自适应网络上的压缩回归

Marco Carpentiero、Vincenzo Matta 和 Ali H. Sayed

arXiv:2304.03638v1 [cs.LG] 7 Apr 2023

摘要

本文推导了分布式智能体网络在自适应地且存在通信约束的情况下求解回归问题所能达到的性能。智能体采用最近提出的 ACTC（自适应-压缩-再组合）扩散策略，其中，相邻智能体在本地交换的信号通过随机差分压缩 算子进行编码。我们详细刻画了均方估计误差，并表明该误差由两部分组成：一部分与智能体在没有通信约束时能够达到的误差有关，另一部分则源于压缩。分析揭示了压缩损失与分布式回归问题的基本属性之间的定量关系，尤其是梯度噪声导致的随机逼近误差以及网络拓扑（通过佩龙特征向量体现）。我们表明，了解这些关系对于在智能体之间优化分配通信资源至关重要，同时还需考虑智能体各自的属性，例如数据质量或其在网络拓扑中的中心性。我们设计了一种优化分配策略，其中优化所需的参数可由智能体在线 学习。示例表明，与盲目（即均匀）资源分配相比，利用所提供的均方误差公式优化资源分配能够实现显著的性能提升。

索引词

分布式优化、自适应、学习、扩散策略、随机量化器。

I. 引言

受数据可用性的指数级增长以及网络化设备系统（例如物联网、移动边缘计算、车联网）的成功启发，完全去中心化策略和联邦策略 [2]–[8] 构成了下一代学习算法的骨干。在本文中，我们聚焦于自适应设置下的完全去中心化场景 [2]–[4]。我们考虑一个在线数据分析问题

MarcoCarpentiero和 VincenzoMatta 任职于萨勒诺大学信息与电气工程及应用数学系（DIEM），地址为 via Giovanni Paolo II, I-84084,Fisciano(SA), Italy，同时任职于意大利国家电信大学间联合联盟（CNIT）（电子邮箱：{ mcarpentiero, vmatta}@unisa.it）。

Ali H. Sayed 任职于瑞士洛桑联邦理工学院（EPFL）工程学院，地址：瑞士洛桑 CH-1015（电子邮箱: ali.sayed@epfl.ch）。

该工作的简短会议版本将于 ICASSP 2023 [1] 发表。

其中，多智能体网络通过局部交互，协作学习未知模型，并观测流式数据。此类问题可通过结合流行的学习算法（例如随机梯度算法，采用恒定步长以便及时跟踪数据漂移）以及分布式协作策略（如一致性 [9]–[12] 或扩散 [2]–[4]）成功解决。

去中心化学习面临的最大挑战之一，是智能体之间往返传输信息所造成的通信瓶颈。必须引入合适的压缩策略，例如量化 [13], ，以应对不可避免的通信限制。数据压缩是一门研究成熟的学科，多年来提供了多种有用工具，并已成功应用于许多推断问题，甚至包括分布式场景 [14]–[19]。然而，本文研究的在线优化问题带来了额外挑战。首先，数据分布取决于一个未知参数（该参数可能随时间变化），而此参数正是优化过程的目标，因此无法采用标准的量化器设计方法。其次，由于梯度类算法具有迭代特性，量化造成的失真可能会随时间累积，从而损害优化算法的收敛性。近期研究表明，可通过随机和/或差分压缩来解决这些问题 [20]–[23]。

随机压缩算子可追溯至关于量化器 [24] 和抖动 [25]的概率分析开创性研究。通过适当的随机机制生成编码输出，这些压缩算子呈现出有利于推断的特性，例如无偏性或均方误差有界性，而且这些特性普遍成立，即不受具体数据分布的影响。得益于这些性质，随机压缩方法近期已应用于分布式优化领域，其应用始于广受欢迎的随机量化器 [20] 和随机稀疏化器 [21], [22]。

差分压缩利用通信系统传统中的技术，例如差分脉冲编码调制（差分脉冲编码调制）和增量调制 [13], 来通过仅考虑连续样本之间的创新 来缩小编码器输入的实际范围。这种方法利用连续数据样本之间存在的相关性，避免在冗余信息的表示中浪费资源。

近年来，一些研究考察了存在随机压缩 [26]–[28], 有时还与差分量化 [29]–[31]相结合的梯度下降和随机梯度下降算法。然而，在这些研究中，我们发现了以下设计选择： i) 采用递减步长的非自适应系统； ii) 使用对称组合策略来建模智能体之间的网络交互； iii)严格要求局部强凸性，即每个智能体的代价函数都必须是强凸的；以及iv)采用一致性技术来合并智能体在网络中共享的信息。

在 [32], [33], 中，我们通过消除上述限制，扩展了通信约束下分布式学习的分析，并转而考虑：(a) 利用具有恒定步长的随机梯度下降的自适应系统；与采用递减步长的实现不同，这些系统能够

及时跟踪数据漂移并持续学习；(b) 用于表示各种现实网络场景的非对称且左随机的组合策略；(c) 仅在全局层面具有强凸性，从而允许局部层面存在凸代价函数和非凸代价函数；以及(d) 扩散（而非一致性）策略，这些策略已被证明能够带来更优的性能和更宽的稳定范围 [2]。我们引入了 ACTC（自适应-压缩-然后组合）扩散策略，该策略借鉴了广受欢迎的ATC（自适应-然后组合）扩散策略的分布式实现 [2], [34], [35] ，并对其进行了扩展，使其能够支持受限通信。借助一类合适的随机化差分压缩 算子，ACTC策略能够以显著节省传输资源（例如量化比特）的方式，收敛到期望解的一个小邻域。

在 [32], [33] 中，我们通过瞬态和均方稳定性分析，详细刻画了 ACTC 策略的学习行为及其收敛保证。我们表明，尽管存在数据压缩，通过调整 *ACTC* 算法的步长和稳定性参数，始终可以实现均方稳定性。我们得出结论，自适应网络特有的学习行为得以保留，呈现出两个瞬态阶段和一个稳态阶段 [2], [34], [35]。

在稳定性分析和瞬态分析之后，完整刻画一种学习算法的第三个关键部分是稳态性能。在这项工作中，我们填补了这一研究空白，重点关注所谓的均方误差（*MSE*）网络 [2]这一重要情形。这些网络由空间上分散部署的智能体组成，它们以在线方式协同求解分布式线性回归问题。这一设定代表了信号处理和统计学习中常用的 LMS（最小均方）算法的分布式对应形式 [37]。关于通信约束下均方误差网络的早期研究之一是 [36],，其中压缩误差被建模为噪声源。相比之下，在这项工作中，我们从压缩算子的设计出发，并考虑相关压缩误差的精确形式。这一扩展给分析带来了显著的额外挑战。

本工作的主要贡献如下。

- 我们给出了 MSE 网络中 ACTC 策略下每个独立智能体均方误差的上界。分析表明，该上界由两部分构成：一部分对应于不存在通信约束时所达到的误差，另一部分是压缩损失。

- 利用所得公式，我们建立了压缩损失与分布式回归问题基本属性之间的定量关系，即由梯度噪声引起的随机逼近误差，以及网络拓扑（通过佩龙特征向量体现）。

- 我们证明，了解这些关系对于在各智能体之间高效分配通信资源至关重要，同时还需考虑它们各自的属性，例如数据质量或在网络中的中心性。我们提出了一种优化分配策略，由一种在线算法提供支持，该算法实时学习优化所需的参数。示例说明

与盲目（即均匀）资源分配相比，性能有显著提升。

记号。粗体字母表示随机变量，而正体字母表示其实现值。大写字母表示矩阵，小写字母表示向量和标量。在某些情况下会违背这一约定，例如用 N 表示网络中智能体的数量。所有向量均约定为列向量。符号 $\mathbb{1}_L$ 表示一个所有元素均等于 1 的 $L \times 1$ 向量，而符号 0 也用于表示全零向量或矩阵。$L \times L$ 单位矩阵记为 I_L。运算 $\mathrm{col}\{x_1, x_2, \ldots, x_L\}$ 通过将向量 x_1、x_2、…、x_L 依次堆叠，生成一个列向量。同样，运算 $\mathrm{blkdiag}\{X_1, X_2, \ldots, X_N\}$ 在主对角线上由矩阵 X_1、X_2、…、X_N 构成一个分块对角矩阵。给定两个矩阵 X 和 Y，记号 $X \succeq Y$ 表示 $X - Y$ 的元素均为非负数。对于方阵，记号 $X \geq Y$ 表示 $X - Y$ 是半正定的。记号 $\|x\|$ 表示向量 x 的 ℓ_2 范数，而 $\|X\|$ 表示矩阵 X 的相应诱导范数。给定一个向量 x 和一个半正定实矩阵 Y，符号 $\|x\|_Y^2$ 表示加权范数 $x^\top Y x$。Kronecker 积记为 $\otimes$，而 $X^\top$ 表示矩阵 X的转置。符号 $\mathbb{E}$ 表示期望算子。最后，给定一个非负函数 $f(\mu)$，记号 $f(\mu) = O(\mu)$ 表示存在一个常数$C > 0$ 和一个值 μ_0，使得对于所有 $\mu \leq \mu_0$，均有 $f(\mu) \leq C\mu$。

II. 背景

我们考虑由一个包含 N 个智能体的网络 k 处理的分布式回归问题。智能体 k 在时刻 i 观测到一个响应 变量 $\boldsymbol{d}_{k,i} \in \mathbb{R}$，以及一个由回归量组成的向量 $\boldsymbol{u}_{k,i} \in \mathbb{R}^M$，其相关矩阵为 $R_{u,k} = \mathbb{E}[\boldsymbol{u}_{k,i}\boldsymbol{u}_{k,i}^\top]$。响应变量与回归量通过以下模型相关联：

$$\boldsymbol{d}_{k,i} = \boldsymbol{u}_{k,i}^\top w^o + \boldsymbol{v}_{k,i}, \tag{1}$$

其中，$\boldsymbol{v}_{k,i}$ 表示一个零均值噪声过程，满足 $\mathbb{E}[\boldsymbol{v}_{k,i}^2] = \sigma_{v,k}^2$，且与回归量相互独立。我们假设过程 $\{\boldsymbol{u}_{k,i}\}$ 和 $\{\boldsymbol{v}_{k,i}\}$ 在时间上以及不同智能体之间相互独立。在式 (1) 中，$w^o \in \mathbb{R}^M$ 是未知的确定性参数向量，也是协作推断过程的核心对象。顺便指出，我们的分析可以推广到不同智能体具有不同模型w_k^o的情形，例如文献 [1], [2], [32], [33]中的做法。本文聚焦于统一模型这一重要情形，因为在多任务框架 [38]下，这种情形更适合研究多个局部模型。

网络中的每个智能体都关联有一个局部风险函数，该函数被定义为二次损失函数的期望，即，

$$J_k(w) \triangleq \mathbb{E}\left(\boldsymbol{d}_{k,i} - \boldsymbol{u}_{k,i}^\top w\right)^2. \tag{2}$$

通过 智能体之间的合作，网络目标是最小化总计 成本 ：

$$J(w) = \sum_{k=1}^{N} p_k J_k(w), \tag{3}$$

其中 $\{p_k\}$ 是总和为一的正系数。我们将在下文说明： $i)$ 这些系数与网络的拓扑性质相关，从而建立网络布局与分布式优化问题之间的有用联系；以及 $ii)$ 如果至少一个智能体的回归量具有非奇异相关矩阵，那么 $J(w)$ 的唯一极小化器将与真实模型参数 w^o一致。

在我们的模型中，智能体排列在一个网络中，其拓扑由 组合矩阵$A=[a_{\ell k}]$描述。矩阵条目 $a_{\ell k}$ 充当网络有向边上的组合权重 。当智能体 ℓ 与 k之间不存在通信链路时，相应的组合权重 $a_{\ell k}$ 和$a_{k\ell}$ 将为零。另一方面，当信息只能沿一个方向流动，例如从 ℓ 流向 k时，我们将有 $a_{\ell k}>0$ 和 $a_{k\ell}=0$。我们将智能体 k 的有向邻域（可能包括自环 $\ell=k$）表示为：

$$\mathcal{N}_k \triangleq \{\ell \,|\, a_{\ell k}>0\}. \tag{4}$$

我们假设组合矩阵满足以下正则性条件。

假设1 (左随机组合矩阵）。 A 的每一列中的元素之和为一。更具体地说，有如下结论。

$$a_{\ell k}\geq 0, \quad \sum_{\ell\in\mathcal{N}_k}^{N} a_{\ell k}=1, \quad a_{\ell k}=0 \text{ for } \ell\notin\mathcal{N}_k. \tag{5}$$

□

假设2 (强连通网络)。 存在一条有向路径将任意节点 ℓ 与任何其他节点 k连接起来，该路径始于 ℓ 并终于 k。此外，至少有一个智能体 m 具有自环 $(a_{mm}>0)$。 □

当假设 1 和 2 得到验证后，组合矩阵 A turnsouttobea 本原矩阵 [2],[37]。根据佩龙-弗罗贝尼乌斯定理， A will have a single 最大特征值1 ，其对应的特征向量为 $\pi=[\pi_1,\pi_2,\ldots,\pi_N]^\top$，称为 佩龙特征向量，其所有元素均严格为正，且总和为 1：

$$A\pi=\pi, \qquad \pi\succ 0, \qquad \mathbb{1}_N^\top\pi=1. \tag{6}$$

A. ACTC 扩散策略

考虑 $J_k(w)$ 在 (2) 中的梯度：

$$\nabla J_k(w)=2\left(R_{u,k}w-r_{du,k}\right), \tag{7}$$

并引入 为智能体 k 在时间 i求出如下瞬时梯度近似：

$$\boldsymbol{g}_{k,i}(w)=2\,\boldsymbol{u}_{k,i}\left[\boldsymbol{u}_{k,i}^\top w-\boldsymbol{d}_{k,i}\right]. \tag{8}$$

在本文中，我们重点研究 [32], [33], 提出的 ACTC 策略，该策略由以下三个步骤组成，并由每个智能体 $k = 1$、2、.....、N针对所有时刻 $i = 1$、2、.....、迭代执行：

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu_k \, \boldsymbol{g}_{k,i}(\boldsymbol{w}_{k,i-1}) \tag{9a}$$

$$\boldsymbol{q}_{\ell,i} = \boldsymbol{q}_{\ell,i-1} + \zeta \, \boldsymbol{Q}_\ell(\boldsymbol{\psi}_{\ell,i} - \boldsymbol{q}_{\ell,i-1}), \;\; \forall \ell \in \mathcal{N}_k \tag{9b}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \boldsymbol{q}_{\ell,i} \tag{9c}$$

式（9）中的递推由三个时变向量刻画：中间更新量$\boldsymbol{\psi}_{k,i}$、差分压缩更新量 $\boldsymbol{q}_{k,i}$以及当前的估计极小点 $\boldsymbol{w}_{k,i}$。在时刻 $i = 0$，每个智能体 k 均初始化为一个任意向量 $\boldsymbol{q}_{k,0}$ （具有有限的二阶矩）。然后，智能体 k 接收其邻居 $\boldsymbol{q}_{\ell,0}$ 的初始状态，并计算初始极小化器 $\boldsymbol{w}_{k,0} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \boldsymbol{q}_{\ell,0}$。

对于每个 $i > 0$，智能体执行以下步骤。

- 适应步(9a)：每个智能体 k 在本地遵循方向 $-\boldsymbol{g}_{k,\,i}(\cdot)$ （即其随机梯度的下降方向），并按步长 $\mu_k > 0$ 更新过去的迭代点 $\boldsymbol{w}_{k,\,i-1}$。此步骤对应于一个自学习阶段。

- 压缩步骤 (9b)：每个智能体 k 对中间更新量 $\boldsymbol{\psi}_{k,\,i}$与先前的压缩状态 $\boldsymbol{q}_{k,\,i-1}$之间的差值进行压缩。压缩通过应用适当的随机算子 $\boldsymbol{Q}_k(\cdot)$来执行，详见第 II-B 节。每个智能体 k 从其邻居 $\ell \in \mathcal{N}_k$ 接收压缩差值 $\boldsymbol{Q}_\ell(\boldsymbol{\psi}_{\ell,\,i} - \boldsymbol{q}_{\ell,\,i-1})$，并通过将接收到的压缩差值[1]加到先前状态 $\boldsymbol{q}_{\ell,\,i-1}$ 上，恢复压缩状态 $\boldsymbol{q}_{\ell,\,i}$，其中 $\ell \in \mathcal{N}_k$。需要指出的是，压缩差值由设计参数 $\zeta \in (0,\, 1)$ 进行缩放，该参数决定 ACTC 策略的稳定性——参见 [32], [33]。

- 组合步骤 (9c)：每个智能体 k 将其压缩状态与邻居更新后的压缩状态相结合，执行社交学习 步骤。

为便于后续使用，我们引入以下记号。ent旨在引入智能体之间的最大步长，以及缩放后的步长 s:

$$\mu \triangleq \max_{k=1,2,\ldots,N} \mu_k, \qquad \alpha_k \triangleq \frac{\mu_k}{\mu}. \tag{10}$$

B. 压缩算子

(9) 中 ACTC 策略的实现依赖于压缩算子 $\boldsymbol{Q}_k$（·）。遵循 [20]–[26], [28], [31]–[33], ，我们聚焦于以下相关的随机化压缩算子。

[1]为了执行压缩步骤 (9b)，智能体 k 必须拥有变量 $\{\boldsymbol{q}_{\ell,i-1}\}_{\ell \in \mathcal{N}_k}$。这是可行的，因为在 $i = 0$时，智能体 k 拥有 $\{\boldsymbol{q}_{\ell,0}\}_{\ell \in \mathcal{N}_k}$，而步骤 (9b) 取决于 $\{\boldsymbol{q}_{\ell,i-1}\}_{\ell \in \mathcal{N}_k}$ 和传输的差值 $\{\boldsymbol{Q}_\ell(\boldsymbol{\psi}_{\ell,\,i} - \boldsymbol{q}_{\ell,\,i-1})\}$。因此，共享初始状态 $\{\boldsymbol{q}_{\ell,0}\}_{\ell \in \mathcal{N}_k}$ 足以让智能体 k 在每个时刻 i 执行 (9b)，只需在内存中存储$\{\boldsymbol{q}_{\ell,i}\}_{\ell \in \mathcal{N}_k}$。

假设 3 (压缩算子）。 随机压缩算子将输入 x 与随机输出 $\boldsymbol{Q}(x)$ 相关联，其分布由条件概率测度 $\mathbb{P}(\cdot|x)$ 决定。关于压缩误差：

$$\boldsymbol{e}(x) \triangleq \boldsymbol{Q}(x) - x, \tag{11}$$

所考虑的压缩算子类满足以下两个条件：

$$\mathbb{E}\,\boldsymbol{e}(x) = 0 \qquad \text{[unbiasedness]} \tag{12}$$

$$\mathbb{E}\,\|\boldsymbol{e}(x)\|^2 \leq \omega\,\|x\|^2 \qquad \text{[variance bound]} \tag{13}$$

其中期望是相对于 $\mathbb{P}(\cdot|x)$的。 □

参数 ω 发挥着关键作用，因为它量化了施加于输入 x的压缩量。较小的 ω 对应于低失真的精细量化数据，例如式 (13) 所示。另一方面，较大的 ω 对应于高失真的粗略量化数据。多种压缩算子属于假设 3 所表示的类别，在接下来的仿真中，我们将使用两个常见实例，即 [20] 中的随机量化器和 [31]中的随机稀疏器。

示例 1 (随机量化器 [20]**)**。 给定输入向量 $x \in \mathbb{R}^M$，随机量化器表示其欧几里得范数 $\|x\|$ ，量化误差可忽略（例如达到机器精度），使用 h 位，并分别量化 x 的每个分量 x_m 。一位用于编码 x_m 的符号，并使用 r 位表示 $\chi_m = x_m/\|x\| \in [0,1]$。区间 $[0,1]$ 被划分为 L 个大小为 $\theta = 1/L$ 的等子区间。相应地，给定 r 位，量化级的数量为 $L = 2^r - 1$。随后，通过随机选择包围每个 χ_m 的两个端点之一对其进行编码。令 $j(\chi_m) = \lfloor \chi_m/\theta \rfloor$ 为下端点的索引，其等于 $y(\chi_m) = j(\chi_m)\,\theta$。随机编码基于以下规则：

$$\boldsymbol{j}_{\text{tx}}(\chi_m) = \begin{cases} j(\chi_m) + 1, & \text{with prob. } \frac{\chi_m - y(\chi_m)}{\theta} \\ j(\chi_m), & \text{otherwise} \end{cases}. \tag{14}$$

最后，第 m个量化分量为：

$$[\boldsymbol{Q}(x)]_m = \|x\|\,\text{sign}(x_m)\,\boldsymbol{j}_{\text{tx}}(\chi_m)\,\theta, \quad m = 1, 2, \ldots, M. \tag{15}$$

它显示为在 [20] 中，随机量化器的压缩参数等于 o:

$$\omega = \min\left\{\frac{M}{(2^r-1)^2}, \frac{\sqrt{M}}{2^r-1}\right\}, \tag{16}$$

且对向量 x 进行编码所需的比特数等于 [20]:

$$h + M \times (r+1). \tag{17}$$

□

例2 (随机稀疏化器 [31])。 给定输入 $x \in \mathbb{R}^M$，随机稀疏器应用此规则

$$\boldsymbol{Q}(x) = \frac{M}{S}\,\boldsymbol{F}_S(x), \tag{18}$$

其中 $\boldsymbol{F}_S(x)$ 随机选择 $S < M$ 个分量中的 x，并将其余部分置零 .

在 [22] 中证明，随机稀疏器的压缩参数等于:

$$\omega = \frac{M}{S} - 1, \tag{19}$$

并且，通信节省量通过维度缩减因子 S/M 来量化，例如在压缩感知或模拟压缩框架的无损情形 [40], [41] 中就是如此。注意到，通过仅对未掩蔽分量的数值和位置进行编码，可以高效地表示稀疏向量，因此还可以使用 h 位来表示数值，并使用 $\log_2(M)$ 位来表示位置，从而计算以比特为单位的成本。 □

本工作考察的压缩算子类别已被证明能够忠实地建模实际压缩方案，并且非常适合学习应用中常见的高维场景 [20], [29]–[31], [40], [41]。例如，在随机稀疏化范式中，通过将高维变量压缩到显著更小的子空间中来节省通信开销，因此通信成本相应地以所实现的降维程度来衡量。同样，示例 1 中的向量量化方案使用许多比特来表示高维向量的范数，然后将少量比特分配给各个向量分量。

我们急于补充说明，此处介绍的压缩算子类别并不全面。还可以采用其他构造方式。在 [39] 中提出的一类有用方法，在方差界（13）中引入了第二个参数。我们指出，我们工作中使用的工具也可用于研究这类其他压缩方案——例如参见 [42]。

C. ACTC 策略中的压缩算子

给定智能体 k 在时间 i的向量 $\boldsymbol{x}_{k,\,i}$，我们用 $\boldsymbol{x}_{0:i}$ 表示一个向量，该向量汇集了所有智能体在截至时间 i的所有时刻的$\boldsymbol{x}_{k,\,i}$。我们还引入 ACTC 策略步骤（9b）中待压缩的差分变量及其对应误差:

$$\boldsymbol{\delta}_{k,i} \triangleq \boldsymbol{\psi}_{k,i} - \boldsymbol{q}_{k,i-1}, \qquad \boldsymbol{e}_{k,i} \triangleq \boldsymbol{Q}_k(\boldsymbol{\delta}_{k,i}) - \boldsymbol{\delta}_{k,i}. \tag{20}$$

其中，智能体可以使用不同的压缩算子。因此，对于给定的输入 x，每个智能体 k 都使用一个由概率测度 $\mathbb{P}_k(\cdot|x)$ 定义的压缩算子 $\boldsymbol{Q}_k(\cdot)$，并满足假设 3，其压缩参数为 ω_k。按照数据压缩分布式优化中的惯例，我们假设压缩算子无记忆，且各智能体之间相互独立。

假设4 (关于压缩算子的条件)。 压缩过程在时间和空间上相互独立，即给定 $\boldsymbol{q}_{0:i-1}$ 和 $\boldsymbol{\psi}_{1:i}$后，智能体 k 的压缩向量 $\boldsymbol{Q}_k(\boldsymbol{\delta}_{k,i})$ 仅取决于输入$\boldsymbol{\delta}_{k,i}$，且压缩向量的生成方式如下：

$$\boldsymbol{\delta}_{k,i} \xrightarrow{\mathbb{P}_k(\cdot|\boldsymbol{\delta}_{k,i})} \boldsymbol{Q}_k(\boldsymbol{\delta}_{k,i}), \tag{21}$$

在各智能体之间相互独立。因此，假设 *3* 意味着：

$$\mathbb{E}\big[\boldsymbol{e}_{k,i} \,|\, \boldsymbol{q}_{0:i-1}, \boldsymbol{\psi}_{0:i}\big] = \mathbb{E}\big[\boldsymbol{e}_{k,i} \,|\, \boldsymbol{\delta}_{k,i}\big] = 0, \tag{22}$$

$$\mathbb{E}\big[\|\boldsymbol{e}_{k,i}\|^2 \,|\, \boldsymbol{q}_{0:i-1}, \boldsymbol{\psi}_{0:i}\big] = \mathbb{E}\big[\|\boldsymbol{e}_{k,i}\|^2 \,|\, \boldsymbol{\delta}_{k,i}\big] \leq \omega_k \|\boldsymbol{\delta}_{k,i}\|^2, \tag{23}$$

$$\mathbb{E}\big[\boldsymbol{e}_{\ell,i}\boldsymbol{e}_{k,i}^{\top} \,|\, \boldsymbol{q}_{0:i-1}, \boldsymbol{\psi}_{0:i}\big] = \mathbb{E}\big[\boldsymbol{e}_{\ell,i}\boldsymbol{e}_{k,i}^{\top} \,|\, \boldsymbol{\delta}_{k,i}\big] = 0, \, \forall \ell \neq k. \tag{24}$$

□

D. 代价函数与梯度噪声的性质

本节陈述局部风险函数和梯度噪声的两项性质，这些性质对于确立我们的结果至关重要。

性质 1 (光滑风险)。 局部风险函数 (2) 二阶可微，且其 *Hessian* 矩阵的特征值关于 w一致有界，即

$$\nabla^2 J_k(w) \leq \eta_k I_M, \qquad \eta_k > 0. \tag{25}$$

聚合风险函数(3) 还具有 ν-强凸性，即

$$\sum_{k=1}^{N} \alpha_k \pi_k \nabla^2 J_k(w) \geq \nu I_M, \qquad \nu > 0, \tag{26}$$

当且仅当至少一个智能体具有非奇异相关矩阵 $R_{u,\,k}$。在 (26) 条件下，全局风险(3)的极小化器唯一，且由真实参数 w^o给出。

证明：关系 (25) 是显然的，因为由 (7) 可知，局部海森矩阵为 $\nabla^2 J_k(w) = 2R_{u,k}$。另一方面，鉴于 α_k 和 π_k 严格为正，当且仅当至少一个局部海森矩阵非奇异，即正定时，关系 (26) 才成立。

■

我们进一步引入梯度噪声，它体现了智能体在适应步骤 (9a) 中使用 (8) 代替实际梯度所引入的误差：

$$\boldsymbol{n}_{k,i}(\boldsymbol{w}_{k,i-1}) \triangleq \boldsymbol{g}_{k,i}(\boldsymbol{w}_{k,i-1}) - \nabla J_k(\boldsymbol{w}_{k,i-1}). \tag{27}$$

我们将其缩放后的版本记为

$$\boldsymbol{s}_{k,i} \triangleq \alpha_k \boldsymbol{n}_{k,i}(\boldsymbol{w}_{k,i-1}). \tag{28}$$

梯度噪声会产生扰动，使式 (3) 的估计极小点持续在其真实值 [2], [32]–[35] 附近波动。在梯度噪声满足标准正则性条件时，ACTC 扩散策略的收敛性已在 [33] 中得到证明。

性质 2 (梯度噪声过程)。 梯度噪声满足以下两个性质：

$$\mathbb{E}\big[\boldsymbol{s}_{k,i}\,|\,\boldsymbol{q}_{0:i-1},\boldsymbol{\psi}_{0:i-1}\big]=0, \tag{29}$$

$$\mathbb{E}\big[\boldsymbol{s}_{k,i}\,\boldsymbol{s}_{\ell,i}^{\top}\,|\,\boldsymbol{q}_{0:i-1},\boldsymbol{\psi}_{0:i-1}\big]=0,\qquad \forall \ell\neq k. \tag{30}$$

此外，如果我们将 $M\times M$ 与智能体 k 相关的梯度噪声的条件协方差矩阵记为

$$\mathbb{E}\left[\boldsymbol{s}_{k,i}\boldsymbol{s}_{k,i}^{\top}\,|\,\boldsymbol{q}_{0:i-1},\boldsymbol{\psi}_{0:i-1}\right]=R_{s,k}(\boldsymbol{w}_{k,i-1}), \tag{31}$$

那么它可表示为

$$\begin{aligned}R_{s,k}(w)\triangleq{}&4\,\alpha_k^2\,\sigma_{v,k}^2\,R_{u,k}\\&+4\,\alpha_k^2\,\mathbb{E}\left[\boldsymbol{U}_{k,i}(w-w^o)(w-w^o)^{\top}\boldsymbol{U}_{k,i}^{\top}\right],\end{aligned} \tag{32}$$

其中 $\boldsymbol{U}_{k,i}=\boldsymbol{u}_{k,i}\boldsymbol{u}_{k,i}^{\top}-R_{u,k}$。此外，在稳态下（较大的 i），在$\boldsymbol{w}_{k,i-1}$ 处评估的协方差矩阵 $R_{s,k}(w)$ 等价于在全局极小点 w^o处评估的协方差矩阵，差异仅为 $O(\mu)$ 扰动，形式化地表示为：

$$\limsup_{i\to\infty}\mathbb{E}\|R_{s,k}(\boldsymbol{w}_{k,i-1})-R_{s,k}(w^o)\|=O(\mu). \tag{33}$$

关系(33) 尤其意味着：

$$\begin{aligned}&\mathrm{Tr}[R_{s,k}(w^o)]-O(\mu)\\&\leq\liminf_{i\to\infty}\mathbb{E}\|\boldsymbol{s}_{k,i}\|^2\leq\limsup_{i\to\infty}\mathbb{E}\|\boldsymbol{s}_{k,i}\|^2\\&\leq\mathrm{Tr}[R_{s,k}(w^o)]+O(\mu).\end{aligned} \tag{34}$$

证明：参见附录 C。 ■

III. ACTC 策略的学习性能

本文的主要结果是刻画稳态区间内当 $i\to\infty$ 时每个智能体 k 的 ACTC 策略的均方误差性能。在陈述说明这一结果的定理之前，我们需要先介绍一些关键量。设 Let

$$\Delta_s\triangleq\zeta\,\mathrm{Tr}\left[\left(\sum_{k=1}^{N}\alpha_k\pi_k R_{u,k}\right)^{-1}\left(\sum_{k=1}^{N}\alpha_k^2\,\pi_k^2\,\sigma_{v,k}^2\,R_{u,k}\right)\right], \tag{35}$$

$$\Delta_\omega\triangleq\frac{2\,\zeta}{\nu}\sum_{k=1}^{N}\alpha_k^2\,\pi_k^2\,\sigma_{v,k}^2\,\omega_k\,\mathrm{Tr}[R_{u,k}]+\frac{c\,\zeta^2}{2\,\nu}, \tag{36}$$

其中 c 是一个与 μ 无关的常数，其值取决于压缩参数 $\{\omega_k\}_{k=1}^N$。设

$$\widetilde{\boldsymbol{w}}_{k,i} = \boldsymbol{w}_{k,i} - w^o, \tag{37}$$

and

$$\mathscr{E}_k^{\inf} = \liminf_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2, \quad \mathscr{E}_k^{\sup} = \limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2. \tag{38}$$

定理 1（稳态性能）。假设至少有一个智能体 m 具有非奇异的相关矩阵$R_{u,\,m}$。在假设 ***1–4*** 下，对于足够小的 μ 和 ζ，使得 *ACTC* 策略均方稳定，[2] 对于所有 ℓ 和 k（包括 $\ell = k$ 的情况），我们有：

$$\limsup_{i\to\infty} \left| \frac{\mathbb{E}\widetilde{\boldsymbol{w}}_{\ell,i}^{\top}\widetilde{\boldsymbol{w}}_{k,i}}{\sqrt{\mathbb{E}\|\widetilde{\boldsymbol{w}}_{\ell,i}\|^2\mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2}} - 1 \right| = O(\mu^{1/2}). \tag{39}$$

此外，每个智能体 k 的均方误差有如下界：

$$\mu\,\Delta_s - O(\mu^{3/2}) \le \mathscr{E}_k^{\inf} \le \mathscr{E}_k^{\sup} \le \mu\,(\Delta_s + \Delta_\omega) + O(\mu^{3/2}). \tag{40}$$

证明：参见本文末尾的附录 B，该证明进一步利用了附录 A 中的通用结果。

■

A. 关于定理 *1* 的评论

关系 (39) 突出了一个基本结果。它表明，在稳态下以及所考虑的小步长范围内，任意两个智能体的迭代量趋于完全相关，即

$$\mathbb{E}\widetilde{\boldsymbol{w}}_{\ell,i}^{\top}\widetilde{\boldsymbol{w}}_{k,i} \approx \sqrt{\mathbb{E}\|\widetilde{\boldsymbol{w}}_{\ell,i}\|^2\mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,i}\|^2}, \tag{41}$$

根据柯西–施瓦茨（不）等式，这意味着 $\widetilde{\boldsymbol{w}}_{\ell,\,i}$ 和 $\widetilde{\boldsymbol{w}}_{k,\,i}$ 近似成比例。另一方面，由于 (39) 对 $\ell = k$ 也成立，我们得出 $\widetilde{\boldsymbol{w}}_{\ell,\,i}$ 和 $\widetilde{\boldsymbol{w}}_{k,\,i}$ 具有近似相同的范数，因此比例常数为1：

$$\widetilde{\boldsymbol{w}}_{\ell,i} \approx \widetilde{\boldsymbol{w}}_{k,i}. \tag{42}$$

这意味着所有智能体的误差以协调的 方式演变。

一旦确定所有智能体都达到相同的误差，我们就希望评估该误差。首先注意到，(40) 中的下界由参数 Δ_s决定，而根据(35)，该参数受回归问题中变异性的主要来源，即梯度噪声，所支配。事实上，该下界与未压缩策略 [2] 所得到的极限均方误差相一致。换言之，压缩策略不可能优于未压缩策略。这似乎是一个显而易见的结果，但仍需要形式化证明。上界通过 (36) 中的项 Δ_ω 体现了压缩的影响。注意，主导项在

[2]关于 μ和 ζ 的 ACTC 策略均方稳定性条件已在 [32], [33]中讨论。

下界和上界并不一致。这一差距源于，根据假设 3，我们只能得到刻画压缩误差的上界。如果能够更精确地刻画该误差，我们就能弥合下界与上界之间的差距。在下文中，我们将使用上界作为均方误差性能的代理，因为它考虑了压缩的影响，因此也有助于优化通信资源的分配。

让我们更仔细地考察式 (40) 中上界的结构。利用式 (35) 和 (36)，可以方便地将其分解如下（忽略高阶 $O(\mu^{3/2})$ 修正项）：

$$\underbrace{\mu\,\zeta\,\mathrm{Tr}\left[\left(\sum_{k=1}^{N}\alpha_k\pi_k R_{u,k}\right)^{-1}\left(\sum_{k=1}^{N}\alpha_k^2\,\pi_k^2\,\sigma_{v,k}^2\,R_{u,k}\right)\right]}_{\text{uncompressed evolution error}} + \underbrace{\frac{2\,\mu\,\zeta}{\nu}\sum_{k=1}^{N}\alpha_k^2\,\pi_k^2\,\sigma_{v,k}^2\,\omega_k\,\mathrm{Tr}\left[R_{u,k}\right]}_{\text{gradient noise compression loss}} + \underbrace{\frac{c\,\mu\,\zeta^2}{2\,\nu}}_{\substack{\text{network error}\\ \text{compression loss}}}. \tag{43}$$

我们重点考察误差的各个分量。

— 未压缩演化误差。该项对应于共享完美信息（即未压缩信息）时扩散策略所达到的均方误差。其结构与经典 ATC 扩散策略的性能相同，涉及推断问题的主要属性，即回归量相关矩阵$R_{u,\,k}$ 以及噪声方差 $\sigma_{v,\,k}^2$，并由佩龙特征向量和用于量化各个步长 [2], [35]之间差异的参数α_k 进行适当加权。

— 压缩损失。由于共享压缩信息，稳态行为会受到额外误差来源的影响。式 (43) 突出了压缩损失的根本来源，即梯度噪声 和网络误差分量。

前一项可以与经典量化系统的行为联系起来，其失真取决于待压缩随机变量的方差。在稳态区间内，压缩步骤(9b)中的创新具有与梯度噪声相关的方差——参见附录B-C2 中的引理 2。因此，我们得到一个包含梯度噪声协方差矩阵的迹的 失真 项，$R_{s,\,k}(w^o)=4\,\alpha_k^2\,\sigma_{v,\,k}^2R_{u,\,k}$。网络误差分量项处理各个智能体与网络朝向公共极小点的协调演化之间的局部差异。在经典的 ATC 扩散策略中，该误差项是高阶修正项 [2], [34], [35], ，而在ACTC扩散策略中，由于压缩误差渗入算法演化，该误差项会增大；现在它与均方误差表达式中的其他项处于同一数量级，即 μ 。

值得注意的是，从式（43）可以看出，压缩参数 ω_k 按佩龙特征向量各分量的平方加权，这表明网络布局与设计之间存在一种有用的关系。

压缩算子。因此，考虑网络结构来设计压缩算子，可以用于调节均方误差性能。我们将在下一节探讨这种可能性。

IV. 说明性示例与优化资源分配

我们将在以下均方误差网络场景中说明定理 1 的影响：观测回归量的维数为 $M = 30$，其特征由对角矩阵 $R_{u,\,k}$ 表示，方差在区间 $(1,\ 4)$ 内均匀且独立地抽取。同样，噪声方差 $\sigma^2_{v,\,k}$ 在区间 $(0.25,\ 1)$ 内均匀抽取。该网络由 $N = 10$ 个智能体组成，按照图 1 所示的拓扑排列；在此基础上，我们采用左随机组合策略，并使用均匀平均规则 [2]。所有智能体均采用相同的步长 $\mu_k = \mu = 10^{-2}$，稳定性参数为 $\zeta = 10^{-1}$。图 1 左侧的图展示了数据压缩如何影响网络 性能，即所有智能体上的均方误差平均值：

$$\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\|\boldsymbol{w}_{k,i} - w^o\|^2. \tag{44}$$

智能体使用示例 1 中介绍的随机量化器对传输的信息进行量化。随着比特率增加，ACTC 性能逐渐接近 ATC 性能 [32], [33], ，这与式 (40) 一致。

A. 优化的资源分配

给定通信资源的总预算，我们希望根据学习性能优化其在各智能体之间的分配。令 $x = [x_1, x_2, \ldots, x_N]$，其中 x_k 表示分配给智能体 k的通信资源。请注意，变量 x_k 的作用取决于所选的具体压缩算子。例如，使用例1中的随机量化器时， x_k 是对输入的缩放分量进行量化所需的比特数。使用例2中的随机稀疏化器时， x_k 表示智能体 k未将其屏蔽为零的输入向量分量数量。我们假设每个 x_k 都必须位于给定区间内，且总预算不得超过规定值 X。在这些约束下，我们可以最小化式 (40) 中的上界。从式 (40) 中去掉与 x无关的项 Δ_s，并忽略 $O(\mu^{3/2})$ 修正项以及网络误差分量压缩损失项 $c\zeta/(2\nu)$；对于所考虑的较小 ζ值，这两项对性能的影响显著降低。由此得到问题：

$$\begin{aligned}
&\min_{x\in\mathbb{R}^N} \sum_{k=1}^{N} \pi_k^2\,\omega_k\,d_k \\
&\text{s.t.} \quad \sum_{k=1}^{N} x_k \le X \\
&\qquad x_{\min} \le x_k \le x_{\max},\ k = 1, 2, \ldots, N.
\end{aligned}$$

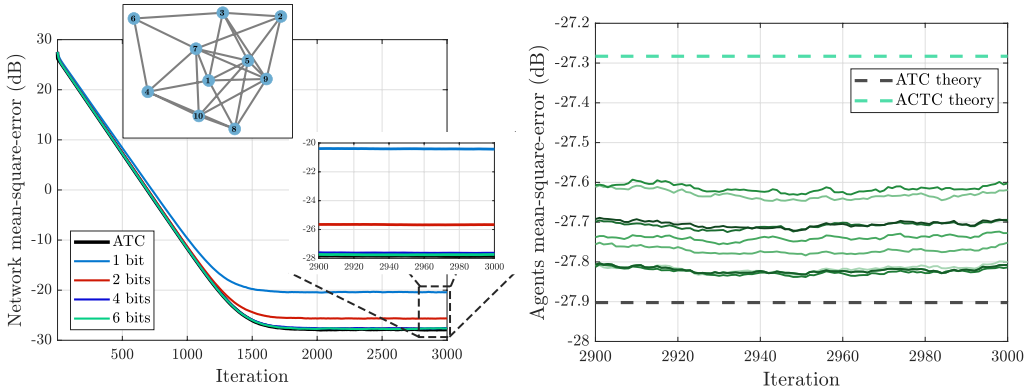


图 1。ACTC 均方误差性能随迭代次数的变化 i。我们参考第 IV 节中的实验设置，其中：(*i*) 回归量 $\boldsymbol{u}_{k,i} \in \mathbb{R}^M$ 服从具有对角协方差矩阵的零均值高斯分布，其方差从 (1,4) 上的均匀分布中独立抽取；(*ii*) 噪声 $\boldsymbol{v}_{k,i} \in \mathbb{R}$ 服从零均值高斯分布，其方差从 (0.25,1) 上的均匀分布中独立抽取。ACTC 策略采用相同步长 $\mu_k = \mu = 10^{-2}$ 和稳定性参数$\zeta = 10^{-1}$运行。左图。智能体应用示例 1 中具有不同比特率的随机量化器时的网络均方误差。插图显示了网络拓扑，其上应用了平均组合策略。所有节点均具有一个自环（为简洁起见未显示）。右图。采用 $r = 6$ 比特时，智能体稳态均方误差的局部放大图（以不同深浅的绿色显示）。黑色虚线表示 ATC 策略 [2], [35] 的均方误差，绿色虚线表示式 (40) 中的均方误差界。均方误差通过 10^3 次蒙特卡洛运行进行估计。

其中我们回顾， ω_k 取决于 x_k ，并令：

$$d_k \triangleq \frac{\mathrm{Tr}[R_{s,k}(w^o)]}{4} = \alpha_k^2 \, \sigma_{v,k}^2 \, \mathrm{Tr}[R_{u,k}]. \tag{45}$$

精确求解（45）需要对 x 施加整数约束，这会导致求解器不实用。与量化理论中的比特分配问题 [13], 通常采用的做法一样，我们在实数域上求解问题（ 45 ），然后将解舍入为满足约束的整数值。

B. 在线资源分配解

我们可以通过应用 Karush-Kuhn-Tucker（KKT）条件来计算（45）的解，参见附录 H。求解需要知道佩龙特征向量的各个分量 π_k 和失真系数 d_k，而这些信息通常无法由智能体获得，但可以通过在线方式进行估计，如下所示。

一种估计佩龙特征向量的方法是，在运行 ACTC 扩散策略的同时，运行一个平均一致性算法 [10]。经过 t 次迭代后，一致性算法的输出可表示为：

$$y_t = (A^\top)^t \, y_0 \xrightarrow{t\to\infty} \sum_{\ell=1}^{N} \pi_\ell y_{\ell,0}, \tag{46}$$

其中 $y_t = [y_{k,t}]$ 是收集智能体 k 在第 t次迭代时输出$y_{k,\,t}$ 的 $N \times 1$ 向量。在假设1下，组合矩阵 A 的幂的各列收敛于佩龙特征向量 [2]。如果智能体 k 是

以值 $y_{k,\,0}=1$ 初始化，而所有其他智能体的值为 0，则所有智能体都将收敛于佩龙特征向量的第 k 个元素，即 π_k。将该过程重复 N 次，即可在所有智能体处得到 π 的估计值。请注意，平均一致性以指数速度收敛 [10], ，因此只需几次迭代即可获得 π 的准确估计。

为了估计 di　　失真项 d_k 我们遵循 [2], [36]。首先，我们注意到，根据 (9a)，我们可以写　e

$$\begin{aligned}\boldsymbol{\psi}_{k,i}-\boldsymbol{w}_{k,i-1}&=\mu_k\,\boldsymbol{g}_{k,i}(\boldsymbol{w}_{k,i-1})\\&=\mu\,\boldsymbol{s}_{k,i}+\mu\,\alpha_k\big(\nabla J_k(\boldsymbol{w}_{k,i-1})-\nabla J_k(w^o)\big),\end{aligned}\tag{47}$$

其中，在第二个等式中，我们加上并减去梯度 $\nabla J_k(\boldsymbol{w}_{k,i-1})$，并利用 $\nabla J_k(w^o)=0$ 这一事实。根据性质 1 中的利普希茨条件，我们可以用范数 $\|\boldsymbol{w}_{k,i-1}-w^o\|^2$ 来界定 $\|\nabla J_k(\boldsymbol{w}_{k,i-1})-\nabla J_k(w^o)\|^2$，其上极限为 $O(\mu)$。因此，利用 (34) 和 (45)，由 (29) 和 (47) 可得：

$$\begin{aligned}d_k-O(\mu)&\le\frac{4}{\mu^2}\liminf_{i\to\infty}\mathbb{E}\|\boldsymbol{\psi}_{k,i}-\boldsymbol{w}_{k,i-1}\|^2\\&\le\frac{4}{\mu^2}\limsup_{i\to\infty}\mathbb{E}\|\boldsymbol{\psi}_{k,i}-\boldsymbol{w}_{k,i-1}\|^2\le d_k+O(\mu).\end{aligned}\tag{48}$$

因此，在稳态下（即对于 i 足够大的情况），并且在所考虑的小步长范围内，我们近似有：

$$\mathbb{E}\|\boldsymbol{\psi}_{k,i}-\boldsymbol{w}_{k,i-1}\|^2\propto d_k\tag{49}$$

对于 (45) 中的优化而言，比例常数显然无关紧要。在遍历性假设下，可随着时间推移，通过计算 $\|\boldsymbol{\psi}_{k,\,i}-\boldsymbol{w}_{k,\,i-1}\|^2$ 的（在线）经验均值来估计 (49) 中的统计均值；甚至可以从自适应的角度，采用遗忘因子为 $\xi_k\in(0,\,1)$ 的平滑滤波器，即 [2], [36]：

$$\boldsymbol{d}_{k,i}=(1-\xi_k)\,\boldsymbol{d}_{k,i-1}+\xi_k\|\boldsymbol{\psi}_{k,i}-\boldsymbol{w}_{k,i-1}\|^2,\tag{50}$$

其中，$\boldsymbol{d}_{k,i}$ 是 d_k 在时间 i的估计值。为了应用附录H 中提供的 KKT解，每个智能体都必须能够访问估计的 $\{\pi_k\}$ 和 $\{d_k\}$。因此，一旦智能体估计出这些量，就会通过网络共享这些值。就通信成本而言：$i)$ d_k 的估计由每个智能体在本地执行（无需通信）；$ii)$ 用于估计佩龙特征向量的一致性算法以及共享估计量所产生的通信成本可以安全地忽略，因为该过程仅发生一次，因此在 ACTC 策略的迭代过程中会被摊薄。

C. 对实际压缩方案的应用

我们首先考察在例1中采用随机量化器、在例2中采用随机稀疏器时方程（45）的解。在以下示例中，我们使用 Bollob´as-Riordan 优先连接模型 [43], [44] 生成图2所示的网络拓扑，该模型可创建自然聚类的拓扑，其中包含邻域较大的枢纽节点。在此网络拓扑上，我们采用通过相对度策略 [2] 获得的组合策略。

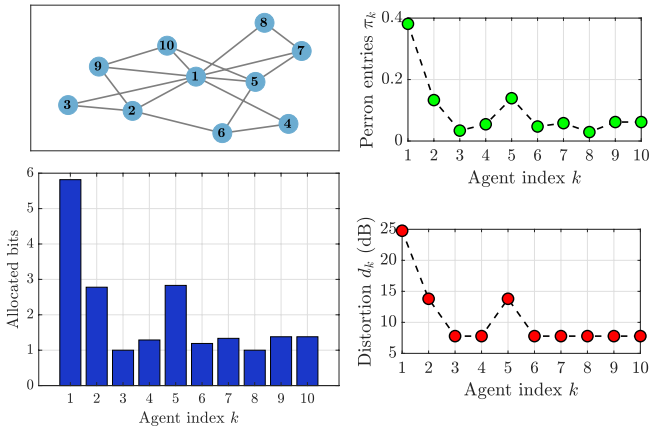


图 2。左上。 使用 Bollob´as-Riordan 模型构建的网络拓扑，其中智能体 1 充当枢纽节点。左下。 使用示例 1 中的随机量化器为 ACTC 策略进行优化的比特分配，其解为附录 H 中的 (45)。右上。采用相对度组合策略的所考虑拓扑的佩龙权重分布 π_k 。右下。(45) 中的失真值 d_k 。我们可以看到，分配给各智能体的比特遵循系数 π_k 和 d_k 的变化趋势。在此示例中，我们将矩阵 $R_{u,k}$ 设为 $\{5I_M, 2I_M, I_M, I_M, 2I_M, I_M, I_M, I_M, I_M, I_M\}$，并将方差 $\sigma^2_{v,k}$ 设为$\{1, 0.2, 0.1, 0.1, 0.2, 0.1, 0.1, 0.1, 0.1, 0.1\}$。

1) 随机量化器资源分配：将表达式 (16) 代入 (45)会导致一个不切实际、不可微且非凸的优化问题。因此，我们考虑如下给出的压缩参数 (16) 的近似版本：

$$\widehat{\omega}_k = \frac{M}{(2^{x_k} - 1)^2}. \tag{51}$$

根据（16），关系（51）是压缩误差（13）的上界；众所周知，在高分辨率 状态下，即在 x_k 取相对较高值时，该上界是紧的。我们通过应用 Karush-Kuhn-Tucker（KKT）条件求解问题（45）——参见附录 H。

在（51）所隐含的高分辨率近似下，问题（45）可以借助标量量化器的最优比特分配经典方法以闭式求解。闭式解将为我们理解优化分配提供有益的诠释。将（51）写成 $\widehat{\omega}_k \approx M2^{-2x_k}$ 并忽略盒约束后，问题（45）与 [13] 中的比特分配问题具有相同的结构。通过拉格朗日乘子，或应用算术-几何平均不等式，可以证明最优分配为 [13, §8.3]:

$$x_k = \bar{x} + \log_2 \frac{\pi_k}{\pi_{\text{av}}} + \frac{1}{2} \log_2 \frac{d_k}{d_{\text{av}}}, \tag{52}$$

其中

$$\bar{x} = \frac{X}{N}, \quad \pi_{\text{av}} = \left(\prod_{k=1}^{N} \pi_k \right)^{\frac{1}{N}}, \quad d_{\text{av}} = \left(\prod_{k=1}^{N} d_k \right)^{\frac{1}{N}}. \tag{53}$$

认识到 $\bar{x}$ 是每个智能体的平均比特分配，这一结果表明，智能体k 可用的资源是对 $\bar{x}$ 的扰动，其取决于佩龙权重 π_k 和失真值 d_k。

观察图 2 示例中问题 (45) 的优化比特分配，我们注意到，大多数可用资源都分配给了佩龙权重较高（即中心性较高）且误差源显著（即更容易出错）的智能体。这一解读与关系 (52) 一致，该关系准确揭示了网络中心性和数据质量如何相互作用，从而决定优化后的资源分配。智能体 k 的中心性越高，π_k 就越大，这意味着当 $\pi_k > \pi_{\text{av}}$时，相对于平均分配 $\bar{x}$，它能获得更多压缩资源。事实上，“中心”智能体会将信息转发给许多邻居，因此在推动网络趋向目标极小化器方面影响深远。这解释了为什么在 (52) 中，较高的佩龙权重 π_k 有利于分配更多比特。同理，为失真 d_k 高于平均失真 d_{av}的智能体分配更多通信资源也是有益的，因为这样可以减少在网络中传播的压缩误差。当需求相互冲突时（例如中心性较低而失真较高），分配规则 (52) 揭示了如何权衡取舍。

图3显示，优化的资源分配显著提升了 ACTC 策略的稳态网络性能，使其更接近参考 ATC 性能，且不增加额外资源开销。左图展示了随机量化器采用优化分配后的效果，其中设置（45）中的 $X = 20$, $x_{\min} = 1$ 和 $x_{\max} = 11$。智能体在进入稳态后，于时刻 $T_{\text{opt}} = 1600$ 求解优化问题。蓝色和红色曲线分别比较均匀分配（对所有 k 均为 $x_k = \bar{x}$ ）与优化分配；后者通过求解 KKT 解，再应用向下取整算子得到整数值。可以看到，网络从优化分配中获益。值得注意的是，时刻 T_{opt} 之后仅经过几步，红色曲线便接近参考 ATC 误差（即未压缩策略的误差）。与均匀分配所达到的性能相比，优化策略将均方误差降低了约2dB。

比较 ACTC 和 ATC 策略的比特开销也很有用。假设机器精度为$h = 32$，并考虑经过 $T = 2000$ 次迭代达到收敛，根据 (17) 可知，ACTC 策略的比特开销为：

$$r_{\text{ACTC}} = T\left(Nh + MX + MN\right) = 2.396 \text{ Mbits}, \tag{54}$$

请注意，比特开销仅取决于总比特预算 X，因此均匀分配和优化分配的比特开销相同。然而，根本区别在于，在开销相同的情况下，优化分配能够达到几乎与作为参照的（未压缩）ATC 策略相同的性能，而后者的比特开销为：

$$r_{\text{ATC}} = T \times N \times M \times h = 19.2 \text{ Mbits}, \tag{55}$$

这意味着可节省超过16兆比特。

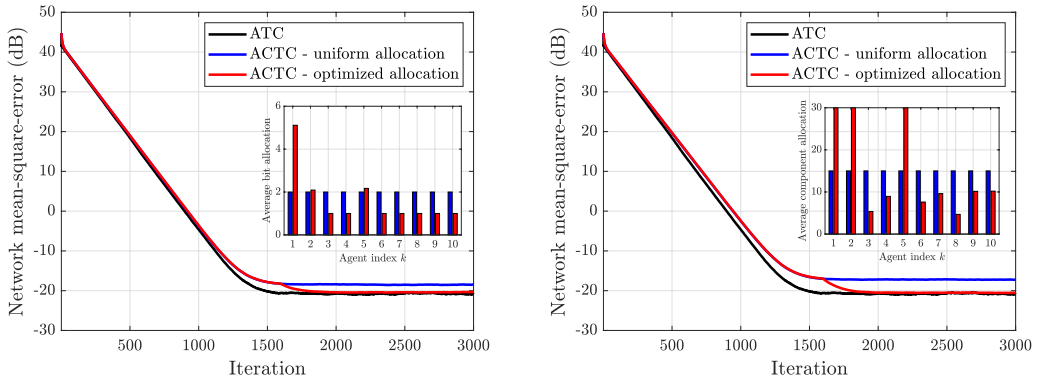


图3。在统一资源分配和优化资源分配下，ACTC均方误差性能随迭代次数 i的变化。网络拓扑如图2所示，并采用相对度组合策略。回归量矩阵$R_{u,\,k}$ 以及噪声方差$\sigma^2_{v,\,k}$ 也在图2的图注中给出。其他系统参数与图1示例中使用的参数相同。左图。智能体在例1中应用随机量化器时的网络均方误差，总资源预算为$X = 20$ ，约束为 $1 \le x_k \le 11$。插图显示了两种分配策略的比特率 x_k 。右图。智能体在例2中应用随机稀疏化器时的网络均方误差，总资源预算为 $X = 150$ ，约束为$1 \le x_k \le M$。插图显示了两种分配策略下未屏蔽分量的数量 x_k 。在两项实验中，智能体均在时间 $T_{\text{opt}} = 1600$计算优化的资源分配。均方误差通过 10^3 次蒙特卡洛运行进行估计。

2）随机稀疏器的资源分配：我们再次应用 KKT条件（见附录 H），利用随机稀疏器的压缩参数 (19) 求解问题 (45)。我们设置的资源总预算为 $X = 150$ ，其中 $x_{\min} = 1$ ，且 $x_{\max} = M$。与上一个示例一样，优化后的资源分配在 $i = 1600$ 处计算。在均匀分配中，每个智能体的输入向量有 $x_k = 150/10 = 15$ 个分量未被屏蔽为零。图 3 右侧面板对比了采用随机稀疏器并进行均匀分配的 ACTC 策略与 KKT 优化解的性能。再一次，优化策略提升了学习性能；与均匀分配所达到的性能相比，均方误差降低了约3 dB。请注意，与所有智能体都传输向量的全部分量的基准 ATC 策略相比，这一结果是在保证降维因子为 $X/M = 1/2$ 的前提下实现的。

V. 结论

本研究聚焦于通信约束下的在线分布式回归问题。我们通过建立每个智能体均方误差的上界，刻画了适应-压缩-组合（ACTC）扩散策略的性能。分析表明，所有智能体以协调的方式进行学习，并达到相同的稳态均方误差。该误差由两部分组成：一部分刻画经典的（即未进行数据压缩的）ATC 扩散策略，另一部分则是额外的压缩损失。所得解析解使我们能够根据网络拓扑和梯度

噪声刻画压缩损失，并能够根据智能体的数据质量及其在网络中的中心性，优化智能体之间的比特分配。为此，我们设计了一套优化流程，并制定了一种在线策略，用于学习优化流程所需的参数。实验分析表明，该方法显著节省了通信资源。未来还有若干值得探索的方向，包括：刻画其他代价函数下的均方误差性能；扩展至其他压缩算子模型和可变速率量化方案；以及扩展至具有子空间约束和多任务设置的学习问题。

APPENDIX A
USEFUL RECURSIONS

在以下推导中，给定一组 $M\times 1$ 向量$\boldsymbol{x}_{k,\,i}$，我们通常会使用扩展$MN\times 1$ 向量来进行分析：

$$\boldsymbol{x}_i \triangleq \mathrm{col}\{\boldsymbol{x}_{1,i},\boldsymbol{x}_{2,i},\ldots,\boldsymbol{x}_{N,i}\}. \tag{56}$$

为了评估迭代序列$\boldsymbol{w}_{k,\,i}$的均方误差，先关注压缩后的迭代序列$\boldsymbol{q}_{k,\,i}$的均方误差会更为方便，尤其是关注中心化向量$\widetilde{\boldsymbol{q}}_{k,\,i}=\boldsymbol{q}_{k,\,i}-w^o$；其对应的 $MN\times 1$ 扩展向量记为 $\widetilde{\boldsymbol{q}}_i$。根据 ACTC 递推式 (9)，不难验证 $\widetilde{\boldsymbol{q}}_i$ 满足以下递推关系（计算细节见第 D-A 节）：

$$\widetilde{\boldsymbol{q}}_i = \mathcal{B}\,\widetilde{\boldsymbol{q}}_{i-1} + \zeta\,\boldsymbol{e}_i - \mu\,\zeta\,\boldsymbol{s}_i, \tag{57}$$

其中， $MN\times MN$ 矩阵 $\mathcal{B}$ 定义于式 (108) 中。在式 (105a) 中，我们还可以得到以下网络级量 $\boldsymbol{\delta}_i$的递推关系

$$\boldsymbol{\delta}_i = [(I_{MN}-\mu\,\mathcal{H})\,\mathcal{A}^{\top}-I_{MN}]\widetilde{\boldsymbol{q}}_{i-1}-\mu\,\boldsymbol{s}_i, \tag{58}$$

其中，矩阵 $\mathcal{H}$ 和 $\mathcal{A}$ 定义于式 (106) 中。

它　同样有必要引入扩展的条件梯度噪声协方差矩阵

$$\begin{aligned}\mathcal{R}_s(\boldsymbol{w}_i) &\triangleq \mathbb{E}\left[\boldsymbol{s}_i\boldsymbol{s}_i^{\top}|\boldsymbol{q}_{0:i-1},\boldsymbol{\psi}_{0:i-1}\right]\\ &= \mathrm{blkdiag}\{R_{s,1}(\boldsymbol{w}_{1,i-1}),\ldots,R_{s,N}(\boldsymbol{w}_{N,i-1})\},\end{aligned} \tag{59}$$

其中，最后一个等式由（30）得出。

同样，我们　为每个智能体引入条件压缩误差协方差矩阵，即　　y,

$$R_{q,k}(\boldsymbol{\delta}_{k,i}) = \mathbb{E}\left[\boldsymbol{e}_{k,i}\boldsymbol{e}_{k,i}^{\top}|\boldsymbol{\delta}_i\right], \tag{60}$$

及其扩展版本

$$\begin{aligned}\mathcal{R}_q(\boldsymbol{\delta}_i) &\triangleq \mathbb{E}\left[\boldsymbol{e}_i\boldsymbol{e}_i^{\top}|\boldsymbol{\delta}_i\right]\\ &= \mathrm{blkdiag}\{R_{q,1}(\boldsymbol{\delta}_{1,i}),\ldots,R_{q,N}(\boldsymbol{\delta}_{N,i})\},\end{aligned} \tag{61}$$

其中，由于式 (24)，最后一个等式成立。

附录 B

定理 1 的证明

证明：证明主要分为三个步骤：

(i) 在 B-A 节中，我们说明，压缩状态 $\widetilde{\boldsymbol{q}}_i$ 的演化可以更方便地由一个近似模型$\boldsymbol{q}_{a,i}$替代。

(ii) 在 B-B 节中，我们将智能体迭代过程之间的相互作用与近似模型的相关矩阵 $\boldsymbol{q}_{a,i}$ 联系起来，以证明 (39)。

(iii) 在 B-C 节中，我们考察梯度噪声和压缩误差，从而明确刻画相关矩阵的迹$\boldsymbol{q}_{a,i}$ ，以证明 (40)。

A. 近似模型

通过采用网络坐标变换（详见附录 D-C），我们可以用协调分量 $\bar{\boldsymbol{q}}_i$ 和误差分量 $\check{\boldsymbol{q}}_i$来表示压缩迭代量 $\widetilde{\boldsymbol{q}}_i$ 的演化。[33] 中对 ACTC 策略的稳定性分析表明，协调分量 $\bar{\boldsymbol{q}}_i$ 起主导作用，而误差分量 $\check{\boldsymbol{q}}_i$ 则充当扰动项——参见附录 D-E。基于这些原因，在 (127) 中研究 $\bar{\boldsymbol{q}}_i$ 的演化时，关注协调分量的递推关系是有意义的：

$$\begin{aligned}\bar{\boldsymbol{q}}_i = &\left(I_M - \mu\zeta\, G_{11}\right)\bar{\boldsymbol{q}}_{i-1} - \mu\zeta\, G_{12}\check{\boldsymbol{q}}_{i-1}\\ &+ \zeta(\pi^\top \otimes I_M)\boldsymbol{e}_i - \mu\zeta(\pi^\top \otimes I_M)\boldsymbol{s}_i,\end{aligned} \tag{62}$$

其中，矩阵 G_{11} 和 G_{12} 定义见 (122)。忽略 (62) 中的误差分量 $\check{\boldsymbol{q}}_{i-1}$ 后，可得近似递推关系：

$$\begin{aligned}\boldsymbol{q}_{a,i} = &\left(I_M - \mu\zeta\, G_{11}\right)\boldsymbol{q}_{a,i-1}\\ &+ \zeta(\pi^\top \otimes I_M)\boldsymbol{e}_i - \mu\zeta(\pi^\top \otimes I_M)\boldsymbol{s}_i.\end{aligned} \tag{63}$$

接下来我们将说明，每个单独向量 $\boldsymbol{q}_{k,\,i}$ 的性能都可以通过 $\boldsymbol{q}_{a,\,i}$来近似。由于我们关注的是均方误差性能，因此一个基本的描述量是 $\boldsymbol{q}_{a,\,i}$的相关矩阵：

$$\Pi_{a,i} \triangleq \mathbb{E}\left[\boldsymbol{q}_{a,i}\boldsymbol{q}_{a,i}^\top\right]. \tag{64}$$

利用（63）展开（64）中的定义，结合（22）和（30），并使用（59）和（61），我们得到

$$\begin{aligned}\Pi_{a,i} = &\, D\,\Pi_{a,i-1}D\\ &+ \mu^2\zeta^2(\pi^\top \otimes I_M)\mathbb{E}\left[\mathcal{R}_s(\boldsymbol{w}_{i-1})\right](\pi \otimes I_M)\\ &+ \zeta^2(\pi^\top \otimes I_M)\mathbb{E}\left[\mathcal{R}_q(\boldsymbol{\delta}_i)\right](\pi \otimes I_M),\end{aligned} \tag{65}$$

其中 $D = (I_M - \mu\zeta G_{11})$ 是一个实对称 $M \times M$ 矩阵。注意，在 ACTC 策略均方稳定的条件下， D 是施尔稳定的 [33]。通过在全局极小点 w^o 处加上并减去由（ 59 ）给出的梯度噪声协方差矩阵，我们可以将（65）写为

$$\begin{aligned}\Pi_{a,i} &= D\,\Pi_{a,i-1}D \\ &+ \mu^2\zeta^2(\pi^\top \otimes I_M)\mathcal{R}_s(\mathbb{1}_N \otimes w^o)(\pi \otimes I_M) \\ &+ \zeta^2(\pi^\top \otimes I_M)\mathbb{E}\left[\mathcal{R}_q(\boldsymbol{\delta}_i)\right](\pi \otimes I_M) + \mu^2\zeta^2 \\ &\times (\pi^\top \otimes I_M)\mathbb{E}\left[\mathcal{R}_s(\boldsymbol{w}_{i-1}) - \mathcal{R}_s(\mathbb{1}_N \otimes w^o)\right](\pi \otimes I_M). \end{aligned} \tag{66}$$

下一个引理表明，在小 μ 范围内，可以忽略（66）式右端的最后一项，因此我们可以通过矩阵 $\Pi^o_{a,\ i}$ 近似真实的相关矩阵 Π_i ，该矩阵刻画了未受扰动的模型：

$$\begin{aligned}\Pi^o_{a,i} &= D\,\Pi^o_{a,i-1}D \\ &+ \mu^2\zeta^2(\pi^\top \otimes I_M)\mathcal{R}_s(\mathbb{1}_N \otimes w^o)(\pi \otimes I_M) \\ &+ \zeta^2(\pi^\top \otimes I_M)\mathbb{E}\left[\mathcal{R}_q(\boldsymbol{\delta}_i)\right](\pi \otimes I_M), \end{aligned} \tag{67}$$

替代出现在(66)中的矩阵$\Pi_{a,i}$ 。更具体地说，下面的引理通过证明原始模型与未扰动模型之间的等价性，表明任意两个智能体 ℓ 和 k 的压缩迭代量 $\mathbb{E}[\widetilde{\boldsymbol{q}}_{\ell,i}\widetilde{\boldsymbol{q}}_{k,i}^\top]$ 之间的相关矩阵（即扩展状态相关矩阵Π_i的每个 $(\ell,\ k)$ 块）在 $O(\mu^{3/2})$ 扰动意义下等价于$\Pi^o_{a,i}$。

引理 1 (近似压缩状态模型)。 对于足够小的步长 μ 和稳定性参数 ζ ， 只要 *ACTC* 策略是均方稳定的，就有。

$$\limsup_{i\to\infty}\left\|\Pi_i - \mathbb{1}_N\mathbb{1}_N^\top \otimes \Pi^o_{a,i}\right\| = O(\mu^{3/2}). \tag{68}$$

关系(68) 尤其意味着

$$\limsup_{i\to\infty}\left|\mathbb{E}[\widetilde{\boldsymbol{q}}_{\ell,i}^\top\widetilde{\boldsymbol{q}}_{k,i}] - \mathrm{Tr}[\Pi^o_{a,i}]\right| = O(\mu^{3/2}). \tag{69}$$

证明：参见附录 E。 ■

B. 智能体交互 — 方程(39)

引理1 揭示了相互智能体的关系可由 $\Pi^o_{a,i}$ 捕获。特别地，我们可以通过$\Pi^o_{a,i}$的迹，近似任意智能体 ℓ 和 k 的压缩迭代量之间的标量积（包括 $\ell = k$ 的情况）。我们可以直接用 (37) 中的迭代量 $\widetilde{\boldsymbol{w}}_{k,i}$ 来重新表述 (69)。事实上，根据组合步骤 (9c)，我们有 $\widetilde{\boldsymbol{w}}_{k,i} = \sum_{\ell=1}^N a_{\ell k}\widetilde{\boldsymbol{q}}_{\ell,i}$，从而得到

$$\widetilde{\boldsymbol{w}}_{\ell,i}^\top\widetilde{\boldsymbol{w}}_{k,i} = \sum_{j=1}^N\sum_{h=1}^N a_{j\ell}a_{hk}\widetilde{\boldsymbol{q}}_{j,i}^\top\widetilde{\boldsymbol{q}}_{h,i}. \tag{70}$$

由于 A是左随机的，这意味着

$$\widetilde{\boldsymbol{w}}_{\ell,i}^{\top}\widetilde{\boldsymbol{w}}_{k,i}-\mathrm{Tr}[\Pi_{a,i}^{o}]=\sum_{j=1}^{N}\sum_{h=1}^{N}a_{j\ell}a_{hk}\left(\widetilde{\boldsymbol{q}}_{j,i}^{\top}\widetilde{\boldsymbol{q}}_{h,i}-\mathrm{Tr}[\Pi_{a,i}^{o}]\right), \tag{71}$$

且（69）意味着

$$\limsup_{i\to\infty}\left|\mathbb{E}[\widetilde{\boldsymbol{w}}_{\ell,i}^{\top}\widetilde{\boldsymbol{w}}_{k,i}]-\mathrm{Tr}[\Pi_{a,i}^{o}]\right|=O(\mu^{3/2}). \tag{72}$$

现在令 $\rho_{\ell k,i}\triangleq\mathbb{E}\widetilde{\boldsymbol{w}}_{\ell,i}^{\top}\widetilde{\boldsymbol{w}}_{k,i}$。由于我们可以写成

$$|\rho_{\ell k,i}-\rho_{kk,i}|\le\left|\rho_{\ell k,i}-\mathrm{Tr}[\Pi_{a,i}^{o}]\right|+\left|\rho_{\ell k,i}-\mathrm{Tr}[\Pi_{a,i}^{o}]\right|, \tag{73}$$

由 (72) 可得

$$\limsup_{i\to\infty}|\rho_{\ell k,i}-\rho_{kk,i}|=O(\mu^{3/2}). \tag{74}$$

另一方面，我们有

$$\begin{aligned}\frac{\rho_{\ell k,i}}{\sqrt{\rho_{\ell\ell,i}\rho_{kk,i}}}-1&=\frac{\rho_{\ell k,i}-\rho_{kk,i}}{\sqrt{\rho_{\ell\ell,i}\rho_{kk,i}}}+\frac{\sqrt{\rho_{kk,i}}-\sqrt{\rho_{\ell\ell,i}}}{\sqrt{\rho_{\ell\ell,i}}}\\&=\frac{\rho_{\ell k,i}-\rho_{kk,i}}{\sqrt{\rho_{\ell\ell,i}\rho_{kk,i}}}+\frac{\rho_{kk,i}-\rho_{\ell\ell,i}}{\sqrt{\rho_{\ell\ell,i}}\left(\sqrt{\rho_{kk,i}}+\sqrt{\rho_{\ell\ell,i}}\right)}.\end{aligned} \tag{75}$$

应用上极限和下极限的性质，由（75）可得如下关系链：

$$\begin{aligned}\limsup_{i\to\infty}\left|\frac{\rho_{\ell k,i}}{\sqrt{\rho_{\ell\ell,i}\rho_{kk,i}}}-1\right|&\le\frac{\limsup_{i\to\infty}|\rho_{\ell k,i}-\rho_{kk,i}|}{\sqrt{\liminf_{i\to\infty}\rho_{\ell\ell,i}\,\liminf_{i\to\infty}\rho_{kk,i}}}\\&+\frac{\limsup_{i\to\infty}|\rho_{kk,i}-\rho_{\ell\ell,i}|}{\sqrt{\liminf_{i\to\infty}\rho_{\ell\ell,i}}\left(\sqrt{\liminf_{i\to\infty}\rho_{kk,i}}+\sqrt{\liminf_{i\to\infty}\rho_{\ell\ell,i}}\right)}\\&\le\frac{O(\mu^{3/2})}{\sqrt{\left(\kappa\mu-O(\mu^{3/2})\right)\left(\kappa\mu-O(\mu^{3/2})\right)}}\\&+\frac{O(\mu^{3/2})}{\sqrt{\left(\kappa\mu-O(\mu^{3/2})\right)}\left(\sqrt{\kappa\mu-O(\mu^{3/2})}+\sqrt{\kappa\mu-O(\mu^{3/2})}\right)}\\&=O(\mu^{1/2}).\end{aligned} \tag{76}$$

这就得到了（39）。

C. 均方误差界 — 方程(40)

由式 (72) 可知，为了表征均方误差，我们需要计算式 (67) 中 $\Pi_{a,i}^{o}$ 的迹。为方便记号，我们引入以下矩阵：

$$\begin{aligned}\bar{R}_s &= (\pi^\top \otimes I_M)\mathcal{R}_s(\mathbb{1}_N \otimes w^o)(\pi \otimes I_M)\\ &= \sum_{k=1}^{N} \pi_k^2 R_{s,k}(w^o), \end{aligned} \tag{77}$$

$$\begin{aligned}\bar{R}_{q,i} &= (\pi^\top \otimes I_M)\mathbb{E}\left[\mathcal{R}_q(\boldsymbol{\delta}_i)\right](\pi \otimes I_M)\\ &= \sum_{k=1}^{N} \pi_k^2 \mathbb{E}\left[R_{q,k}(\boldsymbol{\delta}_{k,i})\right]. \end{aligned} \tag{78}$$

展开（67）中的递推关系可得

$$\begin{aligned}\Pi_{a,i}^{o} = D^i \Pi_{a,0}^{o} D^i + \mu^2\zeta^2 \sum_{j=0}^{i} D^j \bar{R}_s D^j\\ + \zeta^2 \sum_{j=0}^{i} D^j \bar{R}_{q,i-j} D^j. \end{aligned} \tag{79}$$

右端上的第一项随 $i \to \infty$ 消失，因为 D 是施尔稳定的。现在我们刻画另外两项。

1）梯度噪声稳态：对于（79）中右端的第二项，我们可以写成

$$\begin{aligned}\mathrm{Tr}\left[\sum_{j=0}^{\infty} D^j \bar{R}_s D^j\right] &= \mathrm{Tr}\left[\sum_{j=0}^{\infty} D^{2j} \bar{R}_s\right]\\ &= \mathrm{Tr}\left[(I_M - D^2)^{-1}\bar{R}_s\right] \end{aligned} \tag{80}$$

其中，第一个等式利用了迹算子的循环性质，而第二个等式则通过计算施尔稳定矩阵 D 的几何矩阵级数得到。回顾式 (131) 中 D 的定义：

$$\begin{aligned}I_M - D^2 &= (I_M + D)(I_M - D)\\ &= (2I_M - \mu\,\zeta\, G_{11})\,\mu\,\zeta\, G_{11}, \end{aligned} \tag{81}$$

然后

$$\begin{aligned}(I_M - D^2)^{-1} &= \left(I_M - \mu\,\zeta\frac{G_{11}}{2}\right)^{-1}\frac{G_{11}^{-1}}{2\,\mu\,\zeta}\\ &= \frac{G_{11}^{-1}}{2\,\mu\,\zeta} + \left[\left(I_M - \mu\,\zeta\frac{G_{11}}{2}\right)^{-1} - I_M\right]\frac{G_{11}^{-1}}{2\,\mu\,\zeta}, \end{aligned} \tag{82}$$

其中，在第二个等式中，我们加上并减去了单位矩阵 I_M。对于（82）右端的第二项，我们有：

$$\begin{aligned}
&\left[\left(I_M-\mu\,\zeta\frac{G_{11}}{2}\right)^{-1}-I_M\right]\frac{G_{11}^{-1}}{2\,\mu\,\zeta}\\
&=\left[\sum_{j=1}^{\infty}\left(\mu\,\zeta\frac{G_{11}}{2}\right)^{j}\right]\frac{G_{11}^{-1}}{2\,\mu\,\zeta}=\frac{1}{4}\left[\sum_{j'=0}^{\infty}\left(\mu\,\zeta\frac{G_{11}}{2}\right)^{j'}\right]\\
&=\frac{1}{4}\left(I_M-\mu\,\zeta\frac{G_{11}}{2}\right)^{-1}. \qquad (83)
\end{aligned}$$

由于迹和逆算子都是连续算子，我们知道，随着 $\mu\ \to\ 0$ $(I_M\ -\mu\,\zeta G_{11}/2)^{-1}\bar{R}_s$ 的迹收敛于 $\mathrm{Tr}[\bar{R}_s]$，这尤其意味着：

$$\mathrm{Tr}\left[\left(I_M-\mu\,\zeta\frac{G_{11}}{2}\right)^{-1}\bar{R}_s\right]=O(1). \qquad (84)$$

因此，我们可以将 (80) 写为

$$\begin{aligned}
&\mathrm{Tr}\left[\sum_{j=0}^{\infty}D^j\bar{R}_sD^j\right]=\frac{1}{2\mu\zeta}\mathrm{Tr}\left[G_{11}^{-1}\,\bar{R}_s\right]+O(1)\\
&=\frac{1}{4\,\mu\,\zeta}\mathrm{Tr}\left[\left(\sum_{k=1}^{N}\alpha_k\pi_kR_{u,k}\right)^{-1}\left(\sum_{k=1}^{N}\pi_k^2R_{s,k}(w^o)\right)\right]\\
&+O(1), \qquad (85)
\end{aligned}$$

其中，第一个等式是在将（83）代入（82）后由（80）得到的，而第二个等式则由（77）和（123）得到。

*2)*压缩误差稳态：将 (78) 代入 (79) 右端的第三项，我们有

$$
\begin{aligned}
&\mathrm{Tr}\left[\sum_{j=0}^{i} D^j \bar{R}_{q,i-j} D^j\right] \\
&= \mathrm{Tr}\left[\sum_{j=0}^{i} D^j \left(\sum_{k=1}^{N} \pi_k^2 \mathbb{E}\left[R_{q,k}(\boldsymbol{\delta}_{k,i-j})\right]\right) D^j\right] \\
&\overset{(a)}{=} \sum_{j=0}^{i}\sum_{k=1}^{N} \pi_k^2 \,\mathrm{Tr}\left[D^{2j}\mathbb{E}\left[R_{q,k}(\boldsymbol{\delta}_{k,i-j})\right]\right] \\
&\overset{(b)}{\le} \sum_{j=0}^{i}\sum_{k=1}^{N} \pi_k^2 \sum_{m=1}^{M} \lambda_m\big(D^{2j}\big)\lambda_m\big(R_{q,k}(\boldsymbol{\delta}_{k,i-j})\big) \\
&\overset{(c)}{\le} \sum_{j=0}^{i}(1-\mu\,\zeta\,\nu)^{2j} \sum_{k=1}^{N} \pi_k^2 \underbrace{\sum_{m=1}^{M} \lambda_m\big(R_{q,k}(\boldsymbol{\delta}_{k,i-j})\big)}_{=\mathrm{Tr}[R_{q,k}(\boldsymbol{\delta}_{k,i-j})]} \\
&\overset{(d)}{\le} \sum_{j=0}^{i}(1-\mu\,\zeta\,\nu)^{2j} \sum_{k=1}^{N} \pi_k^2 \omega_k \mathbb{E}\|\boldsymbol{\delta}_{k,i-j}\|^2 \\
&\overset{(e)}{=} \sum_{j=0}^{i}(1-\mu\,\zeta\,\nu)^{2j} \mathbb{E}\|\boldsymbol{\delta}_{i-j}\|_{\Omega_\pi}^2,
\end{aligned}
\tag{86}
$$

其中，步骤 (a) 源于迹算子的线性性质和循环性质，而步骤 (b) 则源于冯·诺伊曼迹不等式（由于所涉及的矩阵均为对称矩阵，因此将奇异值替换为特征值），其中我们用 $\lambda_m(\cdot)$ 表示其矩阵自变量 [47]的第 m 个特征值。观察到 D^j 的对角元素均不超过 D^j 的最大特征值（回忆一下，D 是正定的），而最大特征值又不超过$(1-\mu\,\zeta\,\nu)^j$，其中后者由 (132) 得出，因此可得步骤 (c) 。步骤 (d) 源于 (23) 中的方差上界。为得到步骤 (e)，我们回顾：给定一个由 $M\times 1$ 个条目 x_k组成的$MN\times 1$块向量 $x=\mathrm{col}$，以及一组正标量 $\{y_k\}_{k=1}^N$，有

$$
\sum_{k=1}^{N} y_k \|x_k\|^2 = x^\top Y x = \|x\|_Y^2, \tag{87}
$$

其中 $Y=\mathrm{blkdiag}\,\{y_1 I_M, y_2 I_M, \ldots, y_N I_M\}$。计算 (86) 中关于 k 的内层求和，并且由于 $MN\times 1$ 扩展向量 $\boldsymbol{\delta}_i$ 堆叠了 $M\times 1$ 向量 $\boldsymbol{\delta}_{k,\,i}$，我们可以引入对角加权矩阵

$$
\Omega_\pi = \mathrm{blkdiag}\left\{\pi_1^2\omega_1 I_M, \ldots, \pi_N^2\omega_N I_M\right\}, \tag{88}
$$

并由 (87) 得到：

$$
\mathbb{E}\|\boldsymbol{\delta}_{i-j}\|_{\Omega_\pi}^2 = \sum_{k=1}^{N} \pi_k^2\omega_k \mathbb{E}\|\boldsymbol{\delta}_{k,i-j}\|^2. \tag{89}
$$

为了刻画(86)中的最后一项求和，我们利用微分迭代的加权范数的渐近行为。

引理 2 (差分迭代的加权能量)。 对于足够小的步长 μ 和稳定性参数 ζ，使得 *ACTC* 策略均方稳定，差分迭代 $\boldsymbol{\delta}_i$ 的 Ω_π 加权范数满足界

$$\begin{aligned}\limsup_{i\to\infty}\mathbb{E}\|\boldsymbol{\delta}_i\|_{\Omega_\pi}^2 &\le \mu^2\sum_{k=1}^{N}\pi_k^2\omega_k\mathrm{Tr}[R_{s,k}(w^o)]\\ &\quad+\mu^2\zeta\, c+O(\mu^3),\end{aligned}\tag{90}$$

其中 c 是一个 与 μ 无关、取决于压缩参数 $\{\omega_k\}_{k=1}^N$ 的常数 .

证明：见附录 *F*。 ■

接下来我们观察到 tha t如果序列 $\{z_i\}$ 具有等于 z的有限上极限，则成立（见引理3 ）

$$\limsup_{i\to\infty}\sum_{j=0}^{i}\beta^j z_{i-j}\le\frac{z}{1-\beta},\qquad(0<\beta<1).\tag{91}$$

现在，我们将（91）应用于（86）中的最后一个求和，并作如下选择:

$$z_i=\mathbb{E}\|\boldsymbol{\delta}_i\|_{\Omega_\pi}^2,\qquad\beta=(1-\mu\,\zeta\,\nu)^2.\tag{92}$$

进一步利用 (90) 对 z_i的上极限进行上界估计，最终得到:

$$\begin{aligned}&\limsup_{i\to\infty}\mathrm{Tr}\left[\sum_{j=0}^{i}D^j\bar{R}_{q,i-j}D^j\right]\\ &\le\frac{\mu}{\zeta\,\nu\,(2-\mu\,\zeta\,\nu)}\sum_{k=1}^{N}\pi_k^2\omega_k\mathrm{Tr}[R_{s,k}(w^o)]\\ &+\frac{\mu\,c}{\nu\,(2-\mu\,\zeta\,\nu)}+O(\mu^2)\\ &\le\frac{\mu}{2\,\zeta\,\nu}\sum_{k=1}^{N}\pi_k^2\omega_k\mathrm{Tr}[R_{s,k}(w^o)]+\frac{\mu\,c}{2\,\nu}+O(\mu^2),\end{aligned}\tag{93}$$

其中，在第二个不等式中，我们使用了渐近展开式 $1/(2-\mu\,\zeta\,\nu)\;=1/2+O(\mu)$。

将 (85) 和 (93) 代入 (79)，并进一步利用定义 (35) 和 (36)，可得:

$$\begin{aligned}\mu\,\Delta_s-O(\mu^2)&\le\liminf_{i\to\infty}\mathrm{Tr}[\Pi_{a,i}^o]\\ &\le\limsup_{i\to\infty}\mathrm{Tr}[\Pi_{a,i}^o]\\ &\le\mu\,(\Delta_s+\Delta_\omega)+O(\mu^2),\end{aligned}\tag{94}$$

由式（72），证明完毕。 ■

附录 C

性质2的证明

证明：将（7）和（8）代入（28），并进行如下代数运算，即可得到梯度噪声的精确表达式：

$$\begin{aligned}
\boldsymbol{s}_{k,i} &= 2\alpha_k\Big\{\boldsymbol{u}_{k,i}\left(\boldsymbol{u}_{k,i}^{\top}\boldsymbol{w}_{k,i-1}-\boldsymbol{d}_{k,i}\right) \\
&\quad -\left(R_{u,k}\boldsymbol{w}_{k,i-1}-r_{du,k}\right)\Big\} \\
&\overset{(a)}{=} 2\alpha_k\Big\{(\boldsymbol{u}_{k,i}\boldsymbol{u}_{k,i}^{\top}-R_{u,k})\boldsymbol{w}_{k,i-1}-\boldsymbol{u}_{k,i}\boldsymbol{d}_{k,i}+R_{u,k}w^o\Big\} \\
&\overset{(b)}{=} 2\alpha_k\Big\{\boldsymbol{U}_{k,i}\boldsymbol{w}_{k,i-1}-\boldsymbol{u}_{k,i}(\boldsymbol{u}_{k,i}^{\top}w^o+\boldsymbol{v}_{k,i})+R_{u,k}w^o\Big\} \\
&= 2\alpha_k\Big\{\boldsymbol{U}_{k,i}(\boldsymbol{w}_{k,i-1}-w^o)-\boldsymbol{u}_{k,i}\boldsymbol{v}_{k,i}\Big\},
\end{aligned} \tag{95}$$

其中，在步骤(a) 中，我们利用了如下事实：令式（7）等于零，最终得到关系。

$$R_{u,k}w^o = r_{du,k}, \tag{96}$$

并且在步骤 (b) 中，我们调用了 (1)，并使用了定义 $\boldsymbol{U}_{k,i}=\boldsymbol{u}_{k,i}\boldsymbol{u}_{k,i}^{\top}-R_{u,k}$。由此可得

$$\begin{aligned}
&\mathbb{E}[\boldsymbol{s}_{k,i}|\boldsymbol{q}_{0:i-1},\boldsymbol{\psi}_{0:i-1}] = 2\,\alpha_k \\
&\times\Big\{\mathbb{E}[\boldsymbol{U}_{k,i}(\boldsymbol{w}_{k,i-1}-w^o)|\boldsymbol{q}_{0:i-1},\boldsymbol{\psi}_{0:i-1}] \\
&\quad -\mathbb{E}[\boldsymbol{u}_{k,i}\boldsymbol{v}_{k,i}|\boldsymbol{q}_{0:i-1},\boldsymbol{\psi}_{0:i-1}]\Big\} \\
&= \mathbb{E}[\boldsymbol{U}_{k,i}](\boldsymbol{w}_{k,i-1}-w^o)-\mathbb{E}[\boldsymbol{u}_{k,i}\boldsymbol{v}_{k,i}]=0,
\end{aligned} \tag{97}$$

其中，最后一个等式中去掉条件，是因为该对 $(\boldsymbol{u}_{k,i},\boldsymbol{v}_{k,i})$ 与截至时刻 $i-1$的所有过去变量在统计上相互独立（因为过去变量取决于过去的噪声和回归量，并且根据假设，序列 $\boldsymbol{v}_{k,i}$ 和$\boldsymbol{u}_{k,i}$ 随时间相互独立）。最后一个等式源于注意到回归量 $\boldsymbol{u}_{k,i}$与零均值噪声 $\boldsymbol{v}_{k,i}$在统计上相互独立，因此 $\mathbb{E}[\boldsymbol{U}_{k,i}]=\mathbb{E}[\boldsymbol{u}_{k,i}\boldsymbol{u}_{k,i}^{\top}]-R_{u,k}=0$。利用类似的论证，可以直接证明（30）（进一步利用智能体之间的独立性）以及（31）。

T为完成证明，我们需要证明（33）。为此，利用（32）可得：

$$\begin{aligned}
&\|R_{s,k}(w)-R_{s,k}(w^o)\| \\
&= \|\mathbb{E}\left[\boldsymbol{U}_{k,i}(w-w^o)(w-w^o)^{\top}\boldsymbol{U}_{k,i}^{\top}\right]\| \\
&\overset{(a)}{\leq} \mathbb{E}\|\boldsymbol{U}_{k,i}\|^2\,\|(w-w^o)(w-w^o)^{\top}\| \\
&\overset{(b)}{\leq} \mathbb{E}\|\boldsymbol{U}_{k,i}\|^2\,\mathrm{Tr}[(w-w^o)(w-w^o)^{\top}] \\
&= \mathbb{E}\|\boldsymbol{U}_{k,i}\|^2\,\|w-w^o\|^2,
\end{aligned} \tag{98}$$

其中，在步骤 (a) 中，我们对凸函数 $\|\cdot\|$ 应用詹森不等式，并利用算子范数的次乘性；而在步骤 (b) 中，我们应用性质 $\|X\| \leq \mathrm{Tr}[X]$，该性质对任意（对称）半正定矩阵 X均成立。在（98）中令 $w = \boldsymbol{w}_{k,i-1}$ 并取期望，可得

$$\begin{aligned}\mathbb{E}\|R_{s,k}(\boldsymbol{w}_{k,i-1}) - R_{s,k}(w^o)\| \\ \leq \mathbb{E}\|\boldsymbol{U}_{k,i}\|^2\, \mathbb{E}\|\boldsymbol{w}_{k,i-1} - w^o\|^2,\end{aligned} \tag{99}$$

由于从 [33] 我们知道 上极限$_{i\to\infty}\, \mathbb{E}\|\widetilde{\boldsymbol{w}}_{k,\ i-1}\|^2 = O$（$\mu$），式（99）蕴含了（33）中的结论。

最后，我们证明(33)蕴含(34)。利用不等式 $|\mathrm{Tr}[X]| \leq M\|X\|$ (该不等式可视为冯·诺伊曼迹不等式 [47]的推论)，我们有：

$$\begin{aligned}&- M\|R_{s,k}(\boldsymbol{w}_{k,i-1})] - R_{s,k}(w^o)\| \\ &\leq \mathrm{Tr}[R_{s,k}(\boldsymbol{w}_{k,i-1}) - R_{s,k}(w^o)] \\ &\leq M\|R_{s,k}(\boldsymbol{w}_{k,i-1})] - R_{s,k}(w^o)\|.\end{aligned} \tag{100}$$

现在注意到，对（31）取期望可得：

$$\mathbb{E}\|\boldsymbol{s}_{k,i}\|^2 = \mathbb{E}\,\mathrm{Tr}[R_{s,k}(\boldsymbol{w}_{k,i-1})], \tag{101}$$

将其代入（100）可得到（34）。 ■

附录 D

有用背景结果汇集

A. 网络误差动态系統

ACTC 扩散策略的均方性能可以通过中心化变量更方便地进行研究。

$$\widetilde{\boldsymbol{w}}_{k,i} \triangleq \boldsymbol{w}_{k,i} - w^o, \quad \widetilde{\boldsymbol{\psi}}_{k,i} \triangleq \boldsymbol{\psi}_{k,i} - w^o, \quad \widetilde{\boldsymbol{q}}_{k,i} \triangleq \boldsymbol{q}_{k,i} - w^o, \tag{102}$$

将其代入（9）可得

$$\widetilde{\boldsymbol{\psi}}_{k,i} = (I_M - \mu\, H_k)\widetilde{\boldsymbol{w}}_{k,i-1} - \mu\, \boldsymbol{s}_{k,i} \tag{103a}$$

$$\widetilde{\boldsymbol{q}}_{k,i} = \widetilde{\boldsymbol{q}}_{k,i-1} + \zeta\, \boldsymbol{Q}_k(\widetilde{\boldsymbol{\psi}}_{k,i} - \widetilde{\boldsymbol{q}}_{k,i-1}) \tag{103b}$$

$$\widetilde{\boldsymbol{w}}_{k,i} = \sum_{\ell\in\mathcal{N}_k} a_{\ell k}\widetilde{\boldsymbol{q}}_{\ell,i} \tag{103c}$$

其中我们引入了缩放海森矩阵

$$H_k = \alpha_k \nabla^2 J_k(w) = 2\,\alpha_k R_{u,k}, \tag{104}$$

并且，在第一步中，我们使用了（7）、（27）、（28）和（96）。（103）的一个等价表述是

$$\boldsymbol{\delta}_i = [(I_{MN} - \mu\,\mathcal{H})\,\mathcal{A}^\top - I_{MN}]\widetilde{\boldsymbol{q}}_{i-1} - \mu\,\boldsymbol{s}_i \tag{105a}$$

$$\widetilde{\boldsymbol{q}}_i = \widetilde{\boldsymbol{q}}_{i-1} + \zeta\,\boldsymbol{\Omega}(\boldsymbol{\delta}_i) \tag{105b}$$

其中，我们使用了主文附录 A 中引入的 $MN\times 1$向量，并定义扩展矩阵:

$$\mathcal{A} \triangleq A \otimes I_M, \quad \mathcal{H} \triangleq \text{blkdiag}\{H_1, H_2, ..., H_N\}. \tag{106}$$

在（105b）的右端加上并减去 $\zeta\,\boldsymbol{\delta}_i$ 可得

$$\widetilde{\boldsymbol{q}}_i = \widetilde{\boldsymbol{q}}_{i-1} + \zeta\,\boldsymbol{\delta}_i + \zeta\,\boldsymbol{e}_i, \tag{107}$$

其中我们使用了 （20）中的压缩误差定义。将（105a）代入（107）得到 o

$$\widetilde{\boldsymbol{q}}_i = \underbrace{\left\{I_{MN} + \zeta\left[(I_{MN} - \mu\mathcal{H})\mathcal{A}^\top - I_{MN}\right]\right\}}_{\mathcal{B}}\widetilde{\boldsymbol{q}}_{i-1}$$

$$+\,\zeta\,\boldsymbol{e}_i - \mu\,\zeta\,\boldsymbol{s}_i. \tag{108}$$

递推式（108）描述了 ACTC 策略的压缩状态演化，据此可通过（103c）中的组合步骤获得迭代量$\widetilde{\boldsymbol{w}}_{k,\,i}$。

B. 压缩迭代量的相关矩阵

均方误差性能的表征基于对（中心化）压缩迭代量 $\widetilde{\boldsymbol{q}}_i$ 的相关矩阵进行分析，即，

$$\Pi_i \triangleq \mathbb{E}\left[\widetilde{\boldsymbol{q}}_i\widetilde{\boldsymbol{q}}_i^\top\right]. \tag{109}$$

该矩阵有助于表征任意一对智能体 ℓ 和 k 的均方误差性能，因为我们可以写出

$$\begin{aligned}
\mathbb{E}\left[\widetilde{\boldsymbol{q}}_{\ell,i}^\top\widetilde{\boldsymbol{q}}_{k,i}\right] &\overset{(a)}{=} \mathbb{E}\left[\widetilde{\boldsymbol{q}}_i^\top(E_{\ell k}\otimes I_M)\widetilde{\boldsymbol{q}}_i\right] \\
&\overset{(b)}{=} \mathbb{E}\left\{\text{Tr}\left[\widetilde{\boldsymbol{q}}_i\widetilde{\boldsymbol{q}}_i^\top(E_{\ell k}\otimes I_M)\right]\right\} \\
&\overset{(c)}{=} \text{Tr}\left\{\mathbb{E}\left[\widetilde{\boldsymbol{q}}_i\widetilde{\boldsymbol{q}}_i^\top\right](E_{\ell k}\otimes I_M)\right\} \\
&\overset{(d)}{=} \text{Tr}\left\{\Pi_i(E_{\ell k}\otimes I_M)\right\},
\end{aligned} \tag{110}$$

其中， $E_{\ell k}$是一个 $N\times N$选择矩阵，除(ℓ,k)元素外，其余元素均为零。在式 (110) 中，步骤(a) 由应用 $E_{\ell k}$的定义得到；步骤(b) 则由以下等式推出:

$$y^\top X y = \text{Tr}(yy^\top X) \tag{111}$$

该等式对任意维度相容的向量 y 和矩阵 X 均成立；步骤(c) 由迹运算和期望运算的线性性质得到；而(d) 则由定义 (109) 得到。注意，对于 $\ell = k$ ，我们可以从式 (110) 中提取智能体 k 的均方误差。

C. 网络坐标变换

通过对估计迭代量进行坐标变换，对 [33] 中 ACTC 扩散策略的均方分析进行了研究——参见其中的第 V.B 节，该分析基于转置组合矩阵 [46] 的 Jordan 标准分解：

$$A^{\top} \triangleq V^{-1} J_{\mathrm{tot}} V, \tag{112}$$

其中

$$V = \begin{bmatrix} \pi^{\top} \\ V_R \end{bmatrix}, \qquad V^{-1} = \begin{bmatrix} \mathbb{1}_N & V_L \end{bmatrix}, \qquad J_{\mathrm{tot}} = \begin{bmatrix} 1 & 0 \\ 0 & J \end{bmatrix}. \tag{113}$$

因此，网络坐标变换为

$$\mathcal{V} \triangleq V \otimes I_M = \begin{bmatrix} \pi^{\top} \otimes I_M \\ V_R \otimes I_M \end{bmatrix}. \tag{114}$$

回顾 [33]，中的附录 A.A，矩阵 J_{tot} 可表示为

$$J_{\mathrm{tot}} = \mathrm{blkdiag}\{J_1, J_2, \ldots, J_B\}, \tag{115}$$

其中，每个（Jordan 形）块为

$$J_n \triangleq \begin{bmatrix} \lambda_n & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_n \end{bmatrix} = \lambda_n I_{L_n} + \Theta_{L_n}, \tag{116}$$

其中， λ_n 表示矩阵 A[3]的第 n个特征值， L_n 表示第 n个块的维数， Θ_{L_n}表示一个大小为 L_n 的方阵，该矩阵的所有元素均为零，唯有主对角线上方的第一条对角线上的元素等于1。利用（116），我们可以写出

$$J = \Lambda + \Theta. \tag{117}$$

其中

$$\Lambda \triangleq \mathrm{blkdiag}\{\lambda_2 I_{L_2}, \lambda_3 I_{L_3}, \ldots, \lambda_B I_{L_B}\}, \tag{118}$$

and

$$\Theta \triangleq \mathrm{blkdiag}\{\Theta_{L_2}, \Theta_{L_3}, \ldots, \Theta_{L_B}\}. \tag{119}$$

[3]我们不失一般性地假设，特征值按绝对值降序排列。

在以下推导中，我们还将使用扩展矩阵：

$$\mathcal{J}_{\mathrm{tot}} = J_{\mathrm{tot}} \otimes I_M = \begin{bmatrix} I_M & 0 \\ 0 & \mathcal{J} \end{bmatrix}, \quad \mathcal{J} = J \otimes I_M. \tag{120}$$

D. 变换网络坐标中的表示

利用式 (114) 中的网络坐标变换，我们来考察式 (108) 中矩阵 $\mathcal{B}$ 的结构。回顾 [33] 中的式 (117)：

$$(I_{MN} - \mu\,\mathcal{H})\mathcal{A}^\top - I_{MN} = \mathcal{V}^{-1}\,(\mathcal{J}_{\mathrm{tot}} - \mu\,\mathcal{G} - I_{MN})\,\mathcal{V}, \tag{121}$$

其中

$$\mathcal{G} \triangleq \mathcal{V}\,\mathcal{H}\mathcal{A}^\top\,\mathcal{V}^{-1} = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}, \tag{122}$$

其中的矩阵已在 [33] 的附录 A.C 中考察。为后续证明做准备，我们给出 G_{11} 的显式形式，该形式引自 [33]：

$$G_{11} = \sum_{k=1}^{N} \pi_k H_k = 2\sum_{k=1}^{N} \alpha_k \pi_k R_{u,k}. \tag{123}$$

利用式 (121)，矩阵 $\mathcal{B}$ 可写为

$$\mathcal{B} = \mathcal{V}^{-1}\,[I_{MN} + \zeta(\mathcal{J}_{\mathrm{tot}} - \mu\,\mathcal{G} - I_{MN})]\,\mathcal{V}. \tag{124}$$

考察式 (120) 中矩阵 $\mathcal{J}_{\mathrm{tot}}$ 和式 (122) 中矩阵 $\mathcal{G}$ 的结构，可得

$$\mathcal{B} = \mathcal{V}^{-1} \begin{bmatrix} I_M - \mu\,\zeta\,G_{11} & -\mu\,\zeta\,G_{12} \\ -\mu\,\zeta\,G_{21} & \mathcal{J}_\zeta - \mu\,\zeta\,G_{22} \end{bmatrix} \mathcal{V}, \tag{125}$$

其中

$$\mathcal{J}_\zeta \triangleq (1-\zeta) I_{M(N-1)} + \zeta\,\mathcal{J}, \tag{126}$$

将网络坐标变换 (114) 代入 (108) 后，可得变换后的压缩状态：

$$\begin{aligned} \widehat{\boldsymbol{q}}_i &= \begin{bmatrix} \bar{\boldsymbol{q}}_i \\ \breve{\boldsymbol{q}}_i \end{bmatrix} = \begin{bmatrix} (\pi^\top \otimes I_M)\widetilde{\boldsymbol{q}}_{i-1} \\ (V_R \otimes I_M)\widetilde{\boldsymbol{q}}_{i-1} \end{bmatrix} \\ &= \begin{bmatrix} I_M - \mu\,\zeta\,G_{11} & -\mu\,\zeta\,G_{12} \\ -\mu\,\zeta\,G_{21} & \mathcal{J}_\zeta - \mu\,\zeta\,G_{22} \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{q}}_{i-1} \\ \breve{\boldsymbol{q}}_{i-1} \end{bmatrix} \\ &\quad + \zeta \begin{bmatrix} \bar{\boldsymbol{e}}_i \\ \breve{\boldsymbol{e}}_i \end{bmatrix} - \mu\,\zeta \begin{bmatrix} \bar{\boldsymbol{s}}_i \\ \breve{\boldsymbol{s}}_i \end{bmatrix}. \end{aligned} \tag{127}$$

$M \times 1$ 向量 $\bar{\boldsymbol{q}}_i$ 表示网络演化的协调分量，而 $M(N-1) \times 1$向量 $\breve{\boldsymbol{q}}_i$ 表示网络误差分量。在 (127) 中，我们还引入了梯度噪声 $\boldsymbol{s}_i$ 和压缩误差 $\boldsymbol{e}_i$的变换形式，即

$$\widehat{\boldsymbol{s}}_i = \begin{bmatrix} \bar{\boldsymbol{s}}_i \\ \breve{\boldsymbol{s}}_i \end{bmatrix} = \begin{bmatrix} (\pi^{\top} \otimes I_M)\boldsymbol{s}_i \\ (V_R \otimes I_M)\boldsymbol{s}_i \end{bmatrix}, \widehat{\boldsymbol{e}}_i = \begin{bmatrix} \bar{\boldsymbol{e}}_i \\ \breve{\boldsymbol{e}}_i \end{bmatrix} = \begin{bmatrix} (\pi^{\top} \otimes I_M)\boldsymbol{e}_i \\ (V_R \otimes I_M)\boldsymbol{e}_i \end{bmatrix}. \tag{128}$$

利用（121），在应用网络坐标变换后，（105a）中的微分迭代量的演化可写为

$$\widehat{\boldsymbol{\delta}}_i = (\mathcal{J}_{\text{tot}} - I_{MN} - \mu\,\mathcal{G})\,\widehat{\boldsymbol{q}}_{i-1} - \mu\,\widehat{\boldsymbol{s}}_i, \tag{129}$$

并且等价地，根据（120）和（122）（另见 [33] 中的附录 A.C），有

$$\begin{aligned} \widehat{\boldsymbol{\delta}}_i = \begin{bmatrix} \bar{\boldsymbol{\delta}}_i \\ \breve{\boldsymbol{\delta}}_i \end{bmatrix} &= \begin{bmatrix} (\pi^{\top} \otimes I_M)\widetilde{\boldsymbol{q}}_{i-1} \\ (V_R \otimes I_M)\widetilde{\boldsymbol{q}}_{i-1} \end{bmatrix} \\ &= \begin{bmatrix} -\mu\,G_{11} & -\mu\,G_{12} \\ -\mu\,G_{21} & \mathcal{J} - I_{M(N-1)} - \mu\,G_{22} \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{q}}_{i-1} \\ \breve{\boldsymbol{q}}_{i-1} \end{bmatrix} - \mu \begin{bmatrix} \bar{\boldsymbol{s}}_i \\ \breve{\boldsymbol{s}}_i \end{bmatrix}. \end{aligned} \tag{130}$$

最后，我们引入下面的矩阵，该矩阵将在接下来的推导中反复使用：

$$D \triangleq I_M - \mu\,\zeta\,G_{11}, \tag{131}$$

根据 (123)，这是一个实对称矩阵。当步长 μ 和稳定性参数 ζ 足够小时，即 ACTC 策略均方稳定时，矩阵 D 施尔稳定——参见 [33]，中的式 (184)，因此我们有。

$$\|I_M - \mu\,\zeta\,G_{11}\| \le 1 - \mu\,\zeta\,\nu < 1. \tag{132}$$

E. 压缩迭代量的均方误差界

[33] 中的定理 1 表明，对于足够小的 μ 和 ζ 值，ACTC 扩散在均方意义下是稳定的。具体而言，(127) 中变换后的压缩迭代量 $\widehat{\boldsymbol{q}}_i$ 的两个分量满足以下界限 [33,（271）式和（273）式]:

$$\limsup_{i\to\infty} \mathbb{E}\|\bar{\boldsymbol{q}}_i\|^2 = O(\mu), \tag{133}$$

$$\limsup_{i\to\infty} \mathbb{E}\|\breve{\boldsymbol{q}}_i\|^2 = O(\mu^2). \tag{134}$$

附录 E 引理 1 的证明

证明：利用（113）中的逆网络坐标变换，我们可以写出

$$\widetilde{\boldsymbol{q}}_i = \mathbb{1}_N \otimes \bar{\boldsymbol{q}}_i + (V_L \otimes I_M)\,\breve{\boldsymbol{q}}_i, \tag{135}$$

并应用定义（109）：

$$\begin{aligned}\Pi_i = \mathbb{1}_N\mathbb{1}_N^\top \otimes \bar{\Pi}_i &+ \mathbb{E}\left[(V_L \otimes I_M)\,\breve{\boldsymbol{q}}_i\breve{\boldsymbol{q}}_i^\top (V_L \otimes I_M)^\top\right] \\ &+ \mathbb{E}\left\{(\mathbb{1}_N \otimes \bar{\boldsymbol{q}}_i)\,[(V_L \otimes I_M)\,\breve{\boldsymbol{q}}_i]^\top\right\} \\ &+ \mathbb{E}\left\{[(V_L \otimes I_M)\,\breve{\boldsymbol{q}}_i]\,(\mathbb{1}_N \otimes \bar{\boldsymbol{q}}_i)^\top\right\},\end{aligned} \tag{136}$$

其中

$$\bar{\Pi}_i \triangleq \mathbb{E}\left[\bar{\boldsymbol{q}}_i\bar{\boldsymbol{q}}_i^\top\right]. \tag{137}$$

计算（136）的范数可得

$$\begin{aligned}&\|\Pi_i - \mathbb{1}_N\mathbb{1}_N^\top \otimes \bar{\Pi}_i\| \\ &\overset{(a)}{\leq} \left\|\mathbb{E}\left[(V_L \otimes I_M)\,\breve{\boldsymbol{q}}_i\breve{\boldsymbol{q}}_i^\top (V_L \otimes I_M)^\top\right]\right\| \\ &+ 2\left\|\mathbb{E}\left\{(\mathbb{1}_N \otimes \bar{\boldsymbol{q}}_i)\,[(V_L \otimes I_M)\,\breve{\boldsymbol{q}}_i]^\top\right\}\right\| \\ &\overset{(b)}{\leq} \mathbb{E}\|(V_L \otimes I_M)\,\breve{\boldsymbol{q}}_i\|^2 \\ &+ 2\left\|\mathbb{E}\left\{(\mathbb{1}_N \otimes \bar{\boldsymbol{q}}_i)\,[(V_L \otimes I_M)\,\breve{\boldsymbol{q}}_i]^\top\right\}\right\| \\ &\overset{(c)}{\leq} \|(V_L \otimes I_M)\|^2\mathbb{E}\|\breve{\boldsymbol{q}}_i\|^2 \\ &+ \sqrt{\mathbb{E}\|\mathbb{1}_N \otimes \bar{\boldsymbol{q}}_i\|^2\mathbb{E}\|(V_L \otimes I_M)\,\breve{\boldsymbol{q}}_i\|^2},\end{aligned} \tag{138}$$

其中，步骤(a) 由三角不等式得出，步骤(b) 由詹森不等式 $\|\mathbb{E}[\cdot]\| \leq \mathbb{E}\|\cdot\|$和范数的次乘性得出，而步骤$(c)$ 应用了柯西-施瓦茨不等式。对（138）取上极限，结合界（133）和（134），我们得出

$$\limsup_{i\to\infty} \|\Pi_i - \mathbb{1}_N\mathbb{1}_N^\top \otimes \bar{\Pi}_i\| \leq O(\mu^{3/2}). \tag{139}$$

¯现在我们着手证明， Π_i 等于相关矩阵$\Pi_{a,\,i}$ 在（64）中仅相差高阶扰动，进而由于（139）成立，这意味着 Π_i 接近于$\Pi_{a,\,i}$ 根据（127）的第一行以及（63）中$\boldsymbol{q}_{a,\,i}$ 的定义，我们看到

$$\begin{aligned}&\mathbb{E}\|\bar{\boldsymbol{q}}_i - \boldsymbol{q}_{a,i}\|^2 \\ &= \mathbb{E}\|(I_M - \mu\zeta\, G_{11})(\bar{\boldsymbol{q}}_{i-1} - \boldsymbol{q}_{a,i-1}) - \mu\zeta\, G_{12}\,\breve{\boldsymbol{q}}_{i-1}\|^2 \\ &\leq (1 - \mu\zeta\,\nu)\mathbb{E}\|\bar{\boldsymbol{q}}_{i-1} - \boldsymbol{q}_{a,i-1}\|^2 + \frac{\mu\zeta\,\sigma_{12}^2}{\nu}\mathbb{E}\|\breve{\boldsymbol{q}}_{i-1}\|^2,\end{aligned} \tag{140}$$

其中，不等式来自对凸权重 $1-\mu\zeta\nu$ 和 $\mu\zeta\nu$ 应用詹森不等式，而界 (132) 和 (152) 来自 [33]。对表达式 (140) 取上极限，我们得出

$$\begin{aligned}&\limsup_{i\to\infty}\mathbb{E}\|\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i}\|^2\\&\leq(1-\mu\zeta\nu)\limsup_{i\to\infty}\mathbb{E}\|\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i}\|^2+O(\mu^3),\end{aligned}\tag{141}$$

其中，第一个不等式由（134）得出。将右端出现的上极限移至左端，由（141）可得:

$$\limsup_{i\to\infty}\mathbb{E}\|\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i}\|^2=O(\mu^2).\tag{142}$$

根据(137)中相关矩阵的定义，我们有

$$\begin{aligned}\bar{\Pi}_i&=\mathbb{E}\left[\bar{\boldsymbol{q}}_i\bar{\boldsymbol{q}}_i^\top\right]\\&=\mathbb{E}\left[(\boldsymbol{q}_{a,i}+\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i})(\boldsymbol{q}_{a,i}+\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i})^\top\right]\\&=\mathbb{E}\left[\boldsymbol{q}_{a,i}\boldsymbol{q}_{a,i}^\top\right]+\mathbb{E}\left[\boldsymbol{q}_{a,i}(\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i})^\top\right]\\&+\mathbb{E}\left[(\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i})\boldsymbol{q}_{a,i}^\top\right]+\mathbb{E}\left[(\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i})(\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i})^\top\right]\\&=\Pi_{a,i}+\mathbb{E}\left[\bar{\boldsymbol{q}}_i(\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i})^\top\right]+\mathbb{E}\left[(\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i})\bar{\boldsymbol{q}}_i^\top\right]\\&-\mathbb{E}\left[(\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i})(\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i})^\top\right],\end{aligned}\tag{143}$$

其中，在最后一个等式中，我们使用了(64)中相关矩阵的定义。现在，我们考察(143)中最后三项的稳态行为:

$$\begin{aligned}&\limsup_{i\to\infty}\left\|\mathbb{E}\left[\bar{\boldsymbol{q}}_i(\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i})^\top\right]\right\|\\&\overset{(a)}{\leq}\limsup_{i\to\infty}\mathbb{E}\left\|\bar{\boldsymbol{q}}_i(\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i})^\top\right\|\\&\overset{(b)}{\leq}\limsup_{i\to\infty}\mathbb{E}\left[\|\bar{\boldsymbol{q}}_i\|\,\|\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i}\|\right]\\&\overset{(c)}{\leq}\limsup_{i\to\infty}\sqrt{\mathbb{E}\|\bar{\boldsymbol{q}}_i\|^2\mathbb{E}\|\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i}\|^2}=O(\mu^{3/2}),\end{aligned}\tag{144}$$

其中，步骤 (a) 是将詹森不等式应用于凸函数 $\|\cdot\|$；步骤 (b) 根据算子范数的次乘性得出；步骤 (c) 使用了柯西–施瓦茨不等式；最后一个等式使用了（133）和（141）。对于（143）中的第二项，同样的分析也成立。最后，对于第三项，我们可以写成:

$$\begin{aligned}&\limsup_{i\to\infty}\left\|\mathbb{E}\left[(\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i})(\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i})^\top\right]\right\|\\&\leq\limsup_{i\to\infty}\mathbb{E}\left\|(\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i})(\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i})^\top\right\|\\&\leq\limsup_{i\to\infty}\mathbb{E}\|\bar{\boldsymbol{q}}_i-\boldsymbol{q}_{a,i}\|^2=O(\mu^2),\end{aligned}\tag{145}$$

其中，第一个不等式再次应用了詹森不等式，第二个不等式使用了范数的次乘性，最后一步由（142）得出。计算（143）的范数，然后使用（144）和（145），我们得到

$$\limsup_{i\to\infty}\|\bar{\Pi}_i-\Pi_{a,i}\|=O(\mu^{3/2}). \tag{146}$$

现在可以写成

$$\begin{aligned}
&\limsup_{i\to\infty}\|\Pi_i-\mathbb{1}_N\mathbb{1}_N^\top\otimes\Pi_{a,i}\|\\
&=\limsup_{i\to\infty}\|\Pi_i-\mathbb{1}_N\mathbb{1}_N^\top\otimes\bar{\Pi}_i+\mathbb{1}_N\mathbb{1}_N^\top\otimes(\bar{\Pi}_i-\Pi_{a,i})\|\\
&\overset{(a)}{\leq}\limsup_{i\to\infty}\|\Pi_i-\mathbb{1}_N\mathbb{1}_N^\top\otimes\bar{\Pi}_i\|\\
&+\limsup_{i\to\infty}\|\mathbb{1}_N\mathbb{1}_N^\top\otimes(\bar{\Pi}_i-\Pi_{a,i})\|\\
&\overset{(b)}{=}\limsup_{i\to\infty}\|\Pi_i-\mathbb{1}_N\mathbb{1}_N^\top\otimes\bar{\Pi}_i\|\\
&+\limsup_{i\to\infty}\|\mathbb{1}_N\mathbb{1}_N^\top\|\|\bar{\Pi}_i-\Pi_{a,i}\|=O(\mu^{3/2}),
\end{aligned} \tag{147}$$

其中，步骤(a) 由三角不等式得出；步骤(b) 源于以下事实：由2诱导的矩阵范数等于最大奇异值，而 Kronecker 积 $X\otimes Y$的奇异值是 X 和 Y的奇异值之积，其中 X 和 Y是任意两个适当的矩阵。最后一个等式用到了 (139) 和 (146)。

引入以下记号：

$$\widetilde{\mathcal{R}}_{i-1}\triangleq\mathbb{E}\left[\mathcal{R}_s(\boldsymbol{w}_{i-1})\right]-\mathcal{R}_s(\mathbb{1}_N\otimes w^o), \tag{148}$$

由 (66) 和 (67)，$\Pi_{a,\,i}$ 与 $\Pi^o_{a,\,i}$ 之间的差异上界为

$$\begin{aligned}
&\|\Pi_{a,i}-\Pi^o_{a,i}\|\\
&\leq\|D\|^2\|(\Pi_{a,i-1}-\Pi^o_{a,i-1})\|\\
&+\left\|\mu^2\zeta^2\left(\pi^\top\otimes I_M\right)\widetilde{\mathcal{R}}_{i-1}\left(\pi\otimes I_M\right)\right\|\\
&\leq(1-\mu\,\zeta\,\nu)^2\|(\Pi_{a,i-1}-\Pi^o_{a,i-1})\|\\
&+\left\|\mu^2\zeta^2\left(\pi^\top\otimes I_M\right)\widetilde{\mathcal{R}}_{i-1}\left(\pi\otimes I_M\right)\right\|,
\end{aligned} \tag{149}$$

其中，第一个不等式由三角不等式得出，第二步中我们使用了 (132)。(149) 中右端最后一项的范数可分解如下：

$$\begin{aligned}
&\left\|\mu^2\zeta^2(\pi^\top\otimes I_M)\widetilde{\mathcal{R}}_{i-1}\,(\pi\otimes I_M)\right\| \\
&\le \mu^2\zeta^2\|\pi^\top\otimes I_M\|\times\|\widetilde{\mathcal{R}}_{i-1}\|\times\|\pi\otimes I_M\| \\
&\overset{(a)}{=} \mu^2\zeta^2\|\pi\|^2\|\widetilde{\mathcal{R}}_{i-1}\| \\
&\overset{(b)}{=} \mu^2\zeta^2\|\pi\|^2\max_{1\le k\le N}\|\mathbb{E}\left[R_{s,k}(\boldsymbol{w}_{k,i-1})-R_{s,k}(w^o)\right]\| \\
&\overset{(c)}{\le} \mu^2\zeta^2\|\pi\|^2\max_{1\le k\le N}\mathbb{E}\|R_{s,k}(\boldsymbol{w}_{k,i-1})-R_{s,k}(w^o)\|,
\end{aligned}\tag{150}$$

其中，步骤（a）源于这样一个事实：对于任意矩阵 A，$A\otimes I_M$ 的奇异值是A的（重复）奇异值；步骤（b）利用了 $\mathcal{R}_s$（$\boldsymbol{w}_{i-1}$）的对角形状；步骤（$\mathbb{1}_N\otimes w^o$）则是将詹森不等式应用于凸函数 $\|\cdot\|$。由（33）可得：

$$\limsup_{i\to\infty}\left\|\mu^2\zeta^2(\pi^\top\otimes I_M)\widetilde{\mathcal{R}}_{i-1}\,(\pi\otimes I_M)\right\|=O(\mu^3).\tag{151}$$

对（149）取上极限，然后应用（151），得到

$$\limsup_{i\to\infty}\|\Pi_{a,i}-\Pi^o_{a,i}\|=O(\mu^2).\tag{152}$$

利用推导（147）时所采用的同一组不等式链，但将$\Pi^o_{a,\,i}$ 和$\Pi_{a,\,i}$ 分别替换为$\Pi_{a,\,i}$和 Π_i，再应用（147）和（152），即可得到（68）。

下面还需证明（69）。我们首先计算压缩状态之间的相互作用与（67）中的近似模型之间的绝对差：

$$\begin{aligned}
&\left|\mathbb{E}\left[\widetilde{\boldsymbol{q}}^\top_{\ell,i}\widetilde{\boldsymbol{q}}_{k,i}\right]-\mathrm{Tr}\left[\Pi^o_{a,i}\right]\right| \\
&\overset{(a)}{=}\Big|\mathrm{Tr}\left[\Pi_i(E_{\ell k}\otimes I_M)\right] \\
&-\mathrm{Tr}\left[(\mathbb{1}_N\mathbb{1}_N^\top\otimes\Pi^o_{a,i})(E_{\ell k}\otimes I_M)\right]\Big| \\
&=\left|\mathrm{Tr}\left[(\Pi_i-\mathbb{1}_N\mathbb{1}_N^\top\otimes\Pi^o_{a,i})(E_{\ell k}\otimes I_M)\right]\right| \\
&\overset{(b)}{\le}\sum_{j=1}^{MN}\sigma_j(\Pi_i-\mathbb{1}_N\mathbb{1}_N^\top\otimes\Pi^o_{a,i})\,\sigma_j(E_{\ell k}\otimes I_M) \\
&\overset{(c)}{\le}\left\|\Pi_i-\mathbb{1}_N\mathbb{1}_N^\top\otimes\Pi^o_{a,i}\right\|\underbrace{\sum_{j=1}^{MN}\sigma_j(E_{\ell k}\otimes I_M)}_{=M},
\end{aligned}\tag{153}$$

其中，步骤(a)从相关矩阵 Π_i中提取量 $\mathbb{E}[\widetilde{\boldsymbol{q}}^\top_{\ell,i}\widetilde{\boldsymbol{q}}_{k,i}]$，如（110）所示，并利用以下等式：

$$\mathrm{Tr}\left[\Pi^o_{a,i}\right]=\mathrm{Tr}\left[(\mathbb{1}_N\mathbb{1}_N^\top\otimes\Pi^o_{a,i})(E_{\ell k}\otimes I_M)\right].\tag{154}$$

步骤(b) 使用冯·诺伊曼迹不等式 [47], , 其中$\sigma_j(X)$ 表示矩阵 X的第 j 个奇异值。最后，步骤(c) 应用 ℓ_2 诱导矩阵范数的定义，以及矩阵 $E_{\ell k}\otimes I_M$ 只有 M 个非零奇异值且它们均等于1这一事实。[4] 对（153）取上极限，由引理 1 可得（69）。

■

附录 F 引理 2 的证明

证明：考虑变换后的微分迭代 $\boldsymbol{\delta}_i$ 的演化，它可以通过将逆变换 $\mathcal{V}^{-1}$ 应用于式 (129) 中变换后迭代 $\widehat{\boldsymbol{\delta}}_i$ 的演化而得到，即

$$\boldsymbol{\delta}_i = \mathcal{V}^{-1}\underbrace{\left(\mathcal{J}_{\rm tot} - I_{MN} - \mu\,\mathcal{G}\right)}_{\mathcal{T}}\widehat{\boldsymbol{q}}_{i-1} - \mu\,\boldsymbol{s}_i. \tag{155}$$

对 (155) 应用Ω_π 加权范数并取期望，得到

$$\mathbb{E}\|\boldsymbol{\delta}_i\|^2_{\Omega_\pi} = \mathbb{E}\|\mathcal{V}^{-1}\mathcal{T}\,\widehat{\boldsymbol{q}}_{i-1}\|^2_{\Omega_\pi} + \mu^2\mathbb{E}\|\boldsymbol{s}_i\|^2_{\Omega_\pi}, \tag{156}$$

其中，等式成立是因为梯度噪声过程具有零均值性质。

利用矩阵 $\mathcal{V}^{-1}$ 和 $\mathcal{T}$ 的结构（分别见式 (113) 和 (130)），可得

$$\mathcal{V}^{-1}\mathcal{T} = \begin{bmatrix} -\mu\,\bar{G} & \breve{G}_1 - \mu\,\breve{G}_2, \end{bmatrix}, \tag{157}$$

其中定义了

$$\bar{G} \triangleq (\mathbb{1}_N\otimes I_M)\,G_{11} + (V_L\otimes I_M)\,G_{21}, \tag{158}$$

$$\breve{G}_1 \triangleq (V_L\otimes I_M)(\mathcal{J} - I_{M(N-1)}), \tag{159}$$

$$\breve{G}_2 \triangleq (\mathbb{1}_N\otimes I_M)\,G_{12} + (V_L\otimes I_M)\,G_{22}. \tag{160}$$

[4]通过观察到（$E_{\ell k}\otimes I_M$）（$E_{\ell k}\otimes I_M$）$^\top = E_{\ell k}E_{k\ell}\otimes I_M$，以及矩阵$E_{\ell k}E_{k\ell}$ 除（ℓ，ℓ）位置的元素为 1 外其余元素均为零，可以计算出 $E_{\ell k}\otimes I_M$ 的奇异值。

将（157）代入（156），并回想根据（127）有 $\widehat{\boldsymbol{q}}_i = \mathsf{col}\{\bar{\boldsymbol{q}}_i, \breve{\boldsymbol{q}}_i\}$，可得

$$
\begin{aligned}
&\mathbb{E}\|\boldsymbol{\delta}_i\|_{\Omega_\pi}^2 \\
&\overset{(a)}{\leq} 2\,\mu^2\mathbb{E}\|\bar{G}\bar{\boldsymbol{q}}_{i-1}\|_{\Omega_\pi}^2 + 2\,\mathbb{E}\|(\breve{G}_1 - \mu\,\breve{G}_2)\,\breve{\boldsymbol{q}}_{i-1}\|_{\Omega_\pi}^2 \\
&+ \mu^2\mathbb{E}\|\boldsymbol{s}_i\|_{\Omega_\pi}^2 \\
&\overset{(b)}{\leq} 2\,\mu^2\mathbb{E}\|\bar{G}\bar{\boldsymbol{q}}_{i-1}\|_{\Omega_\pi}^2 + 4\,\mathbb{E}\|\breve{G}_1\breve{\boldsymbol{q}}_{i-1}\|_{\Omega_\pi}^2 \\
&+ 4\,\mu^2\mathbb{E}\|\breve{G}_2\breve{\boldsymbol{q}}_{i-1}\|_{\Omega_\pi}^2 + \mu^2\mathbb{E}\|\boldsymbol{s}_i\|_{\Omega_\pi}^2 \\
&\overset{(c)}{=} 2\,\mu^2\|\Omega_\pi^{1/2}\bar{G}\|^2\mathbb{E}\|\bar{\boldsymbol{q}}_{i-1}\|^2 + 4\,\|\Omega_\pi^{1/2}\breve{G}_1\|^2\|\breve{\boldsymbol{q}}_{i-1}\|^2 \\
&+ 4\,\mu^2\|\Omega_\pi^{1/2}\breve{G}_2\|^2\|\breve{\boldsymbol{q}}_{i-1}\|^2 + \mu^2\mathbb{E}\|\boldsymbol{s}_i\|_{\Omega_\pi}^2,
\end{aligned}
\tag{161}
$$

其中，(a) 和(b) 是通过反复应用权重均为 $1/2$的詹森不等式得到的。对于步骤(c)，注意，对于任意正定矩阵 W，我们可以写成

$$
\begin{aligned}
\|Ax\|_W^2 &= x^\top A^\top W A x = x^\top A^\top W^{1/2} W^{1/2} A\, x \\
&= \|W^{1/2}Ax\|^2 \leq \|W^{1/2}A\|^2\,\|x\|^2,
\end{aligned}
\tag{162}
$$

这对任意矩阵 A 和维数相容的向量 x 均成立。对 (161) 取上极限，我们有 at

$$
\begin{aligned}
\limsup_{i\to\infty}\mathbb{E}\|\boldsymbol{\delta}_i\|_{\Omega_\pi}^2 \leq \limsup_{i\to\infty} 4\,\|\Omega_\pi^{1/2}\breve{G}_1\|^2\mathbb{E}\|\breve{\boldsymbol{q}}_{i-1}\|^2 \\
+\limsup_{i\to\infty}\mu^2\sum_{k=1}^{N}\pi_k^2\omega_k\mathbb{E}\|\boldsymbol{s}_{k,i}\|^2 + O(\mu^3),
\end{aligned}
\tag{163}
$$

其中，我们使用了 (133) 和 (134)，以及以下事实：为计算 $\|\Omega_\pi^{1/2}\bar{G}\|$ 和 $\|\Omega_\pi^{1/2}\breve{G}_2\|$所需的矩阵 G_{11}、G_{12}、G_{21}和 G_{22}的范数均可由常数界定，如 [2], [33]所示。此外，由于Ω_π具有对角结构，我们将梯度噪声的加权范数写成了各个范数之和。

为了考察 (163) 右端的第一项，我们可以使用表征式 [33]:

$$
\limsup_{i\to\infty}\mathbb{E}\|\breve{\boldsymbol{q}}_i\|^2 \leq \mu^2\zeta\,c',
\tag{164}
$$

其中，c' 是一个适当的常数，与 μ无关，但其值取决于压缩算子参数 $\{\omega_k\}_{k=1}^N$。c' 的精确定义可由 [33], 中给出的变换后的压缩迭代量 $\widehat{\boldsymbol{q}}_i$ 的非渐近界得到，具体请参见其中的附录 J。将 (164) 代入 (163)，我们得到

$$
\begin{aligned}
\limsup_{i\to\infty}\mathbb{E}\|\boldsymbol{\delta}_i\|_{\Omega_\pi}^2 \leq \limsup_{i\to\infty}\mu^2\sum_{k=1}^{N}\pi_k^2\omega_k\mathbb{E}\|\boldsymbol{s}_{k,i}\|^2 \\
+ \mu^2\zeta\,c + O(\mu^3),
\end{aligned}
\tag{165}
$$

其中 $c \triangleq 4\,\|\Omega_\pi^{1/2}\breve{G}_1\|^2 c'$。

将（34）代入（165）即可完成证明。 ■

APPENDIX G
AUXILIARY LEMMA

引理3（有用的几何求和）。令 $\beta \in (0,1)$。考虑一个序列 $\{z_i\}$，并令:

$$\limsup_{i\to\infty} z_i = z < \infty. \tag{166}$$

于是，有

$$\limsup_{i\to\infty} \sum_{j=0}^{i} \beta^j z_{i-j} \leq \frac{z}{1-\beta}. \tag{167}$$

证明：根据上极限的定义，对于任意 $\epsilon > 0$，存在一个索引 $i_0(\epsilon)$，使得对于所有 $i \geq i_0(\epsilon)$，都有 $z_i < z + \epsilon$。因此，我们可以写成

$$\begin{aligned}\sum_{j=0}^{i} \beta^j z_{i-j} &= \sum_{j=0}^{i-i_0(\epsilon)} \beta^j z_{i-j} + \sum_{j=i-i_0(\epsilon)+1}^{i} \beta^j z_{i-j} \\ &< (z+\epsilon) \sum_{j=0}^{i-i_0(\epsilon)} \beta^j + i_0(\epsilon)\, z_{\sup}\, \beta^{i-i_0(\epsilon)+1},\end{aligned} \tag{168}$$

其中，我们用 $z_{\sup}$ 表示有界序列 $\{z_i\}$ 的上确界。对等式 (168) 两边取上极限，可得

$$\begin{aligned}\limsup_{i\to\infty} \sum_{j=0}^{i} \beta^j z_{i-j} &\leq (z+\epsilon) \underbrace{\limsup_{i\to\infty} \sum_{j=0}^{i-i_0(\epsilon)} \beta^j}_{\leq \frac{1}{1-\beta}} \\ &+ i_0(\epsilon)\, z_{\sup} \underbrace{\limsup_{i\to\infty} \beta^{i-i_0(\epsilon)+1}}_{=0} \leq z + \epsilon,\end{aligned} \tag{169}$$

由于 ϵ 的任意性，结论得证。 ■

附录 H
资源分配问题的解

A. 随机量化器的优化分配

我们首先考虑采用随机化量化器时的资源分配问题，即在（45）中使用（51）。在这种情况下，优化问题的定义域由满足 $x \succ 0$ 的 $x \in \mathbb{R}^N$ 给出。注意，目标函数在该集合上严格凸且可微。与目标函数相关联的拉格朗日函数等于

$$\begin{aligned}\mathcal{L}(x,\lambda) = &\sum_{k=1}^{N} \pi_k^2 d_k \frac{M}{(2^{x_k}-1)^2} + \sum_{k=1}^{N} \lambda_k^- (x_{\min} - x_k) \\ &+ \sum_{k=1}^{N} \lambda_k^+ (x_k - x_{\max}) + \lambda_0 \left(\sum_{k=1}^{N} x_k - X \right),\end{aligned} \tag{170}$$

其中， $\lambda = [\lambda_0, \lambda_1^-, \lambda_2^-, \ldots, \lambda_N^-, \lambda_1^+, \lambda_2^+, \ldots, \lambda_N^+] \succeq 0$ 是与问题约束相关的拉格朗日乘子，而量 d_k 定义见（53）。

与 (45) 相关的 KKT条件，对于 $k = 1$，2，....., N为 [45]:

$$\frac{\partial \mathcal{L}(x,\lambda)}{\partial x_k} = -\frac{M\log(2)\pi_k^2 d_k 2^{x_k+1}}{(2^{x_k}-1)^3} - \lambda_k^- + \lambda_k^+ + \lambda_0 = 0, \tag{171}$$

$$\lambda_k^-(x_{\min} - x_k) = 0, \qquad \lambda_k^+(x_k - x_{\max}) = 0, \tag{172}$$

$$x_{\min} - x_k \le 0, \qquad x_k - x_{\max} \le 0, \tag{173}$$

$$\lambda_k^-, \lambda_k^+ \ge 0, \tag{174}$$

对此，我们还必须补充三个与 k无关的条件，即

$$\lambda_0 \left(\sum_{k=1}^{N} x_k - X \right) = 0, \tag{175}$$

$$\sum_{k=1}^{N} x_k - X \le 0, \tag{176}$$

$$\lambda_0 \ge 0. \tag{177}$$

为了求出满足 KKT条件的解 $(x,\ \lambda)$，我们从条件开始考虑两种不同的情况

(175).

在第一种情况下， $\sum_{k=1}^{N} x_k = X$，因此我们可以令 $\lambda_0 > 0$。接下来，我们可以根据条件 (172) 得到三种求解未知分配x_k的规则：

i) $\lambda_k^- = \lambda_k^+ = 0$，因此 x_k 是方程[5]的解

$$\lambda_0 = \frac{M\log(2)\pi_k^2 d_k 2^{x_k+1}}{(2^{x_k}-1)^3}. \tag{178}$$

ii) $\lambda_k^- = 0$, $x_k = x_{\max}$，从而

$$\lambda_k^+ = \frac{M\log(2)\pi_k^2 d_k 2^{x_{\max}+1}}{(2^{x_{\max}}-1)^3} - \lambda_0. \tag{179}$$

iii) $\lambda_k^+ = 0$, $x_k = x_{\min}$，从而

$$\lambda_k^- = \lambda_0 - \frac{M\log(2)\pi_k^2 d_k 2^{x_{\min}+1}}{(2^{x_{\min}}-1)^3}. \tag{180}$$

在第二种情况下，如果 $\sum_{k=1}^{N} x_k < X$，我们必须设置 $\lambda_0 = 0$，才能使 (175) 成立。重复导出 (178)–(181) 的相同论证，可得唯一可接受的条件为 $\lambda_k^- = 0$, $x_k = x_{\max}$，且

$$\lambda_k^+ = \frac{M\log(2)\pi_k^2 d_k 2^{x_{\max}+1}}{(2^{x_{\max}}-1)^3}. \tag{181}$$

按照规则 (178) 至 (181) 并通过数值方法选取乘子 $\lambda_0 > 0$，使其满足其余约束 (173) 至 (174)，即可得到极小化器 x。

[5]请注意，函数 $2^x/(2^x-1)^3$ 对于 $x > 0$ 单调递减，并且渐近趋于零。因此，方程 (178) 对于 $x > 0$有且仅有一个解。

B. 随机稀疏化器的优化分配

使用随机稀疏化器时的分配问题，可以再次应用 KKT 条件求解，因为式 (19) 中使用压缩参数的目标函数在满足 $x \in \mathbb{R}^N$ 且 $x \succ 0$ 的定义域内严格凸且可微。在这种情况下，拉格朗日函数等于

$$\mathcal{L}(x,\lambda) = \sum_{k=1}^{N} \pi_k^2 d_k \left(\frac{M}{x_k} - 1\right) + \sum_{k=1}^{N} \lambda_k^- (x_{\min} - x_k) + \sum_{k=1}^{N} \lambda_k^+ (x_k - x_{\max}) + \lambda_0 \left(\sum_{k=1}^{N} x_k - X\right). \tag{182}$$

我们得到相同的条件（171）—（177），其中条件（171）采取以下形式

$$\frac{\partial \mathcal{L}(x,\lambda)}{\partial x_k} = -\frac{M\pi_k^2 d_k}{x_k^2} - \lambda_k^- + \lambda_k^+ + \lambda_0 = 0, \quad \forall k. \tag{183}$$

使用解决(171)-(174) 时所采用的相同方法，当 $\sum_{k=1}^{N} x_k = X$ 和 $\lambda_0 > 0$ 时，我们最终得到以下规则

$i)$ $\lambda_k^- = \lambda_k^+ = 0 \Rightarrow x_k = \sqrt{\left(M\pi_k^2 d_k\right)/\lambda_0} ii)$
$\lambda_k^- = 0$, $x_k = x_{\max} \Rightarrow \lambda_k^+ = \left(M\pi_k^2 d_k\right)/x_{\max}^2 - \lambda_0 .iii)$
$\lambda_k^+ = 0$, $x_k = x_{\min}$，从而

$$\lambda_k^- = \lambda_0 - \left(M\pi_k^2 d_k\right)/x_{\min}^2. \tag{184}$$

在 $\sum_{k=1}^{N} x_k < X$ 和 $\lambda_0 = 0$ 的情况下，我们得到条件 $\lambda_k^- = 0$、 $x_k = x_{\max}$ 和

$$\lambda_k^+ = \left(M\pi_k^2 d_k\right)/x_{\max}^2. \tag{185}$$

数值选择使约束(173)至 (174) 得以满足的乘子 $\lambda_0 > 0$ ，给出极小化器 $x_\circ$

参考文献

[1] M. Carpentiero、V. Matta 和 A. H. Sayed，“自适应网络上的压缩分布式回归”，载于*IEEE*国际声学、语音与信号处理会议论文集（*ICASSP*），希腊罗得岛，2023年6月，第1–5页。[2] A. H. Sayed，“网络上的自适应、学习与优化”，机器学习基础与趋势，第7卷，第4–5期，第311–801页，2014年。[3] A. H. Sayed、S. Y. Tu、J. Chen、X. Zhao 和 Z. J. Towfic，“网络自适应与学习的扩散策略”，*IEEE*信号处理杂志，第30卷，第3期，第155–171页，2013年5月。[4] A. H. Sayed，“自适应网络”，*IEEE*会刊，第102卷，第4期，第460–497页，2014年4月。[5] V. Matta 和 A. H. Sayed，“自适应网络上的估计与检测”，载于协作与图信号处理，P. Djuric 和 C. Richard 编，爱思唯尔，2018年，第69–106页。[6] J. B. Predd、S. B. Kulkarni 和 H. V. Poor，“无线传感器网络中的分布式学习”，*IEEE*信号处理杂志，第23卷，第4期，第56–69页，2006年7月。[7] T. Li、A. K. Sahu、A. Talwalkar 和 V. Smith，“联邦学习：挑战、方法与未来方向”，*IEEE*信号处理杂志，第37卷，第3期，第50–60页，2020年5月。

[8] D. B. McMahan、E. Moore、D. Ramage、S. Hampson 和 B. A. Arcas，“利用去中心化数据进行深度网络的高效通信学习”，载于国际人工智能与统计会议（AISTATS）论文集，美国佛罗里达州劳德代尔堡，2017年4月，第1273–1282页。[9] M. H. DeGroot，“达成一致”，《美国统计学会杂志》，第69卷，第345期，第118–121页，1974年。[10] L. Xiao 和 S. Boyd，“用于分布式平均计算的快速线性迭代”，《系统与控制快报》，第53卷，第1期，第65–78页，2004年9月。[11] S. Boyd、A. Ghosh、B. Prabhakar 和 D. Shah，“随机化闲聊算法”，《IEEE信息论汇刊》，第52卷，第6期，第2508–2530页，2006年6月。[12] A. G. Dimakis、S. Kar、J. M. F. Moura、M. G. Rabbat 和 A. Scaglione，“用于分布式信号处理的闲聊算法”，《IEEE会刊》，第98卷，第11期，第1847–1864页，2010年11月。[13] A. Gersho 和 R. M. Gray，《向量量化与信号压缩》。施普林格，2001年。[14] Wai-Man Lam 和 A. R. Reibman，“去中心化估计系统的量化器设计”，《IEEE通信汇刊》，第41卷，第11期，第1602–1605页，1993年11月。[15] J. A. Gubner，“分布式估计与量化”，《IEEE信息论汇刊》，第39卷，第4期，第1456–1459页，1993年7月。[16] S. Marano、V. Matta 和 P. Willett，“去中心化均方误差估计的量化器渐近设计”，《IEEE信号处理汇刊》，第55卷，第11期，第5485–5496页，2007年11月。[17] M. Longo、T. D. Lookabaugh 和 R. M. Gray，“通信受限条件下去中心化假设检验的量化”，《IEEE信息论汇刊》，第36卷，第2期，第241–255页，1990年3月。[18] V. Saligrama、M. Alanyali 和 O. Savas，“具有丢包和有限容量链路的传感器网络中的分布式检测”，《IEEE信号处理汇刊》，第54卷，第11期，第4118–4132页，2006年11月。[19] Te Sun Han 和 S. Amari，“多终端数据压缩下的统计推断”，《IEEE信息论汇刊》，第44卷，第6期，第2300–2324页，1998年10月。[20] D. Alistarh、D. Grubic、J. Z. Li、R. Tomioka 和 M. Vojnovic，“QSGD：通过梯度量化和编码实现通信高效的随机梯度下降”，载于神经信息处理系统会议（NIPS）论文集，美国加利福尼亚州长滩，2017年12月，第1707–1718页。[21] S. U. Stich、J-B. Cordonnier 和 M. Jaggi，“带记忆的稀疏化随机梯度下降”，载于神经信息处理系统会议（NIPS）论文集，加拿大蒙特利尔，2018年12月，第4447–4458页。[22] J. Wangni、J. Wang、J. Liu 和 T. Zhang，“面向通信高效分布式优化的梯度稀疏化”，载于神经信息处理系统会议（NIPS）论文集，加拿大蒙特利尔，2018年12月，第1306–1316页。[23] C.-Y. Lin、V. Kostina 和 B. Hassibi，“差分量化梯度下降”，载于IEEE国际信息论研讨会（ISIT）论文集，澳大利亚维多利亚州墨尔本，2021年7月，第1200–1205页。[24] B. Widrow，“通过奈奎斯特采样理论研究粗幅度量化”，《IRE电路理论汇刊》，第3卷，第4期，第266–276页，1956年12月。[25] R. M. Gray 和 T. G. Stockham，“抖动量化器”，《IEEE信息论汇刊》，第39卷，第3期，第805–812页，1993年5月。[26] J. Wu、W. Huang、J. Huang 和 T. Zhang，“误差补偿量化随机梯度下降及其在大规模分布式优化中的应用”，载于国际机器学习会议（ICML）论文集，瑞典斯德哥尔摩，2018年7月，第5235–5333页。[27] T. Doan、S. T. Maguluri 和 J. Romberg，“随机量化下分布式梯度方法的收敛率：一种随机逼近方法”，《IEEE自动控制汇刊》，第66卷，第10期，第4469–4484页，2021年10月。[28] A. Reisidazeh、A. Mokhtari、H. Hassani 和 R. Pedarsani，“一种精确的量化去中心化梯度下降算法”，《IEEE信号处理汇刊》，第67卷，第19期，第4934–4947页，2019年8月。[29] A. Koloskova、S. U. Stich 和 M. Jaggi，“采用压缩通信的去中心化随机优化与闲聊算法”，载于国际机器学习会议（ICML）论文集，美国加利福尼亚州长滩，2019年6月，第3478–3487页。

[30] A. Koloskova、T. Lin、S. U. Stich 和 M. Jaggi，“采用任意通信压缩的去中心化深度学习”，载于《国际学习表征会议》（ICLR）论文集，埃塞俄比亚亚的斯亚贝巴，2020年4月/5月，第1–22页。[31] D. Kovalev、A. Koloskova、M. Jaggi、P. Richt´arik 和 S. U. Stich，“去中心化优化的线性收敛算法：免费发送更少比特！”，载于《国际人工智能与统计会议》（AISTATS）论文集，美国加利福尼亚州圣迭戈，2021年4月，第4087–4095页。[32] M. Carpentiero、V. Matta 和 A. H. Sayed，“采用压缩通信的自适应扩散”，载于IEEE国际声学、语音与信号处理会议（ICASSP）论文集，新加坡，2022年5月，第5672–5676页。[33] M. Carpentiero、V. Matta 和 A. H. Sayed，“通信约束下的分布式自适应学习”，在线发布于arXiv:2112.02129 [cs.LG]。[34] J. Chen 和 A. H. Sayed，“自适应网络的学习行为——第一部分：瞬态分析”，《IEEE信息论汇刊》，第61卷，第6期，第3487–3517页，2015年6月。[35] J. Chen 和 A. H. Sayed，“自适应网络的学习行为——第二部分：性能分析”，《IEEE信息论汇刊》，第61卷，第6期，第3518–3548页，2015年6月。[36] X. Zhao、S.-Y. Tu 和 A. H. Sayed，“不完善信息交换和非平稳数据条件下网络上的扩散自适应”，《IEEE信号处理汇刊》，第60卷，第7期，2012年4月，第3460–3475页。[37] A. H. Sayed，《数据推断与学习》，3卷，剑桥大学出版社，2022年。[38] R. Nassif、S. Vlaski、C. Richard、J. Chen 和 A. H. Sayed，“图上的多任务学习：一种分布式流式机器学习方法”，《IEEE信号处理杂志》，第37卷，第3期，第14–25页，2020年5月。[39] N. Michelusi、G. Scutari 和 C. -S. Lee，“具有线性收敛性的分布式算法有限比特量化”，《IEEE信息论汇刊》，第68卷，第11期，第7254–7280页，2022年11月。[40] Y. Wu 和 S. Verd´u，“R´ényi信息维数：近乎无损模拟压缩的基本极限”，《IEEE信息论汇刊》，第56卷，第8期，第3721–3748页，2010年8月。[41] G. Alberti、H. B¨olcskei、C. De Lellis、G. Koliander 和 E. Riegler，“无损模拟压缩”，《IEEE信息论汇刊》，第65卷，第11期，第7480–7513页，2019年11月。[42] R. Nassif、S. Vlaski、M. Antonini、M. Carpentiero、V. Matta 和 A. H. Sayed，“子空间约束下去中心化学习的有限比特量化”，载于《欧洲信号处理会议》（EUSIPCO）论文集，塞尔维亚贝尔格莱德，2022年8月/9月，第1851–1855页。[43] B. Bollob´ás、O. Riordan、J. Spencer 和 G. Tusn´ády，“无标度随机图过程的度序列”，《随机结构与算法》，第18卷，第3期，第279–290页，2001年5月。[44] B. Bollob´ás 和 O. Riordan，“关于无标度随机图的数学结果”，载于《图与网络手册：从基因组到互联网》，威利出版社，英国霍博肯，2003年。[45] S. Boyd 和 L. Vandenberghe，《凸优化》，剑桥大学出版社，2004年。[46] R. A. Horn 和 C. R. Johnson，《矩阵分析》，剑桥大学出版社，1985年。[47] R. Bhatia，《矩阵分析》，施普林格，1997年。