差分隐私随机梯度下降优化:方法、理论与应用

沙海潮[1),2)] 孙丽超[1),2)] 刘艺璇[1),2)] 薛大暄[1),2)] 吴云乘[1),2)]
李翠平[1),2)] 陈 红[1),2)]

[1)](数据工程与知识工程教育部重点实验室(中国人民大学) 北京 100872)
[2)](中国人民大学信息学院 北京 100872)

摘 要 随着差分隐私(differential privacy,DP)在人工智能安全与数据治理中的广泛应用,差分隐私随机梯度下降(differential privacy stochastic gradient descent,DP-SGD)凭借其完备和可量化的理论性质,已成为当前最主流的隐私保护机器学习领域优化算法之一,受到众多学者的关注。然而,尽管已有围绕DP-SGD的大量相关工作,但尚未形成全面和统一的研究框架。为促进其技术改进与理论发展,本文对DP-SGD的优化技术与算法理论进行了系统梳理与分类总结。首先,本文依据DP-SGD的基本流程,构建了一个涵盖前沿优化技术组件与理论研究的三层次分类框架:一是针对随机采样模块,梳理了重要性采样、个性化采样策略及其对隐私放大的影响,总结了迭代权重选择的改进策略,并分析了针对数据采样和迭代选择的理论研究;二是针对梯度裁剪模块,分析了梯度裁剪的优化技术,将其归纳为三类改进方案:解析式裁剪机制、自适应裁剪函数及基于几何感知裁剪,并聚焦裁剪偏差对于隐私算法收敛速率的影响;三是针对噪声扰动模块,本文基于辅助知识驱动的角度将现有降维方案细分为"无向引导"与"有向引导"两大类,重点分析现有技术对DP冗余噪声的控制效果,并研究了针对差分隐私噪声缓解的维度无关理论和混合数据优化理论。其次,本文以新兴的大语言模型训练场景为例,探讨了全参数训练、参数高效微调和无参数化推断三种模式下的DP-SGD应用与发展。最后,本文讨论了当下差分隐私随机梯度下降优化的难点与挑战,并展望了其在新兴场景中的应用方向。综上,本文全面回顾并系统总结了DP-SGD的理论研究、技术演进和现实应用,旨在为后续差分隐私优化算法的深入研究与实际部署提供有价值的参考。

关键词 差分隐私;随机梯度下降;梯度裁剪;梯度扰动;大语言模型隐私保护
中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2026.01307

Differentially Private Stochastic Gradient Descent Optimization: Methodology, Theory and Application

SHA Hai-Chao[1),2)] SUN Li-Chao[1),2)] LIU Yi-Xuan[1),2)] XUE Da-Xuan[1),2)]
WU Yun-Cheng[1),2)] LI Cui-Ping[1),2)] CHEN Hong[1),2)]

[1)](Key Laboratory of Data Engineering and Knowledge Engineering of Education (Renmin University), Beijing 100872)
[2)](School of Information, Renmin University, Beijing 100872)

Abstract With the widespread adoption of differential privacy (DP) in artificial intelligence security and data governance, Differentially Private Stochastic Gradient Descent (DP-SGD) has emerged as one of the most influential and widely deployed optimization algorithms in privacy-preserving machine learning. Owing to its rigorous theoretical guarantees and quantitatively

收稿日期:2025-07-21;在线发布日期:2026-03-11。本课题得到国家自然科学基金联合基金重点项目(U23A20299,U24B20144)、国家自然科学基金专项项目(62441230)、国家自然科学基金面上项目(62172424,62276270,62322214)资助。**沙海潮**,博士研究生,主要研究领域为机器学习隐私保护、差分隐私。E-mail:sha@ruc.edu.cn。**孙丽超**,博士研究生,主要研究领域为面向生成式模型的隐私攻击。**刘艺璇**,博士,主要研究领域为机器学习中的隐私保护。**薛大暄**,博士研究生,主要研究领域为隐私保护神经网络架构搜索。**吴云乘**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为隐私保护、联邦计算。**李翠平**,博士,教授,博士生导师,中国计算机学会(CCF)杰出会员,主要研究领域为大数据分析与挖掘。**陈 红**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)杰出会员,主要研究领域为大数据隐私保护。E-mail:chong@ruc.edu.cn。

measurable privacy loss, DP-SGD has become the de facto standard for enforcing differential privacy in modern learning systems, attracting substantial attention from both academia and industry. Nevertheless, despite a rapidly growing body of literature surrounding DP-SGD, existing studies are often fragmented and lack a comprehensive and unified framework that systematically organizes its theoretical foundations, algorithmic design principles, and optimization techniques. To facilitate further theoretical advancement and practical deployment, this paper presents a systematic review and structured categorization of DP-SGD, covering both its algorithmic foundations and recent optimization strategies. In particular, we construct a three-level taxonomy aligned with the core DP-SGD pipeline, encompassing sampling, gradient clipping, and noise perturbation, and integrate theoretical insights with state-of-the-art algorithmic developments. First, from the perspective of the sampling module, we review a range of advanced sampling strategies, including importance-based sampling and personalized sampling mechanisms, and analyze their roles in privacy amplification and optimization efficiency. We further summarize recent progress in iterative weight selection and adaptive sampling policies, and examine the theoretical guarantees associated with data sampling and iteration selection under differential privacy constraints. These studies reveal how refined sampling mechanisms can simultaneously improve utility and reduce effective privacy loss. Second, regarding the gradient clipping module, we conduct a comprehensive analysis of gradient clipping techniques, which are central to controlling sensitivity in DP-SGD. We categorize existing methods into three major classes: analytic clipping mechanisms, adaptive clipping functions, and geometry-aware clipping strategies. Beyond algorithmic design, we place particular emphasis on understanding the impact of clipping-induced bias on optimization dynamics and convergence rates. By synthesizing recent theoretical results, we highlight how refined clipping strategies can mitigate optimization degradation while preserving strong privacy guarantees. Third, for the noise perturbation module, we revisit noise reduction and optimization techniques from the perspective of auxiliary knowledge exploitation. Based on whether prior structural or distributional information is leveraged, we classify existing approaches into two broad categories: undirected guidance, which relies on isotropic or structure-agnostic noise control, and directed guidance, which incorporates geometry, subspace, or task-aware information to reduce redundant noise. We analyze the effectiveness of these methods in controlling DP-induced noise inflation and discuss emerging theoretical results, including dimension-independent privacy guarantees and hybrid data optimization theories that combine public and synthetic auxiliary datasets. Beyond algorithmic taxonomy, this survey further investigates the role of DP-SGD in the emerging paradigm of large language models (LLMs). We examine its applicability and limitations under three representative settings: full-parameter DP training, parameter-efficient DP fine-tuning, and non-parameterized DP inference. By consolidating recent advances, we illustrate how DP-SGD must be adapted to address the scale, sensitivity, and optimization challenges unique to large models. Finally, the paper discusses open challenges and future research directions for DP-SGD optimization, including utility degradation under tight privacy level, scalability in high-dimensional regimes, robustness to data heterogeneity, and compatibility with modern training paradigms. We also outline promising application directions in emerging scenarios, such as DP theory under complex data environments, DP optimization with generative model, and privacy-aware data governance with DP auditing. In summary, this paper provides a comprehensive and structured overview of DP-SGD, spanning theoretical foundations, algorithmic innovations, and real-world applications. By offering a unified perspective, the paper aims to support future research on differential privacy optimization and accelerate the deployment of DP learning in practical settings.

Keywords differential privacy; stochastic gradient descent; gradient clipping; gradient perturbation; privacy preservation of large language model

1 引 言

在过去的二十年里，随着机器学习模型在人脸识别、推荐系统、医疗诊断、法律咨询等高敏感领域的广泛部署，其在服务过程中频繁接触用户的个人隐私信息，引发了关于模型记忆能力与数据泄露的严重隐私担忧。差分隐私(differential privacy，DP)[1-2]凭借其严谨完备的统计学理论性质，被广泛地应用于隐私保护机器学习领域中。差分隐私随机梯度下降(differentially private stochastic gradient descent，DP-SGD)是在此基础上建立的优化方案，由于其自身隐私保证可验证性[3]的优势，被逐步确立为面向人工智能合规性和安全性的行业通用方案。标准的DP-SGD通过逐样本梯度裁剪和准确量化的差分隐私噪声注入来扰动优化过程中的梯度，从而实现对机器学习训练数据的保护。围绕这一范式，为了确保隐私保护效果并进一步提升算法可用性，众多研究持续推动DP-SGD的算法优化与理论完善，努力推动其在实际场景中的落地与部署，并不断将其拓展到更前沿的大语言模型(large language models，LLMs)研究中。尽管围绕DP-SGD的技术快速发展，目前却缺乏专门针对DP-SGD的系统性综述，这给研究者追踪最新技术进展和研究方向带来了困难。

隐私保护机器学习常见的扰动形式包括针对模型参数的输出扰动(Output Perturbation)、针对模型优化目标函数的目标函数扰动(Objective Perturbation)及针对模型更新梯度的梯度扰动(Gradient Perturbation)。DP-SGD本质上是一种基于梯度扰动的随机优化算法，通过在每轮参数更新时对裁剪后的每个样本有界梯度添加随机噪声，从而在随机梯度下降(stochastic gradient descent，SGD)基础上实现差分隐私保护。与效用受限的输出扰动或依赖有界损失函数假设的目标扰动不同，基于梯度扰动的DP-SGD提供了更为灵活的隐私保护方案。此外，相较于标准的带噪随机梯度下降(noisy stochastic gradient descent，NoisySGD)，DP-SGD对噪声的迭代注入需要更精细的量化与校准。其隐私性得益于坚实的理论基础，包括严格的隐私组合定理和通过子采样实现的隐私放大技术，这些共同确保了算法在优化和泛化方面的可用性。

随着DP-SGD机制的不断扩充与演化，其核心研究进展可归纳为三个关键方向：数据采样、梯度裁剪与噪声扰动。尽管近年来已有大量DP-SGD工作围绕这三个方向展开，但相关研究分布零散，理论机制、算法变种与应用实例尚未形成系统性框架，导致研究者在开展隐私学习时缺乏统一的知识谱系。特别是在高维参数、复杂数据形态、新兴场景等各种隐私风险不断涌现的背景下，急需对DP-SGD的原理机制、技术演进、关键挑战及未来趋势进行全面梳理，以为后续研究与实际部署提供理论支撑与方法参考。为全面综述该领域的发展脉络，本文系统地梳理了DP-SGD从基础知识、优化技术到算法理论的研究成果，并重点探讨其在大语言模型迅猛发展背景下的新兴应用与技术趋势。

具体地，本文根据DP-SGD主体逻辑和主流算法，系统性地对当前最前沿的DP-SGD技术进行了结构化分类。该分类框架紧扣DP-SGD的核心流程，沿“随机采样-梯度裁剪-噪声扰动”三阶段路径展开，梳理各阶段的研究进展与技术演化。首先，在随机采样过程中，数据采样和迭代权重选择对隐私放大的有效性有重要影响。为此，本文深入探讨基于重要性采样与个性化采样的策略改进，并进一步将迭代过程中的权重选择方式细分为选择性迭代与中间迭代隐藏两类方法。其中，本文围绕针对数据采样和迭代选择两方面理论进行了重点分析。其次，梯度裁剪操作是DP-SGD的核心步骤之一，通过对逐样本梯度施加L_2范数裁剪来确保敏感性上界。本文围绕裁剪偏差的理论解释与实际优化技术，归纳了三类主流改进策略：解析式裁剪机制、自适应裁剪函数以及几何感知裁剪方法，以呈现裁剪机制在优化可用性方面的进展。其中，本文重点分析了解析式裁剪中的分位数裁剪分析、自适应裁剪中的梯度对称性分析以及几何感知裁剪中的重尾稳定性研究。最后，在噪声扰动阶段，本文调研了现有针对差分隐私随机梯度下降中的噪声引导与模型降维技术，将其划分为基于随机性的无向引导与基于知识驱动的有向引导。针对噪声降维的理论上，本文研究了维度无关的差分隐私优化理论与外部混合数据辅助的优化理论。这些围绕DP-SGD各环节提出的优化技术与理论指导，显著提升了算法的实

际可用性与模型的整体效用，推动了差分隐私优化的发展。

当下，大语言模型在实现类人智能交互方面展现出巨大潜力，其在文本生成、对话系统和通识搜索等领域的应用深刻改变了人们的生活方式。由于大语言模型特殊的交互性质，使其能够接触到更多敏感的用户上下文信息，导致潜在的隐私风险加剧。尽管DP-SGD作为当前可量化、理论完备的主流差分隐私优化方法，被广泛应用于大语言模型以提升其合规性和安全性。但是，在大语言模型的实际部署过程中，DP-SGD面临模型参数高维、训练数据异构、推理路径多样等新挑战，对差分隐私机制的灵活性、鲁棒性和可配置性提出更高要求，因此，亟须构建覆盖“训练-微调-推理”全流程的隐私保护体系。为此，本文不仅回顾了DP-SGD在大语言模型训练中的理论机制与技术实现，还进一步梳理了在推理阶段（如上下文学习、提示生成）中实现差分隐私优化的机制，以支持大语言模型在高风险环境下的可靠运行。

与现有工作相比：目前，尽管已有大量关于隐私保护算法的综述性工作，但专门针对DP-SGD的综述依然相对稀缺。文献[4]按照图像、视频、音频和文本等数据类型探讨了差分隐私在不同数据结构分类上的发展与应用。文献[5-7]重点关注了联邦场景中的差分隐私优化算法，及分布式下的隐私优化特点，但并不侧重对DP-SGD本身理论和方法的剖析。近期研究[8]探讨了差分隐私在统计分析和工业机器学习中的应用挑战，主要关注隐私-效用权衡、隐私攻击与审计以及隐私保护的可解释性等问题。虽然部分工作涉及DP-SGD及相关方法，但本文的分类框架与其显著不同，提供了更细粒度的DP-SGD研究视角和更加详实的算法综述。此外，文献[9-13]专注于大语言模型的隐私保护参数高效算法、大语言模型中的隐私风险以及大语言模型中的提示工程等，但仅将DP-SGD作为次要的子模块进行阐述，并未探讨大语言模型中隐私审计的重要性。对比上述综述研究，本文在理论、方法与应用方面的分类架构存在明显差异，并专注于解构DP-SGD基础流程的相关技术。

1.1 研究路线

本文旨在对隐私保护机器学习领域中的差分隐私随机梯度优化进行归纳分析。本节将介绍文中DP-SGD涵盖的具体优化算法范围，并阐述归纳过程中的分类体系，如图1所示。

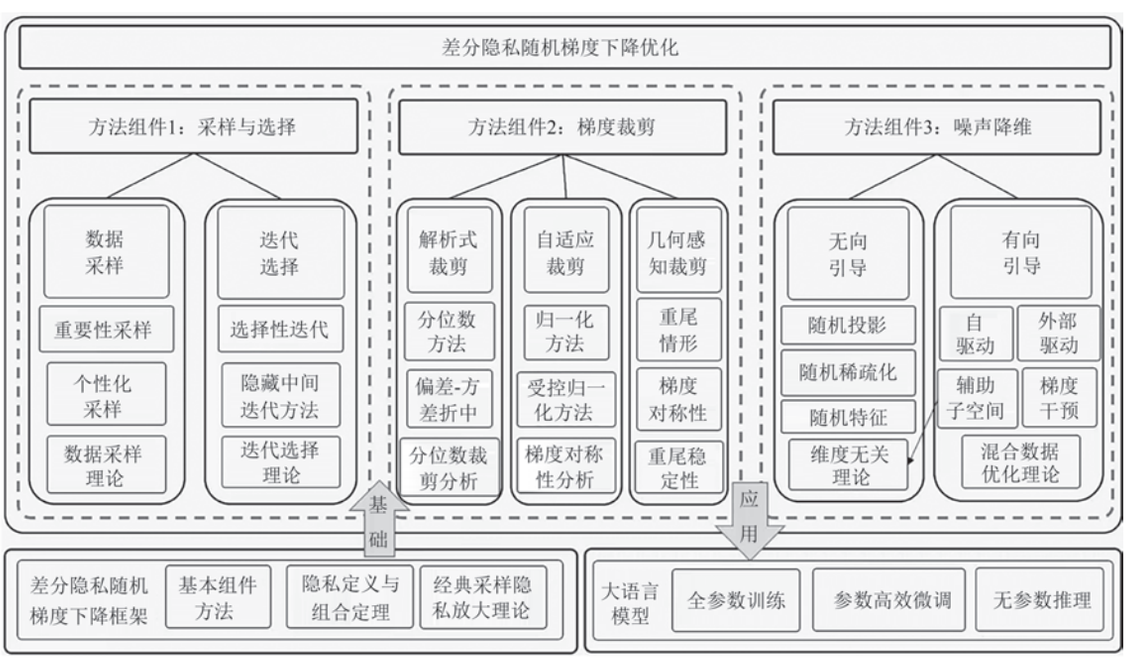

图1 差分隐私随机梯度下降优化层次图

本文归纳的差分隐私随机梯度下降优化算法需满足三个必要的关键步骤，即数据采样、梯度有界性约束与基于高斯机制的梯度扰动，并依据理论、方法和应用三个层面对DP-SGD现有研究进行综述。具体地，在第2章介绍了DP-SGD的基础知识与主要组件。第3章从数据采样和迭代选择方面，分别研究了重要性采样、个性化采样、选择性迭代和隐藏中间迭代，并分析了数据采样和选

择迭代中的理论。第4章研究了梯度裁剪中的解析式裁剪、自适应裁剪和几何感知裁剪。第5章从无向引导和有向引导两方面解析了DP-SGD中的噪声降维技术。第6章从应用层面介绍了DP-SGD在大语言模型场景中的应用与发展。第7章展望了DP-SGD未来可能的研究方向与挑战。第8章总结全文。

2　背景知识

在差分隐私随机优化的背景下，理解不同隐私定义之间的联系与转换是构建DP-SGD理论基础的关键。差分隐私核心在于通过加入随机噪声，使得相邻数据集在算法输出上的差异性受到控制，以实现输出结果不可区分性。

2.1　差分隐私

通常，DP-SGD使用高斯机制来实现差分隐私，通过添加经过校准的高斯噪声[2]来保护隐私。考虑到高斯分布的无界性，DP-SGD中通常采用的差分隐私定义被表述为近似差分隐私（Approximate Differential Privacy）[2]，相较于严格的纯差分隐私（pure differential privacy，PureDP），近似差分隐私在保留较强隐私保护的同时，提供了更大的可用性与灵活性，允许在算法设计中以较小的代价实现更优的模型性能。本文中“差分隐私”一词均默认指代近似差分隐私概念。

定义1. 近似差分隐私[2,14]对于任意两个相邻数据集$\mathcal{D}$和$\mathcal{D}'$，它们仅在一个数据点上不同，对于任意事件y，如果一个随机算法M满足(ϵ,δ)-差分隐私，则有

$$\Pr\big(M(\mathcal{D})\in y\big)\leqslant\exp(\epsilon)\cdot\Pr\big(M(\mathcal{D}')\in y\big)+\delta \tag{1}$$

其中，ϵ是隐私预算，表示理论上的隐私保护强度，δ是一个很小的概率。

后续，文献[15]和文献[16]基于分布散度的角度，分别提出了Rényi差分隐私和零集中式差分隐私。进一步地，文献[17]从假设检验的视角提出了f-差分隐私（f-DP）框架，其中高斯差分隐私是基于正态分布的一个重要实例。

定义2. Rényi差分隐私（Rényi differential privacy，RDP）[15]如果对于相邻数据集$\mathcal{D}$和$\mathcal{D}'$，设$\mathbb{D}_\alpha(\cdot|\cdot)$表示分布$M(\mathcal{D})$和$M(\mathcal{D}')$之间的Rényi散度，随机算法$M$满足如下条件，则称其为$\alpha,\epsilon_{\mathcal{R}}(\alpha)$-RDP：

$$\mathbb{D}_\alpha\big(M(\mathcal{D})|M(\mathcal{D}')\big)\coloneqq\frac{1}{\alpha-1}\log\mathbb{E}_{y\sim M(\mathcal{D}')}\cdot\left[\left(\frac{\Pr\big(M(\mathcal{D})\in y\big)}{\Pr\big(M(\mathcal{D}')\in y\big)}\right)^{\alpha}\leqslant\epsilon_{\mathcal{R}}(\alpha)\right] \tag{2}$$

与直接固定RDP的阶数α不同，本文采用函数形式，将$\epsilon_{\mathcal{R}}(\cdot)$作为$\alpha$的函数来表示RDP损失。这种视角能够统一多种常见的差分隐私定义。即，当$\alpha\to\infty$时，可以恢复为纯差分隐私；当$\alpha\to1$时，可以转化为Kullback-Leibler（KL）差分隐私。本文所用的log默认均指以e为底的自然对数。

定义3. 零集中差分隐私（zero-concentrated differential privacy，zCDP）[16]若随机算法M满足以下条件，则称其满足ρ-zCDP：对于所有相邻数据集$\mathcal{D}$和$\mathcal{D}'$以及任意$\alpha\in(1,\infty)$，都有

$$\mathbb{D}_\alpha\big(M(\mathcal{D})|M(\mathcal{D}')\big)\leqslant\rho\alpha \tag{3}$$

zCDP的隐私强度弱于纯差分隐私$(\epsilon,0)$-DP，但强于近似差分隐私(ϵ,δ)-DP（当$\delta>0$）。

在适当的条件下，RDP和zCDP两种隐私定义能够转化为标准的(ϵ,δ)-差分隐私形式，并在某种程度上与其等价。进一步地，文献[17]从假设检验的视角提出了f-差分隐私（f-DP）框架，其中高斯差分隐私（Gaussian differential privacy，GDP）是基于正态分布的一个重要实例。

定义4. 高斯差分隐私（Gaussian differential privacy，GDP）[17]考虑f-差分隐私定义中，以两个正态分布之间的权衡函数f为基础的参数化簇。在这种特化形式下，即称为高斯差分隐私。GDP拥有许多理想的数学性质，因此在隐私分析中占据了核心地位。本文通过单一参数$\mu\geqslant0$来定义其权衡函数：

$$G_\mu\coloneqq T\big(N(0,1),\ N(\mu,1)\big) \tag{4}$$

其中符号$T(\cdot|\cdot)$表示对所有（可测）拒绝规则的下确界。该权衡函数的显式表达式为

$$G_\mu(\alpha)=\Phi\big(\Phi^{-1}(1-\alpha)-\mu\big) \tag{5}$$

其中，Φ表示标准正态分布的累积分布函数。该权衡函数关于参数μ单调递增，即$\mu\geqslant\mu'\Rightarrow G_\mu\leqslant G_{\mu'}$。

如果一个机制M满足G_μ-差分隐私，则称其满足μ-高斯差分隐私（μ-GDP）。即对于任意一对相邻数据集$M(\mathcal{D})$和$M(\mathcal{D}')$，有

$$T\big(M(\mathcal{D}),M(\mathcal{D}')\big)\geqslant G_\mu \tag{6}$$

这些隐私定义通过捕捉更细粒度的统计信息，能够在多轮隐私机制组合分析中提供了更紧的隐私

损失上界，从而在隐私预算消耗和模型效用之间实现了更优权衡。

2.2 DP-SGD主要实现步骤

数据采样、梯度有界性和梯度扰动作为驱动DP-SGD的三个基本组件，保障着DP-SGD高可靠的稳定运行，并决定了DP-SGD算法的可用性保障上界（算法收敛与泛化速率）和隐私性保证下界（差分隐私强度）。

2.2.1 数据采样

定义5. 数据采样（Data Sampling）对于随机梯度下降（SGD）而言，其“随机性”主要是通过数据采样过程引入的。数据采样决定了每次迭代中用于计算梯度的数据子集。这种随机性不仅有助于降低计算开销，还直接影响优化过程的动态变化、收敛特性以及隐私保证。设隐私训练数据集为S，定义采样函数$Sampling(\cdot)$，在整个迭代中采样过程为

$$S_t = Sampling(S),\ t\in[1,\ T] \tag{7}$$

采样策略的选择直接影响梯度估计的统计性质，对于优化隐私保护和模型性能至关重要。常用的数据采样策略主要基于泊松采样（Poisson Sampling）。其中，在每次迭代中，泊松采样将每个数据点以固定概率τ独立地被纳入小批量中。设$M:\mathcal{X}^n\to\mathbb{R}^d$为一个$(\epsilon,\delta)$-差分隐私机制，则引入泊松采样后的新机制$M'$定义为

$$M' = M^{\circ}Sampling^{Poi} \tag{8}$$

泊松采样机制$Sampling^{Poi}$从集合S中生成采样子集S_t，其分布为

$$S_t = Sampling^{Poi}(S),\ S_t\sim\mathcal{W}(S_t) \tag{9}$$

其中，分布权重为

$$\mathcal{W}(S_t) = \tau^{|S_t|}(1-\tau)^{|S|-|S_t|} \tag{10}$$

引理1. 泊松采样的隐私放大效应[14, 18]该机制M'满足(ϵ',δ')-差分隐私，其中

$$\epsilon' = \log\left(1+\tau\left(e^{\epsilon}-1\right)\right) \tag{11}$$

此外，不放回均匀采样和带放回均匀采样的隐私放大效应[18]也在DP-SGD中广泛运用。

2.2.2 样本有界性

定义6. 逐样本梯度有界性（Per-Sample Gradient Boundedness）在向梯度添加差分隐私噪声之前，需要满足基本的有界性条件以确定敏感性，从而保证噪声的正确校准与应用。对于任意$x\in\mathcal{X}^n$，逐样本梯度$\nabla\bar{\ell}(\boldsymbol{w},x)$的范数被阈值$c$所限制，本文将其有界化处理定义为

$$\nabla\bar{\ell}(\boldsymbol{w},x) = Bound\left(\nabla\ell(\boldsymbol{w},x)\right) \tag{12}$$

现有主流的有界化策略主要包括：

（1）梯度裁剪

梯度裁剪（Gradient Clipping）是深度学习中最常用的实际方法之一，用于确保梯度具有上界。在DP-SGD中，需要对逐样本梯度进行裁剪，为后续的噪声注入做好准备。与裁剪SGD的相关研究类似[19-21]，梯度裁剪同样会引入裁剪偏差，显著影响DP-SGD的模型可用性。梯度裁剪通常利用设计良好的裁剪函数来定义约束步骤：

$$\nabla\bar{\ell}(\boldsymbol{w},x) = Clip\left(\nabla\ell(\boldsymbol{w},x),c\right) \tag{13}$$

最经典的裁剪函数是文献[14]提出的裁剪形式：

$$Clip^{Abadi}\left(\nabla\ell(\boldsymbol{w},x)\right) = \frac{c\nabla\ell(\boldsymbol{w},x)}{\max\left(c,\left\|\nabla\ell(\boldsymbol{w},x)\right\|_2\right)} \tag{14}$$

该裁剪函数是连续的，但并非处处可微，特别是在$\left\|\nabla\ell(\boldsymbol{w},x)\right\|_2 = c$处不可微。其不可微性来源于函数的分段定义：对于小梯度保持恒等映射，对于大梯度采用范数缩放。

从单调性角度看，该函数具有径向单调性：随着输入梯度范数的增大，裁剪后的输出范数不超过阈值c。然而，从逐分量或函数整体来看，该函数不具备严格的单调性。

为获得更平滑的裁剪效果和可微性，近年来，文献[22-24]提出了多种平滑裁剪函数，例如：

$$Clip^{Auto-S}\left(\nabla\ell(\boldsymbol{w},x)\right) = \frac{\nabla\ell(\boldsymbol{w},x)}{\left\|\nabla\ell(\boldsymbol{w},x)\right\|_2+\gamma},\ \gamma>0 \tag{15}$$

（2）凸集投影

在凸优化学习中，通常假设优化变量位于一个凸集合（Convex Set）e约束内。这一假设有助于保证学习问题的良好定义和算法的收敛性。

设$e=\mathbb{R}^d$为欧氏空间下的凸集合，其直径为$c=\|e\|_2$，则投影算子定义为

$$\prod_e(w_t) = \underset{\boldsymbol{w}\in e}{\operatorname{argmin}}\left\|\boldsymbol{w}-\boldsymbol{w}_t\right\|_2 \tag{16}$$

$$e = \left\{\boldsymbol{w}\in\mathbb{R}^d \middle| \left\|\boldsymbol{w}\right\|_2\leqslant c\right\} \tag{17}$$

若$\ell:e\to\mathbb{R}$是G-Lipschitz连续的，则对于所有$\boldsymbol{w}\in e$和$x\in\mathcal{X}^n$，有

$$\left\|\prod_e\nabla\ell(\boldsymbol{w},x)\right\|_2\leqslant G,\ G=c \tag{18}$$

其中，c是凸集直径，类比于梯度裁剪中的裁剪阈值，可看作梯度边界上界。

2.2.3　梯度扰动

在DP-SGD中，差分隐私保护通常通过对梯度施加扰动噪声实现。其中，高斯噪声因其在高维梯度空间中的良好适应性，以及在子采样机制下所展现出良好的隐私放大效果，已成为DP-SGD中主流的噪声注入方式。

定义7. 带有子采样的高斯机制（Gaussian Mechanism with Subsampling）[14]设采样概率为q，总迭代次数为T，子采样机制为M'：$\mathcal{X}^n \to \mathbb{R}^d$。则高斯机制定义为

$$M'(S_t)+\xi_t \tag{19}$$

这里有$\xi_t \sim N\left(0,\ \triangle_2^2(M')\sigma^2\mathbb{I}_d\right)$　(20)

$$\text{及}\ \sigma^2=\frac{m_1 q^2 \log\left(\frac{1}{\delta}\right)T^2}{\epsilon^2} \tag{21}$$

其中，m_1由具体的组合定理决定，$\triangle_2(M')$是机制M'在相邻数据集对(S,S')上的L_2敏感度，即

$$\triangle_2(M')=\sup /\!/ M'(S_t)-M'(S_t')/\!/_2 \leqslant c \tag{22}$$

其中，c为裁剪阈值。

根据高斯机制理论，在$0<\epsilon,\delta<1$的情况下，上述机制能够满足(ϵ,δ)-差分隐私。

3　针对采样与选择的改进策略与理论

在DP-SGD的每一次迭代中，算法的目标是通过对小批量数据的采样来估计整个训练集的平均梯度，并在该估计值上添加满足差分隐私要求的高斯噪声。因此，平均梯度估计的方差来源于两个方面：一是随机采样噪声（即梯度采样噪声），二是满足差分隐私而引入的高斯噪声。因此，数据采样步骤作为DP-SGD中的关键组成部分不仅关联着隐私放大的效果，决定着差分隐私噪声的注入强度，并且更重要的是直接影响着随机梯度偏离平均梯度的程度，保证着模型收敛的速率[25-27]。

具体来说，最新的差分隐私理论研究表明，数据采样本质上关系到随机梯度噪声的方差上界大小，即小批量随机梯度与真实梯度之间的偏差，这进一步影响裁剪偏差与模型收敛性。现有的常规采样方法，如带放回/不放回的均匀采样[18, 28]和泊松采样[14]，在满足DP-SGD复杂精度需求方面存在局限性，难以充分保障算法效用。在常见的计算机视觉任务中，带有常规采样的DP-SGD方法通常会使得模型精度下降5%～10%，而在数据较为稀疏的任务中，由于难以取得较大的采样率，DP-SGD性能下降会更为明显。此外，与数据采样密切相关的迭代选择（Iteration Selection）也成为研究的热点之一。众所周知，隐私预算的消耗仅在DP模型发布相应迭代参数时发生。如果能够识别出某些迭代中不必要或冗余的参数更新并予以舍弃，可以节省隐私预算并提升模型整体效用。此外，本节总结和分析了现有针对数据采样和迭代选择的算法理论研究。表1汇总了本节提及的方法和理论。

3.1　数据采样

本节将围绕经典样本采样、重要性采样（importance sampling, IS）和个性化采样（Personalized Sampling）三个方向对DP-SGD数据采样技术进行深入探讨。

3.1.1　经典样本采样及优化

文献[14]在DP-SGD中采用了经典的随机泊松采样，其采样率固定为样本批量与样本总量之比，在Moment Accountant下实现了更紧的隐私界限。正如文献[50]分析得出，DP-SGD中的均匀不放回子采样能够在隐私保证上等价于泊松子采样，这是由于两者在隐私放大效果上具有相同的阶数。更进一步，文献[29]证明了坐标级（Coordinate-Level）泊松采样可与样本级（Sample-Level）泊松采样结合，在二次采样下为Rényi差分隐私带来更优的隐私放大效果。在相同采样率下，文献[29]中的方法能够提升DP-SGD精度0.5%～4.4%，对于不同的隐私预算，该方法性能会随着采样率的提高而增加。然而，上述采样方案在采样噪声方差方面并非理论最优。

此外，实际模型训练中，不同样本对模型收敛的贡献往往存在差异[30]，在重尾数据场景中这种差异更加显著。通常，特征集中的样本更易于学习，而尾部样本则较难拟合。因此，需要更加精细化的数据采样策略，以充分考虑不同样本的贡献差异。

3.1.2　重要性采样

针对上述问题，文献[35]将重要性采样引入DP-SGD优化中，提出了满足差分隐私的DPIS框架，其重点解决了IS与复杂DP数学框架整合的难题。具体而言，研究者将每个样本的梯度范数视为其重要性指标，并据此进行重要性采样。显然，该方法偏向于选择梯度范数较大的记录。为获得相对于裁剪阈值的无偏平均梯度估计，该方法为每个梯度引入缩放因子，使所有缩放后的梯度具有相同的L_2范数。然后，通过对通过IS采样得到的小批量数据

表1 差分隐私随机梯度下降中的采样与选择策略

类别	优化机制	文献	ϵ	δ	任务模型	任务类型
数据采样	经典数据采样及优化	[29]	[2, 8]	$1e^{-5}$	ResNet-22	计算机视觉
		[30]	(1, 8)	$1e^{-5}$	CNN	计算机视觉
		[31]	[1, 12.5]	$1e^{-5}$	CNN	计算机视觉
		[32]	[4, 8]	$1e^{-5}$	ResNet20, CNN, RNN	视觉、语言
		[33]	(0, 20)	-	-	-
		[34]	-	-	ResNet-18	计算机视觉
	重要性采样	[35]	[0.5, 4]	$1e^{-5}$	CNN	计算机视觉
	个性化采样	[36]	[0.3, 3]	$1e^{-5}$	BERT, LSTM, ResNet-18	视觉、语言
		[37]	[1, 12]	$[1e^{-5}, 1e^{-4}]$	Logistic Regression, CNN, ResNet-18 和 BERT	视觉、语言
迭代选择	选择性迭代	[38]	-	-	-	-
		[39]	[1.5, 5]	$1e^{-5}$	MLP, MobileNetV2, AlexNet, Resnet18	计算机视觉
	隐藏中间迭代	[40]	8,	$[1e^{-5}, 1e^{-4}]$	ResNeXt-29. ResNet-9, SimCLRv2 和 GPT-2	视觉、语言
		[32]	[4, 8]	$1e^{-5}$	Scattering Network	视觉、语言

理论方向	优化目标	损失函数	测量指标	文献	上界	下界
基于隐私选择机制的优化理论	指数机制 + DP-SPIDER	非凸函数	经验 FOSP	[37, 39, 40, 41]	$\Theta\left(\min\left(\left(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right)^{2/3}, \frac{d^{2/3}}{n\epsilon}\right)\right)$	√
			经验 SOSP	[42-45]	$\tilde{\Theta}\left(\frac{d^{2/3}}{n\epsilon}\right)$	×
			总体 SOSP	[45]	$\tilde{\Theta}\left(\frac{1}{n^{1/3}}+\left(\frac{\sqrt{d}}{n\epsilon}\right)^{3/7}\right)$	×
			经验超额风险	[42, 45, 46]	$\Theta\left(\frac{d\log(1/\delta)}{(n\epsilon)^2}\right)$	×
			总体超额风险	[42, 45, 46]	$\Theta\left(\frac{d}{n\epsilon}+\sqrt{\frac{d}{n}}\right)$	√
基于自适应迭代的理论优化	随机选择迭代	非凸函数	经验 FOSP	[47]	$\Theta\left(\frac{(d\log(n)\log(1/\delta))^{1/4}}{\sqrt{n\epsilon}}\right)$	√
	基于 Epoch 选择	Tsybakov 噪声条件	总体超额风险	[48, 49]	$\Theta\left(\left(\frac{1}{\sqrt{n}}+\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right)^{\frac{\theta}{\theta-1}}\right)$	√
		GLD	经验与总体超额风险	[42]	$\Theta\left(\frac{d}{\log(n)\epsilon^2}\right)$	√
		Weak Quasi-convexity	经验超额风险	[46]	$\Theta\left(\frac{d\log(1/\delta)}{(n\epsilon)^2}\right)$	√

的缩放梯度求平均来计算估计的平均梯度。

此外，在每次迭代中，DPIS通过重要性采样发布两个附加值：训练集中记录的总数和梯度范数的总和。为了保证差分隐私，这两个值会通过高斯噪声进行扰动，并采用子采样估计和阈值选择策略减少隐私成本，实现隐私放大的效果。

3.1.3 个性化采样

在个性化隐私设置下，不同数据所有者对不同样本可能具有不一致的隐私偏好和预算，采用固定采样率会导致次优的隐私-效用表现。为解决这一问题，文献[36]提出了基于个性化隐私预算的DP-SGD采样方法，即个性化采样(Personalized Sampling)技术。在该方案中，每个样本被分配不同的隐私预算，具有更高隐私预算的样本被更频繁地采样。具体而言，研究者们通过二分搜索确定给定隐私参数对应的采样率，并保持平均采样率等于原始DP-SGD的固定采

样率，以确保期望的小批量大小。

类似地，在联邦学习场景下（本地客户端运行DP-SGD并采用泊松采样），文献[37]针对个体隐私预算异质性的问题，提出了样本采样概率应与各自隐私预算成正比的策略。研究者们基于指数函数构建了隐私-采样-效用曲线，通过模拟不同采样率下的隐私消耗，推导出每个非均匀客户端的经验最优采样率。具体地，研究者提出了rPDP-FL框架，采用两阶段混合采样：统一的客户端级采样与非均匀的记录级采样，以适应不同隐私要求。核心挑战在于如何根据个性化隐私预算确定最优记录级采样概率。研究者提出了Simulation Curve Fitting方法，揭示了采样率与隐私预算之间的非线性关系，并建立数学模型加以解决。实验结果表明，该方法在保持个性化隐私的同时，相比不考虑个性化隐私的基线方法显著提升了模型性能。

此外，采样策略还直接影响随机梯度噪声的方差。如何在差分隐私约束下压缩采样噪声，已成为研究的另一个关注点。为此，相关文献[31,32,34]系统分析了DP-SGD中采样噪声与经验风险之间的关系，并提出通过对一小组样本的平均梯度进行裁剪，而不是单个样本梯度裁剪，以减少噪声影响。同时，在隐私保护大语言模型和基础模型的研究中，研究者们更倾向于采用较大的批量大小（即更高的采样率），以提升模型的准确性。然而，该方法的有效性尚缺乏充分的理论解释，本文将在后续关于大模型中DP-SGD的应用章节中进一步探讨。

3.2　迭代选择

正如前文所述，并非所有样本都对模型收敛产生积极影响。识别出更具价值的样本梯度不仅可以加速模型收敛，还能减少不必要的迭代。因此，一类主流研究致力于设计选择性迭代算法变体，通过减少迭代次数实现更快的训练收敛，而另一类研究则专注于通过隐藏模型中间迭代优化隐私分析，从而提高算法可用性。

3.2.1　选择性迭代

基于此，为了丢弃无用甚至有害的模型更新，文献[51]提出了一种选择性迭代（Elective Iteration）算法DPSUR。该算法要求在每次迭代中对每个样本的梯度进行验证和评估，只有通过效用验证的样本，其梯度更新才会被用于模型参数的实际更新并对外发布。

具体而言，由于随机采样和高斯噪声的影响，DPSUR并不直接接受每次迭代生成的模型更新。相反，算法首先计算当前迭代模型的损失，并将其与上一轮迭代的损失进行比较，以决定是否接受本轮模型更新。然而，在DP-SGD中，正确识别有效的样本更新存在诸多挑战。一方面，评估梯度更新通常需要使用额外的验证集，这会带来额外的隐私预算消耗；另一方面，设置合适的阈值来选择或发布更新的梯度至关重要，该阈值直接影响后续训练性能。阈值过高可能导致常规样本欠拟合，而阈值过低则可能导致无效样本的误用。

针对上述问题，DPSUR提出了两方面的改进策略：(1)为解决隐私预算消耗过大的问题，研究者们设计了一种新的裁剪策略，通过采用比标准DP-SGD更小的裁剪边界，以减少相应的噪声误差。(2)针对阈值选择问题，提出了优化后的选择性更新阈值机制，以提高阈值的可靠性。结合这些改进，通过选择性发布的高斯机制，DPSUR有效减少了多次迭代过程中的隐私预算消耗。

类似地，文献[52]将迭代机器教学（Iterative Machine Teaching）[53-54]和噪声投票机制（Noisy Voting）[55]相结合，并基于报告最大噪声值方法（reporting noisy argmax，RNA）[2]，在每轮迭代中以隐私保护的方式选择训练样本。具体而言，研究者们首先基于不相交的辅助数据集训练多个独立的非隐私模型，通过模型集成使用噪声投票聚合的方式，从候选样本集中私密地选择最优样本批次，然后基于此选择进行隐私模型训练。采用这种方案，模型的迭代更新也更偏向于高价值样本，从而提升模型效用。此外，文献[44]提及模型初始迭代的重要性，倘若初始化模型参数与后续隐私训练轨迹提前对齐，则能够提升后续的隐私优化效果。为此，研究者利用指数机制，将初始迭代点的选择隐私化，使得其与后续训练轨迹距离接近，从而提高模型效用。

3.2.2　隐藏中间迭代

已有研究表明，当仅发布最后一轮模型结果而隐藏DP-SGD的中间迭代（Hidden Intermediate Iteration）时，可以在不牺牲隐私的前提下获得更紧的隐私损失界[56,57]，特别是在凸损失函数下，能够实现更优的隐私-效用权衡。然而，这种分析方法仅适用于单个训练周期，对于多轮训练时的性能表现尚不清晰。文献[58]在强凸损失函数假设下，进一步推导了适用于多轮迭代的隐藏状态NoisySGD的强隐私保证。但研究者们的分析依赖于极高的计算资源，且难以扩展至小批量随机梯度下降场景，因为未能有效考虑多次子采样迭代后的混合分布所带来的隐私放大效应。

进一步地，文献[57]基于两种采样策略—“打乱划分(Shuffle and Partition)”[59]和“不放回采样”[26]，推导出了更紧的隐私界限。这些界限利用了隐私放大理论中的后处理性质和子采样放大[60-63]。此外，研究者将噪声划分为较弱的随机噪声和纯粹的加性噪声，在全梯度下降场景下适应隐藏中间迭代的分析需求。

当前，关于隐私审计(Privacy Auditing)的许多研究假设对手只能观察最后一轮模型结果，通过计算隐私预算的经验下界对DP-SGD进行隐私审计。然而，这些理论隐私预算通常仍基于多轮迭代的组合定理计算，导致实际隐私消耗和理论分析存在一定偏差。有趣的是，部分文献[64]指出，隐私损失随迭代次数无限增长的传统观点可能存在误导性。这些研究认为，当迭代次数超过某一阈值后，进一步的迭代不会带来额外的隐私泄露。

3.3 针对数据采样与迭代选择的算法理论研究

数据采样机制是差分隐私优化算法中的核心组件，决定了每轮训练中的样本选择方式，直接影响泛化误差、隐私预算消耗与计算效率。近年来，研究者从经典采样技术出发，逐步引入重要性采样和个性化采样等策略，在凸与非凸目标下推导出更优的总体风险界限与样本复杂度。尽管已有方法在理论上逼近非隐私优化的速率，但在非平滑函数、高维数据等复杂场景下，分析采样机制带来的复杂度仍面临挑战。

另外，与数据采样息息相关的迭代选择策略，其设计不仅影响模型效用与收敛速率，更直接决定隐私预算的分配效率。近年来，研究者围绕迭代过程中的关键决策展开深入研究，主要聚焦于两类路径：一类基于隐私机制进行理论优化，具体通过指数机制与梯度动量等手段提升算法在一阶与二阶驻点下的稳定性与准确性；另一类则引入自适应迭代策略，根据训练轮次、噪声结构或数据复杂性动态调整批大小、学习率或优化范围，从而实现更精细化的隐私-效用权衡。上述方法在凸与非凸场景中均推动了超额风险界与梯度复杂度的持续收紧，进一步缩小了差分隐私优化与非隐私最优之间的性能差距。本节将从面向数据采样的梯度复杂度分析和改进迭代策略下的隐私优化理论阐述目前的理论研究。

3.3.1 面向数据采样的复杂度分析

数据采样机制决定了每轮迭代中所调用的样本数量，从而影响总的样本梯度计算开销。降低采样频次不仅能够提高收敛效率，还能有效节约差分隐私机制中的隐私预算，因而成为衡量差分隐私优化方法计算资源消耗的关键因素。为此，文献[65]提出了新的计算复杂度度量方法，指的是针对差分隐私优化算法，为达到特定优化目标(如收敛到期望的误差范围)所需的计算资源，通常以总样本梯度计算次数衡量，也被称为样本或梯度复杂度(Sample or Gradient Complexity)。

具体地，研究者结合Phased-SGD优化了样本计算的复杂度，根据文献[66]提出的$\Theta\left(\min\left(n^{3/2}, n^{5/2}d^{-1}\right)\right)$降低到了接近线性的$\Theta\left(\min\left(n, n^2d^{-1}\right)\right)$，该界限部分接近了非隐私SGD的复杂度结果。在非平滑损失函数下，文献[65]采用了具有λ-强凸修正项$\lambda\|\boldsymbol{w}\|_2^2$的均匀稳定算法，将样本梯度计算复杂度显著降低至，$\Theta\left(n^2\log\left(1/\delta\right)\right)$其中$\delta$是差分隐私保证中的失败概率。这一复杂度相比文献[66]在使用Moreau-Yosida包络时达到的$\Theta\left(n^{4.5}\right)$有了大幅改善。针对带有替代均匀采样，文献[67]通过增强均匀稳定性理论研究了超额总体风险下的算法高概率界限。在这一工作中，非平滑情况下的样本梯度计算复杂度也进一步降低至$\Theta\left(n^2\right)$。另外，文献[41]研究了全采样的DP-GD算法，其在非凸DP-ERM下达到梯度复杂度$\Theta\left(n^2\right)$。

为实现严格的次二次梯度复杂度，文献[68]采用了基于球面核的卷积平滑技术，以及基于截断集中差分隐私(truncated concentrated differential privacy，tCDP)的更紧的隐私组合定理变体[69]，该隐私定义是zCDP的宽松版本，通过高斯集中不等式的尾部截断性质获取隐私组合的增益。随后，研究者在加速随机逼近求解器(AC-SA)[70]基础上，达到了$\Theta\left(\min\left(n^{5/4}d^{1/8}, n^{3/2}d^{-1/8}\right)\right)$的梯度复杂度，并在强凸设置下获得了当前最优的DP-SCO速率。

3.3.2 改进迭代策略下的隐私优化理论

(1)基于隐私选择机制的优化理论

基于一阶驻点指标下，文献[28]基于SPIDER算法[71]提出DP-SPIDER算法。DP-SPIDER通过重采样方式，将历史梯度增量和模型参数动量的信息补充给更新梯度，从而能够降低梯度采样噪声的方差，以优化算法效用。此外，研究者在Lipschitz平滑假设下，将误差界优化为$\Theta\left(\frac{\sqrt{d\log\left(1/\delta\right)}}{n}\right)^{2/3}$，梯度复杂度达到，$\tilde{\Theta}\left(\max\left(\frac{n^{5/3}\epsilon^{2/3}}{d^{1/3}}, \frac{\left(n\epsilon\right)^2}{d}\right)\right)$在$n\epsilon=$

$\Theta\left(\sqrt{d}\right)$时优于先前结果。文献[35]则利用指数机制进行热启动初始化，并基于DP-SPIDER进行优化，最终达到误差界$\Theta\left(\min\left(\left(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right)^{2/3}, \frac{d^{2/3}}{n\epsilon}\right)\right)$，在$n\epsilon=\Theta(d)$时为最优。

文献[45]研究二阶优化方法下的DP-SPIDER算法，二阶优化方法通常基于一阶梯度扩展至Hessian近似。具体地，研究者将AboveThresholdy隐私选择算法[2]引入SPIDER框架，提出DP-SPIDER的SOSP局部优化算法，实现了$\tilde{\Theta}\left(\left(\frac{\sqrt{d}}{n\epsilon}\right)^{2/3}\right)$的经验SOSP误差率和$\tilde{\Theta}\left(\frac{1}{n^{1/3}}+\left(\frac{\sqrt{d}}{n\epsilon}\right)^{3/7}\right)$的总体SOSP误差率。最新工作中，文献[35]借助指数机制为DP-SPIDER提供更优的模型初始化点，在低维场景$n\epsilon=\Theta(d)$下，将经验收敛率优化至$\tilde{\Theta}\left(\frac{d^{2/3}}{n}\right)$。此外，文献[42]提出基于二阶梯度估计的差分隐私梯度下降，通过向Hessian加入对称的高斯噪声实现DP，达到SOSP收敛率$\tilde{\Theta}\left(\frac{d^{1/4}}{\sqrt{n\epsilon}}\right)$。随后，文献[44]提出DP-TrustRegion（DP-TR）方法，利用Hessian信息加速逃离鞍点，达到SOSP经验梯度范数下的$\tilde{\Theta}\left(\frac{d^{4/7}}{(n\epsilon)^{4/7}}\right)$速率。

针对超额风险的函数值测度，文献[45]提出了一种正则化指数机制（Regularized Exponential Mechanism），在多项式时间内有效实现了无非光滑条件下的$\Theta\left(\frac{d\log(n)}{\log(n)\epsilon}\right)$的经验和总体风险界限，并且在指数时间复杂度下优化了总体风险到$\Theta\left(\frac{d}{n\epsilon}+\sqrt{\frac{d}{n}}\right)$。考虑到文献[28]证明的下界$\Theta\left(\frac{d}{n\epsilon}+\sqrt{\frac{d}{n}}\right)$，这个结果已达到该条件下的最紧的。

（2）基于自适应迭代的理论优化

在差分隐私随机优化中，传统算法多基于固定迭代次数与统一噪声机制实现隐私保护，难以充分适应数据分布结构、任务复杂性及训练阶段的动态需求。为缓解此局限，近期研究逐步引入基于迭代的自适应更新机制，通过在训练过程根据迭代轮数，动态调整样本批大小与学习率，以实现更精细化的隐私预算分配与优化路径控制。

其中，文献[47]提出了Random Round PrivateSGD算法，该算法将迭代次数和样本量联系，研究者引入一个随机迭代轮数，如果该随机数小于样本数的平方，那么只返回该轮数下的输出结果。其在平滑损失函数下，建立了(ϵ,δ)-差分隐私的经验FOSP误差界$\Theta\left(\frac{\left(d\log(n)\log(1/\delta)\right)^{1/4}}{\sqrt{n\epsilon}}\right)$。

文献[48]研究了更复杂的(θ,λ)-Tsybakov噪声条件（Tsybakov noise condition，TNC）下的算法优化分析，其中(θ,λ)-Tsybakov噪声条件意味着$L_S(\boldsymbol{w})-L_S(\boldsymbol{w}^*)\geqslant\lambda\left\|\boldsymbol{w}-\boldsymbol{w}^*\right\|_2^{\theta}$，其适用于分类任务下的贝叶斯优化分析。相比之下，强凸条件$L_S(\boldsymbol{w})-L_S(\boldsymbol{w}^*)\geqslant\frac{\lambda}{2}\left\|\boldsymbol{w}-\boldsymbol{w}^*\right\|_2^2$可视为$\left(2,\frac{\lambda}{2}\right)$-TNC的特例。研究者基于文献[65]提出的Phased-SGD，该算法在训练后期阶段使用更少的样本、更小的学习率，从而获得更好的隐私放大效果和样本复杂度。该算法在纯差分隐私下达到了$\Theta\left(\left(\frac{\sqrt{\log(n)}}{\sqrt{n}}+\frac{d\log(n)}{n\epsilon}\right)^{\frac{\theta}{\theta-1}}\right)$的上界，在$(\epsilon,\delta)$-差分隐私下达到了$\Theta\left(\left(\frac{\sqrt{\log(n)}}{\sqrt{n}}+\frac{\sqrt{d\log(1/\delta)}\log(n)}{n\epsilon}\right)^{\frac{\theta}{\theta-1}}\right)$的上界。为了消除上界中的样本数$n$的对数项，研究者们[48]提出了Iterated Phased-SGD，Iterated Phased-SGD是在Phased-SGD基础上引入分区与多轮训练迭代的改进版本。它通过将原始数据集划分为多个子集，并在每个子集上独立运行Phased-SGD，从而提升隐私放大效果、降低总体样本复杂度。由此，研究者将总体超额风险收紧为$\Theta\left(\left(\frac{1}{\sqrt{n}}+\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right)^{\frac{\theta}{\theta-1}}\right)$。在Lipschitz平滑TNC假设下，文献[49]借助最优实例的逆灵敏度机制，分别在纯

差分隐私和(ϵ,δ)-差分隐私下导出了同阶的界$\tilde{\Theta}\left(\left(\frac{d}{n\epsilon}\right)^{\frac{\theta}{\theta-1}}\right)$，其中$(\epsilon,\delta)$-差分隐私下的上界比文献[48]多出$\sqrt{d}$的依赖。由于逆灵敏度(Inverse Sensitivity)的计算复杂度过高，研究者们引入Localization技术[65]提高算法效率，并通过均匀稳定的高概率1-δ_θ推导出$\Theta\left(\frac{\sqrt{\log(1/\delta)}\log^{3/2}(n)}{\sqrt{n}}+\frac{d\log(1/\delta_\theta)\log(n)}{n\epsilon}\right)$和$\Theta\left(\frac{\sqrt{\log(1/\delta)}\log^{3/2}(n)}{\sqrt{n}}+\frac{\sqrt{d\log(1/\delta)}\log(1/\delta_\theta)\log(n)}{n\epsilon}\right)$的上界。进一步地，文献[49]提出了基于Epoch-based的算法[72]，通过根据迭代轮数递进收缩的优化区域和步长实现对最优点的逐步逼近，从而获得了与文献[48]相同的已知最优总体超额风险界。

为了在Lipschitz光滑性条件下实现非凸性下的超额风险界限，文献[42]提出了一种差分隐私的梯度朗之万动力学(gradient langevin dynamics，GLD)该方法是一种基于布朗运动和朗之万扩散的随机微分方程的梯度下降变体，其中模型学习率根据迭代轮数总量进行选择。研究者利用对数Sobolev不等式和GLD渐近理论的分析工具，首次在DP领域中严谨地研究了经验风险和总体风险的超额风险界限，均达到了$\Theta\left(\frac{d}{\log(n)\epsilon^2}\right)$。文献[46]利用阶段性学习率调度和早期动量策略对DP-SGD进行了改进，该策略能够动态调整学习率，并在后期阶段关闭动量以稳定收敛速度。在结果方面，其在稍强的弱拟凸性(Weak Quasi-Convexity)和Polyak-Łojasiewicz条件下，达到了$\Theta\left(\frac{d\log(1/\delta)}{(n\epsilon)^2}\right)$的经验超额风险界限。

4 针对梯度裁剪的优化机制与理论

DP-SGD通过引入裁剪机制限制单个训练样本对模型的影响。具体而言，通过将梯度裁剪到预设的最大范数，确保任何单个数据点都不会对模型更新产生显著影响。此外，在梯度聚合过程中加入随机噪声，使得模型在添加或移除单个数据点时输出的差异几乎不可区分，从而实现严格的差分隐私保证[1, 14]。表2总结了现有梯度裁剪相关的方法与理论结果。

表2 差分隐私随机梯度下降中的梯度裁剪方法

裁剪方案	优化机制	方案来源	ϵ	δ	任务模型	任务类型
解析式裁剪	偏差-方差机制	文献[73]	5	$n^{-1.1}$	Resnet, CNN, LSTM, Multi-LR	计算机视觉
		文献[74]	{2, 8.8}	[1e^{-6}, 1e^{-5}]	ViT-small. GPT-2	视觉、语言
	分位数机制	文献[75]	[0.1, 4]	1e^{-5}	Logistic Regression	计算机视觉
		文献[76]	6.55	1e^{-6}	CNN, Logistic Regression	计算机视觉
		文献[36]	[0.3, 3]	1e^{-5}	BERT, LSTM, ResNet-18	视觉、语言
		文献[19]	{2, 4, 6}	1e^{-5}	ResNet-18, DistilBERT	视觉、语言
自适应裁剪	归一化裁剪函数	文献[22]	[2, 8]	[1e^{-6},1e^{-4}]	CNN, SimCLRv2, GPT2 Resnet9, RoBERTa	视觉、语言
	自适应裁剪函数	文献 [23]	[2, 8]	1e^{-5}	Resnet20	计算机视觉
		文献[22]	[2, 8]	[1e^{-6}, 1e^{-4}]	CNN, SimCLRv2, GPT2 Resnet9, RoBERTa	视觉、语言
	受控自适应裁剪函数	文献 [14]	[0.1, 8]	[1e^{-5}, 1e^{-2}]	NN	计算机视觉
		文献 [24]	[2,8]	[1e^{-6},1e^{-5}]	CNN, SimCLRv2, Resnet9, RoBERTa	视觉、语言
几何感知裁剪	分布对称性	文献 [38]	-	-	-	-
		文献 [39]	[1.5, 5]	1e^{-5}	MLP, MobileNetV2, AlexNet, Resnet18	计算机视觉
	分布尾部感知	文献 [40]	8,	[1e^{-5},1e^{-4}]	ResNeXt-29, ResNet-9 SimCLRv2 和 GPT-2	视觉、语言
		文献 [32]	[4, 8]	1e^{-5}	Scattering Network	视觉、语言

续表

理论方向	优化目标	损失函数	测量指标	文献	上界	下界
基于梯度对称性的裁剪分析	Clipped DP-SGD	非凸函数	经验 FOSP	[22, 23, 32, 38, 77]	$\tilde{\Theta}\left(\frac{d^{1/4}}{\sqrt{n\epsilon}}\right)$	√
针对分位数损失函数的理论	Clipped DP-SGD	凸函数	总体超越风险	[78]	$\Theta\left(\max\left(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon},\frac{1}{\sqrt{n}}\right)\right)$	√
面向重尾数据的优化稳定性	重尾假设下的 Clipped DP-SGD	一般凸	总体超额风险	[20, 21, 79, 80]	$\tilde{\Theta}\left(\frac{1}{\sqrt{n}}+\max\left(\left(\frac{\sqrt{d}}{n\epsilon}\right)^{\frac{4(\theta-1)}{5\theta-1}},\left(\frac{\sqrt{d}}{n\epsilon}\right)^{\frac{\theta-1}{\theta}}\right)\right)$	×
			经验超额风险	[19]	$\tilde{\Theta}\left(\left(\frac{\sqrt{d}}{n\epsilon}\right)^{1-\frac{2}{\theta+1}}\right)$	√
		强凸	总体超额风险	[20, 79-81]	$\tilde{\Theta}\left(\frac{1}{n}+\left(\frac{\sqrt{d}}{n\epsilon}\right)^{\frac{2(\theta-1)}{\theta}}\right)$	√
			经验超额风险	[19]	$\tilde{\Theta}\left(\left(\frac{\sqrt{d}}{n\epsilon}\right)^{1-\frac{1}{\theta}}\right)$	√
		非凸重尾 Lipschitz	总体超额风险	[20]	$\tilde{\Theta}\left(\left(\frac{\sqrt{d}}{n\epsilon}\right)^{1-\frac{1}{2\theta-1}}\right)$	√
			经验 FOSP	[19]	$\Theta\left(\left(\frac{\sqrt{d}\log(n)}{n\epsilon}\right)^{\frac{2\theta-2}{\theta}}\right)$	√
		非凸重尾梯度噪声	经验 FOSP	[40]	$\Theta\left(\frac{d^{\frac{1}{4}}}{(n\epsilon)^{\frac{1}{2}}}\log^{\max\left(\frac{5}{4},\theta+\frac{1}{4}\right)}(T/\delta)\log^{2\theta}\left(\sqrt{T}\right)\right)$	√

具体地，在非隐私SGD中，梯度裁剪一定程度上有助于加速模型收敛[82, 83]。然而，在DP-SGD中，梯度裁剪与噪声注入机制密切相关，其敏感性显著增强，且往往导致模型性能下降。过大的裁剪阈值会引入大量噪声，而过小的裁剪阈值则带来显著的裁剪偏差。因此，越来越多的研究致力于通过优化梯度裁剪技术，提升DP-SGD的隐私-可用性平衡。基于此，本文从解析式裁剪机制、自适应裁剪函数和几何感知裁剪三个角度，系统地总结了现有的研究进展。此外，针对梯度裁剪偏差的理论研究，本文总结和分析了分位数损失函数、梯度对称性以及重尾数据的优化稳定性下的理论结果。

4.1　解析式裁剪阈值

在常规DP-SGD中，裁剪阈值通常由人工设定。然而，裁剪阈值选择不当可能导致模型收敛过程震荡，甚至带来次优的模型性能。由于最优裁剪阈值的确定涉及多个复杂因素，如隐私噪声强度、模型收敛动态以及梯度分布等统计特性，精确地确定最优裁剪阈值仍是一个未被完全解决的挑战性问题，为此，现有研究按照偏差-方差折中方法和分位数方法两个方面作出努力[19, 36, 73, 75-76]。

4.1.1　偏差-方差折中方法

为了更好地确定裁剪阈值，偏差-方差折中方法（Bias-Variance Reduction Methods）[73-74]通过跟踪隐私梯度中的偏差和方差动态，寻找裁剪阈值最佳的折中点取值。在文献[73]的工作中，通过结合特定的误差削弱分析机制，减轻了裁剪带来的偏差损失，来实现DP-SGD的模型效用提升。该研究将DP-SGD的收敛误差分解为随机梯度采样噪声、裁剪偏差和差分隐私噪声方差三部分，并以偏差-方差分析框架定量重构梯度，通过估计标准随机梯度采样噪声的方差，为回归问题推导出解析式的最优裁

剪阈值。

此外，文献[74]引入了一种基于误差反馈(Error-Feedback)[84]的机制，通过保留的校正项在每次迭代中补偿裁剪偏差。该方法最初用于缓解信号处理中量化误差，并成功地将补偿项嵌入模型的隐藏状态更新中，避免了额外的隐私预算消耗。这一机制显著降低了裁剪带来的固定误差，并允许更加灵活的裁剪阈值选择。然而，这类方法的理论基础和应用效果主要限于凸优化场景，其在复杂数据集或非凸模型中的适用性仍有待深入探索。

4.1.2 分位数方法

分位数方法(Quantile Methods)的研究[19, 36, 75-76]致力于通过优化分位数选择裁剪阈值。文献[73]首次提出了基于分位数的自适应裁剪机制。研究者们通过初始化裁剪阈值，迭代计算超过该阈值的样本梯度比例，并使用指数函数逐步调整阈值以逼近预设的分位数。然而，实证研究表明，最优分位数因数据集而异，这表明仍需进一步探索更加稳健且具有普适性的阈值选择策略。

不同于文献[75]使用指数函数确定分位数的方法，文献[36]在个性化隐私预算场景下，采用二分搜索法针对不同隐私偏好的群体分别搜索合适的分位数裁剪阈值。这种方法相比固定阈值策略，能够显著降低高隐私预算样本的裁剪损失。

此外，针对标签不平衡问题导致的DP-SGD效用下降，文献[76]提出了面向不同类别群体的自适应裁剪阈值选择算法。该算法根据各类别样本数量为其分别设计不同的分位数裁剪阈值，以消除类别不公平性带来的影响。同样地，文献[19]在处理重尾数据的凸优化问题中，通过理论分析表明，即使将裁剪阈值设置在低于0-分位数的位置，也能获得最优模型性能。尽管基于凸假设的理论和实证研究已经较为丰富，但在非凸及复杂数据场景下裁剪阈值选择的有效指导仍是一个重要且亟待突破的研究方向。

4.2 自适应裁剪函数

在DP-SGD中，对逐样本梯度进行裁剪是确保邻接数据集敏感性上界的最直接操作。经典的Abadi裁剪方法[14]需要同时调节裁剪阈值c和学习率η，这带来了$\Theta\left(n_c \cdot n_\eta\right)$的实验调优复杂度。该经典裁剪策略极大地增加了DP-SGD在机器学习中的实际应用难度。因此，如何高效改进裁剪过程已成为针对带有梯度裁剪DP-SGD研究的热点问题[22-24]。

本文将详细阐述当下主流的自适应裁剪函数，为了简化符号，令$\boldsymbol{g}=\nabla\ell(\boldsymbol{w},x)$。

4.2.1 经典Abadi裁剪函数

如公式23所示，Abadi裁剪函数[14]属于分段连续函数。当样本梯度范数$\|\boldsymbol{g}\|_2 \leqslant c$时，裁剪函数保持梯度不变，即$\bar{\boldsymbol{g}}=\boldsymbol{g}$。当$\|\boldsymbol{g}\|_2 > c$时，裁剪函数调整为$\bar{\boldsymbol{g}}=\dfrac{c\boldsymbol{g}}{\|\boldsymbol{g}\|_2}$，通过缩放确保梯度范数等于裁剪阈值$c$。具体裁剪函数定义为

$$Clip^{\text{Abadi}}(\boldsymbol{g}):\bar{\boldsymbol{g}}=\boldsymbol{g}/\max\left(1,\frac{\|\boldsymbol{g}\|_2}{c}\right) \tag{23}$$

这表明，小范数梯度能够在裁剪过程中被完整保留，而大范数梯度则可能丢失部分有用信息。由此导致的裁剪偏差对模型收敛性和最终性能产生直接影响，因此，对裁剪函数的改进与优化是提升DP-SGD模型性能的关键所在。

4.2.2 归一化裁剪函数

归一化裁剪(Normalized Clipping Functions)[22-23](如公式24)是一种在非隐私SGD中广泛使用的技术，其核心思想是将每个样本的梯度缩放到统一范数常数1。这种方法能够稳定训练动态，防止梯度爆炸[83, 85-86]，但也可能导致所谓的“惰性区域”(Lazy Region)问题[22]，即梯度缺乏足够的动量，难以跳出局部最优解。与Abadi裁剪不同，归一化裁剪实际上放大了小范数梯度(默认情况下，该归一化裁剪函数近似等价于Abadi裁剪中的$c\to 0$，即裁剪阈值从正区间接近于任意小的值)。归一化裁剪函数定义如下：

$$Clip^{\text{Normalized}}(\boldsymbol{g}):\bar{\boldsymbol{g}}=c\boldsymbol{g}/\|\boldsymbol{g}\|_2 \tag{24}$$

4.2.3 自适应裁剪函数

自适应裁剪函数(Automatic Clipping Functions)[22-23]是在归一化裁剪基础上的改进，如公式25所示。该方法引入了一个校正项γ，通过增加裁剪多样性，为梯度提供了更大的灵活性以跳出鞍点。相关研究表明，γ通常对算法性能不敏感，则不需要额外的调优开销，其值一般设置在0.01至0.0001之间。该裁剪函数定义为

$$Clip^{\text{Auto-S}}(\boldsymbol{g}):\bar{\boldsymbol{g}}=c\boldsymbol{g}/\left(\|\boldsymbol{g}\|_2+\gamma\right) \tag{25}$$

该方法在增加调参效率的同时，在计算机视觉任务和自然语言处理任务中，能够提升经典Abadi裁剪

DP-SGD精度0.1%-1%，减小了与非隐私算法间的差距。

4.2.4　受控自适应裁剪函数

上述两种归一化裁剪方法均属于单调的权重函数，该类型函数一方面对小范数梯度的存在过度放大的现象，对于更小范数的梯度甚至出现无限放大；另一方面，又对大范数梯度产生强抑制的情况。这种不平衡的抑制会扭曲梯度动态，通过过度强调信息量较少的梯度，削弱了更重要梯度的影响，使梯度实际贡献与其被放大的影响不匹配，进而阻碍模型优化过程。

为了缓解这一失衡问题，文献[24]提出了一种非单调的受控自适应裁剪函数(Controlled Adaptive Clipping Functions)(如公式26)，通过引入调节参数γ调控小范数梯度的放大幅度，有效抑制小范数梯度的过度放大，并减少归一化裁剪带来的范数偏差。该裁剪函数定义如下：

$$Clip^{\mathrm{DP-PSAC}}(g):\bar{g}=cg/\left(\|g\|_2+\frac{\gamma}{\|g\|_2+\gamma}\right) \tag{26}$$

该受控自适应裁剪方法能够进一步提升DP-SGD性能，在计算机视觉任务和在自然语言处理任务中，相对于非受控自适应裁剪有着0.1%-0.5%的提高。

4.2.5　非自适应与自适应裁剪的对比

总体而言，经典的Abadi裁剪需要同时调节裁剪阈值c和学习率η，导致调参复杂度为两者参数范围累积。相比之下，自适应裁剪方法通过将裁剪阈值c与学习率η合并为一个统一的参数$\tilde{\eta}=\eta\cdot c$(如公式27)，有效降低了调参复杂度至单个参数范围。

$$g/\max\left(1,\frac{\|g\|_2}{c}\right)\overset{c\to 0}{\approx} cg/\left(\|g\|_2+\gamma\right) \tag{27}$$

其中，γ通常取值为较小的常数。

具体而言，本文重新审视了Abadi裁剪与自适应裁剪之间的关系，发现当梯度范数大于等于1时，自适应裁剪实际上是Abadi裁剪($c=1$)的特例。因此，对于大范数梯度，两种裁剪算法在压缩效果上并无本质区别，而它们的主要差异体现在小范数梯度的处理上。

由此可见，小范数梯度对模型优化的贡献直接决定了裁剪函数的有效性。如果小范数梯度对模型性能的贡献有限，从可用性角度看，自适应裁剪方法未必是最优选择。尽管文献[39]在假设梯度分布对称的前提下，证明了自动裁剪方法能够实现近似最优的DP-SGD收敛速率，该速率同样可以通过Abadi裁剪实现[38]。

然而，考虑到Abadi裁剪在实际调参中存在显著困难，自适应裁剪方法为实践者提供了一种更加便捷的调优手段，能够在不增加复杂性的情况下更容易获得良好的模型性能，从而在实验中展现出更高的可用性和效率。

4.3　几何感知梯度裁剪

基于已有的理论研究[32, 38-77]，越来越多的工作认识到梯度的几何特性在梯度裁剪效果中发挥着关键作用。文献[38]首次提出，当梯度分布具有对称的几何性质时，有助于加速裁剪型DP-SGD的收敛。在其工作中，梯度在L_2范数球面空间内的对称性确保了同一半径上的梯度被裁剪的概率相同，从而更好地约束了裁剪偏差。从根本上来看，该工作揭示了DP-SGD中裁剪偏差主要受随机梯度噪声主导。具体而言，逐样本梯度裁剪概率的不均衡性会引入额外偏差，而这一偏差可通过基本的马尔可夫不等式，直接与随机梯度采样噪声的尾部性质在欧氏范数下建立联系。

在此基础上，文献[29]和文献[87]提出了在本征空间中优化裁剪过程的策略。直观上，这一优化过程可以理解为将大小为$\mathbb{E}\left[\|\mathcal{G}\|_2^2\right]$的梯度噪声期望压缩到其子空间版本，即$\mathbb{E}\|V_kV_k^T\mathcal{G}\|_2$，其中$\mathcal{G}$为随机梯度噪声和$V_k$代表投影子空间，从而减少梯度噪声对于隐私优化的影响。具体来说，这些工作通过在内在空间中利用投影技术，选择更具优势的低秩子空间来进行裁剪优化。由于选取的特征子空间能够有效排除冗余成分，这种方法显著增强了裁剪操作的效果。此外，这一预投影步骤在理论上也得到了严格证明，尤其是在深度学习中梯度特征值衰减迅速的情况下，可以实现维度无关的优化界限。

此外，文献[40]进一步指出，裁剪阈值应当对梯度分布的尾部行为保持敏感。相较于轻尾分布的梯度，在重尾场景下，梯度噪声的方差可能是无界的，导致无法直接使用马尔可夫不等式进行有效界定。因此，需要引入更强的集中不等式工具[88]，以高概率约束重尾噪声，并合理增大裁剪阈值来有效控制裁剪偏差。具体而言，研究者们利用梯度归一化方向信息对同一批次内梯度进行分类，将各个梯度划分为轻尾区域和重尾区域。基于这种分类机制，为不同尾部几何特性的梯度分别设置不同的裁剪阈值。

这一策略在确保隐私性的同时，还显著提升DP-SGD的模型可用性和优化效果。

4.4 针对梯度裁剪偏差的理论分析

非凸DP-SGD不可避免地需要对每个样本梯度进行裁剪，以确保潜在的相邻数据集灵敏度有界性，从而使噪声机制满足差分隐私。然而，与裁剪随机梯度下降（Clipped SGD）类似，梯度裁剪会引入额外的理论偏差，使得非凸DP-SGD理论研究变得尤为困难。已有文献[89]表明，梯度裁剪会直接削弱DP-SGD的优化性能，甚至在较优的损失函数条件下引入$\Omega(1)$的偏差。此外，Clipped SGD的相关研究也表明，梯度裁剪在一定程度上会影响算法的收敛速率[39, 83]。为此，本文分析了目前针对梯度裁剪偏差的理论分析结果。

4.4.1 针对分位数损失函数的理论结果

分位数损失函数（Quantile Loss Function）属于非光滑函数，被广泛应用于支持向量机、分位数回归和DP-SGD中的梯度裁剪选择问题。为了解决DP-SCO中分位数损失函数的非光滑特性，文献[78]采用卷积平滑技术[90]，直接获得了更加平滑的损失函数结构。与文献[65]中使用的Moreau-Yosida包络平滑方法相比，卷积平滑通过选择核函数和带宽的灵活性，而不仅仅局限于光滑参数，从而为DP-SCO提供了一个实现近似最优总体超额风险$\Theta\left(\max\left(\frac{\sqrt{d\log(1/\delta)}}{n},\frac{1}{\sqrt{n}}\right)\right)$的理论上界。

4.4.2 基于梯度几何特性的裁剪分析

实际上，由于使用传统理论工具分析带有裁剪偏差的差分隐私随机梯度下降Clipped DP-SGD极为困难，文献[38]从梯度对称性的角度重新研究了Clipped DP-SGD的FOSP。该对称性假设可以辨识为，$\mathbb{P}(\mathcal{G})=\mathbb{P}(-\mathcal{G})$，其中$\mathcal{G}$为定义的梯度噪声，基于一些实证观测[38]，通过梯度的对称性理论能够简化非凸优化界限的证明逻辑，避免了繁琐的裁剪概率估计和高概率尾界分析。在研究者们的工作中，声明在此假设下，梯度裁剪带来的偏差可以忽略，从而使Clipped DP-SGD能够达到与未包含裁剪分析下的DP-SGD相同的最优收敛率$\tilde{\Theta}\left(\frac{\sqrt{d}}{n\epsilon}\right)$。这一结果非常具有突破性，确认了Clipped DP-SGD的上界与DP-SGD保持一致。类似地，文献[22]在其自适应归一化裁剪的工作中也利用了该对称性假设，并进一步验证了在更强假设下，该界限确实是最优的。然而，值得注意的是，该假设在实际中难以得到保证。因此，在更一般的假设（非对称梯度）下，文献[23]进行了严格分析，并考虑了因裁剪偏差带来的损失，最终推导出Clipped DP-SGD的最优界限为$\tilde{\Theta}\left(\frac{d^{1/4}}{\sqrt{n\epsilon}}\right)$。当$d=\Theta(n)$时，该速率显著低于先前的界限，但当$d\to n^2\epsilon^2$时，该结果反而优于之前的界限，表明在高维模型中，裁剪有可能加速算法收敛。此外，文献[39]聚焦于联邦学习中的DP-SGD。其中，客户端在上传模型更新前必须进行裁剪并加噪，而这种裁剪操作与集中式DP-SGD中的梯度裁剪有所区别。研究者首先通过实验证明，在训练神经网络时，即便数据高度异质化，裁剪后的FedAvg仍能取得良好效果，这部分归因于在多种主流架构下客户端更新分布几何特性相似。基于这一关键观察，研究在对称性假设下进一步给出了联邦下DP-SGD的收敛性分析，揭示了裁剪偏差与客户端更新几何分布之间的关系。

从另一个视角，一系列文献[32, 77]从随机采样方差的角度分析了梯度裁剪对DP-SGD的影响。这些工作在理论上建立了采样噪声和裁剪偏差之间的联系，指出较大的随机采样噪声将导致更多的裁剪偏差。从直观上讲，由于单个样本梯度偏离期望值的存在，裁剪会导致大范数梯度丢失更多信息。因此，梯度的几何特性极大地决定了裁剪的有效性。

4.4.3 面向重尾数据的优化稳定性研究

在非隐私的随机梯度下降中，假设梯度呈现重尾分布的理论分析已受到广泛关注。大量深度学习的研究工作通过实证验证了，在模型训练过程中，梯度往往呈现比次高斯分布（Sub-Gaussian Distributions）更重的尾部[83, 91-94]。然而，这种更重的尾部意味有着更多偏离均值的样本，模型对这些样本的学习情况较为困难，会阻碍算法的收敛性。更具一般性的是，即使对于如MNIST这类标签相对均衡的简单数据集，其在卷积神经网络中的梯度仍表现出重尾特性[92]。此外，近期的差分隐私随机梯度下降研究广泛考察了重尾梯度分布在凸优化中的表现[19-21, 79-81, 95]。然而，与非凸场景下关于重尾现象的充分研究相比，非凸DP优化中的重尾现象仍处于不断探索阶段[19-20, 40]。

在DP-SCO方面，文献[79]针对次指数重尾数据分布的复杂性，通过引入梯度平滑和裁剪技术，在强凸损失函数下实现了高概率的总体超额风

险上界$\tilde{\Theta}\left(\frac{d^2}{n\epsilon^2}\right)$，在一般凸损失函数下实现了$\tilde{\Theta}\left(\frac{d^{2/3}}{(n\epsilon^2)^{1/3}}\right)$。在此基础上，文献[21]基于$\rho$-zCDP隐私定义，假设梯度噪声分布的$\theta$阶矩有界，即$\mathbb{E}\left[\left(\|\nabla l(\boldsymbol{w}_t)-\nabla L(\boldsymbol{w}_t)\|_2/K^{\theta}\right)\right]\leqslant 2$（其中$K$为正常数），在一般凸函数下将总体超额风险提升至$\Theta\left(\frac{d}{n^{1/3}}\right)$。基于相同的重尾假设，文献[81]提出了一种迭代更新方法，将强凸场景下的总体风险率进一步提升为$\tilde{\Theta}\left(\sqrt{\frac{d}{n}}+\sqrt{d}\left(\frac{\sqrt{d}}{n\epsilon}\right)^{1-\frac{1}{\theta}}\right)$。稍有不同的是，文献[20]基于重尾Lipschitz梯度假设$\mathbb{E}_{x\sim S}\left[\|\nabla\ell(\boldsymbol{w}_t,x)\|_2^{\theta}\right]\leqslant G^{\theta}$，采用了隐私版本的AC-SA算法[70]提高了效率，并在一般凸性和强凸性下分别达成了以下总体风险速率：$\tilde{\Theta}\left(\frac{1}{\sqrt{n}}+\max\left(\left(\frac{\sqrt{d}}{n\epsilon}\right)^{\frac{4(\theta-1)}{5\theta-1}},\left(\frac{\sqrt{d}}{n\epsilon}\right)^{\frac{\theta-1}{\theta}}\right)\right)$和$\tilde{\Theta}\left(\frac{1}{n}+\left(\frac{\sqrt{d}}{n\epsilon}\right)^{\frac{2(\theta-1)}{\theta}}\right)$。同样地，针对经验凸优化问题，文献[19]在超越统一Lipschitz条件的基础上，基于Łojasiewicz不等式，推导了更尖锐的凸EMR上界：对于一般凸损失函数为$\tilde{\Theta}\left(\left(\frac{\sqrt{d}}{n\epsilon}\right)^{1-\frac{2}{\theta+1}}\right)$，对于强凸场景为$\tilde{\Theta}\left(\left(\frac{\sqrt{d}}{n\epsilon}\right)^{1-\frac{1}{\theta}}\right)$。同时，该文献针对裁剪范数的选择也提出了具体的取值指导，认为裁剪阈值的选取应与重尾程度强相关。

在非凸优化中，文献[20]基于ρ-zCDP隐私定义，将经典的PL条件扩展到了更广义的非凸正则优化问题（Non-convex Proximal Optimization），在非凸场景下达成了收敛速率$\tilde{\Theta}\left(\left(\frac{\sqrt{d}}{n\epsilon}\right)^{\frac{2\theta-2}{\theta}}\right)$。该结果已接近强凸问题的下界。文献[19]基于Lipschitz梯度的θ阶矩有界假设，推导了经验风险的上界$\tilde{\Theta}\left(\left(\frac{\sqrt{d}}{n\epsilon}\right)^{1-\frac{2}{2\theta-1}}\right)$。这一结果比之前一致的Lipschitz（Uniform Lipschitz）假设更具普遍性，为无约束非凸问题中的DP-SGD提供了新的理论支持。此外，文献[40]在假设梯度服从重尾的次威布尔分布情况下，首次将Clipped DP-SGD中重尾指数的依赖降至多项式级别，其收敛率达到了$\Theta\left(\frac{d^{1/4}}{(n\epsilon)^{\frac{1}{2}}}\log^{\max\left(\frac{5}{4},\,\theta+\frac{1}{4}\right)}(T/\delta)\log^{2\theta}\left(\sqrt{T}\right)\right)$。

5　针对隐私噪声缓解的技术与理论

DP-SGD通常因其不可避免的过量隐私随机噪声，从而导致模型性能显著下降。大量研究工作致力于优化DP噪声过量问题，以推动其在实际中的广泛应用。目前，被广泛采用的解决方案主要通过优化梯度更新方向来减少冗余噪声的影响。具体而言，这类方法通过引导梯度在随机或更有价值方向上更新，从而降低高维梯度中无效噪声的干扰[96]。根据引导策略的差异，本文将现有噪声缓解技术进一步地划分为无向引导（Undirected Guidance）与有向引导（Directed Guidance）两类。需要强调的是，这些噪声缓解机制通常与具体的噪声注入方式或隐私组合定理无关，具备较好的通用性。此外，本节从面向噪声降维的非维度依赖理论和外部数据驱动的优化理论两方面，分析和总结了目前针对隐私噪声缓解技术的理论保障。表3为本节方法和理论的汇总。

表3　差分隐私随机梯度下降中的降噪方法

类别	优化机制	文献	ϵ	δ	任务模型	任务类型
无向引导	随机子空间	[97]	-	$1e^{-5}$	LSTM	情感分析
	随机稀疏化	[98]	{1, 3}	-	CNN	计算机视觉
		[99]	[0.5, 4]	$1e^{-5}$	Resnet18	计算机视觉
		[100]	[2, 10]	$[1e^{-7}, 1e^{-5}]$	Resnet9	计算机视觉
		[101]	[2, 8]	$1e^{-5}$	CNN, Resnet18	计算机视觉
	随机剪枝	[102]	-	-	CNN, VGG-16	计算机视觉

续表

类别	优化机制	文献	ϵ	δ	任务模型	任务类型
		[103]	$[0.5, 4.6]$	$1e^{-5}$	WRN16-4, Resnet50	计算机视觉
有向引导	自驱动方式	[104]	$[2, 6]$	$[1e^{-6}, 1e^{-5}]$	WRN28-4	计算机视觉
		[105]	$[1, 4]$	-	BERT	语言模型
		[106]	$[1, 8]$	$7.8\times 1e^{-7}$	WRN16-4, ViT	计算机视觉
	基于梯度干预的外部驱动	[107]	$[0.84, 3]$	$1e^{-5}$	Logistics Regression LSTM	语言、视觉
		[108]	$\{1, 3\}$	$1e^{-5}$	ResNet-50, CLIP	计算机视觉
		[109]	$[1, 10]$	$1e^{-5}$	WRN16-4	计算机视觉
		[96]	$[2, 8]$	$[1e^{-6}, 1e^{-5}]$	CNN, ResNet-20	计算机视觉
	基于子空间的外部驱动	[110]	$[0.23, 2.41]$	$1e^{-5}$	CNN	计算机视觉
		[87]	$[0.23, 1.3]$	$1e^{-6}$	CNN, Resnet18	计算机视觉
		[108]	$[0.8, 8]$	$1e^{-5}$	ResNet-50, CLIP	计算机视觉
		[111]	$[0.1, 10]$	$1e^{-5}$	CNN	计算机视觉

理论方向	优化目标	损失函数	测量指标	文献	上界	下界
差分隐私梯度下降维度无关理论	投影差分隐私梯度下降	投影非凸	经验 FOSP	[110]	$\tilde{\Theta}\left(\frac{\sqrt{k}}{n\epsilon}\right)$	×
		广义线性模型	总体超额风险	[25, 28, 44, 89]	$\Theta\left(\frac{1}{\sqrt{n}}+\min\left(\frac{\mathrm{rank}^{2/3}}{n\epsilon}, \frac{1}{(n\epsilon)^{3/7}}\right)\right)$	×
		非凸 GLMs	总体 FOSP	[42, 44, 89, 112]	$\Theta\left(\frac{1}{n^{1/2}}+\left(\frac{\sqrt{\mathrm{rank}}}{n\epsilon}\right)^{1/6}\wedge\frac{1}{(n\epsilon)^{3/7}}\right)$	×
		非欧氏对偶函数	总体超额风险	[113]	$\Theta\left(\frac{d^{1/2-1/p}}{\sqrt{n}}+\frac{d^{1-1/p}\log(d)}{n\epsilon}\right)$	√
针对外部数据驱动的优化理论	基于梯度干预的优化方法	公开数据梯度干预	凸总体超额风险	[44, 107, 108, 109, 114-116]	$\Theta\left(\frac{1}{\sqrt{n_{pub}}}, \frac{1}{\sqrt{n+n_{pub}}}+\frac{\sqrt{d}}{(n+n_{pub})\epsilon}\right)$	√
			凸经验超额风险	[117]	$\tilde{\Theta}\left(\min\left(\frac{n}{n+n_{pub}}, \frac{d}{(n+n_{pub})\epsilon}\right)\right)$	×
	基于辅助数据子空间的隐私优化	公开数据子空间	经验 FOSP	[96, 108, 110]	$\tilde{\Theta}\left(\frac{\sqrt{k}}{n\epsilon}+\frac{\sqrt{\log(d)}}{\sqrt{n_{pub}}}\right)$	×
		合成数据子空间	经验 FOSP	[87]	$\tilde{\Theta}\left(\frac{\sqrt{k}}{n\epsilon}+\max\left(\frac{\sqrt{T}}{\sqrt{n_{syn}}}, \frac{1}{n_{pub}^{3/10}}\right)\right)$	×

5.1 基于无向引导的噪声缓解技术

针对噪声缓解的无向引导方法[97-98, 100, 102-103, 105-106]，主要集中于通过随机技术在各个维度上均匀地减少噪声冗余，以降低DP-SGD中整体的噪声影响。随机稀疏化是机器学习领域中最常用的优化技术之一，通常用于模型压缩和优化。文献[98]在DP-SGD中提出，在梯度裁剪之前对逐样本梯度进行预先随机稀疏化，从而压缩逐样本梯度的L_2范数。这一策略能够在保持相同裁剪阈值的情况下减少裁剪偏差，并保留更多的梯度信息。然而，由于不同模型对可用性的需求不同，该方法需要精细调节稀疏率。

类似地，文献[103]将随机稀疏化同时应用于模型参数和梯度空间，在计算机视觉任务的分类模型中，在较低隐私预算下取得了更高的准确率。文献[105]探讨了随机稀疏化在大模型微调中的效果，发现对50%的模型参数进行稀疏化能够有效提升模型的可用性。文献[100]则针对等变卷积层提出了一种新的稀疏化方法，通过在卷积核中引入额外的变换，使网络能够识别旋转和反射等变化不变的特征，因此在减少DP噪声的同时降低了模型的参数需求。文献[102]将稀疏化技术引入隐私保护的联邦学习中，通过在客户端本地运行DP-SGD并对模型进行稀疏化，既降低了大规模联邦模型的通信开销，又减少了额外的隐私损失。此外，文献[101]研究了联邦学习中用户级(User-Level))DP-SGD带来的性能下降问题。现有方法虽然能提供严格的隐私

保证，但往往导致模型精度显著降低。作者通过分析发现，性能退化的核心原因在于未对本地更新范数进行有效约束。为此，研究者提出了两种技术：有界本地更新正则化与本地更新稀疏化（Local Update Sparsification），旨在在执行用户级DP操作前限制本地更新，从而提升模型质量与精度。其实验结果显示，该框架在保证用户级隐私的同时，显著改善了隐私与效用之间的权衡。

上述工作主要基于参数空间或梯度空间进行优化。此外，一些工作通过在本征空间（Intrinsic Space）中采用随机技术也实现了无向引导。文献[97]提出了一种新的差分隐私优化器，用于近似估算每个样本的梯度范数，而无需精确计算。具体来说，研究者们利用Johnson-Lindenstrauss（JL）投影快速地估计每个样本的梯度范数。具体地，JL投影通过将向量投影到一个均匀随机的低维空间，并利用在该投影空间的梯度范数来近似计算原始逐样本梯度的L_2范数。通过多次执行这种投影，实验上可以获得更精确的L_2范数估计。

此外，在缺乏真实图像数据作为指导的情况下，生成模型能够学习到接近真实图像的视觉噪声先验。在这一背景下，文献[106]利用未训练的StyleGAN[118-119]生成的纹理和着色器样式噪声，以及未训练的StyleGAN提取的结构先验，用于隐私训练。通过这种方式生成的随机噪声先验，可用于训练具备公共图像变换不变性的特征模型，通过对合成图像的预训练，获得对背景或结构性视觉特征的隐私训练能力。除了在下游隐私数据集上使用随机初始化，研究者还采用表征学习方法获取“热启动”模型，将随机过程中的先验知识编码进初始模型参数中，从而提高模型的隐私训练效果。

5.2　基于有向引导的噪声缓解技术

有向引导指的是在DP-SGD中，通过优先关注特定的维度或方向进行梯度更新，从而有效缓解噪声过大的问题。不同于无向引导的随机性，有向引导强调识别并利用梯度空间中最具价值的成分，例如历史优化方向或潜在分布类似的子空间，以提升梯度更新的信噪比。通过在这些目标区域集中更新参数，有向引导不仅能够减少冗余噪声的影响，还能显著提升模型的可用性和收敛速率。有向引导可以按照指导来源，更细粒度地分为自驱动和外部驱动两种类型。

5.2.1　自驱动方式

自驱动指导指不依赖任何辅助数据或额外知识，仅基于隐私数据本身，通过利用历史更新信息（如累计梯度或模型权重）来抵消过大的隐私噪声，从而辅助并增强模型训练过程。已有研究发现，在隐私训练中可以重复利用历史梯度信息。文献[104]首先计算历史权重的累计更新量$\triangle_t^w=\boldsymbol{w}_t-\boldsymbol{w}_0$，并对其进行奇异值分解（SVD），获得左奇异矩阵$\boldsymbol{L}_t^w$和右奇异矩阵$\boldsymbol{R}_t^w$。随后，在投影到$\boldsymbol{L}_t^w$和$\boldsymbol{R}_t^w$的训练数据上分别计算分解后的梯度$\partial\boldsymbol{L}_t^w$和$\partial\boldsymbol{R}_t^w$，并分别在各部分添加噪声。最后，通过$\boldsymbol{L}_t^w$和$\boldsymbol{R}_t^w$构建的投影上将$\partial\boldsymbol{L}_t^w$和$\partial\boldsymbol{R}_t^w$重构回原始梯度空间，迭代更新中关键步骤为：

$$\left(\partial\boldsymbol{L}_t^w\right)\boldsymbol{R}_t^w+\boldsymbol{L}_t^w\left(\partial\boldsymbol{R}_t^w\right)-\boldsymbol{L}_t^w\boldsymbol{L}_w^t\left(\partial\boldsymbol{L}_t^w\right)\boldsymbol{R}_t^w \quad (28)$$

类似地，在非隐私随机梯度下降中，锐度感知最小化技术（sharpness-aware minimization，SAM）通过感知梯度二阶信息，来帮助模型更好地找到平坦的最优点，受此启发，文献[120]尝试将SAM与DP-SGD相结合，但由于SAM会带来额外的训练开销负担，会进一步加重隐私训练负担。因此，研究者利用历史梯度的海森矩阵指引模型SAM训练，实现了DP-SGD可用性与效率的提升。考虑到隐私预算往往被反复用于保护已“学到”的历史信息，文献[121]提出了DPDR框架，通过梯度分解与重构（gradient decomposition and reconstruction，GDR）技术识别和回收历史知识，在节省隐私预算的同时加速模型的早期收敛。GDR将当前梯度分解为正交部分和平行部分，优先向正交部分添加噪声，以减少多余的信息增益。该框架采用混合策略，在模型训练早期使用GDR，后期切换回DP-SGD，以最大化隐私预算的利用率。理论分析和实验证明，在相同的隐私水平下，DPDR相比于标准DP-SGD能够实现更快的收敛速度和更高的模型精度。

自知识蒸馏（self-knowledge distillation，SKD）[122-123]是一种通过教师-学生架构促进知识迁移的技术，也可应用于隐私学习场景，有效提升模型性能。不同于传统的知识蒸馏（Knowledge Distillation）[124]，SKD将历史梯度视作教师模型，辅助后续模型训练。SKD的主要挑战在于，在保证隐私的前提下，使学生模型能够有效从教师模型中学习知识。传统蒸馏方法通常同时考虑目标类别和非目标类别的损失，这在隐私和准确性之间的平衡上存在困难。为此，研究者们将隐私保护直接集成到知识迁移过程中。这类方法能够通过存储中间模型检查点构建教师模型，使学生模型通过自身的历史学习过程逐渐充当教师，

从而提升学习效果。同时，该方法将蒸馏损失重新划分为目标类别和非目标类别两部分，有助于学生模型获取更加丰富和清晰的知识，提高模型在隐私保护下的泛化能力。

5.2.2 外部驱动方式

与自驱动方法不同，外部驱动策略依赖有限规模的公共数据或者合成数据，在不违反隐私约束的前提下提升模型性能。正如前文所讨论的，由于隐私数据不可避免地受到差分隐私噪声的影响，模型往往难以跳出困扰收敛的局部最优解。借助辅助数据对于隐私学习的增益，研究者主要从两大类方向优化了DP-SGD：

(1)基于梯度干预的外部驱动

基于梯度干预的DP-SGD典型方法包括热启动初始化(Warm-up Initialization)[44, 114-115, 109]和梯度代理(Gradient Proxies)[107-108, 116]。具体而言，热启动策略[114-115, 125]利用公共的同分布数据生成良好的模型初始化，相当于基于公共数据预训练一个热启动模型。文献[44]证明，当$n_{pri}\epsilon > d$时，热启动初始化可在$\frac{d^{2/3}}{n_{pri}\epsilon}$的收敛速率下引导模型达到更优的稳定点。此外，文献[126]进一步展示，通过采用先进的扩散模型(如Elucidating Diffusion Model[127])，结合判别器指导[128]和多样性增强[129]技术，生成的合成数据能够显著提升模型性能。具体地，相比于DP-SGD，该方法能够基于Vision Transformer模型，在CIFAR-10与CIFAR-100数据集上，超过基准线将近30%-40%的精度，并在低隐私预算下维持该优势。与此同时，合成公共梯度也被用作真实梯度的近似，直接与隐私梯度共同用于训练[107-108]，或用于动态调整DP-SGD中的梯度裁剪阈值[114, 116]。

(2)基于子空间的外部驱动

近期研究还持续关注基于子空间的DP-SGD优化器[96, 104, 108, 110, 130-131]的发展。其中最早且具有重要影响力的是文献[110]提出的投影差分隐私随机梯度下降(projected differentially private stochastic gradient descent，PDP-SGD)，该方法通过将隐私敏感数据的梯度投影到低秩子空间中，有效降低了隐私经验风险。该方法将维度相关误差从$\Theta\left(\frac{d}{n_{pri}\epsilon}\right)$降低至$\Theta\left(\frac{k}{n_{pri}\epsilon}\right)$，$k$为投影维度，即使仅使用1%的公共数据，也能大幅提升模型性能。具体地，由于DP-SGD需要对梯度中的每个维度加噪声，其在低隐私预算条件下的表现十分受制，并且目前主流模型趋向更大规模和维度，使得DP-SGD的精度对比非隐私SGD下降超过10%。相比之下，PDP-SGD即使在高维条件与低隐私预算下，仍能够维持可用性，提升DP-SGD多达5%～20%精度。然而，由于子空间基于有限公共数据学习，存在额外的投影重构误差。

类似地，文献[96]提出了GEP方法，将隐私梯度投影到非敏感的锚点子空间中，然后分别对低维嵌入和残差梯度添加噪声，在收敛率上获得了$\sqrt{k}$的改进。该方法在仅使用3%公共数据的情况下，实现了$\Theta\left(\frac{T\sqrt{k}}{\sqrt{n_{pub}}}\right)$的投影误差。文献[108]则引入了未标注的公共数据(占训练数据的10%)和文本侧信息，利用公共子空间进行小样本训练和跨模态零样本学习，显著提升了计算机视觉任务中的性能。

同时，在联邦框架下，文献[111]针对联邦学习中不同客户端存在异质化隐私需求的问题，提出了联邦框架下的DP-SGD优化方法。传统方法通常假设所有客户端共享相同的隐私预算，但在实际应用中，由于政策或个人偏好差异，不同客户端往往具有不同的隐私要求。文献将隐私预算较大的客户端定义为“公共客户端”，隐私预算较小的定义为“隐私客户端”，并提出Projected Federated Averaging (PFA)方法：通过提取公共客户端更新的主奇异子空间，将隐私客户端的DP-SGD更新梯度投影后再聚合，以更好地利用高质量信息而避免模型偏差。在此基础上，作者进一步提出PFA+，允许隐私客户端直接上传投影后的更新，从而大幅减少通信开销。实验结果表明，PFA与PFA+均能在保证隐私的前提下显著提升模型效用，其中PFA+更是实现了超过99%的上行通信量减少。

综上所述，基于外部数据的梯度干预DP-SGD优化策略与基于子空间的DP-SGD方法具有本质区别。前者直接利用大规模公共或者合成数据(通常超过训练数据10%)计算梯度，用于影响反向传播过程并提升模型效用而后者则通过引入锚点限制梯度下降的子空间，在较低隐私预算下依然保持较强的模型性能。

5.3 面向噪声降维的维度无关与外部优化理论

首先，DP-SGD中的差分隐私噪声通常与模型维度呈现线性相关。高维模型因其在真实场景中的

优异的任务表现而被广泛应用。相比之下，众多理论界限都局限于低维设定[28, 44]，即当样本数量n、隐私预算ϵ与模型维度d满足$n\epsilon=\Theta(d)$时，才能取得比基线DP-SGD更好的收敛速率。同时，已有大量研究致力于在凸优化假设下减少DP-SGD上界对模型维度的依赖[89, 132-133]。因此，在高维非凸设定下研究（近似）无维度依赖的DP-SGD成为一个极有挑战和潜力的方向[96, 110]。

其次，目前研究[56, 115, 134]通过引入公共数据或者合成数据，弥合DP-SGD与非隐私SGD之间的性能差距。已有实验研究表明[115]，基于公共数据预训练的DP微调显著提升了模型性能，进一步强调了辅助数据在改善隐私优化中的潜力。同时，利用大语言模型和其他生成模型生成的合成数据在隐私保护中的应用也日益受到关注[135]。因此，本节将从理论层面探讨了DP-SGD维度无关理论和基于外部数据驱动的隐私优化理论。

5.3.1　差分隐私梯度下降维度无关理论

考虑到DP噪声会遍布整个环境维度，在高维模型中可能引发噪声灾难。为此，文献[110]提出在近似k维低秩子空间中执行DP-SGD，将梯度和DP噪声都投影到预构建的投影子空间上。由于高斯噪声的各向同性性质，该方法线性降低了算法对模型维度d的依赖，从而实现了近似无维度依赖的收敛率$\tilde{\Theta}\left(\frac{\sqrt{k}}{n\epsilon}\right)$。在此基础上，文献[96]进一步改进了该方法，通过将梯度分解为锚点分量和残差分量，锚点部分使用投影梯度下降方法处理，残差部分利用梯度范数界限控制，最终证明了整个算法的上界几乎不依赖于模型维度。

此外，针对广义线性模型（GLMs）的隐私非维度依赖理论也被广泛研究，其通常通过模型参数和特征输入之间的内积关系（或原始输入的映射特征嵌入）的损失函数而构建的模型框架。这一框架为诸如线性回归、逻辑回归和泊松回归等差分隐私数据分析提供了基础。目前，文献[44]中的结果达到了针对GLMs的最优界限，研究者提出了一种热启动Warm-up策略，通过指数机制预选一个具有隐私保护的初始点而非随机初始点，该热启动的初始点更接近DP-SGD训练的轨迹，然后结合DP-SPIDER算法，最终实现了有条件的改进总体超额风险界$\Theta\left(\frac{1}{\sqrt{n}}+\min\left(\frac{\mathrm{rank}^{2/3}}{n\epsilon},\frac{1}{(n\epsilon)^{3/7}}\right)\right)$，该结果意味着模型能够更快地逼近全局最优点。相似地，在非凸GLMs下，文献[112]将Johnson-Lindenstrauss变换应用于高维GLMs的随机投影，确保了Lipschitz假设，GLMs总体的近似维度无关率$\Theta\left(\frac{1}{\sqrt{n}}+\left(\frac{\sqrt{\mathrm{rank}}}{n\epsilon}\right)^{2/3}\wedge\frac{1}{(n\epsilon)^{2/5}}\right)$。在此基础上，文献[44]进一步采用隐私初始点热启动技术，在$d<n\epsilon$条件下，将该速率提升至$\Theta\left(\frac{1}{n^{1/2}}+\left(\frac{\sqrt{\mathrm{rank}}}{n\epsilon}\right)^{1/6}\wedge\frac{1}{(n\epsilon)^{3/7}}\right)$。

上述已讨论的理论结果均基于L_2-几何分析。然而，在非欧氏几何空间（Non-Euclidean L_p-Geometry，$p\geq1$）中，DP梯度下降所需噪声的维度依赖性显著增强（特别是L_1-几何中从$\sqrt{d}$增大至d），给理论分析带来了巨大挑战。文献[46]也在非欧氏范数空间的DP维度无关理论发展中做出了重要贡献，基于随机Frank-Wolfe框架系统性地研究了三种情况：

（1）对于$p=1$，利用Report Noisy Argmax机制[2]满足DP要求，得到了几近维度无关的超额风险界$\Theta\left(\frac{\log(d)}{\sqrt{n\epsilon}}\right)$。

（2）对于$1\leqslant p\leqslant2$，扩展了广义高斯机制，并提出了新的带噪随机镜像下降（stochastic mirror descent，SMD）算法，总体超额风险界为$\Theta\left(\frac{1}{\sqrt{n}}+\frac{\sqrt{d}}{n^{3/4}}\right)$，在低维场景下实现了最优性；

（3）对于$p>2$，进一步给出了最优的总体超额风险界$\Theta\left(\frac{d^{1/2-1/p}}{\sqrt{n}}+\frac{d^{1-1/p}\log(d)}{n\epsilon}\right)$，适用于$n=\Omega(d)$的低维场景。

5.3.2　基于外部数据驱动的隐私优化理论

（1）基于梯度干预的混合隐私优化

大量实验和理论研究探讨了如何通过有限规模的公共数据或由公共来源生成的合成数据进行隐私学习。这一领域的两个核心策略被称为基于梯度干预的DP-SGD，包括热启动初始化方法[44, 114-115, 126]和梯度代理技术[107-108, 116]。这些方法通过直接将较大

规模的公共或合成数据集的梯度引入训练过程，并主动干预反向传播，以提升模型的可用性，相比基于子空间的方法，这类方法涉及的辅助数据规模更大。

文献[117]从理论上研究了结合公共梯度的隐私优化（也称为semi-DP学习）的界限。具体而言，对于强凸Lipschitz场景，研究者们得到了如下的经验超额风险上界：$\tilde{\Theta}\left(\min\left(\frac{n}{n+n_{pub}}, \frac{d}{(n+n_{pub})\epsilon}\right)\right)$，其中，当$d/\epsilon > n$时，这一结果优于仅依赖隐私数据的传统DP-SGD。对于总体超额风险，该工作推导了严格的上界：$\tilde{\Theta}\left(\min\left(\frac{1}{\sqrt{n_{pub}}}, \frac{\sqrt{d}}{(n+n_{pub})\epsilon} + \frac{1}{\sqrt{n+n_{pub}}}\right)\right)$，该结果在$n_{pub}d > (n+n_{pub})^2\epsilon^2$或$n_{pub}=\Theta(n)$条件下表现最优。此外，文献[136]研究了公共梯度辅助的DP-SCO的下界，其速率为$\Omega\left(\frac{1}{\sqrt{n_{pub}}}, \frac{1}{\sqrt{n+n_{pub}}} + \frac{\sqrt{d}}{(n+n_{pub})\epsilon}\right)$，并发现这一下界与文献[117]中最优上界一致。进一步地，研究者们将理论框架扩展到了无标签公共数据的凸GLMs场景，并证明了在充足公共数据$\left(n_{pub}=\Theta(n\epsilon)\right)$的情况下，可以实现几近维度无关的总体风险上界：$\tilde{\Theta}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n\epsilon}}\right)$。

然而，现有文献研究中发现，若公共数据不足，反而可能降低优化效果，这提示了过度依赖小规模公共数据存在潜在风险。目前，关于预热策略的理论分析主要通过指数机制选择初始模型参数，从而提升DP-SPIDER算法的经验风险最优性。尽管预热阶段未直接使用公共数据，但这一工作为基于公共数据的预训练方法提供了理论依据，并强调了预热在隐私学习中的重要性。

对于基于合成数据的梯度干预优化，已有文献[109]利用扩散模型从有限公共数据中合成额外的独立同分布数据，并通过合成梯度代理有效提升隐私训练性能。然而，这一新兴领域仍存在诸多未解难题，包括合成数据的泛化能力、与原生数据的分布偏差等问题，甚至担忧合成数据可能带来的额外隐私风险，这些问题尚缺乏深入的理论探讨和强有力的理论保证。为更好地描述基于梯度干预的方法，本文明确给出其不同的更新过程，其中热启动初始化参数方法中梯度更新方式为公式(29)，以及梯度代理方法如公式(30)所示。

$$\boldsymbol{w}_t = \boldsymbol{w}^*_{pub} \vee \boldsymbol{w}^*_{syn} + \boldsymbol{w}_{t-1} - \eta_t(\boldsymbol{g}_t + \boldsymbol{\xi}_t) \quad (29)$$

$$\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \eta_t\left(\boldsymbol{g}_t + \left(\boldsymbol{g}^{pub}_t \vee \boldsymbol{g}^{syn}_t\right) + \boldsymbol{\xi}_t\right) \quad (30)$$

(2)基于辅助数据子空间的混合隐私优化

鉴于在深度学习中，梯度下降过程往往集中在关键的特征值分量上[115, 137-141]，基于子空间的DP-SGD优化策略得到了广泛研究。相较于基于梯度干预的DP-SGD，基于子空间的方法通常在由公共或合成数据构建的受限投影子空间中进行优化[87, 96, 108, 110, 142]。如前一节所述，基于子空间的方法能够达到近乎不依赖模型维度的收敛率$\tilde{\Theta}\left(\frac{\sqrt{k}}{n\epsilon}\right)$。然而，这种改进依赖于从少量辅助数据中估计其二阶矩，在构建的过程中，辅助数据不可避免地引入了额外的误差——即投影重构误差。为此，一系列工作围绕这些权衡量进行改进与优化。

在采用独立划分策略（即辅助数据不在各迭代间复用）的场景中，文献[110]利用Ahlswede-Winter不等式对误差进行了上界估计，得到了$\tilde{\Theta}\left(\frac{\sqrt{k}}{n\epsilon} + \frac{\sqrt{T}}{\sqrt{n_{pub}}}\right)$的FOSP理论上界。而在允许公共数据复用的场景下，通过泛化工具如泛化链（Generic Chaining）技术[143]，可以得到更紧的一致性上界$\tilde{\Theta}\left(\frac{\sqrt{k}}{n\epsilon} + \frac{\sqrt{\log(d)}}{\sqrt{n_{pub}}}\right)$。相比之下，文献[96]基于公共数据构建的公共子空间中引入了更显著的误差$\tilde{\Theta}\left(\frac{T}{\sqrt{n_{pub}}}\right)$，该误差严重依赖于每轮迭代使用的公共数据量。这些结果表明，随着公共数据规模的增加，投影重构误差会逐渐减小。然而，在实际应用中，尤其是医疗等敏感领域，受限于法律和伦理要求，含有可识别信息的公共数据往往难以获取。因此，当$n_{pub} \ll n$时，这种误差的存在会对隐私优化产生不可忽视的负面影响。

针对这一问题，文献[87]提出了一种基于合成数据的投影增强算法，在不复用数据的情况下，将投

影重构误差降低到了$\tilde{\Theta}\left(\max\left(\frac{\sqrt{T}}{\sqrt{n_{syn}}},\frac{1}{n_{pub}^{3/10}}\right)\right)$。为了进一步分析辅助数据不足带来的风险，该工作定义$n_{pub}=n^{\alpha}$和$n_{syn}=n_{pub}^{\beta}$，用于联合分析各种混合数据场景，并重新校准了基于公共辅助的子空间DP-SGD的EMR收敛率：$\tilde{\Theta}\left(\frac{\sqrt{k}}{\left(n\epsilon'\right)^{\alpha\beta/3}}\right)$，其中，$\epsilon'=\min(\epsilon,1)$。通常情况下，基于子空间的方法中$\alpha$取值一般在(0,1)区间内。因此，该校准界从更统一的视角重新审视了有限公共数据对隐私优化的影响，相比以往依赖$n\epsilon$倒数的界限，这一结果更加合理。此外，对于基于合成数据增强的子空间DP-SGD，研究者们得到了如下收敛率：$\tilde{\Theta}\left(\frac{\sqrt{k}}{\left(n\epsilon'\right)^{\alpha f(\beta)/3}}\right)$，其中$f(\beta)$为线性函数且存在上界，表明当合成数据量超过公共数据时，模型性能会受到积极影响。然而，由于合成数据的优势受限于有限的公共知识，额外的知识无法凭空生成，合成子空间的泛化能力也同样是局限的。基于子空间的方法的更新过程可描述为公式(31)内容：

$$\boldsymbol{w}_t=\boldsymbol{w}_{t-1}-\eta\boldsymbol{V}_{k,t}\boldsymbol{V}_{k,t}^{\mathrm{T}}\left(\boldsymbol{g}_t+\boldsymbol{\xi}_t\right)\tag{31}$$

其中，$\boldsymbol{V}_{k,t}\boldsymbol{V}_{k,t}^{\mathrm{T}}$为公共或者合成数据衍生的子空间。

6　DP-SGD在大语言模型中的应用

随着大语言模型的迅猛发展，它们已深度嵌入日常生活与工业生产，显著提升了智能化服务水平。然而，模型日益增强的记忆与生成能力也带来了前所未有的隐私泄露风险，引发学术界和工业界的高度关注。在理论上，差分隐私作为一项强有力的隐私保护框架，仍可适用于LLMs的优化过程。然而，在实际部署中，将DP-SGD有效应用于大模型训练与推理任务，面临着一系列新兴挑战，包括模型参数空间的高维性、微调机制的异构性，以及内部推理路径的复杂性和不可解释性等。

为应对上述挑战，当前研究已围绕LLMs中的差分隐私训练展开多维度探索，涵盖从全量训练，再到参数高效微调的多种优化范式。同时，考虑到部分实际场景中模型参数不可更新的限制，近年来也涌现出一系列面向非参数化推理过程的差分隐私方法(如DP In-Context Learning)，为模型部署阶段的隐私保护提供了新的思路。为此，本文系统梳理并分析了当前在LLMs优化各阶段中差分隐私机制的应用路径，聚焦可用性、可扩展性与合规性三个目标，以期为未来隐私安全的大模型训练与推理提供理论基础与方法指导。

6.1　针对全参数的隐私模型训练

本节总结了现有面向大模型全参数训练的DP-SGD应用细节与方案。将DP-SGD应用于完整模型训练时面临三大挑战：(1)随着模型维度和数据集复杂度的提升，DP-SGD对模型性能的削弱更加严重；(2)大规模数据处理对存储和计算资源提出了更高要求；(3) DP-SGD的超参数高敏感性在大模型中依然存在，同时下游任务的网络复杂性和多样化优化目标进一步加剧了超参数调优的难度。为此，本文分别研究面向高维参数的高可用隐私训练、面向大规模计算的高效解构训练及针对可调参数的经验指导。

6.1.1　面向高维参数的高可用隐私训练

为缓解高维DP噪声带来的负面影响，文献[144]发现，将批量大小扩展到约200万时，显著提升了DP-SGD在BERT基础大模型[145]上的训练表现。文献[146]则发现，梯度更新的维度并非决定隐私训练性能的关键因素。尽管理论上隐私(凸和非凸)优化的下界存在对维度的依赖，但实证结果表明，预训练模型越大，往往能带来更优的隐私训练效果。此外，通过减少梯度更新维度的参数高效适配方法，并不一定优于对所有模型参数进行微调的基线方法。文献[147]提出了一个基于Hessian矩阵的分析框架，用于研究不同优化器下DP训练的样本损失，并探讨了梯度裁剪、噪声添加和超参数选择对泛化损失的影响。研究者们进一步从理论上解释分析出，相较于梯度裁剪带来的偏差，预训练阶段中的模型参数对差分隐私噪声更为敏感。

6.1.2　面向大规模计算的高效解构训练

为降低计算成本并减少模型复杂性，文献[144]提出了一种动态批量大小策略以提升训练效率。通过采用递增批量大小的调度策略，该方法在维持固定批量精度的同时优化了训练效率。其实验结果表明，该策略可以减少约14%的数据访问量以达到相同的模型精度。文献[146]在DP-SGD上应用了Ghost Clipping技术，通过逐层裁剪策略减少了内存消耗，使大规模Transformer模型能够在DP约束下进行训练，其性能接近非隐私模型且仅需在每次更新中增加一次反向传播。此外，文献[104]提出了重

参数化梯度扰动方法，将原始的高维权重矩阵替换为一对低秩梯度载体和一个残差权重项。通过重参数化，可以保持前向和反向传播的信号不变，这大大降低了DP-SGD训练中的内存消耗，并减少了单独梯度计算和存储的需求。这种方法支持全参数训练，特别适用于大规模参数上的模型训练。

6.1.3 针对可调参数的经验指导

在超参数调优方面，文献[144]观察到DP-SGD与缩放不变层之间存在负向交互，通过调整这些层的权重衰减超参数可有效提升模型性能。文献[146]指出，通过合理调优超参数并对齐下游任务目标，可以在较低的隐私预算下实现对所有模型参数的高效隐私训练，甚至超越非隐私基线模型。此外，从文献[144]可知，在大语言模型中采用更大的批量能够提升模型可用性，基于大批量的梯度能够中和随机梯度与平均梯度的偏移，这种偏移误差原本在大规模训练文本下是占据主导的。同样，在文献[104]的工作中，采用了小批量大小为1,000的超参数配置，可以增强期望梯度的平均效果，并减缓DP噪声的方差影响。

6.2 基于参数高效的隐私微调

参数高效微调（parameter-efficient DP fine-tune，PEFT）是当前活跃的研究方向之一，通过仅更新大模型中的一小部分参数实现高性能模型。这一方法基于机器学习模型存在内在维度性（Intrinsic Dimensionality）的广泛假设，即模型训练所需的最小参数数量远小于其总参数数量。相关文献[138]探讨了大模型的内在维度性，并发现仅通过随机投影选取少量参数进行训练，依然可以达到大部分模型性能。

基于此，本文对DP微调技术（利用DP-SGD训练微调数据）进行了综述，并从以下三个方面讨论了现有研究：高效的模型性能提升方法、宽松的隐私定义与隐私预算调度策略及DP微调中的公平性和隐私审计问题。

6.2.1 高可靠的差分隐私微调范式

为缓解大语言模型带来的显著计算和存储开销，一系列工作采用了参数高效的方法，仅更新原模型的一小部分参数，从而提升了训练的可扩展性。文献[148]引入了额外的可训练参数，这些参数仅占预训练LLMs总参数的很小一部分。在微调过程中，仅对这些参数应用DP-SGD，而保持预训练权重冻结。这种设计确保即便DP噪声导致微调参数高度噪声化，也不会干扰预训练权重中大部分方向编码的信息，从而避免了“遗忘”预训练阶段获得的知识。在实验中，研究者们在MNLI数据集[149]上使用RoBERTa-Base和RoBERTa-Large[150]，以及在E2E自然语言生成数据集[151]上使用GPT-2[152]，分别在LoRA[153]、Adapter[154]和Compacter[155]方法上实现了最优效果。这些结果表明，更大的语言模型通常能实现更高的准确率，并且参数高效的方法在效果上能够超越全参数微调。在差分隐私微调下的自然语言生成任务中，为了实现明确控制并生成带标签样本，文献[156]结合DP-SGD构造了控制代码，例如格式“type|rating”，通过将控制代码前置于每个样本，从而引导生成与原数据集类别分布一致的样本。此外，文献[157]利用基于BERT的LLMs探索了DP-SGD的有效性，表明DP微调不总是遵循“更大模型更优”的通用规律，其结果强调了任务特定的优化和隐私机制调整的必要性。文献[158]考虑到大语言模型中的嵌入模型经常发生梯度稀疏性，研究者们通过选择性地向预选嵌入行添加噪声来提升隐私微调效率。文献[147]进一步对DP机制的影响进行了全面分析，特别是噪声和梯度裁剪在微调阶段的影响，并提出了微调阶段对梯度裁剪更为敏感的理论解释。在优化梯度裁剪策略下，文献[22]和文献[24]探讨了参数高效微调方法（如LoRA和prefix-tuning[159]）的有效性，其实证结果表明这些方法在性能上可以与传统全参数微调相媲美。

6.2.2 灵活的隐私定义和预算调度策略

文献[160]发现，在大模型场景下，句子中的私人信息往往是稀疏的，并非所有属性都需要保护。为此，研究者们提出了一种新的差分隐私定义——选择性差分隐私（selective differential privacy，SDP），该定义旨在仅对记录中的敏感部分提供隐私保护。在此基础上，研究者们将隐私和公共tokens分组，并开发了Selective-DPSGD机制以实现隐私和可用性之间的更好平衡。然而，该机制要求语言模型预先了解如何区分公共和私人tokens，并且未考虑上下文中潜在的敏感信息以及可能推断出的敏感tokens。为了解决这一问题，文献[161]提出了一种两阶段的隐私微调机制来实现SDP。具体而言，研究者们首先编辑下游任务的领域内数据并基于这些数据进行微调，然后再在隐私数据上使用DP-SGD进行微调。这种编辑微调步骤使模型能够直接学习领域知识，为Selective-DPSGD提供更好的初始化。同样，文献[162]提出了一种基于句子序列级别的差分隐私定义，用于在大模型分布式微调

场景中，通过对前向传播的嵌入施加选择性高斯噪声[163]来实现隐私保护。此外，考虑到LLMs微调迭代次数有限，不同于传统的百万轮训练[164]，现有的隐私组合定理可能并不适用。为此，文献[165]引入了Edgeworth Accountant[166]，一种更适合有限轮次的隐私预算累积方法，使得噪声规模更加宽松，从而有效提升LLMs DP微调性能。

6.2.3 高效微调中的公平性和隐私审计

在LLMs微调过程中，DP会放大与公平性相关的偏见，如性别、种族和宗教偏见。因此，相比非DP微调模型，DP模型更容易生成偏见性文本。文献[167]分析了这一现象的原因，认为DP主要通过注入噪声减少了个别样本的影响，但由于刻板印象关联在训练数据中远多于非刻板印象关联，添加的噪声更容易影响后者，导致模型难以学习非刻板印象的关联。为此，研究者提出了反事实数据增强（counterfactual data augmentation，CDA）策略，通过生成反事实句子对（例如："A man/woman works as a doctor"）减少模型对刻板印象的依赖。在差分隐私约束下引入CDA，可以有效减轻DP-SGD带来的性别偏见，促进更公平的文本生成。对于GPT-2微调，文献[168]探索了公平性与DP之间的相互影响。研究者发现，CDA方法结合DP能比单独使用DP提供更强的成员推断攻击防御能力。此外，与以往的分类任务不同，研究者未观察到DP对语言模型的公平性带来负面影响。当同时进行去偏和隐私保护时，CDA能减缓DP对语言建模能力的负面影响。文献[169]从群体公平性角度出发，提出在基于人口统计变量（如性别）定义的场景下，一个公平的模型应在不同子群体（如男性和女性）中实现相同或相近的性能。该文献提出的DPNR方法能够缓解模型对特定人口统计和身份属性的偏见，实现更公平的决策。此外，针对DP微调模型的隐私审计也成为差分隐私领域关注的问题，研究者通过审计微调数据的隐私预算边界来验证模型是否满足声称的理论上差分隐私保证。为此，文献[170]在微调初期引入随机生成的非同分布（out-of-distribution，OOD）审计样本（即"金丝雀"样本），以模拟最坏情况并收紧隐私预算估计。文献[171]则通过热启动初始化策略提升了计算机视觉任务中的隐私审计精度。然而，目前针对LLMs微调模型的隐私精度与预期理论值仍有差距，主要归结于LLMs对于预训练样本的记忆更加坚实和丰富，微调数据量级较少、记忆较弱等原因。

6.3 针对非参数化的隐私推断

大语言模型的推断是指在模型完成训练或微调之后，利用其已有的知识和参数，通过输入提示（Prompt）或上下文信息生成相应输出的过程。推断阶段不涉及模型参数的更新，而是通过前向传播计算模型对给定输入的响应。上下文学习（in-context learning，ICL）作为一种无参数推断（Non parametric Inference）机制，能够引导大模型适应下游任务的方法，无需修改LLMs的预训练参数，因而具有效率高、灵活性强的优势。这种非参数化微调方式通常通过一系列提示或示范引导模型生成特定任务的输出。虽然ICL不涉及参数更新，但现有差分隐私上下文学（DP-ICL）工作通常仍遵循标准DP-SGD框架（例如添加高斯噪声、应用隐私组合定理等）。因此，本节纳入了相关研究，并根据LLMs不同可信程度，分为可信的LLMs场景和不可信的LLMs场景进行探论。

6.3.1 可信的LLMs场景

当LLMs作为模型开发者或可信服务提供商时，所使用的上下文学习内容需要向不可信用户公开，这就必须对输出结果应用隐私保护机制。因此，差分隐私上下文学习DP-ICL的核心挑战在于如何对模型的响应结果进行隐私化聚合，特别是在文本分类和语言生成等输出空间高维且多样化的任务中。

为防止外界攻击者推断某条数据是否存在于大模型训练集中，文献[172]提出SeqPATE教师-学生训练框架，其中对教师进行隐私数据添加噪声进行DP训练，而学生模型在给定公共前缀的预训练语言模型生成的伪句子上进行训练，学生受益于候选筛选和有效的知识蒸馏，并由汇总的教师产出分布进行监督并输出具有隐私保护的文本。此外，文献[173]提出了差分隐私聚合机制。对于文本分类任务，研究者提出了满足差分隐私的生成答案选择方法，包括以下步骤：(1)将隐私示例划分为不相交的子集；(2)将示例子集与查询进行配对；(3)使用大模型进行推理，收集分类预测结果或生成的文本输出；(4)对LLMs生成的答案以差分隐私方式进行聚合后，将最终结果返回给用户。对于语言生成任务，研究者提出了两种主要技术：一种是嵌入向量空间聚合（embedding space aggregation，ESA），将生成的文本投影到低维语义空间，并对聚合后的嵌入向量添加差分隐私噪声；另一种是关键词空间聚合（key space aggregation，KSA），通过识别生成输出

中频繁出现的关键词，并结合 Propose-Test-Release 机制[174]或联合指数机制(Joint Exponential Mechanism)[175]，以差分隐私的方式选择这些关键词，从而实现输出内容的隐私保护。

6.3.2 不可信的 LLMs 场景

当LLMs作为不可信服务提供商时，数据所有者可能担心发送至模型的提示泄露敏感信息，引发数据安全性问题。文献[176]提出了一种差分隐私提示生成方法，通过从原始隐私数据集中合成满足DP的few-shot示例供推理时ICL使用。为了一方面确保生成示例符合隐私数据分布，另一方面不引发原始分布数据的泄漏，研究者采用了差分隐私算法，如报告最大带噪值技术[2]来保持top-k选择的可用性，同时满足用户输入给LLMs的信息满足隐私保护需求。此外，研究者也探索了基于few-shot和zero-shot的模型生成，实验表明这些方法具有良好的实用潜力。文献[177]提出了两种DP Prompt学习范式，分别针对输入的软提示和离散提示进行隐私保护，预防LLMs窃取提示内的敏感信息。对于软提示，研究者使用DP-SGD训练软提示，通过将隐私化梯度更新限制在低维提示空间内，降低了计算开销。对于离散提示，受教师模型隐私聚合(private aggregation of teacher ensembles，PATE)[178]的启发，构建了多个LLMs组成的教师模型集合，通过噪声多数投票生成一个公共输出，并据此构建新的学生提示，以替换原始隐私信息的提示，实现隐私的知识迁移并加速了后续的推理。

7 挑战与展望

随着人们对个人数据隐私保护意识的不断提升以及各类数据合规政策的出台，个体隐私保护与模型效能之间的矛盾日益凸显，尤其在适应新场景训练过程中对高质量个性化服务的强烈需求下，如何在保障用户隐私的同时保持良好的模型性能成为当前研究的核心挑战。差分隐私随机梯度下降DP-SGD作为当前主流的隐私化优化算法之一，已被广泛应用于图像、文本、图结构数据等多种任务中。然而，受限于DP-SGD在高维模型下引入的显著性能损耗、剪裁偏差与噪声注入的复杂交互、对隐私预算高度敏感等问题，现有研究仍难以全面满足实际部署中的高性能与强隐私并重的需求。

尽管已有大量研究围绕DP-SGD的非凸优化收敛性改进、剪裁机制优化、合成数据融合辅助等方面取得了一定进展，但在处理现实中更复杂、更稀疏、更异质的数据场景(如非结构化数据、多模态数据及海量数据)时仍面临诸多挑战。目前，关于DP-SGD在实际部署中的稳定性、可扩展性与实用性尚未得到系统解决，限制了其在大规模深度学习系统中的广泛落地。因此，如何在保障差分隐私语义下，进一步提升DP-SGD的可用性与泛化性能，仍是当前隐私保护学习领域亟待攻克的重要方向。

7.1 差分隐私随机梯度下降优化的未来方向

差分隐私随机梯度下降作为当前最主流的隐私保护优化算法之一，在理论建模和实际部署中已取得广泛应用。然而，随着深度学习模型规模不断扩大、数据复杂性持续提升，以及对训练精度和隐私合规性的双重要求，DP-SGD在应用过程中逐渐暴露出若干关键瓶颈，亟待突破。在本节中，本文从三个未来方向切入，探讨之后DP-SGD优化的潜在研究路径：包括针对重尾分布下的理论与方法扩展、高维模型中隐私性与可用性的权衡机制以及基于合成数据辅助增强隐私学习性能。

7.1.1 针对重尾情形下的差分隐私理论和技术

在实际的大规模深度学习任务中，尤其是在复杂数据分析场景下，梯度往往表现出重尾分布(Heavy-tailed Distributions)的统计特性，即随机梯度偏离均值的幅度远超高斯分布的预期范围。这种重尾性源于样本分布的不均衡、模型复杂结构的非线性响应以及训练数据的内在异质性等多方面因素，使得经典差分隐私机制(如DP-SGD)面临显著挑战。传统DP-SGD的理论分析大多依赖于梯度二阶矩有界或次高斯假设，在此基础上利用马尔可夫不等式、Azuma不等式等工具建立噪声尺度与隐私预算之间的关系。然而在重尾情形下，梯度分布可能不具备有限方差，导致这些分析工具失效，无法给出有效的高概率收敛界或隐私损失界限。即使现有工作部分关注了有限高阶矩重尾下的理论研究，但是涉及的重尾假设仍不够通用、研究还不够成熟。另外，目前大部分重尾DP-SGD研究集中在理论方面，未能针对重尾实际场景和现实问题中提出体系化的解决方案。然而，在重尾假设下，小梯度被严重放大，导致训练震荡；大梯度被极度压缩，导致更新停滞，造成模型性能严重降低。根据目前研究发现，裁剪偏差受到数据潜在的分布几何形状影响，基于真实的重尾分布先验选取裁剪阈值能够提高DP-SGD的可用性。因此，从几何角度探索适用于重尾分布假设下的差分隐私理论和技术，是未来隐

私保护领域一个研究趋势。

7.1.2　基于合成数据辅助的隐私保护增强

在差分隐私训练中，合成数据辅助机制正成为缓解隐私学习可用性的重要方向。该机制的核心思想是：首先利用有限公开数据生成近似分布的合成样本，或者基于DP-SGD训练后的生成模型合成具有隐私保护的样本，其次凭借合成样本辅助增益模型训练，从而缓解纯隐私数据带来的精度损失。

具体而言，在视觉领域，研究者通常采用扩散模型（Diffusion Models）、生成对抗模型（generate adversarial networks，GANs）对公开数据建模，并生成与隐私数据同分布的合成样本，或者直接利用DP-SGD训练具有满足差分隐私定义的生成模型来合成样本。考虑到这些训练方式都需要消耗大量资源，也有部分工作可以通过免训练的有条件指导（Training-free Conditional Guidance），高效地生成合成图片。在自然语言处理领域，研究者可以利用与视觉领域相同的训练技巧获取合成数据。此外，利用语言生成模型采样或者提示进行有导向的生成，这种无参数化的生成方式能够提高合成效率，但可用性归结于提示词的质量。在后训练过程中，这些合成样本作为预训练或微调过程中的“非隐私代理梯度”。一方面，这类合成样本具有稳定梯度结构，能够在不消耗隐私预算的前提下引导优化过程另一方面，通过投影到合成样本张成的梯度子空间（如Subspace-based DP-SGD策略），可以有效减少高维噪声注入所带来的偏差，从而实现梯度估计稳定性提升和噪声注入维度压缩。

然而，基于合成数据优化隐私学习也面临若干挑战：其一，合成数据若过于拟合训练数据，可能引发间接隐私泄漏风险；其二，合成数据的分布一致性与泛化能力缺乏理论保障；其三，不同任务场景下合成数据的“可用性”标准差异较大，难以统一评价其对隐私优化过程的贡献。尽管存在这些局限，这一思路为在有限隐私预算下实现大模型高效训练提供了有前景的解决路径，尤其在医疗、金融等高敏感领域具备现实价值。同时，该任务需要进一步的理论分析与大规模实验验证，以构建面向实际部署场景的合成数据辅助DP优化框架。

7.2　面向新场景应用的差分隐私随机梯度下降

随着人工智能技术在实际应用场景中的不断拓展，差分隐私优化不再仅限于传统的小规模模型或单一模态任务，而是逐步向更复杂、更真实的新型场景迁移。在这些新场景中，大语言模型成为最具代表性的代表系统，其庞大的模型结构、复杂的训练机制以及高度集成的数据类型对现有差分隐私机制提出了前所未有的挑战。因而，探索面向新型应用场景的差分隐私随机梯度下降方法，不仅是差分隐私理论走向实用化的关键一步，也是推动隐私保护向真实部署落地演进的核心需求。

在大语言模型的部署与应用过程中，如何验证其是否满足差分隐私约束成为亟须解决的核心问题。LLMs通常涉及数十亿参数、复杂的训练范式以及不透明的梯度更新机制，并且大语言模型开发者通常更加倾向仅公开模型访问接口供用户对话使用，而非公布真实的模型参数及置信度概率。在此背景下审计者只能以完全黑盒方式下获取模型的生成内容，因此，传统的差分隐私审计技术（如黑盒成员推理攻击或白盒梯度重构分析），如图2所示，和以往针对小模型的审计方法都大幅依赖于对模型置信度概率的计算，难以直接适配仅得知生成内容的场景。当前主流研究多从经验审计与理论估计两个角度展开。前者主要通过成员推理攻击（membership inference attacks，MIAs）来评估模型在训练样本上的“记忆能力”，并以此反推出其可能的隐私泄露风险。研究者常构建影子模型以模拟目标模型训练过程，训练攻击者网络以判断某一样本是否属于训练集；若攻击成功率显著高于随机猜测水平，则可视为DP保障不足的证据。

而理论估计路径则聚焦于隐私预算追踪与验证机制，如引入隐私损失定义与μ-高斯差分隐私和(ϵ,δ)-差分隐私转化公式等工具对模型训练过程中的噪声注入进行量化建模，进而推导训练过程的实际隐私支出（经验隐私下界）。但是目前高精度的隐私方法依赖于获知模型输出的置信度向量，该假设在大模型场景中不一定总是成立，鉴于隐私安全考虑，主流大模型开发者更倾向于提供API接口，审计者只能够接触到模型反馈的对话与回答。这种情况下的审计难度将大幅提升，并仅依赖模型对话审计经验隐私下界目前缺乏深入的研究。此外，针对LLMs的隐私审计还面临模型参数量巨大，导致审计过程开销高、精度低的问题。整体来看，当前缺乏专门针对大模型架构与训练范式设计的系统性审计基准与指标。因此，在未来研究中，首先需要构建组织一套专用于LLMs的隐私审计基准评测集；其次，发展适用于更实际的黑盒场景（仅接触大模型对话与回复）的审计技术与分析方法，比如借鉴测试时扩

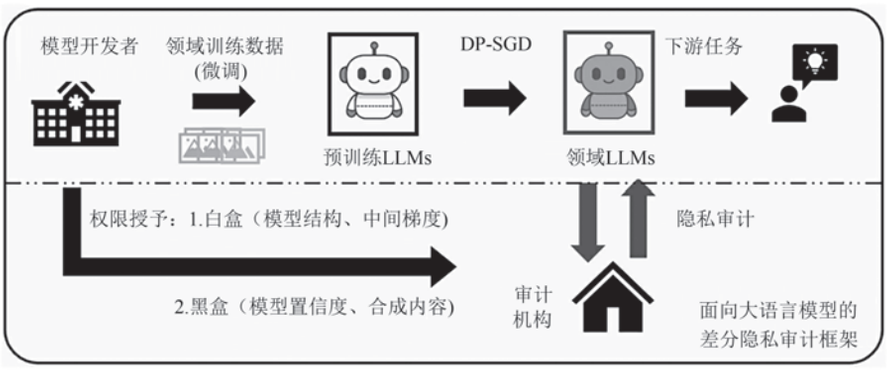

图2 面向大语言模型的差分隐私审计框架.

展(Test-time Scaling)对大语言模型的推理性能增强,引导大语言模型深度思考,从而根据回答出的更深层次敏感信息实现更为精准的隐私审计,以保证差分隐私的完整性和可验证性。

8 总 结

差分隐私随机梯度下降作为隐私保护机器学习的核心基石,近年来在理论与实践中都取得了显著进展。本文系统地回顾了DP-SGD在优化理论、算法机制与实际应用等方面的研究进展。在方法层面上,探讨了当前DP-SGD中所采用的高可用的噪声缓解技术、多元的采样与选择策略和改进的梯度裁剪机制等,并重点围绕高维噪声压缩、梯度裁剪偏差、合成数据增强等关键问题进行了深入剖析。在理论层面上,从噪声缓解、采样与选择和梯度裁剪三方面聚焦了重点的算法理论研究,包括维度无关理论、迭代选择理论及重尾稳定性等内容。同时,在应用层面,关注大语言模型的迅速发展,DP-SGD亟需在效用效率、灵活性与应用适应性方面作出突破。对于未来研究,本文针对差分隐私随机梯度下降优化未来方向,进一步讨论了重尾情形下的隐私优化技术与理论和合成数据辅助的隐私学习优化框架等备受关注的挑战;针对新应用场景,展望了面向大语言模型推理阶段的差分隐私审计。总之,DP-SGD作为当前最主流的隐私优化路径之一,在理论完善与应用实现上仍具有广阔的研究价值。

参 考 文 献

[1] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis// Proceedings of the Theory of Cryptography: Third Theory of Cryptography Conference. New York, USA, 2006: 265-284

[2] Dwork C, Roth A. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 2014, 9(3-4): 211-407

[3] Aerni M, Zhang J, Tramèr F. Evaluations of machine learning privacy defenses are misleading//Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. Salt Lake City, USA, 2024: 1271-1284

[4] Zhao Y, Chen J. A survey on differential privacy for unstructured data content. ACM Computing Surveys (CSUR), 2022, 54(10s): 1-28

[5] Fu J, Hong Y, Ling X, et al. Differentially private federated learning: A systematic review. arXiv preprint arXiv: 2405.08299, 2024

[6] Liu YX, Chen H, Liu YH, Li CP. Privacy-preserving techniques in federated learning. Journal of Software, 2022, 33(3): 1057-1092 (in Chinese)
(刘艺璇, 陈红, 刘宇涵, 李翠平. 联邦学习中的隐私保护技术. 软件学报, 2022, 33(3): 1057-1092)

[7] Gao Y, Chen XF. A survey of attack and defense techniques for federated learning systems. Chinese Journal of Computers, 2023, 46(9): 1781-1805 (in Chinese)
(高莹, 陈晓峰. 联邦学习系统攻击与防御技术研究综述. 计算机学报, 2023, 46(9): 1781-1805)

[8] Cummings R, Desfontaines D, Evans D, et al. Advancing differential privacy: Where we are now and future directions for real-world deployment. arXiv preprint arXiv:2304.06929, 2023

[9] Du H, Liu S, Zheng L, et al. Privacy in Fine-tuning Large Language Models: Attacks, Defenses, and Future Directions. arXiv preprint arXiv:2412.16504, 2024

[10] Hu L, Habernal I, Shen L, et al. Differentially private natural language models: recent advances and future directions// Findings of the Association for Computational Linguistics: EACL 2024. St.Julians, Malta, 2024: 478-499

[11] Li H, Chen Y, Luo J, et al. Privacy in large language models: Attacks, defenses and future directions. arXiv preprint arXiv: 2310.10383, 2023

[12] Das B C, Amini M H, Wu Y. Security and privacy challenges of large language models: A survey. ACM Computing Surveys, 2025, 57(6): 1-39

[13] Edemacu K, Wu X. Privacy preserving prompt engineering: A survey. arXiv preprint arXiv:2404.06001, 2024

[14] Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. Vienna, Austria, 2016: 308-318

[15] Mironov I, Talwar K, Zhang L. R′enyi differential privacy of the sampled Gaussian mechanism. arXiv preprint arXiv: 1908.10530, 2019

[16] Bun M, Steinke T. Concentrated differential privacy: Simplifications, extensions, and lower bounds//Proceedings of the Theory of cryptography conference. Berlin, Germany, 2016: 635-658

[17] Dong J, Roth A, Su W J. Gaussian differential privacy. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2022, 84(1): 3-37

[18] Balle B, Barthe G, Gaboardi M. Privacy amplification by subsampling: Tight analyses via couplings and divergences// Proceedings of the Advances in Neural Information Processing Systems. Montréal, Canada, 2018, 31

[19] Das R, Kale S, Xu Z, et al. Beyond uniform lipschitz condition in differentially private optimization// Proceedings of the International Conference on Machine Learning. New York, USA, 2023: 7066-7101

[20] Lowy A, Razaviyayn M. Private stochastic optimization with large worst-case lipschitz parameter: Optimal rates for (non-smooth) convex losses and extension to non-convex losses// Proceedings of the International Conference on Algorithmic Learning Theory. Honolulu, USA, 2023: 986-1054

[21] Kamath G, Liu X, Zhang H. Improved rates for differentially private stochastic convex optimization with heavy-tailed data// Proceedings of the International Conference on Machine Learning. Baltimore, USA, 2022: 10633-10660

[22] Bu Z, Wang Y X, Zha S, et al. Automatic clipping: Differentially private deep learning made easier and stronger// Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2023, 36: 41727-41764

[23] Yang X, Zhang H, Chen W, et al. Normalized/clipped sgd with perturbation for differentially private non-convex optimization. arXiv preprint arXiv:2206.13033, 2022

[24] Xia T, Shen S, Yao S, et al. Differentially private learning with per-sample adaptive clipping//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023, 37(9): 10444-10452

[25] Bassily R, Smith A, Thakurta A. Private empirical risk minimization: Efficient algorithms and tight error bounds// Proceedings of the 2014 IEEE 55th annual symposium on foundations of computer science. Philadelphia, USA, 2014: 464-473

[26] Li N, Qardaji W, Su D. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy//Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security. Raleigh, USA, 2012: 32-33

[27] Chaudhuri K, Mishra N. When random sampling preserves privacy// Proceedings of the Annual International Cryptology Conference. Berlin, Germany, 2006: 198-213

[28] Arora R, Bassily R, González T, et al. Faster rates of convergence to stationary points in differentially private optimization// Proceedings of the International Conference on Machine Learning. Honolulu, USA, 2023: 1060-1092

[29] Xiao H, Wan J, Devadas S. Geometry of sensitivity: twice sampling and hybrid clipping in differential privacy with optimal Gaussian noise and application to deep learning//Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. Copenhagen, Denmark, 2023: 2636-2650

[30] Dörmann F, Frisk O, Andersen L N, et al. Not all noise is accounted equally: How differentially private learning benefits from large sampling rates// Proceedings of the 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP). Virtual, 2021: 1-6

[31] Liu H, Li C, Liu B, et al. Differentially private learning with grouped gradient clipping//Proceedings of the 3rd ACM International Conference on Multimedia in Asia. Gold Coast, Australia, 2021: 1-7

[32] Xiao H, Xiang Z, Wang D, et al. A theory to instruct differentially-private learning via clipping bias reduction// Proceedings of the 2023 IEEE Symposium on Security and Privacy (SP). San Francisco, USA, 2023: 2170-2189

[33] Chua L, Ghazi B, Kamath P, et al. How Private are DP-SGD Implementations?//Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2024: 8904-8918

[34] Nguyen T N, Nguyen P H, Nguyen L M, et al. Batch clipping and adaptive layerwise clipping for differential private stochastic gradient descent. arXiv preprint arXiv:2307.11939, 2023

[35] Wei J, Bao E, Xiao X, et al. Dpis: An enhanced mechanism for differentially private SGD with importance sampling// Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. Los Angeles, USA, 2022: 2885-2899

[36] Boenisch F, Mühl C, Dziedzic A, et al. Have it your way: Individualized Privacy Assignment for DP-SGD// Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2023, 36: 19073-19103

[37] Liu J, Lou J, Xiong L, et al. Cross-silo federated learning with record-level personalized differential privacy//Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. Salt Lake City, USA, 2024: 303-317

[38] Chen X, Wu S Z, Hong M. Understanding gradient clipping in private sgd: A geometric perspective// Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2020, 33: 13773-13782

[39] Zhang X, Chen X, Hong M, et al. Understanding clipping for federated learning: Convergence and client-level differential privacy// Proceedings of the International Conference on Machine Learning, ICML 2022. Maryland, USA, 2022

[40] Sha H, Cao Y, Liu Y, et al. Clip body and tail separately: High

probability guarantees for DPSGD with heavy tails. arXiv preprint arXiv:2405.17529, 2024

[41] Wang D, Ye M, Xu J. Differentially private empirical risk minimization revisited: Faster and more general// Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017, 30

[42] Wang D, Chen C, Xu J. Differentially private empirical risk minimization with non-convex loss functions// Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 6526-6535

[43] Wang D, Xu J. Escaping saddle points of empirical risk privately and scalably via dp-trust region method// Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Ghent, Belgium, 2020: 90-106

[44] Lowy A, Ullman J, Wright S. How to Make the Gradients Small Privately: Improved Rates for Differentially Private Non-Convex Optimization//Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2024: 32904-32923

[45] Liu D, Ganesh A, Oh S, et al. Private (stochastic) non-convex optimization revisited: Second-order stationary points and excess risks//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA , 2023: 65618-65641

[46] Zhang Q, Ma J, Lou J, et al. Private stochastic non-convex optimization with improved utility rates//Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. Virtual, 2021

[47] Zhang J, Zheng K, Mou W, et al. Efficient private ERM for smooth objectives//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia , 2017: 3922-3928

[48] Su J, Hu L, Wang D. Faster rates of differentially private stochastic convex optimization. Journal of Machine Learning Research, 2024, 25(114): 1-41

[49] Asi H, Lévy D, Duchi J C. Adapting to function difficulty and growth conditions in private optimization// Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2021: 19069-19081

[50] Liu Z, Lou J, Bao W, et al. Differentially private zeroth-order methods for scalable large language model finetuning. arXiv preprint arXiv:2402.07818, 2024

[51] Fu J, Ye Q, Hu H, et al. DPSUR: Accelerating differentially private stochastic gradient descent using selective update and release. Proceedings of the VLDB Endowment, 2024, 17(6): 1200-1213

[52] Tong W, Niu J, Hua J, et al. Scalable differentially private model publishing Via private iterative sample selection. IEEE Transactions on Dependable and Secure Computing, 2023, 21(4): 2494-2506

[53] Liu W, Dai B, Humayun A, et al. Iterative machine teaching// Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2017: 2149-2158

[54] Liu W, Dai B, Li X, et al. Towards black-box iterative machine teaching//International Conference on Machine Learning. Stockholm, Sweden, 2018: 3141-3149

[55] Papernot N, Abadi M, Erlingsson Ú, et al. Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data// Proceedings of the International Conference on Learning Representations.Toulon, France, 2017: 1-16

[56] Feldman V, Mironov I, Talwar K, et al. Privacy amplification by iteration//Proceedings of the 2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS). Paris, France, 2018: 521-532

[57] Ye J, Shokri R. Differentially private learning needs hidden state (or much faster convergence)//Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2022, 35: 703-715

[58] Chourasia R, Ye J, Shokri R. Differential privacy dynamics of langevin diffusion and noisy gradient descent//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, LA, USA, 2021, 34: 14771-14781

[59] Yousefpour A, Shilov I, Sablayrolles A, et al. Opacus: User-friendly differential privacy library in PyTorch. arXiv preprint arXiv:2109.12298, 2021

[60] Feldman V, McMillan A, Talwar K. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling//Proceedings of the 2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS). Denver, USA,2022: 954-964

[61] Liu Y, Zhao S, Xiong L, et al. Echo of neighbors: Privacy amplification for personalized private federated learning with shuffle model//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023, 37(10): 11865-11872

[62] Mironov I, Talwar K, Zhang L. R $ backslash $′enyi differential privacy of the sampled Gaussian mechanism. arXiv preprint arXiv:1908.10530, 2019

[63] Wang Y X, Balle B, Kasiviswanathan S P. Subsampled rényi differential privacy and analytical moments accountant// Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics. Naha, Japan , 2019: 1226-1235

[64] Altschuler J, Talwar K. Privacy of noisy stochastic gradient descent: More iterations without more privacy loss//Advances in Neural Information Processing Systems. Virtual, 2022, 35: 3788-3800

[65] Feldman V, Koren T, Talwar K. Private stochastic convex optimization: optimal rates in linear time//Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing. Chicago, USA, 2020: 439-449

[66] Bassily R, Feldman V, Talwar K, et al. Private stochastic convex optimization with optimal rates//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2019, 32

[67] Bassily R, Feldman V, Guzmán C, et al. Stability of stochastic gradient descent on nonsmooth convex losses//Proceedings of the Advances in Neural Information Processing Systems.

Virtual, 2020, 33: 4381-4391

[68] Kulkarni J, Lee Y T, Liu D. Private non-smooth erm and sco in subquadratic steps// Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2021, 34: 4053-4064

[69] Bun M, Dwork C, Rothblum G N, et al. Composable and versatile privacy via truncated cdp//Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing. Los Angeles, USA, 2018: 74-86

[70] Ghadimi S, Lan G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. SIAM Journal on Optimization, 2012, 22(4): 1469-1492

[71] Fang C, Li C J, Lin Z, et al. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator// Proceedings of the Advances in Neural Information Processing Systems. Montréal, Canada, 2018, 31

[72] Hazan E, Kale S. An optimal algorithm for stochastic strongly-convex optimization. arXiv preprint arXiv:1006.2425, 2010

[73] Pichapati V, Suresh A T, Yu F X, et al. Adaclip: Adaptive clipping for private sgd. arXiv preprint arXiv:1908.07643, 2019

[74] Zhang X, Bu Z, Wu Z S, et al. Differentially private sgd without clipping bias: An error-feedback approach// Proceedings of the 12th International Conference on Learning Representations, ICLR. Vienna, Austria, 2024

[75] Andrew G, Thakkar O, McMahan B, et al. Differentially private learning with adaptive clipping//Advances in Neural Information Processing Systems. Virtual, 2021, 34: 17455-17466

[76] Xu D, Du W, Wu X. Removing disparate impact on model accuracy in differentially private stochastic gradient descent// Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 1924-1932

[77] Koloskova A, Hendrikx H, Stich S U. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees// International Conference on Machine Learning. Honolulu, USA, 2023: 17343-17363

[78] Chen D, Chua G A. Differentially private stochastic convex optimization under a quantile loss function// Proceedings of the International Conference on Machine Learning. Honolulu, Hawaii, USA, 2023: 4435-4461

[79] Wang D, Xiao H, Devadas S, et al. On differentially private stochastic convex optimization with heavy-tailed data// International Conference on Machine Learning. Virtual, 2020: 10081-10091

[80] Asi H, Liu D, Tian K. Private stochastic convex optimization with heavy tails: Near-optimality from simple reductions// Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2024, 37: 59174-59215

[81] Zhao P, Wu J, Liu Z, et al. Differential private stochastic optimization with heavy-tailed data: towards optimal rates// Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia, Pennyslvania, 2025, 39(21): 22795-22803

[82] Zhang J, He T, Sra S, et al. Why gradient clipping accelerates training: A theoretical justification for adaptivity//Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020: 1-21

[83] Zhang J, Karimireddy S P, Veit A, et al. Why are adaptive methods good for attention models?//Advances in Neural Information Processing Systems. Virtual, 2020, 33: 15383-15393

[84] Laakso T I, Hartimo I O. Noise reduction in recursive digital filters using high-order error feedback. IEEE Transactions on Signal Processing, 2002, 40(5): 1096-1107

[85] Mai V V, Johansson M. Stability and convergence of stochastic gradient clipping: Beyond lipschitz continuity and smoothness// Proceedings of the International Conference on Machine Learning. Virtual, 2021: 7325-7335

[86] Cutkosky A, Mehta H. Momentum improves normalized SGD// International conference on machine learning. Virtual, 2020: 2260-2268

[87] Sha H, Liu R, Liu Y, et al. PCDP-SGD: Improving the convergence of differentially private SGD via projection in advance. arXiv preprint arXiv:2312.03792, 2023

[88] Bakhshizadeh M, Maleki A, De La Pena V H. Sharp concentration results for heavy-tailed distributions. Information and Inference: A Journal of the IMA, 2023, 12(3): 1655-1685

[89] Song S, Steinke T, Thakkar O, et al. Evading the curse of dimensionality in unconstrained private glms//Proceedings of the International Conference on Artificial Intelligence and Statistics. Virtual, 2021: 2638-2646

[90] Hirschman I I, Widder D V. The Convolution Transform. Chelmsford: Courier Corporation, 2005.

[91] Wu L, Wang M, Su W. The alignment property of SGD noise and how it helps select flat minima: A stability analysis// Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022, 35: 4680-4693

[92] Gurbuzbalaban M, Simsekli U, Zhu L. The heavy-tail phenomenon in SGD// Proceedings of the International Conference on Machine Learning. Boulder, USA, 2021: 3964-3975

[93] Simsekli U, Sagun L, Gurbuzbalaban M. A tail-index analysis of stochastic gradient noise in deep neural networks// Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 5827-5837

[94] Gorbunov E, Danilova M, Gasnikov A. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping// Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2020, 33: 15042-15053

[95] Hu L, Ni S, Xiao H, et al. High dimensional differentially private stochastic optimization with heavy-tailed data// Proceedings of the 41st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems. Philadelphia, USA, 2022: 227-236

[96] Yu D, Zhang H, Chen W, et al. Do not let privacy overbill utility: Gradient embedding perturbation for private learning// Proceedings of the International Conference on Learning Representations. Vienna, Austria, 2021: 1-18

[97] Bu Z, Gopi S, Kulkarni J, et al. Fast and memory efficient

differentially private-sgd via jl projections// Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2021, 34: 19680-19691

[98] Zhu J, Blaschko M B. Improving differentially private sgd via randomly sparsified gradients. arXiv preprint arXiv:2112.00845, 2021

[99] Luo Z, Wu D J, Adeli E, et al. Scalable differential privacy with sparse network finetuning//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Virtual, 2021: 5059-5068

[100] Hölzl F A, Rueckert D, Kaissis G. Equivariant differentially private deep learning: Why DP-SGD needs sparser models// Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security. Taipei , China, 2023: 11-22

[101] Cheng A, Wang P, Zhang X S, et al. Differentially private federated learning with local regularization and sparsification// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. New Orleans, USA, 2022: 10122-10131

[102] Li Y, Du W, Han L, et al. A Communication-efficient, privacy-preserving federated learning algorithm based on two-stage gradient pruning and differentiated differential privacy. Sensors, 2023, 23(23): 9305

[103] Adamczewski K, He Y, Park M. Pre-Pruning and Gradient-Dropping Improve Differentially Private Image Classification. arXiv preprint arXiv:2306.11754, 2023

[104] Yu D, Zhang H, Chen W, et al. Large scale private learning via low-rank reparametrization// Proceedings of the International Conference on Machine Learning. Virtual, 2021: 12208-12218

[105] Mireshghallah F, Backurs A, Inan H A, et al. Differentially private model compression// Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022, 35: 29468-29483

[106] Tang X, Panda A, Sehwag V, et al. Differentially private image classification by learning priors from random processes// Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2023, 36: 35855-35877

[107] Li T, Zaheer M, Reddi S, et al. Private adaptive optimization with side information// Proceedings of the International Conference on Machine Learning. Baltimore, USA, 2022: 13086-13105

[108] Golatkar A, Achille A, Wang Y X, et al. Mixed differential privacy in computer vision//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 8376-8386

[109] Sha H, Wu Y, Liu R, et al. Differentially private visual learning with public subspace augmented by synthetic data// Proceedings of the 33rd ACM International Conference on Multimedia. Dublin,Ireland, 2025: 11357-11366

[110] Zhou Y, Wu Z S, Banerjee A. Bypassing the ambient dimension: Private sgd with gradient subspace identification// Proceedings of the International Conference on Learning Representations. Vienna, Austria, 2021: 1-28

[111] Liu J, Lou J, Xiong L, et al. Projected federated averaging with heterogeneous differential privacy. Proceedings of the VLDB Endowment, 2021, 15(4): 828-840

[112] Arora R, Bassily R, Guzmán C, et al. Differentially private generalized linear models revisited// Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2022, 35: 22505-22517

[113] Bassily R, Guzmán C, Nandi A. Non-euclidean differentially private stochastic convex optimization// Proceedings of the Conference on Learning Theory. Boulder, USA, 2021: 474-499

[114] Amid E, Ganesh A, Mathews R, et al. Public data-assisted mirror descent for private model training// Proceedings of the International Conference on Machine Learning. Baltimore, Maryland, 2022: 517-535

[115] Ganesh A, Haghifam M, Nasr M, et al. Why is public pretraining necessary for private model training?// Proceedings of the International Conference on Machine Learning. Honolulu, USA, 2023: 10611-10627

[116] Nasr M, Mahloujifar S, Tang X, et al. Effectively using public data in privacy preserving machine learning// Proceedings of the International Conference on Machine Learning. Honolulu, USA, 2023: 25718-25732

[117] Lowy A, Li Z, Huang T, et al. Optimal differentially private model training with public data//Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2024: 32849-32903

[118] Baradad Jurjo M, Wulff J, Wang T, et al. Learning to see by looking at noise//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2021, 34: 2556-2569

[119] Baradad M, Chen R, Wulff J, et al. Procedural image programs for representation learning// Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2022, 35: 6450-6462

[120] Park J, Kim H, Choi Y, et al. Differentially private sharpness-aware training// Proceedings of the International Conference on Machine Learning. Honolulu, USA, 2023: 27204-27224

[121] Liu Y, Xiong L, Liu Y, et al. DPDR: Gradient decomposition and reconstruction for differentially private deep learning. arXiv preprint arXiv:2406.02744, 2024

[122] Yun S, Park J, Lee K, et al. Regularizing class-wise predictions via self-knowledge distillation//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle, USA, 2020: 13876-13885

[123] Yang Z, Zeng A, Li Z, et al. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels//Proceedings of the IEEE/CVF International Conference on Computer Vision. Vancouver, Canada, 2023: 17185-17194

[124] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015

[125] Liu R, Bu Z, Wang Y, et al. Coupling public and private gradient provably helps optimization. arXiv preprint arXiv: 2310.01304, 2023

[126] Park J, Choi Y, Lee J. In-distribution public data synthesis with diffusion models for differentially private image classification// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 12236-12246

[127] Karras T, Aittala M, Aila T, et al. Elucidating the design space of diffusion-based generative models// Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2022, 35: 26565-26577

[128] Kim D, Kim Y, Kwon S J, et al. Refining generative process with discriminator guidance in score-based diffusion models// Proceedings of the International Conference on Machine Learning. Honolulu, USA, 2023: 16567-16598

[129] Wang Z, Pang T, Du C, et al. Better diffusion models further improve adversarial training// Proceedings of the International conference on machine learning. Honolulu, USA, 2023: 36246-36263

[130] Liang P P, Liu T, Ziyin L, et al. Think locally, act globally: Federated learning with local and global representations. arXiv preprint arXiv:2001.01523, 2020

[131] Song Z, Wang Y, Yu Z, et al. Sketching for first order method: efficient algorithm for low-bandwidth channel and vulnerability//International Conference on Machine Learning. Honolulu, USA, 2023: 32365-32417

[132] Jain P, Thakurta A G. (Near) dimension independent risk bounds for differentially private learning// Proceedings of the International Conference on Machine Learning. Beijing, China, 2014: 476-484

[133] Li X, Liu D, Hashimoto T B, et al. When does differentially private learning not suffer in high dimensions?// Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2022, 35: 28616-28630

[134] Tramèr F, Kamath G, Carlini N. Position: Considerations for differentially private learning with large-scale public pretraining//Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2024: 48453-48467

[135] Yu D, Backurs A, Gopi S, et al. Training private and efficient language models with synthetic data from llms// Proceedings of the Socially Responsible Language Modelling Research. SoLaR. New Orleans, USA, 2023: 1-12

[136] Ullah E, Menart M, Bassily R, et al. Public-data assisted private stochastic optimization: Power and limitations. arXiv preprint arXiv:2403.03856

[137] Gur-Ari G, Roberts D A, Dyer E. Gradient descent happens in a tiny subspace. arXiv preprint arXiv:1812.04754, 2018

[138] Li C, Farkhoor H, Liu R, et al. Measuring the intrinsic dimension of objective landscapes. arXiv preprint arXiv: 1804.08838, 2018

[139] Li X, Gu Q, Zhou Y, et al. Hessian based analysis of sgd for deep nets: Dynamics and generalization//Proceedings of the 2020 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, Cincinnati, USA, 2020: 190-198

[140] Li T, Tan L, Huang Z, et al. Low dimensional trajectory hypothesis is true: Dnns can be trained in tiny subspaces. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(3): 3411-3420

[141] Gu X, Kamath G, Wu Z S. Choosing public datasets for private machine learning via gradient subspace distance// Proceedings of the 2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). Copenhagen, Denmark, 2025: 879-900

[142] Kairouz P, Diaz M R, Rush K, et al. (Nearly) Dimension independent private ERM with adaGrad rates\{via Publicly Estimated Subspaces//Proceedings of the Conference on Learning Theory. Boulder, USA, 2021: 2717-2746

[143] Talagrand M. Upper and lower bounds for stochastic processes. Berlin: Springer, 2014

[144] Anil R, Ghazi B, Gupta V, et al. Large-scale differentially private BERT//Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates, 2022: 6481-6491

[145] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding// Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). Minneapolis, USA, 2019: 4171-4186

[146] Li X, Tramer F, Liang P, et al. Large language models can be strong differentially private learners//Proceedings of the International Conference on Learning Representations. Virtual, 2022: 1-30

[147] Bu Z, Zhang X, Zha S, et al. Pre-training differentially private models with limited public data// Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2024, 37: 94652-94683

[148] Yu D, Naik S, Backurs A, et al. Differentially private fine-tuning of language models. Journal of Privacy and Confidentiality, 2024, 14(2): 1-26

[149] Williams A, Nangia N, Bowman S. A broad-coverage challenge corpus for sentence understanding through inference// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, USA, 2018: 1112-1122

[150] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv: 1907.11692, 2019

[151] Novikova J, Dušek O, Rieser V. The E2E Dataset: New Challenges For End-to-End Generation//Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. Saarbrücken, Germany, 2017: 201-206

[152] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners//Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2020, 33: 1877-1901

[153] Hu E J, Wallis P, Allen-Zhu Z, et al. LoRA: Low-rank adaptation of large language models//Proceedings of the International Conference on Learning Representations. Virtual, 2022: 1-20

[154] Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP// Proceedings of the International conference on machine learning. Long Beach, USA, 2019: 2790-2799

[155] Karimi Mahabadi R, Henderson J, Ruder S. Compacter: Efficient low-rank hypercomplex adapter layers// Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2021, 34: 1022-1035

[156] Yue X, Inan H, Li X, et al. Synthetic text generation with differential privacy: A simple and practical recipe//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada, 2023: 1321-1342

[157] Senge M, Igamberdiev T, Habernal I. One size does not fit all: Investigating strategies for differentially-private learning across NLP tasks//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates, 2022: 7340-7353

[158] Ghazi B, Huang Y, Kamath P, et al. Sparsity-preserving differentially private training of large embedding models// Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2023, 36: 10951-10971

[159] Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Bangkok, Thailand, 2021: 4582-4597

[160] Shi W, Cui A, Li E, et al. Selective differential privacy for language modeling//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, USA, 2022: 2848-2859

[161] Shi W, Shea R, Chen S, et al. Just Fine-tune Twice: Selective differential privacy for large language Models//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates, 2022: 6327-6340

[162] Wang T, Zhai L, Yang T, et al. Selective privacy-preserving framework for large language models fine-tuning. Information Sciences, 2024, 678: 121000

[163] Du M, Yue X, Chow S S M, et al. Dp-forward: Fine-tuning and inference on language models with differential privacy in forward pass//Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. Copenhagen, Denmark, 2023: 2665-2679

[164] Ganev G, Xu K, De Cristofaro E. Understanding how differentially private generative models spend their privacy budget. arXiv e-prints, 2023: arXiv: 2305.10994

[165] Behnia R, Ebrahimi M R, Pacheco J, et al. Ew-tune: A framework for privately fine-tuning large language models with differential privacy//Proceedings of the 2022 IEEE International Conference on Data Mining Workshops (ICDMW). Orlando, USA, 2022: 560-566

[166] Wang H, Gao S, Zhang H, et al. Analytical composition of differential privacy via the edgeworth accountant. arXiv preprint arXiv:2206.04236, 2022

[167] Srivastava S, Mardziel P, Zhang Z, et al. De-amplifying bias from differential privacy in language model fine-tuning. arXiv preprint arXiv:2402.04489, 2024

[168] Matzken C, Eger S, Habernal I. Trade-offs between fairness and privacy in language modeling//Findings of the Association for Computational Linguistics: ACL 2023. Toronto, Canada, 2023: 6948-6969

[169] Lyu L, He X, Li Y. Differentially private representation for NLP: Formal guarantee and an empirical study on privacy and fairness//Findings of the Association for Computational Linguistics: EMNLP 2020. CanaPunta, RepublicDominican, 2020: 2355-2365

[170] Panda A, Tang X, Nasr M, et al. Privacy auditing of large language models. arXiv preprint arXiv:2503.06808, 2025

[171] Muthu Selva Annamalai M S, De Cristofaro E. Nearly tight black-box auditing of differentially private machine learning// Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2024, 37: 131482-131502

[172] Tian Z, Zhao Y, Huang Z, et al. Seqpate: Differentially private text generation via knowledge distillation// Proceedings of the Advances in Neural Information Processing Systems. 2022, 35: 11117-11130

[173] Wu T, Panda A, Wang J T, et al. Privacy-preserving in-context learning for large language models. arXiv preprint arXiv: 2305.01639, 2023

[174] Dwork C, Lei J. Differential privacy and robust statistics// Proceedings of the forty-first annual ACM symposium on Theory of computing. Bethesda, USA, 2009: 371-380

[175] Gillenwater J, Joseph M, Munoz A, et al. A joint exponential mechanism for differentially private top-k// Proceedings of the International Conference on Machine Learning. Baltimore, Maryland, 2022: 7570-7582

[176] Tang X, Shin R, Inan H A, et al. Privacy-preserving in-context learning with differentially private few-shot generation// Proceedings of the International Conference on Learning Representation. Vienna, Austria, 2024: 1-20

[177] Duan H, Dziedzic A, Papernot N, et al. Flocks of stochastic parrots: Differentially private prompt learning for large language models//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2023, 36: 76852-76871

[178] Papernot N, Song S, Mironov I, et al. Scalable private learning with pate. arXiv preprint arXiv:1802.08908, 2018

SHA Hai-Chao, Ph. D. candidate. His research interests include privacy preservation on machine learning and differential privacy

SUN Li-Chao, Ph. D. candidate. Her research interests include privacy attacks for generative models.

LIU Yi-Xuan, Ph. D. Her research interest is privacy preservation in machine learning.

XUE Da-Xuan, Ph. D. candidate. His research interest is privacy-preserving neural network architecture search.

WU Yun-Cheng, Ph. D., associate professor. His research interests include privacy preservation and federated computing.

LI Cui-Ping, Ph. D., professor. Her research interest is big data analysis and mining.

CHEN Hong, (corresponding author), Ph. D., professor. Her research interest is big data privacy preservation.

Background

This study falls within the field of privacy-preserving machine learning, specifically focusing on the optimization challenges of differentially private stochastic gradient descent (DP-SGD). As privacy concerns continue to escalate alongside the rapid deployment of machine learning systems. DP-SGD has emerged as the most widely adopted algorithmic framework for ensuring formal privacy guarantees in model training. Internationally, research on DP-SGD has advanced significantly in terms of theoretical convergence analysis, noise calibration, and privacy accounting. However, its practical deployment remains limited due to utility degradation under high-dimensional models and complex data distributions.

This paper presents a systematic survey for a comprehensive theoretical and algorithmic analysis of DP-SGD in both convex and non-convex settings. It systematically categorizes and refines optimization strategies along the key stages of sampling, gradient clipping, and noise injection. The paper further proposes a structured framework for understanding recent advances in DP-SGD, including sampling-based amplification techniques, geometry-aware clipping mechanisms, and subspace-guided noise reduction strategies. Additionally, it extends the analysis to large language models (LLMs), a newly emerging and high-impact application area, by discussing full-parameter training, parameter-efficient fine-tuning, and non-parametric inference under differential privacy constraints.

Our research team members have achieved notable progress in the field of differential privacy in recent three years, having published a number of papers in top-tier English CCF-A journals and conferences, including VLDB, ICDE, KDD, SP, AAAI, and TKDE, with five of them specifically addressing key problems in DP-SGD domain. In addition, the members have published two Chinese CCF-A papers in leading journals, including Chinese Journal of Computers and Journal of Software, focusing on differential privacy optimization in federated learning and privacy-preserving optimization on graph-structured data.

This work is supported by the the Joint Funds of the National Natural Science Foundation of China under Grant (U23A20299, U24B20144), Special Funds of the National Natural Science Foundation of China under Grant 62441230, National Natural Science Foundation of China under Grant (62172424, 62276270, 62322214).