Prior Information as a Natural Path to Regularization

Malik Magdon-Ismail

Caltech 136-93

Pasadena, CA 91125

magdon@cco.caltech.edu

June 2, 1998

Abstract

Bayesian inference requires the knowledge of a prior. The necessity of (non-uniform) prior information is implied by the various No Free Lunch theorems that show that a uniform prior over the targets leads to the inability to distinguish between learning algorithms. We develop a framework for the incorporation of prior information into the learning process that leads to the prior information being represented as a regularization term. We formally define the nature of the prior information as a natural characteristic, and then show how, given a natural characteristic, it is possible to deduce the regularization term. As applications, we derive the Tikhonov regularizers from reasonable priors and also derive the regularization terms implied by other common constraints.

Keywords: Prior information, natural characteristic, Tikhonov regularizer, finite grid, counting functions, random walk.

1 Introduction

The application of Bayesian inference assumes the knowledge of a prior. Prior information can be used to greatly improve upon learning performance, if the *assumed* prior information is correct. Effectively, the search space becomes narrower without sacrificing any approximation capability. Thus, one expects the BIAS to remain constant and the VARIANCE to decrease [Geman and Bienenstock, 1992]. The capacity of the learning model will generally have decreased [Sill, 1998a], and possibly the VC dimension as well [Abu-Mostafa, 1993]. One therefore expects the generalization performance to be closer to the in-sample performance, and since we have not sacrificed any in-sample performance, one expects the generalization performance to be better.

One technique for incorporating prior information about the target function into the learning process is through the use of hints [Abu-Mostafa, 1995]. Hints can be enforced "softly" by using virtual examples (extra examples generated using (for example) a known invariance of the target function) or by introducing a penalty function that is minimized when the hint is obeyed. These techniques have been applied with success in the financial markets (ex. [Abu-Mostafa and Atiya, 1996]) and to medical, bond, and credit data by [Sill and Abu-Mostafa, 1997]. Another way to enforce hints is to have a learning model that contains only functions that obey the hint (ex.[Sill, 1998b]). [Sill and Abu-Mostafa, 1997] shows how a "softly" enforced penalty hint can be interpreted in a Bayesian formalism by assigning a probability distribution to the hint violation. Incorporating prior information into the learning process will improve the performance. On the other hand, if the *assumed* prior information is incorrect, it could be potentially disastrous.

• Is it necessary to incorporate some prior information into the learning process if one is to get anywhere at all?

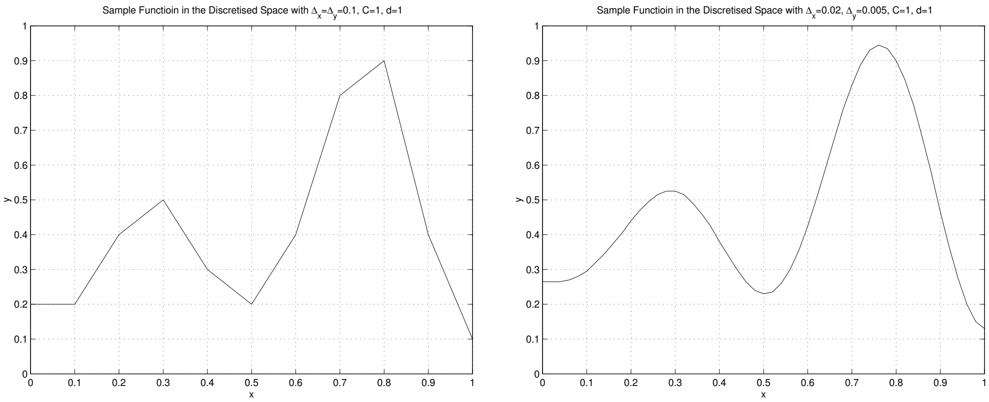
By enforcing hints (and other prior information), one is assigning a prior probability measure to the functions according to whether the hint (other prior information) is obeyed or not. We are subtly doing this by restricting ourselves to a learning model in the first place – a uniform prior over the learning model and zero outside. One expects that the better our prior, the better the performance should be – approximating a sine function in the Fourier basis should certainly do better than (say) using a polynomial basis, with the same amount of effort. Unfortunately, obtaining prior information requires knowledge about the target function, which is the *unknown*. As a result, in many applications, one simply assumes a uniform prior.

In the limit of infinite data, the dependence on the prior usually disappears. In the real world, however, one is dealing with a finite amount of data. In this real world, can one get away with assuming as un-informative a prior as the uniform one? Various no free lunch (NFL) results show that unless one learns with a specific non-uniform prior in mind, all hypotheses are just as likely to yield a given generalization behavior [Wolpert, 1996]. Any non-trivial performance can only be claimed relative to a prior, therefore one is forced to assume a prior.

Thus far we have illustrated why the issue of the prior is important in order to make substantial statements about learning systems. Perhaps the most commonly considered reference prior for Bayesian analysis is the Jeffrey's prior [Jeffreys, 1946] which is invariant under reparametrization and is approximately non-informative [Clarke and Wasserman, 1993], [Clarke and Barron, 1994]. Our goal here is to look at reasonable priors when prior information is available, and to deduce what the optimal learning strategy should be (in the Bayesian sense). We will discuss how priors could be used in the derivation of regularizers, such as the first order Tikhonov regularizer.

We will develop a framework for the incorporation of priors into the learning process. Perhaps it is possible to make some progress without assuming an entire prior over the target function space. Perhaps all we need are certain properties about this prior – we define these properties as natural characteristics (definition 1.1). We will also assume the input and output spaces to be discretized. Given these natural characteristics, we then reduce the problem of incorporating the prior information to the purely combinatorial problem of counting the number of functions that satisfy a certain constraint on a discretized grid. We will formulate and present an analysis of the counting problem for the finite grid that we will shortly discuss. In the asymptotic limit ($\Delta_x \to 0$), we use the central limit theorem to obtain an approximate expression for the solution to the counting problem. Using our technique, we will show that certain well known regularizers naturally arise from reasonable priors.

We begin here by laying the foundations for the uniform prior, that which is almost always assumed, and proceed from there. We will concentrate on compact spaces, and assume that the input space is the unit hypercube in d dimensions. Let the output space be the bounded



(a) (b)

The discretized function space has only a finite number of functions. We show a particular function on different quantization scales. An arbitrary precision can be obtained by making the grid finer and finer.

Figure 1: The discretized function space

interval $[0, C]$. Discretize the input and output spaces as follows. Let the input variable, $\mathbf{x}$, take on values in the set

$$\mathcal{X} = \{0, \Delta_x, 2\Delta_x, \ldots, 1 - \Delta_x, 1\}^d \tag{1}$$

so there are $N_x = 1 + \frac{1}{\Delta_x}$ possible values for each input variable. Let the output variable, y, take a value in the set

$$\mathcal{Y} = \{0, \Delta_y, 2\Delta_y, \ldots, C - \Delta_y, C\} \tag{2}$$

There are $N_y = \frac{C}{\Delta_y} + 1$ possible y values. A function (f) will be a mapping, $f : \mathcal{X} \to \mathcal{Y}$. There are $\Gamma = {N_y}^{{N_x}^d}$ different functions. By making Δ_x, Δ_y as small as we choose, we can achieve an arbitrary precision (for example see figure 1). In all applications, only a finite number of functions are available (computers are finite precision machines). Due to this finiteness, we can define a probability distribution on the function space – a prior over the function space. One natural prior is to take this distribution to be uniform. Therefore to every function we assign a probability of $\frac{1}{\Gamma}$. Many other "natural" priors are plausible.

In what follows, we will use the discretized grid already defined, and restrict ourselves to one input dimension $(d = 1)$. In this case, a function can be viewed as a path on a two dimensional grid, with x monotonically increasing along the path. If we assign a prior probability distribution to the functions, we can then view a function as a random walk on this two dimensional grid, with the distribution of the starting point and the distributions of future jumps being related to that prior.

Let a data set $\mathcal{D}$ be given and assume that the conditional density (likelihood) $P(\mathcal{D}|f)$ is known. Usually, this involves assuming a noise model and then computing the density that a particular noise realization occurred. Taking the Bayesian approach, we would like to find f that maximizes

$$P(f|\mathcal{D}) = \frac{P(\mathcal{D}|f)P(f)}{P(D)} \tag{3}$$

or, equivalently, one can maximize its logarithm, $\mathcal{E}$, given by

$$\mathcal{E} = \log[P(f|\mathcal{D})] = \log[P(\mathcal{D}|f)] + \log[P(f)] - \log[P(D)] \tag{4}$$

The last term can be ignored, and typically $P(f)$ is assumed to be uniform, so the $\log(P(f))$ term is also ignored, resulting in the maximum likelihood method for estimation. In the case of Gaussian noise, this results in having to minimize a sum of squared errors[1]. From the NFL results, we know that assuming $P(f)$ to be uniform is not a fruitful approach.

Suppose we can define some natural characteristic (ρ) of a path. Let $N(\rho)$ be the number of functions with the value of this natural characteristic equal to ρ. Because ρ is implied by knowing the function, we can write

$$P(f) \quad = \quad P(f,\rho) = P(f|\rho)P(\rho) \tag{5}$$

Now suppose that ρ is the *determining* characteristic that governs how the target function was picked. Then it seems reasonable to assume that all functions with the same value for ρ are equally likely:

$$P(f|\rho) = \frac{1}{N(\rho)}\delta_{\rho,\rho(f)} \tag{6}$$

Further if one also stipulates that all ρ's are equally likely ($P(\rho) = const$), then, (4) leads to

[1] Technically, Gaussian noise can not allowed on our grid but we will just use it as an illustrative, familiar example.

the following problem

$$\max_f \{\log[P(\mathcal{D}|f)] - \log[N(\rho(f))]\} \tag{7}$$

The following definition should now be intuitive.

Definition 1.1 *A natural characteristic, ρ, of the target function distribution is a property according to which functions are chosen such that*

- *All target functions with the same value of ρ are equally likely.*

If in addition,

- *All values of ρ are equally likely.*

then we call it a maximally informative characteristic (MIC).

Now, by definition, (7) is the problem we need to solve for MIC's. In general, we solve

$$\max_f \{\log[P(\mathcal{D}|f)] - \log[N(\rho)] + \log[P(\rho)]\} \tag{8}$$

where $\rho = \rho(f)$. All that remains is to find a suitable ρ for the problem at hand. $P(\rho)$ is a prior belief about how ρ is distributed. If $P(\rho) \propto N(\rho)$ then (8) reduces to maximum likelihood estimation. If ρ is a determining characteristic, and we have no reason to believe that any value of ρ is more likely than another, then we get (7). We have broken down the problem into two parts. The calculation of $N(\rho)$ which is purely a combinatorial problem, and the finding of a natural characteristic, the property of the prior on the target function required for (7) and the calculation of $N(\rho)$ to go through. The art is all in the natural characteristic – the ability to choose a natural characteristic that represents the true nature of the learning problem, and one for which $N(\rho)$ can be calculated. Given a problem, one needs some idea about how the target function was picked. We will reverse this process. We will consider a certain set of natural characteristics and from them, determine the objective function that one should optimize. This is a minor beginning toward the development of a catalog of natural characteristics and the objective functions they imply. A practitioner, with this catalog in hand, approaches a problem and tries to decide which natural characteristic to use. Then what remains is to optimize the objective function (7) that is implied by that characteristic. The characteristics that we will consider are variations of the *total variation* of a function, and some other reasonable bound constraints.

Define the variations, $\mathcal{T}_{ij}$, of a function by

$$\mathcal{T}_{ij}(f) = \int dx \left| \frac{\partial^i f}{\partial x^i} \right|^j \tag{9}$$

We will be interested in discretized versions of $\mathcal{T}_{11}, \mathcal{T}_{12}, \mathcal{T}_{22}$, the most commonly used Tikhonov type regularization terms.

2 Counting Functions

Let ρ for a function (path) be defined by

$$\rho(f) = \sum_{i=0}^{N_x - 1} e_i \tag{10}$$

where i indexes the possible x values. For example, if $e_0 = 0$, $e_i = |s_i - s_{i-1}|$ where $s_i = y_i/\Delta_y$, then ρ is the sum of the number of jumps in the path. It is a discrete analog of the integrated magnitude of the gradient, $\mathcal{T}_{11}$. In figure 1 (a), $\rho = 21$. For figure 1 (b), $\rho = 417$ (though this is not obvious from simply looking at the figure). Let $N(\rho)$ be the number of functions with $\sum e = \rho$. Our goal is to calculate $N(\rho)$ for various functional forms $\rho(f)$. We will assume that $e_i \geq 0$.

2.1 Setting up the Combinatorial Problem

Let $B(\rho, N_x, N_y, i, j)$ be the number of paths with $\sum e = \rho$ on the grid of size $N_x \times N_y$ with $s_0 = i$ and $s_{N_x} = j$. If either of the conditions $\rho < 0$, $N_x \leq 0$, $N_y < 0$, then B is defined to be zero. In this section, we will consider ρ of the form

$$\rho = \sum_{i=1}^{N_x} e(s_i, s_{i-1}) \tag{11}$$

when $e(x, y) = |x - y|$, ρ is the total variation. Assume that $e(\cdot, \cdot)$ is a symmetric function. From $B(\rho, N_x, N_y, i, j)$ we can calculate $N(\rho)$ as

$$N(\rho) = \sum_{i,j=0}^{N_y} B(\rho, N_x, N_y, i, j) \tag{12}$$

By symmetry, $B(\rho, N_x, N_y, i, j) = B(\rho, N_x, Ny, j, i)$.

2.2 Counting the Number of Functions

We will present two methods for calculating $B(\rho, N_x, N_y, i, j)$, one based on a recursive argument, and one based on computing the generating function for $B(\rho, N_x, N_y, i, j)$.

2.2.1 Recursive Approach

Proposition 2.1 *Let $N_1 > 0$, $N_2 > 0$ be two integers such that $N_1 + N_2 = N_x$, then*

$$B(\rho, N_x, N_y, i, j) = \sum_{k=0}^{N_y} \sum_{r} B(r, N_1, N_y, i, k) B(\rho - r, N_2, N_y, k, j) \tag{13}$$

PROOF: Fix k, r in the summations. Then, $B(r, N_1, N_y, i, k)B(\rho - r, N_2, N_y, k, j)$ is the number of paths from i to j passing through the point (N_1, k) with $\sum e = r$ in the first N_1 steps and $\sum e = \rho - r$ in the remaining N_2 steps. Summing over the possible values for r, we get the number of paths from i to j passing through the point (N_1, k) with $\sum e = \rho$. Summing over the possible values of k we get the number of paths from i to j with $\sum e = \rho$.
■

Given (13), we can compute $B(\rho, N_x, N_y, i, j)$ by starting at $N_x = 1$ and building upwards. More precisely, note that $B(\rho, 1, N_y, i, j) = \delta_{\rho, e(i,j)}$. Substituting this into (13) we get the following recursion

$$\begin{aligned} B(\rho, 1, N_y, i, j) &= \delta_{\rho, e(i,j)} && (14) \\ B(\rho, N_x + 1, N_y, i, j) &= \sum_{k=0}^{N_y} B(\rho - e(j, k), N_x, N_y, i, k) && (15) \end{aligned}$$

Now, $B(\rho, N_x, N_y, i, j)$ for any values of its arguments can be recursively computed by starting with $N_x = 1$ and increasing N_x by increments of 1 to the desired value. In the case where $e(x, y) = |x - y|$, the recursion becomes

$$B(\rho, 1, N_y, i, j) = \delta_{\rho,|x-y|} \tag{16}$$

$$B(\rho, N_x + 1, N_y, i, j) = \sum_{k=0}^{N_y} B(\rho - |x - y|, N_x, N_y, i, k) \tag{17}$$

2.2.2 Generating Function for $B(\rho, N_x, N_y, i, j)$

We would like to calculate $P(\rho|i, j)$, the probability of obtaining a path from i to j with $\sum e = \rho$, where we take s_k to be *i.i.d.* and $P(s_k = m) = 1/N_y$. Then,

$$B(\rho, N_x, N_y, i, j) = {N_y}^{N_x-2} P(\rho|i, j) \tag{18}$$

We can also calculate $P(\rho)$ from $P(\rho|i, j)$:

$$P(\rho) = \sum_{i,j} P(\rho|i, j)P(i, j) = \frac{1}{{N_y}^2} \sum_{i,j} P(\rho|i, j) \tag{19}$$

and now $N(\rho)$ is given by ${N_y}^{N_x} P(\rho)$. Let $\psi_{ij}(s) \equiv \left\langle e^{\hat{i}s\rho} \right\rangle_{P(\rho|i,j)}$. Since ρ takes on a discrete set of values, we can write

$$\psi_{ij}(s) = \sum_{\rho} P(\rho|i, j) r^{\rho} \tag{20}$$

where $r = e^{\hat{i}s}$. Given ψ_{ij}, we can read off $P(\rho|i, j)$ as the coefficient of r^{ρ}. Therefore we concentrate on ψ_{ij}.

$$\psi_{ij}(s) = \frac{1}{{N_y}^{N_x-2}} \sum_{s_1=0}^{N_y} \cdots \sum_{s_{N_x-1}=0}^{N_y} r^{\left(\sum_{i=1}^{N_x} e(s_i, s_{i-1})\right)} \tag{21}$$

$$= \frac{1}{{N_y}^{N_x-2}} \sum_{s_1 \ldots s_{N_x-1}} r^{e(i,s_1)} r^{e(s_1,s_2)} \ldots r^{e(s_{N_x-1},s_j)} \tag{22}$$

$$= \frac{1}{{N_y}^{N_x-2}} [\mathbf{M}^{N_x-1}]_{ij} \tag{23}$$

where we define an $N_y \times N_y$ "transfer matrix" $\mathbf{M}$ by

$$\mathbf{M}_{ij} = r^{e(i,j)} \tag{24}$$

To get $B(\rho, N_x, N_y, i, j)$ we need to find the coefficient of r^ρ in the expression for ψ_{ij}

$$B(\rho, N_x, N_y, i, j) = [\mathbf{M}^{N_x-1}]_{ij}\Big|_{r^\rho} \tag{25}$$

and for $N(\rho)$ we find the coefficient of r^ρ in the expression for $\psi = \sum_{i,j} \psi_{ij}$

$$N(\rho) = \sum_{i,j=0}^{N_y} [\mathbf{M}^{N_x-1}]_{ij}\Bigg|_{r^\rho} \tag{26}$$

When $e(x,y) = |x-y|$, $\mathbf{M}$ becomes the Toeplitz matrix given by $\mathbf{M}_{ij} = r^{|i-j|}$

$$\mathbf{M} = \begin{pmatrix} 1 & r & r^2 & \dots & r^{N_y-1} \\ r & 1 & r & \dots & r^{N_y-2} \\ \vdots & & \ddots & & \vdots \\ r^{N_y-2} & \dots & r^2 & r & 1 \end{pmatrix} \tag{27}$$

Technical Aside

We note a connection here to counting paths for a bounded random walk with variable step size, each step size having the same probability. The bounded grid is exactly the grid we have been talking about all along. If we want to calculate the probability that, having started at point i and ended at point j, the total distance traveled in the y direction is ρ, then we need to calculate the coefficient of the $e^{\hat{i}s\rho}$ term in $[\mathbf{M}^{N_x-1}]_{ij}$ where $\mathbf{M}$ is given by (27). If N_x is large, then, the largest eigenvalue of $\mathbf{M}$ will exponentially dominate. Diagonalizing $\mathbf{M}$ and ignoring the lower eigenvalues of $\mathbf{M}$ we find that

$$[\mathbf{M}^{N_x-1}]_{ij} \approx \alpha(e^{\hat{i}s}) z_i(e^{\hat{i}s}) z_j(e^{\hat{i}s}) \tag{28}$$

where $\alpha(s)$ is the largest eigenvalue and $\mathbf{z}$ is the corresponding eigenvector. [Grenander and Szegö, 1958, pg 69-75] studies this eigenvalue in some detail, including its limiting form as $N_y \to \infty$.

---

3 Asymptotic Solution to the Counting Problem

A closed form solution to the the recursive approach (13) or the generating function approach (23) has yet to be found for the cases of interest. For reasonable grid sizes, $N(\rho)$ could perhaps be computed numerically using either of the above methods. In practice, $N_x$and N_y are large ($O(2^{53})$), and the asymptotic limits $N_x, N_y \to \infty$ become relevent, so we can make use of the central limit theorem. We will need a version of the central limit theorem that is valid for dependent random variables. For convenience we state it here.

Theorem 3.1 (Central Limit Theorem (CLT), [Billingsley, 1986, pg 375]) *Suppose that $X_1, X_2, \ldots$ is stationary and α-mixing with $\alpha_n = O(n^{-5})$ and that $\langle X_n \rangle = 0$ and $\langle X_n^{12} \rangle < \infty$. If $S_n = X_1 + \cdots + X_n$, then*

$$\frac{Var[S_n]}{n} \to \sigma^2 = \left\langle X_1^2 \right\rangle + 2 \sum_{k=1}^{\infty} \langle X_1 X_{1+k} \rangle \tag{29}$$

where the series converges absolutely. If $\sigma > 0$, then $S_n/\sigma\sqrt{n} \Rightarrow N(0,1)$ ∎

where $N(0,1)$ is the standard normal distribution. Figure 2 illustrates this convergence to a Gaussian for $N_x \sim 100, N_y \sim 50$.

In general, the conditions of the theorem will be satisfied for $X_i = e_i$, so, the probability $P(\rho)$ approaches a Gaussian density. Therefore, to investigate $P(\rho)$ in the asymptotic limits $N_x, N_y \to \infty$ it suffices to calculate $\langle \rho \rangle$ and $Var[\rho]$. Then $N(\rho)$ is given by

$$N(\rho) = {N_y}^{N_x} P(\rho) \to \frac{1}{\sqrt{2\pi Var[\rho]}} \exp\left[-\frac{(\rho - \langle \rho \rangle)^2}{2Var[\rho]} \right] \tag{30}$$

In the remainder of this section we will list $\langle \rho \rangle$ and $Var[\rho]$ for different forms of ρ as $N_x, N_y \to \infty$. The details can be found in appendix 6.1.

1. $e_i = |s_i - s_{i-1}|$ *(Total Variation Constraint $\sim \mathcal{T}_{11}$).*

$$\langle \rho \rangle \to \frac{N_x N_y}{3}, \qquad Var[\rho] \to \frac{N_x {N_y}^2}{15} \tag{31}$$

$N(\rho) = {N_y}^{N_x} N(\langle \rho \rangle, Var[\rho])$.

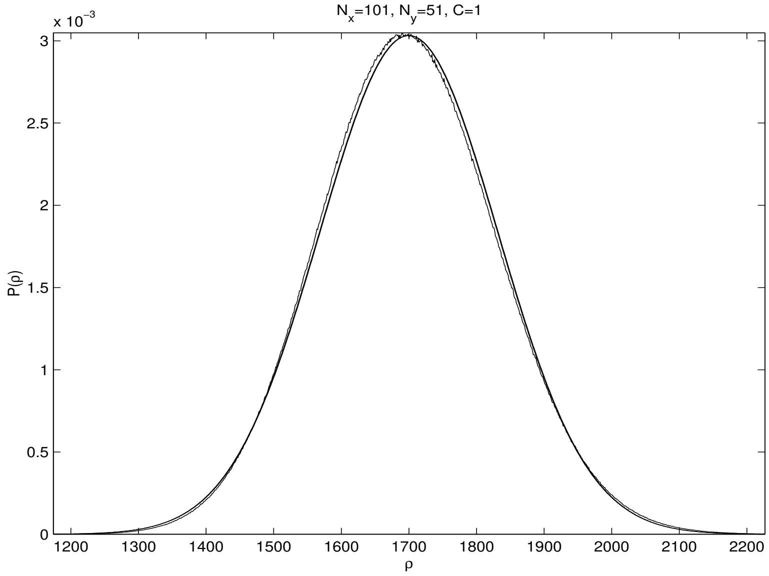


Illustration of the convergence of $N(\rho)$ to a Gaussian for $\rho = |s_i - s_{i-1}|$. Plotted are three! curves, one a montecarlo simulation to obtain $P(\rho)$, the second, a Gaussian with the same mean and variance of ρ that was obtained from the montecarlo simulation and the third is the Gaussian implied by the theoretically computed mean and variance of ρ . The fit is impressive given that N_x was only 101.

Figure 2: $N(\rho)$ and its Gaussian approximation

2. $e_i = (s_i - s_{i-1})^2$ *(Squared Total Variation Constraint $\sim \mathcal{T}_{12}^2$).*

$$\langle\rho\rangle \to \frac{N_x {N_y}^2}{6}, \qquad Var[\rho] \to \frac{N_x {N_y}^4}{20} \tag{32}$$

$N(\rho) = {N_y}^{N_x} N(\langle\rho\rangle, Var[\rho])$.

3. $e_i = (2s_i - s_{i-1} - s_{i+1})^2$ *(Smoothness Constraint $\sim \mathcal{T}_{21}$).*

$$\langle\rho\rangle \to \frac{N_x {N_y}^2}{2}, \qquad Var[\rho] \to \frac{121 N_x {N_y}^4}{180} \tag{33}$$

$N(\rho) = {N_y}^{N_x} N(\langle\rho\rangle, Var[\rho])$.

4. $e_i = g(s_i)$. In this case, $\langle\rho\rangle = N_x\langle g(s_0)\rangle$ and $Var[\rho] = N_x Var[g(s_0)]$

(a) $g(s_i) = s_i/N_x$ *(Mean Constraint).*

$$\langle \rho \rangle \to \frac{N_y}{2}, \qquad Var[\rho] \to \frac{{N_y}^2}{12N_x} \tag{34}$$

$N(\rho) = {N_y}^{N_x} N(\langle \rho \rangle, Var[\rho])$.

(b) $g(s_i) = s_i^2/N_x$ *(Power Constraint).*

$$\langle \rho \rangle \to \frac{{N_y}^2}{6}, \qquad Var[\rho] \to \frac{{N_y}^4}{45N_x} \tag{35}$$

$N(\rho) = {N_y}^{N_x} N(\langle \rho \rangle, Var[\rho])$.

5. $\rho = \max_i\{s_i\}$ *(Maximum Power Constraint)* To get $N(\rho)$, note that of the $(1+\rho)^{N_x}$ paths that are available, the ρ^{N_x} that do not reach the level ρ are excluded, so

$$N(\rho) = (\rho + 1)^{N_x} - \rho^{N_x} \tag{36}$$

6. *Multiple Constraints* Suppose that the natural characteristic is a combination of properties of the path – a vector characteristic ($\vec{\rho}$). One needs to compute the multivariate probability distribution $P(\vec{\rho})$. Once again, in the fine grid limit, we can appeal to the CLT and reduce the computation to computing the parameters of the multivariate Gaussian. These are the mean vector ($\langle \vec{\rho} \rangle$) and the covariance matrix (Σ) given by

$$\Sigma = \left\langle \vec{\rho}\vec{\rho}^{\,T} \right\rangle - \langle \vec{\rho}\, \rangle \langle \vec{\rho}\, \rangle^T \tag{37}$$

Because $\Delta\rho = 1$, $N(\rho)$ is also the number density of functions at ρ, the number of functions with ρ in the interval $[\rho, \rho + \Delta\rho]$. It is this number density that is needed in what follows.

4 Application to Regularization

In the previous section, we discussed the computation of $N(\rho)$. Remember that the presumption is that ρ is a natural characteristic, so $P(f|\rho) = 1/N(\rho)$. ρ was generally a function of

Characteristic	Discrete Version	$N(\rho)$	Limiting Form
$\int_0^1 dx\left\|\frac{dy}{dx}\right\|$	$\frac{C}{N_y}\sum_i \|s_i - s_{i-1}\|$	$N(\frac{N_x C}{3}, \frac{N_x C^2}{15})$	$\propto e^{\frac{5\rho}{C}}$
$\int_0^1 dx(\frac{dy}{dx})^2$	$N_x\left(\frac{C}{N_y}\right)^2\sum_i(s_i - s_{i-1})^2$	$N(\frac{{N_x}^2C^2}{6}, \frac{{N_x}^3C^4}{20})$	$\propto e^{\frac{10\rho}{3N_x C^2}}$
$\int_0^1 dx(\frac{d^2y}{dx^2})^2$	${N_x}^3\left(\frac{C}{N_y}\right)^2\sum_i(2s_i - s_{i-1} - s_{i+1})^2$	$N(\frac{{N_x}^4C^2}{2}, \frac{121{N_x}^7C^4}{180})$	$\propto e^{\frac{90\rho}{121{N_x}^3C^2}}$
$\int_0^1 dx\|y\|$	$\left(\frac{C}{N_y}\right)\frac{1}{N_x}\sum_i s_i$	$N(\frac{C}{2}, \frac{C^2}{15N_x})$	$N(\frac{C}{2}, \frac{C^2}{15N_x})$
$\int_0^1 dx\ y^2$	$\left(\frac{C}{N_y}\right)^2\frac{1}{N_x}\sum_i s_i^2$	$N(\frac{C^2}{6}, \frac{C^4}{45N_x})$	$N(\frac{C^2}{6}, \frac{C^4}{45N_x})$
$\sup\{\|y\|\}$	$\frac{C}{N_y}\max_i\{s_i\}$	$\left(\frac{N_y\rho}{C}+1\right)^{N_x} - \left(\frac{N_y\rho}{C}\right)^{N_x}$	$\propto \rho^{N_x}e^{\frac{C}{\lambda\rho}}$ $\left(\frac{N_y}{N_x}\to\lambda\right)$

In the last column we give the limit as $N_x \to \infty$ and we assume that $P(\rho)$ (8) is non-zero only on some bounded region . Taking the log of this column gives the relevant "regularization" term.

Table 1: Various characteristics and the form for $N(\rho)$ that they lead to.

the s_i's. We are more interested in characteristics defined in terms of $y_i = \Delta_y s_i = Cs_i/N_y$. For example,

$$\int_0^1 dx\left|\frac{dy}{dx}\right| = \sum_i |y_i - y_{i-1}| = \frac{N_y}{C}\sum_i |s_i - s_{i-1}| \tag{38}$$

Table 1 gives some characteristics that might be of interest, and the corresponding discrete versions, along with $N(\rho)$, the number density of functions at ρ.

Given $N(\rho)$, we form the objective function to minimize using (7) by taking the logarithm of $N(\rho)$. Suppose that we assume that the data was created by taking the true targets and adding Gaussian noise, which we assume to have zero mean and variance σ^2. Then up to the addition of a constant, $\log[P(D|f)] = -1/2\sigma^2\sum_l r_l^2$ where we have defined the residuals $r_l = f(x_l) - y_l$, and l indexes the data points. Up to constants, which we ignore, the optimization problem (7) reduces to

$$\min_f\left\{\frac{1}{2\sigma^2}\sum_l r_l^2 + \log[N(\rho(f))]\right\} \tag{39}$$

where we have assumed that every value of ρ is equally likely. The second term can be considered a regularization term, and is given by the logarithm of the final column in table

1. If one has some other prior on the possible values of ρ (the natural characteristic) then this can be incorporated by using (8).

4.1 Derivation of the Tikhonov Regularizer

Many times, one wishes to enforce smoothness constraints on the learned function, so one heuristically introduces Tikhonov type regularization terms. Suppose that ρ is one of the smoothness measures. Then, if one would like to distinguish between target functions on the basis of ρ, and only ρ, one is making the statement that all target functions with the same value of ρ are equally likely, ie. that ρ is a natural characteristic. Thus we can apply the results from table 1 to get the form of the regularization term. We see that for all the smoothness characteristics, the dependence on ρ is given by $\exp(\lambda\rho)$, where λ can be read off from table 1. This leads to a regularization term $\lambda\rho$, which is precisely Tikhonov type regularization for the discrete setup we have described.

4.1.1 The Continuum Limit

Applying these methods to the continuum limit (though practically not necessary) could be interesting from the theoretical point of view. A little bit of care has to be used in passing to the continuum limit because for the derivatives to approach the true derivative, one needs $\Delta_y << \Delta_x$ (better resolution on the $y-$axis than the $x-$axis). We can define a precision to which we would like to measure the derivatives by the ratio Δ_y/Δ_x which we will assume tends to $1/\lambda$.

1. $\rho = \frac{C}{N_y}\sum_i |s_i - s_{i-1}| \to \mathcal{T}_{11}$. The regularization term is $5\mathcal{T}_{11}/C$. This is exactly the Tikhonov regularizer with regularization parameter equal to $5/C$.
2. $N_x\left(\frac{C}{N_y}\right)^2 \sum_i (s_i - s_{i-1})^2 \to \mathcal{T}_{12}$. The regularization term is $(10\mathcal{T}_{12}/3C^2)\lambda\Delta_y$. This is exactly the Tikhonov regularizer with regularization parameter equal to $(10/3C^2)\lambda\Delta_y$ (a similar result holds for $\mathcal{T}_{22}$–the regularization parameter is $(90/121C^2)\lambda^3{\Delta_y}^3$). Note that in these cases, the precision in the derivative and the $y-$quantization appear. We offer the following intuition. Regularization is found most useful in situations where noise is present, and one can view the quantization $\lambda\Delta_y = \Delta_x$ as "artificial noise",

thus as Δ_x decreases, we expect the need for less regularization. In fact, [Bishop, 1995] shows that training with input noise is equivalent to having a regularization term $\sim \mathcal{T}_{12}$.

5 Comments

The role of priors is important. Here we have discussed how to go from a particular prior (natural characteristic) to a regularization term by using counting arguments for paths on a grid. Given a problem, the aim should be to formulate a prior belief about what the most determining feature of the target function is, its natural characteristic (or set of natural characteristics), and then one can compute the relevant regularization term using a Bayesian framework as we have illustrated. One often hears that most functions encountered in practice are smooth. Exactly what does this mean? One is trying to convey the information that two target functions of equal smoothness are equally likely. We have shown that such smoothness natural characteristics lead in general to Tikhonov type regularization terms, the minimization of an error function of the form

$$\mathcal{E} = \frac{1}{N} \sum_{l} r_l^2 + \frac{2\sigma^2\lambda}{N} \mathcal{T}_{ij} \tag{40}$$

where we factored out the $1/2\sigma^2$ in (39). We have given a prescription for the computation of λ, and have computed the regularization terms for a variety of natural characteristics (priors). Note that the term C that appears as a bound for the grid is also a prior.

We have addressed not what regularization terms one should use, but why regularization terms should arise naturally from the choice of a prior, and how to calculate them. Why one should use a certain regularization term is entirely a function of the problem at hand.

6 Appendix

6.1 $N(\rho)$ for various functional forms $\rho(f)$

1. $e_i = |s_i - s_{i-1}|$*(Total Variation Constraint)*$\sim \mathcal{T}_{11}$.

   For the expected value of ρ we have

$$\begin{aligned}
\langle\rho\rangle &= \left\langle \sum_{i=1}^{N_x-1} |s_i - s_{i-1}| \right\rangle = \sum_{i=1}^{N_x-1} \langle |s_i - s_{i-1}| \rangle & (41)\\
&= (N_x-1)\langle |s_1 - s_0| \rangle = \frac{N_x-1}{{N_y}^2} \sum_{s_0,s_1=0}^{N_y-1} |s_1 - s_0| & (42)\\
&= \frac{(N_x-1)(N_y-1)(N_y+1)}{3N_y} & (43)
\end{aligned}$$

   To compute the variance of ρ, first we will compute the expected value of ρ^2

$$\left\langle \rho^2 \right\rangle = \left\langle \sum_{i=1}^{N_x-1} \sum_{j=1}^{N_x-1} |(s_i - s_{i-1})(s_j - s_{j-1})| \right\rangle \tag{44}$$

   We need to consider four cases separately: $j = i$; $j = i+1$;$j = i-1$; neither of the above. Taking these four cases into consideration we have

$$\begin{aligned}
\left\langle \rho^2 \right\rangle &= (N_x-1)\left\langle (s_1-s_0)^2 \right\rangle + 2(N_x-2)\langle |(s_1-s_0)(s_2-s_1)| \rangle + \\
&\quad ((N_x-1)^2 - 3(N_x-1)+2)\langle |s_1-s_0| \rangle^2 & (45)
\end{aligned}$$

   We need to compute these expectations with respect to the probability distribution $P(s_0, s_1) = 1/{N_y}^2$. Then, $Var[\rho]$ is given by $\langle\rho^2\rangle - \langle\rho\rangle^2$. A tedious calculation yields:

$$\begin{aligned}
Var[\rho] &= \frac{(N_x-1)(N_y-1)(N_y+1)({N_y}^2+1)}{15{N_y}^2} - \\
&\quad \frac{(N_y-1)(N_y-2)(N_y+1)(N_y+2)}{90{N_y}^2} & (46)
\end{aligned}$$

In the limit $N_x, N_y \to \infty$,

$$\langle \rho \rangle \to \frac{N_x N_y}{3}, \qquad Var[\rho] \to \frac{N_x {N_y}^2}{15} \tag{47}$$

2. $e_i = (s_i - s_{i-1})^2$ *(Squared Total Variation Constraint* $\sim \mathcal{T}_{12}$*)*

   Following similar techniques as above, we find

$$\begin{aligned} \langle \rho \rangle &= \sum_{i=1}^{N_x - 1} \left\langle (s_i - s_{i-1})^2 \right\rangle & (48) \\ &= \frac{(N_x - 1)(N_y - 1)(N_y + 1)}{6} & (49) \end{aligned}$$

   Once again, for the variance, we need to consider separately the four cases as above. A tedious calculation yields:

$$\begin{aligned} Var[\rho] &= \frac{(N_x - 1)(N_y - 1)(3(N_y - 1)^3 + 12(N_y - 1)^2 + 8(N_y - 1) - 8)}{60} \\ &\quad - \frac{N_y((N_y - 1)^3 + 4(N_y - 1)^2 + (N_y - 1) - 6)}{90} & (50) \end{aligned}$$

   In the limit $N_x, N_y \to \infty$,

$$\langle \rho \rangle \to \frac{N_x {N_y}^2}{6}, \qquad Var[\rho] \to \frac{N_x {N_y}^4}{20} \tag{51}$$

3. $e_i = (2s_i - s_{i-1} - s_{i+1})^2$ *(Smoothness Constraint)* $\sim \mathcal{T}_{21}$

   Again, using similar techniques as above, we find

$$\begin{aligned} \langle \rho \rangle &= \frac{(N_x - 2)(N_y - 1)(N_y + 1)}{2} & (52) \\ Var[\rho] &= \frac{(N_x - 1)(N_y - 1)(121(N_y - 1)^3 + 484(N_y - 1)^2 + 376(N_y - 1) - 216)}{180} \\ &\quad - \frac{N_y(181(N_y - 1)^3 + 724(N_y - 1)^2 + 556(N_y - 1) - 336)}{180} & (53) \end{aligned}$$

   In the computation of the variance, the six cases ($j = i$; $j = i \pm 1$;$j = i \pm 2$; neither of the above) need to be handled separately. The limiting forms are

$$\langle\rho\rangle \rightarrow \frac{N_x {N_y}^2}{2}, \qquad Var[\rho] \rightarrow \frac{121 N_x {N_y}^4}{180} \tag{54}$$

4. $e_i = |2s_i - s_{i-1} - s_{i+1}|$ *(Smoothness Constraint* $\sim \mathcal{T}_{21}$*)* Though tedious, this can also be calculated using the above techniques.

7 Acknowledgments

We would like to thank Yaser Abu-Mostafa and the members of the Caltech Learning Systems Group for helpful comments.

References

[Abu-Mostafa and Atiya, 1996] Y. S. Abu-Mostafa and A. F. Atiya. Introduction to financial forecasting. *Applied Intelligence*, 6:205–213, 1996.

[Abu-Mostafa, 1993] Y. S. Abu-Mostafa. Hints and the VC dimension. *Neural Computation*, 4:278–288, 1993.

[Abu-Mostafa, 1995] Y. Abu-Mostafa. Hints. *Neural Computation*, 4(7):639–671, 1995.

[Billingsley, 1986] P. Billingsley. *Probability and Measure.* Wiley Series in Probability and Mathematical Statistics. Wiley, 1986.

[Bishop, 1995] C. M. Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995.

[Clarke and Barron, 1994] B. S. Clarke and A. R. Barron. Jeffery's prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41:37–60, 1994.

[Clarke and Wasserman, 1993] B. Clarke and L. Wasserman. Noninformative priors and nuisance parameters. *American Statistical Association*, 88(424):1427–1432, 1993.

[Geman and Bienenstock, 1992] S. Geman and E. Bienenstock. Neural networks and the bias variance dilemma. *Neural Computation*, 4:1–58, 1992.

[Grenander and Szegö, 1958] U. Grenander and Szegö. *Toeplitz Forms and their Applications.* California Monographs in Mathematical Sciences. University of California Press, Los Angeles, 1958.

[Jeffreys, 1946] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London*, A(196):453–461, 1946.

[Sill and Abu-Mostafa, 1997] J. Sill and Y. S. Abu-Mostafa. Monotonicity hints. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 9, pages 634–640. Morgan Kaufmann, 1997.

[Sill, 1998a] J. Sill. The capacity of monotonic functions. *Discrete Applied Mathematics*, Special Issue on VC Dimension, 1998.

[Sill, 1998b] J. Sill. Monotonic networks. In *Advances in Neural Information Processing Systems (NIPS)*, volume 10, 1998.

[Wolpert, 1996] D. H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390, 1996.