Tractable, Optimal High Dimensional Bayesian Inference

Madhu Advani Surya Ganguli

Modern neuroscience has entered the era of high dimensional data. It is often the case that we can simultaneously record N = O(100) to O(1000) neurons but only for a limited number P trials per condition, which may be the same order of magnitude as the dimensionality of data (N). This high dimensional data scenario carries with it the curse of dimensionality; in essence it is exceedingly difficult to analyze such limited amounts of high dimensional data by estimating large probabilistic models.

We develop analytical methods and algorithms to combat this curse for a wide variety of (Bayesian) regression problems. For example, consider the problem of estimating functional connectivity between neurons. This is effectively a high dimensional regression problem in which an unknown pattern of N synaptic weights onto a noisy neuron are to be inferred, based on P measured inputs and outputs to the neuron, and where we may have an appropriate prior over synaptic weights.

Maximum likelihood (ML) or maximum a posteriori (MAP) estimation are almost ubiquitous methods for solving this regression problem. We consider instead the optimal tractable estimator; this calculation involves optimizing an arbitrary loss function over the unknown synaptic weights. We used methods from statistical mechanics to find this optimal loss. Intriguingly, we find the optimal function is neither ML, nor MAP, but involves a smoothed version of the log-likelihood function and the prior where the degree of smoothing depends on the ratio P/N. The optimal loss function enjoys substantial improvements in squared error relative to ML and MAP.

These results indicate that widely cherished Bayesian procedures for analyzing data must be modified in the high dimensional setting. In essence, they suggest we should strive not just to be Bayesians, but smooth Bayesians.

1 Additional Detail

Consider the following Bayesian regression problem in which we must estimate an unknown vector w^0 (i.e. a vector of N synaptic weights). The weights obey,

$$Y_i = X_i^T w^0 + \epsilon_i \tag{1}$$

where Y_i are (neuron) outputs, X is our design matrix of (neural) inputs, ϵ_i is noise drawn from a distribution $f(\epsilon)$, and i indexes the P sample data points. Furthermore the weights, or components of w^0 are drawn i.i.d. from a prior distribution $R(w)$.

We consider the class of M-estimators of the form

$$\hat{w} = \arg\min_w \left[\sum_{i=1}^{P} \rho(y_i - \sum_j X_{ij} w_j) + \sum_{j=1}^{N} \sigma(w_j) \right] \tag{2}$$

where ρ is the loss function and σ is a regularizer. If ρ were the logarithm of the distribution of the noise ϵ, and σ were the logarithm of the prior R, then $\hat{w}$ would be the MAP estimate.

Our goal is to find the optimal M-estimator, or the optimal functions ρ_{opt} and σ_{opt} that on average minimize the squared error $\|\hat{w} - w^0\|^2$. Note that the posterior mean of w given the data is the absolute optimal minimum mean squared error estimator in Bayesian regression, but the posterior mean is not an

M-estimator; its computation requires a high dimensional integral over the vector w that is in general intractable. Thus we limit ourselves to the class of M-estimators, whose minimization is much more computationally tractable.

We develop a method, based on the replica theory of the statistical mechanics of disordered systems, to compute the optimal ρ and σ. Our results thus provide a new derivation as well as an important extension of El, Karoui et. al. 2013, who only considered a heuristic derivation of the ML scenario, and did not consider the Bayesian scenario. We find that the optimal ρ_{opt} and σ_{opt} have the analytic forms,

$$\rho_{opt}(x) = \left(\frac{x^2}{2} + \hat{q}_0 \ln(\zeta(x))\right)^* - \frac{x^2}{2} \tag{3}$$

$$\sigma_{opt}(x) = \frac{\hat{q}_0}{\hat{a}} \left[\left(\frac{x^2}{2} + \hat{a} \ln(\xi(x))\right)^* - \frac{x^2}{2}\right] \tag{4}$$

Where $\zeta_{\hat{q}_0} = f * \phi_{\hat{q}_0}$ defines a convolution between the noise and a gaussian ϕ of variance q_0. Similarly $\xi_{\hat{a}} = R * \phi_{\hat{a}}$. Also $.^*$ denotes the Fenchel conjugate operator $g(x)^* = \sup_y \left[xy - g(y)\right]$. Finally $\hat{a}$ and $\hat{q}_0$ are a pair of scalars that are are the solutions to the constrained optimization minimizing q_0 s.t. $aI_{q_0} = \frac{N}{P}$ and $a^2 J_a = a - q_0$, where $I_{q_0} = \int \frac{(\zeta'_{q_0}(x))^2}{\zeta_{q_0}(x)} dx$ and $J_a = \int \frac{(\xi'_a(x))^2}{\xi_a(x)} dx$ are related to the Fisher Information.

Thus ρ_{opt} is closely related to the log likelihood term $\log f$ and σ_{opt} is closely related to the log prior $\log R$, however they are systematically modified and smoothed in a manner that depends on the ratio of the dimensionality of the problem (N) to the amount of data (P). In essence, higher dimensionality necessitates more smoothing.

We expect these optimal estimators will have an important, widespread application to data analysis in neuroscience, as high dimensional Bayesian regression is a fundamental data analytic method relevant for modern neuroscientific datasets, and our results indicate that exceedingly common approaches to do this, like ML and MAP, are suboptimal.