arXiv:2302.03884v2 [cs.LG] 3 Jun 2023

# DIFF2：通过梯度差分实现差分隐私优化的非凸分布式学习

**Tomoya Murata** [1 2] **Taiji Suzuki** [2 3]


## 摘要

本文研究非凸光滑目标函数的差分隐私优化。在以往的工作中，就平方全梯度范数而言，已知的最佳效用界为 $\widetilde{O}(\sqrt{d}/(n\varepsilon_{\mathrm{DP}}))$；差分隐私梯度下降（DP-GD）是实现该界的一种方法，其中 $n$ 为样本量，$d$ 为问题维度，$\varepsilon_{\mathrm{DP}}$ 为差分隐私参数。为改进已知的最佳效用界，我们提出一种新的差分隐私优化框架，称为*DIFF2*（通过梯度差分实现的差分隐私优化）。该框架基于传输的梯度差分而非梯度本身，构造方差可能相当小的差分隐私全局梯度估计器。研究表明，采用梯度下降子程序的 DIFF2 可达到效用界 $\widetilde{O}(d^{2/3}/(n\varepsilon_{\mathrm{DP}})^{4/3})$；就对样本量的依赖关系而言，该界可能显著优于先前结果$n$。据我们所知，这是首个针对非凸目标函数、改进标准效用界 $\widetilde{O}(\sqrt{d}/(n\varepsilon_{\mathrm{DP}}))$ 的基础性结果。此外，我们还将计算和通信效率更高的子程序与 DIFF2 相结合，并给出了相应的理论分析。我们通过数值实验验证了 DIFF2 框架的优越性。


## 1. 引言

隐私保护已成为现代人工智能技术的重要组成部分之一。数据集由浏览历史、签到数据、医疗记录、金融记录等构成，通常包含私人信息；在私有数据集上训练得到的模型或计算出的梯度等统计信息一旦发布，就会导致私人信息泄露。在近期提出的联邦学习框架（Koneˇcnˇy 等人，2016；McMa- han 等人，2017年）中，每个客户端都拥有一个私有本地数据集，并与中央服务器迭代地交换自身模型或梯度，因此，防止恶意客户端窃取本地数据集所包含的私人信息至关重要。

实际上，已有许多针对训练好的模型或梯度窃取私人信息的攻击。主要攻击包括成员推断攻击（Shokri 等人，2017年；Yeom 等人，2018年；Nasr 等人，2019年）和重建攻击（Fredrikson 等人，2015年；Zhu 等人，2019年；Yang 等人，2019年；Wang 等人，2019d；Geiping 等人，2020年；Zhang 等人，2020b）。成员推断攻击旨在根据数据集的共享统计信息，推断已知个人或记录是否存在于数据集中。重建攻击旨在根据梯度等共享信息，重建受害用户或客户端的数据样本。鉴于这些攻击在实践中已取得成功，隐私保护技术已成为必需。

为了保护数据集的隐私，最简单的方法是数据匿名化。粗略地说，数据匿名化试图直接从数据集中移除个人身份信息。最主要的数据匿名化技术是 $k$-匿名化（Samarati 和 Sweeney，1998年）及其变体（Li 等人，2006；Machanavajjhala 等人，2007年）。$k$-匿名化通过抑制/泛化属性和/或添加虚拟记录，确保每条记录在称为准标识符的潜在识别属性上，至少与其他 $k-1$ 条记录相似。然而，由于存在利用辅助信息的去匿名化算法，$k$-匿名化不足以实现隐私保护。另一方面，$k$-匿名化的高级变体往往会损害数据集的效用，导致训练出的模型性能下降，尽管它提高了 $k$-匿名化的安全性。

另一种隐私保护方法是采用差分隐私技术。差分隐私是一种通用

*贡献相同 [1]日本东京 NTT DATA Mathematical Systems Inc. [2]日本东京东京大学信息科学与技术研究生院 [3]日本东京理化学研究所先进智能项目中心。通信作者：Tomoya Murata <murata@msi.co.jp>、Taiji Suzuki< taiji@mist.i.u-tokyo.ac.jp>。

论文集：$40^{th}$ 国际机器学习会议，美国夏威夷州檀香山。PMLR 202，2023年。版权所有 © 2023年，作者所有。

概念，用于量化隐私保护程度（Dwork 等，2006；Dwork 和 Naor，2010；Dwork 等，2010；Dwork，2011；Dwork 等，2014）。为保证机器学习算法的差分隐私，已有多种方法被提出。输出扰动和目标函数扰动是适用于非分布式环境的通用差分隐私方法。在分布式环境中，一种有效的差分隐私保障方法是梯度扰动，即在共享梯度前向其中添加一些噪声（Song 等，2013；Abadi 等，2016；McMahan 等，2018；Geyer 等，2017；Triastcyn 和 Faltings，2019）。例如，在差分隐私梯度下降（DP-GD）算法中，每次迭代都会向梯度添加独立的零均值高斯噪声，并使用得到的含噪梯度进行参数更新。

在差分隐私优化中，比较满足预设差分隐私级别的优化算法的优化准确度（称为效用）至关重要，因为效用与隐私级别之间通常存在权衡关系。近期的差分隐私优化研究不仅给出了差分隐私保证，还从理论上研究了优化算法的效用。

在先前的工作中，一般非凸差分隐私优化已知的最佳效用界为$\widetilde{O}(\sqrt{d}/(n\varepsilon_{\mathrm{DP}}))$[1]，其依据是平方全局梯度范数；DP-GD 是达到该界的一个实例，其中$n$ 表示样本量，$d$ 表示问题维度，$\varepsilon_{\mathrm{DP}}$ 表示差分隐私参数（Zhang 等人，2017年；Wang 等人，2017年）。关于能否改善已知最佳效用界对样本量 $n$的依赖关系（即样本效率），这是一个重要的开放问题。

## 主要贡献

我们提出通过梯度差分进行差分优化（*DIFF2*）框架，以改进非凸差分隐私优化的现有最佳效用界。DIFF2 的主要特点如下：

**算法特性**。 DIFF2 框架的主要算法特性包括：(i) 共享局部梯度差而非梯度本身，以降低差分隐私噪声大小；以及 (ii) 使用聚合梯度差分与先前差分隐私全局梯度估计器之和构造差分隐私全局梯度估计器。所得全局梯度估计量会被送入通用子程序，由其基于该估计量优化目标。

**理论特性**。 采用梯度下降子例程的 DIFF2 可达到效用 $\widetilde{O}(d^{2/3}/(n\varepsilon_{\mathrm{DP}})^{4/3})$，其效果可能 显著优于目前已知的、在样本量 $\widetilde{O}(\sqrt{d}/(n\varepsilon_{\mathrm{DP}}))$ 依赖性方面针对非凸目标函数的 DP-GD 最佳效用 $n$。我们的差分隐私分析依赖于以下事实：对于光滑损失函数，梯度差分的敏感度可能远小于梯度本身的敏感度。为推导出最小差分隐私噪声水平，我们主要采用 R´ényi 差分隐私（RDP）技术。在效用分析中，仔细 评估 *DIFF2* 生成的差分隐私全局梯度估计量的偏差和方差 并确定最优的重启间隔 (详情请参见第 4.2 小节）。我们还考虑了一种更高效的子例程，称为偏差-方差缩减局部随机梯度下降（BVR-L-SGD）例程，并证明采用 BVR-L-SGD 例程的 DIFF2 比采用梯度下降例程的 DIFF2 具有更高的计算效率和通信效率。

从理论角度来看，我们在表1中比较了非凸情形下若干差分隐私算法所能达到的最佳效用、随机梯度复杂度，以及达到该效用所需的通信复杂度。 $N$ 定义为最小值$_{p\in[P]}\, n_p P$。从该表可以看出，在依赖于 $N$的方面，DIFF2-BVR-L-SGD以最佳梯度复杂度和最佳通信复杂度实现了最佳效用。[2] [3]

## 相关工作

本文在此简要回顾相关工作。

凸优化。Chaudhuri 等人（2011）研究了输出扰动和目标函数扰动方法，并考察了它们在凸问题中的效用。随后，有人提出了梯度扰动方法（Song 等人，2013），并研究了其效用（Bassily 等人，2014；Wang 等人，2017；Bassily 等人，2019）。为保证 $(\varepsilon_{\mathrm{DP}}, \delta_{\mathrm{DP}})$-差分隐私，针对非-$\widetilde{\ }$强凸目标函数，得到的效用界以目标差距衡量为 $\widetilde{O}(\sqrt{d}/(n\varepsilon_{\mathrm{DP}}))$；针对强凸目标函数，则为 $O(d/(n^2\varepsilon_{\mathrm{DP}}^2))$。若干论文推导出了与这些上界相匹配的下界，并证明了 DP-(S)GD 在 $\ell_2$有界域（Bassily 等人，2014）、 $\ell_1$ 有界域（Asi 等人，2021）和无约束域（Liu 和 Lu，2021）上的最优性。

[1] $\widetilde{O}$、 $\widetilde{\Theta}$ 和 $\widetilde{\ }\Omega$符号中省略了一个依赖于 $1/\delta_{\mathrm{DP}}$ 的额外多对数因子，以便简洁呈现。

[2] 尽管 DIFF2-BVR-L-SGD 的梯度复杂度有些复杂，但我们可以确认，其对 $N$ 的依赖严格小于 $N^2$。[3] 需要注意的是，要实现更高的效用 $\varepsilon_{\mathrm{opt}}$，通常需要更高的梯度复杂度和通信复杂度。因此，将 SRN-SGD、DP-SRM 和 DIFF2-GD 与 DP-GD 及差分隐私随机梯度下降进行比较并不公平，因为这会偏向 DP-GD 和差分隐私随机梯度下降。不过值得注意的是，DIFF2-BVR-L-SGD 仍然比包括差分隐私随机梯度下降在内的这些方法实现了更优的梯度复杂度和通信复杂度。

**DIFF2：通过梯度差分实现差分隐私优化**

| 算法 | 效用 | 梯度复杂度 | 通信复杂度 |
|---|---|---|---|
| DP-GD | $\frac{\sqrt{d}}{N\varepsilon_{\mathrm{DP}}}$ | $N+\frac{N^2\varepsilon_{\mathrm{DP}}}{\sqrt{d}}$ | 梯度复杂度 |
| 差分隐私随机梯度下降 | $\frac{\sqrt{d}}{N\varepsilon_{\mathrm{DP}}}$ | $\frac{N^2\varepsilon^2_{\text{差分隐私}}}{d}$ | $\frac{N^2\varepsilon^2_{\text{差分隐私}}}{1+\frac{\text{隐私}}{bd}}$ |
| SRN-SGD (Tran 和 Cutkosky, 2022年) | $\frac{d^{\frac{2}{3}}}{(N\varepsilon_{\mathrm{DP}})^{\frac{4}{3}}}+\frac{1}{N}$ | $\frac{N^3\varepsilon^{\frac{7}{3}}_{\text{差分隐私}}\varepsilon^{\frac{4}{3}}_{\text{隐私}}}{d^{\frac{2}{3}}}$ | $\frac{N^3\varepsilon^{\frac{7}{3}}_{\text{差分隐私}}\varepsilon^{\frac{4}{3}}_{\text{隐私}}}{d^{\frac{2}{3}}}4$ |
| 差分隐私随机响应机制 (Wang 等人,2019b) | $\frac{d^{\frac{2}{3}}}{(N\varepsilon_{\mathrm{DP}})^{\frac{4}{3}}}$ | $N+\frac{N^3\varepsilon^{\frac{7}{3}}_{\text{差分隐私}}\varepsilon^{\frac{4}{3}}_{\text{隐私}}}{d^{\frac{2}{3}}}$ | $1+\frac{(N\varepsilon_{\mathrm{DP}})^{\frac{4}{3}}}{d^{\frac{2}{3}}}$ |
| DIFF2-GD (本研究) | $\frac{d^{\frac{2}{3}}}{(N\varepsilon_{\mathrm{DP}})^{\frac{4}{3}}}$ | $N+\frac{N^3\varepsilon^{\frac{7}{3}}_{\text{差分隐私}}\varepsilon^{\frac{4}{3}}_{\text{隐私}}}{d^{\frac{2}{3}}}$ | 梯度复杂度 |
| DIFF2-BVR-L-SGD (本研究) | $\frac{d^{\frac{2}{3}}}{(N\varepsilon_{\mathrm{DP}})^{\frac{4}{3}}}$ | $N+\frac{N^{\frac{11}{6}}\varepsilon^{\frac{4}{3}}_{\mathrm{DP}}}{d^{\frac{1}{3}}}+\frac{N^{\frac{19}{12}}\varepsilon^{\frac{5}{6}}_{\mathrm{DP}}d^{\frac{2}{3}}}{d^{\frac{1}{6}}}+\frac{N^3P_{\varepsilon^3_{\text{差分隐私}}}}{d^{\frac{1}{3}}}$ | $\frac{\text{梯度复杂度}}{N}+\frac{\zeta_2(N\varepsilon_{\mathrm{DP}})^{\frac{4}{3}}}{d^{\frac{2}{3}}}$ |
| $(K=b=\sqrt{N})$ | | | |

*表1。* 非凸目标函数的 $(\varepsilon_{\mathrm{DP}},\ \delta_{\mathrm{DP}})$-差分隐私保证优化算法的效用和随机梯度复杂度的阶比较。“效用”指以梯度范数平方 $\mathbb{E}\|\nabla f(x_{\mathrm{out}})\|^2$表示的可实现的最佳优化误差。“梯度复杂度”指达到最佳效用所需的单次随机梯度计算次数。“通信复杂度”指达到最佳效用所需的通信轮数。 $\zeta_2=O(1)$表示本地数据集的海森异质性。 $N$定义为 $\min_{p\in[P]} n_p P$。 $d$表示参数空间的维度。 Tran 和 Cutkosky （2022年）和 Wang 等人 （2019b）（2023 年更新版本）是与本研究独立开展的工作。

非凸优化。Zhang 等人（2017）证明，对于非凸光滑目标函数，DP-SGD 在梯度范数平方（即一阶最优性）方面的效用为 $\widetilde{O}(\sqrt{d}/(n\varepsilon_{\mathrm{DP}}))$。研究了差分隐私 RMSprop 和差分隐私 Adam，并证明其效用界与 DP-GD 基本相同（Zhou 等人，2020）。Wang 等人（2017）对凸和非凸情形下的 DP-GD 与 DP-SVRG 进行了统一分析；特别是，对于满足 Polyak-Łojasiewicz（PL）条件的目标函数，证明了 DP-GD 在目标差距方面的效用为 $\widetilde{O}(d/(n^2\varepsilon^2_{\mathrm{DP}}))$。Wang 等人（2019a）基于梯度朗之万动力学理论，在不要求满足 PL 条件的情况下，给出了 DP-GD 的新效用界 $\widetilde{O}(1/(n\varepsilon^2_{\mathrm{DP}})+d/n^{\Omega(1)})$，以目标差距衡量。此外，Wang 等人（2019a）还提出了一种 DP-GD 变体，用于寻找二阶最优点，其效用与一阶最优性情形相同。

**计算与通信效率。** 尽管计算与通信效率并非本研究的重点，但许多研究已针对实际应用考察了这些效率。差分隐私随机梯度下降是降低差分隐私梯度下降高昂的每次更新计算成本的理想候选方法。为进一步降低计算成本，已有研究考察了应用方差缩减技术（Johnson &Zhang，2013年；Defazio 等人，2014；Nguyen 等人，2017年；Cutkosky& Orabona，2019年）（Wang 等人，2019b；Bassily 等人，2019年；Asi 等人，2021年）。此外，Kuru 等人（2022年）将差分隐私梯度下降与 Nesterov 加速相结合，以提高计算效率。另一方面，已有多项研究在分布式学习中开发了比差分隐私梯度下降通信效率更高的差分隐私算法。Noble 等人 （2022年）表明，与差分隐私随机梯度下降相比，他们提出的差分隐私联邦平均和差分隐私 SCAFFOLD 在通信轮数方面实现了更高的通信效率，同时保持了差分隐私随机梯度下降的效用。在保持差分隐私随机梯度下降效用的同时，通过梯度压缩降低每次更新的通信成本，也已由多篇论文在差分隐私优化设置中进行了研究（Agarwal 等人，2018年；Zhang 等人，2020a；Ding 等人，2021年；Li 等人，2022年）。

**效用改进。** 已有若干研究尝试从理论和实践两方面改进差分隐私随机梯度下降的效用。 Chourasia 等人 （2021年）基于强凸目标函数的梯度朗之万动力学理论，证明了差分隐私梯度下降中隐私损失的收敛性，并指出先前依赖组合定理的差分隐私分析可能较为宽松。然而，所得效用与已知的最佳结果相同。 Du 等人 （2021年）考虑了在差分隐私随机梯度下降中对差分隐私噪声大小和梯度裁剪半径进行动态调度，以提升效用，并通过实验证明了其有效性。 Wang 等人 （2020a）提出将差分隐私随机梯度下降与拉普拉斯平滑技术相结合，以降低差分隐私噪声，并报告了其实证性能优势。该方法进一步扩展到了联邦学习场景（Liang 等人，2020年）。最近，两项独立研究（Tran 和 Cutkosky，2022年；Wang 等人，2019b）[5] 基于与我们类似的思路，改进了非分布式学习场景下的效用界。

**问题设定的扩展。** Wang 等人（2020b）；Hu 等人（2022年）；Kamath 等人（2022年）尝试将差分隐私保证分析中通常采用的梯度一致有界假设替换为梯度分布的 $k$阶矩有界性，以处理凸优化中的重尾梯度。Das 等人（2022年）考虑将梯度一致有界假设放宽为样本依赖的有界性假设。

[5]Wang 等人(2019b) 已于2023年更新。

Han 等人（2022年）将差分隐私随机梯度下降扩展到了黎曼优化情形。已有多项工作针对集中式乃至去中心化分布式学习，发展了交替方向乘子法（交替方向乘子法）的差分隐私版本（Huang 等人，2019年；Huang & Gong，2020年；Shang 等人，2021年）。Yang 等人（2021年）为成对学习建立了差分隐私随机梯度下降理论。

## 2. 问题设定与假设

本节首先介绍本文中使用的符号。随后说明问题设定，并给出分析中采用的理论假设。

**符号说明**。 $\|\cdot\|$ 表示欧几里得 $L_2$ 范数 $\|\cdot\|_2$： $\|x\| = \sqrt{\sum_i x_i^2}$，其中向量为 $x$。对于矩阵 $X$， $\|X\|$ 表示由欧几里得 $L_2$ 范数诱导的范数。对于自然数 $m$， $[m]$ 表示集合 $\{1, 2, \ldots, m\}$。对于集合 $A$，$\#A$ 表示元素个数，该数可能为 $\infty$。对于任意数 $a$， $b$， $a \vee b$表示 $\max\{a, b\}$，而 $a \wedge b$表示 $\min\{a, b\}$。对于集合 $A$，我们用 $\mathrm{Unif}(A)$ 表示 $A$上的均匀分布。

### 2.1. 问题设定

我们希望在集中式分布式学习设置中，为每个客户端 $p \in [P]$ 寻找目标函数 $f(x) := (1/P)\sum_{p=1}^{P} f_p(x)$ 的一个近似极小化解，并提供差分隐私保证；其中， $f_p(x) := (1/n_p)\sum_{i=1}^{n_p} \ell(x, z_i^{(p)})$ 是客户端 $p$ 在私有数据集 $D_p$ 上的风险。

**威胁模型**。 本文聚焦于中心化差分隐- 私。中心化差分隐私假设存在可信中央服务器和诚实但好奇的客户端。在这种情况下，当服务器汇总从客户端接收的向量并将结果发送给每个客户端时，我们只需关注聚合向量可能导致的客户端信息泄露，而非客户端发送的单个向量。

接下来，需要保证中央服务器输出 $\mathcal{M}(D_1, \ldots, D_P)$ 相对于每个本地数据集 $D_p$ 满足记录级 $(\varepsilon, \delta)$-差分隐私[6]，即 $\mathbb{P}(\mathcal{M}(\mathcal{D}) \in S) \le e^{\varepsilon}\mathbb{P}(\mathcal{M}(\mathcal{D}') \in S) + \delta$；其中，对于任意 $S$ 以及任意相邻本地数据集 $D_p$ 和 $D_p'$， $\mathcal{D} := (D_1, \ldots, D_p, \ldots, D_P)$ 且 $\mathcal{D}' := (D_1, \ldots, D_p', \ldots, D_P)$。这里，如果 $d_{\mathrm{H}}(D, D') = 1$，我们就称数据集 $D = \{z_i\}_{i=1}^n$ 和 $D' = \{z_i'\}_{i=1}^n$ 相邻；其中， $d_{\mathrm{H}}$ 是 $D$ 和 $D'$ 之间的汉明距离，定义为 $d_{\mathrm{H}}(D, D') := \sum_{i=1}^n \mathbb{1}_{x_i \neq x_i'}$。

当中央服务器不可信时，可以采用一些安全聚合技术，例如多方计算（MPC），以确保与可信服务器情形相同的隐私级别。采用安全聚合技术后，服务器（以及客户端）只能访问聚合结果，而不能访问客户端发送的个体向量。

**理论性能度量**。 在本文中，我们关注寻找近似一阶平稳点，因为由于 $f$的非凸性质，通常很难找到 $f$ 的近似全局极小值，甚至局部极小值。然后，给定隐私参数 $\varepsilon_{\mathrm{DP}}$ 和 $\delta_{\mathrm{DP}}$，我们度量一个（$\varepsilon_{\mathrm{DP}}$， $\delta_{\mathrm{DP}}$）-差分隐私算法输出 $x$ 的$\|\nabla f(x)\|^2$。更小的 $\|\nabla f(x)\|^2$ 表示更高的效用。

### 2.2. 理论假设

这里描述了我们分析中所使用的理论假设。

**假设 2.1** （光滑性）。 对于每个 $z \in \cup_{p=1}^P \mathrm{supp}(D_p)$， $\ell(\cdot, z)$是 $L$-光滑的，即：

$$\|\nabla \ell(x, z) - \nabla \ell(y, z)\| \le L\|x - y\|, \forall x, y \in \mathbb{R}^d.$$

我们假设损失 $\ell$ 具有 $L$-光滑性，而不是风险 $f$具有该性质，这对DIFF2框架至关重要。

**假设 2.2** （全局最优解的存在性）。 $f$ 存在一个全局极小化点 $x_* \in \mathbb{R}^d$。

**假设 2.3** （梯度有界性）。 对于每个 $z \in \cup_{p=1}^P \mathrm{supp}(D_p)$， $\ell(\cdot, z)$ 都是 $G$-梯度有界的，即

$$\|\nabla \ell(x, z)\| \le G, \forall x \in \mathbb{R}^d.$$

假设 2.3 略显严格，尤其是在联邦学习中，但这对于我们的效用分析是必要的，也是以往差分隐私优化文献中的标准假设。

## 3. 所提框架：DIFF2

本节将介绍我们所提框架的具体流程。

**核心思想**。DIFF2 的主要过程是基于局部梯度差构造差分隐私全局梯度估计器 $\widetilde{v}_r$。回顾一下，标准的差分隐私算法只是对 $\nabla f(x_{r-1})$ 应用高斯机制，即 $\widetilde{v}_r := \nabla f(x_{r-1}) + \xi_r$，其中 $\xi_r$ 是均值为零的高斯噪声。为满足差分隐私保证， $\xi_r$ 的标准差应与 $\nabla f(x_{r-1})$ 的敏感度成正比。对于 $G$-Lipschitz 损失函数 $\ell$，该敏感度有界于 $2G/(n_{\min}P)$。相比之下，DIFF2 使用一种近似

$$\begin{aligned}\nabla f(x_{r-1}) &= \nabla f(x_{r-1}) - \nabla f(x_{r-2}) + \nabla f(x_{r-2}) \\ &\approx \nabla f(x_{r-1}) - \nabla f(x_{r-2}) + \widetilde{v}_{r-1}\end{aligned}$$

[6]我们关注记录级差分隐私。将我们的分析直接扩展到客户端级差分隐私的情形是很简单的。

**算法 1** DIFF2($x_0$, $\sigma_1$, $\sigma_2$, $C_1$, $C_2$, $T$, $R$, $*$参数)

**# 假设中央服务器是可信的。**
**对于** $r = 1$到 $R$ **执行**
  **对于** 客户端 $p \in [P]$ 并行 **执行**
    **如果** $(r-1)\%T = 0$ **则**
      $d_r^{(p)}$ = 裁剪均值($\{\nabla \ell(x_{r-1}, z)\}_{z \in D_p}$, $C_1$)。
      将 $d_r^{(p)}$发送到中央服务器。
    **else**
      发送 $d_r^{(p)}$ = 裁剪均值($\{\nabla \ell(x_{r-1}, z) - \nabla \ell(x_{r-2}, z)\}_{z \in D_p}$, $C_{2,r}$) to the 中央服务器，其中 $C_{2,r} = C_2 \|x_{r-1} - x_{r-2}\|$。
    **结束 if**
  **结束 for**
  **如果** $(r-1)\%T = 0$，**则**
    设置 $\sigma = \sigma_1$, $C = C_1$ 和 $\widetilde{v}_{r-1} = 0$.
  **else**
    设置 $\sigma = \sigma_2$ 和 $C = C_{2,r}$.
  **结束条件判断**
  设置 $v_r = \frac{1}{P}\sum_{p=1}^{P} d_r^{(p)} + \widetilde{v}_{r-1}$ 和 $\widetilde{v}_r = v_r + \xi_r$, 其中 $\xi_r \sim N(0, \sigma^2 C^2 I)$。
  $x_r$, $x_r^{\text{out}}$ = 通用例程($x_{r-1}$, $\widetilde{v}_r$, $*$args)
  **对于** $p \in [P]$ 并行**执行**
    从中央服务器接收 $x_r$。
  **结束循环**
**结束循环**
**返回：** $x^{\text{out}} = x_{\hat{r}-1}^{\text{out}}$ ($\hat{r} \sim$均匀分布$[R]$)。

**算法 2** 裁剪均值($\{x_i\}_{i \in I}$, $C$)

$\hat{x}_i = \min\left\{\frac{C}{\|x_i\|}, 1\right\} x_i$ 对于 $i \in I$。
**返回：** $\frac{1}{|I|}\sum_{i \in I} \hat{x}_i$。

**算法 3** GD-Routine($x_{r-1}$, $\widetilde{v}_r$, $\eta$)

1: 设置 $x_r = x_{r-1} - \eta \widetilde{v}_r$。
2: **返回：** $x_r$, $x_r$.

并按如下方式构造新的估计器 $\widetilde{v}_r$：

$$\widetilde{v}_r := (\nabla f(x_{r-1}) - \nabla f(x_{r-2}) + \xi_r) + \widetilde{v}_{r-1},$$

其中 $\widetilde{v}_1 := \nabla f(x_0) + \xi_1$。然后，我们观察到，$\xi_r$ 的标准差取决于梯度差分 $\nabla f(x_{r-1}) - \nabla f(x_{r-2})$ 的敏感度；对于 $L$-光滑损失函数 $\ell$，该敏感度有界于 $L\|x_{r-1} - x_{r-2}\|$。因此，如果 $x_{r-1}$ 和 $x_{r-2}$ 接近，梯度差分的敏感度可能小于梯度本身的敏感度 $\nabla f(x_{r-1})$，并且 DIFF2 在第 $r$ 轮的噪声规模会小于标准噪声规模。这就是 DIFF2 的机制，也解释了为什么 DIFF2 的梯度估计器可能比标准梯度估计器具有更低的方差。

**新框架：DIFF2**。我们提出的框架 DIFF2 的具体流程见算法 1。

如上所述，DIFF2 构造了差分隐私全局梯度估计器 $\widetilde{v}_r$。随后，将该估计器输入通用子程序，该子程序基于 $\widetilde{v}_r$ 执行某种优化（可能是局部优化）（第 17 行）。

现在，我们来看看构造 $\widetilde{v}_r$ 的具体过程。在初始轮次，每个工作节点计算局部梯度，并使用裁剪均值（算法 2）确保局部梯度有界（第 5 行），这通常用于以往的差分隐私保证算法。中央服务器聚合裁剪后的局部梯度，并向聚合梯度添加均值为零、方差为 $\sigma_1^2 C_1^2$ 的高斯噪声，其中 $C_1$ 是局部梯度的裁剪半径，从而生成初始的差分隐私全局梯度估计器 $\widetilde{v}_1$。在后续轮次中，服务器聚合裁剪后的局部梯度差分$\{\nabla \ell(x_{r-1}, z) - \nabla \ell(x_{r-2}, z)\}_{z \in D_p}$，而非局部梯度 $\{\nabla \ell(x_{r-1}, z)\}_{z \in D_p}$（第 8 行）。随后，全局梯度估计量 $v_r$ 更新为聚合梯度差分与前一轮估计器 $\widetilde{v}_{r-1}$ 之和（第 16 行）。最后，通过添加均值为零、方差为 $\sigma_2^2 C_{2,r}$ 的高斯噪声，得到差分隐私全局梯度估计器$\widetilde{v}_r$，其中 $C_{2,r}$ 是局部梯度差分的裁剪半径。

重要的是，每经过 $T$ 轮次，我们都会 重新启动 这些过程；我们将 $\widetilde{v}_r$ 重置为标准差分隐私全局梯度估计量$\nabla f(x_{r-1}) + \xi_r$，与初始轮次相同。必须重新启动的原因将在第4节中解释。

**具体算法：DIFF2-GD**。给定全局梯度估计量 $\widetilde{v}_r$，通用子程序最简单的选择是基于 $\widetilde{v}_r$执行梯度下降步。将梯度下降例程（GD-Routine，算法 3）应用于 DIFF2，可得到名为 DIFF2-GD 的新算法。DIFF2-GD 在每一轮仅执行一次梯度下降步。注意，当 $T = 1$时，DIFF2-GD 与标准差分隐私梯度下降相同。

## 4. 差分隐私分析与效用分析

本节对 DIFF2-GD 进行理论分析。我们的分析主要分为两部分：差分隐私保证分析（第4.1节）和效用分析（第4.2节）。

### 4.1. DIFF2-GD 的差分隐私保证分析

在本小节中，我们研究保证满足（$\varepsilon_{\text{DP}}$, $\delta_{\text{DP}}$）-差分隐私所需的最小噪声水平。证明见补充材料第C 节。

得益于梯度差分 $\{\nabla \ell(x_{r-1}, z) - \nabla \ell(x_{r-2}, z)\}_{z \in D_p}$ 的使用，我们可以大幅降低差分隐私

噪声大小，相较于 DP-GD 中每一轮都使用梯度本身而言。这是因为，对于光滑损失函数，梯度差分的敏感度可能远小于梯度本身 $\ell$。这是 $L2$梯度差分的敏感度可以界定为 $L\|x_{r-1}-x_{r-2}\|$；当$x_{r-1}$ 和 $x_{r-2}$ 足够接近时，例如它们处于驻点附近时，利用 $L$的光滑性， $\ell$ 该值 $\|\nabla\ell(x_{r-1},\ z)-\nabla\ell(x_{r-2},\ z)\| \le L\|x_{r-1}-x_{r-2}\|$。相比之下，梯度本身具有较大的 $L2$敏感度 $G$。为推导差分隐私噪声大小，我们的分析依赖 Rényi 差分隐私（RDP）技术（Mironov, 2017年），包括高斯机制的 RDP 保证、多个 RDP 机制的组合定理，以及将 RDP 保证转换为差分隐私保证。

以下命题揭示了满足$(\varepsilon_{\mathrm{DP}},\delta_{\mathrm{DP}})$-差分隐私保证所需的最小噪声水平$\sigma_1$ 和 $\sigma_2$，其值可能远小于 DP-GD 所需的最小噪声水平。

**命题 4.1**(满足$(\varepsilon_{\mathrm{DP}},\delta_{\mathrm{DP}})$-差分隐私)。 *DIFF 2 2α⌈R/T⌉/ $n_p^2P^2\sigma_1^2$ + 2α R −-GD* 中 $\lceil R/T\rceil$ / $n_p^2P^2\sigma_2^2$ + $1/\delta_{\mathrm{DP}}$ / α − 1 $\delta_{\mathrm{DP}}$ 定义的机制 $\{x_r,x_r^{\mathrm{out}},\widetilde{v}_r\}_{r\in[R]}$ 对任意 $\alpha\in(1,\infty)$ 均满足(()()()log()())-差分隐私，其中 $n_{\min}:=\min\{n_p\}_{p=1}^P$。特别地，令 $\alpha:=1+\lceil 2\log(1/\delta_{\mathrm{DP}})/\varepsilon_{\mathrm{DP}}\rceil$，

$$\sigma_1^2=\frac{4u\alpha\lceil\frac{R}{T}\rceil}{n_{\min}^2P^2\varepsilon_{\mathrm{DP}}}=\widetilde{\Theta}\left(\frac{R}{Tn_{\min}^2P^2\varepsilon_{\mathrm{DP}}^2}\right)$$

*and*

$$\sigma_2^2=\frac{\frac{4u}{u-1}\alpha(R-\lceil\frac{R}{T}\rceil)}{n_{\min}^2P^2\varepsilon_{\mathrm{DP}}}=\widetilde{\Theta}\left(\frac{R}{n_{\min}^2P^2\varepsilon_{\mathrm{DP}}^2}\right)$$

足以保证机制$\{x_r,x_r^{\mathrm{out}},\widetilde{v}_r\}_{r\in[R]}$ 对于任意 $u>1$ 满足 $(\varepsilon_{\mathrm{DP}},\delta_{\mathrm{DP}})$ 。

备注 4.2。 与 DP-GD 的噪声大小 $\sigma^2C_1^2=\widetilde{\Theta}(C_1^2R/($ $n_{\min}P\varepsilon_{\mathrm{DP}})^2)$ 相比， $\sigma_1^2C_1^2$小了 $T$ 倍，而且 $\sigma_2^2C_{2,\ r}^2$ 可能小得多，因为 $C_{2,\ r}=C_2\|x_{r-1}-x_{r-2}\|\ll C_1$ 可能成立。

### 4.2. DIFF2-GD 的效用分析

本小节给出了 DIFF2-GD 的效用分析。证明见补充材料的第 D 节。

如第4节所述，在 DIFF2 框架中，可以减小差分隐私噪声大小，从而降低全局梯度估计器 $\widetilde{v}_r$ 的方差；与标准差分隐私梯度下降相比，预期效用也会更好。另一方面，我们还需要考虑偏差 $\widetilde{v}_r$；可以看出， $\widetilde{v}_r$不再是 $\nabla f(x_{r-1})$ 的无偏估计量，因为$\widetilde{v}_r$ 包含了差分隐私噪声的历史信息。通常，重启间隔 $T$越大，$\widetilde{v}_r$的方差越小，但 $\widetilde{v}_r$的偏差越大。因此，主要挑战在于我们的效用分析需要仔细评估 $\widetilde{v}_r$ 的方差和偏差，并确定 $T$ 的最优取值，以控制这两种误差之间的权衡关系。

以下定理指出，在对 $T$ 作出最优选择时，DIFF2 在保证 ( $\varepsilon_{\mathrm{DP}},\delta_{\mathrm{DP}}$)-差分隐私的同时，其效用优于 DP-GD 已知的最佳结果。

定理 *4.3*（效用界）。假设假设 *2.1*、*2.2* 和 *2.3* 成立。设 $C_1\ge G$, $C_1=\Theta(G)$,$C_2\ge L$, $C_2=\Theta(L)$, $f(x_0)-f(x_*)=O(1)$ 且$n_{\min}P=\Omega(G^2\sqrt{d}/(L\varepsilon_{\mathrm{DP}}))$。在命题 *4.1* 中对 $\sigma_1^2$ 和 $\sigma_2^2$ 的选择下，若适当选择 $\eta=\Theta(\min\{1/L,1/(\sqrt{T}L\sigma_2\sqrt{d})\}$ 和 $T=\Theta(\max\{1,\tau R\})$，其中 $\tau:=(G^2\sqrt{d}/(Ln_{\min}P\varepsilon_{\mathrm{DP}}))^{2/3}$，则 *DIFF2-GD* 满足

$$\mathbb{E}\|\nabla f(x^{\mathrm{out}})\|^2\le O\left(\frac{L}{R}\right)+\widetilde{O}\left(\frac{L\sqrt{d}}{n_{\min}P\varepsilon_{\mathrm{DP}}\sqrt{R}}+\varepsilon_{\mathrm{opt}}\right),$$

其中 $\varepsilon_{\mathrm{opt}}:=\widetilde{\Theta}((LGd)^{\frac{2}{3}}/(n_{\min}P\varepsilon_{\mathrm{DP}})^{\frac{4}{3}})$。特别地，令

$$R=\Theta\left(1+\frac{L}{\varepsilon_{\mathrm{opt}}}\right)+\widetilde{\Theta}\left(\frac{L^2d}{n_{\min}^2P^2\varepsilon_{\mathrm{DP}}^2\varepsilon_{\mathrm{opt}}^2}\right)$$

可得到效用 $\mathbb{E}\|\nabla f(x^{\mathrm{out}})\|^2\le\varepsilon_{\mathrm{opt}}$。

注 4.4： 所得效用显著优于已知的最佳效用界 $\widetilde{O}($ $\sqrt{LG}\sqrt{d}/(n_{\min}P\varepsilon_{\mathrm{DP}}))$，当 $n_{\min}P\gg\sqrt{Ld}/(G\varepsilon_{\mathrm{DP}})$ 时尤为如此。此外，当 $n_{\min}P=\widetilde{\Omega}((Ld)^2/(G\varepsilon_{\mathrm{opt}}))$ 时，所需的通信轮数$R$ 变为 $O(1+L/\varepsilon_{\mathrm{opt}})$，且客户端 $p$ 所需的每样本梯度计算次数变为$O(n_p(1+L/\varepsilon_{\mathrm{opt}}))$。

## 5. 更高效的子程序

本节将简要讨论 DIFF2 框架的一种更高效的子程序，同时降低计算和通信成本。这一点并非我们的主要关注点，但对于实际使用 DIFF2 框架非常重要。具体而言，我们提出了一种基于 BVR-L-SGD 算法（Murata 与 Suzuki, 2021年）的局部优化子程序。该子程序继承了 BVR-L-SGD 的偏差-方差降低性质，因此在计算和通信方面具有良好的效率。本节旨在为 DIFF2 推导一种子程序，相较于 DIFF2-GD，它需要更低的计算和通信成本，同时保持与 DIFF2-GD 相同的效用。

**DIFF2-BVR-L-SGD** DIFF2-GD 的一个问题是，它在每一轮中都没有利用随机梯度和局部优化。这可能导致计算成本和通信成本都很高。因此，自然的做法是将局部优化与随机

算法 4 BVR-L-SGD 例程($x_{0,r-1}$, $\widetilde{v}_{1,r}$, $\eta$, $b$, $\sigma_3$,$C_3$, $K$) 1: 设置 $x_{1,r-1} = x_{0,r-1} - \eta\widetilde{v}_{1,r}$. 2: 设置 $p_r = 1 + r\%P$. 3: 客户端 $p_r$ 从中央服务器接收$x_{1,r-1}$ . 4: # 仅客户端 $p_r$ 执行局部优化。 5: 对于 $k = 2$ 到 $K$， 执行 6: 从 $D_{p_r}$ 中无放回抽取大小为 $b$ 的随机子集 $I_{k,r}$ . 7: 设置 $C_{3,k,r} = C_3\|x_{k-1,r-1} - x_{k-2,r-1}\|$. 8: $v_{k,r}$ = 裁剪均值($\{\nabla\ell(x_{k-1,r-1}, z) - \nabla\ell(x_{k-2,r-1}, z)\}_{z\in I_{k,r}}, C_{3,k,r}$) + $\widetilde{v}_{k-1,r}$. 9: 设置 $\widetilde{v}_{k,r}$ = $v_{k,r}$ + $\xi_{k,r}$， 其中 $\xi_{k,r}$ $\sim N(0, \sigma_3^2 C_{3,k,r}^2 I)$. 10: 更新 $x_{k,r-1} = x_{k-1,r-1} - \eta\widetilde{v}_{k,r}$. 11: 结束循环 12: 客户端 $\hat{p}$ 将$x_{K,r-1}$ 和 $x_{\hat{k}-1,r-1}$ 发送到中央$\hat{}$服务器，其中 $k \sim$ 服从 Unif$[K]$. 13: 返回： $x_{K,r-1}$,$x_{\hat{k},r-1}$。

局部梯度，以降低计算和通信成本。受 BVR-L-SGD（4） Murata 与 Suzuki（2021年）启发的 BVR-L-SGD 例程是解决上述两个问题的 GD-Routine 的一种良好替代方案。在每次局部迭代 $k \in [K]$ 中，BVR-L-SGD 例程根据逐样本梯度差异的裁剪均值更新偏差-方差缩减的随机梯度估计器 $v_{k,\ r}$ 其方式与 DIFF2 类似（第 8 行）。随后，将均值为零、方差为 $\sigma_3^2 C_3^2\|x_{k-1,\ r-1} - x_{k-2,\ r-1}\|^2$ 的高斯噪声 $\xi_{k,\ r}$ 添加到 $v_{k,\ r}$ 并根据差分隐私梯度估计器 $\widetilde{v}_{k,\ r} := v_{k,\ r} + \xi_{k,\ r}$ 更新当前参数（第 9 行）。我们将结合 DIFF2 的 BVR-L-SGD 例程称为 DIFF2-BVR-L-SGD。注意，当 $T = 1$ 时，不含差分隐私噪声的 DIFF2-BVR-L-SGD 与 BVR-L-SGD 的特殊情形一致。

与 DIFF2-GD 的分析类似，我们可以推导出 DIFF2-BVR-L-SGD 在 $(\varepsilon_{\rm DP}, \delta_{\rm DP})$-差分隐私保证下的最小噪声水平。重要的是，由于使用了随机梯度，我们需要运用 Wang 等人(2019c) 提出的 Rényi 差分隐私子采样放大技术进行精确分析。详细表述见补充材料中的命题 E.2。

命题 *5.1*（$(\varepsilon_{\rm DP}, \delta_{\rm DP})$-差分隐私的最小噪声水平（简化版））。存在 $\sigma_1^2$、 $\sigma_2^2$ 和 $\sigma_3^2$， 使得 *DIFF2-BVR-L-SGD* 满足 $(\varepsilon_{\rm DP}, \delta_{\rm DP})$-差分隐私，并且 $\sigma_1^2 = \widetilde{\Theta}(R/(Tn_{\min}^2P^2\varepsilon_{\rm DP}^2))$, $\sigma_2^2 = \widetilde{\Theta}(R/(n_{\min}^2P^2\varepsilon_{\rm DP}^2))$ 且 $\sigma_3^2 = \Theta((KR/(n_{\min}^2P\varepsilon_{\rm DP}^2)) \vee (1/(b^2\varepsilon_{\rm DP})))$，其中 $b \le \min\{n_{\min}/(2e\alpha), (4n_{\min}/(\alpha\sigma_3^2))^{1/3}\}$。

为展示 DIFF2-BVR-L-SGD 的通信效率，还需采用以下假设。该假设常用于非差分隐私分布式优化中——（Karimireddy 等人，2021年；Murata 与 Suzuki，2021年）。注意，该假设在 $\zeta := L$ 在假设 2.1 下，并且我们默认 $\zeta \ll L$。

**假设 5.2**（海森异质性）。 $\{f_p\}_{p=1}^P$ 是 $\zeta$-海森异质性，即

$$\left\|\nabla^2 f_p(x) - \nabla^2 f(x)\right\| \le \zeta, \forall x \in \mathbb{R}^d.$$

为了推导 DIFF2-BVR-L-SGD 的效用界，我们需要界定由差分隐私噪声以及局部优化和随机梯度噪声额外造成的偏差和方差。下述定理表明，DIFF2-BVR-L-SGD 能够达到与 DIFF2-GD 相同的效用。详细表述见补充材料中的定理 F.4。

**定理 5.3** （效用界（简化版））。假设定理 *4.3* 的假设成立。此外，假设 *5.2* 成立，且小批量大小 $b$ 满足命题 *5.1*中的条件。令 $C_3 := C_2$。在命题 $\sigma_1^2$、 $\sigma_2^2$ 和 $\sigma_3^2$ *5.1*中给出的选择下，若选取适当的$\eta$ 和 $T$，则 *DIFF2-BVR-L-SGD* 可达到如下效用：

$$\mathbb{E}\|\nabla f(x^{\rm out})\|^2 \le \varepsilon_{\rm opt} := \widetilde{\Theta}\left(\frac{(LGd)^{\frac{2}{3}}}{(n_{\min}P\varepsilon_{\rm DP})^{\frac{4}{3}}}\right)$$

通过设置 $R = \Theta(1) + c_1/\varepsilon_{\rm opt} + c_2/\varepsilon_{\rm opt}^2$， 其中 $c_1 := \Theta(L/K + \zeta + L/(\sqrt{K}b)) + \widetilde{\Theta}(L\sqrt{d}/(\sqrt{Kb}\sqrt{\varepsilon_{\rm DP}}))$ 且 $c_2 := \widetilde{\Theta}(L^2d/(n_{\min}^2P\varepsilon_{\rm DP}^2))$。

备注 5.4（通信效率）。 DIFF2-BVR-L-SGD 的通信效率可能远高于 DIFF2-GD。事实上，当 $n_{\min}P = \widetilde{\Omega}(L^2\sqrt{Pd}/(G\varepsilon_{\rm DP}))$ 时，有 $c_2/\varepsilon_{\rm opt}^2 \le O(1)$，因此 DIFF2-BVR-L-SGD 所需的通信轮数为 $O(1 + c_1/\varepsilon_{\rm opt})$，这在 $K \gg 1$ 和 $\zeta \ll L$ 时远小于 DIFF2-GD 的 $O(1 + L/\varepsilon_{\rm opt})$。

备注 5.5（计算效率）。 DIFF2-BVR-L-SGD 的计算效率可能高于 DIFF2-GD。事实上，客户端 $p$ 达到最佳可能效用所需的单次梯度计算总数为 $O((n_p+Kb)R)$。如上面的备注所示，通信轮数 $R$ 可能远小于 DP-GD，因此当 $Kb = O(n_p)$ 时，总计算成本也可能低于 DP-GD。

## 6. 数值结果

在本节中，我们提供了一些实验结果，以验证第 4 节中的理论发现。具体而言，我们通过实证比较 DIFF2-GD 与 DP-GD 的效用，验证 DIFF2 框架的优越性。我们重点研究了维度相对较低的问题，其中 $n_{\min}P \gg d$，因此效用

$\widetilde{O}(d^{2/3}/(n_{\min}P\varepsilon_{\mathrm{DP}})^{4/3})$ 的 DIFF2-GD 效用预计优于 DP-GD 的 $\widetilde{O}(\sqrt{d}/(n_{\min}P\varepsilon_{\mathrm{DP}}))$。

我们在五个数据集上进行了回归和分类任务：（i）加利福尼亚住房数据集 [7]；（ii）燃气轮机一氧化碳和氮氧化物排放数据集 [8]；（iii）博客反馈数据集 [9]；（iv）KDDCup99 数据集 [10]；以及（v）土地覆盖类型数据集 [11]。我们仅在论文正文中报告回归任务的结果。数据集概述以及分类任务的结果见补充材料第A 节。

**数据准备**。对于每个数据集，我们将原始数据集随机划分为 80% 的训练集和 20% 的测试集。随后，我们对数值属性进行归一化，使其均值为零、标准差为一。此外，我们用整个数据集中目标变量的最大绝对值除目标变量，对其进行归一化。我们将训练集随机划分为大小相等的 10 个子集，并将每个子集分配给 10 个工作节点中相应的工作节点[12]。

**模型**。 我们使用一个包含 10 个隐藏单元并采用 Softplus 激活函数的单隐藏层全连接神经网络进行了实验。损失函数采用平方损失。我们通过从 $[-\sqrt{w_{\mathrm{in}}}, \sqrt{w_{\mathrm{in}}}]$ 中均匀采样来初始化参数，其中 $w_{\mathrm{in}}$ 是输入层中的单元数。

**已实现算法**。 我们实现了非隐私梯度下降、DP-GD 和 DIFF2-GD。对于 DP-GD 和 DIFF2-GD，差分隐私参数设为 $\varepsilon_{\mathrm{DP}} \in \{3.0, 5.0\}$ 和 $\delta_{\mathrm{DP}} = 10^{-5}$。我们使用命题 4.1 来确定 DP-GD（$T = 1, u \to 1$）和 DIFF2-GD（$T \in \{0.003R, 0.01R, 0.03R, 0.1R\}, u = 1.25$）的差分隐私噪声大小。对于每种算法，我们都自动调整了超参数。调参过程的详细信息见补充材料的第 A 节。

**评估**。我们根据三项指标评估了已实现算法随通信轮数变化的表现：训练损失、训练梯度平方范数和测试损失。我们报告了在使各项指标最小化的超参数下，已实现算法的训练损失和训练梯度平方范数。另一方面，我们报告了在使训练损失最小化的超参数下，已实现算法的测试损失。每种算法的通信总轮数固定为 2，000。我们独立重复了实验 5 次，并报告上述指标的均值和标准差。

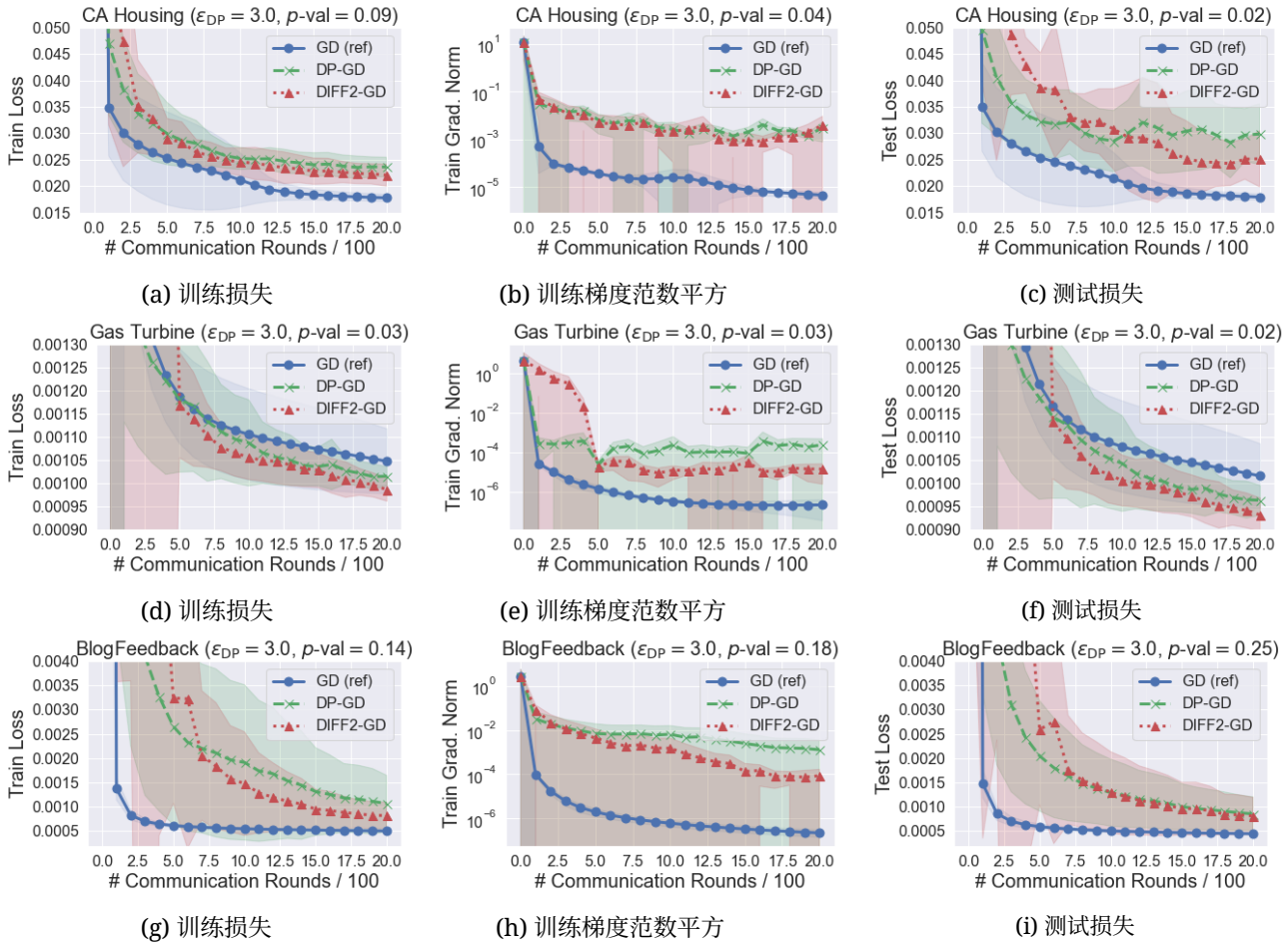
**结果**。图2 展示了 $\varepsilon_{\mathrm{DP}} = 3.0$ 个数据集上 DIFF2-GD 与 DP-GD 的比较。“GD（ref）”表示不添加差分隐私噪声的梯度下降。每幅图标题中的“$p$-值”表示 $p$-值，即针对单侧$t$-检验的 2,000 轮 DIFF2-GD 与 DP-GD 的最小值之差为负这一备择假设下的检验结果；其中一个随机种子对应一个样本，即$t$-分布的自由度为四。我们可以看到，对于每项指标，DIFF2-GD 在最终学习轮次中的表现始终优于 DP-GD，尽管它在初始学习轮次有时收敛速度较慢。此外，$t$-检验结果支持 DIFF2-GD 的优势具有统计显著性。 $\varepsilon_{\mathrm{DP}} = 5.0$ 的情况与 $\varepsilon_{\mathrm{DP}} = 3.0$的情况结果相似，受篇幅限制，相关结果见补充材料。

## 7. 结论与未来工作

在这项工作中，我们提出了一种名为 DIFF2 的新型差分隐私优化框架，以改进此前已知的最佳效用界。DIFF2 传输的是梯度差分，而非梯度本身，并基于梯度差分构造了一个方差可能非常小的差分隐私全局梯度估计器。研究表明，采用梯度下降子程序的 DIFF2 可实现 $\widetilde{O}(d^{2/3}/(n_{\min}P\varepsilon_{\mathrm{DP}})^{4/3})$ 的效用；就对样本量 $n_{\min}P$ 的依赖而言，这一结果可能显著优于此前的结果。我们进行了数值实验，验证了 DIFF2-GD 相较于 DP-GD 的优越性，以及 DIFF2 框架带来的益处。

未来第一项重要工作是推导用于非凸光滑目标函数的 $(\varepsilon_{\mathrm{DP}}, \delta_{\mathrm{DP}})$-差分隐私优化算法的效用下界，并讨论 DIFF2 框架的最优性。第二项重要工作是针对高维问题，为差分隐私优化算法开展新的效用分析；这类问题常见于现代大型深度神经网络的学习。尽管本研究解决了差分隐私优化理论中一个相当基础且重要的问题——效用提升，但我们的理论改进仅限于相对低维的问题。最后，将我们的算法扩展到各种方向也是一项重要的未来工作。例如，跨设备联邦学习通常采用客户端采样，但本文并未讨论这一点。此外，DIFF2 框架

[7] https://www.dcc.fc.up.pt/~ltorgo/ 回归/加利福尼亚_住房.html。[8] https://archive.ics.uci.edu/ml/datasets/ 燃气+涡轮机+一氧化碳+和+氮氧化物+排放+数据+集。我们移除了 NOx 属性，并使用 CO 作为目标变量。[9] https://archive.ics.uci.edu/ml/datasets/ 博客反馈。[10] https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html [11] http://archive.ics.uci.edu/ml/datasets/ 覆盖类型 [12] 这意味着本地数据集以独立同分布方式进行分配。由于我们关注的是 DIFF2-GD 与 DP-GD 的比较，且这些方法不依赖局部优化，因此异质性不会影响优化过程。因此，我们决定采用最简单的独立同分布情形。




(a) 训练损失 (b) 训练梯度范数平方 (c) 测试损失

(d) 训练损失 (e) 训练梯度范数平方 (f) 测试损失

(g) 训练损失 (h) 训练梯度范数平方 (i) 测试损失

图 *1.* 训练损失、训练梯度范数和测试损失随通信轮数（$\varepsilon = 3.0$、$\delta = 10^{-5}$ 和 $R = 2,000$）的变化对比。(a)-(c)展示了加利福尼亚住房数据集上这三项指标的对比，(d)-(f)展示了燃气轮机一氧化碳和氮氧化物排放数据集上的对比，(g)-(i)展示了博客反馈数据集上的对比。DIFF2-GD的表现始终优于DP-GD。

在去中心化拓扑上的应用是一项重要扩展。将我们的差分隐私保证和效用分析扩展到这些场景并非易事。尽管第 5 节简要提到了比梯度下降计算和通信效率更高的子程序，但多种优化算法都可以纳入 DIFF2 框架，包括简单的随机梯度下降、加速梯度方法、自适应梯度方法等。接下来，开展统一的理论分析，并比较它们的理论和实证性能，十分重要。

## 致谢

TS部分得到了JSPS KAKENHI(20H00576)和JST CREST的资助。

## 参考文献

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K. 和 Zhang, L. 使用差分隐私的深度学习。载于*ACM SIGSAC*计算机与通信安全会议，第308–318页，2016年。

Agarwal, N., Suresh, A. T., Yu, F. X. X., Kumar, S., 和 McMahan, B. cpSGD：通信高效且满足差分隐私的分布式随机梯度下降。载于神经信息处理系统进展，第 31 卷，2018年。

Asi, H., Feldman, V., Koren, T., 和 Talwar, K. 私有随机凸优化：$l_1$ 几何下的最优速率。载于国际机器学习会议，第 139 卷，第 393–403 页，2021年。

Bassily, R.、Smith, A. 和 Thakurta, A. 私有经验风险最小化：高效算法与紧致误差

界限。载于 *IEEE Symposiumon Foundationsof Computer Science*，第 55 卷，第 464–473 页，2014 年。

Bassily, R.、Feldman, V.、Talwar, K. 和 Guha Thakurta, A. 具有最优速率的私有随机凸优化。载于《神经信息处理系统进展》，第 32 卷，2019年。

Chaudhuri, K.、Monteleoni, C. 和 Sarwate, A. D. 差分隐私经验风险最小化。《机器学习研究杂志》，第 12 卷，2011年。

Chourasia, R.、Ye, J. 和 Shokri, R. 朗之万扩散和噪声梯度下降的差分隐私动态。载于《神经信息处理系统进展》，第 34 卷，第 14771–14781 页，2021年。

Cutkosky, A. 和 Orabona, F. 非凸随机梯度下降中的基于动量的方差缩减。载于《神经信息处理系统进展》，第 32 卷，2019年。

Das, R.、Kale, S.、Xu, Z.、Zhang, T. 和 Sanghavi, S. 超越差分隐私优化中的一致利普希茨条件。《*arXiv* 预印本 *arXiv:2206.10713*》，2022年。

Defazio, A.、Bach, F. 和 Lacoste-Julien, S. SAGA：一种快速的增量梯度法，适用于非强凸复合目标函数。载于 神经信息处理系统进展，第27卷，第1646–1654页，2014年。

Ding, J.、Liang, G.、Bi, J. 和 Pan, M. 差分隐私与通信高效协同学习。载于*AAAI*人工智能大会，第35卷，第7219–7227页，2021年。

Du, J.、Li, S.、Chen, X.、Chen, S. 和 Hong, M. 动态差分隐私随机梯度下降。*arXiv*预印本 *arXiv:2111.00173*，2021年。

Dwork, C. 为隐私数据分析奠定坚实基础。*ACM*通信，54：86–95，2011年。

Dwork, C. 和 Naor, M. 统计数据库中的信息披露防护难题，或差分隐私的理由。隐私与保密杂志，2：93–107，2010年。

Dwork, C., McSherry, F., Nissim, K., 和 Smith, A. 在隐私数据分析中校准噪声以匹配敏感度。发表于密码学理论会议，第 265–284 页，2006 年。

Dwork, C., Rothblum, G. N., 和 Vadhan, S. 提升与差分隐私。发表于*IEEE*计算机科学基础年会，第 51–60 页，2010 年。

Dwork, C., Roth, A., 等. 差分隐私的算法基础。理论计算机科学基础与趋势®，9:211–407，2014年。

Fredrikson, M., Jha, S., 和 Ristenpart, T. 利用置信度信息的模型反演攻击及基本对策。发表于*ACM SIGSAC*计算机与通信安全会议，第 1322–1333 页，2015年。Geiping, J., Bauermeister, H., Dr¨oge, H., 和 Moeller, M. 反演梯度——在联邦学习中破坏隐私有多容易？发表于神经信息处理系统进展，第 33 卷，第 16937–16947 页，2020年。Geyer, R. C., Klein, T., 和 Nabi, M. 差分隐私联邦学习：客户端层面的视角。*arXiv* 预印本 *arXiv:1712.07557*，2017年。

Han, A., Mishra, B., Jawanpuria, P., 和 Gao, J. 差分隐私黎曼优化。*arXiv* 预印本 *arXiv:2205.09494*，2022年。

Hu, L., Ni, S., Xiao, H., 和 Wang, D. 重尾数据下的高维差分隐私随机优化。发表于*ACM SIGMOD-SIGACT-SIGAI*数据库系统原理研讨会，第 227–236 页，2022年。

Huang, Z. 和 Gong, Y. 面向凸分布式学习的差分隐私交替方向乘子法：通过多步近似提升准确率。*arXiv* 预印本 *arXiv:2005.07890*，2020年。

Huang, Z.、Hu, R.、Guo, Y.、Chan-Tin, E. 和 Gong, Y. DP-交替方向乘子法：基于交替方向乘子法的差分隐私分布式学习。*IEEE*信息取证与安全汇刊，15：1002–1012，2019年。

Johnson, R. 和 Zhang, T. 使用预测方差缩减加速随机梯度下降。载于神经信息处理系统进展，第26卷，第315–323页，2013年。

Kamath, G.、Liu, X. 和 Zhang, H. 针对重尾数据的差分隐私随机凸优化的改进速率。载于国际机器学习会议，第162卷，第10633–10660页，2022年。

Karimireddy, S. P., Jaggi, M., Kale, S., Mohri, M., Reddi, S., Stich, S. U. 和 Suresh, A. T. 打破跨设备联邦学习的集中式障碍。载于神经信息处理系统进展，第 34 卷，第 28663–28676 页，2021年。

Kone˘cn`y, J., McMahan, H. B., Yu, F. X., Richt´arik, P., Suresh, A. T. 和 Bacon, D. 联邦学习：提升通信效率的策略。*arXiv* 预印本 *arXiv:1610.05492*，2016。

Kuru, N.、Birbil, I.、Gurbuzbalaban, M. 和 Yildirim, S. 差分隐私加速优化算法。*SIAM* 优化期刊，32：795–821，2022年。

Li, N.、Li, T. 和 Venkatasubramanian, S. *t*-接近性：超越 *k*-匿名性和 *l*-多样性的隐私保护。载于*IEEE* 国际数据工程会议，第 106–115 页，2006 年。

Li, Z., Zhao, H., Li, B. 和 Chi, Y. SoteriaFL：一种用于隐私保护联邦学习并支持通信压缩的统一框架。发表于神经信息处理系统进展，2022年。即将发表。

Liang, Z., Wang, B., Gu, Q., Osher, S. 和 Yao, Y. 采用拉普拉斯平滑的差分隐私联邦学习。*arXiv* 预印本 *arXiv:2005.00218*，2020年。

Liu, D. 和 Lu, Z. 差分隐私 ERM 的紧下界。*OpenReview* 预印本 *https://openreview.net/forum?id=30nbp1eV0dJ*，2021年。

Machanavajjhala, A., Kifer, D., Gehrke, J. 和 Venkitasub-ramaniam, M. *l*-多样性：超越 *k*-匿名的隐私保护。*ACM*数据知识发现汇刊，1:3–es，2007年。

McMahan, B., Moore, E., Ramage, D., Hampson, S. 和 y Arcas, B. A. 基于去中心化数据的深度网络通信高效学习。载于国际人工智能与统计会议，第 54 卷，第 1273–1282 页，2017年。

McMahan, H. B., Ramage, D., Talwar, K. 和 Zhang, L. 差分隐私循环语言模型的学习。载于国际学习表征会议，第 6 卷，2018年。

Mironov, I. R´enyi 差分隐私。载于*IEEE*计算机安全基础研讨会，第 30 卷，第 263–275 页，2017年。

Murata, T. 和 Suzuki, T. 用于降低异质性的联邦学习的偏差-方差缩减局部随机梯度下降。载于国际机器学习会议，第 139 卷，第 7872–7881 页，2021年。

Nasr, M.、Shokri, R. 和 Houmansadr, A. 深度学习的全面隐私分析：针对集中式学习和联邦学习的被动式与主动式白盒推理攻击。载于 *IEEE*安全与隐私研讨会，第739–753页，2019年。

Nguyen, L. M.、Liu, J.、Scheinberg, K. 和 Tak´aˇc, M. SARAH：一种使用随机递归梯度解决机器学习问题的新方法。载于国际机器学习会议，第70卷，第2613–2621页，2017年。

Noble, M., Bellet, A., 和 Dieuleveut, A. 面向异构数据的差分隐私联邦学习。载于人工智能与统计国际会议，第151卷，第10110–10145页。PMLR，2022年。

Samarati, P. 和 Sweeney, L. 信息披露时的隐私保护：*k*-匿名及其通过泛化和抑制实现的执行机制。技术报告，SRI International，1998年。Shang, F., Xu, T., Liu, Y., Liu, H., Shen, L., 和 Gong, M. 用于机器学习的差分隐私ADMM算法。*IEEE*信息取证与安全汇刊，16：4733–4745，2021年。Shokri, R., Stronati, M., Song, C., 和 Shmatikov, V. 针对机器学习模型的成员推断攻击。载于*IEEE*安全与隐私研讨会，第3–18页，2017年。

Song, S.、Chaudhuri, K. 和 Sarwate, A. D. 采用差分隐私更新的随机梯度下降。载于 *IEEE*全球信号与信息处理大会，第245–248页，2013年。

Tran, H. 和 Cutkosky, A. 面向隐私保护非凸经验风险最小化的动量聚合。载于 神经信息处理系统进展，2022年。即将出版。

Triastcyn, A. 和 Faltings, B. 贝叶斯差分隐私下的联邦学习。载于*2019*年*IEEE*大数据国际会议，第 2587–2596 页，2019年。

Wang, B.、Gu, Q.、Boedihardjo, M.、Wang, L.、Barekat, F. 和 Osher, S. J. DP-LSSGD：一种提升隐私保护经验风险最小化效用的随机优化方法。载于数学与科学机器学习会议，第 107 卷，第 328–351 页，2020a。

Wang, D.、Ye, M. 和 Xu, J. 重新审视差分隐私经验风险最小化：更快且更通用。载于神经信息处理系统进展，第 30 卷，2017年。

Wang, D.、Chen, C. 和 Xu, J. 使用非凸损失函数的差分隐私经验风险最小化。载于国际机器学习会议，第 97 卷，第 6526–6535 页，2019a。

Wang, D., Xiao, H., Devadas, S. 和 Xu, J. 关于重尾数据下差分隐私随机凸优化。载于国际机器学习会议，第 119 卷，页码 10081–10091，2020b。

Wang, L., Jayaraman, B., Evans, D. 和 Gu, Q. 高效的隐私保护随机非凸优化。*arXiv* 预印本 *arXiv:1910.13659*，2019b。

Wang, Y.-X., Balle, B., 和 Kasiviswanathan, S. P. 子采样 Rényi 差分隐私与解析矩会计器。载于国际人工智能与统计会议，第 89 卷，第 1226–1235 页，2019c。

Wang, Z., Song, M., Zhang, Z., Song, Y., Wang, Q., 和 Qi, H. 超越类别代表推断：联邦学习中的用户级隐私泄露。载于*IEEE INFOCOM*计算机通信会议，第 2512–2520 页，2019d。

Yang, Z., Zhang, J., Chang, E.-C., 和 Liang, Z. 通过背景知识对齐，在对抗环境中实现神经网络反演。载于*ACM SIGSAC*计算机与通信安全会议，第 225–240 页，2019年。

Yang, Z., Lei, Y., Lyu, S., 和 Ying, Y. 非光滑损失下成对学习的随机梯度下降的稳定性与差分隐私。载于国际人工智能联合会议，第 130 卷，第 2026–2034 页。PMLR，2021年。

Yeom, S., Giacomelli, I., Fredrikson, M., 和 Jha, S. 机器学习中的隐私风险：分析其与过拟合的关联。载于*IEEE*计算机安全基础研讨会，第 268–282 页，2018年。

Zhang, J., Zheng, K., Mou, W., 和 Wang, L. 面向光滑目标函数的高效私有 ERM。载于国际人工智能联合会议，第 3922–3928 页，2017年。

Zhang, X., Fang, M., Liu, J., and Zhu, Z. 私有且通信高效的边缘学习：一种基于稀疏差分高斯掩蔽的分布式随机梯度下降方法。载于移动网络与移动计算的理论、算法基础及协议设计国际研讨会，第 261–270 页，2020a。

Zhang, Y., Jia, R., Pei, H., Wang, W., Li, B., and Song, D. 秘密揭示者：针对深度神经网络的生成模型反演攻击。载于*IEEE/CVF*计算机视觉与模式识别会议，第 253–261 页，2020b。

Zhou, Y., Chen, X., Hong, M., Wu, Z. S., and Banerjee, A. 私有随机非凸优化：自适应算法与更紧的泛化界。*arXiv*预印本 *arXiv:2006.13501*，2020年。

Zhu, L., Liu, Z. 和 Han, S. 梯度导致的深度泄露。载于神经信息处理系统进展，第 32 卷，2019年。

# A. 数值实验补充材料

在本节中，我们提供　　由于篇幅限制，论文中未给出我们数值实验的其他信息和结果
　　由于篇幅限制，论文中未予展示。

### 超参数调优

非隐私梯度下降的超参数只有学习率 $\eta$。DP-GD 的超参数是学习率 $\eta$ 和裁剪半径 $C_1$。Diff2-GD 的超参数是学习率 $\eta$、重启间隔 $T$、裁剪半径 $C_1$ 和 $C_2$。

为调整裁剪半径和重启间隔，我们使用 $C_1 \in \{1, 3.0, 10.0, 30.0, 100.0\}$ 运行 DP-GD，并使用 $C_1$，$C_2 \in \{1, 3.0, 10.0, 30.0, 100.0\}$ 和 $T \in \{0.003R, 0.01R, 0.03R, 0.1R\}$ 运行 DIFF2-GD，共进行 2，000 轮。为调整学习率 $\eta$，我们使用 $\eta \in \{0.5^i | i \in \{0, \ldots, 9\}\}$运行每种已实现的算法。为缩短执行时间，每 20 轮评估一次训练损失；如果检查训练损失后发现所用学习率不合适，就停止训练。具体而言，出现以下情况时，学习率会被判定为不合适：(i) 训练损失达到 NAN；或 (ii) 耐心计数达到五。若当前训练中的训练损失大于此前最低训练损失的 1.05 倍，耐心计数就加一；若当前训练损失小于此前最佳训练损失，则耐心计数重置为零。如果使用某个学习率 $\eta_*$达到 2，000 轮，我们就终止学习率调优，并将最佳学习率确定为 $\eta_*$，而不再继续尝试更小的学习率，以免浪费计算资源。

对于训练损失和训练梯度平方范数这两个标准，我们选择使标准最小化的最佳超参数，并报告使用最佳超参数时的学习结果。对于测试损失，我们选择使训练损失最小化的最佳超参数，并报告使用最佳超参数时的学习结果。

这些过程针对全部五个随机种子分别独立执行。

### 数据集信息

我们实验中使用的三个数据集的信息总结于表2中。

表 2。 数值实验中使用的数据集摘要。

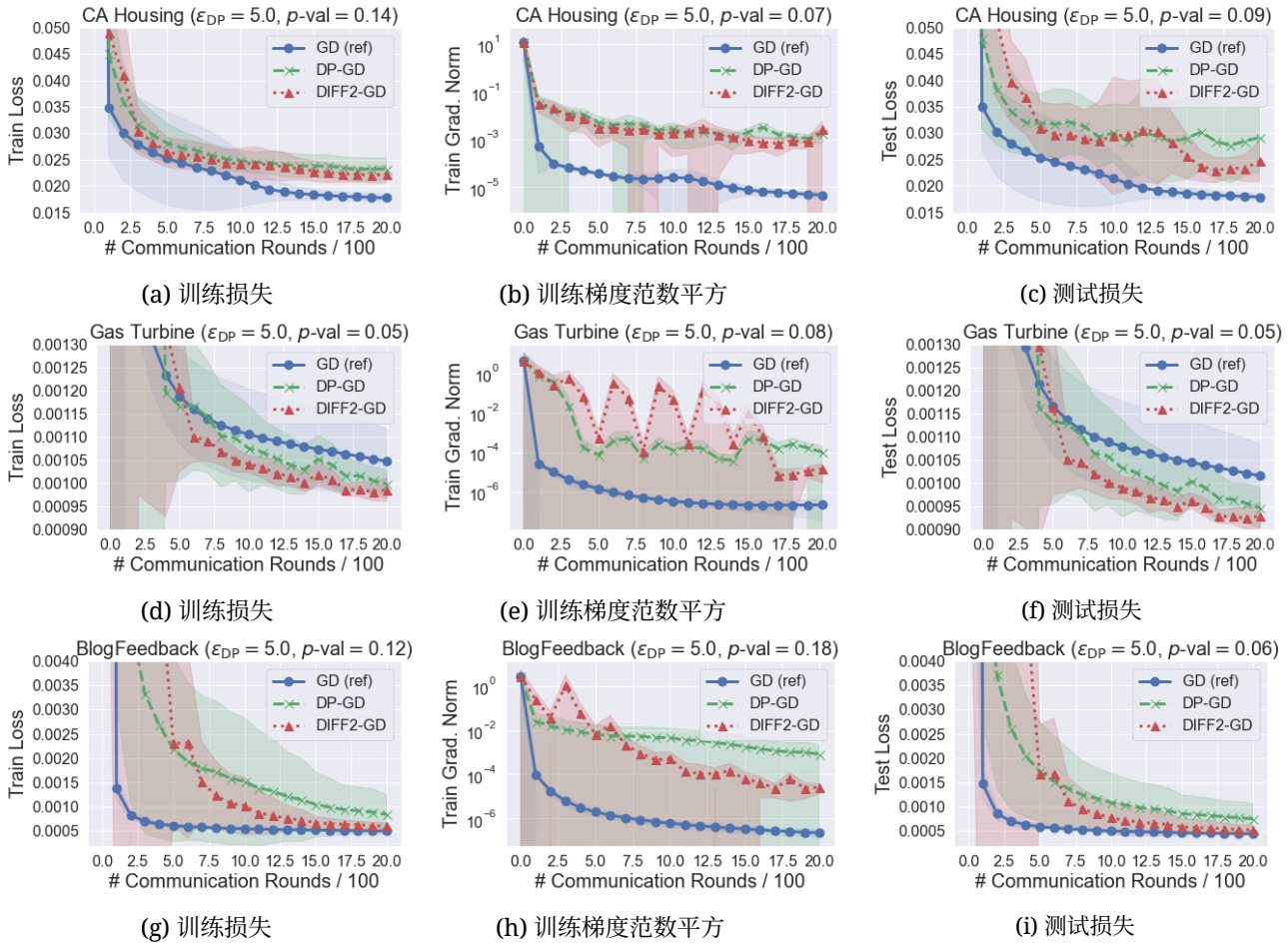
| 数据集 | # 样本 | # 属性 |
|---|---|---|
| 加利福尼亚住房 | 20，640 | 8 |
| 燃气轮机一氧化碳和氮氧化物排放 | 36, 733 | 9 |
| 博客反馈 | 60，021 | 280 |
| 覆盖类型 | 58, 1012 | 54 |
| KDDCup99 | 4,898,431 [13] | 41 |

### 回归任务的补充数值结果

这里，我们提供了回归任务中 $\varepsilon_{\mathrm{DP}} = 5.0$ 情况的更多数值结果。我们确认，结果与在主论文中提供 $\varepsilon_{\mathrm{DP}} = 3.0$ 时的结果相似。

### 计算基础设施

- 操作系统：Ubuntu 16.04.6
- CPU：AMD EPYC 7552 48 核处理器。
- CPU 内存：1.0 TB。
- 编程语言：Python 3.9.12。
- 深度学习框架：Pytorch 1.12.1。[13]在我们的实验中，我们仅使用了 10% 的数据。




(a) 训练损失 (b) 训练梯度范数平方 (c) 测试损失

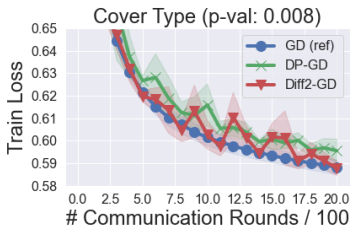
(d) 训练损失 (e) 训练梯度范数平方 (f) 测试损失

(g) 训练损失 (h) 训练梯度范数平方 (i) 测试损失

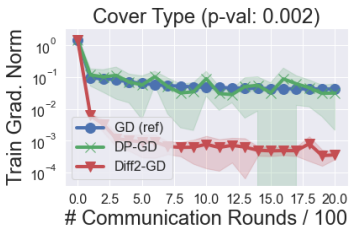
图 *2*。 训练损失、训练梯度范数和测试损失随通信轮数（$\varepsilon = 5.0$、 $\delta = 10^{-5}$ 和 $R = 2,\ 000$）的变化对比。(a)–(c) 展示了加利福尼亚住房数据集上这三项指标的对比，(d)–(f) 展示了燃气轮机一氧化碳和氮氧化物排放数据集上的对比，(g)–(i) 展示了博客反馈数据集上的对比。我们确认，DIFF2-GD 的表现始终优于 DP-GD。

## 分类任务的其他数值结果

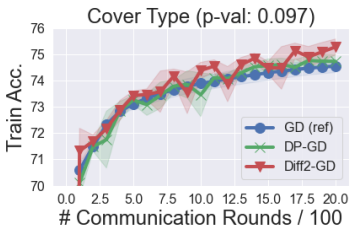
这里，我们提供了关于 $\varepsilon_{\mathrm{DP}} = 3.0$ 分类任务的更多数值结果。我们确认，即使在分类任务中，所提方法也始终表现出明显优势。



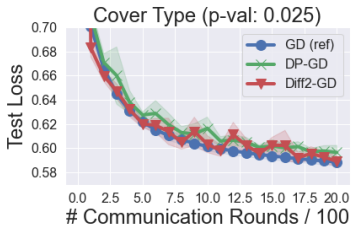

(a) 训练损失



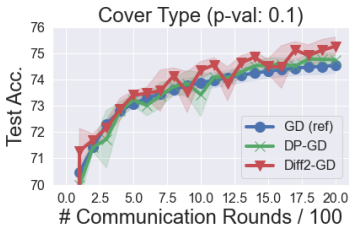

(b) 训练梯度范数平方



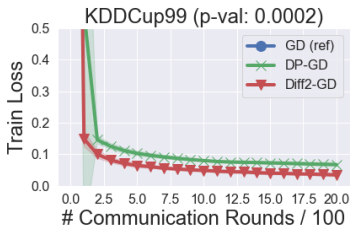

(c) 训练准确率



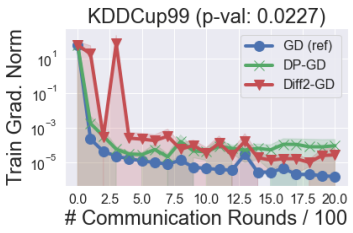

(d) 测试损失



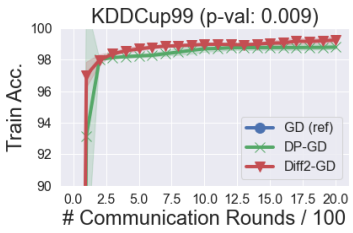

(e) 测试准确率



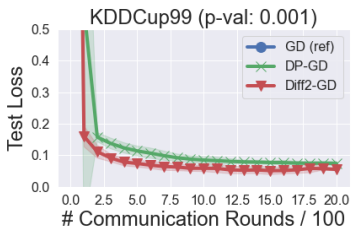

(f) 训练损失



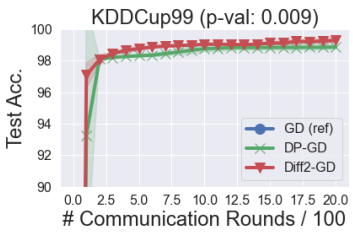

(g) 训练梯度范数平方




(h) 训练准确率




(i) 测试损失




(j) 测试准确率

图 *3.* 训练损失、训练梯度范数和测试损失随通信轮数的变化对比（$\varepsilon = 3.0$、$\delta = 10^{-5}$ 和 $R = 2,\ 000$）。(a)-(e) 展示了土地覆盖类型数据集上三个指标的对比，(f)-(j) 展示了 KDDCup99 数据集上的对比。我们确认，DIFF2-GD 始终优于 DP-GD。

## B. Rényi 差分隐私综述

我们的差分隐私分析依赖于结合子采样和洗牌放大的 Rényi 差分隐私（RDP）技术。本节简要回顾分析中用到的一些已知 RDP 结果。

定义 B.1（（$\alpha, \varepsilon$）-RDP（（Mironov，2017年））。若随机机制 $\mathcal{M}\colon \mathcal{D} \to \mathcal{R}$ 满足（$\alpha, \varepsilon$）-RDP（$\alpha \in (1, \infty)$ 和 $\varepsilon > 0$），则对于任意数据集 $D, D' \in \mathcal{D}$，只要满足 $d_{\mathrm{H}}(D, D') = 1$，就有

$$\frac{1}{\alpha - 1} \log \mathbb{E}_{o \sim \mathcal{M}(D')} \left( \frac{\mathcal{M}(D)(o)}{\mathcal{M}(D')(o)} \right)^{\alpha} \leq \varepsilon,$$

其中，$M(D)(o)$ 表示 $\mathcal{M}(D)$ 在 $o$ 处的密度。

**引理 B.2** (RDP 后处理性质（Mironov，2017年））。设 $\mathcal{M}$： $\mathcal{D} \to \mathcal{R}$ 是$(\alpha,\ \varepsilon)$-*Rényi* 差分隐私，且 $g$：$\mathcal{R} \to \mathcal{R}'$ 是任意函数。则 $g \circ \mathcal{M}$： $\mathcal{D} \to \mathcal{R}'$ 也满足$(\alpha,\ \varepsilon)$-*Rényi* 差分隐私。

引理 B.3（RDP 机制的组合（（Mironov，2017年））。设 $\mathcal{M}_r$: $\mathcal{R}_1 \times \cdots \times \mathcal{R}_{r-1} \times \mathcal{D} \to \mathcal{R}_r$ be $(\alpha, \varepsilon_r)$-RDP for $r \in [R]$。那么， $\mathcal{M}$: $\mathcal{D} \to \mathcal{R}_1 \times \cdots \times \mathcal{R}_R$ defined by $\mathcal{M}(D) := (\mathcal{M}_1(D)\ \mathcal{M}_2(\mathcal{M}_1(D)\ D) \ldots, \mathcal{M}_R(\mathcal{M}_1(D) \ldots, D))$ is $(\alpha, \sum_{r=1}^{R} \varepsilon_r)$-RDP。

**引理 B.4** (从 RDP 到差分隐私 ((Mironov, 2017年)))。 如果随机机制 $\mathcal{M}$ 是 $(\alpha, \varepsilon)$-*RDP*，那么 $\mathcal{M}$ 对于每个 $(\varepsilon + \log(1/\delta)/(\alpha - 1)\ \delta)$-*DP*， $\delta \in (0,1)$。

**定义 B.5** ($l_q$-敏感度)。 $\Delta_q(h) := \sup_{d_{\mathrm{H}}(D,D')=1} \|h(D) - h(D')\|_q$ 称为函数 $h$ 的 $l_q$-敏感度，其中最大值取遍任意相邻数据集 $D, D' \in \mathcal{D}$。

引理 *B.6*（高斯机制（（*Mironov*，*2017*年））。给定一个函数 $h$，高斯机制 $\mathcal{M}(D) := h(D) + \mathcal{N}(0, \sigma_1^2 I)$ 对每个 $\alpha \in (1, \infty)$ 都满足 $(\alpha, \alpha\Delta_2^2(h)/(2\sigma_1^2))$-*Rényi* 差分隐私。

引理 *B.7*（子采样放大（*Wang* 等人，*2019c* 年的定理 *9*））。设 $\mathcal{M}$ 是一个随机机制，它以包含 $b \le n$ 个点的数据集为输入，并且 $\gamma := b/n$。$\mathcal{M} \circ$ 子采样$_\gamma$ 定义如下：（*1*）子采样$_\gamma$：从大小为 $n$ 的输入数据集中无放回地子采样 $\gamma n$ 个点；（*2*）将子采样得到的点作为输入，应用 $\mathcal{M}$。对于每个整数 $\alpha \ge 2$，如果 $\mathcal{M}$ 是 $(\alpha, \varepsilon(\alpha))$-*Rényi* 差分隐私，那么 $\mathcal{M} \circ$ 子采样$_\gamma$ 是 $(\alpha, \varepsilon'(\alpha))$-*Rényi* 差分隐私，其中

$$\varepsilon'(\alpha) \le \frac{1}{\alpha - 1} \log \left( 1 + \gamma^2 \binom{\alpha}{2} \min \left\{ 4(e^{\varepsilon(2)} - 1), e^{\varepsilon(2)} \min\{2, (e^{\varepsilon(\infty)} - 1)^2\} \right\} + \sum_{j=3}^{\alpha} \gamma^j \binom{\alpha}{j} e^{(j-1)\varepsilon(j)} \min\{2, (e^{\varepsilon(\infty)} - 1)^j\} \right).$$

我们可以利用引理 B.7得到一个简单上界。

**引理 B.8** （子采样放大（简单上界））。 在与引理 *B.7*相同的设置下，若 $\varepsilon(\alpha)$ 关于 $\alpha$单调递增，则有

$$\varepsilon'(\alpha) \le \frac{2}{3} \left( 4 + \frac{e}{c} \right) \frac{\gamma^2 \alpha^2 \varepsilon(2)}{\alpha - 1}$$

在 $\varepsilon(\alpha) \le 1/3 \wedge \log(1/(2\gamma\alpha))$ 和 $\gamma \le \varepsilon(2)/(c\alpha)$ 条件下，对某个 $c > 0$ 成立。

证明。 由引理B.7，我们有

$$\varepsilon'(\alpha) \le \frac{1}{\alpha - 1} \log \left( 1 + 4\gamma^2 \binom{\alpha}{2} (e^{\varepsilon(2)} - 1) + 2 \sum_{j=3}^{\alpha} \gamma^j \binom{\alpha}{j} e^{(j-1)\varepsilon(j)} \right).$$

假设 $\varepsilon(\alpha) \le$ lo g$(1/(2\gamma\alpha))$. 那么，我们可以看到 $\gamma\alpha e^{\varepsilon(\alpha)} < 1/2$。第三项可以按如下方式界定 :

$$\begin{aligned} 2 \sum_{j=3}^{\alpha} \gamma^j \binom{\alpha}{j} e^{(j-1)\varepsilon(j)} &\le \frac{1}{3} \sum_{j=3}^{\alpha} (\gamma\alpha e^{\varepsilon(\alpha)})^j \\ &\le \frac{1}{3} \frac{(\gamma\alpha e^{\varepsilon(\alpha)})^3}{1 - \gamma\alpha e^{\varepsilon(\alpha)}} \\ &\le \frac{2(\gamma\alpha e^{\varepsilon(\alpha)})^3}{3} \end{aligned}$$

假设对于某个 $c > 0$，有 $\gamma \le \varepsilon(2)/(c\alpha)$。那么，我们有

$$2 \sum_{j=3}^{\alpha} \gamma^j \binom{\alpha}{j} e^{(j-1)\varepsilon(j)} \le \frac{2\gamma^2\alpha^2 e^{3\varepsilon(\alpha)} \varepsilon(2)}{3c}.$$

因此，我们得到

$$\begin{aligned}\varepsilon'(\alpha) &\le \frac{1}{\alpha-1}\log\left(1+\left(2(e^{\varepsilon(2)}-1)+\frac{2e^{3\varepsilon(\alpha)}\varepsilon(2)}{3c}\right)\gamma^2\alpha^2\right)\\ &\le \left(2(e^{\varepsilon(2)}-1)+\frac{2e^{3\varepsilon(\alpha)}\varepsilon(2)}{3c}\right)\frac{\gamma^2\alpha^2}{\alpha-1}\\ &\le \left(2\varepsilon(2)(1+\varepsilon(2))+\frac{2e\varepsilon(2)}{3c}\right)\frac{\gamma^2\alpha^2}{\alpha-1}.\end{aligned}$$

这里，对于第二个不等式，我们使用了 $\log(1+x)\le x$。最后一个不等式成立，是因为假设 $\varepsilon(2)\le\varepsilon(\alpha)\le 1/3\le 3/2$时，$e^x-1\le x+x^2$ for $x\le 3/2$。由此得到所需结果。 □

## C. DIFF2-GD 的差分隐私分析

在本节中，我们研究 DIFF2（算法 1）使用 GD-Routine（算法 3）时的差分隐私级别 .

首先，我们考虑 $D_p$下 $r\in[R]$ 的 $l_2$-敏感度，其中 $v_r$。

**引理 C.1** ($l_2$-敏感度 $v_r$)**.** 在算法 *1*中，$v_1$ 相对于 $D_p$具有 $l_2$-敏感度 $2C_1/(n_pP)$。此外，给定先前机制 $\{x_{r'},\widetilde{v}_{r'}\}_{r'\in[r-1]}$的输出，$v_r$ 相对于 $D_p$具有 $l_2$-敏感度 $2C_{2,\ r}/(n_pP)$。

证明。 当 $(r-1)\%T=0$ 时，对于相邻本地数据集 $D_p$和 $D'_p$，$v_1=(1/P)$ 的 $l_2$-敏感度 $\sum_{p=1}^P d_1^{(p)}$ 可界定为

$$\frac{1}{n_pP}\left\|\min\left\{\frac{C_1}{\|\nabla\ell(x_{r-1},z)\|},1\right\}\nabla\ell(x_{r-1},z)-\min\left\{\frac{C_1}{\|\nabla\ell(x_{r-1},z')\|},1\right\}\nabla\ell(x_{r-1},z')\right\|\le\frac{2C_1}{n_pP}$$

根据 $d_1^{(p)}$ 的定义，对于某个 $z\ne z'$，因为 $x_{r-1}$ 是固定的。

当 $(r-1)\%T\ne 0$ 时，对于相邻本地数据集 $D_p$和 $D'_p$，$v_r=(1/P)$ 的 $l_2$-敏感度 $\sum_{p=1}^P d_r^{(p)}+\widetilde{v}_{r-1}$ 可界定为

$$\begin{aligned}&\frac{1}{n_pP}\left\|\min\left\{\frac{C_{2,r}}{\|\nabla\ell(x_{r-1},z)-\nabla\ell(x_{r-2},z)\|},1\right\}(\nabla\ell(x_{r-1},z)-\nabla\ell(x_{r-2},z)\right.\\ &\left.-\min\left\{\frac{C_{2,r}}{\|\nabla\ell(x_{r-1},z')-\nabla\ell(x_{r-2},z')\|},1\right\}(\nabla\ell(x_{r-1},z')-\nabla\ell(x_{r-2},z')\right\|\le\frac{2C_{2,r}}{n_pP}\end{aligned}$$

对于某些e $z\ne z'$ 根据 $d_r^{(p)}$ 的定义，因为 $x_{r-1}$、$x_{r-2}$ 和 $\widetilde{v}_{r-1}$ 均为固定值。证明至此完成。 f. □

结合引理 C.1 与引理 B.6 和 B.3，我们可以证明命题 4.1。

**命题 4.1 的证明**

首先，对于每个机制 $\widetilde{v}_{k,\ r}$，我们推导其关于 $D_p$ 的 Rényi 差分隐私界。

当 $(r-1)\%T=0$ 时，根据引理 B.6，我们知道，对于每个 $\alpha\in(1,\infty)$，$\widetilde{v}_r$ 都满足 $(\alpha,2\alpha/(n_p^2P^2\sigma_1^2))$-Rényi 差分隐私，因为 $\xi_1\sim\mathcal{N}(0,\sigma_1^2C_1^2I)$，且根据引理 C.1 的第一部分，$v_1$ 的敏感度为 $l_2$-敏感度 $\Delta_2=2C_1/(n_pP)$。

同理，当 $(r-1)\%T\ne 0$ 时，根据引理 B.6，对于每个 $\alpha\in(1,\infty)$，我们知道 $\widetilde{v}_r$ 是 $(\alpha,2\alpha/(n_p^2P^2\sigma_2^2))$-Rényi 差分隐私，因为 $\xi_r\sim\mathcal{N}(0,\sigma_2^2C_{2,r}^2I)$，且根据引理 C.1 的第二部分，$v_r$ 的敏感度为 $l_2$-敏感度 $\Delta_2=2C_{2,r}/(n_pP)$。

因此，根据 $\{x_{r'},x_{r'}^{\text{out}},\widetilde{v}_{r'}\}_{r'\in[r-1]}$，$\{x_r,x_r^{\text{out}},\widetilde{v}_r\}$ 满足$(\alpha,\ 2\alpha/(n_p^2P^2\sigma^2))$-Rényi 差分隐私，这可由 Rényi 差分隐私的后处理性质（引理B.2）得出，因为 $x_r=x_r^{\text{out}}=x_{r-1}-\eta\widetilde{v}_r$，其中 $\sigma^2=\sigma_1^2$当$(r-1)\%T=0$时，以及 $\sigma^2=\sigma_2^2$当$(r-1)\%T\ne 0$。

现在，根据引理 B. 3, $\{x_r,x_r^{\text{out}},\widetilde{v}_r\}_{r\in[R]}$ $\alpha$ $2\alpha\lceil R/T\rceil/\ n_p^2P^2\sigma_1^2\ +\ 2\alpha\ R\ -$可知，该机制是关于 (, () ()())-Rényi 差分隐私机制。 $\lceil R/T\rceil\ /\ n_p^2P^2\sigma_2^2$ $D_p$

因此，令 $\alpha=1+\lceil 2\quad 1/\delta_{\text{DP}}\ /\varepsilon_{\text{DP}}\rceil$ $\sigma_1^2$ $\sigma_2^2$ $2\alpha\lceil R/T\rceil/\ n_p^2P^2\sigma_1^2\ +\log(\ )$，我们只需找到满足 ( ) ( ) ( )) $2\alpha\ R-\lceil R/T\rceil\ /\ n_p^2P^2\sigma_2^2\quad\le\ \varepsilon/2$ $u\ >\ 1$ $\sigma_1^2\ =$ 的 和。于是可以看出，对于任意常数 ，只需满足

$4u\alpha\lceil R/T\rceil/(n_{\min}^2P^2\varepsilon_{\mathrm{DP}})$ 和 $\sigma_2^2=(4u/(u-1))\alpha(R-\lceil R/T\rceil)/(n_{\min}^2P^2\varepsilon_{\mathrm{DP}})$ 保证机制 $\{x_r,x_r^{\mathrm{out}},\widetilde{v}_r\}_{r\in[R]}$ 满足 $(\varepsilon_{\mathrm{DP}},\delta_{\mathrm{DP}})$-差分隐私。这就是所需结果。

## D. DIFF2-GD 的效用分析

本节将对采用 GD-Routine（算法 1）的 DIFF2（算法 3）进行效用分析。**引理 D.1** (下降 Lemma)。假设条件 *2.1* 成立。在 $\eta\le 1/(2L)$下，*DIFF2*（算法 *1*）满足

$$\mathbb{E}[f(x_r)]\le\mathbb{E}[f(x_{r-1})]-\frac{\eta}{2}\mathbb{E}\|\nabla f(x_{r-1})\|^2-\frac{1}{4\eta}\mathbb{E}\|x_r-x_{r-1}\|^2+\eta\mathbb{E}\left\|\frac{1}{\eta}(x_{r-1}-x_r)-\nabla f(x_{r-1})\right\|^2.$$

对每一轮 $r\in[R]$。

证明。根据 $f$ 的 $L$-光滑性，我们有

$$\begin{aligned}f(x_r)&\le f(x_{r-1})+\langle\nabla f(x_{r-1}),x_r-x_{r-1}\rangle+\frac{L}{2}\|x_r-x_{r-1}\|^2\\&=f(x_{r-1})+\frac{1}{\eta}\left(-\frac{\eta^2}{2}\|\nabla f(x_{r-1})\|^2-\frac{1}{2}\|x_r-x_{r-1}\|^2+\frac{1}{2}\|x_r-x_{r-1}+\eta\nabla f(x_{r-1})\|^2\right)\\&\quad+\frac{L}{2}\|x_r-x_{r-1}\|^2\\&=f(x_{r-1})-\frac{\eta}{2}\|\nabla f(x_{r-1})\|^2-\left(\frac{1}{2\eta}-\frac{L}{2}\right)\|x_r-x_{r-1}\|^2+\frac{\eta}{2}\left\|\frac{1}{\eta}(x_{r-1}-x_r)-\nabla f(x_{r-1})\right\|^2.\end{aligned}$$

对随机性的所有历史取期望，并假设 $\eta\le 1/(2L)$，即可得到所需结果。

□

**命题 D.2.** 假设 *2.1*、*2.2* 和 *2.3* 均成立。假设 $C_1\ge G$且 $C_2\ge L$。那么，若适当选择 $\eta=\Theta(\min\{1/L,1/\sqrt{T\sigma_2^2C_2^2d}\}$，*DIFF2-GD* 满足

$$\mathbb{E}\|\nabla f(x^{\mathrm{out}})\|^2\le O\left(\frac{f(x_0)-f(x_*)}{\eta R}+\sigma_1^2C_1^2d\right).$$

证明。 对引理 D.1 中从 $r=1$到 $R$的不等式取平均，可得

$$\mathbb{E}\|\nabla f(x^{\mathrm{out}})\|^2=\frac{1}{R}\sum_{r=1}^R\mathbb{E}\|\nabla f(x_{r-1})\|^2\le\frac{2(f(x_0)-f(x_*))}{\eta R}-\frac{1}{2\eta^2}\frac{1}{R}\sum_{r=1}^R\mathbb{E}\|x_r-x_{r-1}\|^2+\frac{2}{R}\sum_{r=1}^R\mathbb{E}\left\|\frac{1}{\eta}(x_{r-1}-x_r)-\nabla f(x_{r-1})\right\|^2.\tag{1}$$

这里我们用到了最优解 $x_*$的存在性。

现在，我们考虑式（1）中的最后一项。由于 $x_r-x_{r-1}=-\eta\widetilde{v}_r$，我们有

$$\left\|\frac{1}{\eta}(x_{r-1}-x_r)-\nabla f(x_{r-1})\right\|^2=\|\widetilde{v}_r-\nabla f(x_{r-1})\|^2.$$

首先，我们考虑情形（$r-1$）%$T=0$。由于假设 $C_1\ge G$，因此有（$1/P$）$\sum_{p=1}^P d_r^{(p)}=\nabla f$（$x_{r-1}$），从而 $\widetilde{v}_r=\nabla f$（$x_{r-1}$）$+\xi_r$。于是，我们得到

$$\mathbb{E}\|\widetilde{v}_r-\nabla f(x_{r-1})\|^2=\mathbb{E}\|\xi_r\|^2=\sigma_1^2C_1^2d.$$

接下来，我们考虑情形 $(r-1)\%T \neq 0$。由于假设 $C_2 \geq L$，因此有 $(1/P)\sum_{p=1}^{P} d_r^{(p)} = \nabla f(x_{r-1}) - \nabla f(x_{r-2})$ $\widetilde{v}_r = \nabla f(x_{r-1}) - \nabla f(x_{r-2}) + \widetilde{v}_{r-1} + \xi_r$（），从而（）（）。接着，注意到

$$\begin{aligned}
\mathbb{E}\|\widetilde{v}_r - \nabla f(x_{r-1})\|^2 &= \mathbb{E}\|\widetilde{v}_{r-1} - \nabla f(x_{r-2}) + \xi_r\|^2 \\
&= \mathbb{E}\|\widetilde{v}_{r-1} - \nabla f(x_{r-2})\|^2 + \mathbb{E}\|\xi_r\|^2 \\
&= \sum_{r'=T(r)}^{r} \mathbb{E}\|\xi_{r'}\|^2 \\
&= \sigma_1^2 C_1^2 d + \sigma_2^2 d \sum_{r'=T(r)+1}^{r} C_{2,r}^2,
\end{aligned}$$

其中，$T(r)$ 是满足 $r+1-T \leq T(r) < r$ 且 $(T(r)-1)\%T = 0$ 的整数。这里，我们利用了以下事实：$\widetilde{v}_{T(r)} = \nabla f(x_{T(r)-1}) + \xi_{T(r)}$。利用这一关系，我们得到

$$\begin{aligned}
\frac{1}{R}\sum_{r=1}^{R} \mathbb{E}\|\widetilde{v}_r - \nabla f(x_{r-1})\|^2 &= \sigma_1^2 C_1^2 d + \sigma_2^2 C_2^2 d \frac{1}{R}\sum_{r=1}^{R}\sum_{r=T(r)+1}^{r} \|x_{r-1} - x_{r-2}\|^2 \\
&\leq \sigma_1^2 C_1^2 d + \sigma_2^2 C_2^2 d \frac{1}{R}\sum_{r=1}^{R}\sum_{r=T(r)}^{T(r)+T-1} \|x_{r-1} - x_{r-2}\|^2 \\
&\leq \sigma_1^2 C_1^2 d + T\sigma_2^2 C_2^2 d \frac{1}{R}\sum_{r=1}^{R} \|x_{r-1} - x_{r-2}\|^2.
\end{aligned}$$

假设 $\eta \leq 1/(2\sqrt{T\sigma_2^2 C_2^2 d})$。则有

$$-\frac{1}{2\eta^2}\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|x_r - x_{r-1}\|^2 + \frac{2}{R}\sum_{r=1}^{R}\mathbb{E}\left\|\frac{1}{\eta}(x_{r-1} - x_r) - \nabla f(x_{r-1})\right\|^2 \leq 2\sigma_1^2 C_1^2 d.$$

将其应用于 (1)，可得

$$\mathbb{E}\|\nabla f(x^{\text{out}})\|^2 \leq O\left(\frac{f(x_0) - f(x_*)}{\eta R} + \sigma_1^2 C_1^2 d\right).$$

这就是所需结果。 □

**定理 4.3 的证明**

由命题 D.2，在 $C_1 = \Theta(G)$下，我们知道

$$\mathbb{E}\|\nabla f(x^{\text{out}})\|^2 \leq O\left(\frac{f(x_0) - f(x_*)}{\eta R} + \sigma_1^2 G^2 d\right).$$

假设 $f(x_0) - f(x_*) = O(1)$。

根据命题 4.1，对于 $(\varepsilon_{\text{DP}}, \delta_{\text{DP}})$-差分隐私，关于 $D_p$，所需的噪声水平为 $\sigma_1^2 = \widetilde{\Theta}(R/(Tn_{\min}^2 P^2 \varepsilon_{\text{DP}}^2))$ 和 $\sigma_2^2 = \Theta(R/(n_{\min}^2 P^2 \varepsilon_{\text{DP}}^2))$。分别利用这些结果并代入 $\eta$的定义，可得

$$\begin{aligned}
\mathbb{E}\|\nabla f(x^{\text{out}})\|^2 &\leq O\left(\frac{1}{\eta R}\right) + \widetilde{O}\left(\frac{RG^2 d}{Tn_{\min}^2 P^2 \varepsilon_{\text{DP}}^2}\right) \\
&= O\left(\frac{L + \sqrt{T}L\sigma_2\sqrt{d}}{R}\right) + \widetilde{O}\left(\frac{RG^2 d}{Tn_{\min}^2 P^2 \varepsilon_{\text{DP}}^2}\right) \\
&= O\left(\frac{L}{R}\right) + \widetilde{O}\left(\frac{\sqrt{T}L\sqrt{d}}{n_{\min} P \varepsilon_{\text{DP}}\sqrt{R}} + \frac{RG^2 d}{Tn_{\min}^2 P^2 \varepsilon_{\text{DP}}^2}\right).
\end{aligned}$$

假设 $n_{\min}P = \Omega(G^2\sqrt{d}/(L\varepsilon_{\mathrm{DP}}))$。我们定义

$$T := \Theta\left(1 \vee \left(\frac{G^2\sqrt{d}}{Ln_{\min}P\varepsilon_{\mathrm{DP}}}\right)^{\frac{2}{3}} R\right)$$

其中 $1 \le T \le R$。

然后，我们得到

$$\mathbb{E}\|\nabla f(x^{\mathrm{out}})\|^2 \le O\left(\frac{L}{R}\right) + \widetilde{O}\left(\frac{L\sqrt{d}}{n_{\min}P\varepsilon_{\mathrm{DP}}\sqrt{R}} + \frac{(LGd)^{\frac{2}{3}}}{(n_{\min}P\varepsilon_{\mathrm{DP}})^{\frac{4}{3}}}\right).$$

最后，令

$$R = \Theta\left(1 \vee \frac{L}{\varepsilon_{\mathrm{opt}}}\right) \vee \widetilde{\Theta}\left(\frac{L^2 d}{n_{\min}^2 P^2 \varepsilon_{\mathrm{DP}}^2 \varepsilon_{\mathrm{opt}}^2}\right)$$

其中 $\varepsilon_{\mathrm{opt}} := \Theta\left(\frac{(LGd)^{\frac{2}{3}}}{(n_{\min}P\varepsilon_{\mathrm{DP}})^{\frac{4}{3}}}\right)$ 即可得到所需结果。

## E. DIFF2-BVR-L-SGD 的差分隐私分析

本节对采用 BVR-L-SGD-Routine（算法 1）的 DIFF2（算法 4）进行差分隐私分析。我们研究满足 $\sigma_1^2$， $\sigma_2^2$和 $\sigma_3^2$所需的最小噪声水平 $(\varepsilon_{\mathrm{DP}}, \delta_{\mathrm{DP}})$-差分隐私，相对于 $D_p$，对于 $p \in [P]$。

由于我们可以利用引理 C.1 来确定 $l_2$ 的敏感度，因此我们重点关注 $v_r$的 $l_2$-敏感度，$v_{k,\,r}$ 对于 $k \in [K]$ 和 $r \in [R]$。**引理 E.1** ($l_2$-敏感度：$v_{k,\,r}$)。 给定先前机制 $\{x_{r-1}, \widetilde{v}_r\} \cup \{x_{k',r-1}\}_{k' \in [k-1]}$的输出，$l_2$由算法 $v_k$定义的 $r$ 的敏感度为 *4* 如下

$$\begin{cases} \frac{2C_{3,k,r}}{b} & (p = p_r), \\ 0 & (p \ne p_r). \end{cases}$$

相对于 $D_p$，对于 $k \ge 2$。

证明。 $l_2$相邻本地数据集 $D_p$ 和 $D_p'$ 下，$v_k$的敏感度可界定为$r$：

$$\begin{aligned} &\frac{1}{b}\Bigg\|\min\left\{\frac{C_{3,k,r}}{\|\nabla\ell(x_{k-1,r-1},z) - \nabla\ell(x_{k-2,r-1},z)\|}, 1\right\}(\nabla\ell(x_{k-1,r-1},z) - \nabla\ell(x_{k-2,r-1},z) \\ &- \min\left\{\frac{C_{3,k,r}}{\|\nabla\ell(x_{k-1,r-1},z') - \nabla\ell(x_{k-2,r-1},z')\|}, 1\right\}(\nabla\ell(x_{k-1,r-1},z') - \nabla\ell(x_{k-2,r-1},z')\Bigg\| \le \frac{2C_{3,k,r}}{b} \end{aligned}$$

对于某个 $z \ne z'$，根据 $v_{k\,r}^{(p)}$ 的定义，由于$x_{k-1,\,r-1}$、 $x_{k-2,\,r-1}$ 和 $\widetilde{v}_{k-1,\,r}$ 均为固定值。当 $p \ne p_r$时，$v_{k,\,r}$ 关于 $p$ 的敏感度显然为零。证明完毕。 □

结合引理 C.1 和 E.1 以及引理 B.6 和 B.3，可得以下命题：

命题 *E.2*（$(\varepsilon_{\mathrm{DP}}, \delta_{\mathrm{DP}})$-差分隐私的最小噪声水平）。对于任意整数 $\alpha \ge 2$，*DIFF2-BVR-L-SGD* 中定义的机制 $\{x_r, x_r^{\mathrm{out}}, \widetilde{v}_r\}_{r\in[R]}$ 满足 $(2\alpha\lceil R/T\rceil/(n_p^2P^2\sigma_1^2)+2\alpha(R-\lceil R/T\rceil)/(n_p^2P^2\sigma_2^2)+K\lceil R/P\rceil\varepsilon'(\alpha)+\log(1/\delta_{\mathrm{DP}})/(\alpha-1)$ $\delta_{\mathrm{DP}})$-差分隐私，其中 $\varepsilon'(\alpha)$ 在引理 *B.7* 中定义，且 $\varepsilon(\alpha) := 2\alpha/(b^2\sigma_3^2)$， $n_{\min} := \min\{n_p\}_{p=1}^P$。特别地，令$\alpha := 1 + \lceil 2\log(1/\delta_{\mathrm{DP}})/\varepsilon_{\mathrm{DP}}\rceil$，

$$\sigma_1^2 = \frac{4u_1\alpha\lceil\frac{R}{T}\rceil}{n_{\min}^2P^2\varepsilon_{\mathrm{DP}}},\ \sigma_2^2 = \frac{4u_2\alpha(R-\lceil\frac{R}{T}\rceil)}{n_{\min}^2P^2\varepsilon_{\mathrm{DP}}} \text{ and } 2K\left\lceil\frac{R}{P}\right\rceil\varepsilon'(\alpha) \le \left(1 - \frac{1}{u_1} - \frac{1}{u_2}\right)\varepsilon_{\mathrm{DP}}$$

是保证任意常数 $u_1$、 $u_2 > 1$ 下满足$(\varepsilon_{\mathrm{DP}}, \delta_{\mathrm{DP}})$-差分隐私的充分条件。

此外，存在满足上述条件的 $\sigma_1^2$、 $\sigma_2^2$ 和 $\sigma_3^2$，使得

$$\sigma_1^2 = \widetilde{\Theta}\left(\frac{R}{Tn_{\min}^2P^2\varepsilon_{\mathrm{DP}}^2}\right), \sigma_2^2 = \widetilde{\Theta}\left(\frac{R}{n_{\min}^2P^2\varepsilon_{\mathrm{DP}}^2}\right) \text{ and } \sigma_3^2 = \widetilde{\Theta}\left(\frac{KR}{n_{\min}^2P\varepsilon_{\mathrm{DP}}^2} \vee \frac{1}{b^2\varepsilon_{\mathrm{DP}}}\right)$$

在 $b \le \min\{n_{\min}/(2e\alpha), (4n_{\min}/(\alpha\sigma_3^2))^{1/3}\}$。

证明。 首先，根据引理 C.1 的论证，我们知道 $\widetilde{v}_{1,\ r} = \widetilde{v}_r$ 因此 $x_{1,\ r-1}$ 在 $(r-1)\%T = 0$ 时是 $(\alpha,\ 2\alpha/(n_p^2P^2\sigma_1^2))$-Rényi 差分隐私，而在 $(r-1)\%T \neq 0$ 时是 $(\alpha,\ 2\alpha/(n_p^2P^2\sigma_2^2))$-Rényi 差分隐私；对于每个 $\alpha \in (1,\ \infty)$，以上结论均在给定先前机制的输出时成立。

接下来，若 $p = p_r$，根据引理 B.6 和 E.1，对于子采样小批量，有 $\widetilde{v}_{k,r}$ 对于 $k \ge 2$ 是 $(\alpha, \varepsilon(\alpha))$-Rényi 差分隐私，其中 $\varepsilon(\alpha)$: $= 2\alpha/(b^2\sigma_3^2)$ 对每个 $\alpha \in (1, \infty)$ 均成立，因为 $\xi_{2,r} \sim \mathcal{N}(0, \sigma_3^2C_{3,k,r}^2I)$。然后，应用引理 B.7 可知，对于整数 $\alpha \ge 2$，$\widetilde{v}_{k,r}$ 是 $(\alpha, \varepsilon'(\alpha))$-Rényi 差分隐私，其中 $\varepsilon'(\alpha)$ 根据给定的先前机制的输出，按引理 B.7 定义。由此，根据引理 B.2，在给定先前机制的输出时，$\{x_{k,r-1}, \widetilde{v}_{k,r}\}$ 对于 $k \ge 2$ 也满足 $(\alpha, \varepsilon'(\alpha))$-Rényi 差分隐私。

当 $p \neq p_r$, $\widetilde{v}_{k,\ r}$ 因此 $x_{k,\ r-1}$ 关于 $D_p$ 对于 $k \ge 2$，显然是 $(\alpha,\ 0)$-Rényi 差分隐私。

现在，根据引理 B. 3, $\{x_{k,r-1}\}_{k\in[K]\cup\{0\}\ r\in[R]} \cup \{\widetilde{v}_{k,r}\}_{k\in[K]\ r\in[R]}$ $\alpha$ $2\alpha\lceil R/T\rceil/\ n_p^2P^2\sigma_1^2\ +$可知，机制 ,,关于 $D_p$满足 (, $^{2\alpha}$ ( ) $(R - \lceil R/T\rceil)/(n_p^2P^2\sigma_2^2) + K\lceil R/P\rceil\varepsilon'^{(\alpha}$ ))-Rényi 差分隐私。

然后，利用引理 B. 4, $\{x_{k,r-1}\}_{k\in[K]\cup\{0\}\ r\in[R]}\cup\{\widetilde{v}_{k,r}\}_{k\in[K]\ r\in[R]}$ $2\alpha\lceil R/T\rceil/\ n_p^2P^2\sigma_1^2\ +$可以看出，

$2\alpha\ R - \lceil R/T\rceil\ /\ n_p^2P^2\sigma_2^2\ + K\lceil R/P\rceil\varepsilon'\ \alpha\ +\ 1/\delta_{\mathrm{DP}}\ /\ \alpha - 1\ \ \delta_{\mathrm{DP}}$ $\alpha \ge 2$

对于每个整数，机制 ,,满足 ( ( ) ( ) ( ) ( ) log( ) ( ) )-差分隐私。注意 $\{x_r, x_r^{\mathrm{out}}, \widetilde{v}_r\}_{r\in[R]} \subset \{x_{k,r-1}\}_{k\in[K]\cup\{0\},\ r\in[R]} \cup \{\widetilde{v}_{k,r}\}_{k\in[K],\ r\in[R]}$。

现在，选择 $\alpha := 1 + \lceil 2\log(1/\delta_{\mathrm{DP}})/\varepsilon_{\mathrm{DP}}\rceil$,

$$\sigma_1^2 = \frac{4u_1\alpha\lceil\frac{R}{T}\rceil}{n_{\min}^2P^2\varepsilon_{\mathrm{DP}}} = \widetilde{\Theta}\left(\frac{R}{Tn_{\min}^2P^2\varepsilon_{\mathrm{DP}}^2}\right),$$

$$\sigma_2^2 = \frac{4u_2\alpha(R - \lceil\frac{R}{T}\rceil)}{n_{\min}^2P^2\varepsilon_{\mathrm{DP}}} = \widetilde{\Theta}\left(\frac{R}{n_{\min}^2P^2\varepsilon_{\mathrm{DP}}^2}\right),$$

and

$$2K\left\lceil\frac{R}{P}\right\rceil\varepsilon'(\alpha) \le \left(1 - \frac{1}{u_1} - \frac{1}{u_2}\right)\varepsilon_{\mathrm{DP}}$$

对于任意常数 $u_1$, $u_2 > 1$，机制 $\{x_r, x_r^{\mathrm{out}}, \widetilde{v}_r\}_{r\in[R]}$ 可保证 $(\varepsilon_{\mathrm{DP}}, \delta_{\mathrm{DP}})$。

最后，我们推导出 $\sigma_3^2$的一个简单上界。根据引理B.8，我们知道

$$\varepsilon'(\alpha) \le \frac{2}{3}(4 + e)\frac{\gamma^2\alpha^2\varepsilon(2)}{\alpha - 1}$$

在 $\varepsilon(\alpha) \le 1/3 \wedge \log(1/(2\gamma\alpha))$ 和 $\gamma \le \varepsilon(2)/\alpha$[14] 条件下。

假设

$$\sigma_3^2 \ge \frac{8}{3}(4 + e)\frac{\alpha^2}{\alpha - 1}\frac{6K\lceil\frac{R}{P}\rceil}{n_{\min}^2\varepsilon_{\mathrm{DP}}} \vee \frac{6\alpha}{b^2}$$

and

$$b \le \frac{n_{\min}}{2e\alpha} \wedge \left(\frac{4n_{\min}}{\alpha\sigma_3^2}\right)^{\frac{1}{3}}.$$

[14]在此，我们针对简单表示在引理 $c = 1$B.8 中设定 $c < 1$ 利用 $b$ 有助于放宽 $b$ 这一条件，但代价是噪声水平会升高。

由于 $2\gamma\alpha \le 1/e$，可得 $\log(1/(2\gamma\alpha)) \ge 1$。因此，满足 $\varepsilon(\alpha) = 2\alpha/(b^2\sigma_3^2) \le 1/3 \wedge \log(1/(2\gamma\alpha))$。接下来，注意到 $\gamma \le b/n_{\min} \le 4/(\alpha b^2\sigma_3^2) = \varepsilon(2)/\alpha$ 成立。最后，

$$\begin{aligned}\varepsilon'(\alpha) &\le \frac{2}{3}(4+e)\frac{\gamma^2\alpha^2\varepsilon(2)}{\alpha-1}\\ &\le \frac{8}{3}(4+e)\frac{\alpha^2}{\alpha-1}\frac{1}{n_{\min}^2\sigma_3^2},\end{aligned}$$

因此，满足 $6K\lceil R/P\rceil\varepsilon'(\alpha) \le \varepsilon_{\mathrm{DP}}$。

因此，

$$\sigma_3^2 \ge \widetilde{\Theta}\left(\frac{KR}{n_{\min}^2 P\varepsilon_{\mathrm{DP}}^2} \vee \frac{1}{b^2\varepsilon_{\mathrm{DP}}}\right).$$

足以满足 $6K\lceil R/P\rceil\varepsilon'(\alpha) \le \varepsilon_{\mathrm{DP}}$。

至此，证明完毕。

□

## F. DIFF2 与 BVR-L-SGD 例程的效用分析

本节将分析带有 BVR-L-SGD 例程（算法1）的 DIFF2（算法 4）的效用。

首先，我们着重分析 BVR-L-SGD 例程（算法 4）。

**引理 F.1**（下降引　理）。假设条件 *2.1* 成立。在 $\eta \le 1/(2L)$ 下，*BVR-L-SGD* 例程满足　　*s*

$$\begin{aligned}\mathbb{E}[f(x_{k,r-1})] \le \mathbb{E}[f(x_{k-1,r-1})] &- \frac{\eta}{2}\mathbb{E}\|\nabla f(x_{k-1,r-1})\|^2 - \frac{1}{4\eta}\mathbb{E}\|x_{k,r-1} - x_{k-1,r-1}\|^2\\ &+ \eta\mathbb{E}\|\widetilde{v}_{k,r} - \nabla f(x_{k-1,r-1})\|^2,\end{aligned}$$

对于每一轮 $r \in [R]$和迭代 $k \in [K]$。

证明。为便于简单表示，$x_k$、$\xi_k$、$p$、$v_k$ 和 $\widetilde{v}_k$ 在本证明中分别表示 $x_{k,\,r-1}$、$\xi_{k,\,r}$、$p_r$ $v_{k,\,r}$ 和 $\widetilde{v}_{k,\,r}$。令 $\eta$ 在 $k \in \{2,\dots,K\}$ 时为 $\eta$，在 $k = 1$时为 $\eta$。

根据 $f$ 的 $L$-光滑性，我们有

$$\begin{aligned}f(x_k) &\le f(x_{k-1}) + \langle\nabla f(x_{k-1}), x_k - x_{k-1}\rangle + \frac{L}{2}\|x_k - x_{k-1}\|^2\\ &= f(x_{k-1}) + \frac{1}{\eta}\left(-\frac{\eta^2}{2}\|\nabla f(x_{k-1})\|^2 - \frac{1}{2}\|x_k - x_{k-1}\|^2 + \frac{1}{2}\|x_k - x_{k-1} + \eta\nabla f(x_{k-1})\|^2\right)\\ &\quad + \frac{L}{2}\|x_k - x_{k-1}\|^2\\ &= f(x_{k-1}) - \frac{\eta}{2}\|\nabla f(x_{k-1})\|^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\|x_k - x_{k-1}\|^2 + \frac{\eta}{2}\|\widetilde{v}_k - \eta\nabla f(x_{k-1})\|^2.\end{aligned}$$

对等式两边关于第 $k$ 次迭代中的随机性取期望，

$$\mathbb{E}[f(x_k)] \le f(x_{k-1}) - \frac{\eta}{2}\|\nabla f(x_{k-1})\|^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\mathbb{E}\|x_k - x_{k-1}\|^2 + \frac{\eta}{2}\mathbb{E}\|\widetilde{v}_k - \eta\nabla f(x_{k-1})\|^2.$$

对所有历史随机性取期望，在 $\eta \le 1/(2L)$ 下，我们得到

$$\mathbb{E}[f(x_k)] \le \mathbb{E}[f(x_{k-1})] - \frac{\eta}{2}\mathbb{E}\|\nabla f(x_{k-1})\|^2 - \frac{1}{4\eta}\mathbb{E}\|x_k - x_{k-1}\|^2 + \frac{\eta}{2}\mathbb{E}\|\widetilde{v}_k - \nabla f(x_{k-1})\|^2.$$

这就是所需的结果。□

命题 *F.2.* 假设条件 *5.2* 和 *2.1* 成立。那么，如果 $C_3 \geq L$ *with* $C_3 = \Theta(L)$ *and* $\eta \leq \min\{1/(2L)\ 1/(2\sqrt{2e(K^2\eta^2 + KL^2/b + KC_3^2\sigma_3^2 d)})\} = \Theta(\min\{1/L, 1/(K\zeta)\ \sqrt{b}/(\sqrt{KL})\ 1/(\sqrt{KL}\sigma_3\sqrt{d})\})$*, BVR- L-SGD-Routine*（算法 *4*）满足

$$\mathbb{E}\|\nabla f(x_r^{\text{out}})\|^2 \leq \frac{2(\mathbb{E}[f(x_{r-1})] - \mathbb{E}[f(x_r)])}{\eta K} - \frac{1}{4\eta^2}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|x_{k,r-1} - x_{k-1,r}\|^2 + 2e\mathbb{E}\|\widetilde{v}_r - \nabla f(x_{r-1})\|^2.$$

对于每一轮 $r \in [R]$。

证明。 为便于简单表示， $x_{k,r-1}$、 $\xi_{k,r}$、 $p_r$、 $v_{k,r}$ 和 $\widetilde{v}_{k,r}$分别表示$x_{k,r-1}$、 $\xi_{k,r}$、 $p_r$ $v_{k,r}$ 和$\widetilde{v}_{k,r}$ 在本证明中，将引理F.1 从 $k = 1$到 $K$ 中的不等式求和可得

$$\mathbb{E}\|\nabla f(x_{\hat{k}-1})\|^2 = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\nabla f(x_{k-1})\|^2 \leq \frac{2(\mathbb{E}[f(x_0)] - \mathbb{E}[f(x_K)])}{\eta K} - \frac{1}{2\eta^2}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|x_k - x_{k-1}\|^2 + \frac{2}{K}\sum_{k=1}^{K}\mathbb{E}\|\widetilde{v}_k - \nabla f(x_{k-1})\|^2. \tag{2}$$

现在，我们界定 $\mathbb{E}\|\widetilde{v}_k - \nabla f(x_{k-1})\|^2$ 对于 $k \geq 2$的范围。由于我们假设 $C_2 \geq L$，因此在 $L$-光滑性条件下，对于 $\ell$始终有 $C_2\|x_{k-1} - x_{k-2}\|/\|\nabla\ell(x_{k-1,} z_l) - \nabla\ell(x_{k-2,} z_l)\| \geq 1$，从而有

$$v_k = g_k(x_{k-1}) - g_k(x_{k-2}) + \widetilde{v}_{k-1},$$

其中， $g_k(x_{k-1}) := \frac{1}{b}\sum_{z\in I_{k,r}}\nabla\ell(x_{k-1,} z)$，以及 $g_k(x_{k-2}) := \frac{1}{b}\sum_{z\in I_{k,r}}\nabla\ell(x_{k-2,} z)$。因此，我们有

$$\begin{aligned}
&\mathbb{E}\|\widetilde{v}_k - \nabla f(x_{k-1})\|^2 \\
=&\mathbb{E}\|v_k - \nabla f(x_{k-1})\|^2 + \mathbb{E}\|\xi_k\|^2 \\
=&\mathbb{E}\|g_k(x_{k-1}) - g_k(x_{k-2}) + \widetilde{v}_{k-1} - \nabla f(x_{k-1})\|^2 + \mathbb{E}\|\xi_k\|^2 \\
=&\mathbb{E}\|\nabla f_p(x_{k-1}) - \nabla f_p(x_{k-2}) + \widetilde{v}_{k-1} - \nabla f(x_{k-1})\|^2 \\
&+ \mathbb{E}\|g_k(x_{k-1}) - g_k(x_{k-2}) - \nabla f_p(x_{k-1}) + \nabla f_p(x_{k-2})\|^2 + \mathbb{E}\|\xi_k\|^2 \\
\leq&(1 + 1/K)\mathbb{E}\|\widetilde{v}_{k-1} - \nabla f(x_{k-2})\|^2 \\
&+ (1 + K)\mathbb{E}\|\nabla f_p(x_{k-1}) - \nabla f_p(x_{k-2}) - \nabla f(x_{k-1}) + \nabla f(x_{k-2})\|^2 \\
&+ \mathbb{E}\|g_k(x_{k-1}) - g_k(x_{k-2}) - \nabla f_p(x_{k-1}) + \nabla f_p(x_{k-2})\|^2 \\
&+ \mathbb{E}\|\xi_k\|^2.
\end{aligned}$$

第二项可按如下方式进行界定：

$$\begin{aligned}
&\mathbb{E}\|\nabla f_p(x_{k-1}) - \nabla f_p(x_{k-2}) - \nabla f(x_{k-1}) + \nabla f(x_{k-2})\|^2 \\
=&\mathbb{E}\|\nabla f_p(x_{k-1}) - \nabla f_p(x_{k-2}) - \nabla f(x_{k-1}) + \nabla f(x_{k-2})\|^2 \\
=&\mathbb{E}\left\|\int_0^1 \nabla^2 f_p((1-\theta)x_{k-1} + \theta x_{k-2})(x_{k-1} - x_{k-2})d\theta - \int_0^1 \nabla^2 f((1-\theta)x_{k-1} + \theta x_{k-2})(x_{k-1} - x_{k-2})d\theta\right\|^2 \\
\leq&\mathbb{E}\int_0^1 \left\|(\nabla^2 f_p((1-\theta)x_{k-1} + \theta x_{k-2}) - \nabla^2 f((1-\theta)x_{k-1} + \theta x_{k-2}))(x_{k-1} - x_{k-2})\right\|^2 d\theta \\
\leq&\zeta^2\mathbb{E}\|x_{k-1} - x_{k-2}\|^2.
\end{aligned}$$

这里，对于第二在不等式中，我们使用了中值定理。最后一个不等式由假设5成立。 2.

第三项可按如下方式界定：

$$\begin{aligned}&\mathbb{E}\|g_k(x_{k-1}) - g_k(x_{k-2}) - \nabla f_p(x_{k-1}) + \nabla f_p(x_{k-2})\|^2\\ =\ &\frac{1}{b}\mathbb{E}[\mathbb{E}_{z\sim D_p}\|\nabla \ell(x_{k-1}, z) - \nabla \ell(x_{k-2}, z) - \nabla f_p(x_{k-1}) + \nabla f_p(x_{k-2})\|^2]\\ \le\ &\frac{L^2}{b}\mathbb{E}\|x_{k-1} - x_{k-2}\|^2.\end{aligned}$$

由于 $\xi_k \sim \mathcal{N}(0,\ \sigma_3^2 C_3^2\|x_{k-1} - x_{k-2}\|^2 I)$，在给定迭代 $k-1$ 的条件下，我们有 $\mathbb{E}\|\xi_k\|^2 = C_3^2\sigma_3^2\|x_{k-1} - x_{k-2}\|^2 d$。因此，我们得到

$$\begin{aligned}\mathbb{E}\|\widetilde{v}_k - \nabla f(x_{k-1})\|^2 \le (1 + 1/K)\mathbb{E}\|\widetilde{v}_{k-1} - \nabla f(x_{k-2})\|^2 + \left((K+1)\zeta^2 + \frac{L^2}{b}\right)\mathbb{E}\|x_{k-1} - x_{k-2}\|^2\\ + C_3^2\sigma_3^2 d\mathbb{E}\|x_{k-1} - x_{k-2}\|^2.\end{aligned}$$

递归地应用此不等式可得

$$\begin{aligned}\mathbb{E}\|\widetilde{v}_k - \nabla f(x_{k-1})\|^2 &\le (1 + 1/K)^{k-1}\mathbb{E}\|\widetilde{v}_1 - \nabla f(x_0)\|^2\\ &\quad + \left(K\zeta^2 + \frac{L^2}{b} + C_3^2\sigma_3^2 d\right)\sum_{k'=2}^{k-1}(1 + 1/(K-1))^{k-1-k'}\mathbb{E}\|x_{k'-1} - x_{k'-2}\|^2\\ &\le e\mathbb{E}\|\widetilde{v}_1 - \nabla f(x_0)\|^2 + e\left(K\zeta^2 + \frac{L^2}{b} + C_3^2\sigma_3^2 d\right)\sum_{k=2}^{K}\mathbb{E}\|x_{k-1} - x_{k-2}\|^2.\end{aligned}$$

将此不等式与（2）结合可得

$$\begin{aligned}\mathbb{E}\|\nabla f(x_{\hat{k}-1})\|^2 = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\nabla f(x_{k-1})\|^2 \le \frac{2(\mathbb{E}[f(x_0)] - \mathbb{E}[f(x_K)])}{\eta K} - \frac{1}{2\eta^2}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|x_k - x_{k-1}\|^2\\ + 2e\mathbb{E}\|\widetilde{v}_1 - \nabla f(x_0)\|^2\\ + 2e\left(K^2\zeta^2 + \frac{KL^2}{b} + KC_3^2\sigma_3^2 d\right)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|x_k - x_{k-1}\|^2.\end{aligned}$$

假设 $\eta \le 1/(2\sqrt{2e(K^2\zeta^2 + KL^2/b + K\sigma_3^2 C_3^2 d)})$，这意味着

$$\frac{1}{4\eta^2} \ge 2e\left(K^2\zeta_2^2 + \frac{KL^2}{b} + KC_2^2\sigma_2^2 d\right).$$

那么，我们有

$$\mathbb{E}\|\nabla f(x_{\hat{k}-1})\|^2 \le \frac{2(\mathbb{E}[f(x_0)] - \mathbb{E}[f(x_K)])}{\eta K} - \frac{1}{4\eta^2}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|x_k - x_{k-1}\|^2 + 2e\mathbb{E}\|\widetilde{v}_1 - \nabla f(x_0)\|^2.$$

这就是预期的结果。 □

命题 *F.3*。假设假设 *2.1*、*2.2*、*2.3* 和 *5.2* 成立。如果 $C_1 \ge G$ *with* $C_1 = \Theta(G)$，$C_2$，$C_3 \ge L$ *with* $C_2$，$C_3 = \Theta(L)$，并且我们适当选择 $\eta = \Theta(\min\{1/L, 1/(K\zeta)\ \sqrt{b}/(\sqrt{KL})\ 1/(K\sqrt{TL}\sigma_2\sqrt{d})\ 1/(\sqrt{KL}\sigma_3\sqrt{d})\}$，则 *DIFF2-BVR-L-SGD* 满足

$$\mathbb{E}\|\nabla f(x^{\text{out}})\|^2 \le O\left(\frac{f(x_0 - f(x_*)}{\eta KR} + \sigma_1^2 C_1^2 d\right).$$

证明。 根据命题 F.2，我们知道

$$\mathbb{E}\|\nabla f(x_r^{\text{out}})\|^2 \leq \frac{2(\mathbb{E}[f(x_{r-1})] - \mathbb{E}[f(x_r)])}{\eta K} - \frac{1}{4\eta^2}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|x_{k,r-1} - x_{k-1,r}\|^2 + 2e\mathbb{E}\|\widetilde{v}_r - \nabla f(x_{r-1})\|^2.$$

将该不等式从 $r=1$ 到 $R$ 取平均，可得

$$\begin{aligned}\mathbb{E}\|\nabla f(x^{\text{out}})\|^2 = \frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\nabla f(x_{\hat{r}-1})\|^2 \leq \frac{2(f(x_0) - \mathbb{E}[f(x_*)])}{\eta K R} - \frac{1}{4\eta^2}\frac{1}{KR}\sum_{r=1}^{R}\sum_{k=1}^{K}\mathbb{E}\|x_{k,r-1} - x_{k-1,r}\|^2 \\ + \frac{2e}{R}\sum_{r=1}^{R}\mathbb{E}\|\widetilde{v}_r - \nabla f(x_{r-1})\|^2.\end{aligned}$$

我们对最后一项进行界定。类似于命题 D.2 证明中的论证，我们可以证明

$$\frac{1}{R}\sum_{r=1}^{R}\left\|\widetilde{v}_r - \nabla f(x_{r-1})\right\|^2 \leq \sigma_1^2 C_1^2 d + T\sigma_2^2 C_2^2 d\frac{1}{R}\sum_{r=1}^{R}\|x_{r-1} - x_{r-2}\|^2.$$

此外，由于 $\|x_r - x_{r-1}\|^2 \leq K\sum_{k=1}^{K}\|x_{k,\ r-1} - x_{k-1,\ r-1}\|^2$，我们有

$$\begin{aligned}\mathbb{E}\|\nabla f(x^{\text{out}})\|^2 = \frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\nabla f(x_{\hat{r}-1})\|^2 \leq \frac{2(f(x_0) - f(x_*))}{\eta K R} - \frac{1}{4\eta^2}\frac{1}{KR}\sum_{r=1}^{R}\sum_{k=1}^{K}\mathbb{E}\|x_{k,r-1} - x_{k-1,r}\|^2 \\ + 2e\left(\sigma_1^2 C_1^2 d + K^2 T\sigma_2^2 C_2^2 d\frac{1}{KR}\sum_{r=1}^{R}\sum_{k=1}^{K}\mathbb{E}\|x_{k,r-1} - x_{k-1,r}\|^2\right).\end{aligned}$$

因此，在 $\eta \leq 1/(2\sqrt{2e\overline{K^2 T\sigma_2^2 C_2^2 d}})$ 下，这意味着

$$\frac{1}{4\eta^2} \geq 2eK^2 T\sigma_2^2 C_2^2 d$$

利用这一事实，并将上述不等式从 $r=1$ 到 $R$ 取平均，我们得到

$$\mathbb{E}\|\nabla f(x^{\text{out}})\|^2 \leq \frac{2(f(x_0 - f(x_*))}{\eta K R} + 2e\sigma_1^2 C_1^2 d.$$

这就是所需的结果。 □

定理 *F.4*（效用界）。假设假设 *2.1*、*2.2*、*2.3* 和 *5.2* 成立。假设 $C_1 \geq G$ *with* $C_1 = \Theta(G)$、$C_2$、$C_3 \geq L$ *with* $C_2$、$C_3 = \Theta(L)$、$f(x_0) - f(x_*) = \Theta(1)$ 以及 $n_{\min}P = \Omega(G^2\sqrt{d}/(L\varepsilon_{\text{DP}}))$。此外，假设命题 *E.2* 中的 $b$ 条件成立。在命题 *E.2* 中对 $\sigma_1^2$、$\sigma_2^2$ 和 $\sigma_3^2$ 进行选择后，若我们适当地选择 $\eta = \Theta(\min\{1/L, 1/(K\zeta)\ \sqrt{b}/(\sqrt{KL})\ 1/(K\sqrt{TL}\sigma_2\sqrt{d})\ 1/(\sqrt{KL}\sigma_3\sqrt{d})\}$ 和 $T = \Theta(\max\{1, \tau R\})$，其中 $\tau := (G^2\sqrt{d}/(Ln_{\min}P\varepsilon_{\text{DP}}))^{2/3}$，则 *DIFF2-BVR-L-SGD* 满足

$$\mathbb{E}\|\nabla f(\widetilde{x}^{\text{out}})\|^2 \leq \left(O\left(\frac{L}{K} + \zeta + \frac{L}{\sqrt{Kb}}\right) + \widetilde{O}\left(\frac{L\sqrt{d}}{\sqrt{Kb}\sqrt{\varepsilon_{\text{DP}}}}\right)\right)\frac{1}{R} + \widetilde{O}\left(\frac{L\sqrt{d}}{n_{\min}\sqrt{P\varepsilon_{\text{DP}}}}\right)\frac{1}{\sqrt{R}} + \widetilde{O}\left(\frac{(LGd)^{\frac{2}{3}}}{(n_{\min}P\varepsilon_{\text{DP}})^{\frac{4}{3}}}\right)$$

具体而言，令

$$R = 1 + \left(\Theta\left(\frac{L}{K} + \zeta + \frac{L}{\sqrt{Kb}}\right) + \widetilde{\Theta}\left(\frac{L\sqrt{d}}{\sqrt{Kb}\sqrt{\varepsilon_{\text{DP}}}}\right)\right)\frac{1}{\varepsilon_{\text{opt}}} + \widetilde{\Theta}\left(\frac{L^2 d}{n_{\min}^2 P\varepsilon_{\text{DP}}^2}\right)\frac{1}{\varepsilon_{\text{opt}}^2}$$

导致效用

$$\mathbb{E}\|\nabla f(x^{\text{out}})\|^2 \leq \varepsilon_{\text{opt}} := \widetilde{\Theta}\left(\frac{(LGd)^{\frac{2}{3}}}{(n_{\min}P\varepsilon_{\text{DP}})^{\frac{4}{3}}}\right).$$

证明。 根据命题 F.3，我们知道

$$\mathbb{E}\|\nabla f(\widetilde{x}^{\text{out}})\|^2 \leq O\left(\frac{1}{\eta KR} + G^2\sigma_1^2 d\right). \tag{3}$$

因为 $C_1 = \Theta(G)$。

根据命题 E.2，针对 $(\varepsilon_{\text{DP}},\ \delta_{\text{DP}})$-差分隐私，相对于 $D_p$，所需的噪声水平为 $\sigma_1^2 = \widetilde{\Theta}(R/(Tn_{\min}^2P^2\varepsilon_{\text{DP}}^2))$、$\sigma_2^2 = \Theta(R/(n_{\min}^2P^2\varepsilon_{\text{DP}}^2))$ 和 $\sigma_3^2 = KR/(n_{\min}^2P\varepsilon_{\text{DP}}^2) + 1/(b^2\varepsilon_{\text{DP}})$ 分别为。将 $\sigma_1^2$的定义代入式 (3)，可得

$$\mathbb{E}\|\nabla f(\widetilde{x}^{\text{out}})\|^2 \leq O\left(\frac{1}{\eta KR}\right) + \widetilde{O}\left(\frac{RG^2d}{Tn_{\min}^2P^2\varepsilon_{\text{DP}}^2}\right).$$

将 $\eta$、 $\sigma_2^2$、 $\sigma_3^2$ 的定义代入所得结果，可得

$$\begin{aligned}
&\mathbb{E}\|\nabla f(\widetilde{x}^{\text{out}})\|^2 \\
&\leq O\left(\frac{L + K\zeta + \frac{\sqrt{K}}{\sqrt{b}}L}{KR}\right) + \widetilde{O}\left(\frac{K\sqrt{T}L\sigma_2\sqrt{d} + \sqrt{K}L\sigma_3\sqrt{d}}{KR} + \frac{RG^2d}{Tn_{\min}^2P^2\varepsilon_{\text{DP}}^2}\right) \\
&= O\left(\frac{\frac{L}{K} + \zeta + \frac{L}{\sqrt{Kb}}}{R}\right) + \widetilde{O}\left(\frac{\sqrt{T}L\sigma_2\sqrt{d} + \frac{L\sigma_3\sqrt{d}}{\sqrt{K}}}{R} + \frac{RG^2d}{Tn_{\min}^2P^2\varepsilon_{\text{DP}}^2}\right) \\
&= O\left(\frac{\frac{L}{K} + \zeta + \frac{L}{\sqrt{Kb}}}{R}\right) + \widetilde{O}\left(\frac{\frac{\sqrt{TR}L\sqrt{d}}{n_{\min}P\varepsilon_{\text{DP}}} + \frac{L\sqrt{R}\sqrt{d}}{n_{\min}\sqrt{P}\varepsilon_{\text{DP}}} + \frac{L\sqrt{d}}{\sqrt{Kb}\sqrt{\varepsilon_{\text{DP}}}}}{R} + \frac{RG^2d}{Tn_{\min}^2P^2\varepsilon_{\text{DP}}^2}\right) \\
&= \left(O\left(\frac{L}{K} + \zeta + \frac{L}{\sqrt{Kb}}\right) + \widetilde{O}\left(\frac{L\sqrt{d}}{\sqrt{Kb}\sqrt{\varepsilon_{\text{DP}}}}\right)\right)\frac{1}{R} + \widetilde{O}\left(\frac{\sqrt{T}L\sqrt{d}}{n_{\min}P\varepsilon_{\text{DP}}} + \frac{L\sqrt{d}}{n_{\min}\sqrt{P}\varepsilon_{\text{DP}}}\right)\frac{1}{\sqrt{R}} + \widetilde{O}\left(\frac{RG^2d}{Tn_{\min}^2P^2\varepsilon_{\text{DP}}^2}\right)
\end{aligned}$$

假设 $n_{\min}P = \Omega(G^2\sqrt{d}/(L\varepsilon_{\text{DP}}))$。定义

$$T := \Theta\left(1 \vee \left(\frac{G^2\sqrt{d}}{Ln_{\min}P\varepsilon_{\text{DP}}}\right)^{\frac{2}{3}} R\right)$$

其中 $1 \leq T \leq R$。

于是，我们有

$$\mathbb{E}\|\nabla f(\widetilde{x}^{\text{out}})\|^2 \leq \left(O\left(\frac{L}{K} + \zeta + \frac{L}{\sqrt{Kb}}\right) + \widetilde{O}\left(\frac{L\sqrt{d}}{\sqrt{Kb}\sqrt{\varepsilon_{\text{DP}}}}\right)\right)\frac{1}{R} + \widetilde{O}\left(\frac{L\sqrt{d}}{n_{\min}\sqrt{P}\varepsilon_{\text{DP}}}\right)\frac{1}{\sqrt{R}} + \widetilde{O}\left(\frac{(LGd)^{\frac{2}{3}}}{(n_{\min}P\varepsilon_{\text{DP}})^{\frac{4}{3}}}\right)$$

因此，令

$$R = 1 + \left(\Theta\left(\frac{L}{K} + \zeta + \frac{L}{\sqrt{Kb}}\right) + \widetilde{\Theta}\left(\frac{L\sqrt{d}}{\sqrt{Kb}\sqrt{\varepsilon_{\text{DP}}}}\right)\right)\frac{1}{\varepsilon_{\text{opt}}} + \widetilde{\Theta}\left(\frac{L^2d}{n_{\min}^2P\varepsilon_{\text{DP}}^2}\right)\frac{1}{\varepsilon_{\text{opt}}^2}$$

取 $\varepsilon_{\text{opt}} := \widetilde{\Theta}\left(\frac{(LGd)^{\frac{2}{3}}}{(n_{\min}P\varepsilon_{\text{DP}})^{\frac{4}{3}}}\right)$，即可得到所需结果。 □