On impossibility of learning in a reparameterization covariant way

Timothy Holy[1] **and Ilya Nemenman**[2]
[1]Department of Anatomy and Neurobiology
Washington University Medical School
660 S. Euclid Avenue, St. Louis, MO 63110
holy@pcg.wustl.edu
[2]Kavli Institute for Theoretical Physics
University of California, Santa Barbara, CA 93106
nemenman@kitp.ucsb.edu

The problem of inferring a probability density $Q(x)$ from a finite number N of observations has been discussed in many contexts. Arguably, the most interesting question here is the inference of the density when only some assumptions about its smoothness properties can be made. This has been analyzed in the framework of Bayesian statistics using functional analysis methods of Quantum Field Theory [BCS96, Hol97, NB02]. If $Q(x) = 1/l_0 \exp\phi(x)$ [BCS96, NB02], or $Q(x) = \phi^2(x)$ [Hol97], then the methods prescribe priors of the form

$$\mathcal{P}[\phi(x)] = \frac{1}{Z} \exp\left\{-\frac{\ell^{2\eta-1}}{2}\int dx \left(\frac{\partial^\eta \phi}{\partial x^\eta}\right)^2\right\} \delta\left[\int dx\, Q(\phi(x)) - 1\right] . \qquad (1)$$

Here Z is the normalization constant and δ-function enforces the normalization of the density Q. Further, the exponential term punishes for rapid variability (non–smoothness) of ϕ and, therefore, of Q. Larger values of η and ℓ correspond to smoother assumptions about Q, and Ref. [NB02] suggested a way of self-consistently estimating the correct values of them from the data.

A few issues still remain [NB02], but, conceptually, the problem seems to be well understood. The biggest hurdle was observed in Ref. [Per97]. If $Q(x)$ is a probability density, then $Q(x)dx$ is a measure invariant under reparameterizations. Thus under a coordinate transformation $z = z(x)$, Q must transform as follows:

$$Q(z) = Q(x(z)) \left|\frac{dx}{dz}\right| \qquad (2)$$

Unfortunately, since the prior, Eq. (1), is not reparameterization–invariant, this transformation condition does not hold. Refs. [Per97, Per99] tried to correct for this (there were similar attempts for non–Bayesian frameworks as well). The aim of the current note is to present a series of related arguments that show that reparameterization–covariant inference is impossible in principle. We first start with a general discussion, then give an explicit and unexpectedly simple example, and finally show how approximate reparameterization covariance is still possible.

First note that, for any $K = \dim x$, we can write the probability density as $Q(x) = \sqrt{|g(x)|}\,\widetilde{Q}(x)$, where $|g|$ is the determinant of the metric tensor. $\sqrt{|g|}\,d^K x$

forms an invariant volume element, thus $\sqrt{|g|}$ transforms like a scalar density similarly to Q. Therefore, $\widetilde{Q}$ is a pure scalar. One can, of course, write a reparameterization–invariant regularizing prior for the scalar portion of the probability density using the metric tensor [Per97, Per99]. However, what is to be used to enforce the smoothness of the metric itself in a metric–independent way? Indeed, in one dimension all differential–geometric properties are determined by embedding (equivalently, parameterization), and there is no intrinsic curvature commonly used to identify more complex solutions prone to overfitting [Vap98, BNT01]. However, the metric part of the p. d. f. enters multiplicatively and must be regularized for successful learning. Therefore, at least in one dimension, existence of a reparameterization–covariant and at the same time regular inference mechanism is unlikely. For higher dimensions, there is, for example, intrinsic curvature, but it is doubtful that it will be enough to properly regularize allowable metric tensors.

This intuition can be made very precise. Let us describe learning with an operator L that maps observed data, $\{x_i\}$, $i = 1 \dots N$, onto probability densities $Q(x)$. Further, we introduce a reparameterization operator R_z, which acts as follows: $R_z x = z(x)$, $R_z Q(x) = Q(x(z))J(z)$, where $Q(x)$ is non–singular, and $J^{-1}(z) = |dx/dz|$ is the Jacobian of the reparameterization. Then reparameterization covariance means that

$$[R_z, L] = 0\,. \tag{3}$$

That is, reparameterization commutes with learning. Suppose now that we chose a particular reparameterization $z = z(x)$ that keeps the data unchanged $R_z x_i = z_i \equiv x_i$. Then we get $LR_z\{x_i\} = L\{x_i\} \equiv Q(x)$. On the other hand, $RL\,\{x_i\} = R\,Q(x) = J(z)Q(z)$. Therefore, for such reparameterizations,

$$[R, L] = (J - 1)L\,. \tag{4}$$

This is zero for the trivial reparameterization, $z = x$, but nonzero for any other reparameterization and continuous $Q(x) = L\{x_i\}$. The only fully reparameterization–covariant solution is a singular, unregularized learning machine, $L\{x_i\} = 1/N \sum \delta(x - x_i)$, but this overfits hopelessly [Vap98].

The problem is, of course, equivalent to the fact that there are infinitely many ways to reparameterize any $\{x_i\}$ into equally spaced $\{z_i\}$. Thus without some a priori constraints on particular coordinates used, the data are completely uninformative.

One should not assume that probability densities are special in having this problem: any quantity that transforms non–trivially will have its own analogue of Eq. (4). The need for knowing a parameterization has been noticed many times in the previous literature, but usually set aside as a technical issue. For example, Cucker and Smale [CS01] note that the total error of learning consists of the sample and the approximation errors. The latter is finite only under a condition that the true (unknown) measure on x is absolutely continuous with respect to an assumed one, e. g., J that maps one measure onto the other is bounded. This is a manifestly parameterization–dependent assumption.

While Ref. [CS01] deals specifically with quadratic regressions, it is clear that this particular part of their results generalizes easily. Indeed, learning is usually represented as minimizing the risk $\mathcal{R}$, which is the expected value of some loss $\mathcal{L}$,

$$\mathcal{R} = \int dx Q(x) \mathcal{L}(Q, x)\,, \tag{5}$$

If $Q(x)$ is allowed to be very small, then there always can be an interval of x that has an (infinitely) large risk, but a vanishingly small measure, so that it is unobservable for any $N < \infty$. Without constraining possible densities, one cannot guarantee asymptotic consistency in the sense developed in Ref. [Vap98].

Now when we have firmly established that it is impossible to learn in a covariant fashion, we should investigate how badly non-covariant approaches perform. Is it possible to estimate mistakes made by the learning process? If $\phi(x)$ is the quantity being estimated (such as $\phi(x) = \log Q(x)/l_0$ in [BCS96]), and its (unknown) true value is $\phi_0(x)$, then we can define $\psi(x) = \phi(x) - \phi_0(x)$ as the error of the estimator. For most learning scenarios (see, for example, [BCS96])

$$\operatorname{Var} \psi(x) \propto \frac{1}{N^\alpha P^\beta(x)}\,, \tag{6}$$

where $P(x)$ is the (unknown) value of the true probability density, and α, β depend on the assumptions made about the smoothness of ϕ, $0 < (\alpha, \beta) < 1$. In particular, in Ref. [BCS96], $\alpha = \beta = 1/2$, and other possibilities were discussed in Ref. [NB02]. Note that Eq. (6) does not bound the error variance uniformly. Even worse, a simple reparameterization (stretching of x) can make P vanishingly small everywhere, and then the variance diverges. This is clearly because the smoothness properties, Eq. (1), are imposed onto Q is some coordinates x. However, there is a close relation between the coordinates and the densities [cf. Eq. (2)], and choosing the coordinates limits the set of probability distributions that can be learned reliably.

To see this notice that $1/P(x)$ is the Radon–Nikodym derivative of the measure uniform on x, $U(x)$, with respect to the true, unknown measure, $\pi(x)$. If $P(x) \geq P_0 > 0$ [equivalently, $U(x)$ is absolutely continuous with respect to $\pi(x)$], then

$$\operatorname{Var} \psi(x) \lesssim \frac{1}{N^\alpha P_0^\beta}\,. \tag{7}$$

Here the $\lesssim$ reminds us that Eq. (7) is only an asymptotic series in $1/N$.

Eqs. (6, 7) mean that, in order to bound possible errors, one should make an assumption that the coordinate system chosen is "reasonable": the Jacobian that maps $\pi(x)$ into $U(x)$ is never singular, or $P(x)$ is bounded away from zero. Surely, the assumption must not necessarily be hard and may be enforced smoothly by a prior, but it is needed.

Finally, we notice that P_0 describes how parameterization–dependent the learning machine is: when P_0 decreases, the set of probability densities that can be learned well grows. If one were able to guarantee uniformly small errors for $P_0 \to 0$, the learning machine would succeed for all densities and would be reparameterization–covariant. Thus Eq. (7) bounds a combination of two errors (estimation of ϕ and non-covariance) by a function that falls as a power law in the number of data. The details (such as α, β, and the choice of $\operatorname{Var} \psi$ as the error measure) certainly depend on a particular problem setup and on assumptions about the smoothness of ϕ and the allowed values of the Radon–Nikodym derivative. However, it is clear that the general result should be common: one can trade better approximate covariance for better non-covariant estimation, and N constraints the balance. It is interesting whether Occam style arguments [Mac92, Bal97, NB02] can be used to find the optimum tradeoff between the two errors.

Acknowledgments

We thank Vijay Balasubramanian, William Bialek, Curtis Callan and Vipul Periwal for many stimulating discussions.

References

[Bal97] Vijay Balasubramanian. Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Comp.*, 9:349–368, 1997.

[BCS96] William Bialek, Curtis Callan, and Steve Strong. Field theories for learning probability distributions. *Phys. Rev. Lett.*, 77:4693–4697, 1996.

[BNT01] William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neur. Comp.*, 13:2409–2463, 2001.

[CS01] Felipe Cucker and Steve Smale. On the mathematical foundations of learning theory. *Bull. (New Ser.) Amer. Math. Soc.*, 39(1):1–49, 2001.

[Hol97] Timothy E. Holy. Analysis of data from continuous probability distributions. *Phys. Rev. Lett.*, 79:3545–3548, 1997.

[Mac92] D. J.C̃. MacKay. Bayesian interpolation. *Neural Comp.*, 4:415–447, 1992.

[NB02] Ilya Nemenman and William Bialek. Occam factors and model independent Bayesian learning of continuous distributions. *Phys. Rev. E*, 65, 2002.

[Per97] Vipul Periwal. Reparameterization invariant statistical inference and gravity. *Phys. Rev. Lett.*, 78:4671–4674, 1997.

[Per99] Vipul Periwal. Geometric statistical inference. *Nucl. Phys. B*, 554 [FS]:719–730, 1999.

[Vap98] Vladimir Vapnik. *Statistical learning theory.* John Wiley & Sons, New York, 1998.