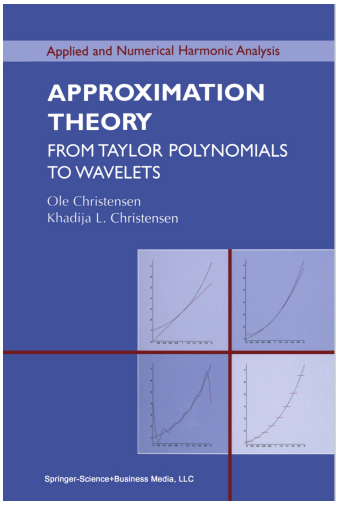
Approximation Theory: From Taylor Polynomials to Wavelets

http://freemind.pluskid.org/books/approximation-theory-from-taylor-polynomials-to-wavelets

posted on Free Mind on October 13, 2014
generated with pandoc on December 3, 2015
category: Books

tags: Approximation, Sparsity



这是由 Ole Christensen 和 Khadija L. Christensen 两人合写的一本小册子，全书加上封面一共才 166 页，我想这应该是我能迅速把它看完的一大原因。:p 当然书本身是非常有趣的，虽然我已经忘记了自己最初是在哪里碰到它的了。

全书从 Weierstrass Theorem 引出，该定理讲的是在有界闭区间上的任何一个连续函数可以由一个多项式函数进行 uniform approximation。记得很早以前跟数学系学长聊天的时候说我其实对数学很感兴趣，结果他冷不丁问我知不知道 Weierstrass' Theorem。所以我想这应该是数学里的美妙结论的一个代表吧——当然话说回来数学里的美妙定理几乎是数也数不完的。这里的美妙之处就在于我们都知道多项式很“简单”，这样一来所有问题都可以近似地化归为多项式的问题来解决，简直就跟在说“所有的问题都已经解决了”一样。

当然如果是这样的话这本书就没得写了。实际上 Weierstrass 虽然美妙，但是实际用起来却还是有很多问题。比如定理只提到了存在性却并没有直接给出一个方便的近似构造方式，再比如我们虽然可以得到一列多项式函数一致地收敛到给定的函数，但是我们并不知道收敛速度是多少，更没有一个显式的误差估计，这些在实际问题中都是非常重要的。

当然天下没有免费的午餐，想要得到更多的结论，通常就需要对需要近似的函数加更多的限制和假设。书里紧接着给的例子就是在假设 $f(x)$ 光滑（无限次可导）并且任意阶导数可以通过一个常数 C 控制住的情况下，使用泰勒级数给出了 $f(x)$ 在一个区间 I 上的一个显式近似表达式以及误差估计项：对任意整数 N 和任意 $x \in I$：

$$\left| f(x) - \sum_{n=0}^{N} \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n \right| \leq \frac{C}{(N+1)!}|x-x_0|^{N+1}$$

其中 x_0 是 I 中的一个点。这个结论其实可以很容易通过数学分析中学到的带余项的泰勒展开得到。以上的两个例子都是处理一致收敛或者说 uniform approximation 的情况，实际中我们经常还会碰到各种其他的收敛或者近似。比一致收敛弱一点的情况是 pointwise approximation，也就是说可以保证函数在每一点都收敛或者近似，但是却不能让它们达到

步调一致，这比一致收敛要求要低很多，不过 pointwise 的结果当限制到一个紧集上的时候同样可以得到 uniform 的结果。再往下还有更弱的，比如积分收敛：函数列 $g_n(x)$ 可能在任何一点 x 上都不收敛到 $f(x)$，但是当考虑两个函数之差 $|f(x)-g_n(x)|$ 的时候，可以可以保证这个差值函数的积分趋向于零。直观来说就是尽管我不能保证每个点都很老实，但是只要集体行为（积分的尺度）从总体上来说可以控制就好了。这其实是非常常见的一种收敛形式，只要换一个形式就可以看到。考虑由如下内积定义的区间 I 上的函数的 Hilbert Space：

$$\langle f,g\rangle = \int_I f(x)\overline{g(x)}\,dx$$

对应的 norm 就是

$$\|f\|^2 = \int_I |f(x)|^2\,dx$$

此时考虑 Hilbert Space 中的一列向量 $\{g_n\}$ 收敛到点 f，其实就是说

$$0 \leftarrow \lim_{n\to\infty}\|f-g_n\|^2 = \lim_{n\to\infty}\int_I |f-g_n|^2\,dx$$

也就是我们刚才所说的积分收敛，这个特殊情况也成为 least square 收敛。如果在 L_p 空间中考虑，就会对应不同形式的积分收敛。

在做一些无穷级数的准备知识介绍之后，书的内容就转向 Fourier 级数，也就是通过正弦余弦函数的组合来逼近一个周期函数，这里考虑 2π 为周期的情况。类似地这里我们会分别得到 least square 收敛和 pointwise 收敛等不同情况的结论。比如，考虑 least square 的情况，当 f 在 $[-\pi,\pi]$ 上可积时，我们可以得到 least square 收敛，并且有如下 Parsevel 恒等式：

$$\|f\|^2 = \sum_{n=-\infty}^{\infty}\left\|\hat{f}(n)\right\|^2$$

其中 $\hat{f}(n)$ 是 f 的第 n 个 Fourier 系数，从这里我们可以看到，对 f 做 least square 近似可以通过只保留 N 个最大的 Fourier 系数的方式来实现。当然计算和比较所有的 Fourier 系数是不太现实的，但是如果我们进一步的有关于 f 的光滑性的前提的话，可以非常简单地直接保留前 N 个 Fourier 系数即可——因为 f 的光滑性实际上对应了其 Fourier 系数的衰减性质。例如，如果 f 是二阶连续可导的，那么其 Fourier 系数则以 $O(1/|n|^2)$ 的速率衰减，因此当 n 很大时其系数可以忽略不计，所以根据上面的 Parsevel 恒等式直接 drop 掉所有 n 很大的对应项也不会有太大的损失。

不过 Fourier 级数有一个问题就是不能处理非周期函数。虽然 Fourier 变换作为 Fourier 级数的推广可以处理非周期函数了，但是级数求和变

成了求积分之后，从 Approximation 的角度来看就不那么直观了。于是书的后面部分开始介绍 wavelet，还提到了 wavelet 的发展历史以及其中的一些趣事。

在接近末尾的 Best N-term approximation 一小节中，书里举了一个例子来说明 wavelet 带来的比 Fourier 级数更好的稀疏性的好处：如果保留前 N 项系数的话，那么通过 Fourier 级数和通过 wavelet 的方式来进行近似其 square error 的 decay rate 都是差不多的 $1/N$，但是如果考虑 wavelet 系数的稀疏性，同样保留 N 个系数，但并不一定限制是前 N 个的话，那么近似的 square error 可以达到 2^{-N} 的 decay rate。当然无论是 wavelet 也好还是 approximation 也好，都是非常大的 topic，这里只能是浮光掠影地提到一些，书里的很多结论虽然都给出了比较严格的叙述，但是几乎都没有给出证明——否则也不会这么轻易读完吧。:D

最后，我想说的是，这是一本相当有趣的书。实际上 Approximation 的问题在 machine learning 中也是无处不在。比如在优化的时候所谓 gradient descent 或者 Newton Method 之类的，其实就是在局部对函数做一阶或者二阶的近似。再考虑 Learning 问题本身，所谓的 Empirical Risk Minimization (ERM) 其实整个就是在对 (Unknown) True Risk 进行近似之后再最小化。而这个近似的 pointwise convergence 是由传统的大数定理或者具体的 Chernoff Bound 保证的，但是为了保证 ERM 结果的有效性，pointwise convergence 是不够的，还需要该近似在整个 Learning 用的函数空间 $\mathcal{H}$ 上的 uniform convergence 性质才行，而 VC 维 之类的工具就是用来分析是否能从 pointwise convergence 得到 uniform convergence 的。

此外，Learning 本身为了保证可学习性，把搜索限制在一个函数空间 $\mathcal{H}$ 内（比如如果是在做 classification 的话我们会要求 H 的 VC 维是有限的），但是当 true function $f^\dagger$ 不在 $\mathcal{H}$ 内的时候，我们实际上会有一个 approximation error：

$$\inf_{h\in\mathcal{H}} \|f^\dagger - h\|$$

在不改变 $\mathcal{H}$ 的情况下，我们的 learning 算法再好也无法达到比这个更小的 error。所以这里存在一个 trade-off：一方面我们希望 $\mathcal{H}$ 尽量 “简单” 以达到更好的 uniform convergence 保证，另一方面我们又希望 $\mathcal{H}$ 尽量 “复杂” 以实现对 $f^\dagger$ 的更好的近似降低 approximation error。就好像 coin 的两面，前者方面的研究主要是借助于 VC 维、Rademacher Complexity 之类的工具，而后者方面的研究似乎知名度比较低一点，但是也是同样重要的内容。如果感兴趣的话 Ding-Xuan Zhou 有一本书叫做《Learning Theory: An Approximation Theory Viewpoint》似乎有讲这方面的内容。

后记：想读的书的 list 不管是专业书还是非专业书列表都在不停地增长，虽然经常都会做出努力去开始读一本书——并且是饶有兴致地，但是大部分情况都是在一两个星期之后由于其他的事情不知不觉间被打断了，很多时候往往经常是发现另一本届时觉得更加有趣或者更加急需阅

读的书。结果就是不停地从图书馆借书，但是真正读完的却还是寥寥无几。所以为了督促我自己好好地读书，我想在读完一本书之后写一点书评或者简短的总结。专业类的书就贴到这个技术博客里好了。