机器学习研究学报18（2018）1-52提交11/16;修订12/17;发表于4月18日

超带：基于匪盗的方法
Hyperparameter Optimization

Lisha Li LISHAL@CS.CMU.EDU
Carnegie Mellon University，匹兹堡，PA 15213 Kevin Jamieson Jamieson@cs.washington.edu

华盛顿大学，西雅图，WA 98195 Giulia desalvo giuliad@google.com

谷歌研究，纽约，NY 10011 AFShin Rostamizadeh Rostami@Google.com

谷歌研究，纽约，纽约10011 ameettalwalkar talwalkar@cmu.edu

卡内基梅隆大学，匹兹堡，PA 15213确定

Editor: Nando de Freitas

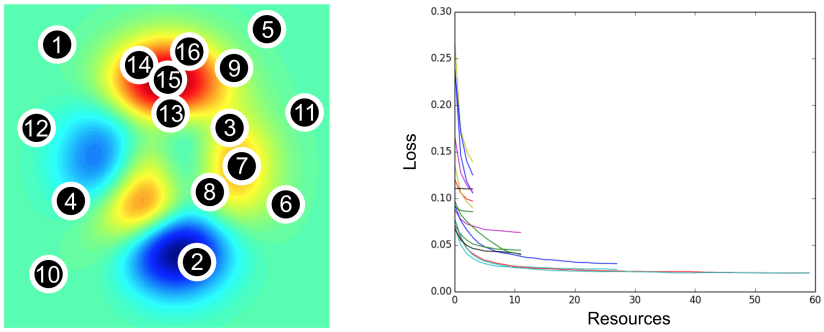
Abstract

机器学习算法的性能均批判性地识别出良好的超参数集。虽然最近的方法使用贝叶斯优化来自适应地选择配置，但我们专注于通过自适应资源分配和早期停止加快随机搜索。我们将HyperParameter优化制定为纯探索非随机无限无限武装强盗问题，其中迭代，数据样本或特征等预定义资源被分配给随机采样的配置。我们介绍了一种新颖的算法，超带，用于该框架并分析其理论属性，提供了几种理想的保证。此外，我们在套件的超参数优化问题上比较具有流行贝叶斯优化方法的超额带。我们观察到超接管可以在我们的竞争对手上设置多个级别加速，在各种深度学习和基于内核的学习问题上设置。

关键词：封路数据计优化，型号选择，无限武装匪徒，在线
optimization, deep learning

1. Introduction

近年来，机器学习模型在交错的计算成本的价格中爆炸了复杂性和表现性。此外，难以通过标准优化技术来设定与这些模型相关的越来越多的调谐参数。这些“超级参数”被输入到机器学习算法，该算法控制算法的性能如何推广到新的，看不见的数据;超参数的例子包括影响模型架构，正则化量和学习率的那些。预测模型的质量批判性地取决于其封路数据配置，但是，这些超参数如何互相交互以影响所得模型。



(a) Configuration Selection (b) Configuration Evaluation

图1：（a）Heatmap在二维搜索空间上显示验证错误

红色对应于验证误差较低的区域。配置选择方法自适应地选择要列车的新配置，以顺序方式按照数字所示进行。（b）绘图显示验证误差作为分配给每个配置的资源的函数（图绘图中的每行）。配置评估方法将更多资源分配给有希望的配置。

因此，从业者往往默认为蛮力方法，如随机搜索和网格搜索（Bergstra和Bengio，2012）。

努力开发更有效的搜索方法，封锁的问题

优化最近由贝叶斯优化方法主导（Snoek等，2012; Hutter等，2011;
Bergstra等，2011），专注于优化HyperParameter配置选择。这些方法旨在通过以自适应方式选择配置，比标准基线更快地识别良好的配置;见图1（a）。现有的经验证据表明，这些方法优于随机搜索（Thornton等，2013;
Eggensperger等，2013;
Snoek等，2015b）。然而，这些方法解决了同时拟合和优化具有未知光滑度的高维，非凸起功能的基本上具有挑战性问题，并且可能嘈杂的评估。

超级计数器优化的正交方法侧重于加速配置 -

比例评估;见图1（b）。这些方法在计算中是自适应的，在快速消除较差的时，将更多资源分配给承诺的封路数据配置。资源可以采取各种形式，包括培训集的大小，功能数量或迭代算法的迭代次数。通过自适应地分配资源，这些方法旨在检查比统一训练所有配置的方法的级别的级别更多的封路数据配置，从而快速识别良好的超参数。虽然存在将贝叶斯优化与自适应资源分配相结合的方法（Swersky等，2014,2014; Domhan等，2015;
Klein等，

2017A），我们专注于加快随机搜索，因为它提供了一个简单且理论上的原理发射点（Bergstra和Bengio，2012）.1

我们通过制定QuandExameter开发一种新颖的配置评估方法

优化作为纯粹探索的自适应资源分配问题，解决了如何在随机采样的超级计数器配置中分配资源.2我们的程序超细依赖于分配资源的原则性的早期停止策略，允许它更好地评估幅度的级别比贝叶斯优化方法等黑盒程序。超细是一种通用技术，这是与现有的配置评估方法不同的假设（Domhan等，2015;
Swersky等，2014;Gy ū orgy和Kocsis，2011; Agarwal等，2011; Sparks等人。，2015; jamieson和talwalkar，2015）。

我们的理论分析表明超接管适应未知的能力

收敛速率和验证损失的行为作为普遍参数的函数。此外，超频线比流行的贝叶斯优化算法快5×至30倍，基于各种深度学习和基于内核的学习问题。这项工作的理论贡献是引入了非随机设置中的纯探索，无限武装的强盗问题，超接管是一种解决方案。当超接管应用于特殊情况随机设置时，我们表明该算法在无限（Carpentier和Valko，2015）和有限K武装强盗设置中的已知下限的日志因子内（Kaufmann等，2015）。

本文的结构如下。第2节总结了两个领域的相关工作：

（1）封立参数优化，和（2）纯粹勘探强盗问题。第3节介绍超带，并通过详细示例为算法提供直觉。在第4节中，我们提出了广泛的经验结果与最先进的竞争对手比较的超额带。第5节框架框架作为无限武装的强盗问题，并总结了超带的理论结果。最后，第6节讨论了超频的可能扩展。

2. Related Work

在第1节中，我们简要讨论了普遍存在计优化文献中的相关工作。在这里，我们提供了更全面的工作覆盖范围，并且还总结了对匪徒问题的显着相关工作。

2.1 Hyperparameter Optimization

贝叶斯优化技术模拟了配置在评估度量Y（即，测试精度）上的配置性能的条件概率p（y | λ），给定一组高达参数λ。

1.无论平滑度还是，随机搜索都将渐近地收敛到最佳配置。

通过简单的覆盖参数进行优化功能的结构。虽然随机搜索的收敛速度取决于光滑度，并且在搜索空间中的尺寸的数量中是指数，但对于没有额外的结构假设的贝叶斯优化方法也是如此（Kandasamy等，2015）。

2.李等人出现了这项工作的初步版本。（2017）。我们将前一篇论文延长了一个

超频彻底的理论分析;具有应用于随机无限武装匪的算法的无限地平线版本;超接带的额外直觉和讨论，以方便在实践中使用;和117多级模型选择任务的集合的其他结果。

基于序列模型的算法配置（SMAC），树结构Parzen估计器（TPE）和留棉是三种良好的方法（Feurer等，2014）。SMAC使用随机森林来模拟p（y | λ）作为高斯分布（Hutter等，2011）。TPE是一种基于树结构偏移密度估计器的非标准贝叶斯优化算法（Bergstra等，2011）。最后，Spearmint使用高斯进程（GP）来模拟p（y | λ），并通过GP的HyperParameters执行切片采样（Snoek等，2012）。

以前的工作比较了这些贝叶斯搜索者的相对表现（Thornton等，2013年; EggenSperger等人。，2013; Bergstra等，2011年; Snoek等人。，2012; Feurer等人。，2014年，2015）。通过EggenSperger等，对这三种方法进行了广泛的调查。（2013）介绍了一个名为HPOLIB的超参数优化的基准库，我们用于我们的实验。 Bergstra等。（2011）和Thornton等。（2013）显示贝叶斯优化方法在少数基准任务上经验上垂直随机搜索。然而，对于高维问题，标准贝叶斯优化方法类似于随机搜索（Wang等人，2013）。最近专门为高维问题设计的方法假设问题的较低有效维度（Wang等，2013）或目标功能的添加剂分解（Kandasamy等，2015）。但是，可以预期，这些方法的性能对所需输入敏感;即有效维度（Wang等，2013）或添加剂组分的数量（Kandasamy等，2015）。

使用信心绑定的强盗设置也研究了高斯进程收购函数（GP−UCB），相关的Sublinear遗憾界限（Srinivas等，2010;Gr ¨ unew ¨ alder等，2010）。Wang等人。（2016）通过删除控制勘探和剥削的参数来改善GP−UCB。Contal等人。（2014）通过使用相互信息采集函数来派生比GP−UCB的更严格的遗憾。然而，van der Vaart和Van Zanten（2011）表明，GPS的学习率对先前通过前面的例子的定义对先前的示例敏感，其中学习率在观察的数量中从多项式到对数劣化。另外，没有GP的协方差矩阵的结构假设，拟合后部是O（n3）（Wilson等，2015）。因此，Snoek等人。（2015A）和Springenberg等人。（2016）采用贝叶斯神经网络提出，用n线性缩放，以模拟后部。

自适应配置评估不是一个新的想法。Maron和Moore（1997）和Mnih和audibert（2008）考虑了一个设置，其中训练时间相对便宜（例如，k离邻邻分类）和大验证集的评估通过评估验证集的增加的子集来加速，停止提前配置表现不佳。由于验证集的子集提供了对其预期性能的无偏见估计，这是多武装匪徒随机最佳武器识别问题的一个例子（参见Jamieson和2014年的工作，用于简要调查）。

相比之下，我们地址评估时间相对便宜的设置目标是通过在完整验证集上评估部分训练的模型来早期停止长期运行的培训程序。此设置中的先前方法需要强烈的假设或使用启发式来执行自适应资源分配。gy ¨ orgy和kocsis（2011）和agarwal等。（2011）对培训算法的收敛行为进行的参数假设，在这些假设下提供理论性能保证。不幸的是，这些假设通常很难验证，并且经验性能可以

当他们被侵犯时会受到彻底的痛苦。Krueger等。（2015）基于顺序分析提出了一种启发式，以确定用于越来越多的数据子集的训练配置的停止时间。然而，这种方法的理论正确性和经验性能高度依赖于用户定义的“安全区”。

几种混合方法结合了自适应配置选择和评估

还介绍（Swersky等，2014年，2014年; Domhan等，2015; Kandasamy等，2016; Klein等，2017a; Golovin等，2017）。 Swersky等人提出的算法。
（2013）使用GP学习相关任务之间的相关性，并要求子任务作为输入，但是在没有先验知识的情况下，目标任务具有高信息的有效子任务是未知的。类似于Swersky等人的工作。（2013），Klein等。
（2017A）使用内核将条件验证误差作为高斯进程建模，该过程将协方差与下采样率捕获，以允许自适应评估。 swersky等。（2014），Domhan等人。（2015）和Klein等人。
（2017A）对学习曲线的收敛性进行参数假设，以进行早期停止。相比之下，Golovin等人。
（2017）根据从非参数GP模型的预测性能设计了早期停止规则，其中内核旨在测量性能曲线之间的相似性。最后，Kandasamy等人。
（2016）扩展GP−UCB以允许通过定义单调改善的资源单调改善的子特派来进行自适应配置评估。

在另一项工作中，Sparks等人。（2015）提出了一项减半匪盗算法

这不需要明确的收敛行为，jamieson和talwalkar（2015）分析了karnin等人最初提出的类似算法。（2013）对于不同的环境，提供理论担保并鼓励实证结果。不幸的是，这些减半风格算法遭受了“N与B / N”问题，我们将在第3.1节中讨论。超接地解决了这个问题，并提供了一种用于超参数优化的强大，理论上是原则上的早期停止算法。

我们注意到超频带可以与任何覆盖物采样方法相结合

并且不依赖于随机抽样;理论结果仅假设采样的超参数配置的验证损失是从一些静止分布中汲取的。事实上，随后我们提交，Klein等人。（2017B）使用贝叶斯神经网络来模拟学习曲线的超频带组合自适应配置选择，并仅选择具有高预测性能的配置来输入到超带。

2.2 Bandit Problems

纯粹的勘探强盗问题旨在最大限度地减少简单的遗憾，定义为在任何给定的设置中尽快从最佳解决方案的距离。纯粹探索多武装强盗问题在随机设置中具有悠久的历史（偶数甚至，2006; Bubeck等，2009），最近延伸到贾姆斯森和塔尔沃尔卡尔的非随机环境（2015）。相关的是，Carpentier和Valko（2015）研究了随机纯勘探无限武装的强盗问题，其中每个臂的拉动我产生I.I.D.
[0,1]中的样本具有期望νi，其中νi是从累积分布函数的分布中汲取的损失，当然，νi的值是未知的，因此推断其价值的唯一方法是拉臂很多次。
Carpentier和Valko（2015）提出了一种随时算法，并在其错误中突出了一个紧密（达到彩色因子）上限，假设我们将参考F的β−参数化为

第5.3.2节。然而，它们的算法专门用于F的β−参数化，而且，它们必须在运行算法之前估计β，限制了算法的实际适用性。此外，该算法假设来自臂的随机损耗，因此已知收敛行为;因此，它不适用于我们的封路计优化设置.3两个相关的工作线都使用底层度量空间是高斯过程优化（Srinivas等，2010）和X武装匪徒（Bubeck等，2011年）或在公制空间上定义的匪徒。然而，这些作品要么承担随机奖励，要么需要了解底层功能的内容（例如，适当的内核或平滑度水平）。

相比之下，为非随机设置和自动设计超带

适应未知的f而不制定任何参数假设。因此，我们认为我们的工作是一般适用的无限武装匪徒的纯粹勘探算法。据我们所知，这也是第一个在实际应用中测试这种算法的工作。

3. Hyperband Algorithm

在本节中，我们介绍了超带算法。我们为算法提供直觉，通过一个简单的示例来突出显示主要想法，该示例是使用迭代作为自适应分配的资源，并在实践中展示了一些关于部署超带的指南。

3.1 Successive Halving

HyperBand扩展了Jamieson和Talwalkar（2015）所提出的逐次验证算法，并将其称为子程序。原始的逐步验算法背后的想法直接从其名称遵循：统一地将预算分配给一组高考配置，评估所有配置的性能，丢失最差的一半，然后重复直到一个配置仍然存在。该算法将指数增加的资源分配给更有前途的配置。不幸的是，ChecientiveHalvent需要配置数量N作为算法的输入。给定一些有限预算b（例如，选择一个小时训练时间来选择一个高级参数配置），B /
N资源在整个配置中平均分配。但是，对于固定的B，无论我们应该（a）是否应该考虑许多配置（大n），都不清楚，并不清楚具有小的平均培训时间;或（b）考虑少量配置（小n），平均培训时间更长。

我们使用一个简单的例子来更好地了解这个权衡。图2显示了验证

丢失作为分配两个配置的总资源的函数，终端验证损耗ν1和ν2。阴影区域与终端验证丢失的中间损耗的最大偏差绑定，并且将被称为“包络”功能.4可以在信封不再重叠时区分两种配置。简单的算术表明，当信封的宽度小于ν2−1，即时，当中间损耗保证小于ν2−ν1时，这就会发生这种情况

3.查看Jamieson和Talwalkar（2015）的工作，以便进行详细的讨论，激励非随机
setting for hyperparameter optimization.

4.保证这些信封功能存在;请参阅第5.2节中的讨论，我们正式定义
these envelope (or γ) functions.

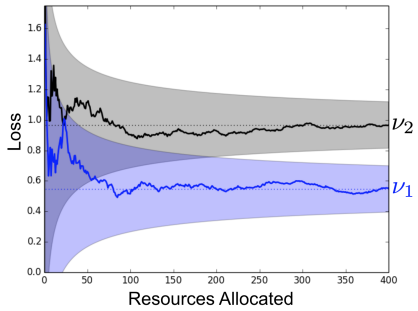


图2：验证丢失作为分配两个配置的总资源的函数

显示。ν1和ν2表示收敛时的终端验证损耗。阴影区域绑定了中间损耗从终端验证丢失的最大距离，并与资源单调减少。

终端损失。此观察结果中有两个外卖：当（1）包络函数更宽或（2）时，终端损耗在一起时，需要更多资源来区分两种配置。

但是，在实践中，最佳分配策略是未知的，因为我们没有了解信封功能，也不知道终端损失的分配。因此，如果在配置之前需要更多资源可以在质量方面差异化（例如，如果迭代训练方法为给定的数据集或者如果随机选择的HyperParameter配置相似地收敛），则工作是合理的具有少量配置。相反，如果在少量资源之后通常会揭示配置的质量（例如，如果迭代训练方法为给定的数据集很快收敛，或者如果随机选择的超参数配置具有高概率的低质量），则n是瓶颈，我们应该选择n大。

当然，如果元数据或以前的经验表明某个权衡可能在实践中运作良好，应该利用该信息分配给该权衡的大部分资源。但是，如果没有这种补充信息，从业人员被迫制定了这个权衡，严重阻碍了现有配置评估方法的适用性。

3.2 Hyperband

通过考虑用于固定B的几个可能的N个可能的值，以算法1所示，以算法1所示，解决了这个“n与b / n”问题，实质上执行n的可行值。与每个值的关联是在一些被丢弃之前分配给所有配置的最小资源R;较大的N值对应于较小的R，因此更具侵略性的早期停止。超带有两个组件;（1）内循环调用连续哈欠的N和R（第3–9行）的固定值（2）外环迭代N和R（线1–2）的不同值迭代。我们将在HyperBand内称为“括号”中的每个这样的连续展示。每个括号都旨在使用大约B总资源，并对应于n之间的不同权衡

算法1：超参数优化超频算法。

input : R, η (default $\eta = 3$)
initialization : $s_{\max} = \lfloor \log_\eta(R) \rfloor$, $B = (s_{\max} + 1)R$
for $s \in \{s_{\max}, s_{\max} - 1, \ldots, 0\}$ **do**
  $n = \frac{B}{R} \frac{\eta^s}{(s+1)}$, $r = R\eta^{-s}$
  //与（n，r）内循环开始逐次
  $T =$ `get_hyperparameter_configuration`(n)
  for $i \in \{0, \ldots, s\}$ **do**
    $n_i = \lfloor n\eta^{-i} \rfloor$
    $r_i = r\eta^i$
    $L = \{$`run_then_return_val_loss`$(t, r_i) : t \in T\}$
    $T =$ `top_k`$(T, L, \lfloor n_i/\eta \rfloor)$
  end
end
11返回配置迄今为止看到最小的中间损失。

和b / n。因此，单一执行超频带采用（Smax + 1）B的有限预算;我们建议无限期地重复它。

HyperBand需要两个输入（1）R，最大资源量
分配给单个配置，（2）η，一个控制在每轮连续哈欠中丢弃的配置比例的输入。这两个输入决定了考虑了多少个不同的括号;具体地，用Smax = Log η（R）×考虑n的Smax + 1不同值。超细从最具侵略性的括号S = SMAX开始，该SMAX设置为最大化探索，但受到至少一个配置被分配的R资源的约束。每个后续支架在最终括号S =
0的情况下减小n大约η的因子，其中每个配置被分配R资源（这个括号只是执行经典随机搜索）。因此，HyperBand在每个配置的平均预算中执行几何搜索，并消除以比连续持续的单个值连续持续的工作的成本以近似Smax
+的成本选择n的需要。通过这样做，HyperBand能够利用自适应分配运行良好的情况，同时在需要更保守分配的情况下保护自身。

超带需要为任何给定的学习问题定义以下方法：

获取HyperParameter配置（n）- 返回一组n i.i.d的函数。
来自超参数配置空间定义的一些分布的示例。在这项工作中，我们假设从预定义空间（即，带有Min的HyperCube的HyperCube的HyperParameters均匀地采样，其立即产生一致性保证。然而，分布的比对更高的高质量超参数（即，有用的先验），更好的超接管将执行（参见第6节以进行进一步讨论）。

	$s=4$		$s=3$		$s=2$		$s=1$		$s=0$	
i	n_i	r_i	n_i	r_i	n_i	r_i	n_i	r_i	n_i	r_i
0	81	1	27	3	9	9	6	27	5	81
1	27	3	9	9	3	27	2	81		
2	9	9	3	27	1	81				
3	3	27	1	81						
4	1	81								

表1：与各种相对应的超频带括号的NI和RI的值

s的值，当r = 81和 η = 3时。

运行然后返回val丢失（t，r）– 采用hyperparameter配置的函数 –

汇率t和资源分配r为输入，返回培训分配资源的配置后返回验证丢失。

top k（配置，损失，k）– 一种函数，也需要一组配置

作为其相关损失并返回顶部k执行配置。

3.3示例应用程序作为资源的迭代：Lenet

我们接下来存在一个具体示例，以提供关于超带的进一步直观。我们使用Mnist数据集，并优化使用迷你批量随机梯度下降（SGD）.5的Lenet卷积神经网络的超级参数，我们的搜索空间包括网络两层的学习速率，批量大小和内核数量随着封面（详细信息见附录A中的表2）。

我们将分配给每个配置的资源定义为SGD的迭代次数，

使用一个资源单位，对应于一个时代，即通过数据集进行全传递。我们将r到81设置并使用 η = 3的默认值，导致smax = 4，因此在n和b / n之间的不同权衡延伸的5括号。在每个括号内分配的资源显示在表1中。

图3显示了70个试验中的平均测试误差的经验比较

超带的各个括号单独运行以及标准超带。在实践中，我们不知道一个先验的括号S ∈ {0，。。。，4}将最有效地识别良好的超参数，并且在这种情况下，最多（S = 4）也不是最不攻击性（S = 0）设置是最佳的。但是，我们注意到超接线几乎与最佳括号（S = 3）差异，并且优于基线均匀分配（即随机搜索），这相当于括号S = 0。

3.4不同类型的资源

虽然前面的示例专注于迭代作为资源，但是HyperBand自然地推广到各种类型的资源：

5.使用的算法的代码和描述在http://deeplearning.net/tutorial/lenet.html上获得。

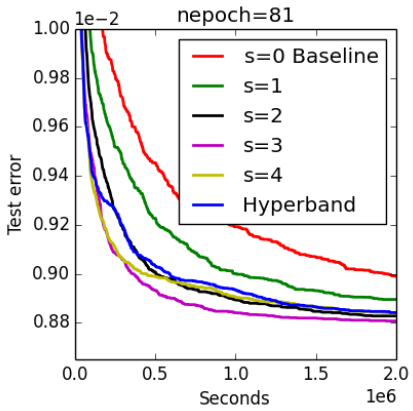


图3：单个括号S和超带的性能。

当各种封路计时，时间可能是优选的时间早期停止

配置在培训时间和从业者的主要目标有所不同，是在固定的挂钟时间内找到一个良好的封路数据设置。例如，培训时间可以用作快速终止分布式计算环境中的争吵作业的资源。

数据集分级 – 在这里，我们考虑一个黑匣子批量培训的设置

将数据设置为输入并输出模型的算法。在此设置中，我们将资源视为与R
R对应于完整数据集大小的R的随机子集的大小。使用超频带的子采样数据集大小，尤其是对于核心方法等超线性训练时间的问题，可以提供大量的加速。

特征子采样 – 随机特征或类似NYSTR ¨ OM的方法很受欢迎

用于近似机器学习应用内核的方法（Rahimi和RECHT，2007）。在图像处理中，特别是深度学习应用，滤波器通常随机采样，滤波器数量对性能产生影响。下采样功能的数量是手工调整超参数时使用的常用工具;超细可以形式化这种启发式。

3.5 Setting R

资源r和η（我们地址接下来）是超带所需的输入。如第3.2节所述，R表示可以分配给任何给定配置的最大资源量。在大多数情况下，在每个配置的最大预算上有一个自然的上限，这些配置通常由资源类型决定（例如，训练集规模为数据集缩小采样;基于内存约采样的内存约束的限制;关于数量的拇指规则迭代训练神经网络时的时期。如果R的一个可能的r值为R，则较小的R将使结果更快（因为每个括号的预算B是R的倍数），但较大的R将提供更好地保证配置之间的配置。

此外，对于其中R未知或不需要的设置，我们提供了一个
第5节超频的无限视野版本。此版本的算法双打

随时间的预算，B∈{2,4,8,16，。。。}，并且对于每个B，尝试所有可能的n∈
。对于B和N的每个组合，算法运行
（无限地平线）逐次校正算法的一个实例，其隐含地设置r = b
2 log2（n），从而增加R作为B增加。无限的主要区别
Horizo
n算法和算法1是唯一支架的数量随着时间的推移而增长，而不是使用每个外环保持恒定。我们将在第5节中更详细地分析此版本的超带，并将其作为标准（有限地平线）超界的理论分析的发射点。

请注意，R也是执行的括号中评估的配置数
最探索，即s =
smax。在实践中，可以希望n≤nmax限制与培训许多配置相关的开销，即小预算，即与初始化，加载模型和验证相关联的成本。在这种情况下，设置Smax = log η（nmax）。

或者，可以重新定义一个资源单位，使得R是人为更小的（即，如果所需的最大迭代为100K，将一个资源单位定义为100次迭代将给出r = 1,000，而定义一个单元是1k迭代将给出r =
100）。因此，可以将一个资源单位解释为最小期望的资源和R作为最大资源和最小资源之间的比率。

3.6 Setting η

η的值是可以基于实际用户约束进行调谐的旋钮。η的较大值对应于更具侵略性的消除计划，从而减少较少的消除;具体地，每个圆形保留1 / η的（n） +
1轮消除n个配置。如果希望在次优渐近常数的成本上更快地接收结果，则可以增加η以减少每个支架的预算B = (log η（r）≠+
1）r。我们强调结果对η的选择不是很敏感。如果我们的理论界限进行了优化（参见第5节），他们建议选择η=e≈.2.718，但在实践中，我们建议将η拍摄等于3或4。

调谐η还将改变括号的数量，从而改变不同的数量
超频尝试的权衡。通常，可能的括号范围相当受到约束，因为括号的数量是R;即，存在（ log η（r） + 1）= smax
+1括号。对于我们在第4节中的实验中，我们选择η为指定的R提供5个括号;对于大多数问题，5是探索的合理数量的n与b / n权衡。然而，对于大R，使用η=
3或4可以提供比所需的更多括号。可以通过几种方式控制括号数。首先，如前一节所述，如果R太大并且开销是一个问题，则可以通过将最大配置数限制为nmax来控制开销，从而限制SMAX。如果开销不是一个关注和侵略性的探索，可以（1）增加η以减少括号的数量，同时保持r作为最探索性支架中的最大配置数，或（2）仍然使用η= 3或4但只尝试括号拨射的基线探索，即设置Nmin，并且只尝试从Smax到S
= Log η（nmin） 的括号。对于具有长期训练时间和高维搜索空间的计算密集问题，我们推荐后者。直观地，如果可以在合理的时间内完成可以培训的配置（即，使用R资源训练）的配置数量是在搜索空间的维度的尺寸和维度下不是指数的

在不使用N和B / N之间的攻击性探索权衡，不可能找到良好的配置。

3.7理论结果概述

通过一个例子，最好证明超接管的理论特性。假设存在n个配置，每个配置都有给定的终端验证错误 ν i为i = 1，。。。 。不损失普遍性，索引配置的配置，使得 ν 1对应于最佳执行配置， ν 2至第二个，等等。现在考虑识别最佳配置的任务。最佳策略将分配给每个配置我从 ν 1区分其所需的最低资源，即足够的信封函数（参见图2）将中间损耗绑定为小于 ν i− ν 1

2远离终端值。相比之下，天真的统一分配

将B /
N分配给每个配置的策略必须分配到每个配置，以区别于 ν 1的任何ARM ν i所需的最大资源。值得注意的是，逐次持有所需的预算只是最佳的一个小因素，因为它会利用易于区分 χ 1的配置。

统一分配和逐次持平所需预算的相对大小

取决于信封函数偏离终端损耗的偏差以及绘制 ν i的分布。当3.1节中讨论的最佳N与B /
N权衡时，逐次持续的预算较小，需要每个配置的资源较少。因此，如果信封函数随着资源分配的函数快速拧紧，或者端子损耗之间的平均距离大，则成功的阶段可以比均匀分配更快。这些直觉在第5节中正式化，并提供了相关的定理/转义式，以考虑到信封函数和绘制的分布。

在实践中，我们没有了解信封功能或分发

ν i的，这两者都是整体的，在表征连续的持续预算中。利用超带，我们通过对待我们的侵略性来解决这种缺点。我们在第5.3.3节中展示了HyperBand，尽管不了解信封功能，也不知道 ν i的分发，需要预算，只有日志因素大于连续哈维文。

4. Hyperparameter Optimization Experiments

在本节中，我们评估了具有三种不同资源类型的超带的经验行为：迭代，数据集址和特征样本。对于所有实验，我们使用其默认设置比较具有三个众所周知的贝叶斯优化算法，TPE和Spearmint的HyperBand。当搜索空间中有条件的超参数时，我们从比较集中排除留空，因为它没有本地支持它们（EggenSperger等，2013）。我们还显示出同性恋的结果，该结果对应重复最探索的超频带括号，以便为侵略性的早期停止提供基线，作为衡量标准基线的标准基线

6.这不是在第4.2.1节中的实验完成的，因为最具侵略性的括号因数据集而异
与培训点数的数据集。

所有加速，我们考虑随机搜索和“随机2×”，随机搜索的变种，其中包含其他方法的两倍。在第2节中描述的混合方法，我们与使用Domhan等人提出的早期终止标准的SMAC的变体进行比较。（2015）在第4.1节中描述的深度学习实验中。我们认为Hyiin等人最近引入的超接管到更复杂的混合方法的比较。（2017A）和Kandasamy等人。（2017）是未来工作的富有成效的方向。

在下面的实验中，我们在确定如何遵循这些宽松的准则
configuration HYPERBAND:

1.鉴于问题，最大资源r应合理，但理想情况下很大
足以让早期停止是有益的。

2. η应该取决于r并选择含有最小值的≈5括号
3括号。这是为了保证超带将使用基线程度
早期停止并防止N VS B Tressoff的Grid太粗糙。

4.1深度学习的早期迭代算法

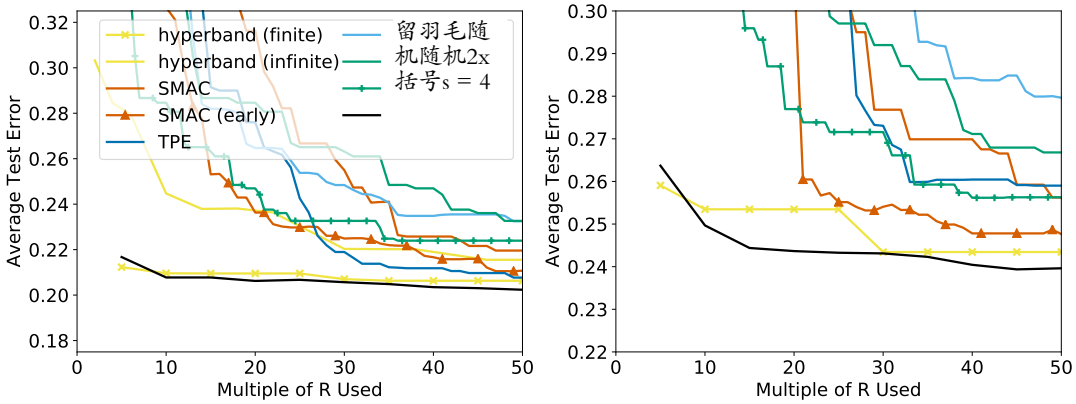
对于该基准测试，我们调整了一个与Snoek等人使用相同的架构的卷积神经网络7。（2012）和Domhan等人。（2015）。两个先前作品中使用的搜索空间不同，我们使用类似于Snoek等人的搜索空间。（2012）对于随机渐变体积的6个超参数，以及2个响应标准化层的超参数（有关详细信息，请参阅附录A）。符合前两个工作，我们使用批量为100的所有实验。

数据集：我们考虑了三种图像分类数据集：CiFar−10（Krizhevsky，
2009年），旋转Mnist与背景图像（MRBI）（Larochelle等，2007）和街头视图房屋号（SVHN）（Netzer等，2011）。CIFAR−10和SVHN包含32×32
RGB图像，而MRBI包含28×28灰度图像。每个数据集被分成培训，验证和测试集：（1）Cifar−10具有40k，10k和10k实例;（2）MRBI拥有10K，2K和50K实例;（3）SVHN分别接近600K，6K和26K，分别进行培训，验证和测试。对于所有数据集，对原始图像执行的唯一预处理正在逐步侦查。

超频配置：对于这些实验，一个资源单位对应
到100个迷你批处理迭代（批量大小为100的10K示例）。对于CiFar−10和MRBI，R设定为300（或30k总迭代）。对于SVHN，R设置为600（或60k的总迭代）以适应更大的训练集。对于这些实验给出了R，我们设置了η=4以产生5个超接头的连续阶层括号。

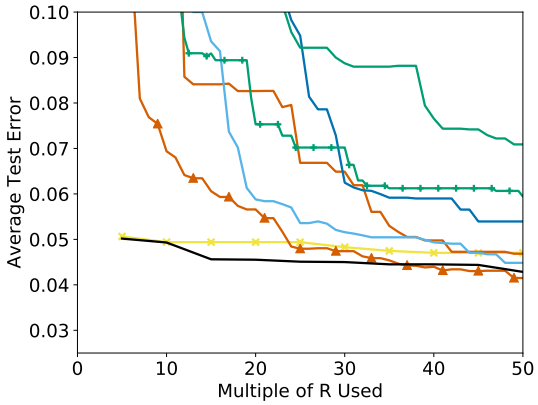
结果：每个搜索者每次试验的总预算为50r，以返回最佳效果
可能的HyperParameter配置。对于超接线，预算足以运行外循环两次（总共10个连续的括号括号）。对于SMAC，TPE和随机搜索，预算对应于训练50种不同的配置完成。为每个搜索者执行十项独立试验。实验相当于在Amazon EC2 G2.8xlarge实例上提供的NVIDIA网格K520卡上的GPU小时超过1年。我们为总的预算制约规定了

7.型号规范可在http://code.google.com/p/cuda−convnet/提供。



(a) CIFAR-10

(b) MRBI



(c) SVHN

图4：10个试验中的平均测试错误。标签“SMAC（早）”对应于SMAC随着Domhan等人提出的早期停止标准。（2015）和标签“括号S = 4”对应于重复最探险的超频型托架。

迭代而不是计算时间来使比较硬件独立.8比较迭代的进度而不是时间忽略开销成本，例如，忽略开销成本。贝叶斯方法和模型初始化和超频验证成本的配置选择成本。虽然开销是硬件依赖的，但超接头的开销低于5％的EC2 G2.8xLARGE机器，因此通过时间通过的进展情况不会显着改变结果。

对于CiFar-10，图4（a）中的结果表明超接线超过订单 -
比其竞争对手快得多。对于MRBI，HyperBand超过了一个订单 -

大多数试验在亚马逊EC2 G2.8xlarge实例上运行，但在不同的机器上运行了一些试验由于这些实验的大量计算需求。

幅度比标准配置选择的速度快，比SMAC（早）快5倍。对于SVHN，HyperBand找到更好的配置速度更快，贝叶斯优化方法具有竞争力和SMAC（早期）优于超大带。SMAC（早）的性能表明了结合早期停止和自适应配置选择的优点。

在三个数据集中，超带和SMAC（早）是唯一的两种方法

始终如一地优于随机2×。在这些数据集上，超接线比随机搜索快20倍，而SMAC（早）比评估窗口内的随机搜索快≤7×。实际上，使用5R预算后的超带返回的第一个结果通常与其他搜索者使用50R之后的结果竞争。此外，超细比跨试验的其他搜索者的变量减少，这在实践中非常希望（参见附录A用于误差栏的绘图）。

如第3.6节所述，用于高维的计算昂贵的问题

搜索空间，只需重复最探索性括号可能是有意义的。同样，如果元数据有关于问题或者已知配置在分配少量资源之后可以看到的配置的质量，那么就应该重复最探索的括号。实际上，对于这些实验，括号S = 4大大优于CIFAR−10和MRBI上的所有其他方法，并且在SVHN上首先与SMAC（早）几乎与SMAC（早期）捆绑在一起。

虽然我们为这些实验设置了r，以便于与贝叶斯方法的比较

和随机搜索，使用无限的地平线超频带来的最大资源也是合理的，直到达到所需的性能。我们使用 $\eta = 4$和$B = 2R$的起始预算评估CIFAR−10上的无限地平线超频带。图4（a）显示无限的地平线超接管与其他方法具有竞争力，但在50R预算限制内没有执行有限的地平超频带。Infinite Horizon算法最初表现不佳，因为它必须调整最大资源r并以不太积极的早期停止速率开始。这展示了在已知最大资源的场景中，最好使用有限的地平线算法。因此，我们专注于我们的实证研究剩余时间的超带的有限范围版本。

最后，CiFar−10是一种非常流行的数据集和最先进的模型实现了很多

较低的误差率比图4所示。性能的差异主要是归因于更高的模型复杂性和数据操作（即使用反射或随机裁剪以人为地增加数据集大小）。如果我们限制与使用相同架构的公布结果的比较并排除数据操作，则数据集的最佳人类专家结果是18％错误，并且Snoek等人的最佳超参数优化结果为15.0％。（2012）Domhan等人的9和17.2％。（2015）。这些结果超越了CiFar−10，因为他们通过包括验证集的数据培训25％，并为更多时期培训。当我们培训300时期的组合训练和验证数据上的超接管发现的最佳模型时，该模型达到了17.0％的测试误差。

9.即使在收到最佳的超参数后，我们也无法重现这种结果
作者通过个人沟通。

4.2 Data Set Subsampling

我们研究了两种不同的覆盖物搜索优化问题，其中超带使用数据集副页作为资源。首先采用Feurer等人提供的广泛框架。（2015）试图自动化预处理和模型选择。由于框架的某些限制，从根本上限制了数据集下采样的影响，我们使用内核分类任务进行了第二个实验。

4.2.1 117 Data Sets

我们使用了Feurer等人介绍的框架。（2015）探索了由15分类器组成的结构化的超参数搜索空间，14个特征预处理方法，共有4个数据预处理方法，总共110个超参数。我们排除了Feurer等人的元学习组件。（2015）用来使用有前途的配置加热贝叶斯方法，以便与随机搜索和超带进行公平的比较。类似于Feurer等人。（2015），我们强加了3GB内存限制，为每个HyperParameter配置和一小时时间窗口进行6分钟超时，以评估每个数据集的每个搜索者。每个搜索者的二十个试验在每个数据集执行，并且聚合中的所有试验都花费了Google Cloud Compute的N1标准−1实例上的一年内的CPU时间。有关我们实验框架的其他详细信息，请参见附录A.

数据集：Feurer等。（2015）使用了140个二进制和多字符分类数据集

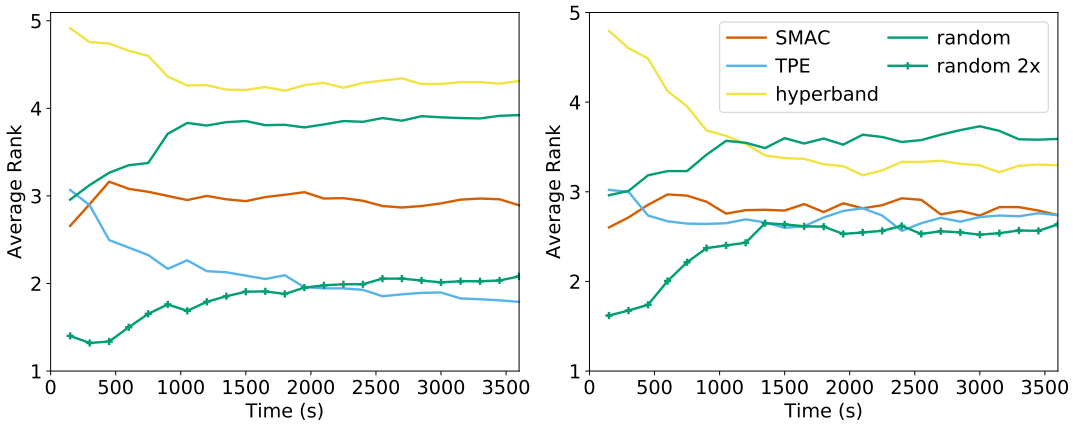
来自OpenML，但其中23个与最新版本的OpenML插件（Feurer，2015）不兼容，因此我们使用其余117个数据集。由于实验设置的局限性（附录A中讨论），我们还分别考虑了这些数据集中的21个，这至少展示了由于附带的最小（尽管仍然是Sublinear）训练加速。具体地，由于100个随机选择的超参数配置，这两个数据集中的每一个在平均至少为3倍的加速度时显示为8倍。

超带配置：由于数据集大小宽，有一些数据集

具有少于10K的训练点，我们用 $\eta =$
3运行超接管，以允许至少3个括号，而不会在小型数据集上的下采样时过于侵略性。R设置为每个数据集的完整训练集大小，并且任何括号的任何括号的最大配置都限于nmax = max {9，r /
1000}。这确保了超接管的最探索性括号将至少两次下降。如第3.6节所述，当指定NMAX时，运行算法时的唯一差异是SMAX = $\log_\eta$（nmax） 而不是 $\log_\eta$（r） 。

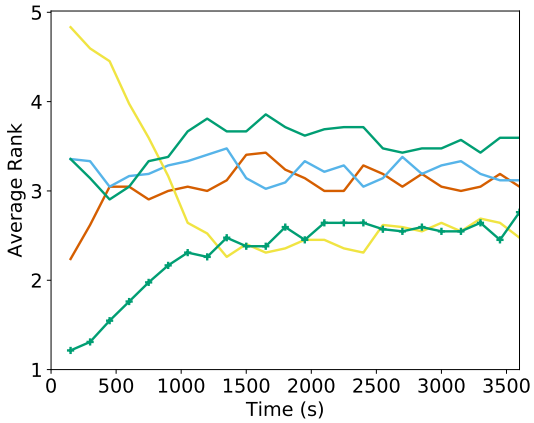
结果：图5（a，b）中所有117个数据集的结果显示超带

尽管在验证错误等级中表现更差，但在测试错误等级中优于随机搜索。贝叶斯方法在测试错误性能方面胜过超频带和随机搜索，而且还表现出对验证集的过度录制的迹象，因为它们以验证错误等级上的较大余量优于超大频率。值得注意的是，随机2×优于所有其他方法。但是，对于21个数据集的子集，图5（c）显示超带优先于测试错误等级上的所有其他搜索者，包括随机2×通过非常小的边距。虽然这些结果更有希望，但在这个实验框架中受到超频带的有效性;对于较小的数据集，启动开销是



（a）117数据集的验证错误

（b）在117数据集上测试错误



（c）21数据集的测试错误

图5：每个搜索者的所有数据集的平均排名。对于每个数据集，搜索者根据20项试验的平均验证/测试错误排名。

相对于总训练时间高，而对于较大的数据集，只有少数配置可以在小时窗口中培训。

我们注意到，虽然平均等级图，如图5所示是一种有效的方法在许多搜索者和数据集中聚合信息，它们不提供关于方法性能之间的差异的幅度。图6，图表6，其中每个搜索者的测试错误与所有117个数据集的随机搜索之间的差异突出显示搜索者跨搜索误差幅度的少量差异。

这些结果并不令人惊讶;如第2.1节所述，香草贝叶斯优化方法类似地执行高维搜索空间中的随机搜索。Feurer等人。（2015）表明，使用META学习来加强贝叶斯优化方法，在这种高维设置中提高了性能。使用元学习来识别

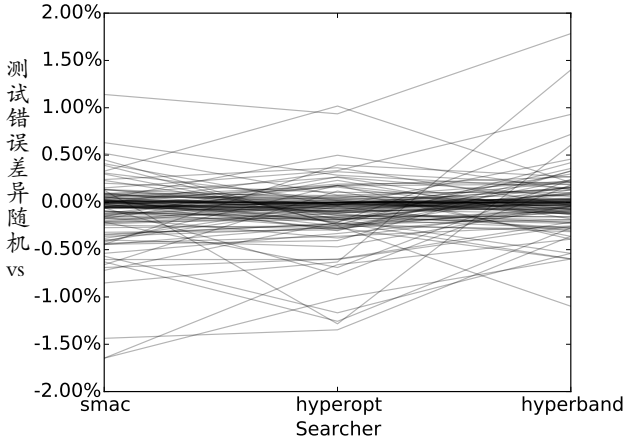


图6：每个线条图，对于单个数据集，测试错误与随机的差异

搜索每个搜索，较低的是更好的。几乎所有的线条都落在−0.5％和0.5％的乐队内，除了几个异常值外，这些线条大多是平的。

有希望的分布从它作为输入到超带的输入是未来工作的方向。

4.2.2内核正规化最小二乘分类

对于此基准，我们调整了CIFAR−10上基于内核的分类器的超级参数。我们使用了多级规则化最小二乘分类模型，该分类模型与SVMS（Rifkin和Klautau，2004;
Agarwal等，2014）具有相当的性能，但可以明显培训.10在搜索空间中考虑的近似参数包括预处理方法，正则化，内核类型，内核长度尺度和其他内核特定的超参数（有关详细信息，请参阅附录A）。对于超带，我们设置了r = 400，每个资源单位表示100个数据点，$\eta = 4$产生总共5括号。在Amazon EC2
M4.2XLARGE实例上运行每个超级计数器优化算法为10个试验;对于给定的试验，允许超细运行两个外环，重复括号S = 4次，所有其他搜索者都运行12小时。

图7显示完成后超频带返回了良好的配置

首先连续20分钟的支架;即使在整个12小时之后，其他搜索者也无法平均达到此错误率。值得注意的是，超界能够在连续阶层的第一个括号中评估超过250个配置，而竞争对手能够在相同的时间内仅评估三种配置。因此，超接线比贝叶斯优化方法快30倍，比随机搜索快70倍。括号S = 4瞄准优于超大带但终端

10. Scikit−Learn中的默认SVM方法是单核，需要数小时才能在CiFar−10上训练，而a
块坐标抑结最小二乘求解器在8个核心机器上少于10分钟。

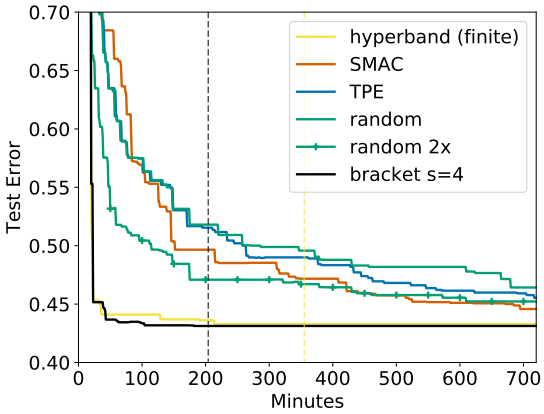


图7：最好的Ker的平均测试错误 -

NEL规则化最小二乘范围由CIFAR-10上的每个搜索者发现的模型。颜色编码的虚线表示给定搜索者的最后一次试验完成。

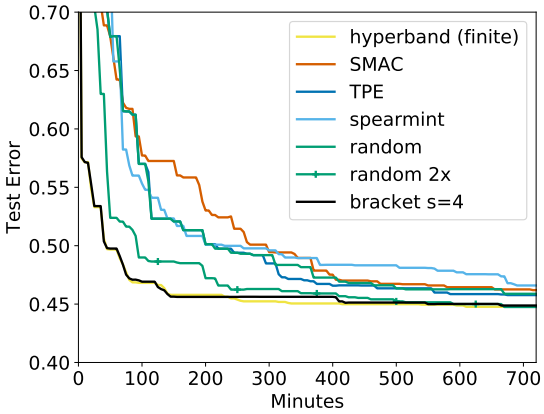


图8：最佳平均测试错误

Cifar-10上每个搜索者发现的随机功能模型。在每次评估中而不是支架末端时计算超带和支架S = 4的测试误差。

两种算法的性能相同。随机2×与SMAC和TPE有竞争力。

4.3特征子采样以加快近似内核分类

接下来，在随机功能内核近似任务上使用功能时，我们会检查超带的性能。使用Rahimi和RECHT（2007）中描述的方法随机生成特征，以近似RBF内核，然后将这些随机功能用作脊回归分类器的输入。HyperParameter搜索空间包括预处理方法，内核长度尺度和L2罚款。虽然使用无限的地平超频似乎自然，但由于近似的保真度以更随机的特征改善，在实践中，可用机器存储器的量施加了特征数量的自然上限。因此，我们使用了有限的水平超频带，最大资源为100k随机特征，可舒适地安装在具有60GB内存的机器中。此外，我们将一个资源单位设置为100个功能，因此r = 1000。再次，我们设置 η = 4以产生5括号的连续阶段。我们对每个搜索者进行了10个试验，每次试验在Google Cloud Compute的N1标准-16计算机上持续12小时。图8中的结果表明，超过贝叶斯方法和随机搜索的超接线速度约为6倍。 HyperBand类似地执行括号S = 4.随机2×优于贝叶斯优化算法。

4.4 Experimental Discussion

虽然我们的实验结果显示超接管是一个有前途的普通优化优化算法，但许多问题自然出现了：

1.影响超带提供的加速度的是什么？

2.为什么连续哈维文似乎擅长超额带？

3.依赖资源的超参数呢？

我们依次解决每个问题。

4.4.1影响超带性能的因素

对于给定R，超带执行的最探索性连续校正轮换使用（ Log η（R）+
1）R的预算来评估R配置，其在随机搜索中给出了电位加速的上限。如果培训时间与资源线性缩放，则与随机搜索相比超接带提供的最大加速

（ log η（r） +
1）。对于我们实验中使用的 η 和r的值，随机搜索的最大加速度约为30×给定线性训练时间。但是，我们将一系列从6倍的加速度比随机搜索快6倍。实现加速的差异可以通过三个因素来解释：

1.如何使用给定的资源培训时间缩放。在培训时间的情况下

超连线作为资源的函数，超带可以提供更高的加速。例如，如果培训尺度如程度的P>
1，则随机搜索的超频的最大加速度大约是 η p−1

η p−1 r.在内核最不正方形中

分类器实验在第4.2.2节中讨论，培训时间与资源的函数相当缩放，这解释了为什么实现了70×的实现加速度高于给定线性缩放的最大预期加速度。

2.与培训相关的开销费用。总评估时间也取决于固定的

与评估每个HyperParameter配置相关的开销成本，例如，初始化模型，恢复先前培训的模型，并计算验证错误。例如，在第4.2.1节中提出的117数据集的下采样实验中，超接线没有提供显着的加速，因为在几秒钟内可以训练许多数据集，并且初始化成本相对于训练时间很高。

3.找到良好配置的难度。封锁率优化问题

可能有所不同。例如，“简单”问题是随机采样配置可能导致高质量模型的问题，因此我们只需要评估少量配置以找到一个良好的设置。相比之下，“硬”问题是任意配置可能是坏的，在这种情况下，必须考虑许多配置。超频带利用下采样以提高评估的配置的数量，因此更适合“硬”问题，其中实际上需要更多的评估来找到一个良好的设置。通常，问题的难度与搜索空间的维度相比缩放。对于低维问题，通过随机搜索和贝叶斯方法评估的配置的数量在尺寸的数量中是指数的，所以可以实现良好的覆盖范围。例如，第4.3节中的特征回顾实验中的低维（D = 3）搜索空间有助于解释为什么超带是

比随机搜索快6倍。相比之下，对于4.1节中的神经网络实验，我们假设由于搜索空间的维度更高，因此观察到超频的更快的加速。

4.4.2 Comparison to SuccessiveHalving

除了LENET实验（第3.3节）和117个数据集经验（第4.2.1节），在我们所有实验中都是连续高度超时的超级带带的最具侵略性的括号。在后智，我们应该只运行支架S =
4，因为积极的早期停止为许多基准测试任务提供了大量的加速。然而，如前所述，它未知一个先验，即括号S
=
4会表现最佳，这就是我们必须通过带有超带的所有可能括号的原因。另一个问题是当一个人进一步增加时会发生什么，而不是4轮消除，为什么不超过相同的最大资源r？在我们的情况下，S =
4是最具侵略性的括号，我们可以给定每个对先前实验所施加的每个配置限制的最低资源。然而，对于较大的数据集，可以扩展S的可能值的范围，在这种情况下，如果更具侵略性的早期停止有助于或者如果最具侵略性的括号，则超接管可以提供更快的加速度基本上是一次性的。

我们相信现有知识，对任务的知识可能特别有用，以限制

超带探索的括号范围。在我们的经验中，积极的早期停止对于神经网络任务通常是安全的，并且更具侵略性的早期停止对于较大的数据集和更长的训练视野可能是合理的。然而，当通过增加S推动早期停止程度时，必须考虑与检查更多型号相关的额外的开销成本。因此，利用元学习的一种方法是使用学习曲线收敛速率，不同搜索空间的难度，以及相关任务的开销成本来确定由超带考虑的括号。

4.4.3 Resource Dependent Hyperparameters

在某些情况下，给定的HyperAmameter的设置应取决于分配的资源。例如，随机林的最大树深度正则化超参数应更高，具有更多数据和更多功能。但是，最大树深度和资源之间的最佳权衡是未知的，可以是特定于数据集的。在这些情况下，收敛速率通常慢，因为较小资源上的性能不指示更大的资源。因此，对于超接管来说，这些问题特别困难，因为可以静止早盘的益处。同样，虽然超接线只有比随着最佳的早期停止率的逐步较慢的小因素，但是如果可能，我们建议删除HyperParameter对资源的依赖。对于随机森林示例，替代正则化HyperParameter是每片叶子的最小样本，其依赖于训练集大小。另外，可以通过简单的归一化除去依赖性。例如，我们内核最小二乘实验的正则化术语是通过训练集大小标准化，以维持平均平衡误差与正则化术语之间的持续权衡。

5. Theory

在本节中，我们介绍了纯粹探索的非随机无限武装强盗（NIAB）问题，这是一个非常普遍的环境，包括我们兴趣的普遍参数优化问题。正如我们将展示的那样，超细实际上适用于远远超出HyperParameter优化的问题。我们首先正式地形成了Quand参数优化问题，然后将其缩减为纯粹探索的IAIB问题。我们随后在无限和有限的地平设置中详细分析超接线。

5.1 Hyperparameter Optimization Problem Statement

设X表示有效的超参数配置的空间，其可以包括可以包括以任意方式相互彼此约束的连续，离散或分类变量（即，不需要限制为[0,1]的子集。d）。对于k = 1,2，。。。让 k：x→[0,1]是X
X序列序列。对于任何HyperParameter配置x∈X， k（x）表示使用x与k资源单位的模型的验证错误（例如迭代）。另外，对于一些R∈n∪{∞}，定义 ⋆ = limk→r k和ν⋆
=infx∈x ⋆（x）。请注意，对于所有K∈N， ⋆（·），ν⋆都是未知的 k（·）算法先验。特别是，不确定 k（x）作为任何固定k的x的函数的函数如何变化，以及如何快速 k（x）→ ⋆（x）作为任何固定x x x的k的函数。

我们假设HyperParameter配置从已知的概率随机进行采样 -
ITY分布P（x）：x→[0，∞），支持x。在我们的实验中，p（x）只是均匀的分布，但算法可以与任何采样方法一起使用。如果x∈X是从该概率分布的随机样本，那么 ⋆（x）是一个随机变量，其分布是未知的，因为 ⋆（·）未知。另外，由于未知 k（x）作为x或k的函数变化，因此不一定地推断出任何J∈n，y x
x的 j（y）的知识的任何关于 k（x）的任何内容。，我们将HyperParmeter优化问题降低到更简单的问题，忽略了HyperParameters的所有基础结构：我们只通过其丢失序列 k（x）与k =
1,2的丢失序列 k（x）相互作用。。。。随着这种减少，x∈X的特定值不仅仅是索引或唯一标识损耗序列。

没有了解快速 k（·）→ ⋆（·）或如何分配 ⋆（x），目标
超细是识别一个超参数配置x∈X，它通过根据可能的少量总资源绘制多个随机配置来最小化 ⋆（x）− ν⋆。

5.2纯探索非随机无限武装炸药问题

我们现在正式定义了感兴趣的强盗问题，并将其与QuandParameter优化问题相关联。NIAB游戏中的每个“ARM”与从分布通过序列随机绘制的序列相关联。如果我们“拉动”iTh绘制的臂，我们会观察损失 i，k。同时，玩家可以绘制新的臂（序列）或额外地拉动先前绘制的臂。可以绘制的武器数量没有限制。我们假设武器只能通过他们的索引来识别

我（即，我们没有手臂的侧面知识或特征表示），我们还提出了以下两个额外假设：

每个I∈n的假设1限制斜唇→∞∞i，k存在，等于νi.11

假设2每个νi是一个有界的i.i.d.随机变量，累积分布函数f。

NIAIA问题的目的是使用尽可能少的总拉动来识别小ν　的ARM　。我们有兴趣表征ν　作为来自所有武器的总量的函数。显然，上面描述的超参数优化问题是NiaIb问题的实例。在这种情况下，ARM I对应于配置Xi X X，具有　i，k＝　k（xi）;假设1相当于要求存在νi=　⋆（xi）;假设2从武器被绘制i.i.d.根据分布函数p（x）的x。f仅仅是　⋆（x）的累积分布函数，其中x是从分布p（x）x的随机变量。注意，由于手臂绘制是独立的，因此νi也是独立的。同样，这并不是说验证损失不依赖于超参数的设置;验证损耗很可能与某些超参数相关，但是这不是在算法中使用的，并且没有关于相关结构的假设。

为了分析NIAB设置中超带的行为，我们必须定义一些额外的物体。让ν⋆= INF {M：P（ν≤M）>0}>−∞，因为分布F的域被界定。因此，累积分布函数f满足

$$\mathbb{P}\left(\nu_i - \nu_* \leq \epsilon\right) = F(\nu_* + \epsilon) \tag{1}$$

并且让F −1（Y）= INFX {x：f（x）≤y}。定义γ：n→r为点亮最小，单调减少功能令人满意

$$\sup_i |\ell_{i,j} - \ell_{i,*}| \leq \gamma(j)\ , \quad \forall j \in \mathbb{N}. \tag{2}$$

随着迭代j的序列增加，该功能γ被保证通过假设1存在，并将与极限值的偏差绑定。对于HyperParameter Optimization，这是　k均匀地收敛到所有x∈X的事实中。此外，γ可以被解释为验证在资源子集上培训的配置的验证误差的偏差与最大数量分配资源。最后，将R定义为第一索引，使得γ（r）= 0如果存在，则设置r =∞。对于Y≥0，使用γ−1（0）：= r的惯例，γ−1（y）= min {j≠n：γ（j）≤y}可以是无限的。

如前所述，存在许多真实世界的情景，其中R是有限的已知。例如，如果使用完整数据集的额外子集用作资源，则最大资源数量不能超过完整数据集大小，因此γ（k）= 0，其中r是R是R的（已知）数据集的全尺寸。在其他情况下，例如迭代训练问题，可能不想或知道如何绑定R.我们将这两个设置分开到有限的地平线设置，其中R是有限和已知的，而无限的地平线

11.我们可以始终定义　i，k，以便保证收敛，即取得序列的最10。

逐步的（无限地平线）输入：预算b，n武器，其中ki，k表示从ith手臂的kth损失

Initialize: $S_0 = [n]$.

For $k = 0, 1, \ldots,$ $\log_2(n)$ 1

将每个臂拉到sk for rk = b $\overline{{}_2(n)\rceil}\rfloor$ times.

在rkth观察到的损失中保持最佳 | sk | /2 武器作为sk + 1。

Ou $\hat{\imath}, \lfloor \frac{B/2}{\lceil \log_2(n) \rceil} \rfloor$ $\lceil \log_2(n) \rceil$

HYPERBAND (Infinite horizon)

Input: None

For $k = 1, 2, \ldots$

For $s \in \mathbb{N}$ s.t. $k - s \geq \log_2(s)$

$B_{k,s} = 2^k$, $n_{k,s} = 2^s$

$\hat{\imath}_{k,s}, \ell_{\hat{\imath}_{k,s}, \lfloor \frac{2^{k-1}}{s} \rfloor} \leftarrow$ SUCCESSIVEHALVING $(B_{k,s}, n_{k,s})$

图9：（顶部）在jamieson中提出和分析的逐次验证算法

和Talwalkar（2015）为非随机设置。注意本算法最初提出用于Karnin等人的随机设置。（2013）。（底部）无限地平线设置的超频算法。超带呼叫作为子程序连续呼叫。

r的r是已知r绑定的，并且假设是无限的。虽然我们的经验结果表明，有限地平线可能更实际地对封锁率优化问题更加相关，但无限的地平线案例与文献有自然的联系，我们首先通过分析此设置。

5.3 Infinite Horizon Setting ($R = \infty$)

考虑图9的超频算法。该算法使用逐次使用（图9）作为子程序，该子程序采用有限组臂作为输入，输出该组中最佳执行臂的估计。首先为特定的一组限制νi分析逐步的（SH），然后考虑根据F的随机绘制时SH的性能。然后，我们分析超带算法。我们注意到，图9的算法最初由Karnin等人提出。（2013）为随机设置。然而，Jamieson和Talwalkar（2015）分析了非随机环境中，并在实践中找到了很好的工作。延长Jamieson和Talwalkar（2015）的结果我们有以下定理：

定理1修复n武器。让 ν i= lim →∞ ∞i，t并假设n1・・・≤nn。对于任何e> 0 etet

$$z_{SH} = 2\lceil \log_2(n) \rceil \max_{i=2,\ldots,n} i\,(1+\gamma^{-1} \max \tfrac{\epsilon}{4}, \tfrac{\nu_i-\nu_1}{2})$$
$$\leq 2\lceil \log_2(n) \rceil \big(n + \sum_{i=1,\ldots,n} \gamma^{-1} \left(\max\left\{\tfrac{\epsilon}{4}, \tfrac{\nu_i-\nu_1}{2}\right\}\right)\big)$$

如果使用任何预算B> ZSH运行图9的逐次验证算法，则返回ARM ，该ARM 满足ν − ν1≤ε/2。而且，|

下一个技术引理将用于表征问题依赖项

当序列从概率分布中汲取时。

$$\mathbf{H}(F,\gamma,n,\delta,\epsilon) := 2n \int_{\nu_*+\epsilon/4}^{\infty} \gamma^{-1}(\tfrac{t-\nu_*}{4}) dF(t) + \left(\tfrac{4}{3}\log(2/\delta) + 2nF(\nu_*+\epsilon/4)\right)\gamma^{-1}\left(\tfrac{\epsilon}{16}\right)$$

and $\mathbf{H}(F,\gamma,n,\delta) := \mathbf{H}(F,\gamma,n,\delta,4(F^{-1}(p_n)-\nu_*))$ *so that*

$$\mathbf{H}(F,\gamma,n,\delta) = 2n\int_{p_n}^{1} \gamma^{-1}(\tfrac{F^{-1}(t)-\nu_*}{4})dt + \tfrac{10}{3}\log(2/\delta)\gamma^{-1}\left(\tfrac{F^{-1}(p_n)-\nu_*}{4}\right).$$

对于具有限制的N臂 ν1≤・・・・≤νN，然后

$$\nu_1 \leq F^{-1}(p_n) \quad \textit{and} \quad \sum_{i=1}\gamma^{-1}\left(\max\left\{\tfrac{\epsilon}{4}, \tfrac{\nu_i-\nu_1}{2}\right\}\right) \leq \mathbf{H}(F,\gamma,n,\delta,\epsilon)$$

对于具有概率至少1 − δ的任何ε≥4（F−1（PN）− ν⋆）。

在定理1中设置ε= 4（F−1（PN）− ν⋆）并使用引理2的结果
ν⋆≤ν1≤ν⋆+（f−1（pn）− ν⋆），我们立即获得以下推论。

Corollary 3 *Fix* δ $(0,1)$ *and* ϵ $4(F^{-1}(\frac{}{n})$ ν *). Let* $B = 4\ \log_2(n)\ \mathbf{H}(F,\gamma,n,\delta,\epsilon)$
其中H（f，γ，n，δ，ε）在Lemma
2中定义。如果图9的连续阶段算法使用根据f随机绘制的指定的b和n臂配置，则ARM ∈[n]被返回，使得具有至少1 − Δ的概率，我们有ν − ν⋆≤

+ε/2。特别地，如果b =4 log2（n） （f，γ，n，δ）和

$\frac{}{n}$ $*$ $\hat{\imath}$ $*$ $\frac{}{n}$ $*$ 概率至少为1 − δ。

请注意，对于任何固定的n∈N，我们具有任何δ0

$$\mathbb{P}(\min_{i=1,\ldots,n} \nu_i - \nu_* \geq \Delta) = (1-F(\nu_*+\Delta))^n \approx e^{-nF(\nu_*+\Delta)}$$

which implies $\mathbb{E}[\min_{i=1,\ldots,n}\nu_i \quad \nu\,] \quad F^{-1}$（n）− ν⋆。也就是说，n需要足够大
因此，可能有可能采样良好的限制。另一方面，对于任何固定的N，推杆3表明，总资源预算B需要足够大，以克服γ描述的序列的收敛速度。接下来，我们将SH联系到一个天真的方法，使得将资源统一地分配给固定的n个臂。

5.3.1 Non-Adaptive Uniform Allocation

非自适应均匀分配策略作为输入B和N臂，将B /
N分配给每个臂，并拾取具有最低损耗的臂。以下结果允许我们与续存相比。

命题4假设我们从f绘制n随机配置，j = min {b / n，r}迭代，并让 = arg mini =
1，...，n j（xi）。不损失一般性假设ν1≤。。。≤νn。如果

$$B \geq n\gamma^{-1} \quad \tfrac{1}{2}(F^{-1}(\tfrac{\log(1/\delta)}{n}) - \nu_*) \tag{3}$$

然后概率至少为1 - Δ我们有ν - ν⋆≤2 $F^{-1} \quad \frac{\log(1/\delta)}{n} \quad - \nu_*\Big)$. *In contrast,*
存在满足F和γ的函数序列，使得如果

$$B \leq n\gamma^{-1}\left(2(F^{-1}(\tfrac{\log(c/\delta)}{n+\log(c/\delta)}) - \nu_*)\right)$$

然后概率至少δ，我们有ν - ν⋆≥2（f-1（log（c / δ） $\bar{)} - \nu_*)$, *where c is a*
常数取决于F的规律性。

对于任何固定的N和足够大的B，推子3显示了逐次的
输出满足ν - ν⋆ f-1（log（2 / δ）的 ∈[n] n）- △⋆具有至少1 - δ的概率。
此保证与命题4的结果类似。然而，随后的持续验证达到其保证

$$B \simeq \log_2(n)\left[\log(1/\delta)\gamma^{-1}\left(F^{-1}(\tfrac{\log(1/\delta)}{n}) - \nu_*\right) + n\int_{\frac{\log(1/\delta)}{n}}^{1} \gamma^{-1}(F^{-1}(t) - \nu_*)dt\right], \tag{4}$$

并且该样本复杂性可以基本上小于EQ中所示的均匀分配所需的预算。（3）命题4.基本上，方程式中的第一个术语。（4）表示分配给恒定数量的武器数量的预算νi≈f-1（log（1 / δ）

第二术语描述了在丢弃之前对次优臂进行采样的次数。下一节使用F和γ的特定参数化，以帮助更好地说明均匀分配的样本复杂度与连续阶层（等式4）的样本复杂度之间的差异。

5.3.2 A Parameterization of F and γ for Interpretability

要获得一些直觉并将结果与现有文献相关联，我们对F和γ进行了显式参数假设。我们强调我们的所有结果都按照先前所说的一般F和γ保留，并且此参数化只是一种提供直觉的工具。首先假设存在常数α > 0

$$\gamma(j) \simeq \left(\frac{1}{j}\right)^{1/\alpha}. \tag{5}$$

12.我们说F G如果存在常量c，c'，使得cg（x）≤f（x）≤c'g（x）。

注意，α的大值意味着 i，k→i的收敛非常慢。

我们将首先考虑两个可能的F.参数化，假设存在正面
constants β such that

$$F(x) \simeq \begin{cases} (x-\nu_*)^\beta & \text{if } x \geq \nu_* \\ 0 & \text{if } x < \nu_* \end{cases}. \tag{6}$$

这里，大量的β意味着它非常罕见地绘制接近最佳值的极限ν⋆。在Carpentier和Valko（2015年）中研究了相同的模型。修复一些Δ> 0.如前一节所述，如果n = log（1 / δ）f（ν⋆+Δ） Δ−β对数（1 / δ）臂绘制

F然后概率至少为1 − Δ，我们具有mini = 1，...，n νi≤ ν⋆+Δ。可预测，两者

统一的分配和逐次的输出满足ν −ν⋆ 的ν

$$\overline{\qquad\qquad}$$
$$n$$

概率至少为1 − δ提供其测量预算足够大。因此，如果n≥Δ−β对数（1
/ δ）和均匀分配的测量预算（等式3）和逐次地（等式4）满足

$$\begin{aligned} &\text{Uniform allocation} && B \simeq \Delta^{-(\alpha+\beta)} \log(1/\delta) \\ &\textsc{SuccessiveHalving} && B \simeq \log_2(\Delta^{-\beta}\log(1/\delta)) \left[\Delta^{-\alpha}\log(1/\delta) + \frac{\Delta^{-\beta}-\Delta^{-\alpha}}{1-\alpha/\beta}\log(1/\delta)\right] \\ & && \quad \simeq \log(\Delta^{-1}\log(1/\delta))\log(\Delta^{-1})\ \Delta^{-\max\{\beta,\alpha\}}\log(1/\delta) \end{aligned}$$

然后也满足诸如概率至少1 − Δ.13的概率至少为1 − Δ.13的预算比例，如Δ−max
{α，β}，这可以明显小于均匀分配的Δ−（α+β）。然而，由于α和β在实践中未知，因此任何方法都不知道如何选择最佳N或B以达到这种Δ的精度。在第5.3.3节中，我们展示了超接地解决了这个问题的情况。

F的第二个参数化是以下离散分布：

$$F(x) = \frac{1}{K}\sum_{j=1}^{K} \mathbf{1}\{x \leq \mu_j\} \qquad \text{with} \quad \Delta_j := \mu_j - \mu_1 \tag{7}$$

对于一些独特的标量μ1<μ2<・・<μk。请注意，通过让k→∞这个离散的CDF可以近似任何分段连续的CDF到任意精度。特别地，该模型可以具有多种方式采用相同的值，使得α质量在μ1和1 −
α质量上为μ2>μ1，捕获jamieson等人的随机无限武装炸药模型。（2016）。在此设置中，均匀分配和逐次的输出在顶部日志内的ν （1 / δ）

如果它们的概率至少为1 − δ的k臂的n个馏分
预算足够大。因此，设Q> 0是这样n Q−1日志（1
/ δ）。然后，如果统一分配的测量预算（等式3）和逐次持续

13.这些数量是中间导致第5.3.3节定理的证明。

(Equation 4) satisfy

$$\text{Uniform allocation} \quad B \simeq \log(1/\delta) \begin{cases} K \max_{j=2,\dots,K} \Delta_j^{-\alpha} & \text{if } q = 1/K \\ q^{-1}\Delta_{\lceil qK \rceil}^{-\alpha} & \text{if } q > 1/K \end{cases}$$

$$\text{SuccessiveHalving} \quad B \simeq \log(q^{-1}\log(1/\delta))\log(1/\delta) \begin{cases} \Delta_2^{-\alpha} + \sum_{j=2}^{K} \Delta_j^{-\alpha} & \text{if } q = 1/K \\ \Delta_{\lceil qK \rceil}^{-\alpha} + \frac{1}{qK} \sum_{j=\lceil qK \rceil}^{K} \Delta_j^{-\alpha} & \text{if } q > 1/K, \end{cases}$$

返回处于最佳Q级的臂的臂，即 /k≈Q和 ν − ν⋆ Δ max{2，qk} ，具有至少1 − Δ。这表明每个手臂用于均匀分配的平均资源是需要将顶部Q分数与最佳区分区分，而这对于连续阶层来说是区分臂所需的平均资源的一个小倍数;在实践中最大和平均值之间的差异非常大。我们谨慎地仔细选择了推子3中的 ε 的价值，以使连续的预算和保证解决。另请注意，一个永远不会采取Q <1 / k，因为q = 1 / k足以返回最佳臂。

5.3.3 Hyperband Guarantees

图9的超频算法通过执行所谓的“加倍伎俩”的二维版本来解决武器N的数量与每个被拉动的平均次数之间的折衷。对于每个固定的B，我们不适应地将N个间距的预定网格视为几何上的几何上分开，使得识别“最佳”设置的损失需要预算不超过日志（b）乘以预算，如果最佳设置n的时间提前。然后，我们连续双B使得到达必要的B所需的测量次数不超过2b。这一想法是，即使我们不知道B，n的最佳设置，以达到一些所需的错误率，希望就是通过以特定顺序尝试不同的值，我们不会浪费太多的努力。

FIX δ ∈（0,1）。对于图9的超频算法中定义的所有（k，s）对，让

$\delta_{k,s}$ $\overline{2k^3}$

$$\mathcal{E}_{k,s} := \{B_{k,s} > 4\lceil \log_2(n_{k,s}) \rceil \mathbf{H}(F, \gamma, n_{k,s}, \delta_{k,s})\} = \{2^k > 4s\mathbf{H}(F, \gamma, 2^s, 2k^3/\delta)\}$$

然后由推论3我们有

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} \bigcup_{s=1}^{k} \{\nu_{\hat{\imath}_{k,s}} - \nu_* > 3\big(F^{-1}\big(\tfrac{\log(4k^3/\delta)}{2^s}\big) - \nu_*\big)\} \cap \mathcal{E}_{k,s} \right) \le \sum_{k=1}^{\infty}\sum_{s=1}^{k} \delta_{k,s} = \sum_{k=1}^{\infty} \frac{\delta}{2k^2} \le \delta.$$

对于足够大的K，我们将有 k s = 1 ek，s = ，假设b = 2k足够大。

让 b是从图9的超接头的圆形kb = log2（b） 的逐次持续性最佳性能的臂输出，并且让Sb≤kb是最大的值

$\mathcal{E}_{k_B,s_B}$ holds. Then

$$\nu_{\hat{i}_B} - \nu_* \leq 3\ F^{-1}(\frac{\log(4\lfloor\log_2(B)\rfloor^3/\delta)}{2^{s_B}}) - \nu_*\big) + \gamma(\lfloor\frac{2^{\lfloor\log_2(B)\rfloor-1}}{\lfloor\log_2(B)\rfloor}\rfloor).$$

另请注意，在最多的舞台k上

$$_1 iB_{i,1} \quad k \quad \overset{k}{_{i=1}} B_{i,1} \quad 2kB_{k,s} = 2\log_2(B_{k,s})B_{k,s}$$

已经采取了总样本。虽然该保证持有一般f，γ，Sb的值，因此难以解释。以下推论分别考虑F和γ的β，α参数化部分，分别为5.3.2节以便更好地解释。

定理5假设第5.2节的假设1和2保持，并且采样损耗序列遵守公式5和6的参数假设。固定Δ∈（0,1）。对于任何T∈N，让 t是从耗尽来自所有轮次的T总预算之后从图9的最后一轮k的逐渐持续的经验最佳性能的臂输出。

$$\nu_{\hat{i}_T} - \nu_* \leq c\left(\frac{\overline{\log}(T)^3\overline{\log}(\log(T)/\delta)}{T}\right)^{1/\max\{\alpha,\beta\}}$$

对于某些常量C = exp（o（max{α，β}））），其中log（x）$\overline{=}$log（x）日志log（x）。

通过对证明的直接修改，可以证明如果统一分配
用于代替超接头的职业职位，统一分配版本

achieves $\nu_{\hat{i}_T} \quad \nu \quad c\ \frac{\log(T)\overline{\log}(\log(T)/\delta)}{T}\ ^{1/(\alpha+\beta)}$。我们将上述定理应用于随机的
无限武装的强盗设置在以下推论中。

导体6
[随机无限武装匪徒]对于任何步骤K，S中的任何步骤K，S在具有NK，S武器的无限地平线超频算法中，考虑到第j族的第j次拉动导致随机丢失yi，j∈[0，1]这样E [yi，j] = νi和p（ νi − ν⋆≤ ε ）= C−1

$$_j \quad s = 1\ yi$$，然后s概率至少1 − δ / 2我们具有≥k≥1，0≤s≤k，1≤
$i \leq n_{k,s}, 1 \leq j$

$$|\nu_i - \ell_{i,j}| \leq \sqrt{\frac{\log(B_k n_{k,s}/\delta_{k,s})}{2j}} \leq \sqrt{\log(\frac{16B_k}{\delta})}\left(\frac{2}{j}\right)^{1/2}.$$

因此，如果在B总拉动之后，我们将 νB定义为从最后完全完成的圆形k的经验最佳臂输出的均值，然后具有至少1 − δ

$$B \quad *$$

该必要性的结果与木匠的4.3节的任何时间结果相匹配
Valko（2015）何种算法专门为随机臂的情况和EQ中定义的F的β参数化构成。（6）。值得注意的是，该结果也匹配了较好的下限，从而达到多对数因子，显示出在这个重要的特殊情况下几乎紧张。但是，我们注意到，此早期的工作对固定预算设置更加仔细分析。

定理7假设第5.2节的假设1和2保持并且采样损耗序列遵守等式5和7的参数假设。对于任何T∈N，可以让 t是从最后一轮连续哈欠的经验最佳性能的臂输出耗尽了从所有轮次耗尽的总预算后，超频带的K的k。FIXΔ∈（0,1）和Q≥（1/k，1），让zq=log（q−1）（Δ−α

t=ω（zq log（zq）log（1/δ））通过超带制造的总拉动我们具有 νt− ν⋆≤Δ max{2，qk} ，概率至少为1 − δω（·）隐藏日志日志（·）因子。

吸引转义转型6的随机设置，以至于α=2，我们得出结论

足以识别最佳Q比例的臂的样本复杂度与概率1 − δ，达到日志因子，像日志（1/δ）log（q−1）（Δ−α）

将此结果解释为Bubeck等人的分布依赖性纯探索结果的扩展。（2009）;但是在我们的情况下，当拉动的数量可能远小于武器K时，我们的界限保持。当Q=1/k时，这意味着最佳臂被识别为关于日志（1/Δ）log（k）{δ−2

上限Karnin等人。（2013）;jamieson等。（2014）和下限Kaufmann等。（2015）达到日志因素。因此，对于随机k武装强盗问题，超频带恢复许多已知的样本复杂度导致对数因子。

5.4 Finite Horizon Setting ($R < \infty$)

在本节中，我们分析了第3节中描述的算法，即有限地平线超频带。我们在5.3节中为无限的地平线超频带来了类似的理论保证，幸运的是，大部分分析将被回收。我们说明图10中的连续哈维文和超频算法的有限范围版本。

有限的地平线设置以两种主要方式不同。首先，至少在每个支架中

一个手臂将被拉动R次，但没有臂将被拉出超过R次。其次，括号的数量，每个代表在n和b之间的不同权衡逐步逐步地固定在logη（r）+
1。因此，由于我们是随机IID的采样序列，因此随着时间的推移，增加B即可乘以武器的数量在每个支架中常数，仅通过小常数影响性能。

定理8修复n武器。让νi= i，r和假设ν1≤···≤νn。对于任何ε>0让

$$z_{SH} = \eta(\log_\eta(R)+1)\Big[n + \sum_{i=1}^{n} \min\left\{R, \gamma^{-1}\left(\max\left\{\tfrac{\epsilon}{4}, \tfrac{\nu_i-\nu_1}{2}\right\}\right)\right\}\Big]$$

如果图10的连续减半算法使用任何预算B≥ZSH运行，则返回ARM ，该ARM 满足ν − ν1≤ε/2。

召回此设置中的γ（r）=0，并且定义静脉≥0γ−1（y）≤r。注意

LEMMA 2仍然适用于此设置，就像上面我们获得以下推论。

Corollary 9 *Fix* $\delta \in (0,1)$ *and* ϵ $4(F^{-1}(\frac{\log}{}n) - \nu\star)$。假设h（f，γ，n，δ，ε）如定义在Lemma 2和B=ηlogη（R）中（n+max
{r，h（f，γ，n，δ，ε）））。如果图10的逐次验证算法使用指定的B和N ARM配置运行，则随机绘制

逐步的（有限地平线）输入：预算b，n臂，其中 i，k表示从iTh臂，最大尺寸r， η≥2（默认情况下 η = 3）的kth损失。

^{0} $^{\eta}$

For $k = 0, 1, \ldots, s$

Set $n_k =\ n\eta^{-k}$, $r_k =\ R\eta^{k-s}$

将每只手臂拉到SK的rk时间。

在rkth观察到的损失为SK + 1方面保持最佳 n η −（k + 1） 武器。

Ou $\hat{\imath},R$ $i\in S_{s+1}$ i,R

超频（有限地平线）输入：预算B，最大尺寸R， η≥2（默认情况下 η = 3）

$^{\max}$

For $k = 1, 2, \ldots$

For $s = s_{\max}, s_{\max} - 1, \ldots, 0$

$B_{k,s} = 2^k$, $n_{k,s} = \lceil \frac{2^k \eta^s}{R(s+1)} \rceil$

$\hat{\imath}_s, \ell_{\hat{\imath}_s,R} \leftarrow$ SuccessiveHalving $(B_{k,s}, n_{k,s}, R, \eta)$

图10：启发了有限的地平线和超频算法

通过他们无限的地平线对应物，可以处理实际的限制。超带呼叫作为子程序连续呼叫。

根据F，返回ARM ∈[n]，使得具有至少1 − Δ的概率，我们有 ν − ν⋆≤ + ε / 2。特别地，如果b =4 log2（n） （f， γ，n， δ）和 $\overline{\quad}$ n $*$ $\hat{\imath}$ $*$ $\overline{\quad}$ n $*$ 概率至少为1 − δ。

如第5.3.2节中，我们可以应用 α， β 参数化以获得可解释性，与抑制抑制≥0 γ −1（y）≤r的附加约束使 γ（j） 1j<r $\frac{1}{j}^{1/\alpha}$. Note that the在eq中给出的继承alving的近似样本复杂性。（4）仍然适用于有限地平线算法。

修复一些 Δ > 0， Δ ∈（0,1），并应用EQ的参数化。（6）我们认识到即，如果n ∈ Δ − β 对数（1 / δ）和足够的采样预算（处理 η 作为绝对常数）的均匀分配（等式3）和连续哈欠（方程式（4））满足

$$\text{Uniform allocation} \qquad B \simeq R\Delta^{-\beta}\log(1/\delta)$$

$$\text{SuccessiveHalving} \qquad B \simeq \log(\Delta^{-1}\log(1/\delta))\log(1/\delta)\left[R + \Delta^{-\beta}\frac{1-(\alpha/\beta)R^{1-\beta/\alpha}}{1-\alpha/\beta}\right]$$

然后，两个也满足概率至少1 − δ 的 ν − ν⋆ δ。回顾更大的 α 意味着较慢的收敛性，并且较大的 β 意味着采样良好的限制难度难度，注意当 α / β <1 的逐步的预算如r + δ − β 对数（1 / δ）相同时 α / β → ∞ 预算渐近r δ − β 对数（1 / δ）。

我们还可以应用EQ的离散CDF参数化。（7）。对于任何q（0,1），如果
n q−1 log（1 / δ ）和均匀分配（等式3）的测量预算和逐次地（等式4）满足

$$\text{Uniform allocation:} \qquad B \simeq \log(1/\delta) \begin{cases} K \min\left\{R, \max\limits_{j=2,\ldots,K} \Delta_j^{-\alpha}\right\} & \text{if } q = 1/K \\ q^{-1} \min\{R, \Delta_{\lceil qK \rceil}^{-\alpha}\} & \text{if } q > 1/K \end{cases}$$

SuccessiveHalving:

$$B \simeq \log(q^{-1}\log(1/\delta))\log(1/\delta) \begin{cases} \min\{R, \Delta_2^{-\alpha}\} + \sum\limits_{j=2}^{K} \min\{R, \Delta_j^{-\alpha}\} & \text{if } q = 1/K \\ \min\{R, \Delta_{\lceil qK \rceil}^{-\alpha}\} + \frac{1}{qK} \sum\limits_{j=\lceil qK \rceil}^{K} \min\{R, \Delta_j^{-\alpha}\} & \text{if } q > 1/K \end{cases}$$

然后返回处于最佳Q级臂的臂，即 /k≈Q和 ν − ν⋆ Δ max{2，Qk} ，具有至少1 − δ 。我们再次遵守均匀分配和连续阶层之间的显着差异，特别是当 Δ − α 时

values of j $1, \ldots, n$.

拥有推导性推导9，所有关于第5.3.3节的讨论前定值5持有

对于有限情况（R
<∞）也是如此。可预见的类似定理也适用于有限的地平线设置，但它们的特定形式（具有Polylog因子）不提供超出足以逐步持续成功的样本复杂性的额外洞察，立即给出。

重要的是要注意，在有限的地平线设置中，对于所有足够大的B（例如，

B> 3R）和所有分布F，逐次持续的预算B应与N Δ − β 对数（1
/ δ ）线性缩放为 Δ→0.将其与其中B到N的比率相比基于 α ， β 为 δ →0的值变得无界。该观察结果的一个结果是在有限的地平线设置中，它足以使B足以识别 Δ − 良好的臂，只需恒定概率，例如持续概率，比如1/10，然后重复逐步的m次以提升这种恒定概率1 − （9

虽然在这种超接管的理论治疗中，我们随着时间的推移而增长B，但在实践中，我们建议将B固定为正如我们在第3节中所做的那样的倍数。由于恒定的有限预算版本的固定预算版本更适合实际应用。时间，而不是指数时间，在每个外环中训练到完成的配置之间。

6. Conclusion

我们通过讨论与分布式计算的并行化超频有关的三个潜在扩展，调整具有不同收敛速率的培训方法，以及与非随机采样方法相结合的超接头。

分布式实现。超接管具有自平行化的可能性

武器是独立的，随机抽样。最简单的并行化方案是将连续阶层的单个括号分配给不同的机器。这可以是异步完成的，并且可以释放机器，可以启动新的括号

用不同的武器。人们还可以并行化单个支架，以便每轮连续持续运行速度更快。这种方法的一个缺点是，如果R可以在一台机器上计算，则当臂呈现下来，任务数量指数逐渐减小，因此必须管理更复杂的作业优先级队列。利用HyperBand的平行概括，有效地利用大规模分布式集群，同时最小化开销成本是未来工作的有趣途径。

调整不同的收敛速率。第二个开放挑战涉及Gen−

将超频背后的想法释放到配置，其中配置具有急剧不同的收敛速率。如果它们具有冲击收敛的超参数（例如，用于具有不同数量的层次或隐藏单元的神经网络的学习速率衰减），以及/或者对应于不同的模型系列（例如，深网络（例如，深网络与决策树）。核心问题发生在速度较慢的收敛速率最终导致更好的模型。为了解决这些问题，我们应该能够调整分配给每个配置的资源，以便在消除时可以进行公平的比较。

包含非随机抽样。最后，超接口可以受益于不同 −

校友采样方案除了简单的随机搜索之外。在Bercstra和Bengio（2012年）研究的Sobol或Latin Hypercub等准随机方法可以通过提供更好地覆盖搜索空间来改善超频带的性能。或者，元学习可用于定义以前的实验通知的智能前瞻（Feurer等，2015）。最后，如第2节所述，探索与自适应配置选择策略相结合超频的方法是一个非常有前途的未来方向。

Acknowledgments

KJ由ONR奖励N00014−15−1−2620和N00014−13−1−0129支持。AT由谷歌教师奖和AWS在教育研究赠款奖中得到支持。

附录A.额外的实验结果

下面提供了第3节和第4节的实验的其他细节。

A.1 LeNet Experiment

第3.3节中讨论的Lenet示例的搜索空间如表2所示。

Hyperparameter	Scale	Min	Max
Learning Rate	log	1e-3	1e-1
Batch size	log	1e1	1e3
Layer-2 Num Kernels (k2)	linear	10	60
Layer-1 Num Kernels (k1)	linear	5	k2

表2：第3.3节Lenet应用的超参数空间。请注意
第1层中的内核数量是由第2层中的核数的上限定。

A.2使用Alex Krizhevsky的CNN架构实验

对于第4.1节中讨论的实验，所使用的确切架构是CUDA−CONCNET的18％型号，用于CIFAR−10.14

搜索空间：用于实验的搜索空间如表3所示

学习速率缩短超级计表明在最大迭代窗口中，学习速率减少了多少次。例如，在具有30,000的最大迭代的CiFar−10上，2的学习速率降低对应于每10,000次迭代减少学习，总共超过30,000次迭代窗口。所有的超参数，除了学习率衰减

减少，与snoek等人的重叠。（2012）。Snoek等人的两个封上参数。（2012）被排除在我们的实验之外：（1）由于Caffe框架的限制和（2）被排除了响应归一化层的宽度，因为它与动态资源分配不兼容，则排除了时期的数量。

数据分割：对于CiFar−10，培训（40,000实例）和验证（10,000

实例从数据批量中采样，具有平衡类别的数据批次。原始测试集（10,000实例）用于测试。对于MRBI，培训（10,000实例）和验证（2,000实例）从具有平衡类的原始培训集中采样。原始测试集（50,000实例）用于测试。最后，对于SVHN，火车，验证和测试分裂是使用与Sermanet等人相同的过程创建的。（2012）。

与早期停止比较：Domhan等人。（2015）提出了早期停止

神经网络的方法并将其与SMAC结合起来加速Quand参数光学率。如果配置击败当前最佳的概率低于指定阈值，则其方法停止训练配置。通过推断学习曲线适合配置的中间验证误差损耗来估计这种概率。

14.模型规范可在http://code.google.com/p/cuda−convnet/提供。

Hyperparameter	Scale	Min	Max
Learning Parameters			
初始学习速率log 5 ⋆ 10−5 5			
Conv2 L_2 Penalty	log	$5 * 10^{-5}$	5
Conv3 L_2 Penalty	log	$5 * 10^{-5}$	5
FC4 L_2 Penalty	log	$5 \quad 10^{-3}$	500
学习速率缩减整数0 3			
Local Response Normalization			
Scale	log	$5 * 10^{-6}$	5
Power	linear	0.01	3

表3：三层卷积网络的超参数和相关范围。

如果提前终止配置，则从估计的学习曲线中预测的终端值被用作传递给超参数优化算法的验证错误。因此，如果学习曲线拟合差，则可能会影响配置选择算法的性能。虽然这种方法是启发性的，但它可以在实践中运作得很好，因此我们将超接带与早期终止进行了比较（图11中标记的SMAC（早期））。我们使用默认参数的保守终止标准，每400次迭代记录验证损耗，并在训练期内评估终止标准（CIFAR−10和MRBI的每8K迭代以及SVHN的每16K迭代）.15比较性能由于R的总迭代的数量是保守的，因为它不考虑拟合学习曲线的时间以检查终止标准。

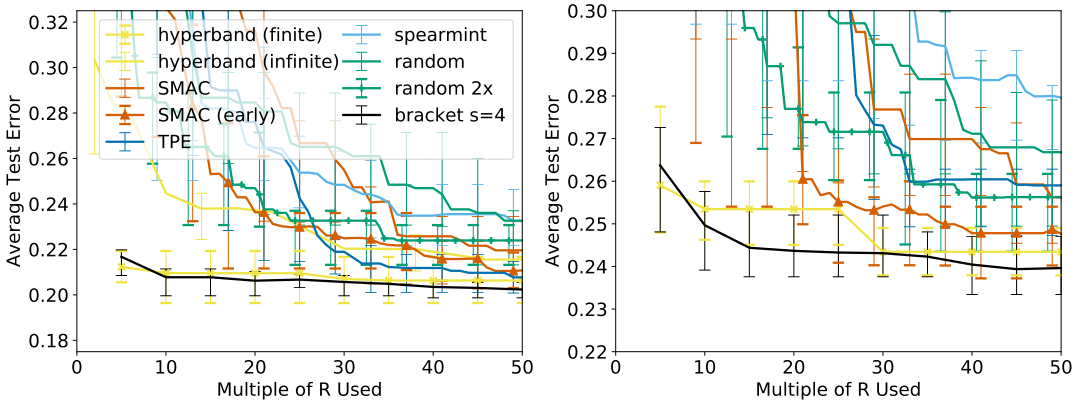
A.3 117 Data Sets Experiment

对于第4.2.1节中讨论的实验，我们遵循Feurer等。（2015）并强加了3GB内存限制，每个HyperParameter配置的6分钟超时，一个小时时间窗口，以评估每个数据集的每个搜索者。此外，我们通过在所有数据集中聚合结果并报告每个方法的平均等级来评估每个搜索者的性能。具体而言，小时训练窗口被分成30个间隔，并且在每个时间点，该模型当时使用最佳验证错误的模型用于计算每个试验的平均错误（数据集搜索者）一对。然后，每个搜索者的性能由数据集排序并在所有数据集中取平均。在Google
Cloud中执行所有实验，在US−Central1−F区域中使用1 CPU和3.75GB内存中的US−Central1−F区域进行。

数据分裂：Feurer等。（2015）将每个数据分成2/3培训和1/3测试集，

虽然我们介绍了一个验证集，以避免对测试数据过度接受。我们还使用了2/3的数据进行培训，但将其余数据拆分为两个同样大小的验证和测试集。我们报告了验证和测试数据的结果。而且，我们进行了

15.我们使用了Https://github.com/utoml/pylearningcurvepredictor提供的代码。



(a) CIFAR-10

(b) MRBI

(c) SVHN

图11：在所有图中都显示了10个试验中的平均测试错误。错误栏表示
顶部和底部四分位数的测试错误与模型相对应的最佳验证错误

每个（数据集搜索者）对的20个试验，以及Feurer等人。（2015）我们在跨试验中保持相同的数据拆分，同时在每次试验中使用不同的随机种子。

实验设置的缺点：该基准含有各种各样的品种
训练设定尺寸和特征尺寸16导致随机搜索能够在某些数据集上测试600个配置，但其他数据集只是数十种。利用数据集大小至少线性地缩放的隐式假设，设计了超接头。对于在秒为单位培训的非常小的数据集，初始化开销主导了计算和限制不提供计算益处。另外，许多正在考虑的分类器和预处理方法返回内存错误，因为它们需要存储Quadratic在特征数（例如，协方差矩阵）或观察次数（例如，内核方法）中。这些错误通常立即发生（从而浪费时间）;但是，它们经常发生在完整数据集上，而不是在限制数据集上。像使用子采样数据集的超带带有的搜索者都可以在试图在完整数据集上训练时，仅在子样本上花费大量时间训练。

A.4 Kernel Classification Experiments

表4显示了第4.2.2节中讨论的内核最小二乘分类实验中考虑的超参数和相关范围。

Hyperparameter	Type	Values
preprocessor	Categorical	min/max, standardize, normalize
内核分类RBF，多项式，乙状体		
C	Continuous	log $[10^{-3}, 10^{5}]$
gamma	Continuous	log $[10^{-5}, 10]$
degree	if kernel=poly	integer [2, 5]
coef0	if kernel=poly, sigmoid	uniform [-1.0, 1.0]

表4：内核正常化最小二乘分类问题的封锁率空间

discussed in Section 4.2.2.

成本术语C由样本数量除以，以便

当资源增加正则术语λ等于缩放成本术语C的倒数。另外，在图12中显示了跨越10试验的顶部和底部四分位数的平均测试误差。

表5显示了随机考虑的超参数和相关范围

在4.3节中讨论了内核近似分类实验。这

正则术语λ由特征的数量除以，使得平方误差和L2惩罚之间的权衡将随着资源的增加而保持恒定。另外，图13中显示了10个试验中的顶部和底部四分位数的平均测试误差。

16.训练设定尺寸范围为670至73,962个观察，而且特性数量为1到10,935。

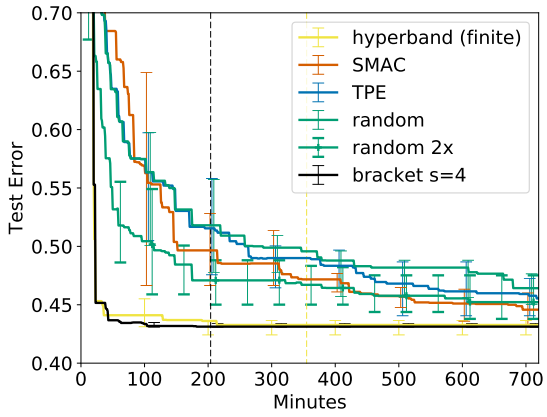


图12：最佳内核的平均测试错误正常化最小方形分类模型

每个搜索者都发现了Cifar−10。颜色编码的虚线表示给定搜索者的最后一次试验完成。误差条对应于10项试验中测试错误的顶部和底部四分位数。

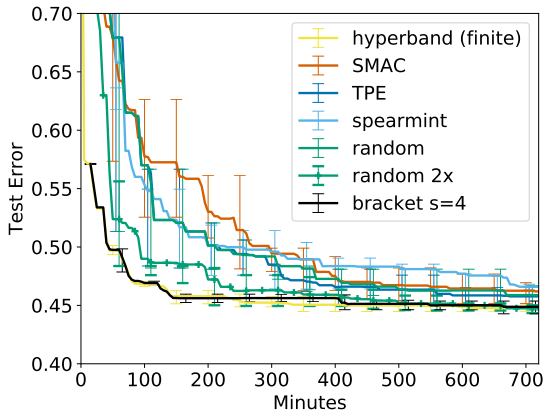


图13：每个搜索者找到的最佳随机功能模型的平均测试错误

CiFar−10。在每次评估中而不是支架末端时计算超带和支架S = 4的测试误差。误差条对应于10项试验中测试错误的顶部和底部四分位数。

Hyperparameter	Type	Values
预处理器分类无，min / max，标准化，标准化		
λ	Continuous	$\log\,[10^{-3}, 10^{5}]$
gamma	Continuous	$\log\,[10^{-5}, 10]$

表5：随机特征内核近似分类的HyperParameter空间

第4.3节中讨论的问题。

Appendix B. Proofs

在本节中，我们提供了第5节中提出的定理的证据。

B.1 Proof of Theorem 1

证明首先，我们验证了算法从未采用总数的样本
exceeds the budget B:

$$\sum_{k=0}^{\lceil \log_2(n)\rceil -1} |S_k| \left\lfloor \frac{B}{|S_k|\lceil \log(n)\rceil} \right\rfloor \leq \sum_{k=0}^{\lceil \log_2(n)\rceil -1} \frac{B}{\lceil \log(n)\rceil} \leq B\,.$$

出于符号，让 i，j：= j（xi）。同样，对于每个I∈[n]：= {1，。。。我们假设的n}
LIMIT LIMK→∞∞i，k存在，等于νi。作为提醒，γ：n→r被定义为点亮最小，单调减少功能令人满意

$$\max_i |\ell_{i,j} - \nu_i| \leq \gamma(j)\ , \quad \forall j \in \mathbb{N}. \tag{8}$$

注意γ是通过存在的存在而存在的，并且随着迭代序列的序列增加而界定与极限值的偏差。

不损失一般性，使终端损耗顺序使得ν1≤ν2≤···≤νN。认为
b≥zsh。然后我们为每一轮k提供

$$\begin{aligned}
r_k &\geq \frac{B}{|S_k|\lceil \log_2(n)\rceil} - 1 \\
&\geq \frac{2}{|S_k|} \max_{i=2,\ldots,n} i\left(1 + \gamma^{-1}\big(\max\big\{\frac{\epsilon}{4}, \frac{\nu_i - \nu_1}{2}\big\}\big)\right) - 1 \\
&\geq \frac{2}{|S_k|}\,(\lfloor |S_k|/2 \rfloor + 1)\left(1 + \gamma^{-1}\big(\max\big\{\frac{\epsilon}{4}, \frac{\nu_{\lfloor |S_k|/2\rfloor +1} - \nu_1}{2}\big\}\big)\right) - 1 \\
&\geq \left(1 + \gamma^{-1}\big(\max\big\{\frac{\epsilon}{4}, \frac{\nu_{\lfloor |S_k|/2\rfloor +1} - \nu_1}{2}\big\}\big)\right) - 1 \\
&= \gamma^{-1}\big(\max\big\{\frac{\epsilon}{4}, \frac{\nu_{\lfloor |S_k|/2\rfloor +1} - \nu_1}{2}\big\}\big),
\end{aligned}$$

其中第四线遵循 |sk|/2≥|sk|/2 − 1。

首先，我们展示了所有T≥II的 i，t− 1，t> 0：=γ−1 i−ν1 。鉴于γ的定义，

我们拥有所有我的∈[n] | i，t − νi|≤γ（t）≤νi−ν2在最后的不等式适用于t≥τi。

因此，对于我们的T≥τi

$$\begin{aligned}\ell_{i,t} - \ell_{1,t} &= \ell_{i,t} - \nu_i + \nu_i - \nu_1 + \nu_1 - \ell_{1,t} \\ &= \ell_{i,t} - \nu_i - (\ell_{1,t} - \nu_1) + \nu_i - \nu_1 \\ &\geq -2\gamma(t) + \nu_i - \nu_1 \\ &\geq -2\frac{\nu_i - \nu_1}{2} + \nu_i - \nu_1 \\ &= 0.\end{aligned}$$

在这种情况下，我们将在手臂1之前消除臂我，因为在每个围绕臂被其经验损失排序，上半部分被丢弃。注意，假设νI限制在I中是非减小的，因此τi值在I中是非增加的。

修复圆形k并假设1∈sk（注意，1∈s0）。以上计算显示

$$t \geq \tau_i \implies \ell_{i,t} \geq \ell_{1,t}. \tag{9}$$

Consequently,

$$\begin{aligned}\{1 \in S_k,\ \ 1 \notin S_{k+1}\} &\iff \left\{ \sum_{i \in S_k} \mathbf{1}\{\ell_{i,r_k} < \ell_{1,r_k}\} \geq \lfloor |S_k|/2 \rfloor \right\} \\ &\implies \left\{ \sum_{i \in S_k} \mathbf{1}\{r_k < \tau_i\} \geq \lfloor |S_k|/2 \rfloor \right\} \\ &\implies \left\{ \sum_{i=2}^{\lfloor |S_k|/2 \rfloor + 1} \mathbf{1}\{r_k < \tau_i\} \geq \lfloor |S_k|/2 \rfloor \right\} \\ &\iff \ r_k < \tau_{\lfloor |S_k|/2 \rfloor + 1} \ .\end{aligned}$$

在算法的定义下，第一行的定义如下，第二由等式9，以及第三个由τi是非增加的（对于所有I<j我们有τi≥τj，并且因此，1{rk<τi}≥1{RK<τj}所以第一个不包括1的第一个指标将在任何其他我在sk [n]洒在整个[n]之前。这意味着

$$\{1 \in S_k,\ \ r_k \geq \tau_{\lfloor |S_k|/2 \rfloor + 1}\} \implies \{1 \in S_{k+1}\}. \tag{10}$$

Recalling that $r_k \quad \gamma^{-1} \ \max \ \frac{\epsilon}{4}, \frac{\nu_{\lfloor |S_k|/2 \rfloor + 1} - \nu_1}{2}\}\big)$ and $\tau_{\lfloor |S_k|/2 \rfloor + 1} = \gamma^{-1}\big(\frac{\nu_{\lfloor |S_k|/2 \rfloor + 1} - \nu_1}{2}\big)$,
我们检查以下三种详尽案例：

- **Case 1:** $\frac{\nu_{\lfloor |S_k|/2 \rfloor + 1} - \nu_1}{2} \geq \frac{\epsilon}{4}$ and $1 \in S_k$

  In this case, $r_k \geq \gamma^{-1}\big(\frac{\nu_{\lfloor |S_k|/2 \rfloor + 1} - \nu_1}{2}\big) = \tau \quad |\,\text{sk}\,|/2 \quad + 1$。通过等式10我们有那个 $1 \in S_{k+1}$ since $1 \in S_k$.

- **Case 2:** $\frac{\nu_{\lfloor |S_k|/2 \rfloor + 1} - \nu_1}{2} < \frac{\epsilon}{4}$ and $1 \in S_k$

  In this case r_k γ^{-1} $\frac{\epsilon}{4}$ but γ^{-1} $\frac{\epsilon}{4}$ $< \tau$ | sk | /2 + 1。等式10表明它可能
  1 ∈ sk但1 / ∈ sk + 1可以。在1 ∈ sk + 1的好情况下，算法在下一个圆周上持续，也可以是2例。所以假设1 / ∈ sk + 1。在这里，我们显示{1 ∈ sk，1 / ∈ sk + 1} = maxi ∈ sk+1 ν i⩽ ν 1+ ε / 2。

  因为1 ∈ S0，这保证了逐次使用ARM i= 1或一些臂 i，满足 ν i⩽ ν 1+ ε / 2。

  4}。请注意，P> | SK | /2 + 1由标准的标准

$$r_k \geq \gamma^{-1} \frac{\epsilon}{4} \geq \gamma^{-1} \frac{\nu_i - \nu_1}{2} = \tau_i, \quad \forall i \geq p.$$

  因此，通过等式9（t⩾ τ i= i，t⩾ 1，t）我们具有臂i⩾p将始终具有 i，rk⩾1，rk并以前或同时消除ARM 1，推测1 ∈ sk。总之，如果扶手1被消除，因此通过P的定义，Maxi ń sk+ 1 ν i⩽maxi<p ν i< ν 1+ ε / 2。

- **Case 3:** $1 \notin S_k$

  由于1 S0，存在一些R <K，使得1 ÷ SR和1 / ∈ SR+ 1。对于该R，只有案例1可以才能增殖1‰SR + 1。但是，在案例2下，如果1 / ∈ sr + 1则maxi ∈ sr+1 ν i⩽ ν 1 ± 1 ± 1 ± 1 ± 1。

  因为1 ∈ s0，我们要么保留在sk中（可能在案件之间交替
对于所有K，直到算法用最佳臂1退出，或者存在一些K，使得壳体3是真实的，并且算法用臂离开，使得 ν i⩽ ν 1 ε / 2。证明是通过注意到的

$$i, \lfloor \frac{B/2}{\lceil \log_2(n) \rceil} \rfloor \quad 1 \qquad i, \lfloor \frac{B/2}{\lceil \log_2(n) \rceil} \rfloor \quad \widehat{i} \qquad \widehat{i} \quad 1$$

通过三角形不等式，因为B⩾2 log2（n） Γ −1（ ε / 4）通过假设。

第二，宽松，但也许更可取的Zsh形式遵循这一事实
γ −1（x）在x中是非增加的

$$\max_{i=2,\ldots,n} i\,\gamma^{-1}\left(\max\left\{\tfrac{\epsilon}{4}, \tfrac{\nu_i - \nu_1}{2}\right\}\right) \leq \sum_{i=1,\ldots,n} \gamma^{-1}\left(\max\left\{\tfrac{\epsilon}{4}, \tfrac{\nu_i - \nu_1}{2}\right\}\right).$$

■

B.2 Proof of Lemma 2

Proof Let $p_n = \frac{\log(2/\delta)}{n}$, $M = \gamma^{-1}\left(\frac{\epsilon}{16}\right)$, and $\mu = \mathbb{E}[\min\{M, \gamma^{-1}\left(\frac{\nu_i - \nu_*}{4}\right)\}]$. Define the events

$$\xi_1 = \{\nu_1 \leq F^{-1}(p_n)\}$$

$$\xi_2 = \left\{\sum_{i=1}^{n} \min\{M, \gamma^{-1}\left(\tfrac{\nu_i - \nu_*}{4}\right)\} \leq n\mu + \sqrt{2n\mu M \log(2/\delta)} + \tfrac{2}{3} M \log(2/\delta)\right\}$$

Note that $\mathbb{P}(\xi_1^c) = \mathbb{P}(\min_{i=1,\ldots,n} \nu_i > F^{-1}(p_n)) = (1-p_n)^n \leq \exp(-np_n) \leq \frac{\delta}{2}$. Moreover, $\mathbb{P}(\xi_2^c) \leq$ 2由伯恩斯坦的不等式

$$\mathbb{E}\ \min\{M, \gamma^{-1}\ \tfrac{\nu_i-\nu_*}{4}\ \}^2\ \leq \mathbb{E}\ M\min\{M, \gamma^{-1}\ \tfrac{\nu_i-\nu_*}{4}\ \}\ = M\mu.$$

因此，P（ξ1∩2）≥1 − δ如下，如下所示，这些事件持有。

首先，我们表明，如果ν⋆≤ ν1≤f−1（上），我们将参考等式（⋆），然后

ma

Case 1: $\frac{\epsilon}{4} \leq \frac{\nu_i-\nu_1}{2}$ and $\epsilon \geq 4(F^{-1}(p_n) - \nu_*)$.

$$\frac{\nu_i-\nu_1}{2} \overset{(*)}{\geq} \frac{\nu_i-\nu_*+\nu_*-F^{-1}(p_n)}{2} = \frac{\nu_i-\nu_*}{4} + \frac{\nu_i-\nu_*}{4} - \frac{F^{-1}(p_n)-\nu_*}{2} \overset{(*)}{\geq} \frac{\nu_i-\nu_*}{4} + \frac{\nu_i-\nu_1}{4} - \frac{F^{-1}(p_n)-\nu_*}{2} \overset{\text{Case 1}}{\geq} \frac{\nu_i-\nu_*}{4}.$$

Case 2: $\frac{\epsilon}{4} > \frac{\nu_i-\nu_1}{2}$ and $\epsilon \geq 4(F^{-1}(p_n) - \nu_*)$.

$$\frac{\nu_i-\nu_*}{4} = \frac{\nu_i-\nu_1}{4} + \frac{\nu_1-\nu_*}{4} \overset{\text{Case 2}}{<} \frac{\epsilon}{8} + \frac{\nu_1-\nu_*}{4} \overset{(*)}{\leq} \frac{\epsilon}{8} + \frac{F^{-1}(p_n)-\nu_*}{4} \overset{\text{Case 2}}{<} \frac{\epsilon}{4}$$

这显示了所需的结果。

因此，对于我们拥有的任何ε≥4（f−1（pn）− ν⋆）

$$\begin{aligned}
\sum_{i=1}^n \gamma^{-1}\left(\max\left\{\tfrac{\epsilon}{4}, \tfrac{\nu_i-\nu_1}{2}\right\}\right) &\leq \sum_{i=1}^n \gamma^{-1}\left(\max\left\{\tfrac{\epsilon}{4}, \tfrac{\nu_i-\nu_*}{4}\right\}\right) \\
&\leq \sum_{i=1}^n \gamma^{-1}\left(\max\left\{\tfrac{\epsilon}{16}, \tfrac{\nu_i-\nu_*}{4}\right\}\right) \\
&= \sum_{i=1}^n \min\{M, \gamma^{-1}\left(\tfrac{\nu_i-\nu_*}{4}\right)\} \\
&\leq n\mu + \sqrt{2n\mu M\log(1/\delta)} + \tfrac{2}{3}M\log(1/\delta) \\
&\leq \left(\sqrt{n\mu} + \sqrt{\tfrac{2}{3}M\log(2/\delta)}\right)^2 \leq 2n\mu + \tfrac{4}{3}M\log(2/\delta).
\end{aligned}$$

A direct computation yields

$$\begin{aligned}
\mu &= \mathbb{E}[\min\{M, \gamma^{-1}\left(\tfrac{\nu_i-\nu_*}{4}\right)\}] \\
&= \mathbb{E}[\gamma^{-1}\left(\max\left\{\tfrac{\epsilon}{16}, \tfrac{\nu_i-\nu_*}{4}\right\}\right)] \\
&= \gamma^{-1}\left(\tfrac{\epsilon}{16}\right) F(\nu_* + \epsilon/4) + \int_{\nu_*+\epsilon/4}^{\infty} \gamma^{-1}(\tfrac{t-\nu_*}{4}) dF(t)
\end{aligned}$$

so that

$$\begin{aligned}
&\sum_{i=1}^n \gamma^{-1}\left(\max\left\{\tfrac{\epsilon}{4}, \tfrac{\nu_i-\nu_1}{2}\right\}\right) \leq 2n\mu + \tfrac{4}{3}M\log(2/\delta) \\
&\quad = 2n\int_{\nu_*+\epsilon/4}^{\infty} \gamma^{-1}(\tfrac{t-\nu_*}{4}) dF(t) + \left(\tfrac{4}{3}\log(2/\delta) + 2nF(\nu_*+\epsilon/4)\right)\gamma^{-1}\left(\tfrac{\epsilon}{16}\right)
\end{aligned}$$

完成证明。 ■

B.3 Proof of Proposition 4

我们将命令分解为上下边界，并分别证明它们。

B.4 Uniform Allocation

命题10假设我们从F的N随机配置中汲取N，每一个预算J，17，让 = arg mini = 1，...，n j（xi）。让νi= ⋆（xi）且不丧失一般性假设ν1⩽。。。⩽νn。如果

$$B \geq n\gamma^{-1} \quad \tfrac{1}{2}(F^{-1}(\tfrac{\log(1/\delta)}{n}) - \nu_*) \tag{11}$$

然后概率至少为1 − Δ我们有ν − ν⋆⩽2 $\left(F^{-1}\left(\frac{\log(1/\delta)}{n}\right) - \nu_*\right)$.

证明请注意，如果我们从f和i⋆ = arg mini = 1，...，n ⋆（xi）然后

$$\begin{aligned}\mathbb{P}\left(\ell_*(X_{i_*}) - \nu_* \leq \epsilon\right) &= \mathbb{P}\Big(\bigcup_{i=1}^{n}\{\ell_*(X_i) - \nu_* \leq \epsilon\}\Big) \\ &= 1 - (1 - F(\nu_* + \epsilon))^n \geq 1 - e^{-nF(\nu_*+\epsilon)},\end{aligned}$$

这相当于用概率至少1−δ，χ⋆（xi⋆）− ν⋆⩽f−1（log（1/δ）/n）− ⋆ ⋆。此外，如果每个配置训练用于J迭代，那么概率至少为1 − Δ

$$\begin{aligned}&\ell_*(X_i) - \nu_* \leq \ell_j(X_i) - \nu_* + \gamma(j) \leq \ell_j(X_{i_*}) - \nu_* + \gamma(j) \\ &\leq \ell_*(X_{i_*}) - \nu_* + 2\gamma(j) \leq F^{-1}\left(\frac{\log(1/\delta)}{n}\right) - \nu_* + 2\gamma(j).\end{aligned}$$

如果我们的测量预算B受约束，则B = NJ然后根据B和N求解J，产生结果。

以下命题表明均匀误差上的上限

命题4中的分配策略实际上是紧张的。也就是说，对于任何分发F和功能γ，存在需要在EQ中描述的预算的损耗序列。（3）为了避免高概率超过ε的损失。

命题11修复任何δ ∈（0,1）和n∈n.对于任何C∈（0,1]，让Fc表示连续累积分布函数f满意的18infx∈[ν⋆，1−ν⋆] infΔ ∈[0,1−x]

$\frac{(x+\Delta/2)}{-F(x)} \geq$

C。让γ表示在N的单调减小功能的空间。对于任何F∈Fc和γ ∈ γ，存在概率分布μ通过x和函数序列 j：x→r j ∈n∈n∈⋆：= limj→∞∞j， ν⋆ =infx∈x ⋆（x）这样

17.这里可以是有界（有限地平线）或无限的（无限地平线）。18.请注意，每当F凸起时满足该条件。此外，如果f（ν⋆ + ε）= C−1 1Eβ然后很容易

2

supx∈x| j（x）− ⋆（x）|≤γ（j）和pμ（ ⋆（x）− ν⋆≤ε）=
f（ε）。此外，如果n配置x1，。。。，Xn由μ和 =argmini∈1，...，n b/ n（xi）绘制，然后至少δ

$$\ell_*(X_{\hat{i}}) - \nu_* \geq 2(F^{-1}(\tfrac{\log(c/\delta)}{n+\log(c/\delta)}) - \nu_*)$$

whenever $B \leq n\gamma^{-1}\left(2(F^{-1}(\tfrac{\log(c/\delta)}{n+\log(c/\delta)}) - \nu_*)\right)$.

证明设定x = [0,1]， ⋆（x）= f −1（x），μ是均匀分布在[0,1]上。定义 ν = f −1（log（c / δ）

$$^{j} \qquad \ell_*(x) \qquad \text{otherwise.}$$

基本上，如果 ⋆（x）在1之内 $\gamma(j)$ of $\nu + 2\gamma$（j）然后我们设置 j（x）等于 ⋆（x）反映
横跨2 +γ（j）。显然，| j（x）− ⋆（x）|≤γ（j）全x∈X.

由于每个 ⋆（xi）根据f分发，我们有

$$\mathbb{P}\bigcap_{i=1}^{n}\{\ell_*(X_i) - \nu_* \geq \epsilon\} = (1 - F(\nu_* + \epsilon))^n \geq e^{-nF(\nu_*+\epsilon)/(1-F(\nu_*+\epsilon))}.$$

设置大于或等于δ / c的右侧并求解ε，找到ε⋆ + ε ≥f−1（log（c / δ）

Define $I_0 = [\nu_*, \nu)$, $I_1 = [\nu, \nu + \frac{}{2}\gamma(B/n))$ and $I_2 = [\nu + \frac{}{2}\gamma(B/n), \widehat{\nu} + \gamma(B/n)]$. Furthermore, for $j \quad 0, 1, 2$ define $N_j = \sum_{i=1}^{n} \mathbf{1}_{\ell\ (X_i)\ \ I_j}$. Given $N_0 = 0$ (which occurs with 概率至少Δ / c），如果n1 = 0那么 ⋆（x ）− ν⋆≥f−1（log（c / Δ） $\overline{\ }) + \frac{1}{2}\gamma(B/n)$ and the

下面我们将显示，如果n2> 0，每当n1> 0时，索赔也是如此。我们现在
表明，当任何M> 0的N1 + N2 = M时，至少概率c发生这种情况。观察

$$\begin{aligned}\mathbb{P}(N_2 > 0 | N_1 + N_2 = m) &= 1 - \mathbb{P}(N_2 = 0 | N_1 + N_2 = m) \\ &= 1 - (1 - \mathbb{P}(\nu_i \in I_2 | \nu_i \in I_1 \cup I_2))^m \geq 1 - (1-c)^m \geq c\end{aligned}$$

since

$$\mathbb{P}(\nu_i \in I_2 | \nu_i \in I_1 \cup I_2) = \frac{\mathbb{P}(\nu_i \in I_2)}{\mathbb{P}(\nu_i \in I_1 \cup I_2)} = \frac{\mathbb{P}(\nu_i \in [\widehat{\nu} + \frac{1}{2}\gamma, \widehat{\nu} + \gamma])}{\mathbb{P}(\nu_i \in [\nu, \nu + \gamma])} = \frac{F(\widehat{\nu} + \gamma) - F(\widehat{\nu} + \frac{1}{2}\gamma)}{F(\nu + \gamma) - F(\nu)} \geq c\,.$$

因此，每当n0 = 0和n2> 0时，每当n1> 0发生概率至少Δ / c・c
= δ时，事件的概率可能在下面的情况下假设这是这种情况。

由于N0 = 0，对于所有J∈N，每个XI必须落入三种情况之一：

1. $\ell_*(X_i) > \widehat{\nu} + \gamma(j) \iff \ell_j(X_i) > \widehat{\nu} + \gamma(j)$
2. $\widehat{\nu} \leq \ell_*(X_i) < \widehat{\nu} + \frac{1}{2}\gamma(j) \iff \widehat{\nu} + \frac{1}{2}\gamma(j) < \ell_j(X_i) \leq \widehat{\nu} + \gamma(j)$

3. $\nu + \frac{1}{2}\gamma(j) \le \ell_*(X_i) \le \nu + \gamma(j) \iff \nu \le \ell_j(X_i) \le \nu + \frac{1}{2}\gamma(j)$

第一个案例以来，在该制度内，我们有 j（x）= ⋆（x），而最后两个案例持有，因为他们认为 j（x）=2 ν + γ（j）− ⋆（X）。因此，对于任何我，这使得 ⋆（xi）∈i2必须是 j（xi）∈i1的情况，反之亦然。导致N2≥N1> 0，我们得出结论，如果 = argmini b/ n（xi）那么 b/ n（x ）∈i1和

∗ î 2 î ∗ ∗ 2 $\overline{n+\log(c/\delta)}$ ∗ 2

we wish $\nu_{\hat{i}} - \nu_* \le 2(F^{-1}(\mathrm{n} + \log(\mathrm{c}/\delta)) - \nu\star)$ 具有至少δ然后我们需要的

$B/n = j \ge \gamma^{-1}\left(2(F^{-1}(\frac{\log}{n+\log(c/\delta)}) - \nu_*)\right.$. ■

B.5 Proof of Theorem 5

证明步骤1：简化H（f，γ，n，δ）。我们首先简化H（f，γ，n，δ）

只是n，δ，α，β。在如下，我们使用可能与下一个不等式不同的常数C，但仍然是依赖于α，β的绝对常数。让pn = log（2 / δ）

$$\gamma^{-1}\left(\tfrac{F^{-1}(p_n)-\nu_*}{4}\right) \le c\left(F^{-1}(p_n) - \nu_*\right)^{-\alpha} \le c\, p_n^{-\alpha/\beta}$$

and

$$\int_{p_n}^1 \gamma^{-1}(\tfrac{F^{-1}(t)-\nu_*}{4})dt \le c\int_{p_n}^1 t^{-\alpha/\beta}dt \le \begin{cases} c\log(1/p_n) & \text{if } \alpha = \beta \\ c\frac{1-p_n^{1-\alpha/\beta}}{1-\alpha/\beta} & \text{if } \alpha \ne \beta. \end{cases}$$

We conclude that

$$\begin{aligned}\mathbf{H}(F,\gamma,n,\delta) &= 2n\int_{p_n}^1 \gamma^{-1}(\tfrac{F^{-1}(t)-\nu_*}{4})dt + \tfrac{10}{3}\log(2/\delta)\gamma^{-1}\left(\tfrac{F^{-1}(p_n)-\nu_*}{4}\right) \\ &\le c p_n^{-\alpha/\beta}\log(1/\delta) + cn\begin{cases}\log(1/p_n) & \text{if } \alpha=\beta \\ \frac{1-p_n^{1-\alpha/\beta}}{1-\alpha/\beta} & \text{if } \alpha\ne\beta.\end{cases}\end{aligned}$$

步骤2：根据δ求解（BK，L，NK，L）。修复Δ> 0。我们的策略是在δ方面描述NK，L。特别地，参数化NK，L使得PNK，L = C日志（4K3 / Δ）

$n_{k,l} = c\Delta^{-\beta}\log(4k^3/\delta)$ so

$$\begin{aligned}\mathbf{H}(F,\gamma,n_{k,l},\delta_{k,l}) &\le c p_{n_{k,l}}^{-\alpha/\beta}\log(1/\delta_{k,l}) + cn_{k,l}\begin{cases}\log(1/p_{n_{k,l}}) & \text{if } \alpha=\beta \\ \frac{1-p_{n_{k,l}}^{1-\alpha/\beta}}{1-\alpha/\beta} & \text{if } \alpha\ne\beta.\end{cases} \\ &\le c\log(k/\delta)\left[\Delta^{-\alpha} + \begin{cases}\Delta^{-\beta}\log(\Delta^{-1}) & \text{if } \alpha=\beta \\ \frac{\Delta^{-\beta}-\Delta^{-\alpha}}{1-\alpha/\beta} & \text{if } \alpha\ne\beta\end{cases}\right] \\ &\le c\log(k/\delta)\min\{\tfrac{1}{|1-\alpha/\beta|}, \log(\Delta^{-1})\}\Delta^{-\max\{\beta,\alpha\}}\end{aligned}$$

最后一行所遵循的地方

$$\Delta^{\max\{\beta,\alpha\}}\frac{\Delta^{-\beta}-\Delta^{-\alpha}}{1-\alpha/\beta} = \beta\frac{\Delta^{\max\{0,\alpha-\beta\}}-\Delta^{\max\{0,\beta-\alpha\}}}{\beta-\alpha}$$
$$= \beta\begin{cases}\frac{1-\Delta^{\beta-\alpha}}{\beta-\alpha} & \text{if } \beta>\alpha\\ \frac{1-\Delta^{\alpha-\beta}}{\alpha-\beta} & \text{if } \beta<\alpha\end{cases} \le c\min\{\tfrac{1}{|1-\alpha/\beta|},\log(\Delta^{-1})\}.$$

使用上行 log（nk，1） ≤clog（log（k/Δ）Δ−1）≤clog（log（k/δ））log（Δ−1）并让zδ= log（δ−1）2Δ−max{β，α}，我们得出结论

$$\begin{aligned}B_{k,l} &< \min\{2^k : 2^k > 4\lceil\log(n_{k,l})\rceil\mathbf{H}(F,\gamma,n_{k,l},\delta_{k,l})\}\\ &< \min\{2^k : 2^k > c\log(k/\delta)\log(\log(k/\delta))z_\Delta\}\\ &\le cz_\Delta\log(\log(z_\Delta)/\delta)\log(\log(\log(z_\Delta)/\delta))\\ &= cz_\Delta\overline{\log}(\log(z_\Delta)/\delta).\end{aligned}$$

步骤3：计算测量总数。此外，输出 k，l之前的测量总数是上限

$$T=\sum_{i=1}^{k}\sum_{j=l}^{i}B_{i,j}\le k\sum_{i=1}^{k}B_{i,1}\le 2kB_{k,1}=2B_{k,1}\log_2(B_{k,1})$$

我们在哪里雇用所谓的“双倍诡计”： k

$$_1\,B_{i,1}=\textstyle\sum_{i=1}^{k}2^i\le 2^{k+1}=2B_{k,i}.$$

Δ Δ Δ Δ

$$^{\beta,\alpha}\}\overline{\log}(\Delta^{-1})^3\overline{\log}(\log(\Delta^{-1})/\delta)$$

根据T获得求解δ

$$\Delta=c\left(\frac{\overline{\log}(T)^3\overline{\log}(\log(T)/\delta)}{T}\right)^{1/\max\{\alpha,\beta\}}.$$

因为输出臂只是经验最好的，所以使用经验估计存在一些错误。返回的手臂返回（k，1）被拉动 2k−1

BK，L / LOG（BK，L）时间，因此可能的错误受γ（BK，L / LOG（BK，L））≤C $\frac{k,l}{B_{k,l}}$ ≤

c $\frac{\log(B)^2\log(\log(B))}{B}$ 这是由上述δ的值占主导地位的。 ■

B.6 Proof of Theorem 7

证明步骤1：简化H（f，γ，n，δ，ε）。我们首先简化H（f，γ，n，δ，ε）

只需n，δ，α，β。如前所述，我们使用可能与下一个不等式不同的常数C，但仍然是绝对常数。让pn = log（2/δ） 首先，我们通过注意到这一点来解决ε

如果ν − ν⋆<Δ2，我们识别最好的手臂。因此，如果ν − ν⋆≤ F $(p_n)-\nu_*$ $+\epsilon/2$ then we set

$$\epsilon=\max\ \ 2\ \Delta_2-\ \ F^{-1}(p_n)-\nu_*\ \ ,4\ \ F^{-1}(p_n)-\nu_*\big)\big\}$$

so that

$$\nu_i - \nu_* \leq \max\left\{3\big(F^{-1}(p_n) - \nu_* \ , \Delta_2 \quad = \Delta_{\lceil \max\{2, cKp_n\}\rceil}.\right.$$

我们在3时对待案件 $\big(F^{-1}(p_n) - \nu_*\big)$ Δ δ 2和替代方案分别。

First assume 3 $F^{-1}(p_n) - \nu_*\big) > \Delta_2$ $, \gamma, n, \delta, \epsilon) =$ $\mathbf{H}(F, \gamma, n, \delta)$. We also have

$$\gamma^{-1}\left(\tfrac{F^{-1}(p_n)-\nu_*}{4}\right) \leq c\left(F^{-1}\left(p_n\right) - \nu_*\right)^{-\alpha} \leq c\,\Delta^{-\alpha}_{\lceil p_n K\rceil}$$

and

$$\int_{p_n}^{1} \gamma^{-1}(\tfrac{F^{-1}(t)-\nu_*}{4})dt = \int_{F^{-1}(p_n)}^{1} \gamma^{-1}(\tfrac{x-\nu_*}{4})dF(x) \leq \frac{c}{K}\sum_{i=\lceil p_n K\rceil}^{K} \Delta_i^{-\alpha}$$

so that

$$\begin{aligned}\mathbf{H}(F, \gamma, n, \delta) &= 2n\int_{p_n}^{1} \gamma^{-1}(\tfrac{F^{-1}(t)-\nu_*}{4})dt + \tfrac{10}{3}\log(2/\delta)\gamma^{-1}\left(\tfrac{F^{-1}(p_n)-\nu_*}{4}\right) \\ &\leq c\Delta^{-\alpha}_{\lceil p_n K\rceil}\log(1/\delta) + \frac{cn}{K}\sum_{i=\lceil p_n K\rceil}^{K}\Delta_i^{-\alpha}.\end{aligned}$$

现在考虑3案3 $\big(F^{-1}(p_n) - \nu_*\big) \leq \Delta_2$. In this case $F(\nu_* + \epsilon/4) = 1/K$, $\gamma^{-1}\ \ \tfrac{\epsilon}{16}\ \ \leq c\Delta_2^{-\alpha}$, and $\ \ \int_{\nu_*+\epsilon/4}^{\infty}\gamma^{-1}(\tfrac{t-\nu_*}{4})dF(t) \leq c\sum_{i=2}^{K}\Delta_i^{-\alpha}$ so that

$$\begin{aligned}\mathbf{H}(F, \gamma, n, \delta, \epsilon) &= 2n\int_{\nu_*+\epsilon/4}^{\infty} \gamma^{-1}(\tfrac{t-\nu_*}{4})dF(t) + \left(\tfrac{4}{3}\log(2/\delta) + 2nF(\nu_* + \epsilon/4)\right)\gamma^{-1}\left(\tfrac{\epsilon}{16}\right) \\ &\leq c(\log(1/\delta) + n/K)\Delta_2^{-\alpha} + \frac{cn}{K}\sum_{i=2}^{K}\Delta_i^{-\alpha}.\end{aligned}$$

步骤2：根据δ求解（BK，L，NK，L）。注意，只有3个PNK，L≤1/ k没有改善

≤Δ2。也就是说，当PNK，L≤1/ k算法时

已经找到了最好的手臂，但将继续无限制地采取样品。因此，我们只考虑q = 1 / k和q> 1 / k时的情况。修复Δ> 0.我们的策略是根据Q描述NK，L。特别地，参数化NK，L使得PNK，L = C日志（4K3 /Δ）

$n_{k,l} = cq^{-1}\log(4k^3/\delta)$ so

$$\begin{aligned}\mathbf{H}(F, \gamma, n_{k,l}, \delta_{k,l}, \epsilon_{k,l}) &\leq c\begin{cases}(\log(1/\delta_{k,l}) + \frac{n_{k,l}}{K})\Delta_2^{-\alpha} + \frac{n_{k,l}}{K}\sum_{i=2}^{K}\Delta_i^{-\alpha} & \text{if } 5\big(F^{-1}(p_{n_{k,l}}) - \nu_*\big) \leq \Delta_2 \\ \Delta^{-\alpha}_{\lceil p_{n_{k,l}}K\rceil}\log(1/\delta_{k,l}) + \frac{n_{k,l}}{K}\sum_{i=\lceil p_{n_{k,l}}K\rceil}^{K}\Delta_i^{-\alpha} & \text{if otherwise}\end{cases} \\ &\leq c\log(k/\delta)\begin{cases}\Delta_2^{-\alpha} + \sum_{i=2}^{K}\Delta_i^{-\alpha} & \text{if } q = 1/K \\ \Delta^{-\alpha}_{\lceil qK\rceil} + \frac{1}{qK}\sum_{i=\lceil qK\rceil}^{K}\Delta_i^{-\alpha} & \text{if } q > 1/K.\end{cases} \\ &\leq c\log(k/\delta)\Delta^{-\alpha}_{\lceil \max\{2, qK\}\rceil} + \frac{1}{qK}\sum_{i=\lceil \max\{2, qK\}\rceil}^{K}\Delta_i^{-\alpha}\end{aligned}$$

使用上行 log（nk，1） ≤clog（log（k / Δ ）q−1）≤clog（log（k / δ ））log（q−1）并让zq = log（q−1）（ Δ − α
i），我们应用了确切的步骤顺序

如在定理5的证明中获得

$$T \leq cz_q\overline{\log}(\log(z_q)/\delta)\overline{\log}(z_q\log(\log(z_q)/\delta))$$

因为输出臂只是经验最好的，所以使用经验估计存在一些错误。圆形（k，1）返回的手臂被拉动 2k−1

因此，可能的误差由γ（bk，1 / log（bk，1））≤c $\frac{k,l}{B_{k,l}}$ c $\frac{}{T}$ $/\alpha$.
这由Δ max{2，qk}为主，用于以上述计算规定的T值，完成证明。

B.7 Proof of Theorem 8

证明让S表示最后阶段的索引，以后确定。如果 rk=rηk−s和 nk=nη−k，那么rk = rk 和nk = nk 那么

$k=0$ $k=0$

从定义，S = min {t∈N：nr（t + 1）η−t≤b，t≤logη（min {r，n}）}。这是

直接验证B≥ZSH是否确保R0≥1和NS≥1。

定理1的证明在这里具有几种修改。首先，我们得出了较低的
如果B≥ZSH具有广义消除速率η的每轮/臂Rk的资源绑定

$$\begin{aligned}
r_k &\geq \frac{B}{|S_k|(\lfloor \log_\eta(n)\rfloor + 1)} - 1 \\
&\geq \frac{\eta}{|S_k|}\max_{i=2,\ldots,n} i\left(1 + \min\left\{R, \gamma^{-1}\left(\max\left\{\frac{\epsilon}{4}, \frac{\nu_i - \nu_1}{2}\right\}\right)\right\}\right) - 1 \\
&\geq \frac{\eta}{|S_k|}\left(\lfloor |S_k|/\eta\rfloor + 1\right)\left(1 + \min\left\{R, \gamma^{-1}\left(\max\left\{\frac{\epsilon}{4}, \frac{\nu_{\lfloor |S_k|/2\rfloor+1} - \nu_1}{2}\right\}\right)\right\}\right) - 1 \\
&\geq \left(1 + \min\left\{R, \gamma^{-1}\left(\max\left\{\frac{\epsilon}{4}, \frac{\nu_{\lfloor |S_k|/2\rfloor+1} - \nu_1}{2}\right\}\right)\right\}\right) - 1 \\
&= \min\left(R, \gamma^{-1}\left(\max\left(\frac{\epsilon}{4}, \frac{\nu_{\lfloor |S_k|/2\rfloor+1} - \nu_1}{2}\right)\right)\right) .
\end{aligned}$$

另外，请注意γ（r）= 0，因此如果最小值是活动的，则 i，r
= νi，我们知道真正的损失。其余的证据与η的定理1代替2的定理1。

另外，我们注意到了

$$\max_{i=n_s+1,\ldots,n} i\,\gamma^{-1}\left(\max\left\{\tfrac{\epsilon}{4}, \tfrac{\nu_i-\nu_1}{2}\right\}\right) \leq n_s\gamma^{-1}\left(\max\left\{\tfrac{\epsilon}{4}, \tfrac{\nu_{n_s+1}-\nu_1}{2}\right\}\right) + \sum_{i>n_s}\gamma^{-1}\left(\max\left\{\tfrac{\epsilon}{4}, \tfrac{\nu_i-\nu_1}{2}\right\}\right).$$

■

References

A. Agarwal，J. Duchi，P. L. Bartlett和C. Levrard。Oracle不等式计算
预算模型选择。在学习理论（Colt）的会议上，2011年。

A. Agarwal，S. Kakade，N.Karampatziakis，L.歌曲和G. Valiant。重新访问的最小二乘：
多级预测的可扩展方法。在国际机会学习会议中（ICML），第541–549页，2014年。

J. Bergstra和Y. Bengio。随机搜索超参数优化。刊中
Machine Learning Research, 13:281–305, 2012.

J.Bergstra，R. Bardenet，Y.Bengio，以及B. Kegl。超参数OPTI的算法 –
mization。在神经信息处理系统（NIPS）中，2011。

S. Bubeck，R. Munos和G. Stoltz。多武装匪徒问题的纯粹探索。在
2009年算法学习理论国际会议（ALT）。

S. Bubeck，R. Munos，G.Stoltz和C.Szepesvari。X武装匪徒。机器杂志
Learning Research, 12:1655–1695, 2011.

A. Carpentier和M. Valko。对于无限许多武装匪徒来说，简单的遗憾。在国际上
机器学习会议（ICML），2015。

E.常数，V. perchet和N.Vayatis。高斯流程优化与相互信息 –
灰。在国际机会学习会议中（ICML），2014年。

T. Domhan，J.T.Pringenberg和F. Hutter。加快自动覆盖物
学习曲线外推深神经网络的优化。在国际人工智能联席会议中（IJCAI），2015年。

K.Egenensperger等。迈向评估贝叶斯优化的实证基础 –
近似参数。在神经信息处理系统（NIPS）贝叶斯优化研讨会上，2013年。

E.偶尔，S. Mannor和Y. Mansour。行动消除和停止条件
对于多武装匪徒和加强学习问题。机床学习研究，7：1079–1105,2006。

M. Feurer. Personal communication, 2015.

M.捕捞者，J.Pringenberg和F. Hutter。使用元学习初始化贝叶斯
优化超参数。在Meta–Learning and算法选择的ECAI研讨会中，2014年。

M.Furemer，A.Klein，K.Eggensperger，J.Pringenberg，M. Blum和F. Hutt。高效的
坚固自动化机器学习。在神经信息处理系统（NIPS）中，2015。

D. Golovin, B. Sonik, S. Moitra, G. Kochanski, J. Karro, and D.Sculley. Google Vizier: A
用于黑匣子优化的服务。在知识发现和数据挖掘（KDD），2017。

S.Gr ¨ unew ¨ alder，J. audibert，M. Opper和J. Shawe−Taylor。高斯遗憾的界限
过程强盗问题。在人工智能国际会议中
Statistics (AISTATS), 2010.

A.Gy ¨ orgy和L. Kocsis。有效的本地搜索算法的多启动策略。杂志
人工智能研究，41：407−444，2011。

F. Hutter，H.Hoos和K.Leyton−Brown。基于序列模型的基因 −
ERAL算法配置。在国际学习与智能优化会议中（狮子），2011年。

K. jamieson和R. Nowak。多武装匪徒的最佳臂识别算法
固定的置信度设置。在信息科学与系统（CISS）会议上，第1−6页。IEEE，2014。

K. jamieson和A. Talwalkar。非转机最佳臂识别和熟王
ETER优化。在国际人工智能与统计（AISTATS）国际会议中，2015年。

K. jamieson，M. Malloy，R. Nowak和S. Bubeck。Lil'ucb：最佳探索
多武装匪算法。在学习理论（COLT）的会议上，第423−439,2014页。

K. G. Jamieson，D. Haas和B. RECHT。识别统计时适应性的力量
备择方案。在神经信息处理系统（NIPS）中，页面775−783,2016。

K. Kandasamy，J. Schneider和B. P'oczos。高维贝叶斯优化和
通过添加剂模型的匪徒。在国际机会学习会议（ICML），2015年。

K.Kandasamy，G. dasarathy，J.B.B.L.Oliva，J.G.Schneider和B. P'oczos。高斯过程
带有多保真评估的强盗优化。在神经信息处理系统（NIPS），2016。

K. Kandasamy，G. dasarathy，J.Schneider和B. P'oczos。多保真贝叶斯optimiza−
连续近似。在国际机会学习会议中（ICML），2017年。

Z. Karnin，T.Koren和O. Somekh。多武装匪徒几乎最佳探索。
在国际机会学习会议中（ICML），2013年。

E.Kaufmann，O. Capp'e和A. Garivier。论最佳臂识别的复杂性
多武装匪盗模型。机器学习研究杂志，16：1−42，2015。

A. Klein，S. Falkner，S. Bartels，P. Hennig和F. Hutter。快速贝叶斯优化
在大型数据集中学习超参数的机器。在国际人工智能和统计（AISTATS）的国际会议中，2017年。

A. Klein，S. Falkner，J.T.Pringenberg和F. Hutter。学习曲线预测
贝叶斯神经网络。在国际学习代表性会议中
(ICLR), 2017b.

A. Krizhevsky。从小图像学习多层特征。在技术报告中，
2009年多伦多大学计算机科学系。

T.Krueger，D.Panknin和M. Braun。通过顺序测试快速交叉验证。杂志
机器学习研究，16：1103−1155,2015。

H. larochelle等。对许多人的深层架构的实证评价
变异因素。在国际机会学习会议中（ICML），2007年。

L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: Bandit-
基于覆盖计优化的基于配置评估。在国际赋予学习代表（ICLR），2017年。

O. Maron和A. Moore。赛车算法：懒惰学习者的模型选择。人造的
Intelligence Review, 11:193–225, 1997.

V. MNIH和J.−Y.audibert。经验伯恩斯坦停止。在国际会议上
Machine Learning (ICML), 2008.

Y. Netzer等。在具有无监督的特征学习中读取自然图像中的数字。在神经中
2011年深度学习和无监督特征学习的信息处理系统（NIPS）研讨会。

A. Rahimi和B. RECHT。大型内核机器的随机功能。在神经中
Information Processing Systems (NIPS), 2007.

R. Rifkin和A. Klautau。捍卫一vs−all分类。机器学习杂志
Research, 5:101–141, 2004.

P. Sermanet，S. chintala和Y. Lecun。卷积神经网络适用于房屋
数字数字分类。在国际模式识别（ICPR）的国际会议中，2012年。

J.Snoek，H. Larochelle和R. Adams。实用的贝叶斯机器学习优化
算法。在神经信息处理系统（NIPS）中，2012。

J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. Patwary, Prabhat, and
R. Adamst。使用深神经网络可扩展的贝叶斯优化。在国际机会学习会议中（ICML），2015A。

J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. Patwary, M. Prabhat,
和亚当斯。贝叶斯优化使用深神经网络。在国际机会学习会议中（ICML），2015B。

E. Sparks，A.Talwalkar，D. Haas，M. J. Franklin，M. I. Jordan和T. Kraska。自动化
模型搜索大型机器学习，在云计算（SoCC）的ACM研讨会中，2015年。

J. Springenberg，A.Klein，S. Falkner和F. Hutter。贝叶斯优化雄厚
贝叶斯神经网络。在神经信息处理系统（NIPS），2016。

N.Srinivas，A.Krause，M. Seeger和S. M. Kakade。高斯过程优化
强盗设置：无后悔和实验设计。在2010年的机器学习（ICML）国际会议中。

K. Swersky，J.Snoek和R. Adams。多任务贝叶斯优化。在神经信息 –
tion Processing Systems (NIPS), 2013.

K. Swersky，J.Snoek和R. P. Adams。冻融贝叶斯优化。arxiv预印刷品
arXiv:1406.3896, 2014.

C. Thornton等自动WEKA：组合选择和封锁优化
分类算法。知识发现和数据挖掘（KDD），2013。

A.Van der Vaart和H. Van Zanten。非参数高斯过程的信息率
方法。机器学习研究杂志，12：2095−2119，2011。

Z.Wang，M. Zoghi，F. Hutter，D. Matheson和N. de Freitas。贝叶斯优化in.
通过随机嵌入的高尺寸。在国际人工智能联席会议中（IJCAI），2013年。

Z. Wang，B. Zhou和S. Jegelka。优化作为高斯过程的估计
在匪徒设置中。在2016年人工智能与统计（AISTATS）国际会议中。

A. G.Wilson，C. Dann和H. Nickisch。关于大型可扩展高斯过程的思考。
arXiv:1511.01870, 2015.