drawn from an ensemble in which this "kinetic energy" is finite. Second, addition of a constant to $\phi(x)$ can be absorbed in a redefinition of the arbitrary ℓ_0, so it makes sense to insist on $\phi \to \phi + {\rm const}$ as a symmetry. Finally, addition of other terms to the action cannot change the asymptotic behavior at large N unless these terms are relevant operators in the ultraviolet. Thus many different priors $P[Q(x)]$ will exhibit the same convergence properties, indexed by a single exponent $\mu(\alpha)$.

We thank V. Balasubramanian, R. Koberle, S. Omohundro, and W. Smith for helpful discussions and for comments on the manuscript.

[1] A nice account of the history is given by J. Earman, *Bayes or Bust?: A Critical Examination of Bayesian Confirmation Theory* (MIT Press, Cambridge, 1992).

[2] There is much literature on the finite dimensional problem. The discussion closest to our own is probably that of A. R. Barron and C.-H. Sheu, Ann. Stat. **19**, 1347 (1991). Even in the finite case it is crucial to have penalties for models which might provide a better fit to the data but should be rejected as overly complex; see J. Rissanen, IEEE Trans. Inf. Theory **30**, 629 (1984); Ann. Stat. **14**, 1080 (1986); *Stochastic Complexity and Statistical Inquiry* (World Scientific, Singapore, 1989); A. R. Barron and T. M. Cover, IEEE Trans. Inf. Theory **37**, 1034 (1991). For the more general mathematical problem of quantifying complexity, see M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications* (Springer-Verlag, New York, 1993). Neural networks provide an explicit framework for many different learning problems, and there are connections between the learning rules for these networks, Bayesian approaches, and complexity penalties; see D. J. C. MacKay, Neural Comput. **4**, 415 (1992); **4**, 448 (1992). These ideas have been translated into physicists' language and extended by V. Balasubramanian, Report No. cond-mat/9601030 [Neural Comput. (to be published)].

[3] Direct integration of Eq. (10) shows that, for x outside the region sampled by the $\{x_i\}$, $Q_{\rm cl}(|x| > \max|x_i|) = 2\ell/Nx^2$, so that $\partial\phi(x)$ does vanish at infinity. Note also that the total weight assigned to the unobserved region vanishes as N increases. This explicit integration can be

continued into the sampled region, reducing the problem to quadratures. At the end, however, one has to solve the matching conditions at each point x_i, and this leads to complicated expressions. Since we are interested in showing that the solution converges to the correct answer, we find it more efficient to study perturbation theory around this answer and see if it is consistent.

[4] We keep track of terms that contribute positive powers of N in the effective action; the zero mode that we need to exclude gives at most $\ln N$.

[5] S. Coleman, *Aspects of Symmetry* (Cambridge University Press, Cambridge, 1975). This is the one dimensional case corresponding to ordinary quantum mechanics, discussed in the first appendix to Coleman's Chap. 7. There is a further simplification because we can use a WKB approximation to evaluate the relevant wave functions.

[6] With the asymptotic behavior of $Q_{\rm cl}(x)$ noted in [3], the WKB integral does not converge. More precisely the WKB approximation breaks down far outside the region where we have observed the samples $\{x_i\}$. But the total weight in this region declines with N, so proper treatment of these terms generates subleading terms in the effective action.

[7] This scaling behavior is also found by Barron and Sheu in Ref. [2]. They studied the learning of probability distributions for which $\phi(x)$ lives in an m-dimensional linear space, and then they choose m to optimize convergence. The exponents derived here correspond to the case where $\phi(x)$ is continuous and has a first derivative that is square integrable, which has a clear connection to our choice of prior. W. Smith has emphasized to us that this scaling can be derived most simply for piecewise constant approximations to $Q(x)$—conventional binning procedures, with the bin size chosen to balance systematic and statistical errors. In the field theoretic language binning is a lattice cutoff. We find it attractive that these results can be recovered in a Bayesian formulation that has no explicit cutoff.

[8] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1949).