2021-03-30

1 Introduction

This week, we will consider the case of functions $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^m$ where either n is large (today) or m is large (Thursday). In both cases, we seek a hidden low-dimensional structure that makes the function in some way easier to handle.

Our main setting today will be scalar-valued functions ($m = 1$ and n large) on a compact set Ω. We assume that our function has some type of regularity; the simplest case is assuming $f \in C^p(\Omega, \mathbb{R})$, i.e. f has p continuous derivatives. In this case, the optimal rate of uniform approximation as a function of sample points N is $\|f - \hat{f}\|_\infty = O(N^{-p/n})$. Hence, for functions where all we know in advance is bounded smoothness, the number of samples required to reach a given error tolerance grows exponentially with the dimension n. This is one instance of the *curse of dimensionality.* The same basic limit on approximation creates similar bottlenecks for global optimization and for quadrature (the computation of definite integrals).

Fortunately, many functions we encounter in practice have additional structure, such as high degrees of smoothness (at least in some directions, or away from some creases or singular points) or an effective low-dimensional dependence on the input, or a sparse sum of univariate (or low-dimensional) functions. Many adaptive algorithms for function approximation work because they are look for signs of these types of structure, and exploit it when it seems to be present. We focus today on a specific case where $f(x)$ is effectively low-dimensional.

2 Active subspaces

We consider the case where our high-dimensional function $f(x)$ looks behind the scenes like $f(x) = g(Ax + \epsilon(x))$ where $A \in \mathbb{R}^{k \times n}$ and $g : \mathbb{R}^k \to \mathbb{R}$ for some modest k. The span of the rows of A is an *active subspace* associated with the directions that matter for f, and the dependence of f on the orthogonal complement matters little (if at all). That is, almost all the dependence on x is mediated by the *active variables* $y = Ax$. The idea is closely related to the notion of *sufficient dimension reduction* in statistics.

Why should we ever expect to see this structure? We could say as a purely empirical matter that this type of low-dimensional subspace comes up often enough that we should check for it, and there is some value to this perspective. In other cases, scaling principles or physical invariants suggest possible ways to reduce the dimensionality (e.g. by *non-dimensionalizing* a physical problem before applying numerical methods). But in some of the situations where active subspaces are used, there is an *a priori* reason to expect that even though it may not be easy to tease out by analytic techniques like forming dimensionless groups, such a dimensionality reduction is possible.

For example, one place that active subspaces have been used is in uncertainty quantification involving coefficients appearing in partial differential equations. In these cases, the vector x really corresponds to a discretization of a function space (and so may be very high-dimensional indeed). But particularly for many of the equilibrium equations of mathematical physics, the property of *elliptic regularity* (known by other names as well, such as St. Venant's principle in mechanics) means that small localized changes to an equation have little impact on the broad shape of the solution.

To make this a little more concrete, consider the case of a function

$$f(E) = h((A+E)^{-1}b)$$

where h has some amount of regularity (e.g. a known Lipschitz constant), and the domain $\Omega \subset \mathbb{R}^{n \times n}$ consists of componentwise bounded entries. We can rewrite this as

$$f(E) = h((I + A^{-1}E)^{-1}\hat{u}), \quad \hat{u} = A^{-1}b;$$

and if $A^{-1}E$ has sufficiently clustered eigenvalues[1], we expect that there will be good approximations to $A^{-1}E$ in small Krylov spaces generated by $A^{-1}E$ and $\hat{u}$.

3 Explicit construction

Suppose we believe that

$$f(x) = g(Ax + \epsilon(x))$$

[1] As should happen if A and E are discretizations of a differential operator and a relatively compact perturbation, for those with some functional analysis background

and we are able to compute gradients of f. By the chain rule, we have

$$\nabla f(x) = A^T \nabla g(Ax)$$

The gradient based approach to computing the active subspace is to form

$$U \Sigma V^T = \frac{1}{\sqrt{M}} \begin{bmatrix} \nabla f(x_1) & \ldots \nabla f(x_M) \end{bmatrix},$$

or (equivalently)

$$U \Sigma^2 U^T = \frac{1}{M} \sum_{j=1}^{M} \nabla f(x_j) \nabla f(x_j)^T$$

where $x_1, \ldots, x_M$ are $\alpha k \log(n)$ independent samples from some density ρ (here k is the anticipated rank and α is an oversampling factor of about 2-10). This is a randomized approximation to

$$C = \int (\nabla f(x))(\nabla f(x))^T \rho(x) \, dx.$$

The active coordinates of interest are then computed from the leading singular vectors U_k, where one hopes for a significant spectral gap between σ_k and σ_{k+1}. The process *does* depend on the choice of the density Ω; typical choices include a uniform measure over a compact domain Ω or a Gaussian distribution chosen to cover "usual" choices of parameters.

When explicit gradients are not available, one can approximate them by computing local linear models (using finite differences or a local least squares fit).

4 Using active subspaces

When we have an active subspace, what can we do with it? Briefly, the answer might be "anything we would be happy to do with a low-dimensional function." Specifically, if we believe $f(x) = g(Ax)$ and we have computed an appropriate A, we can sample g (and $\nabla g(x)$ if we have $\nabla f(x)$ available) and use it for

- Computing a *surrogate* or *response surface* used to approximate the function at new points. These often involve kernel methods of the type we will discuss next week.

- Do sample-efficient *quadratures* (e.g. expected values and variances)
- Solve *optimization problems*, which are generally easier in lower-dimensional spaces.

The observent reader might notice that all three of these tasks (function approximation, quadrature, and optimization) are examples we gave earlier in the notes of the "curse of dimensionality," so it is not surprising that dimension reduction on the input space would be useful in each case.

5 Implicit construction

So far, we have described how we might *explicitly* compute a low-dimensional subspace that describes variation in a function. But if we believe such a structure exists, we can also attempt to use it *implicitly.* This is the idea behind the REMBO (Random Embedding for Bayesian Optimization) and REGO (Random Embeddings for Global Optimization) algorithms. In the REGO algorithm, for example, one solves a sequence of problems of the form

$$\text{minimize } f(Ay + p)$$

where $A \in \text{n} \times \text{k}$ is a random Gaussian matrix and p is a random "base point" (which might be chosen to be the same from iteration to iteration). If the problem does admit a k-dimensional active subspace, the $n-k$-dimensional level sets of f (including the affine space(s) of global optimizers) intersect any k-dimensional random subspace with probability one — though we would typically prefer to avoid the possibility of optimizing too far away from the origin, which is one reason for possibly solving with more than one random projection.

An important point about this implicit approach is that the projected subproblems do *not* necessarily require gradient information, and can be solved (for k sufficiently small) using derivative-free methods.

6 Beyond subspaces

In the past couple years, there have been a few papers that push beyond the idea of finding an active linear subspace to finding a nonlinear version (referred to by one group of authors[2] as an active manifold, and another[3] as a kernel-based active subspace). In both cases, instead of writing $f(x) \approx g(Ax)$ one writes $f(x) \approx g(\phi(x))$ where $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^k$. In the case of the kernel-based active subspace, ϕ is learned by lifting the original x vector into a higher-dimensional (redundant) feature space, then applying the active subspace technique there. We will see this idea in different guises next week. For the active manifolds idea, we refer to the paper.

One of the distinct advantages of the active subspace approach is that it's easy to analytically write down the (approximate) level sets. This is significantly more complicated in the nonlinear variants of the problem.

[2] Active Manifolds: a non-linear analogue to Active Subspaces, Bridges Gruber, Felder, Verma, Hoff, 2019

[3] Kernel-based Active Subspaces with application to CFD parametric problems using Discontinuous Galerkin method, Romor, Tezzele, Lario, Rozza, 2020