现代机器学习实践与经典偏差-方差权衡的协调

Mikhail Belkin[a,b,1] , Daniel Hsu[c] , Siyuan Ma[a] , and Soumik Mandal[a]

[a] 俄亥俄州立大学计算机科学与工程系，俄亥俄州哥伦布市，邮编 43210；[b] 俄亥俄州立大学统计学系，俄亥俄州哥伦布市，邮编 43210；以及[c] 哥伦比亚大学计算机科学系和数据科学研究所，纽约州纽约市，邮编 10027

由加州大学伯克利分校 Peter J. Bickel 编辑，2019 年 7 月 2 日批准（2019 年 2 月 21 日收到审稿）

机器学习领域的突破正在迅速改变科学和社会，然而我们对这一技术的基本理解却远远落后。事实上，该领域的核心原则之一--偏差-方差权衡，似乎与现代机器学习实践中观察到的方法行为相悖。偏差-方差权衡意味着模型应在欠拟合和过拟合之间取得平衡：既要足够丰富以表达数据中的潜在结构，又要足够简单以避免拟合虚假模式。然而，在现代实践中，神经网络等非常丰富的模型被训练成完全拟合（即插值）数据。从经典上讲，这种模型会被认为是过度拟合的，但它们在测试数据上却往往能获得很高的准确率。这一明显的矛盾引发了人们对机器学习的数学基础及其对实践者的意义的质疑。在本文中，我们用一条统一的性能曲线调和了经典理解和现代实践。这条 "双下降 "曲线包含了教科书上的 U 型偏差-方差权衡曲线，它显示了在插值点之外增加模型容量是如何提高性能的。我们提供的证据表明，在各种模式和数据集中都 存在双下降现象，而且这种现象无处不在，我们还提出了双下降现象出现的机制。机器学习模型的性能和结构之间的这种联系划定了经典分析的局限，对机器学习的理论和实践都有影响。

机器学习 | 偏差-方差权衡 | 神经网络

机器学习已成为科学、技术和商业领域重要应用的关键。机器学习的重点是预测问题：给定一组来自 $\mathbb{R} \times \mathbb{R}$ 的训练示例 $(x_1, y_1), \ldots, (x_n, y_n)$，从 $\mathbb{R}^d$ 到 $\mathbb{R}$，我们学习一个预测器 $\hat{h}_n : \mathbb{R}^d \to \mathbb{R}$，用来预测标签 y 的新点 x，该点在训练中从未见过。

预测因子 $\hat{h}_n$ 通常选自某个函数类别 H，如具有特定结构的神经网络，使用经验风险最小化（ERM）及其变体。在 ERM 中，预测器是一个函数 $h \in H$，经验（或培训）风险 $\frac{1}{n}\sum_{i=1}^{n} l(h(x_i), y_i)$ 它能最小化，其中 l 是损失函数，如平方损失 $l(\hat{y}, y) = (\hat{y} - y)^2$，为回归或 0-1 损失 $l(\hat{y}, y) = 1_{\{\hat{y} \neq y\}}$ 进行分类。

机器学习的目标是找到在训练中未见的新数据上表现良好的 $\hat{h}_n$。要研究在新数据上的性能（称为泛化），我们通常假设训练示例是从 $\mathbb{R}^d \times \mathbb{R}$ 上的概率分布 P 中随机抽样的，并在新的测试示例 (x, y) 上评估 $\hat{h}_n$。

即经验风险较大），因此对新数据的预测能力较差。2) 如果经验风险过大，经验风险最小化器可能会过度拟合训练数据中的虚假模式，从而导致对新示例的预测准确性较差（经验风险较小，但真实风险较大）。

经典的思路是在欠拟合和过拟合之间找到 "甜蜜点"。对函数类容量的控制可以是显性的，通过选择（如选择神经网络架构），也可以是隐性的，使用正则化（如早期停止）。图 *1A* 所示的经典 U 型风险曲线概括了这一点，该曲线被广泛用于指导模型选择，甚至被认为可 以 描述人类决策的各个方面 (3)。该曲线的教科书推论是 "训练误差为零的模型与训练数据拟合过度，通常泛化效果较差"（参考文献 2，第 221 页），这一观点至今仍被广泛接受。

不过，从业人员通常会使用现代机器学习方法，如大型神经网络和其他非线性预测器，它们的训练风险非常低或为零。尽管这些预测器具有较高的函数类容量，而且与训练数据的拟合度近乎完美，但它们往往能对新数据做出非常准确的预测。事实上，这种行为指导了深度学习中选择神经网络架构的最佳实践，具体来说，网络应该足够大，以允许对训练数据进行不费力的零损失训练（称为插值）。(4).此外，作为对偏差-方差权衡哲学的直接挑战，最近的经验证据表明，神经

重要意义

虽然机器学习和人工智能领域的突破正在改变社会，但我们的基本认识却落后了。传统观点认为，应避免将模态与训练数据完全匹配，因为这会导致在未见数据上表现不佳。然而，功能强大的模态分类器在训练中往往具有近乎完美的拟合效果，这种脱节现象引发了近期关于理论是否能提供实用见解的深入研究和争论。在这项工作中，我们展示了经典理论和现代实践如何在一条统一的性能曲线中得到调和，并提出了其产生的内在机制。我们相信，这种将学习架构的结构和性能联系起来的前所未知的模式将有助于设计和理解学习算法。

机器学习的传统智慧建议，在偏差-方差权衡的基础上，通

过平衡欠拟合和过拟合来控制函数类的容量（参见参考文献 1 和 2）：1) 如果 H 太小，H 中的所有预测因子都可能对训练结果拟合不足。

作者贡献：M.B.、D.H.、S. Ma 和 S. Mandal 设计研究、进行研究、分析数据并撰写论文。

作者声明无利益冲突。本文为 PNAS 直接投稿。

[1]收信人电子邮件：mbelkin@cse.ohio-state.edu.

本文的辅助信息见 www.pnas.org/lookup/suppl/doi:10. 1073/pnas.1903070116/-/DCSupplemental。

2019 年 7 月 24 日在线发表。

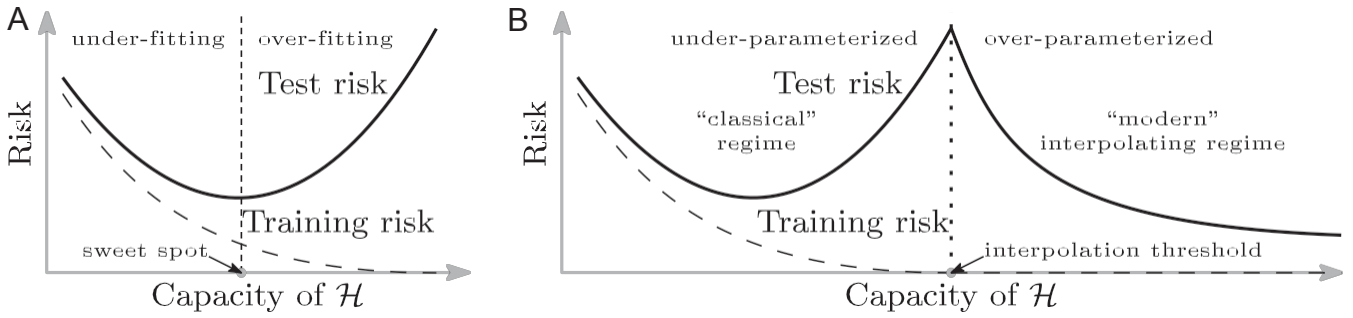


图 1.训练风险曲线（虚线）和测试风险曲线（实线）。(*A*) 由偏差-方差权衡产生的经典 U 型风险曲线。(*B*) *双*下降风险曲线，它包含了 U 型风险曲线（即 "经典 "机制）和使用高容量函数类时观察到的行为（即 "现代 "插值机制），由插值阈值分隔。插值阈值右侧的预测因子的训练风险为零。

即使在训练数据被高水平噪声干扰的情况下，经过内插训练的网络和核机器也能获得接近最优的测试结果（5、6）。

这项工作的主要发现是，未见数据的性能如何取决于模型能力及其产生的机制。图 *1B* 所示的 "双下降 "风险曲线概括了这一依赖关系，包括神经网络在内的重要模型类别和一系列数据集都证明了这一点。该曲线将图 *1A* 中经典的 U 型风险曲线延伸到插值点之外，从而将其包含在内。

当函数类容量低于 "插值阈值 "时，学习到的预测结果会呈现图 *1A* 中的经典 U 型曲线。(在本文中，函数类容量是指在函数类中指定一个函数所需的参数数量）。U 形曲线的底部是一个甜点，它兼顾了与训练数据的拟合和过度拟合的可能性：在甜点的左边，预测因子拟合不足，而在右边，预测因子拟合过度。当我们将函数类容量提高到足够高时（例如，通过增加特征数量或神经网络架构的大小），学习到的预测结果就会达到（接近）"甜蜜点"。
与训练数据完美拟合，即插值。虽然在插值阈值下获得的预测结果通常具有较高的风险，但我们证明，增加函数
超过这一点的分级能力会导致风险下降，通常会低于 "经典 "制度中甜蜜点所达到的风险。

插值阈值右侧的所有已学预测因子都完全符合训练数据，经验风险为零。那么，为什么有些预测因子--尤其是来自更丰富功能类别的预测因子--的测试风险会低于其他预测因子呢？答案是，函数类别的容量并不一定反映预测器与当前问题所需的归纳偏差的匹配程度。对于我们所考虑的学习问题（一系列真实世界数据集以及合成数据），似乎合适的归纳偏差是函数的规则性或平滑性，这是由一定的函数空间规范来衡量的。选择与观测数据完全吻合的最平滑函数是奥卡姆剃刀的一种形式，与观测数据相匹配的最简单的解释应优先考虑（参见奥卡姆剃刀）。

如果能保留更多与数据兼容的候选预测因子，我们就能找到常模更小、因而 "更简单 "的插值函数。因此，增加函数类容量可以提高分类器的性能。

边际理论（7、9、10）中也考虑了相关的观点，即一个

但这并不适用于回归，也不能预测超出插值阈值的第二次下降。最近，人们逐渐认识到，某些插值预测器（不是基于 ERM）确实可以证明是统计最优或接近最优的（11，12），这与我们在插值机制中的经验观察是一致的。

在本文的其余部分，我们将讨论双曲线的经验证据及其产生机制，最后提出一些看法和结束语。

神经网络

在本节中，我们将结合神经网络讨论双下降风险曲线。

随机傅立叶特征我们首先考虑一类流行的非线性参数模型--随机傅立叶特征（RFF）（13），它可以看作是一类在第一层具有固定权重的双层神经网络。具有 N 个（复值）参数的 RFF 模型族 H_N 由函数 $h : \mathbb{R}^d \to \mathbb{C}$ 组成，其形式为

$$h(x) = \sum_{k=1}^{N} a_k \varphi(x ; v_k) \text{ 其中 } \varphi(x ; v) := e^{\sqrt{-1}\langle v_k, x\rangle},$$

向量 $v_1, \ldots, v_N$ 从 $\mathbb{R}^d$ 中的标准正态分布中独立采样。(我们认为 H_N 我们将 N 视为一类具有 $2N$ 个*实值*参数的实值函数，分别取实部和虚部）。注意 H_N 是一个随机函数类，但当 $N \to \infty$ 时，函数类会越来越接近与高斯核相对应的重现核希尔伯特空间（RKHS），表示为 H_∞。虽然可以直接使用[如核机器 (14)]，但随机类 当样本量 n 较大但参数数 N 较小时，使用随机类 N 在计算上很有吸引力。与 n 相比

我们使用 H_N 的学习过程如下。给定来自 $\mathbb{R} \times \mathbb{R}$ 的数据 $(x_1, y_1), \ldots, (x_n, y_n)$ from $\mathbb{R}^d \times \mathbb{R}$，我们通过带有 squared loss 的 ERM 找到预测器 $h_N \in H_N$。也就是说，我们要最小化经验值。

风险目标 $\frac{1}{n}\sum_{i=1}^{n}(h(x_i) - y_i)^2$ over all functions $h \in H_N$.

较大的函数类可能允许发现一个具有较大边际的分类器。虽然边际理论可用于研究分类，但它并不

当最小值不唯一时（$N cn$ 时总是这种情况），我们选择系数 $(a1, \ldots, aN)$ 最小的 $l2$ 准则。选择此规范的目的是为了近似 RKHS 规范 $\|h\|_{\infty}$，对于 $_N$ 中的任意函数，RKHS 规范通常很难计算。对于有多个输出的问题（如多类分类），我们使用有向量值输出的函数和每个输出的损失平方之和。

在图 2 中，我们展示了在名为 MNIST 的流行手写数字数据集子集上使用 N 学习的预测器的测试风险。图 2 还显示了函数系数的 l_2 准则以及训练风险。我们可以看到，对于较小的 N 值，测试风险呈现出与偏差-方差权衡相一致的经典 U 型曲线，峰值出现在插值阈值 $N = n$ 时。$n \log n$ 才能获得良好的测试风险保证 (15)。

用最少的参数（$N = n$ 个随机特征）实现插值的模型类得到的预测结果最不准确（事实上，它对分类没有预测能力），但随着特征数量的增加，超过 n 个特征后，预测结果的准确性显著提高，超过了 U 型曲线底部对应的预测结果。该图还显示，由 (在任何有限 N 的情况下，都优于来自 N 的预测器。

What structural mechanisms account for the double-descent shape? When the number of features is much smaller than the sample size, $N \quad n$, classical statistical arguments imply that the training risk is close to the test risk. Thus, for small N, adding more features yields improvements in both the training and the test risks. However, as the number of features approaches n (the interpolation threshold), features not present or only weakly present in the data are forced to fit the training data nearly perfectly. This results in classical overfitting as predicted by the bias–variance trade-off and prominently manifested at the peak of the curve, where the fit becomes exact.

在插值阈值右侧，所有函数类别都足够丰富，以实现零训练风险。对于我们所考虑的 N 类，对于任何有限的 N，我们都无法保证与训练数据一致的最规则、最小规范预测器（即 $h_{n,}$，在 ）包含在 N 类中。但是，增加 N 可以让我们逐步构建出更好的的近似值。因此，我们希望在插值阈值处学习到的预测器具有最大的规范，并且 $h_{n,N}$ 的规范随着 N 的增加而单调递减，从而解释曲线的第二个下降段。这就是我们在图 2 中观察到的情况，在任何有限 N 的情况下，$h_{n,\infty}$ 的确比所有 $h_{n,N}$ 都更准确。事实证明，在 MNIST 及其他真实和合成数据集上，偏好小规范插值预测因子是一个强大的诱导偏差（6）。对于无噪声数据，我们将在 *SI 附录*中以数学方式精确说明这一点。

*SI 附录*中还提供了使用其他数据集进行相同双下降行为的其他经验证据。例如，我们展示了整流线性单元（ReLU）随机特征模型的双下降，这是一类与 RFF 设置类似的 ReLU 神经网络。我们还描述了一个简单的合成模型，该模型可视为 RFF 模型的一维版本，我们在该模型中观察到了相同的双下降行为。

神经网络和反向传播。在一般的多层神经网络（RFF 或 ReLU 随机特征模型之外）中，学习算法将调整所有权重以适应训练数据，通常使用随机梯度下降（SGD），并通过反向传播来计算部分导数。这种灵活性提高了神经网络的表征能力，但也使 ERM 通常更难实施。尽管如此，如图 3 所示，我们观察到，增加全连接双层神经网络的参数数会导致与 RFF 模型相似的风险曲线。测试风险在超过互补阈值后得到改善，这与神经网络常用训练算法的 "小规范 "归纳偏差（16、17）的猜想是一致的。我们注意到，文献 18-21 以前也观察到了神经网络从参数不足到参数过剩的这种转变。18-21 特别是，参考文献 21 将其与粒子系统中的 "干扰 "物理现象联系起来。

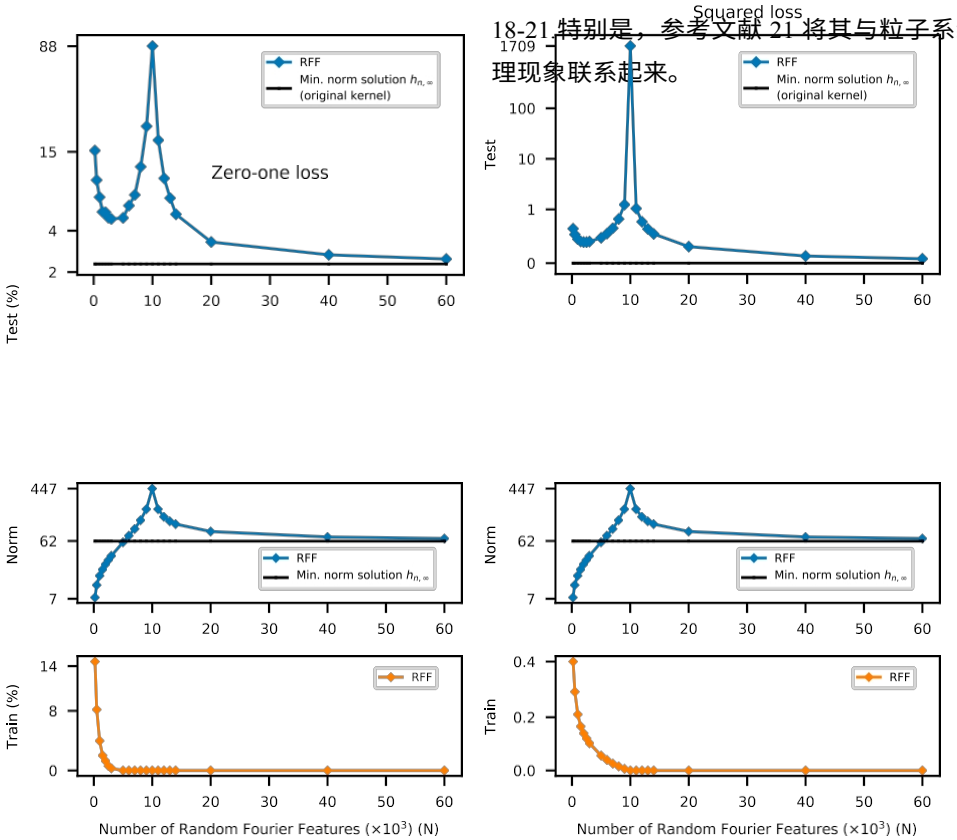


图 2.RFF 模型在 MNIST 上的双下降风险曲线。图中显示的是在 MNIST 子集（$n = 10^4$ ， 10 个类别）上学习的 RFF 模型预测器 $h_{n,N}$ 的测试风险（对数标度）、系数 $\mathcal{A}_2$ 规范（对数标度）和训练风险。插值阈值在 $N = 10^4$ 时达到。

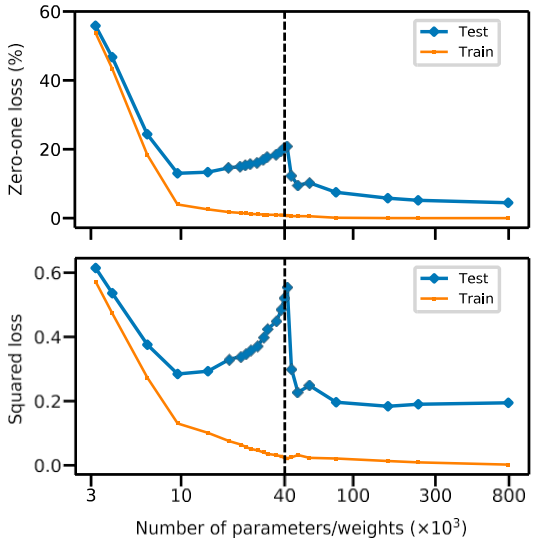


图 3. 全连接神经网络在 MNIST 上的双下降风险曲线。图中显示的是在 MNIST 子集（$n = 4 \cdot 10^3$，$d = 784$，$K = 10$ 个类别）上学习的单层 H 隐藏单元网络的训练和测试风险。参数数为 $(d + 1) \cdot H + (H + 1) \cdot K$。插值阈值（黑色虚线）为 $n \cdot K$。

The computational complexity of ERM with neural networks makes the double-descent risk curve difficult to observe. Indeed, in the classical underparameterized regime ($N \quad n$), the nonconvexity of the ERM optimization problem causes the behavior of local search-based heuristics, like SGD, to be highly sensitive to their initialization. Thus, if only suboptimal solutions are found for the ERM optimization problems, increasing the size of a neural network architecture may not always lead to a corresponding decrease in the training risk. This suboptimal behavior can lead to high variability in both the training and test risks that masks the double-descent curve.

使用参数数量极多的神经网络是很常见的 (22)。但是，要实现单个输出（回归或两类分类）的插值，预计所需的参数至少与数据点的数量相当。更重要的是，如果预测问题有一个以上的输出（如多类分类），那么所需的参数数量应该乘以输出的数量。图 3 所示的神经网络就是这 种 情况。因此，举例来说，像 ImageNet (23) 这样大的数据集（有 10 个^6 示例和 10 个^3 类别），可能需要有 10 个^9 参数的网络才能实现插值；这比 ImageNet (22) 的许多神经网络模型都要大。在这种情况下，"U "形风险曲线的类学机制更适合理解泛化。对于较小的数据集来说，这些大型神经网络会牢牢地处于过参数化状态，而简单地进行训练以获得零训练风险往往会带来良好的测试性能（5）。

有关神经网络的其他结果见 *SI 附录*。

决策树和集合方法

除了神经网络之外，双下降风险曲线在其他预测方法中是否也有体现？我们给出的经验证据表明，在插值阈值之前和之后，决策树和随机森林的提升所探索的函数族也表现出与神经网络类似的泛化行为。

最近，参考文献 24 在插值分类中对 AdaBoost 和随机森林进行了研究。在部分

特别是，他们提供的经验证据表明，当 AdaBoost 和随机森林与最大（插值）决策树一起使用时，拟合方法的灵活性产生的插值预测结果比刚性、非插值方法（如 AdaBoost 或浅树随机森林）产生的预测结果更能抵御训练数据中的噪声。据说这反过来会产生更好的泛化效果。对（近似）插值树进行平均，可确保生成的函数比任何单个树都要平滑，这与符合许多实际问题的归纳偏差是一致的。

我们可以从双梯度风险曲线的内涵中理解这些灵活的拟合方法。请注意，决策树的大小（由树叶数量控制）是参数化函数类容量的一种自然方法：只有 2 片叶子的树对应于具有轴对齐边界的 2 片常数函数，而有 n 片叶子的树可以插入 n 个训练实例。一个经典的观察结果是，当以这种方式考虑类容量时，U 型偏差-方差权衡曲线会在许多问题中表现出来 (2)。（在很多情况下，少于 n 个样本就能达到插值阈值，但 n 显然是一个上限）。为了进一步扩大函数类，我们考虑了多个插值函数的集合（平均值）。

树。因此，在插值阈值之外，我们使用此类树的数量来表示类容量。当我们把风险曲线看作是以这种混合方式定义的类容量的函数时，我们会看到双下降曲线出现，就像神经网络一样（图 4 和 *SI 附录*）。我们在使用另一种流行的集合方法 $_{L2}$ boosting 时也观察到了类似的现象（26, 27）；结果见 *SI 附录*。

结束语

本文介绍的双下降风险曲线是对偏差-方差权衡所预测的 U 形曲线和现代机器学习实践中使用的丰富模型的观察行为的重新整合。双升风险曲线出现的假设机制基于常见的归纳偏差，因此可以解释双升风险曲线在机器学习应用中的出现（以及我们认为的普遍性）。

最后，我们提出一些结束语。

历史缺失。由于一些文化和实践上的障碍，双下降行为可能在历史上被忽视了。观察双下降曲线需要一个具有任意复杂度函数的空间参数族。经典统计学中广泛研究的线性设置通常假定有一个小的、固定的特征集，因此有固定的拟合能力。更丰富的函数族通常用于非参数统计中，平滑和正则化几乎总是在非参数统计中使用 (28)。各种形式的正则化既能防止插值，又能改变函数类的有效容量，从而减弱或掩盖插值峰值。

The RFF model is a popular and flexible parametric family. However, these models were originally proposed as a computationally favorable alternative to kernel machines. This computational advantage over traditional kernel methods holds only for $N \quad n$, and hence models at or beyond the interpolation threshold are typically not considered.

一般多层神经网络的情况略有不同，涉及的问题也更多。由于 ERM 优化问题的非凸性，经典的欠参数化机制中的解对初始化非常敏感。

* 这些树的训练方法与随机森林中提出的方法相同，只是没有引导重采样。这与参考文献 25 中的 PERT 方法类似。

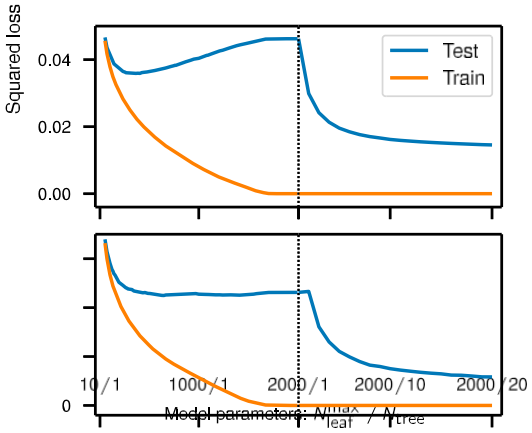


图 4. 随机森林在 MNIST 上的双下降风险曲线。在 MNIST 子集（$n = 10^4$，10 个类别）上训练的随机森林的模型复杂度不断增加，双下降风险曲线也随之增加。其复杂度由树的数量 N_{tree} 和每棵树允许的最大叶子数 $N_{叶子}^{max}$ 控制。

此外，正如我们所见，插值阈值处的峰值是在一个较窄的参数范围内观察到的。如果对参数空间进行取样而忽略了这一范围，可能会导致错误的印象，即增加网络规模就能提高性能。最后，在实践中，一旦（估计的）测试风险未能改善，神经网络的训练通常会立即停止。如上所述，这种过早停止训练的做法会产生强烈的正则化效应，导致难以观察到插值峰值。

归纳偏差。在本文中，我们讨论了几种选择插值解的方法。对于随机傅立叶特征，可以通过特征空间中的最小矩规范线性回归明确构建解。当特征数量趋于无穷大时，它们就会接近再现核希尔伯特空间中的最小函数规范解，即在插值约束条件下最大化函数平滑性的解。对于神经网络而言，诱导偏差归因于所使用的特定训练程序，即通常是 SGD。当除最后一层以外的所有网络层都固定时（如 RFF 模型），初始化为零的 SGD 也会趋向于最小规范解。虽然对更一般的神经网络来说，SGD 的行为还不完全清楚，但有大量经验和一些理论证据（如参考文献 16）表明，存在类似的最小规范归纳偏差。与平均化相关的另一种归纳偏差也用于随机森林。对可能不平滑的内插树进行平均，可以得到平滑度更高的内插解决方案；这种平均解决方案比任何单个内插树的性能都要好。

值得注意的是，对于核机器，这三种方法都能得到相同的最小规范解。事实上，最小规范间分类器 $_{hn,\infty}$ 可以通过显式规范最小化（求解显式线性方程组）、SGD 或高斯过程轨迹平均（计算后验均值 (29)）直接获得。

优化和实际考虑。在我们的实验中，适当选择的 "现代 "模型在测试集上的表现通常优于最优的经典模型。但过参数化模型的另一个重要实际优势在于优化。越来越多的人认识到，大型模型 "容易 "优化，因为局部方法（如 SGD）会收敛到超参数化状态下训练风险的全局最小值（如参考文献 30）。因此，大型内插模型可以具有较低的测试风险，同时又易于优化，特别是 SGD（31）。插值峰左侧的模型可能与右侧的模型在优化特性上有本质区别，这种区别具有重要的实际意义。

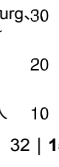
展望。经典的 U 型偏差-方差权衡曲线塑造了我们对模型选择和学习算法在实践中定向应用的看法。本研究对模型性能的理解划定了经典分析的界限，并为研究和比较机器学习中经典和现代机制的计算、统计和数学特性开辟了新的研究方向。我们希望这一观点反过来能帮助实践者选择最佳性能的模型和算法。

致谢。M.B. 得到了国家自然科学基金 RI-1815697 号资助。D.H. 获得国家自然科学基金 CCF-1740833 和斯隆研究奖学金的资助。

1. S.Geman, E. Bienenstock, R. Doursat, Neural networks and the bias/variance dilemma. *Neural Comput.***4**, 1-58 (1992).
2. T.Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning* (Springer, 2001), vol. 1.
3. G. Gigerenzer、H. Brighton，《启发式智人》（Homo heuristicus）：为什么有偏见的头脑能做出更好的推论？*Top.1, 107-143 (2009).***1**, 107-143 (2009).
4. R.Salakhutdinov, Deep learning tutorial at the Simons Institute, Berke- ley. https://simons.berkeley.edu/talks/ruslan-salakhutdinov-01-26-2017-1.Accessed 28 December 2018 (2017).
5. C.Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, "Understanding deep learning requires rethinking generalization" in *Proceedings of International Conference on Learning Representations* (International Conference on Learning Representations, 2017).
6. M.Belkin, S. Ma, S. Mandal, "To understand deep learning we need to under- stand kernel learning" in *Proceedings of the 35th International Conference on Machine Learning,* J. Dy, A. Krause, Eds.(机器学习研究论文集》，瑞典斯德哥尔摩，2018 年）
，第 80 卷，第 541-549 页。
7. V.V. N. Vapnik，《*统计学习理论的本质*》（施普林格出版社，1995 年）。
8. A.Blumer, A. Ehrenfeucht, D. Haussler, M. K. Warmuth, Occam's razor.*Inf.24, 377-380 (1987).***24**, 377-380 (1987).
9. P.L. Bartlett, The sample complexity of pattern classification with neural networks：权重的大小比网络的大小更重要。*IEEE Trans.Inf.Theory* **44**, 525-536 (1998).
10. R.E. Schapire, Y. Freund, P. Bartlett, W. S. Lee, Boosting the margin：投票方法有效性的新解释。*Ann.***26**, 1651-1686 (1998).
11. M.Belkin, D. Hsu, P. Mitra, "Overfitting or perfect fitting?插值分类和回归规则的风险边界"，《*神经信息研究进展*》（*Advances in Neural Information*），2009 年。

处理系统》，S. Bengio、H. Wallach、H. Larochelle、K. Grauman、N. Cesa-Bianchi、R.Garnett, Eds.(Curran Associates, Inc., 2018），第 2300-2311 页。

12. M.M.贝尔金、A.拉赫林、A.B.齐巴科夫：《数据*插值*是否与统计最优性相矛盾？M.Sugiyama, Eds.(机器学习研究论文集，2019 年》，第 89 卷，第 1611-1619 页。
13. A.Rahimi, B. Recht, "Random features for large-scale kernel machines" in *Advances in Neural Information Processing Systems*, J. C. Platt, D. Koller, Y. Singer, S. T. Roweis, Eds.(Curran Associates, Inc., 2008), pp.
14. B.E. Boser, I. M. Guyon, V. N. Vapnik, "A training algorithm for optimal margin clas- sifiers" in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (ACM, New York, NY, 1992), pp.
15. A.Rudi, L. Rosasco, "Generalization properties of learning with random features" in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg、S.Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, Eds.(Curran Associates, Inc., New York, NY, 2017），第 3215-3225 页。
16. S.Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, N. Srebro, "Implicit regularization in matrix factorization" in *Advances in Neural Information Processing Systems,* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan、R.Garnett, Eds.(Curran Associates, Inc., New York, NY, 2017），第 6151-6159 页。
17. Y.Li, T. Ma, H. Zhang, "Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations" in *Proceedings of the 31st Conference On Learning Theory,* S. Bubeck, V. Perchet, P. Rigollet, Eds. (Proceedings of the Machine Learning Research, 2018, vol. 75, pp.(机器学习研究论文集》，2018 年），第 75 卷，第 2-47 页。
18. S.Bo¨ s, M. Opper, "Dynamics of training" in *Advances in Neural Information Processing Systems, M.* C. Mozer, M. I. Jordan, T. Petsche, Eds.(麻省理工学院出版社，1997 年），第 141-147 页。



19. M.M. S. Advani, A. M. Saxe, 神经网络泛化误差的高维动态.ArXiv:1710.03667（2017 年 10 月 10 日）.
20. B.尼尔 *等人*，神经网络中偏差-方差权衡的现代视角。A modern take on the bias-variance tradeoff in neural networks. ArXiv:1810.08591 (25 January 2019).
21. S.Spigler *等人* ，从欠参数化到过参数化的干扰过渡影响损失景观和泛化.ArXiv:1810.09665（2018 年 10 月 22 日）.
22. A.Canziani, A. Paszke, E. Culurciello, 面向实际应用的深度神经网络模型分析.ArXiv:1605.07678（2016 年 5 月 24 日）.
23. O.Russakovsky *等人*，ImageNet 大规模视觉识别挑战。*Int.J. Comput.Vis.***115**, 211-252 (2015).
24. A.J. Wyner、M. Olson、J. Bleich、D. Mease，解释作为插值分类器的 adaboost 和随机森林的成功。*J. Mach.Learn.***18**, 1-33 (2017).
25. A.Cutler, G. Zhao, Pert-perfect random tree ensembles.*Comput.***33**, 490-497 (2001).
26. J. H. Friedman, Greedy function approximation：A gradient boosting machine.*Ann.***29**, 1189-1232 (2001).
27. P.Bu¨ hlmann, B. Yu, Boosting with the l_2 loss：Regression and classification.*J. Am.J. Am. Stat.***98**, 324-339 (2003).
28. L.Wasserman, *All of Nonparametric Statistics* (Springer, 2006).
29. C.E. Rasmussen, "Gaussian processes in machine learning" in *Advanced Lectures on Machine Learning,* O. Bousquet, U. von Luxburg, G. Ra¨ tsch, Eds.(Springer, Berlin, 2004), pp.
30. M.Soltanolkotabi, A. Javanmard, J. D. Lee, Theoretical insights into the optimization landscape of the over-parameterized shallow neural networks.*IEEE Trans.Inf.Theory* **65**, 742-769 (2018).
31. S.Ma, R. Bassily, M. Belkin, "The power of interpolation：Understanding the effectiveness of SGD in modern over-parametrized learning" in Proceedings of the *35th International Conference on Machine Learning, Proceedings of Machine Learning Research*, J. Dy, A. Krause, eds. (PMLR, Stockholmsma¨ ssan, Stockholm, Sweden, 2018), vol. 80, pp.