arXiv:2302.07049v1 [math.OC] 14 Feb 2023

多层无目标函数优化及其在神经网络训练中的应用

Serge Gratton*、Alena Kopani˘c´akov´a†、Philippe L. Toint‡

2023年2月
3日

摘要

本文提出一类用于无约束非线性优化的多层算法，该类算法无需计算目标函数。其中，无动量 AdaGrad 方法是一个特例（单层实例）。不计算目标函数的设计旨在降低该类算法对噪声的敏感性，而多层特性则旨在减少其计算成本。本文分析了这些算法的评估复杂度，并在监督学习应用中的深度神经网络训练背景下，展示了它们在存在噪声时的表现。

关键词：非线性优化、多层方法、无目标函数优化（OFFO）、复杂度、神经网络、深度学习。

1 引言

在许多情况下，涉及大量变量的优化问题确实呈现出某种结构，例如导数的稀疏性 [3, 5, 17, 37, 48, 49], 特定的不变性[10, 19, 21, 28, 40] 或隐含的谱性质 [4, 20, 26, 27, 36, 39, 41], ，这里仅举出三个值得关注的例子。如果问题源于某个底层无限维情形的离散化，那么人们早已知晓：使用不同网格尺寸对同一问题进行不同的离散化，并将这些离散化谨慎地用于所谓的多层或多重网格算法，可以带来显著的计算效益。在这种情况下，出色的数值性能通常是通过利用这些离散化之间的自然层次结构，逐步消除误差（或残差）的各种频率分量 [6]同时利用这样一个事实：与细离散相比，函数和导数在粗离散上的计算通常成本更低。多重网格方法如今已成为数值分析中研究十分充分的领域，并被视为求解由椭圆型偏微分方程的求解所产生的线性系统和非线性系统的关键工具。类似思想也已应用于非线性优化，其中 MG-Opt [39, 41] 和 RMTR [26] 多层框架均被设计用于利用相同的性质，并且往往取得了非常成功的效果（例如参见 [38, 25],）。

近年来，人们对大规模问题的另一种优化方法进行了广泛探索，推动了极其简单的一阶方法的应用（参见众多相关文献，例如 [14, 47, 33, 45, 52,24]）。这些方法每次迭代的计算成本很低，但通常需要进行（有时非常）多次迭代。它们之所以广受欢迎，原因有好几个。首先，

*图卢兹大学，INP，IRIT，图卢兹，法国。部分研究经费由图卢兹人工与自然智能研究所（ANITI）提供，该研究所隶属法国“投资未来计划 - PIA3”，资助协议编号为 ANR-19-PI3A-0004。电子邮箱：serge.gratton@enseeiht.fr。

†布朗大学应用数学系，美国普罗维登斯。电子邮箱：alena kopanicakova@brown.edu。

‡那慕尔大学那慕尔复杂系统中心（naXys），比利时那慕尔。电子邮箱：philippe.toint@unamur.be。

可以证明，这些方法能够收敛，且其全局收敛速率有时可与更复杂的方法相媲美。其次，它们通过避免计算目标函数值，以及通常避免计算梯度以外的其他导数，来实现简洁性（因此得名）。这也使它们在函数及其导数存在噪声时具有很强的鲁棒性 [22]，当问题规模很大，以至于这些量只能通过抽样等方式进行实际估算，而无法精确计算时，这一特点尤为重要。采用计算函数值的方式进行优化的情境有时称为 OFFO（无目标函数优化）。人们已经研究了大量一阶 OFFO 方法，但本文将聚焦于 AdaGrad [14]，它是这一类方法中最著名且已证明收敛的方法之一。

本文旨在证明，将多层算法与 *OFFO* 算法相结合在理论上是严谨可行的，在实践中也是高效的。为实现这些目标，我们分三步阐述本文的贡献。

- 首先，我们介绍一类新颖的多层 OFFO 算法（AdaGrad 可视为其中一种单层实现）（第 2 节）。
- 随后，我们分析此类算法的全局收敛速度，并展示了与当前最先进成果相当的结果（第 3 节和第 4 节）。
- 最后，我们以监督学习应用中深度神经网络（深度神经网络）的训练所产生的含噪优化问题为背景，说明所提方法的用法及其优势（第 5 节）。

最后，我们在第 6 节简要总结。

记号。符号 $\|\cdot\|$ 表示标准欧几里得范数。若 x 是一个向量，则 $|x|$ 是其第 j 个分量为 $|x_j|$的向量。矩阵 M 的奇异值记为 $\sigma_i[M]$。

2 多层 OFFO 算法类别

现在，我们将介绍一个理想化的多层 OFFO 框架，该框架融合了 [23]和 [26]中的思想，并对其进行分析。我们考虑的问题是光滑无约束优化的一种结构化形式，即

$$\min_{x\in\mathbb{R}^n} f(x) \tag{1}$$

对于一个二阶连续可微函数 f，它将 IR^n 映射到 IR。该问题具有特定结构，因为我们假设已知一组函数，它们提供了目标函数 f的一系列“层次化”近似。如上所述，当考虑连续问题离散化后的函数时，通常会出现这种情况（此时层次由不同的离散化网格构成）；或者当目标函数涉及一个图，而该图的描述精细程度可能不同时，也会出现这种情况。更具体地说，我们假设已知一组函数$\{f_\ell\}_{\ell=1}^r$，使得每个 f_ℓ 都是一个二阶连续可微函数，将 IR$^{n_\ell}$ 映射到 IR；它们与原问题的联系在于，对于所有 $x \in$ IR^n，$n_r = n$ 和 $f_r(x) = f(x)$。我们还假设，对于每个 $\ell = 2, \ldots, r$，求值或最小化 f_ℓ 的成本都高于 $f_{\ell-1}$。这可能是因为 f_ℓ 的变量多于 $f_{\ell-1}$（如果 f_ℓ 表示同一个无限维目标函数越来越精细的离散化，通常就会如此），也可能是因为 f_ℓ 的结构（就部分可分性、稀疏性或特征结构而言）比 $f_{\ell-1}$更复杂，或者出于其他任何原因。为统一术语，我们将某个特定的 ℓ 称为一个层。然而，若要让 $f_{\ell-1}$ 在最小化 f_ℓ时发挥作用，这两个函数的变量之间就应当存在某种联系。因此，我们假设，对于每个 $\ell = 2, \ldots, r$，都存在一个满秩线性算子 R_ℓ，它将 IR$^{n_\ell}$ 映射到 IR$^{n_{\ell-1}}$（限制），以及另一个满秩算子P_ℓ，它将 IR$^{n_{\ell-1}}$ 映射到 IR$^{n_\ell}$（延拓），使得

$$\omega P_\ell = R_\ell^T \tag{2}$$

其中某个已知常数为* $\omega > 0$。在多重网格算法的语境中， P_ℓ 和 R_ℓ 被解释为细网格与粗网格之间的限制和延拓（例如参见 [6, 41, 26, 25], ）。

在进一步讨论细节之前，我们先约定一条重要的索引规则。由于有时需要同时标识一个层、算法的一次迭代以及一个向量的分量，我们用索引 ℓ 表示层级，用 i 表示迭代，用 j 表示分量。例如，$x_{\ell,\,i,\,j}$ 表示层 ℓ中第 j次迭代时向量 x 的第 i 个分量。每次进入一个层时，迭代索引 i 都会重置为零†。

由于我们旨在将 [23] 的 ASTR1 无目标函数框架扩展到多层情境，因此现在回顾该算法的主要概念，并建立一些记号。正如名称中的 TR 所示，ASTR1 是一种信赖域优化算法。众所周知，这类算法在理论上是可靠的（深入介绍见 [11] ，较新的综述见[51] ），在实践中也非常高效。与所有信赖域方法一样，在层 ℓ 上，下一迭代点通过在模型可信的区域内最小化一个依赖于层的目标函数模型 h_ℓ来确定。在我们的多层框架中，该模型可以是（可能为二次的）泰勒型

$$m_{\ell,i}(s) = g_{\ell,i}^T s + \tfrac{1}{2} s^T B_{\ell,i} s, \tag{3}$$

其中 $g_{\ell,\,i} \overset{\text{定义}}{=} \nabla_x^1 h_\ell(x_{\ell,\,i})$，且$B_{\ell,\,i}$ 是有界的海森矩阵近似，或由下式定义的低层模型

$$h_{\ell-1}(x_{\ell-1,0} + s_{\ell-1}) \overset{\text{def}}{=} f_{\ell-1}(x_{\ell-1,0} + s_{\ell-1}) + v_{\ell-1}^T s_{\ell-1}, \tag{4}$$

其中

$$v_{\ell-1} = R_\ell g_{\ell,i} - \nabla f_{\ell-1}(x_{\ell-1,0}). \tag{5}$$

按照惯例，令 $v_r = 0$，因此，对于所有 s_r，

$$h_r(x_{r,0} + s_r) = f_r(x_{r,0} + s_r) = f(x_{r,0} + s_r) \text{ and } g_{r,k} = \nabla_x^1 h_r(x_{r,k}) = \nabla_x^1 f(x_{r,k}). \tag{6}$$

因此，模型 h_ℓ 对应于通过一个线性项对 f_ℓ 进行修正，以满足“线性相容性”关系

$$g_{\ell-1,0} = \nabla_x^1 h_{\ell-1}(x_{\ell-1,0}) = R_\ell g_{\ell,i}. \tag{7}$$

这种一阶修正 (4) 常用于全近似格式 [6], 中的多重网格应用，也用于其他场景 [16, 41, 1, 38, 26, 34]。我们称其为“线性相容性”，因为它关键地确保了 h_ℓ 和 $h_{\ell-1}$ 的一阶行为在$x_{\ell,\,i}$ 和$x_{\ell-1,\,0}$的邻域内分别相容。为说明这一点，只需验证：若 s_ℓ 和 $s_{\ell-1}$满足 $s_\ell = P_\ell s_{\ell-1}$，则利用 (2) 和 (7)，

$$g_{\ell,i}^T s_\ell = g_{\ell,i}^T P_\ell s_{i-1} = \frac{1}{\omega} R_\ell g_{\ell,i}^T s_{\ell-1} = \frac{1}{\omega} g_{\ell-1,0}^T s_{\ell-1}. \tag{8}$$

定义/选择模型后，信赖域方法的典型迭代会在以当前迭代点为中心的球内将其最小化，该球的半径由算法根据过去的性能自适应计算得到。这个球可以用不同的范数定义，但本文将重点讨论“无穷范数”的一种缩放形式，其中分别测量向量各分量的绝对值。在 ASTR1 语境下，信赖域半径使用当前梯度的大小和一个逐分量严格为正的 权重向量计算得到。我们暂不对其作完整定义，后续分析中将对其加以说明。由于我们的多层算法是递归的，因此还必须在某一层强制终止：当整体步长延拓到前一层后的欧几里得范数过大时，也就是说，当不等式

$$\|P_{\ell+1}(x_{\ell,i} - x_{\ell,0})\| \le \delta_\ell \tag{9}$$

*为简便起见，我们选择令 ω 独立于 ℓ，这总是可以通过缩放实现。

†我们很清楚，这会造成一些歧义，因为如果层 ℓ ($\ell < r$) 被使用多次，索引序列 ℓ, i 可能会出现多次，这意味着该层存在多个初始迭代点。这种歧义可根据上下文消除。

不成立时，其中， $\delta_\ell \geq 0$ 是层级为 $\ell+1$ 时步长范数的界；当 $\ell < r$ 时取此界，否则取 $+\infty$。

为了指定算法，我们最后定义权重向量 w_ℓ （其大小为 n_ℓ）对应的对角矩阵

$$D(\{w_\ell\}) \stackrel{\rm def}{=} \mathrm{diag}\left(\frac{1}{w_{\ell,1}}, \ldots, \frac{1}{w_{\ell,n_\ell}}\right). \tag{10}$$

我们假设，对于 $j \in \{1, \ldots, n_\ell\}$，存在一个常数 $\varsigma_j \in (0, 1],$，使得 $w_{\ell,\ j} \geq \varsigma_j$ 对每个 $\ell \in \{1, \ldots, r\}$ 均成立。多层目标函数无关信赖域（MOFFTR）算法将在下一页给出。

通过调用以下内容即可求解原问题 (1)

$$\mathsf{MOFFTR}(\, r,\, f,\, x_{r,0},\, \epsilon_r,\, i_r^{(\max)},\, +\infty,\, \varsigma\,), \tag{11}$$

其中， $i_r^{(\max)}$ 是顶层迭代的最大次数， ς 表示权重下界向量。为统一术语，我们将算法的步骤 4计算步长的迭代称为泰勒迭代，而将步长由递归调用 (17) 得出的迭代称为递归迭代。

在此阶段，一些说明有助于进一步解释并阐明该算法的细节。

1. 请注意，若达到依赖于层的精度阈值 ϵ_ℓ ，或达到依赖于层的最大迭代次数 $i_\ell^{(\max)}$ ，则任意层上的迭代都会在步骤 1 中终止。另请注意，(12) 保证 (9) 成立。

2. 分量式信赖域半径定义见 (13)。请注意，这种选择会阻止对应梯度分量为零时步长中出现非零分量。还请注意，(13) 避免了可能导致 (12) 对 $i+1$ 不成立的大步长。事实上，若 $\ell < r$，则由 (13) 和 (19) 可得

$$\|P_{\ell+1}(x_{\ell,i+1} - x_{\ell,i})\| = \|P_{\ell+1}s_{\ell,i}\| \leq 2\delta_\ell.$$

因此，任何会违反该不等式的 $x_{\ell,\ i+1}$ 都不满足 (12)（对于 $i+1$），因为此时

$$\|P_{\ell+1}(x_{\ell,i+1} - x_{\ell,0})\| \geq \|P_{\ell+1}(x_{\ell,i+1} - x_{\ell,i})\| - \|P_{\ell+1}(x_{\ell,i} - x_{\ell,0})\| > 2\delta_\ell - \delta_\ell = \delta_\ell,$$

其中，我们利用 (12)（对于 i）推导出了最后一个不等式。

模型的选择（步骤 2 末尾）并未正式规定，而是留给用户决定。在多重网格文献中称为“V 循环”的一种典型模式中，给定层级 ℓ 上要执行的任务如下。首先执行一组标准的泰勒迭代。然后，如果$\ell > 1$ （也就是说，当前层不是最低层），且在层 $\ell-1$ 上有望取得显著进展（即条件 (16) 不成立），则递归地在层 $\ell-1$调用该算法。随后在层 ℓ再执行第二组泰勒迭代。该名称所暗示的“V”形，源于在所有层级递归应用这一模式。虽然通常会指定两组泰勒迭代的次数（用多重网格术语来说，分别称为“预平滑”和“后平滑”），但 MOFFTR 并不要求这样做。事实上，该算法允许采用各种固定或自适应的迭代模式。

3. 接下来，我们回顾步骤 3 的机制。首先注意，$\delta_{\ell,\ i}$ 是迭代（$\ell,\ \ i$）时允许采用的步长的欧几里得范数，前提是这次迭代是一次泰勒迭代（见（13））。随后，我们选取一组权重 $w_{\ell-1,\ 0}$ 供较低层级使用。除了必须分别以下界 ς_j 为限外，这些权重还必须满足另外两个条件。第一个条件（14）目前以一种非常通用的方式表述，并确保这些权重

算法 2.1： $\boxed{x_+ = \mathsf{MOFFTR}(\ell, h_\ell, x_{\ell,0}, \epsilon_\ell, i_\ell^{(\max)}, \delta_\ell, w_{\ell,0})}$

步骤 0：初始化。给定常数 $\kappa_{\rm R} \in (0,\ 1)$、$\alpha \geq 1$、$\tau \in (0,\ 1]$、$\kappa_{\rm B} \geq 1$ 和 $\varsigma_j \in (0,1]$($j \in \{1,\ldots,n_\ell\}$)。令 $i = 0$。

步骤 1：终止测试。如果 $\ell < r$ 且

$$\|P_{\ell+1}(x_{\ell,i} - x_{\ell,0})\| > \delta_\ell \tag{12}$$

则返回 $x_+ = x_{\ell,\ i-1}$。否则，计算 $g_{\ell,\ i} \overset{\text{定义}}{=} \nabla_x^1 h_\ell(x_{\ell,\ i})$。如果$\|g_{\ell,\ i}\| \leq \epsilon_\ell$ 或 $i = i_\ell^{(\max)}$，则返回 $x_+ = x_{\ell,\ i}$。

步骤 2：定义信赖域。设置

$$\widehat{\Delta}_{\ell,i} = D(\{w_{\ell,i}\})|g_{\ell,i}| \ \text{ and } \ \Delta_{\ell,i} = \begin{cases} \widehat{\Delta}_{\ell,i} & \text{if} \quad \ell = r, \\ \min\left[\dfrac{2\delta_\ell}{\|P_{\ell+1}\|\,\|\widehat{\Delta}_{\ell,i}\|}, 1\right] \widehat{\Delta}_{\ell,i} & \text{if} \quad \ell < r. \end{cases} \tag{13}$$

如果 $i > 0$，定义 $w_{\ell,\ i} \in \mathbb{R}^{n_\ell}$，使得 $w_{\ell,\ i,\ j} \geq \varsigma_j$，对于 $j \in \{1,\ldots,n_\ell\}$。如果在迭代 i时需要泰勒步，则转到步骤4。

步骤 3：递归步。选择 $w_{\ell-1,\ 0} \in \mathbb{R}^{n_{\ell-1}}$，使得 $w_{\ell-1,\ 0,\ j} \geq \varsigma_j$ 对于 $j \in \{1,\ldots,n_{\ell-1}\}$,

$$\boxed{\text{“the lower-level weights are large enough”}} \tag{14}$$

and

$$\|D(\{w_{\ell-1,0}\})\,|R_\ell g_{\ell,i}|\| \leq \frac{\alpha\|\Delta_{\ell,i}\|}{\|P_\ell\|}. \tag{15}$$

如果 $\ell = 1$ 或

$$\sum_{j=1}^{n_\ell - 1} \frac{[R_\ell g_{\ell,i}]_j^2}{w_{\ell-1,0,j}} < \kappa_{\rm R} \sum_{j=1}^{n_\ell} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}} \tag{16}$$

则转到步骤4。否则（即，如果 $\ell > 1$ 且 (16) 不成立），计算

$$s_{\ell,i} = P_\ell\Big[\mathsf{MOFFTR}(\ell-1, h_{\ell-1}, R_\ell x_{\ell,i}, \epsilon_{\ell-1}, i_{\ell-1}^{(\max)}, \alpha\|\Delta_{\ell,i}\|, w_{\ell-1,0}) - R_\ell x_{\ell,i}\Big], \tag{17}$$

其中 $h_{\ell-1}$ 由 (4) 给出。

步骤4：泰勒步。选择一个对称的海森矩阵近似 $B_{\ell,\ i}$，使得

$$\|B_{\ell,i}\| \leq \kappa_{\rm B}. \tag{18}$$

计算步长 $s_{\ell,\ i}$，使得

$$|s_{\ell,i,j}| \leq \Delta_{\ell,i,j} \quad (j \in \{1,\ldots,n_\ell\}), \tag{19}$$

and

$$g_{\ell,i}^T s_{\ell,i} + \tfrac{1}{2} s_{\ell,i}^T B_{\ell,i} s_{\ell,i} \leq \tau\left(g_{\ell,i}^T s_{\ell,i}^Q + \tfrac{1}{2}(s_{\ell,i}^Q)^T B_{\ell,i} s_{\ell,i}^Q\right), \tag{20}$$

其中

$$s_{\ell,i,j}^L = -\text{sign}(g_{\ell,i,j})\Delta_{\ell,i,j} \quad (j \in \{1,\ldots,n_\ell\}), \tag{21}$$

$$s_{\ell,i}^Q = \gamma_{\ell,i} s_{\ell,i}^L, \quad \text{with} \quad \gamma_{\ell,i} = \begin{cases} \min\left[1, \dfrac{|g_{\ell,i}^T s_{\ell,i}^L|}{(s_{\ell,i}^L)^T B_{\ell,i} s_{\ell,i}^L}\right] & \text{if} \quad (s_{\ell,i}^L)^T B_{\ell,i} s_{\ell,i}^L > 0, \\ 1 & \text{otherwise.} \end{cases} \tag{22}$$

步骤5：更新。设置

$$x_{\ell,i+1} = x_{\ell,i} + s_{\ell,i}, \tag{23}$$

将 i 加一，然后返回步骤 1。

当第 ℓ 层的权重较大时，它不可能很小。如何实现这一点取决于权重的具体选择，正如下文将要看到的。第二个是看似晦涩的条件 (15)，它只是确保第 $\ell-1$ 层总步长的欧几里得范数全局界足够大，从而允许在较低层级至少进行一次迭代。该条件指出，较低层级的第一步 P_ℓ 的延拓，其欧几里得长度至多为第 ℓ层（假设）步长欧几里得长度的某个倍数 $\alpha \geq 1$ 。接着，条件 (16) 比较 $h_{\ell-1}$ 在 $R_\ell x_{\ell,\,i}$ 处的线性近似与 h_ℓ 在 $x_{\ell,\,i}$处的线性近似所产生的下降。如果前者小于后者的某个比例 κ_{R} ，这表明“较低层级不太可能取得显著进展”，于是我们会继续在当前层级进行最小化。如果有可能取得显著进展，我们便选择在较低层级对 $h_{\ell-1}$ 进行最小化（递归迭代），使用权重 $w_{\ell-1,\,0}$，从 $R_\ell x_{\ell,\,i}$ 开始，并在半径为 $\alpha\|\Delta_{\ell,\,i}\|$的欧几里得球内进行。

我们注意到，(15) 很容易满足。事实上，可以直接验证，只要

$$w_{\ell-1,0,j} \geq \max\left[\varsigma_j, \frac{\|P_\ell\|\,|R_\ell g_{\ell,i}|_j}{\alpha\|\Delta_{\ell,i}\|}\right]. \tag{24}$$

4. 泰勒迭代中的步长在步骤4中使用一种借鉴自 ASTR1 算法 [23]的技术进行计算。熟悉信赖域理论的读者会认出，其中$s_\ell^{Q}{}_{,i}$ 是通过在负梯度张成空间与信赖域的交集上最小化二次模型而得到的“柯西点”的一种变体（参见 [11, 第6.3节2]），而 $s_{\ell,i}^{L}$ 则是在信赖域内对更简单的线性模型 $g_\ell^T{}_{,i}s$ 进行最小化。

5. 目标函数 f_r 及其近似值 $\{f_\ell\}_{\ell=1}^{r-1}$ 从未被计算，因此将它们结合起来的优化方法确实是“无目标函数”的。

3 收敛性分析

我们的收敛性分析基于以下标准假设。

AS.1: 对于每个 $\ell \in \{1,\ldots,r\}$，函数 f_ℓ 连续可微。

AS.2: 对于每个 $\ell \in \{1,\ldots,r\}$，梯度 $\nabla_x^1 f_\ell(x)$ 是利普希茨连续的，利普希茨常数为 $L \geq 0$，即

$$\|\nabla_x^1 f_\ell(x) - \nabla_x^1 f_\ell(y)\| \leq L\|x-y\|$$

对于所有 $x,\ y \in \mathrm{IR}^n$ 以及所有 $\ell \in \{1,\ldots,r\}$。

AS.3: 存在一个常数 f_{low}，使得对于所有 x，$f(x) \geq f_{\mathrm{low}}$。

下文中，我们使用以下记号

$$\Gamma_0 \stackrel{\mathrm{def}}{=} f(x_0) - f_{\mathrm{low}} \tag{25}$$

表示起始点处的目标函数与其下界之间的差距。请注意，我们并未假设 f_ℓ 的梯度保持有界。

在考虑权重的更具体选择之前，我们先推导泰勒步的一项基本性质，并加强 [23, 引理 2.1]，对泰勒迭代的“线性下降”（即目标函数简单线性模型中的下降量）进行量化。

引理 3.1 假设 AS.1 和 AS.2 成立。考虑层级为 ℓ 的一次泰勒迭代 i 。那么

$$g_{\ell,i}^T s_{\ell,i} \leq -\frac{\tau \varsigma_{\min}}{2\kappa_{\rm B}} \sum_{j=1}^{n_\ell} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}} + \frac{\kappa_{\rm B}}{2} \|\Delta_{\ell,i}\|^2 \tag{26}$$

其中 $\varsigma_{\min} \stackrel{\text{定义}}{=} \min_{j\in\{1,\ldots,n_\ell\}} \varsigma_j \in (0,1]$。

证明。 首先注意到，由于（21）以及 $w_{\ell,i,j}$ 的定义，

$$|g_{\ell,i}^T s_{\ell,i}^L| = \sum_{j=1}^{n_\ell} \frac{w_{\ell,i,j} g_{\ell,i,j}^2}{w_{\ell,i,j}^2} \geq \sum_{j=1}^{n_\ell} \frac{\varsigma_j g_{\ell,i,j}^2}{w_{\ell,i,j}^2} \geq \varsigma_{\min} \|s_{\ell,i}^L\|^2. \tag{27}$$

现在假设 $(s_{\ell,i}^L)^T B_{\ell,i} s_{\ell,i}^L > 0$ 且 $\gamma_{\ell,i} < 1$。那么，根据（20）、（22）、（27）和（18），

$$g_{\ell,i}^T s_{\ell,i}^Q + \tfrac{1}{2}(s_{\ell,i}^Q)^T B_{\ell,i} s_{\ell,i}^Q = \gamma_{\ell,i} g_{\ell,i}^T s_{\ell,i}^L + \tfrac{1}{2}\gamma_{\ell,i}^2 (s_{\ell,i}^L)^T B_{\ell,i} s_{\ell,i}^L = -\frac{(g_{\ell,i}^T s_{\ell,i}^L)^2}{2(s_{\ell,i}^L)^T B_{\ell,i} s_{\ell,i}^L} \leq -\frac{\varsigma_{\min} |g_{\ell,i}^T s_{\ell,i}^L|}{2\kappa_{\rm B}}.$$

将此不等式与（22）结合可得

$$g_{\ell,i}^T s_{\ell,i}^Q + \tfrac{1}{2}(s_{\ell,i}^Q)^T B_{\ell,i} s_{\ell,i}^Q \leq -\frac{\varsigma_{\min}}{2\kappa_{\rm B}} \sum_{j=1}^{n_\ell} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}}. \tag{28}$$

或者，假设 $(s_{\ell,i}^L)^T B_{\ell,i} s_{\ell,i}^L \leq 0$ 或 $\gamma_{\ell,i} = 1$。然后，利用（22）、（21）以及界$\kappa_{\rm B} \geq 1$ 和 $\varsigma_{\min} \leq 1$,

$$g_{\ell,i}^T s_{\ell,i}^Q + \tfrac{1}{2}(s_{\ell,i}^Q)^T B_{\ell,i} s_{\ell,i}^Q = g_{\ell,i}^T s_{\ell,i}^L + \tfrac{1}{2}(s_{\ell,i}^L)^T B_{\ell,i} s_{\ell,i}^L \leq \tfrac{1}{2} g_{\ell,i}^T s_{\ell,i}^L \leq -\frac{\varsigma_{\min}}{2\kappa_{\rm B}} \sum_{j=1}^{n_\ell} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}}. \tag{29}$$

因此，由 (28)、(29) 和 (20) 可得

$$g_{\ell,i,}^T s_{\ell,i} + \tfrac{1}{2} s_{\ell,i}^T B_{\ell,i} s_{\ell,i} \leq -\frac{\tau \varsigma_{\min}}{2\kappa_{\rm B}} \sum_{j=1}^{n_\ell} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}}.$$

由此可见，由 (28)、(29)、(20) 和 (18) 可得

$$g_{\ell,i}^T s_{\ell,i} \leq -\tau \varsigma_{\min} \sum_{j=1}^{n_\ell} \frac{g_{\ell,i,j}^2}{2\kappa_{\rm B} w_{\ell,i,j}} + \tfrac{1}{2} |s_{\ell,i}^T B_{\ell,i} s_{\ell,i}| \leq -\tau \varsigma_{\min} \sum_{j=1}^{n_\ell} \frac{g_{\ell,i,j}^2}{2\kappa_{\rm B} w_{\ell,i,j}} + \frac{\kappa_{\rm B}}{2} \sum_{j=1}^{n_\ell} s_{\ell,i,j}^2,$$

而 (26) 则由 (19) 得出。 □

我们还证明以下一个简单引理。

引理 3.2 考虑 MOFFTR 算法执行过程中的迭代 (ℓ, i)。则

$$\|s_{\ell,i}\| \leq \alpha \|D(\{w_{\ell,i}\}) |g_{\ell,i}|\|. \tag{30}$$

证明。 如果迭代 (ℓ, i) 是泰勒迭代，则 (30) 可由 (13)、(19) 和界$\alpha \geq 1$得出。如果它是递归迭代，则由 (17) 和 (19) 可得

$$\|s_{\ell,i}\| = \|P_\ell(x_{\ell-1,i} - x_{\ell-1,0})\| \leq \delta_{\ell-1} = \alpha \|D(\{w_{\ell,i}\}) |g_{\ell,i}|\|,$$

得到式（30）。 □

现在考虑递归迭代时可能发生的情况。

引理 3.3 考虑一个递归迭代（ℓ, i）。则

$$\|\Delta_{\ell-1,0}\| \leq \frac{\alpha}{\|P_\ell\|}\|\Delta_{\ell,i}\|, \tag{31}$$

迭代点 $x_{\ell-1,1}$ 在算法的步骤 1 中被接受，并且在层 $\ell-1$ 至少完成一次迭代。

证明。 由 (13)、(15) 以及MOFFTR的调用顺序可知,

$$\|\Delta_{\ell-1,0}\| \leq \|\widehat{\Delta}_{\ell-1,0}\| = \|D(\{w_{\ell-1,0}\})|g_{\ell-1,0}|\| = \|D(\{w_{\ell-1,0}\})|R_\ell g_{\ell,i}|\| \leq \frac{\alpha}{\|P_\ell\|}\|\Delta_{\ell,i}\|$$

因此，利用 (19)，可得

$$\|P_\ell(x_{\ell-1,1} - x_{\ell-1,0})\| \leq \|P_\ell\|\,\|x_{\ell-1,1} - x_{\ell-1,0}\| \leq \|P_\ell\|\,\|\Delta_{\ell-1,0}\| \leq \alpha\|\Delta_{\ell,i}\| = \delta_{\ell-1}.$$

因此，(12) 在迭代（$\ell-1$,1）处不成立，于是迭代点 $x_{\ell-1,1}$ 被接受，所需结论随即成立。□

引理 3.4 考虑一个递归迭代 (ℓ, i)，并假设算法在第 $\ell-1$ 层已完成 $i_{\ell-1} \geq 1$ 次迭代。那么

$$\left| g_{\ell,i}^T s_{\ell,i} - \frac{1}{\omega}\sum_{k=0}^{i_{\ell-1}-1} g_{\ell-1,k}^T s_{\ell-1,k} \right| \leq \frac{2i_{\ell-1}^{(\max)} L\delta_{\ell-1}^2}{\omega\sigma_{\min}[P_\ell]^2}. \tag{32}$$

证明。 利用 (17) 和 (2)（另见 (8)），我们推得

$$g_{\ell,i}^T s_{\ell,i} = g_{\ell,i}^T \sum_{k=0}^{i_{\ell-1}-1} P_\ell s_{\ell-1,k} = \frac{1}{\omega}\sum_{k=0}^{i_{\ell-1}-1} g_{\ell,i}^T R_\ell^T s_{\ell-1,k} = \frac{1}{\omega}\sum_{k=0}^{i_{\ell-1}-1} g_{\ell-1,0}^T s_{\ell-1,k}. \tag{33}$$

Now

$$g_{\ell-1,0}^T s_{\ell-1,k} = g_{\ell-1,k}^T s_{\ell-1,k} + (g_{\ell-1,0} - g_{\ell-1,k})^T s_{\ell-1,k} \stackrel{\text{def}}{=} g_{\ell-1,k}^T s_{\ell-1,k} + \nu_{\ell-1,k}, \tag{34}$$

其中，利用柯西-施瓦茨不等式、(4)、AS.2 和 (9),

$$\begin{aligned} |\nu_{\ell-1,k}| &\leq \|g_{\ell-1,0} - g_{\ell-1,k}\|\,\|s_{\ell-1,k}\| \\ &= \|\nabla_x^1 f_\ell(x_{\ell-1,0}) - \nabla_x^1 f_\ell(x_{\ell-1,k})\|\,\|s_{\ell-1,k}\| \\ &\leq L\|x_{\ell-1,k} - x_{\ell-1,0}\|\,\|s_{\ell-1,k}\| \\ &\leq \frac{L}{\sigma_{\min}[P_\ell]^2}\|P_\ell(x_{\ell-1,k} - x_{\ell-1,0})\|\,\|P_\ell s_{\ell-1,k}\| \\ &\leq \frac{2L\delta_{\ell-1}^2}{\sigma_{\min}[P_\ell]^2}. \end{aligned} \tag{35}$$

结合式 (33)、(34)、(35) 和界 $i_{\ell-1} \leq i_{\ell-1}^{(\max)}$，即可得到式 (32)。 □

这个关键引理使我们能够量化递归迭代中的“线性下降”，并根据引理 3.1，进一步量化 MOFFTR 算法的所有迭代中的“线性下降”。

引理 3.5 假设 AS.1 和 AS.2 成立。那么，对于所有 $\ell \in \{1, \ldots, r\}$ 和所有 $i \geq 0$,

$$g_{\ell,i}^T s_{\ell,i} \leq -\beta_{1,r} \sum_{j=1}^{n} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}} + \beta_{2,r} \sum_{j=1}^{n} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}^2} \tag{36}$$

对于某些与 ℓ 和 i 无关的常数 $\beta_{1,\,r} > 0$ 和 $\beta_{2,\,r} > 0$ 。

证明。 我们首先应用引理 3.4，推得（32）成立。我们还应用引理 3.3，得出（31）成立，并且 $i_{\ell-1} \geq 1$。首先，假设迭代（ℓ，i）是一次递归迭代，并且 ℓ 是调用（17）中的 MOFFTR 所达到的最低层级的索引加一。于是，MOFFTR 算法在层 $\ell - 1$ 上的每次迭代都是一次泰勒迭代，因此引理 3.1 中的不等式（26）适用。由此，利用（32）和（13），我们得到

$$\begin{aligned} g_{\ell,i}^T s_{\ell,i} &= -\frac{\tau \varsigma_{\min}}{2\kappa_{\rm B}\omega} \sum_{k=0}^{i_{\ell-1}-1} \sum_{j=1}^{n_{\ell-1}} \frac{g_{\ell-1,k,j}^2}{w_{\ell-1,k,j}} + \frac{\kappa_{\rm B}}{2\omega} \sum_{k=0}^{i_{\ell-1}-1} \|\Delta_{\ell-1,k}\|^2 + \frac{2i_{\ell-1}^{(\max)} L \delta_{\ell-1}^2}{\omega \sigma_{\min}[P_\ell]^2} \\ &\leq -\frac{\tau \varsigma_{\min}}{2\kappa_{\rm B}\omega} \sum_{k=0}^{i_{\ell-1}-1} \sum_{j=1}^{n_{\ell-1}} \frac{g_{\ell-1,k,j}^2}{w_{\ell-1,k,j}} + \frac{\kappa_{\rm B}}{2\omega} \sum_{k=0}^{i_{\ell-1}-1} \frac{4\delta_{\ell-1}^2}{\sigma_{\min}[P_\ell]^2} + \frac{2i_{\ell-1}^{(\max)} L \delta_{\ell-1}^2}{\omega \sigma_{\min}[P_\ell]^2} \\ &= -\frac{\tau \varsigma_{\min}}{2\kappa_{\rm B}\omega} \sum_{k=0}^{i_{\ell-1}-1} \sum_{j=1}^{n_{\ell-1}} \frac{g_{\ell-1,k,j}^2}{w_{\ell-1,k,j}} + \frac{2i_{\ell-1}^{(\max)} (\kappa_{\rm B} + L)}{\omega \sigma_{\min}[P_\ell]^2} \delta_{\ell-1}^2. \end{aligned} \tag{37}$$

现在考虑到 $i_{\ell-1} \geq 1$，忽略右侧第一项求和中关于 $k \in \{1, \ldots, i_{\ell-1}-1\}$ 的项，并使用调用 (17) 中 δ_ℓ 的定义，以及 (16) 和 (13) 不成立这一事实，可得

$$\begin{aligned} g_{\ell,i}^T s_{\ell,i} &\leq -\frac{\tau \varsigma_{\min}}{2\kappa_{\rm B}\omega} \sum_{j=1}^{n_{\ell-1}} \frac{g_{\ell-1,0,j}^2}{w_{\ell-1,0,j}} + \frac{2i_{\ell-1}^{(\max)} (\kappa_{\rm B} + L)}{\omega \sigma_{\min}[P_\ell]^2} \alpha^2 \|\Delta_{\ell,i}\|^2 \qquad (38) \\ &\leq -\frac{\tau \varsigma_{\min} \kappa_{\rm R}}{2\kappa_{\rm B}\omega} \sum_{j=1}^{n_\ell} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}} + \frac{2\alpha^2 i_{\ell-1}^{(\max)} (\kappa_{\rm B} + L)}{\omega \sigma_{\min}[P_\ell]^2} \|\widehat{\Delta}_{\ell,i}\|^2 \\ &\leq -\frac{\tau \varsigma_{\min} \kappa_{\rm R}}{2\kappa_{\rm B}\omega} \sum_{j=1}^{n_\ell} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}} + \frac{2\alpha^2 i_{\ell-1}^{(\max)} (\kappa_{\rm B} + L)}{\omega \sigma_{\min}[P_\ell]^2} \sum_{j=1}^{n_\ell} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}^2}. \qquad (39) \end{aligned}$$

或者，如果迭代 (ℓ, i) 是泰勒迭代，则由 (26) 和 (13) 可得

$$g_{\ell,i}^T s_{\ell,i} \leq -\frac{\tau \varsigma_{\min} \kappa_{\rm R}}{2\kappa_{\rm B}} \sum_{j=1}^{n_\ell} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}} + \kappa_{\rm B} \sum_{j=1}^{n_\ell} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}^2}. \tag{40}$$

Com合并（39）和（40），可得，对于层 ℓ 的所有迭代（递归迭代和泰勒迭代） ，

$$\begin{aligned} g_{\ell,i}^T s_{\ell,i} \leq & -\frac{\kappa_{\rm R}}{\max[\omega, 1]} \left[\frac{\tau \varsigma_{\min}}{2\kappa_{\rm B}}\right] \sum_{j=1}^{n_\ell} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}} \\ & + \max\left\{[\kappa_{\rm B}], \frac{2\alpha^2 i_{\ell-1}^{(\max)}}{\max[\omega, 1]\, \sigma_{\min}[P_\ell]^2} \Big([\kappa_{\rm B}] + L \Big) \right\} \sum_{j=1}^{n_\ell} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}^2}. \end{aligned}$$

请注意，方括号中的项对应于将 ℓ 替换为 $\ell-1$ 后的界（40）（因为层 $\ell-1$ 仅包含泰勒迭代）。接下来，我们可以递归地定义

$$\beta_{1,1} \stackrel{\rm def}{=} \frac{\tau\varsigma_{\min}}{2\kappa_{\rm B}} \quad \text{and} \quad \beta_{1,\ell+1} \stackrel{\rm def}{=} \frac{\kappa_{\rm R}}{\max[\omega,1]}\,\beta_{1,\ell}, \tag{41}$$

$$\beta_{2,1} \stackrel{\rm def}{=} \kappa_{\rm B} \quad \text{and} \quad \beta_{2,\ell+1} \stackrel{\rm def}{=} \max\left\{\beta_{2,\ell}, \frac{2\alpha^2 i_{\ell-1}^{(\max)}}{\max[\omega,1]\,\sigma_{\min}[P_\ell]^2}\Big(\beta_{2,\ell}+L\Big)\right\} \tag{42}$$

对于 $\ell \in \{1,\ldots,r\}$，并得到所需的结论。 □

现在可以推导出目标函数下降量的一个核心界。

引理 3.6 假设 AS.1 和 AS.2 成立。那么，对于 $\ell \in \{1,\ldots,r\}$,

$$h_\ell(x_{\ell,i}) - h_\ell(x_{\ell,i+1}) \geq \sum_{j=1}^{n_\ell} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}}\left[\beta_{1,r} - \frac{\beta_{2,r}+\frac{1}{2}\alpha^2 L}{w_{\ell,i,j}}\right]. \tag{43}$$

证明。 依次使用 AS.1、AS.2、(30)、(13) 和 (36)，可得

$$\begin{aligned}
h_\ell(x_{\ell,i+1}) &\leq h_\ell(x_{\ell,i}) + g_{\ell,i}^T s_{\ell,i} + \tfrac{1}{2}L\|s_{\ell,i}\|^2 \\
&\leq h_\ell(x_{\ell,i}) + g_{\ell,i}^T s_{\ell,i} + \tfrac{1}{2}L\alpha^2\|D(\{w_{\ell,i}\})|g_{\ell,i}|\|^2 \\
&= h_\ell(x_{\ell,i}) + g_{\ell,i}^T s_{r,i} + \tfrac{1}{2}\alpha^2 L\sum_{j=1}^{n_\ell}\frac{g_{\ell,i,j}^2}{w_{\ell,i,j}^2} \\
&\leq h_\ell(x_{\ell,i}) - \beta_{1,r}\sum_{j=1}^{n_\ell}\frac{g_{\ell,i,j}^2}{w_{\ell,i,j}} + \big(\beta_{2,r}+\tfrac{1}{2}\alpha^2 L\big)\sum_{j=1}^{n_\ell}\frac{g_{\ell,i,j}^2}{w_{\ell,i,j}^2}
\end{aligned} \tag{44}$$

从而得到 (43)。 □

3.1 发散权重

为了使我们的方法具有一致性和实用性，现在必须说明如何选择权重，并进一步明确条件 (14)。我们首先考虑如下定义的“发散权重”。在本节中，我们假设权重 $w_{i,\,k}$ 的选取满足：对于某个幂参数 $0<\nu\leq\mu<1$，所有 $i\in\{1,\ldots,n\}$ 以及某些常数 $\varsigma_i\in(0,1]$,

$$\max[\varsigma_i, v_{i,j}]\,(i+1)^\nu \leq w_{r,i,j} \leq \max[\varsigma_i, v_{i,j}]\,(i+1)^\mu \quad (j\geq 0), \tag{45}$$

其中，对于每个 i，$v_{i,\,j}$ 满足

$$v_{i+1,j} > v_{i,j} \quad \text{implies that} \quad v_{i+1,j} \leq |g_{r,i+1,j}| \tag{46}$$

and

$$v_{i,j} \geq |g_{r,i,j}|/a(i) \tag{47}$$

对于某个仅取决于 i的正函数 $a(i)$。使用如下形式的权重

$$v_{i,j} = \max_{t\in\{0,\ldots,i\}} |g_{r,t,j}| \tag{48}$$

在单层情形下应用于含噪示例时，取得了良好的数值表现（见 [23]）。这一特定选择称为MAXGI更新规则，满足（46）和（47）（其中 $a(i)=1$）。相关条件（14）现在规定为

$$\min_{j\in\{1,\ldots,n_{\ell-1}\}} w_{\ell-1,0,j} \geq \min_{j\in\{1,\ldots,n_\ell\}} w_{\ell,i,j}. \tag{49}$$

考虑到 (24)，我们可以看到，该定义

$$w_{\ell-1,0,j} = \max\left[\varsigma_j, \frac{\|P_\ell\|\,|R_\ell g_{\ell,i}|_j}{\alpha\|\Delta_{\ell,i}\|}, \min_{j\in\{1,\ldots,n_\ell\}} w_{\ell,i,j}\right]$$

同时蕴含 (15) 和 (49)。注意，对于层 r，我们只需要 (45)–(47)；较低层级的权重所需的增长则由 (49) 保证。

利用引理 3.5 和 (47)，可以立即推导出在层 r 的每次迭代中目标函数值变化的下界。

引理 3.7 假设 AS.1 和 AS.2 成立。那么

$$f(x_i) - f(x_{i+1}) = h_r(x_{r,i}) - h_r(x_{r,i+1}) \geq -n(\beta_{2,r} + \tfrac{1}{2}\alpha^2 L)a(i)^2 \tag{50}$$

对所有 $i \geq 0$ 均成立。

证明。忽略（43）中的负项，并结合 $\ell = r$、（45）和（47），我们推得

$$\begin{aligned} h_r(x_{r,i+1}) &\leq h_r(x_{r,i}) + \left(\beta_{2,r} + \tfrac{1}{2}\alpha^2 L\right) \sum_{j=1}^{n} \frac{g_{r,i,j}^2}{w_{r,i,j}^2} \\ &\leq h_r(x_{r,i}) + \left(\beta_{2,r} + \tfrac{1}{2}\alpha^2 L\right) \sum_{j=1}^{n} \frac{g_{r,i,j}^2}{\max[\varsigma, v_{i,j}]^2 (j+1)^\nu} \\ &\leq h_r(x_{r,i}) + \left(\beta_{2,r} + \tfrac{1}{2}\alpha^2 L\right) \sum_{j=1}^{n} \frac{g_{r,i,j}^2}{v_{i,j}^2 (j+1)^\nu} \end{aligned}$$

（50）由（6）可得。 □

现在，我们可以利用（45）—（47）和（49）陈述采用 MOFFTR 算法时的主要结果。

定理 3.8 假设 AS.1–AS.3 成立，并且将 MOFFTR 算法应用于问题（1）的一次调用（形式如（11））， 其中权重 $w_{\ell,\,i,\,j}$ 按照（45）—（47）选取，且条件（14）具体化为（49）。那么，对于任意 $\vartheta \in (0, \beta_{1,\,r})$， 存在一个子序列 $\{i_t\} \subseteq \{i\}_{i_\vartheta}^{\infty}$，使得

$$\min_{k\in\{i_\varsigma+1,\ldots,i_t\}} \|\nabla_x^1 f(x_{r,k})\|^2 = \min_{k\in\{i_\varsigma+1,\ldots,i_t\}} \|g_{r,k}\|^2 \leq \kappa_\diamond \frac{(i_t+1)^\mu}{i_t - i_\vartheta} \leq \frac{2\kappa_\diamond(i_\vartheta+1)}{i_t^{1-\mu}} \tag{51}$$

其中

$$i_\vartheta \stackrel{\text{def}}{=} \left(\frac{\beta_{2,r} + \frac{1}{2}\alpha^2 L}{\varsigma_{\min}(\beta_{1,r} - \vartheta)}\right)^{\frac{1}{\nu}} - 1, \qquad i_\varsigma \stackrel{\text{def}}{=} \left(\frac{2(i_\vartheta+1)\kappa_\diamond}{\varsigma_{\min}}\right)^{\frac{1}{1-\mu}} \tag{52}$$

and

$$\kappa_\diamond \stackrel{\text{def}}{=} \frac{2}{\vartheta}\left[f(x_0) - f_{\text{low}} + n(\beta_{2,r} + \tfrac{1}{2}\alpha^2 L)\sum_{k=0}^{i_\vartheta} a(k)^2\right].$$

证明。 见附录 A。 □

有必要对这一结果作几点说明。

1. 定理 3.8 提供了有关 MOFFTR 算法在索引为 i_ς 的迭代之后收敛速率的有用信息，而该索引可以先验地计算。事实上， i_ς 仅取决于 μ、ν 和问题常数。若 $\{i_t\} = \{i\}_{i_\varsigma}^{\infty}$，则达到满足精度要求 $\|g_k\| \leq \epsilon$ 的迭代所需的复杂度界为

$$\mathcal{O}\left(\epsilon^{-\frac{2}{1-\mu}}\right) + i_\varsigma(\mu, \nu), \tag{53}$$

对于较小的 ν 和 μ，该值可以接近 $\mathcal{O}(\epsilon^{-2})$（尽管代价是较大的 i_ς）。单层 ASTR1 算法的某些变体也能达到这一速率（见[23, 定理 4.1]）。若 $\{i_t\} \subset \{i\}_{i_\varsigma}^{\infty}$，则可能需要等到 (53) 之后的 $\{i_t\}$ 中下一次迭代，才能达到梯度界。请注意，右端项 (51) 的衰减速率取决于索引 i_t （完整序列中的索引），而非 t （子序列索引）。有趣的是，可以证明 $\{i_t\} = \{i\}_{i_\vartheta}^{\infty}$，若假设目标函数的梯度一致有界（见 [23, 定理 4.1]）。

2. 与所有最坏情况界一样，界 (51) 是悲观的。在此情形下尤其如此，因为我们只考虑了这样一种情况：在 i_ϑ 之前的所有迭代都使目标函数增加了假设所允许的最大幅度。在实践中，这种情况极不可能发生。

3. 还可以放宽条件 (49)，只要求右侧至少为左侧的一个固定比例。相关论证基本不变，但证明中会引入另一个常数，并贯穿整个证明。为避免增加记号负担，我们没有纳入这种可能性。

4. [23, 定理 4.2] 中已证明，对于单层情形，上述复杂度界是紧的。因此，对于多层情形，该界也同样是紧的。

5. 请注意，要求 (45) 允许对权重作多种选择。下一节将从数值角度探讨具体选择 (48)。

3.2 AdaGrad 类权重

我们不再关注 (45)，而是考虑一种受广受欢迎的 AdaGrad 方法启发的权重选择方式，通过累积梯度分量的平方来实现权重大小的必要增长。有趣的是，这将使我们能够证明完整迭代点序列的复杂度结果（不涉及任何子序列）。更具体地说，给定 $\varsigma \in (0, 1]$ and $\mu \in (0,1)$，我们对所有 $\ell \in \{1, \ldots, r\}$、所有 $i \geq 0$ 和 $j \in \{1, \ldots, n_\ell\}$ 定义权重如下

$$w_{\ell,i,j} \stackrel{\text{def}}{=} \left(\varsigma + \sum_{i=0}^{j} g_{\ell,i,j}^2\right)^{\mu}. \tag{54}$$

当 $\mu = \frac{1}{2}$ 时，可得到 AdaGrad 权重。此时，条件 (14) 被具体规定为要求

$$\|w_{\ell-1,0}\| \geq \|w_{\ell,i}\|. \tag{55}$$

再次考虑 (24)，我们验证定义

$$w_{\ell-1,0,j} = \max\left[1, \frac{\|w_{\ell,i}\|}{\|\widehat{w}_{\ell-1,0}\|}\right] \widehat{w}_{\ell-1,0,j} \quad \text{where} \quad \widehat{w}_{\ell-1,0,j} = \max\left[\varsigma_j, \frac{\|P_\ell\| \, |R_\ell g_{\ell,i}|_j}{\alpha \|\Delta_{\ell,i}\|}\right] \tag{56}$$

足以确保 (15) 和 (55) 均成立。

现在我们给出采用 (54) 和 (55) 的 MOFFTR 算法变体的复杂度结果。该结果与 [23, 定理 3.2]类似，但采用了由引理 3.6 给出的更复杂的线性下降多层版本。

定理 3.9 假设 AS.1–AS.3 成立，且在以 (11) 形式调用时，将 MOFFTR 算法应用于问题 (1)，其中权重按照 (54) 选取，并将 (14) 实例化为 (55)。则

$$\operatorname*{average}_{k\in\{0,\ldots,i\}} \|\nabla_x^1 f(x_{r,k})\|^2 = \operatorname*{average}_{k\in\{0,\ldots,i\}} \|g_{r,k}\|^2 \leq \frac{\kappa_*}{i+1}, \tag{57}$$

其中

$$\kappa_* \stackrel{\rm def}{=} \begin{cases} \max\left[\varsigma, \left(\frac{4n(\beta_{2,r}+\frac{1}{2}L)}{\beta_{1,r}(1-2\mu)}\right)^{\frac{1}{\mu}}, \frac{1}{2}\left(\frac{(1-2\mu)\Gamma_0}{n(\beta_{2,r}+\frac{1}{2}L)}\right)^{\frac{1}{1-2\mu}}\right] & \text{if} \quad 0<\mu<\frac{1}{2}, \\ \max\left[\varsigma, \frac{1}{2}e^{\frac{2\Gamma_0}{n(\beta_{2,r}+\frac{1}{2}L)}}, \frac{\varsigma\psi^2}{2}\left|W_{-1}\left(-\frac{1}{\psi}\right)\right|^2\right] & \text{if} \quad \mu=\frac{1}{2}, \\ \max\left[\varsigma, \left[\frac{2^\mu}{\beta_{1,r}}\left(\Gamma_0+\frac{n(\beta_{2,r}+L)\varsigma^{1-2\mu}}{2\mu-1}\right)\right]^{\frac{1}{1-\mu}}\right] & \text{if} \quad \frac{1}{2}<\mu<1, \end{cases} \tag{58}$$

其中 $\beta_{1,r}$ 和 $\beta_{2,r}$ 分别是由 (41) 和 (42) 定义的常数，

$$\psi \stackrel{\rm def}{=} \frac{4\max[\frac{3}{2}\beta_{r,1}, n(\beta_{r,2}+\frac{1}{2}L)]}{\beta_{1,r}\sqrt{\varsigma}} \tag{59}$$

并且 W_{-1} 是朗伯函数 [12]的第二分支。

证明。 参见附录 B。 □

我们以几条关于定理 3.9 的简短评论结束本节分析。

1. 由该定理可知，对于任意 $\epsilon>0$，至多需要进行 $\mathcal{O}(\epsilon^{-2})$ 次 MOFFTR算法迭代，即可将 $\|g_{r,k}\|$ 降至 ϵ 以下。因此，除非 $\{i_t\}=\{i\}_{i_\vartheta}^{\infty}$，该结果比定理 3.8 的结论更强。其阶数也与单层信赖域算法已知的结果相同（见 [7, 定理 2.3.7]）。

2. 通过选择 $r=1$ 和 $\mu=\frac{1}{2}$，可得到 AdaGrad 的无动量精确版本。因此，定理 3.9 为该方法的单层和多层版本都提供了收敛性分析。

3. [23, 定理 3.4] 已证明，使用 AdaGrad 类权重的单层（$r=1$）算法的全局收敛速度不可能优于 $\mathcal{O}(1/\sqrt{i})$。因此，$r\geq 1$ 也有同样的结论。

4. 如果愿意假设梯度一致有界，则可以改进界 (57)（尽管阶数不会提高）。单层情形的证明请参见 [23]。

5. 如最后一篇参考文献所述，利用不等式，可以用一个较弱但更明确的界替代涉及朗伯函数的界

$$\left|W_{-1}(-e^{-x-1})\right| \leq 1+\sqrt{2x}+x \quad \text{for} \quad w>0 \tag{60}$$

[9, 定理 1]。记住，对于 γ_1 和由（83）给出的 γ_2，$\log\left(\frac{\gamma_2}{\gamma_1}\right) \geq \log(3) > 1$，并在（60）中令 $x = \log\left(\frac{\gamma_2}{\gamma_1}\right) - 1 > 0$，则可得

$$\left|W_{-1}\left(-\frac{\gamma_1}{\gamma_2}\right)\right| \leq \log\left(\frac{\gamma_2}{\gamma_1}\right) + \sqrt{2\left(\log\left(\frac{\gamma_2}{\gamma_1}\right) - 1\right)}.$$

6. 由 (54) 所蕴含的权重严格单调性也可以放宽，以提供更大的算法灵活性。事实证明，(54) 可以替换为

$$w_{\ell,i,j} \in [\chi\, v_{\ell,i,j}, v_{\ell,i,j}] \;\text{ with }\; v_{\ell,i,j} \stackrel{\rm def}{=} \left(\varsigma + \sum_{i=0}^{j} g_{\ell,i,j}^2\right)^{\mu}$$

对于某个（固定的） $\chi \in (0, 1]$，而不改变定理 3.9 的性质，只需修改常数 κ_*，使其显式包含 χ。同样，单层情形的证明见 [23]。

4 扩展

上述理论可以通过多种方式加以扩展，这些方式在实践上有用和／或在理论上有趣。

4.1 依赖于迭代的算法要素

与第 5 页中的表述相比，MOFFTR 算法的实现具有很大的灵活性，部分原因是我们将某些算法参数视为常数。尽管在许多情况下这是一项优势，但在特定问题中，让这些参数在固定范围内变化也可能提升性能。例如，α就是如此，它是较低层级信赖域半径可超过上层信赖域半径的倍数。f_ℓ 的定义也属于这种情况，我们的分析中从未明确使用该定义；(13) 式第二部分右侧的因子 2 也是如此。因此，可以说，我们的分析涵盖了一整类可能的算法。

4.2 利用较低层级的迭代

到目前为止，我们所介绍的多层理论仅限于利用每一层的第一次迭代（见（37）与（38）之间的过渡）。从某种意义上说，这种方法相当粗略，因为它没有说明为什么在较低层级执行多于一次的迭代可能有利于收敛。为了加深理解，我们需要进一步说明 $h_{\ell-1}$ 如何提供 h_ℓ 的近似。在迭代 $(\ell,\ i)$ 中，$\Delta_{\ell,\,i}$ 旨在表示以 $x_{\ell,\,i}$ 为中心的球的半径，在该球内，一阶泰勒模型能够足够好地近似 h_ℓ。如果 $f_{\ell-1}$ 反过来以某种方式近似 f_ℓ，我们预计 $h_{\ell-1}$ 中的线性下降（即 $\sum_{j=1}^{n_{\ell-1}} g_{\ell-1,\,k,\,j}^2 / w_{\ell-1,\,k,\,j}$）应当与层级 ℓ 上的线性下降（即 $\sum_{j=1}^{n_\ell} g_{\ell-1,\,k,\,j}^2 / w_{\ell-1,\,k,\,j}$）保持一致，其中所考虑的球的延拓半径为 $\Delta_{\ell,\,i}$。下文中，我们将讨论：如果假设（或强制）条件（16）对于所有迭代 $(\ell-1,\ k)$ 都不成立，而不仅仅是对于 $(\ell-1,\ 0)$ 不成立，即如果

$$\sum_{j=1}^{n_{\ell-1}} \frac{g_{\ell-1,k,j}^2}{w_{\ell-1,k,j}} \geq \kappa_{\rm R} \sum_{j=1}^{n_\ell} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}} \quad \text{for} \quad k \in \{0, \ldots, i_{\ell-1}\}. \tag{61}$$

‡在本讨论中，我们忽略了这样一个事实：在较低层级计算梯度通常比在上层计算梯度所需的计算成本显著更低，而这一优势在实践中有时至关重要。

（与（16）相比，我们可能需要使用一个更小的 $\kappa_{\mathrm{R}\circ}$ ）回到引理 3.5，在这一强化假设下，我们得到以下结果。

引理 4.1 假设 AS.1、AS.2 和 (61) 成立。那么，对于所有 $\ell \in \{1, \ldots, r\}$ 和所有 $i \geq 0$,

$$g_{\ell,i}^T s_{\ell,i} \leq -i_\ell^{(\mathrm{low})} \beta_{1,r} \sum_{j=1}^{n} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}} + \beta_{2,r} \sum_{j=1}^{n} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}^2} \tag{62}$$

对于某些常数 $\beta_{1,r} > 0$ 和 $\beta_{2,r} > 0$ ，它们与 ℓ 和 i 无关；其中 $i_\ell^{(\mathrm{low})}$ 是从迭代 (ℓ, i) 到（但不包括）迭代 $(\ell, i+1)$ 的泰勒迭代总数。

证明。我们首先沿用引理 3.5 的证明，直到 (37)。然后，我们不再像得到 (38) 那样忽略各项，而是保留它们，并利用 (61) 和 (13) 推导出

$$\begin{aligned} g_{\ell,i}^T s_{\ell,i} &\leq -i_{\ell-1} \frac{\tau \varsigma_{\min}}{2\kappa_{\mathrm{B}}\omega} \sum_{j=1}^{n_{\ell-1}} \frac{g_{\ell-1,0,j}^2}{w_{\ell-1,0,j}} + \frac{2 i_{\ell-1}^{(\max)} (\kappa_{\mathrm{B}} + L)}{\omega \sigma_{\min}[P_\ell]^2} \alpha^2 \|\Delta_{\ell,i}\|^2 \\ &\leq -i_{\ell-1} \frac{\tau \varsigma_{\min} \kappa_{\mathrm{R}}}{2\kappa_{\mathrm{B}}\omega} \sum_{j=1}^{n_\ell} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}} + \frac{2\alpha^2 i_{\ell-1}^{(\max)} (\kappa_{\mathrm{B}} + L)}{\omega \sigma_{\min}[P_\ell]^2} \|\widehat{\Delta}_{\ell,i}\|^2 \\ &\leq -i_{\ell-1} \frac{\tau \varsigma_{\min} \kappa_{\mathrm{R}}}{2\kappa_{\mathrm{B}}\omega} \sum_{j=1}^{n_\ell} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}} + \frac{2\alpha^2 i_{\ell-1}^{(\max)} (\kappa_{\mathrm{B}} + L)}{\omega \sigma_{\min}[P_\ell]^2} \sum_{j=1}^{n_\ell} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}^2}, \end{aligned} \tag{63}$$

其中，与引理 3.5 一样，我们利用 (16) 和 (13) 不成立这一条件，得到了最后两个不等式。与引理 3.5 一样，若迭代 (ℓ, i) 是一次泰勒迭代，则 (40) 仍然成立（此时 $i_{\ell-1} = 0$）。结合 (63) 和 (40)，可得，对于层 ℓ 上的所有迭代，

$$\begin{aligned} g_{\ell,i}^T s_{\ell,i} \leq & -\max[1, i_{\ell-1}] \frac{\kappa_{\mathrm{R}}}{\max[\omega, 1]} \left[\frac{\tau \varsigma_{\min}}{2\kappa_{\mathrm{B}}} \right] \sum_{j=1}^{n_\ell} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}} \\ & + \max \left\{ [\kappa_{\mathrm{B}}], \frac{2\alpha^2 i_{\ell-1}^{(\max)}}{\max[\omega, 1]\, \sigma_{\min}[P_\ell]^2} \Big([\kappa_{\mathrm{B}}] + L \Big) \right\} \sum_{j=1}^{n_\ell} \frac{g_{\ell,i,j}^2}{w_{\ell,i,j}^2}. \end{aligned}$$

于是，我们可以利用 (41) 和 (42) 递归地定义 $\beta_{1,\ell}$ 和 $\beta_{1,\ell}$ ，最后由 $i_\ell^{(\mathrm{low})}$ 的定义可得 (62)。□

注意，(62) 与 (36) 相同，只是现在 $\beta_{1,r}$ 乘以了 $i_\ell^{(\mathrm{low})}$ 。这一修改贯穿所有证明，使 (52) 和 (58) 中的[§] 常数得到改进。最后我们指出，要求 (61) 可以视为改善 (62) 中负项与正项之间平衡的一种方法，但未必是唯一的方法。

4.3 弱相干性

高层级与低层级模型之间的线性相容性要求可以适当放宽。考察上述理论可知，该要求仅在 (33) 中用到。若我们假设 $\omega P_\ell^T = R_\ell + E_\ell$ 而非 (2)，那么很容易验证，满足以下条件的误差矩阵 E_ℓ

$$\|E_\ell g_{\ell,i}\| \leq \kappa_E \delta_{\ell-1} \tag{64}$$

[§]特别是，通过抵消$\beta_{2,r}$中 $i_{\ell-1}^{(\max)}$ 常数的影响。

对于某个固定的 $\kappa_E \geq 0$，可确保对（33）中的每个 t，

$$g_{\ell,i}^T P_\ell s_{\ell-1,t} = \frac{1}{\omega}(R_\ell g_{\ell,i})^T s_{\ell-1,t} + \frac{\kappa_E \delta_{\ell-1}\|s_{\ell-1,t}\|}{\omega} \leq \frac{1}{\omega} g_{\ell-1,i}^T s_{\ell-1,t} + \frac{2\kappa_E}{\omega\sigma_{\min}[P_\ell]}\,\delta_{\ell-1}^2,$$

这里我们使用 (9) 推导出最后一个不等式。这在 (32) 的右侧添加了一个 $\mathcal{O}(\delta_{\ell-1}^2)$ 项，因此可以用不同的常数继续论证。条件 (64) 是可实施的，因为在迭代 $(\ell,\ i)$ 中使用 R_ℓ 或 P_ℓ 之前，$\delta_{\ell-1}$ 已经是已知的；而在 $\|g_{\ell,\ i}\|$ 仍然相对较大的早期迭代中，该条件可以相当宽松。线性相容性通常只需近似保持，这一点在算法应用于梯度含噪的问题时尤为重要。在这种情况下，坚持精确的线性相容性只会将较高层级的梯度误差传递到较低层级，这显然是不可取的。

5 数值结果

本节将展示所提出的MOFFTR 算法在深度神经网络训练中的数值表现，尤其关注监督学习应用。设 $\mathcal{D} = \{(y_s,\ c_s)\}_{s=1}^{n_s}$ 为一个带标签数据集，其中 $y_s \in \mathrm{IR}^{n_{in}}$ 表示输入特征，$c_s \in \mathrm{IR}^{n_{out}}$ 表示期望目标。我们的目标是学习深度神经网络的参数，使其能够针对给定的 y_s逼近 c_s。下文中，我们采用深度神经网络的连续深度方法，其前向传播可解释为对非线性常微分方程（ODE）[44, 8, 50]的离散化。这种方法使我们能够以相当自然的方式构建MOFFTR 框架所需的多层级结构和传递算子（见第 5.1 节）。

采用连续深度方法，监督学习问题可表述为以下连续最优控制问题 [29]:

$$\begin{aligned} \min_{Q,q,\theta,W_T,b_T} \quad & \frac{1}{n_s}\sum_{s=1}^{n_s} \hbar(\mathcal{P}(W_T q_s(T) + b_T), c_s) + \int_0^T \mathcal{R}(\theta(t))\ dt + \mathcal{S}(W_T, b_T), \\ \text{subject to} \quad & \partial_t q_s(t) = \mathcal{F}(q_s(t), \theta(t)), \qquad \forall t \in (0,T), \\ & q_s(0) = Q y_s, \end{aligned} \tag{65}$$

其中，q 表示从 IR 到 $\mathrm{IR}^{n_{fp}}$ 的时变状态，θ 表示从 IR 到 IR^{n_c}的时变控制参数。(65) 中的约束将输入特征 y_s连续变换为最终状态 $q_s(T)$，该状态定义于时间 T。这一过程分两步完成。首先，将输入 y_s 映射到动力系统的维度，得到 $q_s(0) = Qy_s$，其中 $Q \in \mathrm{IR}^{n_{fp}\times n_{in}}$。其次，使用一个“残差块” $\mathcal{F}$，将特征从 $\mathrm{IR}^{n_{fp}} \times \mathrm{IR}^{n_c}$ 非线性变换到 $\mathrm{IR}^{n_{fp}}$。我们考虑两种此类残差块：全连接型和卷积型。全连接残差块定义为 $\mathcal{F}(q_s(t)\ \theta(t)) := \sigma(W(t)q_s(t) + b(t))$，其中 $\theta(t) = (\mathrm{flat}(W(t)), \mathrm{flat}(b(t)))$，$\sigma$ 是从 $\mathrm{IR}^{n_{fp}}$ 到 $\mathrm{IR}^{n_{fp}}$的激活函数，$b(t) \in \mathrm{IR}^{n_{fp}}$是“偏置”，而 $W(t) \in \mathrm{IR}^{n_{fp}\times n_{fp}}$ 是一个稠密矩阵。卷积残差块的形式为 $\mathcal{F}(q_s(t)\ \theta(t)) := \sigma(\mathrm{BN}(t, W(t)q_s(t) + b(t)))$，其中 $W(t)$现在表示一个稀疏卷积算子，BN 表示连续时间批归一化 [32, 44]，其将输入从 IR $\times \mathrm{IR}^{n_{fp}}$映射到 $\mathrm{IR}^{n_{fp}}$。

(65) 中的目标函数定义为：最小化期望目标 c_s 与预测输出 $\hat{c}_s \in \mathrm{IR}^{n_{out}}$ 之间的偏差，该偏差由损失函数 $\hbar$（从 $\mathrm{IR}^{n_{out}\times n_{out}}$ 映射到 IR）度量。这里，预测输出按如下方式获得：$\hat{c}_s := \mathcal{P}(W_T q_s(T) + b_T) \in \mathrm{IR}^{n_{out}}$，其中 $\mathcal{P}$ 表示一个从 $\mathrm{IR}^{n_{out}}$ 映射到 $\mathrm{IR}^{n_{out}}$ 的假设函数。线性算子 $W_T \in \mathrm{IR}^{n_{out}\times n_{fp}}$ 和 $b_T \in \mathrm{IR}^{n_{fp}}$ 用于对提取特征 $q_s(T)$ 进行仿射变换，即在时间 T 处由动力系统输出所得到的特征。参数为 β_1, $\beta_2 > 0$ 的正则化项 $\mathcal{R}$ 和 $\mathcal{S}$ 定义如下。使用蒂霍诺夫正则化来惩罚

W_T 和 b_T 的大小，即 $\mathcal{S}(W_T, b_T) := \frac{\beta_1}{2}\|W_T\|_F^2 + \frac{\beta_1}{2}\|b_T\|^2$，其中 $\|\cdot\|_F$ 表示 Frobenius 范数。对于随时间变化的控制参数，我们使用 $\mathcal{R}(\theta(t)) := \frac{\beta_1}{2}\|\theta(t)\|^2 + \frac{\beta_2}{2}\|\partial_t\theta(t)\|^2$，其中第二项确保参数随时间平滑变化，详情参见 [29] 。

为对问题 (65) 进行数值求解，我们采用先离散后优化方法。离散化使用等距网格 $0 = \tau_0 < \cdots < \tau_{K-1} = T$，其中包含 K 个点。随后，在给定时刻 τ_k ，分别将状态和控制近似为 $q_k \approx q(\tau_k)$ 和 $\theta_k \approx \theta(\tau_k)$。请注意，现在每个 θ_k 和 q_k 都对应于深度神经网络第 k 层相关的参数和状态。我们的时间离散化采用前向欧拉格式，由此得到具有恒等跳跃连接的著名 ResNet 架构 [31]。或者，也可以采用更先进且数值稳定性可能更好的时间积分格式，例如参见 [29]。对于此处考虑的显式欧拉格式，我们通过采用足够小的时间步长 $\Delta_t = T/(K-1)$ 来确保数值稳定性。

5.1 实现与算法设置

我们的残差网络实现基于深度学习库 Keras [15], ，而 MOFFTR框架则使用 NumPy 库实现。我们考虑四种不同的一阶 MOFFTR 算法变体（即 $B_{\ell,\ i} = 0$ 对所有 $(\ell,\ i)$ ）。第一种变体采用发散权重，其定义由式（45）和（48）给出的 MAXGI 更新规则指定，其中 $\mu = 0.1$。其余变体均采用 AdaGrad 类权重，由式（54）和（56）指定，其中 $\mu \in \{0.1, 0.5, 0.9\}$。所选更新规则用于更新所有层级的权重。MOFFTR 算法采用 V 循环实现，包含一个预平滑步骤和零个后平滑步骤。对于采用密集残差块的残差网络，我们在最低层执行 10 次迭代，即 $i_1^{(\max)} = 10$ ，并采用 $\kappa_R = 0.01$ 和$\alpha = 5$。对于卷积残差网络，我们采用 $i_1^{(\max)} = 5$ 、 $\kappa_R = 0.001$ 和 $\alpha = 25$。所有数值示例中的参数 ν和 ς均设为 $\nu = 0.1$，以及 $\varsigma = 0.01$ 。此外，我们考虑了第 4.3 节中的讨论；由于在子采样噪声存在的随机设置下应用MOFFTR 框架，因此不施加一阶相干关系（7）。在我们的 MOFFTR 算法实现中，所有 $(\ell,\ i)$ 均使用子采样导数。

通过使用不同的离散化参数 Δ_t 对问题 (65) 进行离散化，可得到 MOFFTR 框架所需的目标函数层级 $\{f\}_{\ell=1}^r$ 。随后，每个 f_ℓ 都对应一个深度不同的网络。除非另有说明，以下所有数值示例均采用三个层级，并通过因子为二的均匀细化得到。算子 $\{P\}_{\ell=1}^{r-1}$ 通过一维（时间方向）上的分段线性插值构造，详情参见 [18, 35] 。请注意，在 [30, 13, 8] 的多层参数初始化背景下，也采用了类似方法来构造延拓算子。我们根据 (2) 定义限制算子 $\{R\}_{\ell=1}^{r-1}$ ，对于具有全连接残差块和卷积残差块的网络，分别选取 $\omega = \frac{1}{2}$ 和 $\omega = 1$ 。

为评估 MOFFTR 方法的性能，我们将其与单层ASTR1 方法进行比较；后者是通过调用相应的 MOFFTR 算法并使用 $r = 1$ 得到的。比较中还包括基线随机梯度（SGD）[46], ADAM [33]和 AdaGrad¶ 方法。所有方法的学习率均通过充分的超参数搜索选定，并针对每个数据集、网络和批量大小分别进行搜索。具体而言，我们考察集合 $\{0.0001, 0.00025, 0.0005, 0.00075, 0.001, 0.0025, 0.005, 0.0075$,**0.01**, **0.025**, **0.05**, **0.075**, **0.1**, **0.25**, **0.5**, **0.75**, $1.0\}$ 中的学习率。随后，在所呈现的数值实验中，采用能够取得最佳泛化结果（对五次独立运行取平均）的学习率。对于 MOFFTR 方法，所有层级均采用相同的学习率。

下文通过分析单层方法和多层方法的主要计算成本（即与梯度评估相关的成本）来进行比较。设C_r 为使用完整数据集 $\mathcal{D}$在最上层进行一次梯度评估的计算成本。根据 C_r 的定义，并利用反向传播的成本与层数和样本数呈线性关系这一事实，我们

¶通过调用 MOFFTR 算法并采用 $r = 1$、由（54）给出的权重以及 $\mu =$，即可得到 AdaGrad $\frac{1}{2}$.

将总计算成本 C 定义如下：

$$C = \sum_{\ell=1}^{r} 2^{\ell-r} \sharp_\ell \, C_r, \tag{66}$$

其中，缩放因子[‖] $2^{\ell-r}$ 用于体现与层 ℓ和层 r相关的成本差异。符号 $\sharp_\ell$ 表示在使用完整数据集 $\mathcal{D}$的层 ℓ上执行的梯度评估次数。例如，如果我们在使用 n_b 个样本的层 ℓ 上评估梯度三次，那么 $\sharp_\ell = 3n_b/|\mathcal{D}|$。

所有实验均使用瑞士国家超级计算中心（CSCS）的 Piz Daint 超级计算机上的 XC50计算节点（Intel Xeon E5-2690 v3处理器、NVIDIA Tesla P100显卡）完成。

5.2 数值示例

我们通过分类和回归领域中的三个数值示例，研究所提出的MOFFTR 算法的收敛性质和有效性。

5.2.1 使用 Indian Pines 数据集进行高光谱图像分割

我们的第一个示例是利用 Indian Pines 数据集 [2] 提供的高光谱图像进行土壤分割。输入数据由机载可见光/红外成像光谱仪（AVIRIS）传感器采集，包含 145×145 个像素，每个像素有 200 个光谱波段，波长范围为 400 至 2500 nm。在所有像素中，只有 10，249 个带标签数据，我们将其分为 7，174 个训练样本和 3，075 个验证样本。每个像素都被分配到 16 个类别之一，用以表示土地覆盖类型，例如玉米、大豆等。分割采用 ResNet 完成，其中使用宽度为 50 的密集残差块和 $ReLu$ 激活函数 σ。参数 K，T，β_1，β_2设为 $K = T = 3$ 和 $\beta_1 = \beta_2 = 10^{-3}$。此外，我们使用 softmax 假设函数和交叉熵损失函数，其定义为 $h(\hat{c}_s, c_s) := c_s^T \log(\hat{c}_s)$。为训练 ResNet，我们运行不带动量的 MOFFTR 、 ASTR1 和 随机梯度下降 方法的变体。当 $\text{acc}_{\text{train}} > 0.98$ 或 $\text{acc}_{\text{val}} > 0.98$时，所有方法都会立即终止。当$\sum_{i=1}^{15}(\text{acc}_{\text{train}})_e - (\text{acc}_{\text{train}})_{e-i} < 0.001$或 $\sum_{i=1}^{15}(\text{acc}_{\text{val}})_e - (\text{acc}_{\text{val}})_{e-i} < 0.001$时，也会终止，其中 $(\text{acc}^{\text{train/val}})_e$定义为给定训练轮次 e下，训练/验证数据集中正确分类样本数与训练/验证数据集样本总数之比。

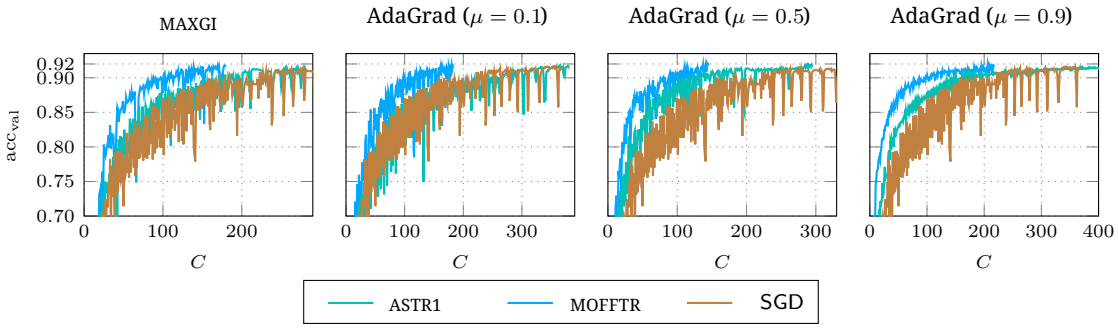

图1：Indian Pines 示例：验证准确率随总计算成本 C 的变化情况。报告的是 10 次独立运行中验证准确率最高的一次。实验使用的批量大小为 1,024。

表1报告了随着噪声增大（即批量大小减小），所有求解策略所需的总计算成本和验证准确率。如表中所示，所有

[‖]假设在一维中按二倍因子进行均匀粗化。

方法		批量大小（批量大小/$\lvert\mathcal{D}\rvert$）							
		512（6%）		1,024（12%）		4,096（50%）		8,100（100%）	
		C	准确率$_{\text{val}}$	C	准确率$_{\text{val}}$	C	准确率$_{\text{val}}$	C	准确率$_{\text{val}}$
SGD		359	92.5%	369	91.7%	1,230	92.2%	3,093	92.1%
ASTR1	AdaGrad($\mu = 0.1$)	230	92.6%	379	91.8%	979	92.3%	2,748	92.4%
	AdaGrad($\mu = 0.5$)	143	92.4%	295	91.8%	853	92.3%	3,171	91.9%
	AdaGrad($\mu = 0.9$)	232	91.4%	404	91.6%	976	92.0%	2,626	91.7%
	MAXGI	195	92.4%	281	91.7%	884	92.2%	2,719	91.8%
MOFFTR	AdaGrad($\mu = 0.1$)	125	92.4%	183	91.7%	422	92.3%	1,683	92.3%
	AdaGrad($\mu = 0.5$)	113	92.1%	145	92.1%	412	92.2%	1,394	92.3%
	AdaGrad($\mu = 0.9$)	118	91.5%	217	91.9%	453	92.1%	1,760	91.9%
	MAXGI	96	92.3%	174	91.7%	334	91.9%	1,292	92。4%

表1：Indian Pines 示例：训练 ResNet 所需的总计算成本 C 和验证准确率 acc_{val}。报告 10 次独立运行中验证准确率方面的最佳结果。

求解策略在给定批量大小下都能达到相近的验证准确率。然而，MOFFTR 所有变体的计算成本都低于其单层对应方法和随机梯度下降 方法，另见图1。与 ASTR1相比，加速因子在 1.2 到 2.6 之间波动。与 随机梯度下降 方法相比，加速更明显，范围为 1.7 到 3.7。此外，我们的结果表明，即使批量大小较小，也能始终观察到加速效果。这从经验上证实，所提出的算法框架既保留了多层方法更快的收敛特性，同时又像大多数 OFFO 方法一样，不受子采样噪声的影响。

5.2.2 参数化中子扩散-反应（NDR）的代理建模)

我们的第二个示例研究了如何为具有空间变化系数和外部源项的参数化中子扩散-反应问题构建代理模型。目标是构建一个代理模型，以预测给定参数集下的平均中子通量。为此，我们生成了一个包含 3,000 个样本的数据集，并将其划分为 2,600 个训练样本和 400 个测试样本。按照 [43]，计算域 $\Omega = (0{,}170)^2$ 是非均匀的，由四个不同的材料区域组成，分别记为 $\Omega_1, \ldots, \Omega_4$（见图 2 左图）。

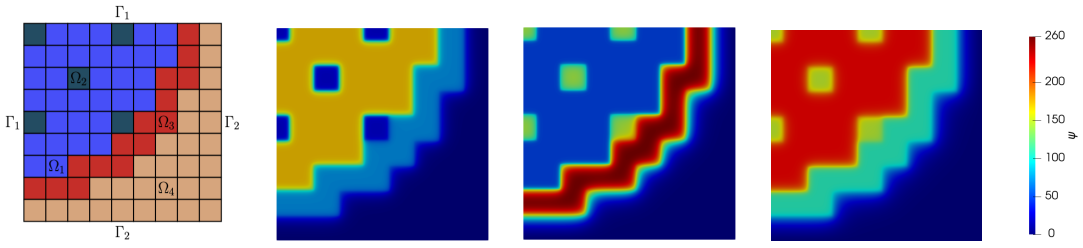

图 2：左：用于构建 NDR 数据集的计算域。每个子域以不同颜色表示。中至右：NDR 数据集中的样本示例。

该问题的强形式表示为

$$\begin{aligned} \nabla \cdot [D(x)\nabla\psi(x)] + \alpha(x)\nabla\psi(x) &= q(x), && \text{in } \Omega, \\ \psi(x) &= 0, && \text{on } \Gamma_1 := [0,170]\times\{1\}\cup\{0\}\times[0,170], \\ D(x)\nabla\psi(x)\cdot n(x) &= 0, && \text{on } \Gamma_2 := [0,170]\times\{0\}\cup\{1\}\times[0,170], \end{aligned} \tag{67}$$

其中， x 表示空间坐标， ψ 表示从 Ω 到 IR 的中子通量。函数 D、 α、 q 分别定义为 $D(x) = \sum_{i=1}^{4} \mathbb{1}_{\Omega_i}(x) D_i$、 $q(x) = \sum_{i=1}^{3} \mathbb{1}_{\Omega_i}(x) q_i$ 和 $\alpha(x) = \sum_{i=1}^{4} \mathbb{1}_{\Omega_i}(x) \alpha_i$（$\mathbb{1}_{\Omega_i}$ 表示区域 Ω 的示性函数）。随后，使用 11 个参数对问题 (67) 进行参数化，这些参数从均匀分布 $\mathcal{U}(a, b)$ 中采样，其取值由下界 (a) 和上界 (b) 指定。具体而言，扩散系数 $\{D_i\}_{i=1}^{3}$ 从 $\mathcal{U}(0.15, 0.6)$ 中采样，而 D_4 从 $\mathcal{U}(0.2, 0.8)$ 中采样。反应系数 α_1、.....、 α_4 分别取自 $\mathcal{U}(0.0425, 0.17)$、$\mathcal{U}(0.065, 0.26)$ 和 $\mathcal{U}(0.04, 0.16)$、 $\mathcal{U}(0.005, 0.02)$。源项 $\{q_i\}_{i=1}^{3}$ 的取值从 $\mathcal{U}(5, 20)$ 中采样，而 q_4 的值设为 0。对于每组参数/输入特征，我们通过使用有限元法在四边形网格上（两个空间维度均为 500 个单元）数值求解 (67)，生成目标值 $c_s = \int_\Omega \psi(x) dx / \int_\Omega dx$。

为构建所需的代理模型，我们使用宽度为 10 的密集残差块、双曲正切激活函数 σ 以及参数 $T = K = 3$ 和 $\beta_1 = \beta_2 = 10^{-4}$来训练 ResNet。恒等假设函数和均方损失泛函定义为 $\hbar(\hat{c}_s, c_s) := \|c_s - \hat{c}_s\|_2^2$。

与回归问题中的常见做法一样，我们使用带动量的 MOFFTR、ASTR1、随机梯度下降和 ADAM 方法的变体来训练 ResNet。有关如何在多层框架中处理动量的细节，请参见 [35, 附录 A]。训练在固定的计算预算下进行，即所有求解策略在 $C > C_{\max}$ 时立即终止，其中我们将 $C_{\max}$ 设为 2,000。

方法		批量大小（批量大小/$\|\mathcal{D}\|$）			
		256（25%）		1,024（100%）	
		$f_{\text{train}}(\times 10^3)$	$f_{\text{val}}(\times 10^3)$	$f_{\text{train}}(\times 10^3)$	$f_{\text{val}}(\times 10^3)$
SGD		0.61 ± 0.30	3.57 ± 1.44	1.87 ± 1.32	6.87 ± 1.09
ADAM		0.39 ± 0.23	5.77 ± 0.79	0.47 ± 0.50	5.35 ± 0.23
ASTR1	AdaGrad($\mu = 0.1$)	0.56 ± 0.13	5.41 ± 0.72	1.01 ± 0.37	5.72 ± 1.18
	AdaGrad($\mu = 0.5$)	0.47 ± 0.32	4.62 ± 2.71	0.79 ± 0.12	7.04 ± 4.93
	AdaGrad($\mu = 0.9$)	0.54 ± 0.17	7.43 ± 2.48	0.86 ± 0.10	4.93 ± 2.56
	MAXGI	0.53 ± 0.35	5.67 ± 1.23	1.04 ± 0.29	4.98 ± 1.28
MOFFTR	AdaGrad($\mu = 0.1$)	0.48 ± 0.29	0.89 ± 0.31	0.48 ± 0.73	1.02 ± 0.42
	AdaGrad($\mu = 0.5$)	0.39 ± 0.31	0.95 ± 0.29	0.63 ± 0.45	1.29 ± 0.53
	AdaGrad($\mu = 0.9$)	0.45 ± 0.25	1.06 ± 0.36	0.62 ± 0.41	1.36 ± 0.49
	MAXGI	0.51 ± 0.14	0.91 ± 0.30	0.73 ± 0.47	1.38 ± 0.45

表2：NDR 示例：通过训练 ResNet 2，000 个训练轮次后得到的训练损失（f_{train}）和验证损失（f_{val}）的均值与标准差。统计结果由 10 次独立运行得出。

我们从表2中观察到，除随机梯度下降外，所有求解策略得到的训练损失值（f_{train}）都相近。有趣的是，我们还注意到，MOFFTR 算法的所有变体都具有更好的泛化能力，即其验证损失（f_{val}）低于所有单层方法。这一现象也可以在图3中观察到。表2中的结果表明，MOFFTR 方法在存在噪声时仍能保持良好的泛化特性；这一性质对于代理建模及其他需要可靠解的科学应用尤为重要。

5.2.3 使用 SVNH 数据集进行图像分类

我们最后一个数值示例与使用 SVNH数据集 [42]进行图像分类有关。每幅图像由 32×32 像素表示，并包含从 0 到 9 的相互重叠的数字。该数据集包含 99,289 个样本，其中 73, 257 个用于训练，26, 032 个用于测试。我们通过对图像进行标准化来预处理所有样本，使像素值处于范围 $[0, 1]$, , 并从每个像素中减去均值。此外，我们使用标准的数据增强技术，特别是图像旋转、水平和垂直平移以及水平翻转。图像分类使用 ResNet 完成，该网络包含卷积残差块（32

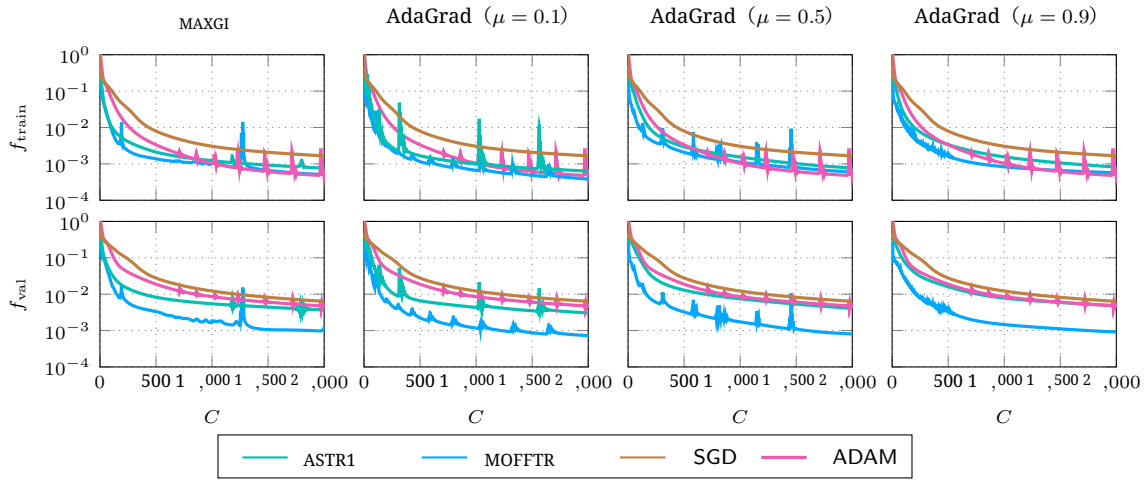

图3：NDR 示例：使用批量大小为 1,024（完整数据集）训练 ResNet 2,000 个训练轮次期间，训练损失函数（f_{train}）和验证损失函数（f_{val}）的变化历程。报告 10 次独立运行中最低f_{val} 的结果。

个滤波器）、$ReLu$ 激活函数和参数 $T = 3$、 $K = 5$、 $\beta_1 = 10^{-3}$、 $\beta_2 = 0.005$。此外，我们使用 soft max 假设函数和交叉熵损失函数。

ResNet 的训练采用 512 的批量大小，并使用与第 5.2.1 节 Indian Pines 示例相同的停止准则。在此实验中，我们考察不使用动量的MOFFTR、ASTR1 和 随机梯度下降 方法。表3报告了所有方法的计算成本和验证准确率。研究针对不断增加的细化层级数进行**。

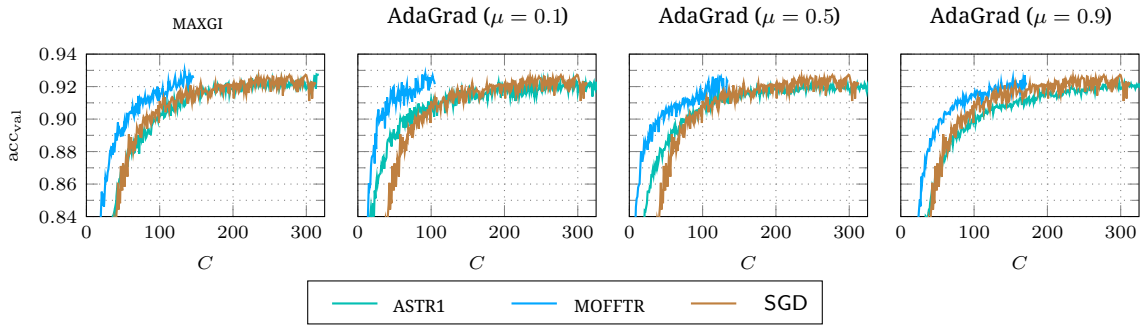

图4：SVNH 示例：训练具有 17 个残差块（三个细化层级）的 ResNet 时，验证准确率随总计算成本变化的情况。从 10 次独立运行中选取验证准确率最高的一次。

这些结果表明，增加网络深度能够增强其表征能力，从而证明更高的计算成本是合理的。我们还注意到，MOFFTR方法的所有变体在达到相近准确率的同时，所需的计算成本均低于单层求解策略。所获得的计算收益从 1.7 倍到 3.4 倍不等。无论细化层级如何，都能观察到这一点，这表明 MOFFTR 算法具有明显的潜力，可加速大规模网络的训练，例如实际应用中使用的网络。我们还注意到，在 MOFFTR 框架的所有变体中，选择 MAXGI 权重所获得的验证准确率高于 AdaGrad 类变体的验证准确率

**MOFFTR 算法使用的层级数量随细化层级线性增加。

方法		层级（残差块）					
		2 (9)		3 (17)		4 (33)	
		C	准确率$_{\text{val}}$	C	准确率$_{\text{val}}$	C	准确率$_{\text{val}}$
SGD		280	92.4%	313	92.7%	336	92.9%
ASTR1	AdaGrad($\mu = 0.1$)	325	92.4%	361	92.6%	341	92.8%
	AdaGrad($\mu = 0.5$)	334	92.3%	333	92.4%	328	92.7%
	AdaGrad($\mu = 0.9$)	351	92.1%	358	92.4%	369	92.6%
	MAXGI	313	92.3%	317	92.7%	325	92.8%
MOFFTR	AdaGrad($\mu = 0.1$)	134	92.3%	105	92.7%	139	92.9%
	AdaGrad($\mu = 0.5$)	148	92.2%	135	92.6%	144	92.8%
	AdaGrad($\mu = 0.9$)	169	92.1%	171	92.6%	151	92.8%
	MAXGI	158	92.5%	146	92.9%	143	93.1%

表3：SVNH 示例：训练深度逐渐增加的 ResNet（细化层级）所需的总计算成本 C 和验证准确率 acc_{val}。报告了 10 次独立运行中验证准确率最高的结果。

。最后我们观察到，在所有 AdaGrad 类变体中，采用 $\mu = 0.1$的配置取得了最高的验证准确率，同时计算成本最低。对应的单层 ASTR1 方法也呈现出类似结果，这表明在实践中，探索传统 $\mu = 0.5$ 之外的 μ 取值可能会有所裨益。

6 结论

我们提出了一类用于无约束最小化的多层算法，这些算法无需计算目标函数的值。我们还表明，在深度神经网络训练中，由子采样引起的噪声存在时，这类算法的性能具有竞争力。我们分析了两个有趣子类的收敛理论，并证明其达到了当前最先进的水平。我们的实验还表明，尽管目前尚未纳入理论分析，但加入动量后，多层方法的优势仍然得以保留。作者充分认识到，只有持续开展实验，才能揭示本研究方案的真正实用价值；不过，他们指出，初步数值测试结果令人鼓舞。

致谢

A. Kopani˘c´akov´a 衷心感谢瑞士国家科学基金会通过项目“DeepONets 的多层训练 — 多尺度与多物理应用”（资助编号：206745）提供的支持，并感谢先进科学计算平台（PASC）通过 EXATRAIN 项目提供的部分支持。Ph. L. Toint 感谢 ANITI 一如既往且友好的部分支持。

参考文献

[1] N. M. Alexandrov 和 R. L. Lewis。一阶模型管理在工程优化中的概述。Optimization and Engineering，2:413–430，2001。[2] M. F. Baumgardner、L. L. Biehl 和 D. A. Landgrebe。220 波段 AVIRIS 高光谱图像数据集：1992 年 6 月 12 日 Indian Pine TeSite 3。普渡大学研究资料库，10(7):991，2015。[3] A. Beck 和 N. Hallak。涉及群稀疏项的优化问题。Mathematical Programming，2018。在线。[4] A. Borzi 和 K. Kunisch。椭圆最优控制问题多重网格求解的全局化策略。Optimization Methods and Software，21(3):445–459，2006。[5] A. Bouaricha 和 R. B. Schnabel。大型稀疏非线性方程组的张量方法。Mathematical Programming，82(3):377–412，1998。

[6] W. L. Briggs、V. E. Henson 和 S. F. McCormick。多重网格教程。SIAM，美国费城，第 2 版，2000。[7] C. Cartis、N. I. M. Gould 和 Ph. L. Toint。非凸优化算法的评估复杂度。MOS-SIAM 优化系列第 30 卷。SIAM，美国费城，2022 年 6 月。[8] B. Chang、L. Meng、E. Haber、F. Tung 和 D. Begert。从动力系统视角看多层残差网络。arXiv:1710.10348，2017。[9] I. Chatzigeorgiou。朗伯函数的界及其在用户协作中断分析中的应用。IEEE Communications Letters，17(8):1505–1508，2013。[10] A. R. Conn、N. I. M. Gould 和 Ph. L. Toint。LANCELOT：用于大规模非线性优化的 Fortran 软件包（A 版）。Springer 计算数学系列第 17 卷。Springer Verlag，海德堡、柏林、纽约，1992。[11] A. R. Conn、N. I. M. Gould 和 Ph. L. Toint。信赖域方法。MOS-SIAM 优化系列第 1 卷。SIAM，美国费城，2000。[12] R. M. Corless、G. H. Gonnet、D. E. Hare、D. J. Jeffrey 和 D. E. Knuth。关于 Lambert W 函数。Advances in Computational Mathematics，5:329—359，1996。[13] E. C. Cyr、S. G¨unther 和 J. B. Schroder。层并行深度神经网络训练的多层初始化。International Journal of Computing and Visualization in Science and Engineering，2021。[14] J. Duchi、E. Hazan 和 Y. Singer。用于在线学习和随机优化的自适应次梯度方法。Journal of Machine Learning Research，12，2011 年 7 月。[15] F. Chollet 等。Keras，2015。 https://keras.io.[16] M. Fisher。变分资料同化的最小化算法。载于《大气建模数值方法的最新进展》，第 364–385 页，英国雷丁，1998。欧洲中期天气预报中心。[17] K. Fujisawa、M. Kojima 和 K. Nakata。在半定规划的原始-对偶内点方法中利用稀疏性。Mathematical Programming，B 辑，79(1–3):235–253，1997。[18] L. Gaedke-Merzh¨auser、A. Kopaniˇc´akov´a 和 R. Krause。深度残差网络的多层最小化。载于 ESAIM。Proceedings and Surveys。71:131-144，2021。[19] D. M. Gay。自动发现并利用非线性规划问题中的部分可分结构。技术报告，贝尔实验室，默里山，新泽西州，美国，1996。[20] E. Gelman 和 J. Mandel。关于优化问题的多层迭代方法。Mathematical Programming，48(1):1–17，1990。[21] D. Goldfarb 和 S. Wang。适用于一元、可分解及部分可分优化的部分更新牛顿方法。SIAM Journal on Optimization，3(2):383–397，1993。[22] S. Gratton、S. Jerad 和 Ph. L. Toint。无目标函数优化正则化算法的收敛性质，包括一个 $\mathcal{O}(\epsilon^{-3/2})$ 复杂度界。arXiv:2203.09947，2022。[23] S. Gratton、S. Jerad 和 Ph. L. Toint。一阶无目标函数优化算法及其复杂度。arXiv:2203.01757，2022。[24] S. Gratton、S. Jerad 和 Ph. L. Toint。一类一阶类 Adagrad 算法的参数化复杂度分析。arXiv:2203.01647，2022。[25] S. Gratton、M. Mouffe、A. Sartenaer、Ph. L. Toint 和 D. Tomanos。递归信赖域方法用于多层非线性有界约束优化的数值经验。Optimization Methods and Software，25(3):359 – 386，2010。[26] S. Gratton、A. Sartenaer 和 Ph. L. Toint。多尺度非线性优化的递归信赖域方法。SIAM Journal on Optimization，19(1):414–444，2008。[27] S. Gratton 和 Ph. L. Toint。近似不变子空间与拟牛顿优化方法。Optimization Methods and Software，25(4):507–529，2010。[28] A. Griewank 和 Ph. L. Toint。部分可分函数的无约束优化。载于 M. J. D. Powell 编，《Nonlinear Optimization 1981》，第 301–312 页，伦敦，1982。Academic Press。[29] E. Haber 和 L. Ruthotto。深度神经网络的稳定架构。Inverse Problems，34 014004，2017 1( ): 。[30] E. Haber、L. Ruthotto、E. Holtham 和 S.-H. Jun。跨尺度学习——卷积神经网络的多尺度方法。载于第三十二届 AAAI 人工智能会议，2018。[31] K. He、X. Zhang、S. Ren 和 J. Sun。深度残差网络中的恒等映射。载于欧洲计算机视觉会议，第 630–645 页，海德堡、柏林、纽约，2016。Springer Verlag。[32] S. Ioffe 和 Ch. Szegedy。批归一化：通过减少内部协变量偏移加速深度网络训练。arXiv:1502.03167

，2015.

[33] D. Kingma 和 J. Ba。Adam：一种随机优化方法。载于《国际学习表征会议》（ICLR）论文集，2015年。[34] A. Kopani˘c´akov´a。关于在多层最小化方法中使用混合粗层模型。arXiv:2211.15078，2022年。[35] A. Kopani˘c´akov´a 和 R. Krause。深度残差网络的全局收敛多层训练。SIAM 科学计算杂志，0:S254–S280，2022年。[36] R. Kornhuber。针对某些非光滑优化问题的自适应单调多重网格方法。载于 R. Glowinski、J. P´eriaux、Z. Shi 和 O. Widlund 编，《科学与工程中的区域分解方法》，第177–191页。约翰·威立父子出版公司，奇切斯特，英格兰，1997年。[37] J. B. Lasserre。稀疏多项式优化中的收敛半定松弛。技术报告，LAAS-CNRS，法国图卢兹 Colonel Roche 大道7号，邮编31077，2005年11月。[38] M. Lewis 和 S. G. Nash。VLSICAD 多尺度优化方法的实践问题。载于 Jason Cong 和 Joseph R. Shinnerl 编，《多尺度优化与 VLSI/CAD》，第265–291页，多德雷赫特，荷兰，2002年。Kluwer 学术出版社。[39] M. Lewis 和 S. G. Nash。由微分方程支配的系统的多重网格优化模型问题。SIAM 科学计算杂志，26(6):1811–1837，2005年。[40] J. Mare˘cek、P. Richt´arik 和 M. Tak´a˘c。用于最小化部分可分函数的分布式块坐标下降。技术报告，爱丁堡大学数学与统计学系，爱丁堡，苏格兰，2014年。[41] S. G. Nash。离散化优化问题的多重网格方法。Optimization Methods and Software，14:99–116，2000年。[42] Y. Netzer、T. Wang、A. Coates、A. Bissacco、B. Wu 和 A.Y. Ng。利用无监督特征学习在自然图像中识别数字，2011年。[43] Z. M. Prince 和 J. C. Ragusa。将适当广义分解应用于中子扩散的参数不确定性量化。《国际工程数值方法杂志》，119 921，2019 9 899( ): – 。[44] A. F. Queiruga、N. B. Erichson、D. Taylor 和 M. W. Mahoney。连续深度神经网络。arXiv:2008.02389，2020年。[45] S. Reddi、S. Kale 和 S. Kumar。关于 Adam 及其后续方法的收敛性。载于《国际学习表征会议》（ICLR）论文集，2018年。[46] H. Robbins 和 S. Monro。一种随机逼近方法。《数学统计年刊》，22(3):400—407，1951年。[47] T. Tieleman 和 G. Hinton。第6.5讲——RMSPROP。COURSERA：机器学习中的神经网络，2012年。[48] Ph. L. Toint。关于利用稀疏性的拟牛顿方法的说明。Mathematical Programming，21(2):172– 181，1981年。[49] Ph. L. Toint。迈向高效的利用稀疏性的牛顿最小化方法。载于 I. S. Duff 编，《稀疏矩阵及其应用》，第57–88页，伦敦，1981年。Academic Press。[50] E. Weinan。关于通过动力系统实现机器学习的提议。《数学与统计学通讯》，5(1):1–11，2017年。[51] Y. Yuan。信赖域算法的最新进展。Mathematical Programming，A辑，151 281，2015 1 249( ): – 。[52] D. Zhou、J. Chen、Y. Tang、Z. Yang、Y. Cao 和 Q. Gu。关于非凸优化中自适应梯度方法的收敛性。arXiv:2080.05671，2020年。

定理 3.8 的证明

我们首先回顾一个有用的技术结果。

引理 A。1 [23, 引理 3.5] 考虑任意的 $i \in \{1, \ldots, n\}$，并假设存在一个 j_ς，使得

$$\min\left[\frac{g_{i,j}^2}{\varsigma_i}, \frac{g_{i,j}^2}{v_{i,j}}\right] \leq \varsigma_i \quad \text{for} \quad j \geq j_\varsigma. \tag{68}$$

Then

$$\min\left[\frac{g_{i,j}^2}{\varsigma_i}, \frac{g_{i,j}^2}{v_{i,j}}\right] \geq \frac{g_{i,j}^2}{2\varsigma_i} \quad \text{for} \quad j \geq j_\varsigma. \tag{69}$$

定理 3.8 的证明 由 (43) 和 (6) 可得，对于 $i \geq i_\vartheta$,

$$f(x_{r,i_\vartheta}) - f(x_{r,i+1}) \geq \frac{1}{2} \sum_{k=i_\vartheta+1}^{i} \sum_{j=1}^{n} \frac{g_{r,k,j}^2}{w_{r,k,j}} \left[\beta_{1,r} - \frac{\beta_{2,r} + \frac{1}{2}\alpha^2 L}{w_{\ell,k,j}} \right].$$

此外，(45) 和 (52) 也确保，由于 $i \geq i_\vartheta$,

$$\beta_{1,r} - \frac{\beta_{2,r} + \frac{1}{2}\alpha^2 L}{w_{\ell,k,j}} \geq \vartheta$$

对所有 $j \in \{1, \ldots, n\}$ 均成立，因此

$$\begin{aligned}
f(x_{r,i_\vartheta}+1) - f(x_{r,i+1}) &\geq \frac{\vartheta}{2} \sum_{k=i_\vartheta+1}^{i} \sum_{j=1}^{n} \frac{g_{r,k,j}^2}{w_{r,k,j}} \\
&\geq \frac{\vartheta}{2} \sum_{k=i_\vartheta+1}^{i} \sum_{j=1}^{n} \frac{g_{r,k,j}^2}{\max[\varsigma_j, v_{r,k}](k+1)^\mu} \\
&\geq \frac{\vartheta}{2(i+1)^\mu} \sum_{k=i_\vartheta+1}^{i} \sum_{j=1}^{n} \min\left[\frac{g_{r,k,j}^2}{\varsigma_j}, \frac{g_{r,k,j}^2}{v_{r,k}} \right] \\
&\geq \frac{\vartheta(i-i_\vartheta)}{2(i+1)^\mu} \min_{k \in \{i_\vartheta, \ldots, i\}} \left(\sum_{j=1}^{n} \min\left[\frac{g_{r,k,j}^2}{\varsigma_j}, \frac{g_{r,k,j}^2}{v_{r,k}} \right] \right).
\end{aligned}$$

但是，利用引理 3.7 中的 (50),

$$f(x_0) - f(x_{i_\vartheta+1}) = f(x_{r,0}) - f(x_{i_\vartheta+1}) \geq -n(\beta_{2,r} + \tfrac{1}{2}\alpha^2 L) \sum_{k=0}^{i_\vartheta} a(k)^2. \tag{70}$$

将这两个不等式与 AS.3 结合，可得

$$\begin{aligned}
\min_{k \in \{i_\vartheta, \ldots, i\}} \left(\sum_{j=1}^{n} \min\left[\frac{g_{r,k,j}^2}{\varsigma_j}, \frac{g_{r,k,j}^2}{v_{r,k}} \right] \right) &\leq \frac{2(i+1)^\mu}{\vartheta(i-i_\vartheta)} \Big(f(x_{i_\vartheta+1}) - f(x_{i+1}) \Big) \\
&= \frac{2(i+1)^\mu}{\vartheta(i-i_\vartheta)} \Big(f(x_0) - f(x_{i+1}) + f(x_{i_\vartheta+1}) - f(x_0) \Big) \\
&\leq \frac{2(i+1)^\mu}{\vartheta(i-i_\vartheta)} \Big(\Gamma_0 + n(\beta_{2,r} + \tfrac{1}{2}\alpha^2 L) \sum_{k=0}^{i_\vartheta} a(k)^2 \Big)
\end{aligned}$$

由此可得，存在一个子序列 $\{i_t\} \subseteq \{i\}_{i_\vartheta}^{\infty}$，使得

$$\sum_{j=1}^{n} \min\left[\frac{g_{r,i_t,j}^2}{\varsigma_j}, \frac{g_{r,i_t,j}^2}{v_{r,i_t}} \right] \leq \frac{2(i_t+1)^\mu}{\vartheta(i_t-i_\vartheta)} \left[\Gamma_0 + n(\beta_{2,r} + \tfrac{1}{2}\alpha^2 L) \sum_{k=0}^{i_\vartheta} a(k)^2 \right]. \tag{71}$$

Now,

$$\frac{(i_t+1)^\mu}{i_t - i_\vartheta} < \frac{2^\mu i_t^\mu}{i_t - i_\vartheta} < \frac{2i_t^\mu}{i_t - i_\vartheta} = \frac{2i_t^\mu i_t}{(i_t - i_\vartheta) i_t} = \frac{i_t}{i_t - i_\vartheta} \cdot \frac{2}{i_t^{1-\mu}} \leq \frac{2(i_\vartheta+1)}{i_t^{1-\mu}}, \tag{72}$$

这里用到了以下事实：$\mu < 1$，且对于 $i_t \geq i_\vartheta + 1$，$\frac{i_t}{i_t - i_\vartheta}$ 是一个递减函数。因此，由（71）可得

$$\sum_{j=1}^{n} \min\left[\frac{g_{r,i_t,j}^2}{\varsigma_j}, \frac{g_{r,i_t,j}^2}{v_{r,i_t}}\right] \leq \frac{4(i_\vartheta+1)}{\vartheta i_t^{1-\mu}}\left[\Gamma_0 + n(\beta_{2,r} + \tfrac{1}{2}\alpha^2 L)\sum_{k=0}^{i_\vartheta} a(k)^2\right].$$

现在定义

$$i_\varsigma \stackrel{\text{def}}{=} \left(\frac{4(i_\vartheta+1)\left[\Gamma_0 + \frac{1}{2}n(\beta_{2,r} + \frac{1}{2}\alpha^2 L)\sum_{k=0}^{i_\vartheta} a(k)^2\right]}{\vartheta \varsigma_{\min}}\right)^{\frac{1}{1-\mu}},$$

我们看到，对于所有 $i_t \geq i_\varsigma$，

$$\min\left[\frac{g_{r,i_t,j}^2}{\varsigma_j}, \frac{g_{r,i_t,j}^2}{v_{r,i_t}}\right] \leq \varsigma_{\min}.$$

随后应用引理 A.1，由 (71) 可推得，对于所有 $i_t \geq i_\varsigma$，

$$\sum_{j=1}^{n} \frac{g_{r,i_t,j}^2}{2\varsigma_j} \leq \frac{2(i_t+1)^\mu}{\vartheta(i_t - i_\vartheta)}\left[\Gamma_0 + n(\beta_{2,r} + \tfrac{1}{2}\alpha^2 L)\sum_{k=0}^{i_\vartheta} a(k)^2\right]$$

因此，由于 $\varsigma_j \leq 1$，有

$$\|g_{r,i_t}\|^2 \leq \frac{(i_t+1)^\mu}{i_t - i_\vartheta}\left(\frac{2}{\vartheta}\right)\left[\Gamma_0 + n(\beta_{2,r} + \tfrac{1}{2}\alpha^2 L)\sum_{k=0}^{i_\vartheta} a(k)^2\right]$$

对于所有 $i_t \geq i_\varsigma$。由此得到 (51)，其中最后一个不等式由 (72) 和 (6) 得出。 □

B 定理 3.9 的证明

我们再次先回顾一个有用的技术引理。

引理 B。1 [23, 引理 3.1] 设 $\{a_k\}_{k\geq 0}$ 为一个非负序列，$\alpha > 0$、$\xi > 0$，并对每个 $k \geq 0$ 定义 $b_k = \sum_{t=0}^{k} a_t$。则

$$\sum_{k=0}^{i} \frac{a_k}{(\xi + b_k)^\alpha} \leq \begin{cases} \frac{1}{(1-\alpha)}((\xi+b_i)^{1-\alpha} - \xi^{1-\alpha}) & \text{if } \alpha \neq 1, \\ \log\left(\frac{\xi+b_i}{\xi}\right) & \text{if } \alpha = 1. \end{cases} \tag{73}$$

定理 3.9 的证明 我们首先注意到，(54) 意味着

$$\varsigma^\mu \leq \max_{j\in\{1,\ldots,n\}} w_{r,i,j} \leq \left(\varsigma + \sum_{k=0}^{i} \|g_{r,k}\|^2\right)^\mu \tag{74}$$

对所有 $i \geq 0$ 均成立。我们还根据引理 3.6 中的 (6) 和 (43) 推得，利用 $\ell = r$ 可知，对于 $i \geq 0$，

$$f(x_{i+1}) \leq f(x_0) - \beta_{1,r}\sum_{k=0}^{i} \frac{\|g_{r,k}\|^2}{\max\limits_{j\in\{1,\ldots,n\}} w_{r,k,j}} + (\beta_{2,r} + \tfrac{1}{2}L)\sum_{j=1}^{n}\sum_{k=0}^{i} \frac{g_{r,k,j}^2}{w_{r,k,j}^2}. \tag{75}$$

情形 (i)。首先假设 $\mu \in (0, \frac{1}{2})$。于是，对每个 $j \in \{1, \ldots, n\}$，我们将引理 B.1 中的 (73) 应用于 $a_k = g_{r,k,j}^2$, $\xi = \varsigma$ 和 $\alpha = 2\mu < 1$，并由 (13) 和 (54) 得到，

$$\sum_{k=0}^{i} \frac{g_{r,k,j}^2}{w_{r,k,j}^2} \leq \frac{1}{1-2\mu} \left[\left(\varsigma + \sum_{k=0}^{i} g_{r,k,j}^2 \right)^{1-2\mu} - \varsigma^{1-2\mu} \right] \leq \frac{1}{1-2\mu} \left(\varsigma + \sum_{k=0}^{i} g_{r,k,j}^2 \right)^{1-2\mu}. \tag{76}$$

现在将此界代入 (75)，并利用 AS.3 可得

$$\beta_{1,r} \sum_{k=0}^{i} \frac{\|g_{r,k}\|^2}{\max\limits_{j \in \{1,\ldots,n\}} w_{r,k,j}} \leq \Gamma_0 + \frac{n(\beta_{2,r} + \frac{1}{2}L)}{1-2\mu} \left(\varsigma + \sum_{k=0}^{i} \|g_{r,k}\|^2 \right)^{1-2\mu}. \tag{77}$$

现在假设

$$\sum_{k=0}^{i} \|g_{r,k}\|^2 \geq \max \left\{ \varsigma, \frac{1}{2} \left[\frac{(1-2\mu)\Gamma_0}{n(\beta_{2,r} + \frac{1}{2}L)} \right]^{\frac{1}{1-2\mu}} \right\}, \tag{78}$$

这意味着

$$\varsigma + \sum_{k=0}^{i} \|g_{r,k}\|^2 \leq 2 \sum_{k=0}^{i} \|g_{r,k}\|^2 \text{ and } \Gamma_0 \leq \frac{n(\beta_{2,r} + \frac{1}{2}L)}{1-2\mu} \left(2 \sum_{k=0}^{i} \|g_{r,k}\|^2 \right)^{1-2\mu}.$$

于是，利用 (77) 和 (74)，

$$\frac{\beta_{1,r}}{2^\mu \left[\sum_{k=0}^{i} \|g_{r,k}\|^2 \right]^\mu} \sum_{k=0}^{i} \|g_{r,k}\|^2 \leq \frac{2^{2(1-\mu)} n(\beta_{2,r} + \frac{1}{2}L)}{1-2\mu} \left(\sum_{k=0}^{i} \|g_{r,k}\|^2 \right)^{1-2\mu}.$$

对 $\sum_{k=0}^{i} \|g_{r,k}\|^2$ 求解此不等式，并利用 $2^{2(1-\mu)} < 4$ 这一事实，可得

$$\sum_{k=0}^{i} \|g_{r,k}\|^2 < \left[\frac{4n(\beta_{2,r} + \frac{1}{2}L)}{\beta_{1,r}(1-2\mu)} \right]^{\frac{1}{\mu}}$$

因此

$$\underset{k \in \{0,\ldots,i\}}{\text{average}} \|g_{r,k}\|^2 < \left[\frac{4n(\beta_{2,r} + \frac{1}{2}L)}{\beta_{1,r}(1-2\mu)} \right]^{\frac{1}{\mu}} \cdot \frac{1}{i+1}. \tag{79}$$

或者，如果（78）不成立，则

$$\underset{k \in \{0,\ldots,i\}}{\text{average}} \|g_{r,k}\|^2 < \max \left\{ \varsigma, \frac{1}{2} \left[\frac{(1-2\mu)\Gamma_0}{n(\beta_{2,r} + \frac{1}{2}L)} \right]^{\frac{1}{1-2\mu}} \right\} \cdot \frac{1}{i+1}. \tag{80}$$

结合（79）、（80）和（6），可得对于 $0 < \mu < \frac{1}{2}$，式（57）成立。_

情形（ii）。现在考虑 $\mu = \frac{1}{2}$ 的情形。对于每个 $j \in \{1, \ldots, n\}$，我们在引理 B.1 中应用（73），其中 $a_k = g_{r,\,k,\,j}^2$、$\xi = \varsigma$ 和 $\alpha = 2\mu = 1$，从而得到

$$\sum_{j=1}^{n} \sum_{k=0}^{i} \frac{g_{r,k,j}^2}{w_{r,k,j}^2} \leq \sum_{j=1}^{n} \log \left(\frac{1}{\varsigma} \left(\varsigma + \sum_{k=0}^{i} g_{r,k,j}^2 \right) \right) \leq n \log \left(1 + \frac{1}{\varsigma} \sum_{k=0}^{i} \|g_{r,k}\|^2 \right)$$

将此界代入式（75）即可得到

$$\beta_{1,r} \sum_{k=0}^{i} \frac{\|g_{r,k}\|^2}{\max\limits_{j \in \{1,\ldots,n\}} w_{r,k,j}} \leq \Gamma_0 + \tfrac{1}{2} n(\beta_{2,r} + \tfrac{1}{2}L) \log \left(1 + \frac{1}{\varsigma} \sum_{k=0}^{i} \|g_{r,k}\|^2 \right).$$

现在假设

$$\sum_{k=0}^{i} \|g_{r,k}\|^2 \geq \max\left[\varsigma, e^{\frac{2\Gamma_0}{n(\beta_{2,r}+\frac{1}{2}L)}}\right], \tag{81}$$

这意味着

$$\varsigma + \sum_{k=0}^{i} \|g_{r,k}\|^2 \leq 2\sum_{k=0}^{i} \|g_{r,k}\|^2 \text{ and } \Gamma_0 \leq \tfrac{1}{2}n(\beta_{r,2}+\tfrac{1}{2}L)\log\left(\frac{2}{\varsigma}\sum_{k=0}^{i}\|g_{r,k}\|^2\right).$$

将 (74) 用于 $\mu = \frac{1}{2}$，于是得到

$$\frac{\beta_{1,r}}{\sqrt{2}\sqrt{\sum_{k=0}^{i}\|g_{r,k}\|^2}}\sum_{k=0}^{i}\|g_{r,k}\|^2 \leq n(\beta_{r,k}+\tfrac{1}{2}L)$$

因此

$$\beta_{r,1}\sqrt{\sum_{k=0}^{i}\|g_{r,k}\|^2} \leq 2\sqrt{2}\max[3\beta_{r,1}, n(\beta_{r,2}+\tfrac{1}{2}L)]\log\left(\sqrt{\frac{2}{\varsigma}\sum_{k=0}^{i}\|g_{r,k}\|^2}\right). \tag{82}$$

现在定义

$$\gamma_1 \stackrel{\text{def}}{=} \beta_{r,1}\sqrt{\frac{\varsigma}{2}}, \quad \gamma_2 \stackrel{\text{def}}{=} 2\sqrt{2}\max[\tfrac{3}{2}\beta_{r,1}, n(\beta_{r,2}+\tfrac{1}{2}L)] \quad \text{and} \quad u \stackrel{\text{def}}{=} \sqrt{\frac{2}{\varsigma}\sum_{k=0}^{i}\|g_{r,k}\|^2} \tag{83}$$

并注意到，根据构造，$\gamma_2 > 3\gamma_1$，因为 $\varsigma \leq 1$。于是，不等式 (82) 可改写为

$$\gamma_1 u \leq \gamma_2 \log(u). \tag{84}$$

记 $\psi(u) \text{ 定义为 } = \gamma_1 u - \gamma_2 \log(u)$。由于 $\gamma_2 > 3\gamma_1$，方程 $\psi(u) = 0$ 有两个根 $u_1 \leq u_2$，且当 $u \in [u_1, u_2]$ 时，(84) 成立。根据 u_2 的定义可得

$$\log(u_2) - \frac{\gamma_1}{\gamma_2}u_2 = 0.$$

令 $z = -\frac{\gamma_1}{\gamma_2}u_2$，可得

$$ze^z = -\frac{\gamma_1}{\gamma_2}.$$

因此 $z = W_{-1}(-\frac{\gamma_1}{\gamma_2}) < 0$，其中 W_{-1} 是朗伯函数在 $[-\frac{1}{e}, 0)$ 上定义的第二分支。由于 $-\frac{\gamma_1}{\gamma_2} \geq -\frac{1}{3}$，$z$ 有定义，因此

$$u_2 = -\frac{\gamma_2}{\gamma_1}z = -\frac{\gamma_2}{\gamma_1}W_{-1}\left(-\frac{\gamma_1}{\gamma_2}\right) > 0 = -\psi\, W_{-1}\left(-\frac{1}{\psi}\right) > 0,$$

其中 $\psi = \frac{\gamma_2}{\gamma_1}$ 由 (59) 给出。因此，由 (84) 和 (83) 可得

$$\sum_{k=0}^{i}\|g_{r,k}\|^2 = \frac{\varsigma}{2}u_2^2 = \frac{\varsigma\psi^2}{2}\left|W_{-1}\left(-\frac{1}{\psi}\right)\right|^2$$

and

$$\underset{k\in\{0,\ldots,i\}}{\text{average}}\ \|g_{r,k}\|^2 \leq \frac{\varsigma\psi^2}{2}\left|W_{-1}\left(-\frac{1}{\psi}\right)\right|^2 \cdot \frac{1}{i+1}. \tag{85}$$

如果 (81) 不成立，则有

$$\underset{k\in\{0,\ldots,i\}}{\text{average}}\ \|g_{r,k}\|^2 < \max\left[\varsigma, e^{\frac{2\Gamma_0}{n(\beta_{2,r}+\frac{1}{2}L)}}\right] \cdot \frac{1}{i+1}. \tag{86}$$

结合 (85)、(86) 和 (6)，得到 $\mu = \frac{1}{2}$ 对应的 (57)。

情形 (iii)。最后，假设 $\frac{1}{2} < \mu < 1$。我们再次对每个满足 $a_k = g_{r,k,j}^2$, $\xi = \varsigma$ 和 $\alpha = 2\mu > 1$ 的 $j \in \{1, \ldots, n\}$ 应用引理 B.1 中的 (73)，得到对于 $j \in \{1, \ldots, n\}$,

$$\sum_{k=0}^{i} \frac{g_{r,k,j}^2}{w_{r,k,j}^2} \leq \frac{1}{1-2\mu} \left(\left(\varsigma + \sum_{k=0}^{i} g_{r,k,j}^2 \right)^{1-2\mu} - \varsigma^{1-2\mu} \right) \leq \frac{\varsigma^{1-2\mu}}{2\mu - 1}. \tag{87}$$

将界（87）代入（75），并利用（74）和 AS.3，可得

$$\beta_{r,1} \sum_{k=0}^{i} \frac{1}{(\varsigma + \sum_{t=0}^{i} \|g_{r,t}\|^2)^{\mu}} \|g_{r,k}\|^2 \leq \Gamma_0 + \frac{n(\beta_{2,r} + \frac{1}{2}L)\varsigma^{1-2\mu}}{2\mu - 1}.$$

现在假设

$$\sum_{k=0}^{i} \|g_{r,k}\|^2 \geq \varsigma, \tag{88}$$

then

$$\underset{k \in \{0,\ldots,i\}}{\text{average}} \|g_{r,k}\|^2 \leq \left[\frac{2^{\mu}}{\beta_{r,1}} \left(\Gamma_0 + \frac{n(\beta_{2,r} + \frac{1}{2}L)\varsigma^{1-2\mu}}{2\mu - 1} \right) \right]^{\frac{1}{1-\mu}} \cdot \frac{1}{i+1}. \tag{89}$$

若（88）不成立，则可推出

$$\underset{k \in \{0,\ldots,i\}}{\text{average}} \|g_{r,k}\|^2 \leq \frac{\varsigma}{(i+1)}. \tag{90}$$

因此，（89）、（90）和（6）最终推出对 $\frac{1}{2} < \mu < 1$ 而言（57）成立。 □