面向多用户 MIMO 系统的基于优化的用户调度框架

Victoria Palhares 和 Christoph Studer

arXiv:2603.02998v1 [cs.IT] 3 Mar 2026

摘要—**资源分配是多用户（MU）多输入多输出（MIMO）无线系统中的关键因素，有助于为所有用户设备（UE）提供高质量的服务。在拥塞场景中，用户设备调度可将用户设备分布到不同的时间、频率或空间资源上，以减轻用户设备间干扰。许多现有的用户设备调度方法依赖贪心算法，无法从全局角度处理资源分配问题。本文提出了一种适用于多用户 MIMO 无线系统的用户设备调度框架，可近似求解一个从全局角度处理调度的非凸优化问题。该用户设备调度框架确定在给定资源时隙中应同时传输的用户设备子集，并且具有灵活性，因为它（i）支持多种目标函数（例如，均衡后均方误差、容量和可实现和速率），且（ii）能够精确控制用户设备应占用的最少和最多资源数。我们展示了该用户设备调度框架在毫米波大规模多用户 MIMO 和 sub-6 GHz 无蜂窝大规模多用户多输入多输出系统中的有效性，并表明与现有调度算法相比，该框架性能更优，同时接近穷举搜索的性能。**

索引术语—**无蜂窝、大规模多用户（MU）多输入多输出（MIMO）、毫米波（mmWave）、优化、资源分配、调度。**

I. 引言

毫米波（mmWave）和无蜂窝大规模多用户（MU）多输入多输出（MIMO）无线系统是下一代无线系统中极具前景的关键技术 [2], [3]。将毫米波与大规模多用户多输入多输出相结合，不仅有望获得较大的波束成形增益，以对抗毫米波频段的高路径损耗，还能在同一时频资源内向多个用户设备（UE）实现高带宽数据传输。在无蜂窝系统中，通过密集部署执行联合基带处理的接入点（AP），可以获得优于传统蜂窝网络的覆盖 [4]。

为了最大化网络中所有用户设备的服务质量（QoS），资源分配策略，例如功率控制和用户设备调度是必要的。功率控制可缓解向同一基站（BS）传输的用户设备之间的远近问题。用户设备调度会在时间、频率和/或空间上分配用户设备的请求，以减少多个用户设备的信道冲激响应相似时场景中的用户设备间干扰——在许多用户设备于相近位置传输的拥塞场景中，就可能出现这种情况。

用户设备调度领域的文献广泛关注贪心算法，即每次选择一个用户设备加入已调度用户设备集合 [5]–[23]。这类方法 (i) 在跨多个资源时隙调度用户设备，或对每个资源块中的用户设备数量施加上限和下限时，难以应对；(ii) 通常仅适用于一种特定的成本函数；(iii) 未能从全局角度考虑用户设备调度问题，因此与对所有可能调度方案进行穷举搜索 (ES) 相比，可能导致服务质量次优。

A. 贡献

我们提出了一种适用于多用户多输入多输出无线系统的新型用户设备调度框架，以联合方式在所有用户设备和所有时隙上全局求解用户设备调度问题。我们首先将用户设备调度建模为一个离散优化问题，该问题支持多种成本函数，例如后线性最小均方误差（LMMSE）均衡均方误差（MSE）、信道容量以及后LMMSE均衡可达速率和。此外，我们的框架能够精确控制每个用户设备应占用的资源数量下限和上限。为了高效地求解离散用户设备调度问题的近似解，我们首先将组合优化问题松弛为一个非凸但光滑的优化问题，并使用前向-后向分裂（FBS） [24]对其进行近似求解。为了应用前向-后向分裂，我们为所提出的每个目标函数提供梯度以及近端算子。在我们的框架中，近端算子对约束集执行正交投影；该约束集被定义为两个由不等式描述的单纯形的交集。用户设备调度问题中的不等式表示用户设备应占用的资源数量下限和上限。为了推导能够找到位于这两个单纯形交集内解的近端算子，我们提出了一种基于卡鲁什-库恩-塔克（KKT）条件计算该投影的算法。在完成前向-后向分裂后，我们提出了一种量化算法，在确保满足约束的同时强制得到二值解。我们使用来自商业光线追踪器的毫米波大规模多用户多输入多输出和6 GHz以下无蜂窝大规模多用户多输入多输出系统的信道向量，来展示我们方法的有效性。

VP就职于德国斯图加特的诺基亚贝尔实验室；CS就职于瑞士苏黎世联邦理工学院；电子邮件：victoria.palhares@nokia-bell-labs.com, studer@ethz.ch。VP和CS的工作部分得到苏黎世联邦理工学院研究基金的支持。我们感谢Haochuan Song和Seyed Hadi Mirfarshbafan分别就前向-后向分裂和毫米波信道进行的讨论，并感谢Gian Marti和Sueda Taner对本文的评论。我们还感谢Remcom提供Wireless InSite射线追踪软件的许可证。本文的简短版本已在IEEE SPAWC 2022 [1]上发表。新贡献如下：(i) 通过不等式约束改进对用户设备活动的控制，(ii) 一个新的目标函数，(iii) 所有所提目标函数的详细梯度，(iv) 带不等式约束的单纯形投影的Karush-Kuhn-Tucker条件的详细推导，以及(v) 在无小区场景中的性能评估。Th

e MATLAB code will be released on GitHub after the review process.

我们的结果表明，与一系列基线方法相比，所提出的框架在每用户设备未编码比特错误率（BER）、每用户设备硬输出互信息（HMI）、每用户均方误差（MSE）和每用户设备可达速率方面均表现更优，并且通常能够非常接近对所有可能调度方案执行穷举搜索（ES）时的性能。

B. 符号说明

大写和小写粗体字母分别表示矩阵和向量。我们用 $A_{i,j}$ 表示$\mathbf{A}$ 的第 i 行、第 j 列元素，用$\mathbf{a}_j$ 表示$\mathbf{A}$ 的第 j 列，并用 a_i 表示向量$\mathbf{a}$ 的第 i 个元素。我们将$\mathbf{I}_M$、 $\mathbf{1}_{L\times M}$和 $\mathbf{0}_{L\times M}$分别定义为 $M \times M$ 单位矩阵、 $L \times M$ 全一矩阵和$L \times M$ 全零矩阵。上标 $(\cdot)^{\mathrm{H}}$表示共轭转置。主对角线上元素为$\mathbf{a}$ 的对角矩阵记为 $\mathbf{diag(a)}$；矩阵$\mathbf{A}$ 的行列式记为 $\mathbf{det(A)}$。绝对值、欧几里得范数和弗罗贝尼乌斯范数分别记为 $|\cdot|$、 $\|\cdot\|_2$和 $\|\cdot\|_{\mathrm{F}}$。e期望记为 $\mathbb{E}[\cdot]$，而 $\leq$ 表示逐元素小于等于。

C. 论文结构

本文其余部分安排如下。第二节回顾相关研究。第三节介绍系统模型和调度问题。第四节介绍调度框架。第五节列出目标函数及其梯度。第六节详细介绍约束集中的正交投影。第七节介绍所提出的量化算法。第八节展示毫米波和无小区大规模多用户多输入多输出场景的仿真结果。第九节讨论局限性和未来研究方向。第十节作出总结。

II. 相关先前研究

用户设备调度已在大规模多用户多输入多输出系统 [16], [17], [22], [25]–[31], 毫米波通信 [5]–[15], [32]–[39], 无蜂窝网络 [18]–[21], [23],[40]–[45], 以及非正交多址接入 (NOMA) [6],[9], [12], [15], [46]中得到广泛研究。本文聚焦于全数字大规模多用户多输入多输出、毫米波和无蜂窝网络中的用户设备调度。下文中，“用户设备集合”指分配至某个时隙的用户设备簇、组、分区或用户设备总体。

A. 贪心调度算法

在大规模多用户多输入多输出系统中，我们重点介绍 [16], [17] 中的贪心用户设备调度算法。文献 [16] 提出了半正交用户选择 (SUS) 方法，通过近似正交的信道构建用户设备集合；文献 [17] 则将 SUS 扩展至分布式多输入多输出。

在毫米波系统中，我们重点介绍基于信道结构的调度（CSS）、贪心最大和速率（贪心）以及基于弦距离的调度（弦距离调度） [5]。在 CSS 中，首先将信干噪比（SINR）最大的用户设备加入集合 $\mathcal{S}$，然后通过半正交性对其进行优化 $\mathcal{K}$。弦距离调度将用户设备分隔在角域中。 [6]–[15]还提出了其他贪心毫米波方法。

在无小区网络中，我们重点介绍 [18]–[20]，中的贪心方法，其中 [18], [19] 依次调度小规模用户设备子集以最大化和速率； [20] 中提出的方法会排除病态用户设备。

总的来说，上述所有方法都通过迭代选择用户设备，以贪心方式优化目标函数，但往往会陷入局部最优。相比之下，我们的框架通过单个优化问题全局求解用户设备调度问题，联合确定哪些用户设备应在每个时隙中传输——这种方法往往能得到出色的用户设备调度方案，并优于贪心方法（见第八节）。

B. 毫米波中的用户设备与波束关联

在毫米波系统中，毫米波频段波传播的高度方向性意味着，许多先前研究所称的“用户设备调度[1]”实际上是用户设备与波束关联 [6]–[12], [32], [35]–[37]。相比之下，我们关注的是用户设备与时隙关联。我们的框架可以扩展到用户设备与波束关联，这留待未来研究。

C. 以用户设备为中心的无小区网络

在无小区网络中，每个用户设备通常只靠近少数几个接入点，而远离其他接入点，从而导致信道稀疏性 [47]。先前的研究提出了以用户设备为中心的架构 [48]，利用这一特性，让用户设备仅向选定的接入点传输数据。在 [43], [44],中，作者针对以用户设备为中心的无小区大规模多用户多输入多输出系统的下行链路，构建了联合用户设备调度、功率分配和波束成形问题。在这项工作中，首先将每个用户设备分配给特定的接入点，为每个接入点形成用户设备集合；随后，各接入点在其天线数量的限制下独立进行调度。在 [21],中，作者提出了一种贪心的联合接入点-用户设备聚类策略，以最大化同时采用时间调度和频率调度的无小区可见光网络的吞吐量。其他频率调度方法包括 [23], [41],，分别旨在减轻用户设备集间干扰和最大化吞吐量。

相比之下，我们考虑的是一种完全协作的无小区场景，其中所有用户设备都与所有接入点通信，并在全部接入点天线和用户设备之间进行全局调度。此外，我们专注于时间调度，将频率调度留待未来研究。

D. 其他调度算法

虽然大多数先前研究依赖贪心算法，但也有一些基于机器学习（ML）的例外 [30], [31],[33], [36]。在 [30], [31], 中，信道状态信息（CSI）被投影到低维嵌入中，并根据相似性进行聚类。在 [36], 中，自监督神经网络根据方位角相似性，将接收相同数据的用户设备划分为由同一波束服务的子集。在 [33], 中，通过 k-均值聚类进行用户设备分类，以最小化用户设备间干扰。

作为贪心方法和基于机器学习的方法的一种低复杂度替代方案，参考文献 [38] 提出了低保真调度（LoFi，LoFi++）。LoFi 会进行多次随机调度试验，并根据给定目标选出最佳方案，而 LoFi++

[1]用户设备调度可用于不同场景，例如时间调度、频率调度，或将用户设备调度到特定波束。

则通过将最关键的用户设备重新分配到另一组来测试性能是否提升，从而引入类似贪心的行为。

这类基于机器学习的方法和启发式方法无法对资源分配进行明确而精确的控制，因为它们无法对每个用户设备的资源设置上下界。相比之下，我们的框架提供了精确的控制和灵活性，例如，可以指定每个用户设备对应的时隙数，以及每个时隙中的用户设备数。

E. 资源分配

如上所述，大规模多用户多输入多输出、毫米波和无蜂窝领域的大多数调度算法在资源分配方面缺乏灵活性。许多算法无法保证每个用户设备都属于至少一个用户设备集合 [5], [8], [13]–[20],[22], [23], [25], [28], [30], [32], [35], [42], [45]。此外，大多数方法无法限制每个集合中的用户设备数以及每个用户设备所属集合数的最小值和最大值 [5]–[23], [25]–[33], [35]–[46]。最后，一些方法不允许用户设备同时属于多个集合（例如，在多个时隙中获得服务） [6], [7], [9], [11], [12], [15], [21], [23], [26], [29]–[31], [33]–[36], [38], [39], [46]。

相比之下，我们的框架会限制每个集合中的用户设备数量以及每个用户设备所属集合的数量，并确保每个用户设备至少属于一个集合，同时允许用户设备集合之间存在交集，因此用户设备可以按需属于多个集合。

F. 成本函数

除了资源分配之外，大规模多用户多输入多输出、毫米波和无蜂窝领域的许多调度算法缺乏目标函数灵活性。现有方法通常只处理问题的一个方面：有些方法强制实现用户设备正交性 [5], [16], [17], ，而另一些则优化特定指标，例如最大化信干噪比 [5], [38], 可达速率 [5]–[8], [13], [18]–[21], [23], [25], [28], [30], [31], [35],[42]–[45], 能效 (EE) [9], [12],频谱效率 (SE) [12], [15], [26], 信道容量 [29], 信道能量[11], [14], [46],，或最小化干扰、误码率和信道相关性 [10], [22], [27], [33], [39]–[41]。

相比之下，我们基于优化的用户设备调度框架具有通用性，因为它支持任何可用用户设备集合表示的可微成本函数。为展示其灵活性，我们提出一种基于 LMMSE均衡器的均衡后均方误差目标函数，并同时采用传统的信道容量和可达速率目标函数。

G. 全数字基站架构

除了依赖贪心算法之外，大多数毫米波调度研究都关注混合模拟-数字 [5], [7], [8],[10]–[15], [32]–[37], [39] 或全模拟 [6], [9] 波束成形，其中每个基站的数据转换器数量少于天线数量，并在模拟域中完成部分波束成形。

相比之下，我们考虑全数字基站和接入点，其中每根天线均配备一对模数转换器，用于对同相和正交基带信号进行采样。这类架构在简化基带处理任务（例如信道估计、数据检测、多用户波束成形和用户设备调度）的同时，可以达到与混合系统相当的能效 [49]–[51] 。

III. 预备知识

A. 系统模型

我们考虑全数字大规模多用户多输入多输出系统的上行链路，其中 U 个单天线用户设备向一台配备 B 根接收天线的全数字基站发送数据，每根接收天线均配有专用的一对数据转换器。根据场景的不同，接收天线采用不同的部署方式。在毫米波场景中，我们考虑一台接收基站（$L = 1$），其配备一个具有 n_{A}根天线的均匀线性阵列（ULA）。在6 GHz以下的无小区场景中，我们考虑 L 个接入点，每个接入点配备一个具有 n_{A}根天线的均匀线性阵列。在这两种情况下，$B = Ln_{\mathrm{A}}$ 均表示接收天线的数量。我们考虑一个块衰落场景，其中包含 $t = 1, \ldots, T$个时隙，信道在频率上保持平坦[2]，并采用标准的输入输出关系：

$$\mathbf{y}[t] = \mathbf{H}\mathbf{s}[t] + \mathbf{n}[t]. \tag{1}$$

此处，$\mathbf{y}[t] \in \mathbb{C}^B$ 表示时隙 t中的接收向量，$\mathbf{s}[t] \in \mathbb{C}^U$ 包含所有 U 个用户设备的发送符号，而$\mathbf{n}[t] \in \mathbb{C}^B$ 表示噪声，其各分量为方差为 N_0的独立同分布圆对称复高斯变量。式 (1) 中的有效信道矩阵 $\mathbf{H} = \sqrt{\frac{\rho_u}{N_0}}\,\mathbf{H}\boldsymbol{\Delta}$ 综合了用户设备发射功率 ρ_u、噪声功率 N_0、上行链路多输入多输出信道矩阵 $\mathbf{H} \in \mathbb{C}^{B\times U}$ 以及功率控制矩阵$\boldsymbol{\Delta} = \mathrm{diag}(\delta_1, \ldots, \delta_U)$的影响，其各项由 [47] 给出

$$\delta_u^2 = \frac{1}{\|\overline{\mathbf{h}}_u\|_2^2} \min\left\{ \|\overline{\mathbf{h}}_u\|_2^2, 10^{\frac{\eta}{10}} \min_{u'=1,\ldots,U} \|\overline{\mathbf{h}}_{u'}\|_2^2 \right\}. \tag{2}$$

这里，$\overline{\mathbf{h}}_u$ 表示$\mathbf{H}$ 的第 u列，而 $\eta \geq 0$决定了接收功率最弱和最强的用户设备之间以分贝（dB）为单位的最大动态范围。

下文中，我们假设采用正交信道估计，即用户设备在专用训练阶段发送正交导频。采用这一训练方案，我们可以将信道估计矩阵建模为$\hat{}\mathbf{H} = \mathbf{H} + \mathbf{E}$，其中$\mathbf{E} \in \mathbb{C}^{B\times U}$ 是一个噪声矩阵，其元素为独立同分布的圆对称复高斯随机变量，方差为 $\frac{N_0}{U}$。在此训练方案的基础上，对于从毫米波大规模多用户多输入多输出系统中获得的信道向量，我们使用BEACHES算法 [52],，该算法对$\mathbf{H}$的每一列分别进行$\hat{}$去噪。对于从6 GHz 以下无蜂窝大规模多用户多输入多输出系统中获得的信道向量，我们则直接采用最小二乘信道估计。

B. 用户设备调度

为形式化定义用户设备调度问题，我们定义一个二值用户设备调度矩阵$\mathbf{C} \in \{0,1\}^{U\times T}$，其中$C_{u,t} = 1$和$C_{u,t} = 0$ 分别表示用户设备 u 在时隙 t 处于活动和非活动状态。此外，我们定义一个对角掩码矩阵$\mathbf{D}_{\mathbf{C}}[t] = \mathrm{diag}(\mathbf{c}_t) \in \{0,1\}^{U\times U}$，其中$\mathbf{c}_t$是$\mathbf{C}$的第$t$列。通过与$\mathbf{s}[t]$相乘，该掩码矩阵描述了哪些用户设备在时隙 t发送符号，哪些用户设备处于空闲状态。

我们的目标是开发一个框架，通过最小化给定的目标函数 $F(\mathbf{C}) \in \mathbb{R}$ 来确定用户设备调度矩阵 $\mathbf{C}$ ，同时满足规定用户设备可占用资源数量上下限的约束

[2]我们的框架可通过正交频分复用（OFDM）或正交频分多址（OFDMA）应用于频率选择性信道，这也使频率调度成为可能。

。我们将掩码矩阵的影响并入有效信道矩阵，此时有效信道矩阵取决于 t，如下所示：

$$\mathbf{H}[t] = \mathbf{H}\mathbf{D}_{\mathbf{C}}[t] = \hat{\mathbf{H}}\mathbf{D}_{\mathbf{C}}[t] - \mathbf{E}\mathbf{D}_{\mathbf{C}}[t]. \tag{3}$$

此处，$\hat{}\mathbf{H}[t] = \hat{}\mathbf{H}\mathbf{D}_{\mathbf{C}}[t]$ 和$\mathbf{E}[t] = \mathbf{E}\mathbf{D}_{\mathbf{C}}[t]$；这两个量都将用于计算第五节中所述的某些目标函数和梯度，以及第VIII-A 节中所述的性能指标。

IV. 用户设备调度框架

A. 用户设备调度作为一个优化问题

所提出的用户设备调度框架需要一个特定于应用的目标函数 $F(\mathbf{C})$，以及一组用于确定时间和空间资源利用率的约束。基于这些信息，我们希望求解以下优化问题：

$$\underset{\mathbf{C}\in\{0,1\}^{U\times T}}{\text{minimize}} \; F(\mathbf{C}) \text{ subject to } \mathbf{C} \in \mathcal{C}_U \cap \mathcal{C}_T, \tag{4}$$

以及两个约束集

$$\mathcal{C}_U = \left\{U_{\min}\mathbf{1}_{1\times T} \overset{e}{\leq} \mathbf{1}_{1\times U}\mathbf{C} \overset{e}{\leq} U_{\max}\mathbf{1}_{1\times T}\right\}, \text{ and} \tag{5}$$

$$\mathcal{C}_T = \left\{T_{\min}\mathbf{1}_{U\times 1} \overset{e}{\leq} \mathbf{C}\mathbf{1}_{T\times 1} \overset{e}{\leq} T_{\max}\mathbf{1}_{U\times 1}\right\}. \tag{6}$$

(5) 中的集合 $\mathcal{C}_U$ 确定每个时隙允许同时传输的用户设备的最小数量 $U_{\min}$ 和最大数量 $U_{\max}$；(6) 中的集合 $\mathcal{C}_T$ 确定每个用户设备允许传输的最小时隙数 $T_{\min}$和最大小时隙数 $T_{\max}$。由于用户设备调度矩阵$\mathbf{C}$具有离散性，(4) 中的优化问题属于组合优化问题。例如，在一个包含 $U = 32$ 个用户设备的场景中，若前 16 个用户设备在第一个时隙传输，其余 16 个用户设备在第二个时隙传输，穷举搜索将需要测试超过 6 亿个调度矩阵。显然，需要采用计算效率高的方法来近似求解 (4) 中的问题。

B. 问题松弛

为了得到高效的用户设备调度算法，我们将$\mathbf{C}$ 中的二值条目 $\{0,1\}$ 松弛到连续区间 $[0,1]$,，即$\mathbf{C} \in [0,1]^{U\times T}$。尽管这种松弛使我们能够使用计算效率较高的基于梯度下降的方法，但它不再保证解属于 $\{0,1\}$。为缓解这一问题，我们在松弛优化问题的目标函数中加入正则化项 $R(\mathbf{C})$，以促使$\mathbf{C}$中的条目取二值（参见第五节-D）。考虑正则化项后，我们将（4）修改为以下松弛优化问题：

$$\underset{\mathbf{C}\in[0,1]^{U\times T}}{\text{minimize}} \; F(\mathbf{C}) + R(\mathbf{C}) \text{ subject to } \mathbf{C} \in \mathcal{C}_U \cap \mathcal{C}_T. \tag{7}$$

使用 F（$\mathbf{C}$）$+ R$（$\mathbf{C}$）作为新的目标函数后，我们现在能够部署数值优化方法，高效地计算问题（7）松弛形式的二值解。由于增广目标函数通常是非凸的（即使原目标函数F（$\mathbf{C}$）在松弛域内是凸的），数值方法可能只能收敛到局部最小值。因此，为了提高解的质量，可以执行 K_{init} 随机初始化（见表 I），并使用代价最低的候选解 $\mathbf{C}^*$，其代价为 F（$\mathbf{C}^*$）。

C. 通过前向-后向分裂进行近似调度

为了近似求解集合$\mathbf{C} \in [0,1]^{U\times T}$对应的（7），我们采用 FBS [53]。该方法在迭代 $i = 1, 2, \ldots, I_{\max}$ 时执行以下步骤：

$$\mathbf{C}^{(i+1)} = \underset{\mathcal{X}_{\mathcal{C}_U\cap\mathcal{C}_T}}{\text{prox}} \Big(\mathbf{C}^{(i)} - \tau^{(i)}\big(\nabla F(\mathbf{C}^{(i)}) + \nabla R(\mathbf{C}^{(i)})\big)\Big). \tag{8}$$

这里，$\text{prox}^{\mathcal{X}_{\mathcal{C}_U\cap\mathcal{C}_T}}(\cdot)$ 是集合 $\mathcal{C}_U \cap \mathcal{C}_T$ 的指示函数的近端算子，对应于向集合 $\mathcal{C}_U \cap \mathcal{C}_T$ 的正交投影。指示函数定义如下：

$$\mathcal{X}_{\mathcal{C}}(\mathbf{X}) = \begin{cases} 0, & \text{for } \mathbf{X} \in \mathcal{C}, \\ \infty, & \text{otherwise.} \end{cases} \tag{9}$$

近端算子详见第六节。第 i 次迭代的步长 $\tau^{(i)}$ 按照 [53], [54] 中的说明选取。此外，在 (8) 中，我们有 $\nabla F(\cdot)$ 和 $\nabla R(\cdot)$，它们分别是 $F(\cdot)$ 和 $R(\cdot)$ 的梯度，详见第五节。FBS 以一个从均匀随机分布中抽取的矩阵 $\mathbf{C}^{(1)} \in [0,1]^{U\times T}$ 作为初始值。最后一次迭代中的矩阵 $\mathbf{C}^{(I_{\max})}$ 的元素会被量化为二值，同时满足约束集 $\mathcal{C}_U$ 和 $\mathcal{C}_T$；详情见第七节。将 $\mathbf{C}^{(I_{\max})}$ 的元素量化为二值，目的是确定用户设备 u 在时隙 t 中是处于激活状态还是非激活状态，即确定 $C^{(I_{\max})}_{u,\,t}$ 是否等于 1 或 0。

V. 目标函数与梯度

为展示我们用户设备调度框架的灵活性，现给出 $F(\mathbf{C})$ 的三种选项及其梯度 $\nabla F(\mathbf{C})$。我们还详细介绍促使解为二值解的正则化项 $R(\mathbf{C})$ 及其梯度 $\nabla R(\mathbf{C})$。请注意，梯度 $\nabla F(\mathbf{C}) \in \mathbb{R}^{U\times T}$和 $\nabla R(\mathbf{C}) \in \mathbb{R}^{U\times T}$分别是偏导数矩阵，其元素为 $\partial F(\mathbf{C})/\partial C_{u,t}$和 $\partial R(\mathbf{C})/\partial C_{u,t}$，其中$u = 1, \ldots, U$和 $t = 1, \ldots, T$。

A. LMMSE均衡后均方误差

1) 目标函数：对于第一个目标函数，我们考虑后LMMSE均衡均方误差，即原始发送符号向量 $\mathbf{s}[t]$ 与估计发送符号向量 $\hat{}\mathbf{s}[t] = \mathbf{W}[t]^{\mathrm{H}}\mathbf{y}[t]$ 之间的均方误差，其中 $\mathbf{W}[t]$ 是一个 LMMSE 均衡器。我们将该目标函数定义为

$$F(\mathbf{C}) = \sum_{t=1}^{T} \mathbb{E}\left[\left\|\mathbf{D}_{\mathbf{C}}[t]\,\mathbf{s}[t] - \mathbf{D}_{\mathbf{C}}[t]\,\mathbf{W}[t]^{\mathrm{H}}\,\mathbf{y}[t]\right\|_2^2\right], \tag{10}$$

其中，期望是对$\mathbf{s}[t]$ 和$\mathbf{n}[t]$ 以及[3]取的

$$\mathbf{W}[t] = \left(\hat{\mathbf{H}}\mathbf{D}_{\mathbf{C}}[t]^2\,\hat{\mathbf{H}}^{\mathrm{H}} + N_0\mathbf{I}_B\right)^{-1}\hat{\mathbf{H}}\mathbf{D}_{\mathbf{C}}[t] \tag{11}$$

是时隙 t中的LMMSE均衡矩阵。

[3]在我们的仿真中，我们通过进行一次 $U \times U$ 求逆运算，降低了LMMSE均衡器计算的复杂度。不过，为了简化推导过程的表述，我们保留了（11）的原始形式。

2）梯度：式（10）中的后LMMSE均衡均方误差目标函数对调度矩阵$\mathbf{C}$ 中某个元素的偏导数为

$$\frac{\partial F(\mathbf{C})}{\partial C_{i,t}} = 2C_{i,t}\big(1 + \hat{\mathbf{h}}_i^{\mathrm{H}}\mathbf{L}_1[t]\,\mathbf{L}_2[t]\,\mathbf{L}_1[t]\,\hat{\mathbf{h}}_i - 2C_{i,t}^2\hat{\mathbf{h}}_i^{\mathrm{H}}\mathbf{L}_1[t]\,\hat{\mathbf{h}}_i\big), \tag{12}$$

其中，$\hat{}\mathbf{h}_i \in \mathbb{C}^{B\times 1}$ 是矩阵 $\hat{}\mathbf{H}$ 的第 i 列，

$$\mathbf{L}_1[t] = \left(\hat{\mathbf{H}}\mathbf{D}_{\mathbf{C}}[t]^2\,\hat{\mathbf{H}}^{\mathrm{H}} + N_0\mathbf{I}_B\right)^{-1}, \tag{13}$$

且 $\mathbf{L}_2[t] = \hat{}\mathbf{H}\mathbf{D}_{\mathbf{C}}[t]^4\,\hat{}\mathbf{H}^{\mathrm{H}}$。[4]

B. 信道容量

1) 目标函数：对于第二个目标函数，我们考虑信道容量。为简单起见，我们取估计信道 $\hat{}\mathbf{H}[t]$ 的容量，并将其视为真实信道$\mathbf{H}[t]$的容量。我们将此目标函数定义为

$$F(\mathbf{C}) = -\sum_{t=1}^{T}\log_2\left(\det\left(\mathbf{I}_B + \frac{1}{N_0}\hat{\mathbf{H}}\mathbf{D}_{\mathbf{C}}[t]^2\,\hat{\mathbf{H}}^{\mathrm{H}}\right)\right). \tag{14}$$

注意，此目标函数是信道容量的负值，因为目标是最大化该量。

2) 梯度：信道容量目标函数 (14) 对调度矩阵$\mathbf{C}$ 中一个元素的偏导数为

$$\frac{\partial F(\mathbf{C})}{\partial C_{i,t}} = -2\frac{1}{N_0}C_{i,t}(\log_2 e)\,\hat{\mathbf{h}}_i^{\mathrm{H}}\mathbf{L}_3[t]^{-1}\,\hat{\mathbf{h}}_i, \tag{15}$$

其中 e 为欧拉数，且$\mathbf{L}_3[t] = \mathbf{I}_B + \frac{1}{N0}\,\hat{}\mathbf{H}\mathbf{D}_{\mathbf{C}}[t]^2\,\hat{}\mathbf{H}^{\mathrm{H}}$。

C. 后*LMMSE*均衡可达速率和

1) 目标函数：对于第三个目标函数，我们考虑后LMMSE均衡可达速率和，即接收向量$\mathbf{y}[t]$ 经过(11)中的LMMSE均衡器$\mathbf{W}[t]$之后，所有用户设备和时隙上的可达速率之和。与(14)中的目标函数类似，我们希望最大化可达速率之和，因此最小化该表达式的负值。我们将该目标函数定义为

$$F(\mathbf{C}) = -\sum_{t=1}^{T}\sum_{u=1}^{U}\log_2(1 + \mathit{SINR}_u(\mathbf{c}_t)), \tag{16}$$

其中，时隙 t 中第 u个用户设备的均衡后信干噪比由$\mathit{SINR}_u(\mathbf{c}_t) = \frac{\phi(\mathbf{c}_t)}{\Psi(\mathbf{c}_t)}$给出。在该表达式中，我们定义

$$\phi(\mathbf{c}_t) = \left|\mathbf{w}_u[t]^{\mathrm{H}}\,\hat{\mathbf{h}}_u[t]\right|^2, \tag{17}$$

其为期望信号功率，以及分母

$$\Psi(\mathbf{c}_t) = \varphi(\mathbf{c}_t) + \varrho(\mathbf{c}_t) + \Phi(\mathbf{c}_t)\,. \tag{18}$$

[4]在上述表达式中，由于各项均为 $\{0,1\}$，因此可以省略$\mathbf{D}_{\mathbf{C}}[t]$ 中的指数；这种简化对性能的影响微乎其微。

在式 (18) 中，我们将干扰、信道估计误差和噪声功率定义如下：

$$\varphi(\mathbf{c}_t) = \sum_{j=1,j\neq u}^{U}\left|\mathbf{w}_u[t]^{\mathrm{H}}\hat{\mathbf{h}}_j[t]\right|^2, \tag{19}$$

$$\varrho(\mathbf{c}_t) = \|\mathbf{w}_u[t]\|_2^2\sum_{j=1}^{U}C_{j,t}^2\Gamma_j, \text{ and} \tag{20}$$

$$\Phi(\mathbf{c}_t) = N_0\|\mathbf{w}_u[t]\|_2^2\,. \tag{21}$$

在上述表达式中，$\mathbf{w}_u[t]$ 是(11)中$\mathbf{W}[t]$ $\hat{}$的第 u列，$\mathbf{h}_u[t]$ 是信道估计矩阵$\hat{}\mathbf{H}[t] = \hat{}\mathbf{H}\mathbf{D}_{\mathbf{C}}[t]$的第 u列。此外，在(20)中，$\mathbb{E}\left[\mathbf{e}_u\mathbf{e}_u^{\mathrm{H}}\right] = \Gamma_u\mathbf{I}_B$，对于 $u = 1$，.....，U，其中$\mathbf{e}_u$是信道估计误差矩阵$\mathbf{E}$的第 u列。对于毫米波大规模多用户多输入多输出场景，我们将 Γ_u 近似为第III-A节中提到的BEACHES算法 [52]的每用户均方误差输出。对于6 GHz以下无蜂窝大规模多用户多输入多输出场景， $\Gamma_u = N_0/U$。我们强调，$\mathrm{SINR}_u(\mathbf{c}_t)$包含$\mathbf{c}_t$ ，因为$\mathbf{W}[t]$ 和$\mathbf{H}[t]$ 依赖于$\mathbf{c}_t$。

2) 梯度：式(16)中后LMMSE均衡可达速率和目标关于调度矩阵$\mathbf{C}$的一个条目的偏导数由下式给出

$$\frac{\partial F(\mathbf{C})}{\partial C_{i,t}} = -\sum_{t=1}^{T}\sum_{u=1}^{U}\frac{(\log_2 e)}{(1+\mathit{SINR}_u(\mathbf{c}_t))}\frac{\partial}{\partial C_{i,t}}\mathit{SINR}_u(\mathbf{c}_t), \tag{22}$$

其中

$$\frac{\partial}{\partial C_{i,t}}\mathit{SINR}_u(\mathbf{c}_t) = \frac{\frac{\partial}{\partial C_{i,t}}[\phi(\mathbf{c}_t)]\,\Psi(\mathbf{c}_t) - \phi(\mathbf{c}_t)\,\frac{\partial}{\partial C_{i,t}}[\Psi(\mathbf{c}_t)]}{\left(\Psi(\mathbf{c}_t)\right)^2}. \tag{23}$$

式(17)中期望信号功率、式(19)中干扰功率、式(20)中信道估计功率以及式(21)中噪声功率的偏导数分别由式(24)、(25)、(26)和(27)给出。在所有这些表达式中，矩阵$\mathbf{L}_1[t]$ 均按式(13)定义。

D. 二值化正则项

1) 目标函数：对于促进$\mathbf{C}$中二值条目的正则化项 $R(\mathbf{C})$，我们采用 [55]中提出的方法。与式(14)和式(16)中的目标函数类似，我们希望最大化该表达式。因此，我们将正则化项定义如下：

$$R(\mathbf{C}) = -\sum_{t=1}^{T}\sum_{u=1}^{U}\alpha|C_{u,t} - 0.5|^2\,, \tag{28}$$

其中 $\alpha \geq 0$ 是正则化参数，其值越大，对二值解的约束越严格（见表 II）。

2) 梯度：式（28）中的正则化项关于矩阵$\mathbf{C}$中某个元素的偏导数为

$$\frac{\partial R(\mathbf{C})}{\partial C_{i,t}} = -\sum_{t=1}^{T}\sum_{u=1}^{U}\frac{\partial}{\partial C_{i,t}}\alpha|C_{u,t} - 0.5|^2\,, \tag{29}$$

其中

$$\frac{\partial}{\partial C_{i,t}}\alpha|C_{u,t} - 0.5|^2 = \begin{cases}\alpha(2C_{u,t} - 1)\,, & \text{for } i = u,\\ 0, & \text{for } i \neq u.\end{cases} \tag{30}$$

$$\frac{\partial \phi(\mathbf{c}_t)}{\partial C_{i,t}} = \begin{cases} 4C_{u,t}^3 \left(\hat{\mathbf{h}}_u^{\mathrm{H}} \mathbf{L}_1[t]\, \hat{\mathbf{h}}_u\right)^2 \left(1 - C_{u,t}^2 \left(\hat{\mathbf{h}}_u^{\mathrm{H}} \mathbf{L}_1[t]\, \hat{\mathbf{h}}_u\right)\right), & \text{for } i = u, \\ -4C_{u,t}^4 C_{i,t} \left(\hat{\mathbf{h}}_u^{\mathrm{H}} \mathbf{L}_1[t]\, \hat{\mathbf{h}}_u\right) \left(\hat{\mathbf{h}}_i^{\mathrm{H}} \mathbf{L}_1[t]\, \hat{\mathbf{h}}_u \hat{\mathbf{h}}_u^{\mathrm{H}} \mathbf{L}_1[t]\, \hat{\mathbf{h}}_i\right), & \text{for } i \neq u. \end{cases} \tag{24}$$

$$\frac{\partial \varphi(\mathbf{c}_t)}{\partial C_{i,t}} = \begin{cases} \sum_{j=1, j\neq u}^{U} 2C_{u,t} C_{j,t}^2 \left(\hat{\mathbf{h}}_j^{\mathrm{H}} \mathbf{L}_1[t]\, \hat{\mathbf{h}}_u \hat{\mathbf{h}}_u^{\mathrm{H}} \mathbf{L}_1[t]\, \hat{\mathbf{h}}_j - 2C_{u,t}^2 \hat{\mathbf{h}}_j^{\mathrm{H}} \left(\mathbf{L}_1[t]\, \hat{\mathbf{h}}_u \hat{\mathbf{h}}_u^{\mathrm{H}}\right)^2 \mathbf{L}_1[t]\, \hat{\mathbf{h}}_j\right), \\ \text{for } i = u, \\ \sum_{j=1, j\neq u}^{U} 2C_{j,t} C_{u,t}^2 \left(\hat{\mathbf{h}}_j^{\mathrm{H}} \mathbf{L}_1[t]\, \hat{\mathbf{h}}_u \hat{\mathbf{h}}_u^{\mathrm{H}} \mathbf{L}_1[t]\, \hat{\mathbf{h}}_j - 2C_{j,t}^2 \hat{\mathbf{h}}_u^{\mathrm{H}} \left(\mathbf{L}_1[t]\, \hat{\mathbf{h}}_j \hat{\mathbf{h}}_j^{\mathrm{H}}\right)^2 \mathbf{L}_1[t]\, \hat{\mathbf{h}}_u\right), \\ \text{for } i = j, \\ -2C_{u,t}^2 C_{j,t}^2 C_{i,t} \left(\hat{\mathbf{h}}_j^{\mathrm{H}} \mathbf{L}_1[t]\, \hat{\mathbf{h}}_i \hat{\mathbf{h}}_i^{\mathrm{H}} \mathbf{L}_1[t]\, \hat{\mathbf{h}}_u \hat{\mathbf{h}}_u^{\mathrm{H}} \mathbf{L}_1[t]\, \hat{\mathbf{h}}_j + \hat{\mathbf{h}}_j^{\mathrm{H}} \mathbf{L}_1[t]\, \hat{\mathbf{h}}_u \hat{\mathbf{h}}_u^{\mathrm{H}} \mathbf{L}_1[t]\, \hat{\mathbf{h}}_i \hat{\mathbf{h}}_i^{\mathrm{H}} \mathbf{L}_1[t]\, \hat{\mathbf{h}}_j\right), \text{ for } i \neq u, i \neq j. \end{cases} \tag{25}$$

$$\frac{\partial \varrho(\mathbf{c}_t)}{\partial C_{i,t}} = \begin{cases} \left(2C_{u,t} \left[\hat{\mathbf{h}}_u^{\mathrm{H}} \mathbf{L}_1[t]\, \mathbf{L}_1[t]\, \hat{\mathbf{h}}_u - 2C_{u,t}^2 \hat{\mathbf{h}}_u^{\mathrm{H}} \mathbf{L}_1[t]\, \mathbf{L}_1[t]\, \hat{\mathbf{h}}_u \hat{\mathbf{h}}_u^{\mathrm{H}} \mathbf{L}_1[t]\, \hat{\mathbf{h}}_u\right]\right) \\ \left(\sum_{j=1}^{U} C_{j,t}^2 \Gamma_j\right) + \left(\|\mathbf{w}_u[t]\|_2^2\right) (2C_{u,t} \Gamma_u), \text{ for } i = u, \\ \left(-2C_{u,t}^2 C_{i,t} \left(\hat{\mathbf{h}}_i^{\mathrm{H}} \mathbf{L}_1[t]\, \mathbf{L}_1[t]\, \hat{\mathbf{h}}_u \hat{\mathbf{h}}_u^{\mathrm{H}} \mathbf{L}_1[t]\, \hat{\mathbf{h}}_i + \hat{\mathbf{h}}_u^{\mathrm{H}} \mathbf{L}_1[t]\, \mathbf{L}_1[t]\, \hat{\mathbf{h}}_i \hat{\mathbf{h}}_i^{\mathrm{H}} \mathbf{L}_1[t]\, \hat{\mathbf{h}}_u\right)\right) \left(\sum_{j=1}^{U} C_{j,t}^2 \Gamma_j\right) \\ + \left(\|\mathbf{w}_u[t]\|_2^2\right) (2C_{i,t} \Gamma_i), \text{ for } i \neq u. \end{cases} \tag{26}$$

$$\frac{\partial \Phi(\mathbf{c}_t)}{\partial C_{i,t}} = \begin{cases} 2N_0 C_{u,t} \left(\hat{\mathbf{h}}_u^{\mathrm{H}} \mathbf{L}_1[t]\, \mathbf{L}_1[t]\, \hat{\mathbf{h}}_u - 2C_{u,t}^2 \hat{\mathbf{h}}_u^{\mathrm{H}} \mathbf{L}_1[t]\, \mathbf{L}_1[t]\, \hat{\mathbf{h}}_u \hat{\mathbf{h}}_u^{\mathrm{H}} \mathbf{L}_1[t]\, \hat{\mathbf{h}}_u\right), \text{ for } i = u, \\ -2N_0 C_{u,t}^2 C_{i,t} \left(\hat{\mathbf{h}}_i^{\mathrm{H}} \mathbf{L}_1[t]\, \mathbf{L}_1[t]\, \hat{\mathbf{h}}_u \hat{\mathbf{h}}_u^{\mathrm{H}} \mathbf{L}_1[t]\, \hat{\mathbf{h}}_i + \hat{\mathbf{h}}_u^{\mathrm{H}} \mathbf{L}_1[t]\, \mathbf{L}_1[t]\, \hat{\mathbf{h}}_i \hat{\mathbf{h}}_i^{\mathrm{H}} \mathbf{L}_1[t]\, \hat{\mathbf{h}}_u\right), \text{ for } i \neq u. \end{cases} \tag{27}$$

VI. 约束集上的正交投影

A. 道格拉斯-拉赫福德分裂（*DRS*）

现在，我们详细说明式 (8) 中 $\mathrm{prox}^{\mathcal{X}_{\mathcal{C}_U \cap \mathcal{C}_T}}(\cdot)$ 的实现方式，它对应于两个单纯形交集上的正交投影。为求得解，我们采用 Douglas-Rachford 分裂法（DRS） [56]。DRS 会交替投影到集合 $\mathcal{C}_U$ 和 $\mathcal{C}_T$ 上，直至找到解。具体而言，我们求解以下优化问题：

$$\mathrm{prox}_{\mathcal{X}_{\mathcal{C}_U \cap \mathcal{C}_T}}(\mathbf{Z}) = \arg\min_{\mathbf{X}} \psi(\mathbf{X}, \mathbf{Z}) + \xi(\mathbf{X}, \mathbf{Z}), \tag{31}$$

使用这两个函数

$$\psi(\mathbf{X}, \mathbf{Z}) = \frac{\beta}{2} \|\mathbf{X} - \mathbf{Z}\|_{\mathrm{F}}^2 + \mathcal{X}_{\mathcal{C}_U}(\mathbf{X}), \text{ and} \tag{32}$$

$$\xi(\mathbf{X}, \mathbf{Z}) = \frac{\beta}{2} \|\mathbf{X} - \mathbf{Z}\|_{\mathrm{F}}^2 + \mathcal{X}_{\mathcal{C}_T}(\mathbf{X}). \tag{33}$$

此处，$\beta \geq 0$ 是可用于加速收敛的调节参数。对集合 $\mathcal{C}_U$ 和 $\mathcal{C}_T$ 的投影通过指示函数 $\mathcal{X}_{\mathcal{C}_U}$ 和 $\mathcal{X}_{\mathcal{C}_T}$强制实现，如（9）所示。

DRS 迭代执行以下两个步骤 [56]:

$$\mathbf{V}^{(k+1)} = \mathrm{prox}_{\psi}\left(\mathbf{G}^{(k)}\right) = \arg\min_{\mathbf{X} \in \mathcal{C}_U} \left\| \mathbf{X} - \frac{\beta \mathbf{Z} + \mathbf{G}^{(k)}}{\beta + 1} \right\|_{\mathrm{F}}^2, \tag{34}$$

$$\begin{aligned} \mathbf{G}^{(k+1)} &= \mathrm{prox}_{\xi}\left(2\mathbf{V}^{(k+1)} - \mathbf{G}^{(k)}\right) + \mathbf{G}^{(k)} - \mathbf{V}^{(k+1)} \\ &= \arg\min_{\mathbf{X} \in \mathcal{C}_T} \left\| \mathbf{X} - \frac{\beta \mathbf{Z} + \left(2\mathbf{V}^{(k+1)} - \mathbf{G}^{(k)}\right)}{\beta + 1} \right\|_{\mathrm{F}}^2 + \mathbf{G}^{(k)} \\ &\quad - \mathbf{V}^{(k+1)}. \end{aligned} \tag{35}$$

我们以 $\mathbf{G}^{(1)} = \mathbf{0}_{U \times T}$ 初始化该过程。收敛或达到最大迭代次数 $K_{\max}$ 后，DRS 输出使 (31) 最小化的矩阵，$\mathbf{V}^{(K_{\max})}$。

B. 带不等式约束的单纯形上的正交投影

由于（34）和（35）将输入矩阵的每一列投影到受不等式约束的单纯形上，我们需要一个由不等式约束指定的一般单纯形上的正交投影。给定向量$\mathbf{q} \in \mathbb{R}^M$，我们的目标是找到最接近$\mathbf{q}$ 的解向量$\mathbf{p}^* \in \mathcal{C}^M$，如下所示：

$$\underset{\mathbf{p} \in [0,1]^M}{\text{minimize}} \frac{1}{2} \|\mathbf{p} - \mathbf{q}\|_2^2 \text{ subject to } \ell_{\min} \leq \sum_{i=1}^{M} p_i \leq \ell_{\max}. \tag{36}$$

在附录A中，我们为上述问题提供了KKT条件 [57, 第5.5节 3]。KKT条件是一组必要条件，有助于识别约束优化问题的潜在解。我们利用这些条件来开发一种计算上述投影的算法。

根据附录A中推导出的KKT条件，式(36)的解$\mathbf{p}^*$ 可以逐元素写为

$$p_i^* = \max\{\min\{q_i + \lambda^*, 1\}, 0\}, \text{ for } i = 1, \ldots, M, \tag{37}$$

其中 $\lambda^* \in \mathbb{R}$。因此，求解式(36)的关键在于找到一个值 λ^*，使所得向量$\mathbf{p}^*$ 满足求和约束，同时在欧氏距离意义下尽可能接近$\mathbf{q}$ 。

现在我们概述寻找λ^*的步骤：

1) 我们按降序对$\mathbf{q}$进行排序，得到 $\tilde{\mathbf{q}}$。

2) 我们将 $\tilde{\mathbf{q}}$ 的最大元素从所有元素中减去，得到新的向量 $\hat{\mathbf{q}}$ ，使第一个元素为 0，其余元素均为负数。3) 我们构建向量 $\boldsymbol{\chi}=\left[-\hat{\mathbf{q}}^T, \mathbf{1}_M^T-\hat{\mathbf{q}}^T\right]^T$，使得 $\mathbf{p}=\max\{\min\{\hat{\mathbf{q}}+\chi_i, 1\}, 0\}\in[0,1]^M$，其中 $i=1,\ldots,2M$。4) 我们考虑$\boldsymbol{\chi}$的连续取值之间的区间，例如 $[\chi_i, \chi_{i+1}]$。我们的目标是确定在哪些区间内满足关于 $\mathbf{p}$的约束，即 $\ell_{\min} \leq \sum_{i=1}^{M}\max\{\min\{\hat{q}_i+\lambda,1\}, 0\} \leq \ell_{\max}$，其中$\lambda\in[\chi_i,\chi_{i+1}]$。5) 找到这样的搜索区间 $[\chi_i,\chi_{i+1}]$ 后，利用附录A中给出的KKT条件，计算 λ^* 的候选值，并选出使式（36）中的均方误差最小的值。

VII. 量化为离散约束集

如第IV-C节所述，矩阵 $\mathbf{C}^{(I_{\max})}$ 中的值位于范围 $[0,1]$，内，因此需要将其量化为集合 $\{0,1\}$中的二值。首先，我们通过检查所有元素是否已经属于 $\{0,1\}$，来判断是否需要量化。如果是，则直接将最终量化矩阵设为 $\mathbf{C}^{\text{quant}}=\mathbf{C}^{(I_{\max})}$。否则，我们采用以下量化策略，在满足约束 $\mathcal{C}_U\cap\mathcal{C}_T$的同时，将各元素映射到 $\{0,1\}$。

A. 投影到下界

量化的第一步是满足下界约束：每个时隙中必须传输的 UE 的最小数量（$U_{\min}$），以及每个 UE 必须传输的最小时隙数（$T_{\min}$），分别如式 (5) 和式 (6) 所定义。为此，我们首先令 $\mathbf{C}=\mathbf{C}^{(I_{\max})}$，并将集合 $\mathcal{U}=1,\ldots,U$初始化为包含所有 UE，同时初始化矩阵$\mathbf{C}^{\text{lower}}=\mathbf{0}_{U\times T}$。对于每个时隙 t，我们选取第 t 列 $\mathbf{c}_t$中数值最高的 $U_{\min}$ 个 UE 对应的索引，并将矩阵 $\mathbf{C}^{\text{lower}}$ 中的相应元素设为 1。然后，我们检查哪些 UE 已在 $\mathbf{C}^{\text{lower}}$中至少被调度了 $T_{\min}$ 个时隙。将这些 UE 从 $\mathcal{U}$中移除后，继续处理下一个时隙。重复此过程，直至处理完全部 T 个时隙。完成此下界投影后，我们计算目标函数 $F(\mathbf{C}^{\text{lower}})$，为下一步满足上界约束做准备。

B. 投影到上界

量化的第二步强制满足上界约束：每个时隙允许传输的用户设备最大数量（$U_{\max}$），以及每个用户设备允许传输的最大时隙数（$T_{\max}$），分别如式 (5) 和 (6) 所定义。为此，我们首先设置 $\mathbf{C}=\mathbf{C}^{(I_{\max})}$、$\mathbf{C}^{\text{upper}}=\mathbf{C}^{\text{lower}}$，并将所有用户设备初始化为集合 $\mathcal{U}=1,\ldots,U$的元素。然后，我们计算一个时隙中还可以调度的用户设备数量，记为 $U_{\text{gap}}=U_{\max}-U_{\min}$。接下来，我们找出在时隙 t 中尚未被调度且仍属于集合 $\mathcal{U}$的用户设备。从中选出在 $\mathbf{C}^{\text{upper}}$、$\mathbf{c}_t^{\text{upper}}$ 的第 t 列中取值最高的 U_{gap} 个用户设备，并将其加入候选集 $\mathcal{S}$。我们定义测试矩阵 $\mathbf{C}^{\text{test}}=\mathbf{C}^{\text{upper}}$。对于每个用户设备 $u\in\mathcal{S}$，我们评估将矩阵 $\mathbf{C}^{\text{test}}$ 中对应元素量化为 1 是否会降低目标函数 $F(\mathbf{C}^{\text{test}})$。如果会，则将 $\mathbf{C}^{\text{upper}}$ 中对应的元素更新为 1。如果此次更新使用户设备 u 达到其允许的最大传输次数（$T_{\max}$），则将其从 $\mathcal{U}$中移除。对所有 T 个时隙重复此过程。上界投影完成后，最终量化矩阵为 $\mathbf{C}^{\text{upper}}$。

VIII. 结果

现在，我们展示所提出的用户设备调度框架在两个仿真场景中的有效性：(i) 集中式毫米波大规模多用户多输入多输出场景，以及 (ii) 6 GHz以下无蜂窝大规模多用户多输入多输出场景。除非另有说明，其余参数在两个场景中均相同。每个用户设备的功率控制动态范围设为 $\eta=6$dB。对于噪声功率，我们采用 $N_0=W\times k_B\times T_0\times NF$ [W] [4],，其中 $W=100$MHz 为带宽，$k_B=1.381\times10^{-23}$ J/K 为玻尔兹曼常数，$T_0=290$K 为噪声温度，$NF=9$dB 为噪声系数。我们使用 Remcom 的 Wireless InSite [58]生成的信道向量来评估该框架。为了生成一次信道实现，U 个用户设备位置以均匀且相互独立的方式随机选取。我们进行 10^3 次信道实现，并且每次信道实现执行 $I_{\text{MC}}=10^5$ 次蒙特卡洛试验。对于FBS算法（参见第IV-C节），初始步长设为 $\tau^{(1)}=0.01$，梯度步的最大迭代次数设为$I_{\max}=100$。对于 DRS算法（参见第六节-A），最大迭代次数设为 $K_{\max}=2000$。

A. 性能指标

对于每用户未编码误比特率和每用户均方误差，我们以每用户累积分布函数（CDF）的形式报告结果；对于每用户HMI和每用户设备可达速率，我们使用互补累积分布函数（CCDF）。我们还通过考察CDF和CCDF的第 90 百分位数来汇总部分结果，分别将其定义为$\Pr[X\leq x]=0.9$ 和$\Pr[X>x]=0.9$。所有考虑的性能指标均在使用LMMSE均衡器之后进行评估。接下来，我们将详细说明每用户未编码误比特率、每用户HMI、每用户均方误差和每用户设备可达速率这些性能指标。

1) 未编码的每用户比特错误率：第 u个用户设备的未编码误码率定义为

$$BER_u=\frac{\sum_{i=1}^{I_{\text{MC}}}\sum_{t=1}^{T_u}\varepsilon_{i,t}^{u}}{QI_{\text{MC}}T_u}, \tag{38}$$

其中 T_u 是第 u个用户设备处于活动状态的时隙数（$T_{\min}\leq T_u\leq T_{\max}$），$\varepsilon_{i,\,t}^{u}$ 是第 u个用户设备在试验 i 和时隙 t中的比特错误数，Q 是每符号比特数。例如，对于四相相移键控和16阶正交振幅调制这两种调制方式，我们分别有 $Q=2$ 和 $Q=4$ 比特。

2）每用户*HMI*：第 u 个用户设备的 HMI 定义为

$$I_u=\sum_{i=1}^{Q}I(b_u^i,\hat{b}_u^i)=\sum_{i=1}^{Q}H(b_u^i)-H(b_u^i|\hat{b}_u^i), \tag{39}$$

其中 $H(b_u^i) = -\sum_{b_u^i\in\{0,1\}} p(b_u^i)\log_2(p(b_u^i))$ 是 b_u^i 的熵，并且

$$H(b_u^i|\hat{b}_u^i) = -\sum_{b_u^i\in\{0,1\}}\sum_{\hat{b}_u^i\in\{0,1\}} p(b_u^i,\hat{b}_u^i)\log_2\left(\frac{p(b_u^i,\hat{b}_u^i)}{p(b_u^i)}\right) \quad (40)$$

是 b_u^i 在给定 $\hat{}\, b_u^i$ 时的条件熵。在上述表达式中，Q 是每符号比特数，b_u^i 是第 i 个 UE 在第 u 个时隙传输的比特，$\hat{}\, b_u^i$ 是第 u 个 UE 在第 i 个时隙传输的第 p 个估计比特，$b_u^i = 0(=p)\ b_u^i = 1(=1/2)\ p$，而 $b_u^i(,\ \hat{}\ u)$ 是和 $\hat{}$的联合概率分布，我们通过经验计算得出。对于联合概率 (，$\hat{}$) 的经验计算，我们仅考虑 UE 处于活动状态的情况。

3）每用户均方误差：每用户均方误差定义为

$$MSE_u = \frac{\sum_{i=1}^{I_{\mathrm{MC}}}\sum_{t=1}^{T_u}\left|s_u^i[t]-\hat{s}_u^i[t]\right|^2}{I_{\mathrm{MC}}T_u}, \quad (41)$$

其中，$s_u^i[t]$ 是 UE u 在第 i 次试验和第 t 个时隙中的发射符号，$\hat{}\, s_u^i[t]$ 是 UE u 在第 i 次试验和第 t 个时隙中的估计发射符号。回顾一下，$\hat{}\mathbf{s}^i[t] = \mathbf{W}[t]^{\mathrm{H}}\mathbf{y}^i[t]$，其中$\mathbf{y}^i[t]$ 是第 i 次试验和第 t 个时隙的接收向量。

4) 每用户设备可达速率：每用户设备可达速率定义为

$$Rate_u = \sum_{t=1}^{T}\log_2(1+\Omega_u(\mathbf{c}_t)), \quad (42)$$

其中 $\Omega_u(\mathbf{c}_t)$ 由 (43) 给出，不应与第五节-C中所述的量相混淆。在该表达式中，$\mathbf{e}_j[t]$ 是噪声矩阵 $\mathbf{E}[t] = \mathbf{E}\mathbf{D}_{\mathbf{C}}[t]$ 的第 j 列。

B. 现有方法与基线算法

在我们的仿真结果中，我们将后LMMSE均衡均方误差、信道容量以及后LMMSE均衡可达速率和分别称为“MSE-LMMSE”、“信道容量”和“可达速率”。

为了将我们的框架与现有方法进行基准对比，我们还仿真了“SUS” [16], “CSS” [5],和“贪心” [5]算法。对于这些方法，我们在第一个时隙调度用户设备，然后在后续时隙中继续调度剩余未调度的用户设备，直到所有用户设备请求被分配到 T 个可用时隙中。此外，我们将我们的方法与 [38]中提出的低复杂度方法“LoFi (K=1)”、“LoFi (K=4)”、“LoFi++ (K=1)”和“LoFi++(K=4)”进行比较。最后，我们还研究了两种额外的基线方法：“随机”和“所有用户设备激活”（简称“AUA”）。“随机”基线以均匀随机方式选取用户设备子集，“AUA”基线则认为所有用户设备在所有时隙中均被调度。

与我们的方法不同，在基准算法之一“随机”以及现有方法“LoFi (K=1)”、“LoFi (K=4)”、“LoFi++ (K=1)”、“LoFi++ (K=4)”、“SUS”、“CSS”和“贪心”中，我们考虑每个用户设备仅允许在一个时隙内传输，即 $T_{\min} = T_{\max} = 1$，并且允许 U/T 个用户设备在每个时隙传输，即$U_{\min} = U_{\max} = U/T$。做出这一考虑的原因在于，对于上述每一种技术，我们都需要测试所有调度可能性，同时考虑 $T_{\min}$ 与 $T_{\max}$ 之间的范围以及 $U_{\min}$ 与 $U_{\max}$之间的范围。因此，我们将创建文献中不存在的新算法，而这并非我们的目标。

表 I 毫米波与无蜂窝大规模多用户多输入多输出系统的评估场景

场景	B	L	n_{A}	U	T	$T_{\min}$	$T_{\max}$	$U_{\min}$	$U_{\max}$	K_{init}
S1	16	1	16	16	2	1	2	8	10	80
S2	32	1	32	32	2	1	2	16	20	10
S3	32	1	32	64	4	1	2	16	20	3
S4	16	8	2	16	2	1	2	8	10	80
S5	32	16	2	32	2	1	2	16	20	10
S6	32	16	2	64	4	1	2	16	20	3

C. 毫米波大规模多用户多输入多输出场景

1) 仿真设置：我们考虑一个载波频率为 60 GHz 的毫米波大规模多用户多输入多输出系统。在此场景中，我们将每个用户设备的发射功率设为 $\rho_u = 100$ [mW]，用户设备发送采用格雷映射的四相相移键控符号。在基站处，我们采用天线间距为半波长的均匀线性阵列。基站和用户设备的天线均为全向天线，高度分别为 15 m 和 1.65 m。我们在一个 150 m $\times$ 200m 的区域内，为 30,351 个用户设备位置生成信道向量。

表 I 列出了毫米波大规模多用户多输入多输出场景的三种不同系统规模 S1、S2 和 S3，其中我们改变接收天线数量 B，即基站处的天线数量 n_{A}、待服务的用户设备数量 U 以及可用时隙数量 T。此外，考虑到已服务用户设备的数量和可用时隙数量，我们分别定义每个用户设备的最小时隙数和最大时隙数，以及每个时隙的最小用户设备数和最大用户设备数，即 $T_{\min}$、$T_{\max}$、$U_{\min}$ 和 $U_{\max}$。在表格的最后一列中，我们定义了所述各场景下用户设备调度算法的随机初始化次数（参见第 IV-B 节）。表 II 列出了表 I 所述各场景 S1、S2 和 S3 对应的正则化参数 α，该参数来自式 (28)。需要指出的是，由于计算复杂度较高，我们仅在最小系统（即场景 S1）上评估 ES 基线。在该基线中，我们测试每一种可能的调度矩阵 $\mathbf{C}$，并选择使给定目标函数最小化的矩阵。我们将使均方误差最小化的穷举搜索算法称为“ES-MSE”。我们将使信道容量和可达速率之和最大化的穷举搜索算法分别称为“ES-容量”和“ES-速率”。

2）结果：图 1 展示了场景 S1 的结果，其参数列于表 I。图 1(a)–(c) 显示，对于 90% 的用户设备，“MSE-LMMSE”的性能最佳，此外还包括“ES-MSE”、“ES-容量”和“ES-速率”。正如预期，“AUA”基线在每用户设备误码率、HMI 和均方误差指标方面表现最差，如图 1(a)–(c) 所示；但在每用户设备可达速率方面表现最佳，如图 1(d) 所示。需要强调的是，我们的“ MSE-LMMSE”、“信道容量”和“可达速率”方法均未针对所展示的指标进行优化，

$$\Omega_u(\mathbf{c}_t) = \frac{\left|\mathbf{w}_u[t]^{\mathrm{H}}\hat{\mathbf{h}}_u[t]\right|^2}{\sum_{j=1,j\neq u}^{U}\left|\mathbf{w}_u[t]^{\mathrm{H}}\hat{\mathbf{h}}_j[t]\right|^2+\sum_{j=1}^{U}\left|\mathbf{w}_u[t]^{\mathrm{H}}\mathbf{e}_j[t]\right|^2+N_0\|\mathbf{w}_u[t]\|_2^2} \tag{43}$$

表 II 毫米波和无蜂窝大规模多用户多输入多输出系统中由 (28) 得出的 α 取值

场景	MSE-LMMSE	信道容量	可达速率
S1	10^{-1}	10	1
S2	10^{-1}	5	1
S3	1	1	10^{-1}
S4	10	5	5
S5	10^{-5}	20	1
S6	10^{-3}	1	10^{-2}

这些指标包括每用户设备误码率、HMI、均方误差和可达速率。例如，对于“MSE-LMMSE”，我们优化的是所有用户设备和时隙上的均方误差之和，但展示的是每用户均方误差的结果。因此，我们无法保证“MSE-LMMSE”在每用户均方误差方面表现最佳，也无法保证“可达速率”在每用户设备可达速率方面表现最佳。我们将在第九节讨论这一问题。

表 III 总结了场景 S2（左）和场景 S3（右）的结果。这些场景的参数见表 I。两种情况下，我们关注性能指标的第 90 百分位数（参见第 VIII-A 节），包括每用户设备误码率、HMI、均方误差和可达速率。

在表 III（左）中，我们看到，整体性能最佳的方法“MSE-LMMSE”使 90% 的用户设备达到低于 20.3% 的误码率、不低于 0.546 比特/符号的 HMI，以及比该指标下表现最佳的现有方法“CSS”和“LoFi++(K=1)”低 0.319 dB 的均方误差，并且每用户设备可达速率至少为 0.379 比特/信道使用。正如预期，“AUA”基线在每用户设备误码率、HMI 和均方误差指标上的表现最差，但每用户设备可达速率表现最佳。

在表 III（右）中，我们展示了在误码率、HMI 和均方误差方面表现最佳的方法“可达速率”：它使 90% 的用户设备达到低于 19.1% 的误码率、不低于 0.6 比特/符号的 HMI，以及比该指标下表现最佳的现有方法“CSS”低 0.284 dB 的均方误差，并且每用户设备可达速率至少为 0.206 比特/信道使用。正如预期，“AUA”基线在每用户设备误码率、HMI 和均方误差指标上的表现最差，但每用户设备可达速率表现最佳。“AUA”较高的每用户设备可达速率表现也在 [1]，中得到了观察；在低信噪比（SNR）区域，其平均每用户设备速率表现良好。

D. 无小区大规模多用户多输入多输出场景

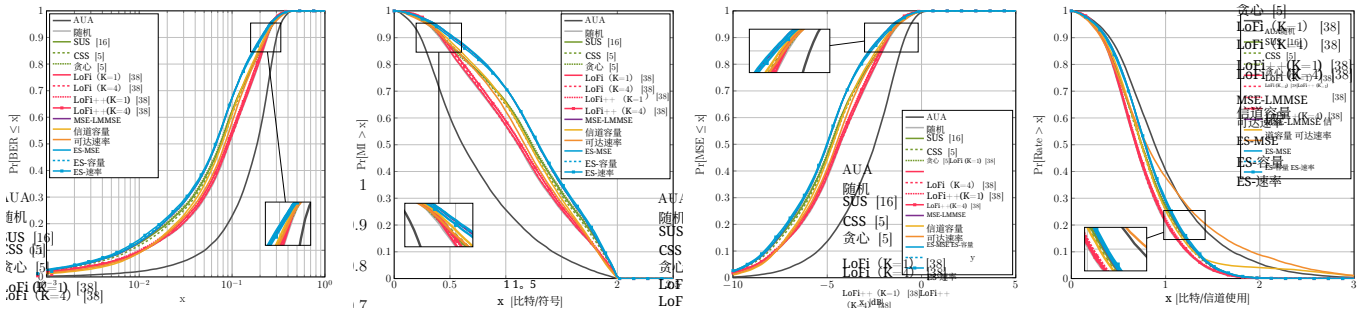
1）仿真设置：我们考虑一个工作在 1.9 GHz 载波频率下的无蜂窝大规模多用户 MIMO 系统。在该系统中，我们将每个用户设备的发射功率设为 $\rho_u = 10$ [mW] 用户设备发送采用格雷映射的 64-QAM 符号。在接入点处，我们采用天线间距为半波长的均匀线性阵列。接入点和用户设备的天线均为全向天线，高度分别为 15 m 和 1.65 m。信道向量针对面积为 200 m $\times$ 300m 的区域内 26,934 个用户设备位置生成。

在表 I 中，我们列出了无小区大规模多用户多输入多输出场景的三种不同系统规模：S4、S5 和 S6。其中，我们改变接入点数量 L、每个接入点的天线数量 n_{A}、待服务的用户设备数量 U 以及可用时隙数量 T。此外，考虑到已服务的用户设备数量和可用时隙数量，我们分别定义了每个用户设备所占时隙数的最小值和最大值，以及每个时隙内用户设备数的最小值和最大值，即 $T_{\min}$、$T_{\max}$、$U_{\min}$ 和$U_{\max}$。表格的最后一列给出了针对所述各场景运行用户设备调度算法（参见第 IV-B 节）时所采用的随机初始化次数。在表 II 中，我们列出了式 (28) 中的正则化参数 α，对应于表 I 所述的各个场景 S4、S5 和 S6。需要指出的是，由于计算复杂度较高，我们仅在规模最小的系统，即场景 S4 中测试 ES 基线。在该基线中，我们测试每一种可能的调度矩阵 $\mathbf{C}$，并选取使给定目标函数最小的矩阵。对于无小区大规模多用户多输入多输出场景中的 ES 算法，我们采用与第 VIII-C1 节相同的命名方案。

2) 结果：在图 2 中，我们展示了场景 S4 的结果，其参数见表 I。在图 2(a)–(d) 中可以看到，对于 90% 的用户设备而言，“MSE-LMMSE”表现最佳，其次是“ES-MSE”、“ES-容量”和“ES-速率”。正如预期，“AUA”基线在每用户设备误码率、HMI、均方误差和可达速率性能指标方面表现最差，如图 2(a)–(d) 所示。需要强调的是，正如第 VIII-C2 节所述，我们的方法——“MSE-LMMSE”、“信道容量”和“可达速率”——均未针对所呈现的指标（每用户设备误码率、HMI、均方误差和可达速率）进行优化。

在表 IV 中，我们分别总结了场景 S5（左）和场景 S6（右）的结果。这些场景的参数见表 I。在这两种情况下，我们考察了性能指标（每用户设备误码率、HMI、均方误差和可达速率）的第 90 百分位数（参见第 VIII-A 节）。

在表 IV（左）中，我们展示了总体性能最佳的方法“MSE-LMMSE”：对于 90% 的用户设备，该方法能实现低于 0.2% 的误码率、至少 5.873 比特/符号的 HMI、比在这一指标上表现最佳的现有方法“LoFi++ (K=4)”低 0.642 dB 的均方误差，以及至少 3.625 比特/信道使用的每用户设备可达速率。正如预期，“AUA”基线在每用户设备误码率、HMI、均方误差和可达速率指标上的表现最差。在表 IV（右）中，我们展示了表现最佳的方法



(a) 每用户未编码误比特率　(b) 每用户HMI　(c) 每用户均方误差　(d) 每用户设备可达速率

图 1. 场景 S1 中毫米波大规模多用户多输入多输出系统的误码率 (a)、HMI (b)、均方误差 (c) 和每用户设备可达速率 (d) 性能： $B = 16$ 接收天线、 $L = 1$ 接入点、 $n_{\mathrm{AP}} = 16$ 每个接入点的天线、 $U = 16$ 用户设备、 $T = 2$ 时隙、 $U_{\min} = 8$ 用户设备、 $U_{\max} = 10$ 用户设备、 $T_{\min} = 1$ 时隙和 $T_{\max} = 2$ 时隙。

表 III 两个毫米波大规模多用户多输入多输出系统的第 90 百分位性能。两个系统均具有 $B = 32$ 接收天线， $L = 1$ 接入点，$n_{\mathrm{AP}} = 32$ 每个接入点的天线， $U_{\min} = 16$ 用户设备， $U_{\max} = 20$ 用户设备，$T_{\min} = 1$ 时隙和 $T_{\max} = 2$ 时隙。

	场景 S2： $U = 32$ 用户设备， $T = 2$ 时隙				场景 S3： $U = 64$ 用户设备， $T = 4$ 时隙			
方法	BER	HMI [比特/符号]	均方误差 [dB]	速率 [比特/信道]	BER	HMI [比特/符号]	均方误差 [dB]	速率 [比特/ch]
AUA	0.292	0.260	-1.026	**0.432**	0.334	0.163	-0.675	**0.281**
随机	0.232	0.438	-1.835	0.333	0.219	0.484	-2.075	0.184
SUS [16]	0.230	0.444	-1.840	0.329	0.213	0.505	-2.130	0.181
CSS [5]	0.225	0.462	-1.922	0.338	0.208	0.525	-2.205	0.186
贪心 [5]	0.233	0.432	-1.794	0.326	0.219	0.483	-2.042	0.177
LoFi (K=1) [38]	0.231	0.440	-1.843	0.336	0.219	0.486	-2.083	0.183
LoFi (K=4) [38]	0.230	0.444	-1.864	0.335	0.218	0.487	-2.090	0.183
LoFi++ (K=1) [38]	0.226	0.459	-1.922	0.344	–	–	–	–
LoFi++(K=4) [38]	0.227	0.452	-1.896	0.342	–	–	–	–
MSE-LMMSE	**0.203**	**0.546**	**-2.241**	0.379	0.198	0.580	-2.430	0.206
信道容量	0.234	0.431	-1.791	0.319	**0.191**	0.593	-2.485	0.209
可达速率	0.227	0.455	-1.888	0.330	**0.191**	**0.600**	**-2.489**	0.206

“MSE-LMMSE”和“可达速率”总体性能最佳：对于 90% 的用户设备，两种方法分别能实现低于 0.6% 和 0.5% 的误码率、至少 5.759 和 5.801 比特/符号的 HMI、比在这一指标上表现最佳的现有方法“CSS”分别低 0.919 和 0.821 dB 的均方误差，以及至少 1.739 和 1.739 比特/信道使用的每用户设备可达速率。正如预期，“AUA”基线在每用户设备误码率、HMI、均方误差和可达速率指标上的表现最差。“AUA”在每用户设备可达速率方面的表现不佳，也在 [1]，中有所体现：其在高信噪比区域的平均每用户设备速率表现不佳。

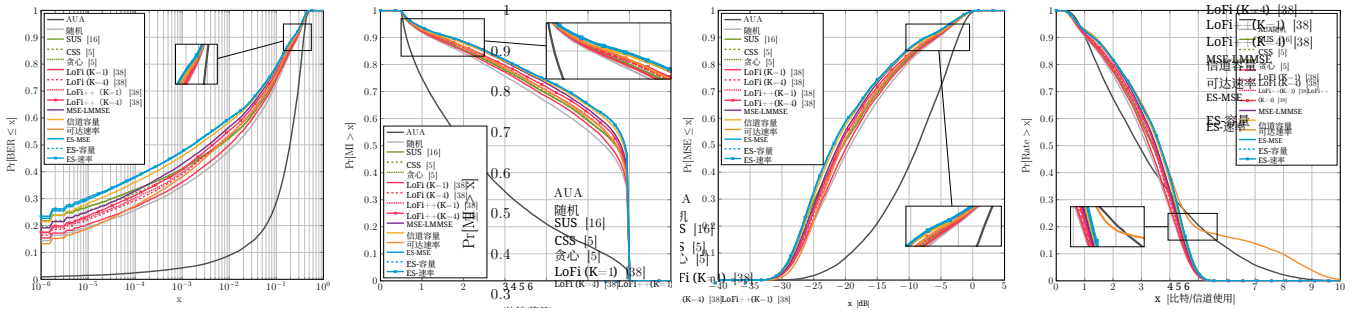
E. 总结

根据这些结果，我们得出以下结论。即使在过载场景（S3 和 S6）中，我们的用户设备调度方法也分别取得了与 S2 和 S5 相当的性能。尽管 S3 和 S6 中的用户设备数量多于接收天线数量，但更多可用时隙仍能保证良好的服务质量。这表明， B 与 U 之间的比值不如 B 与活跃用户设备数量之间的比值重要；在 S2、S3、S5 和 S6 中，对于 $B = 32$ 根接收天线，该比值被限制在 $U_{\min} = 16$ 至$U_{\max} = 20$ 之间。

毫米波大规模多用户多输入多输出场景中的性能相对较差，这是因为这些频段的信道条件不利，导致系统在低信噪比下运行；采用用户设备侧波束成形可以缓解这一问题，但会带来额外的复杂度代价——这留待未来工作研究。相比之下，无小区大规模多用户多输入多输出场景的性能显著更好，因为用户设备很可能至少靠近一个接入点，从而能够在比集中式基站部署更高的信噪比工作区域运行。

优化目标会显著影响得到的用户设备调度方案。例如，“MSE-LMMSE”目标倾向于在每个时隙中调度最少数量的用户设备，即 $U_{\min}$，以减少估计误差；而“信道容量”则受益于更多的同时传输，因此倾向于 $U_{\max}$。“可达速率”体现了在为每个用户设备分配更多时隙与限制每个时隙内的用户设备间干扰之间的权衡。

与低复杂度方法“LoFi”和“LoFi++”相比， [38], 我们的方法在毫米波和无蜂窝大规模多用户多输入多输出系统中均取得了显著的性能提升。此外，“LoFi”和“LoFi++”在 [38] 中针对单一目标函数提出，在资源分配方面缺乏灵活性；而我们的框架支持多个目标，并能灵活设置每个时隙中的用户设备数量以及每个用户设备占用的时隙数量。它们的目标函数也忽略了信道估计误差。尽管“LoFi++”对“LoFi”进行了改进，但它仍有其他局限，包括超过两个时隙后可扩展性降低（$T > 2$），且不支持并行化。



(a) 每用户未编码误比特率　(b) 每用户设备 HMI (c) 每用户设备均方误差　(d) 每用户设备可达速率

图 2. 场景 S4 中无小区多用户多输入多输出系统的误码率 (a)、HMI (b)、均方误差 (c) 和每用户设备可达速率 (d) 性能： $B = 16$ 接收天线， $L = 8$ 接入点， $n_{\mathrm{AP}} = 2$ 每个接入点的天线， $U = 16$ 用户设备， $T = 2$ 时隙， $U_{\min} = 8$ 用户设备， $U_{\max} = 10$ 用户设备， $T_{\min} = 1$ 时隙，以及 $T_{\max} = 2$ 时隙。

表 IV 两个无小区大规模多用户多输入多输出系统的第 90 百分位性能。两个系统均具有 $B = 32$ 接收天线、 $L = 16$ 接入点、$n_{\mathrm{AP}} = 2$ 每个接入点的天线、 $U_{\mathrm{MIN}} = 16$ 用户设备、 $U_{\mathrm{MAX}} = 20$ 用户设备、$T_{\mathrm{MIN}} = 1$ 时隙和 $T_{\mathrm{MAX}} = 2$ 时隙。

	场景 S5： $U = 32$ 用户设备 , $T = 2$ 时隙				场景 S6： $U = 64$ 用户设备 , $T = 4$ 时隙			
方法	BER	HMI [比特/符号]	均方误差 [dB]	速率 [bits/ch]	误码率	HMI [比特/符号]	均方误差 [dB]	速率 [bits/ch]
AUA	0.275	1.257	−5.742	2.03	0.421	0.517	−1.647	0.555
随机	0.013	5.394	−19.209	3.199	0.021	5.131	−18.314	1.525
SUS [16]	0.009	5.581	−19.96	3.321	0.008	5.592	−20.033	1.663
CSS [5]	0.008	5.602	−20.044	3.339	0.008	5.593	−20.034	1.664
贪心 [5]	0.009	5.558	−19.863	3.304	0.009	5.557	−19.883	1.652
LoFi (K=1) [38]	0.009	5.555	−19.85	3.301	0.016	5.315	−18.932	1.575
LoFi (K=4) [38]	0.005	5.735	−20.747	3.45	0.011	5.502	−19.64	1.633
LoFi++(K=1) [38]	0.005	5.731	−20.74	3.448	–	–	–	–
LoFi++ (K=4) [38]	0.004	5.8	−21.171	3.522	–	–	–	–
MSE-LMMSE	**0.002**	**5.873**	**−21.813**	**3.625**	0.006	5.759	**−20.953**	**1.739**
信道容量	0.005	5.745	−20.821	3.461	0.015	5.349	−19.061	1.586
可达速率	0.003	5.82	−21.349	3.585	**0.005**	**5.801**	−20.855	**1.739**

IX. 局限性与未来工作

我们总结了所提出的用户设备调度框架的局限性。尽管我们基于优化的调度方法相较于现有方法具有明显优势，例如支持灵活的约束和目标函数，并能通过全局求解联合分配用户设备至各时隙，但也存在显而易见的缺点。首先，灵活性与计算效率之间存在权衡：与贪心或启发式方法相比，我们的框架可能具有更高的复杂度；比较见 [38, 表 I] 。由于我们的用户设备调度框架在复杂度增加的代价下，性能优于所有纳入比较的基线（ES 除外）及现有方法，因此我们的方法可用于为基于机器学习的调度方法生成高质量训练数据 [30], [31], [33], [36]。此外，由于我们的框架最小化的是非凸目标函数，基于梯度下降的方法可能仅收敛至局部最小值，且无法提供最优性保证。因此，为评估我们方法的性能，我们将其与进化策略方法进行了比较。

另一点是，此处考虑的目标函数优化的是所有用户设备和时隙上的指标，但未明确考虑公平性。更实用的设计可以优先考虑信道条件较差的用户设备，例如采用比例公平，在瞬时可达速率与长期平均吞吐量之间取得平衡 [15], [59], [60]。将此类目标函数纳入我们的框架，是未来的研究方向。

最后，尽管以用户设备为中心的架构在无蜂窝大规模多用户多输入多输出系统中较为常见，但我们的框架并非针对这一设置开发，不过对其进行扩展似乎是可行的。频率调度已在先前的工作中得到研究，也是另一项自然的扩展。此外，我们的方法未纳入考虑用户设备调度的联合功率控制方案；鉴于联合优化功率控制与调度的难度，这同样留待未来研究。

X. 结论

我们提出了一种新颖的基于优化的用户设备调度框架，用于多用户多输入多输出系统。我们的框架支持一系列用于规定用户设备资源分配的目标函数和约束。我们详细介绍了问题表述的各个方面，包括目标函数及其相应梯度的呈现，以及近端算子的推导。我们使用真实的毫米波和 6 GHz 以下无蜂窝大规模多用户多输入多输出信道向量，对框架的性能进行了评估。

在具有 $B = 16$和 $U = 16$的小型 MU-MIMO 毫米波场景中，我们观察到，所提出方法的性能通常接近穷举搜索，并且在每用户设备误码率、HMI 和均方误差方面优于所有纳入考量的基准算法以及现有方法。在具有 $U = B$的大型系统中，我们证明了

所提出方法在每用户设备误码率、HMI 和均方误差方面优于所有纳入考量的基准算法以及现有方法。在 $U > B$ 的过载场景中，我们证明，如果将用户设备的请求分散到更多时隙，性能可以与 $U = B$ 时相近。

在包含 $B = 16$ 和 $U = 16$ 的小规模无小区场景中，我们发现，所提出的方法的性能最接近穷举搜索，并且在每用户设备误码率、HMI、均方误差和可达速率方面均优于基准算法和现有方法。在包含 $U = B$ 的较大系统中，我们已证明，所提出的方法在每用户设备误码率、HMI、均方误差和可达速率方面均优于基准算法和现有方法。在 $U > B$ 的过载场景中，我们已证明，如果将用户设备的请求分散到更多时隙中，性能可以与 $U = B$ 时相近。此外，正如在采用用户设备调度的无小区大规模多用户多输入多输出场景中所见，我们发现，将用户设备分散到更多时隙中，同时将接入点分布到更大的区域，可以带来性能提升，这归因于用户设备与接入点之间的邻近性。

附录A
式（36）的卡鲁什-库恩-塔克（KKT）条件

式（36）所述问题的KKT条件为：

1) $p_i - q_i - \mu_i + \upsilon_i - \gamma + \Gamma = 0, i = 1, \ldots, \theta,$
2) $\mu_i p_i = 0,\ \upsilon_i(p_i - 1) = 0, \gamma\left(\ell_{\min} - \sum_{i=1}^{M} p_i\right) = 0, \Gamma\left(\sum_{i=1}^{M} p_i - \ell_{\max}\right) = 0, i = 1, \ldots, M,$
3) $p_i \geq 0, p_i \leq 1, \sum_{i=1}^{M} p_i \geq \ell_{\min},\ \sum_{i=1}^{M} p_i \leq \ell_{\max},\ i = 1, \ldots, M,$
4) $\mu_i \geq 0,\ \upsilon_i \geq 0,\ \gamma \geq 0,\ \Gamma \geq 0,\ i = 1, \ldots, M_0$。

变量 p_i 对于 $i = 1, \ldots, M$，可能属于以下三种情况之一：$p_i = 1$、$p_i \neq 0, 1$，以及 $p_i = 0$。对于 $p_i \neq 0, 1$，我们得到 $\mu_i = 0$ 和 $\upsilon_i = 0$。因此，根据第三个条件可得：

1) $\ell_{\min} - p_i < \sum_{j=1, j\neq i}^{M} p_j < \ell_{\max} - p_i$ leading to $\gamma = 0$, $\Gamma = 0$, and $p_i = q_i$,
2) $\sum_{j=1, j\neq i}^{M} p_j = \ell_{\min} - p_i$ leading to $\Gamma = 0$ and $p_i = q_i + \gamma = 0$,
3) $\sum_{j=1, j\neq i}^{M} p_j = \ell_{\max} - p_i$ leading to $\gamma = 0$ and $p_i = q_i - \Gamma = 0$.

根据 [61]，对于 $q_i > q_j$，若 $p_i = 0$，则$p_j = 0$。类似地，在此情况下，对于 $q_i < q_j$，若 $p_i = 1$，则 $p_j = 1$。将$\mathbf{q}$ 按降序排列，可得$\mathbf{q} = [q_1, q_2, \ldots, q_M]$，其中 $q_1 > \cdots > q_M$。若按照$\mathbf{q}$对$\mathbf{p}$进行排序，可得

$$\mathbf{p} = \left[\underbrace{1, \ldots, 1}_{M_1}, \underbrace{p_{M_1+1}, \ldots, p_{M-M_0}}_{M_1+1, \ldots, M-M_0}, \underbrace{0, \ldots, 0}_{M_0}\right], \tag{44}$$

其中，M_1 和 M_0 分别表示$\mathbf{p}$ 中等于 1 和 0 的元素个数。

当 $p_i \neq \{0, 1\}$ 时重新表述问题，并将其应用于不等式约束，可得问题的解如下：

1) $\ell_{\min} < p_i + \sum_{j=1, j\neq i}^{M} p_j < \ell_{\max}$, which leads to $p_i = q_i$,
2) $p_i + \sum_{j=1, j\neq i}^{M} p_j = \ell_{\min}$ which leads to $p_i = q_i + \frac{1}{M - M_0 - M_1}\left(\ell_{\min} - M_1 - \sum_{j=M_1+1}^{M-M_0} q_j\right)$,
3) $p_i + \sum_{j=1, j\neq i}^{M} p_j = \ell_{\max}$ which leads to $p_i = q_i + \frac{1}{M - M_0 - M_1}\left(\ell_{\max} - M_1 - \sum_{j=M_1+1}^{M-M_0} q_j\right)$,

其中 $i = M_1 + 1, \ldots, M - M_0$。

参考文献

[1] V. Palhares和C. Studer，“面向毫米波大规模多用户多输入多输出系统的基于优化的用户调度框架”，载于*IEEE*无线通信信号处理进展国际研讨会（*SPAWC*），2022年7月，第1–5页。[2] A. L. Swindlehurst、E. Ayanoglu、P. Heydari和F. Capolino，“毫米波大规模 MIMO：下一次无线革命？”*IEEE*通信杂志，第52卷，第9期，第56–62页，2014年9月。[3] J. Zhang、E. Björnson、M. Matthaiou、D. W. K. Ng、H. Yang和D. J. Love，“面向超越5G的前瞻性多天线技术”，*IEEE*通信选定领域期刊，第38卷，第8期，第1637–1660页，2020年8月。[4] H. Q. Ngo、A. Ashikhmin、H. Yang、E. G. Larsson和T. L. Marzetta，“无小区大规模MIMO与小小区的比较”，*IEEE*无线通信汇刊，第16卷，第3期，第1834–1850页，2017年3月。[5] J. Choi、G. Lee和B. L. Evans，“采用低分辨率模数转换器的毫米波混合波束成形系统用户调度”，*IEEE*无线通信汇刊，第18卷，第4期，第2401–2414页，2019年3月。[6] J. Cui、Y. Liu、Z. Ding、P. Fan和A. Nallanathan，“毫米波NOMA系统的最优用户调度与功率分配”，*IEEE*无线通信汇刊，第17卷，第3期，第1502–1517页，2017年12月。[7] S. He、Y. Wu、D. W. K. Ng和Y. Huang，“毫米波通信中模拟波束与用户调度的联合优化”，*IEEE*通信快报，第21卷，第12期，第2638–2641页，2017年8月。[8] E. Kim、J. Kwak和S. Chong，“毫米波网络中子阵列架构的虚拟波束成形与用户调度”，*IEEE*通信快报，第23卷，第1期，第168–171页，2018年10月。[9] S. Lee和J. H. Lee，“采用随机波束成形的节能毫米波NOMA系统中用户调度与功率分配的联合设计”，载于*IEEE*秋季车辆技术会议（*VTC-Fall*），2018年8月，第1–5页。[10] P. Paul、H. Wu和C. Xin，“BOOST：面向波束成形毫米波网络的用户关联与调度框架”，*IEEE*移动计算汇刊，第20卷，第10期，第2924–2935页，2020年5月。[11] H. Wu、D. Liu、W. Wu、C. Na和M. Liu，“面向毫米波大规模 MIMO 混合波束成形系统的低复杂度两阶段用户调度方案”，载于*IEEE*计算机与通信国际会议（*ICCC*），2017年12月，第945–951页。[12] J. Zhu、Q. Li、Z. Liu、H. Chen和H. Vincent Poor，“混合毫米波MIMO-NOMA系统的增强型用户分组与功率分配”，*IEEE*无线通信汇刊，第1–1页，2021年9月。[13] P. Zhao和Z. Wang，“采用双层PS网络的多用户毫米波系统中用户调度与混合预编码的联合设计”，载于*IEEE*全球通信会议（*GLOBECOM*），2018年12月，第1–6页。[14] H. Xu、T. Zhao、S. Zhu、D. Lv和J. Zhao，“混合波束成形架构下毫米波大规模 MIMO 的凝聚式分组调度”，载于国际通信技术会议（*ICCT*），2018年10月，第347–351页。[15] M. Zhang、Y. Guo、L. Salaün、C. W. Sung和C. S. Chen，“下行毫米波多用户MISO-NOMA系统的比例公平调度”，*IEEE*车辆技术汇刊，第71卷，第6期，第6308–6321页，2022年6月。[16] T. Yoo和A. Goldsmith，“采用迫零波束成形的多天线广播调度最优性”，*IEEE*通信选定领域期刊，第24卷，第3期，第528–541页，2006年3月。[17] G. Lee和Y. Sung，“大规模多用户MIMO广播信道中用户调度的新方法”，*IEEE*通信汇刊，第66卷，第4期，第1481–1495页，2017年12月。[18] S. Mashdour、R. C. De Lamare和J. a. P. S. H. Lima，“面向多小区与无小区大规模MIMO系统、采用增强型贪心技术的多用户调度”，载于*IEEE*春季车辆技术会议（*VTC-Spring*），2022年6月，第1–5页。[19] S. Mashdour、R. C. de Lamare和J. P. S. H. Lima，“集群化无小区大规模MIMO系统中的增强型子集贪心多用户调度”，*IEEE*通信快报，第27卷，第2期，第610–614页，2023年2月。

[20] F. Riera-Palou和G. Femenias，“无小区大规模 MIMO 网络中的权衡：预编码、功率分配与调度”，载于国际先进技术、系统与电信服务会议（TELSIKS），2019年10月，第158–165页。[21] H. Liu、X. Gong、M. Huang、Y. Chen、X. Yuan、K. Chen和S. Yang，“无小区可见光通信网络中面向干扰管理的接入点分组与用户聚类联合设计”，《光学与激光技术》，第156卷，第108465页，2022年12月。[22] K. Ko、I. Byun、J. Lee和W. Shin，“多用户 MIMO 系统中面向误码率最小化的功率分配与调度联合技术”，IEEE Access，第9卷，第66 675–66 686页，2021年4月。[23] X. Gong和G. Wu，“无小区大规模MIMO中考虑用户满意率的动态用户调度”，载于IEEE中国国际通信会议（ICCC），2022年8月，第100–105页。[24] N. Parikh和S. Boyd，《近端算法》。Now Foundations and Trends，2014年。[25] A. Farsaei、A. Alvarado、F. M. J. Willems和U. Gustavsson，“采用最大-最小功率控制的视距大规模 MIMO 改进剔除算法”，IEEE通信快报，第23卷，第6期，第1109–1112页，2019年4月。[26] H. Yang，“大规模 MIMO 中的用户调度”，载于IEEE无线通信信号处理进展国际研讨会（SPAWC），2018年6月，第1–5页。[27] X. Meng、Y. Chang、Y. Wang和J. Wu，“大规模 MIMO 上行网络中基于多用户分组的调度算法”，载于IEEE计算机与通信国际会议（ICCC），2018年12月，第409–413页。[28] Z. Jiang、S. Chen、S. Zhou和Z. Niu，“基于波束的大规模 MIMO 下行链路中用户调度与波束选择优化的联合设计”，IEEE无线通信汇刊，第17卷，第4期，第2190–2204页，2018年1月。[29] B. Hu、C. Hua、C. Chen、X. Ma和X. Guan，“MUBFP：面向MIMO 系统总容量最大化的多用户波束成形与分区”，IEEE车辆技术汇刊，第66卷，第1期，第233–245页，2016年3月。[30] C.-H. Hsu和Z. Ding，“宽带 MU-MIMO 系统中面向资源分配与用户调度的无监督学习”，IEEE通信学会开放期刊，第5卷，第2240–2256页，2024年4月。[31] C. Feres和Z. Ding，“大规模多天线系统中用户调度的无监督学习范式”，IEEE无线通信汇刊，第22卷，第5期，第2932–2945页，2023年5月。[32] S. M. Hosseini、S. Shahsavari和C. Rosenberg，“混合毫米波大规模 MIMO 系统中的多用户调度”，arXiv，第abs/2109.05681卷，2021年9月。 [在线]。可获取：https: //arxiv.org/abs/ 2109 05681 。Z. Zou、S. Zhao、G. Huang和D. Tang，[33]“下行毫米波通信中用户调度与模拟波束选择的新设计”，IEEE物联网期刊，第9卷，第6期，第4168–4178页，2022年3月。[34] Z. Sha、S. Chen和Z. Wang，“毫米波蜂窝网络中近乎无干扰的时空用户调度”，IEEE无线通信汇刊，第21卷，第8期，第6372–6386页，2022年8月。[35] X. Gao、X. Wu、Z. Zhang和D. Liu，“毫米波大规模 MIMO 系统中的低复杂度联合用户调度与混合波束成形”，载于IEEE个人、室内与移动无线电通信国际研讨会（PIMRC），2020年8月，第1–6页。[36] N. Chukhno、O. Chukhno、S. Pizzi、A. Molinaro、A. Iera和G. Araniti，“毫米波网络中D2D辅助多播调度的无监督学习”，载于IEEE宽带多媒体系统与广播国际研讨会（BMSB），2021年8月，第1–6页。[37] J. Lee和E. Ekici，“存在大量移动用户的毫米波网络中的用户调度与波束对准”，IEEE无线通信汇刊，第20卷，第10期，第6481–6492页，2021年10月。[38] A. Gallyas-Sanhueza、G. Marti、V. Palhares、R. Wiesmayr和C. Studer，“多用户 MIMO 无线系统的LoFi用户调度”，载于IEEE声学、语音与信号处理国际会议（ICASSP），2024年4月，第9261–9265页。[39] D. Peng、A. Bandi、Y. Li、S. Chatzinotas和B. Ottersten，“天地一体化通信中的混合波束成形、用户调度与资源分配”，IEEE车辆技术汇刊，第70卷，第9期，第8868–8882页，2021年7月。[40] S. Eddine Hajri、J. Denis和M. Assaad，“通过空间用户分组增强无小区大规模MIMO中的有利传播”，载于IEEE无线通信信号处理进展国际研讨会（SPAWC）
，2018年6月，第 1 页 –5.
[41] J. Denis 和 M. Assaad，《提升无小区大规模 MIMO 网络性能：一种用户调度方法》，IEEE无线通信汇刊，第20卷，第11期，第7360–7374页，2021年11月。[42] J. Chen 和 Z. Wang，《无小区可见光通信网络中用于干扰管理的用户调度与预编码联合设计》，载于 IEEE全球通信会议（GLOBECOM），2019年12月，第1–6页。[43] H. A. Ammar、R. Adve、S. Shahbazpanahi、G. Boudreau 和 K. V. Srinivas，《面向用户中心化无小区MIMO网络的分布式资源分配优化》，IEEE无线通信汇刊，第21卷，第5期，第3099–3115页，2022年5月。[44] ——，《采用用户中心化聚类的多用户无小区MIMO网络下行资源分配》，IEEE无线通信汇刊，第21卷，第3期，第1482–1497页，2022年3月。[45] F. Göttsch、N. Osawa、I. Kanno、T. Ohseki 和 G. Caire，《用户中心化无小区大规模 MIMO 无线网络中的公平性调度》，IEEE无线通信汇刊，第23卷，第9期，第11 942–11 957页，2024年9月。[46] F. Fang、H. Zhang、J. Cheng、S. Roy 和 V. C. M. Leung，《不完美信道状态信息下能效NOMA系统的用户调度与功率分配联合优化》，IEEE通信选定领域期刊，第35卷，第12期，第2874–2885页，2017年11月。[47] H. Song、T. Goldstein、X. You、C. Zhang、O. Tirkkonen 和 C. Studer，《无小区大规模 MU-MIMO 系统中的信道估计与数据检测联合处理》，IEEE无线通信汇刊，第1–1页，2021年11月。[48] S. Buzzi 和 C. D’Andrea，《无小区大规模 MIMO：用户中心化方法》，IEEE无线通信快报，第6卷，第6期，第706–709页，2017年12月。[49] S. Dutta、C. N. Barati、D. Ramirez、A. Dhananjay、J. F. Buckwalter 和 S. Rangan，《毫米波数字波束成形的应用价值》，IEEE无线通信汇刊，第19卷，第2期，第756–770页，2020年2月。[50] K. Roth、H. Pirzadeh、A. L. Swindlehurst 和 J. A. Nossek，《面向多用户及不完美信道状态信息，采用低分辨率 ADC 的混合波束成形与数字波束成形比较》，IEEE信号处理选定主题期刊，第12卷，第3期，第484–498页，2018年6月。[51] P. Skrimponis、S. Dutta、M. Mezzavilla、S. Rangan、S. H. Mirfarshbafan、C. Studer、J. Buckwalter 和 M. Rodwell，《移动毫米波和亚太赫兹接收机的功耗分析》，载于 6G无线峰会（6G SUMMIT），2020年3月，第1–5页。[52] S. H. Mirfarshbafan、A. Gallyas-Sanhueza、R. Ghods 和 C. Studer，《大规模MIMO毫米波系统的波束空间信道估计：算法与超大规模集成电路设计》，IEEE电路与系统汇刊I，第67卷，第12期，第5482–5495页，2020年9月。[53] T. Goldstein、C. Studer 和 R. G. Baraniuk，《前向-后向分裂及其 FASTA 实现指南》，arXiv，第abs/1411.3406卷，2016年12月。 [在线]。可访问：http: //arxiv.org/abs/ 1411 3406 。B. Zhou、L. Gao 和 Y.-H. Dai，《具有自适[54]应步长的梯度方法》，计算优化与应用，第35卷，第1期，第69–86页，3月。2006 [ ] 10 1007 10589 006 6446 0 。在[55]线。可访问：https://doi.org/ . /s - - - O. Castañeda、T. Goldstein 和 C. Studer，《大型 SIMO 无线系统中信道估计与数据检测联合处理的 VLSI 设计》，IEEE电路与系统汇刊I，第65卷，第3期，第1120–1132页，2018年10月。[56] J. Douglas 和 H. H. Rachford，《关于二维和三维空间变量下热传导问题的数值解》，美国数学会汇刊，第82卷，第2期，第421–439页，1956年7月。[57] S. Boyd 和 L. Vandenberghe，《凸优化》。剑桥大学出版社，2004年。[58] （2021）。[在线]。可访问：https://www.remcom.com/ wireless-insite-em-propagation-software/
[59] H. Kushner 和 P. Whiting，《一般条件下比例公平共享算法的收敛性》，IEEE无线通信汇刊，第3卷，第4期，第1250–1259页，2004年7月。[60] F. P. Kelly、A. K. Maulloo 和 D. K. H. Tan，《通信网络中的速率控制：影子价格、比例公平与稳定性》，《运筹学会杂志》，第49卷，第3期，第237–252页，1998年4月。[61] J. Duchi、S. Shalev-Shwartz、Y. Singer 和 T. Chandra，《用于高维学习的 l_1-球高效投影》，载于国际机器学习会议，ICML ’08 系列。美国纽约州纽约市：计算机协会，2008年7月，第272–279页。 [在线]。可访问：https://doi.org/10.1145/1390156.1390191