arXiv:2302.04972v2 [cs.LG] 9 Jun 2023

面向光滑非凸经验风险最小化的差分隐私优化

高昌宇*和 Stephen J. Wright†

摘要

我们开发了简单的差分隐私优化算法，沿着（期望的）下降方向移动，以求得非凸经验风险最小化问题的近似二阶解。我们采用线搜索、小批量处理和两阶段策略，以提高算法的运行速度和实用性。数值实验展示了这些方法的有效性。

1 引言

隐私保护已成为机器学习算法中的核心议题，而差分隐私[Dwork 和 Roth， 2014] 是量化隐私的一种严格且广受欢迎的框架。我们提出了一种差分隐私优化算法，用于求解（可能是非凸的）经验风险最小化问题的近似二阶解。我们还提出了若干技术来提升该方法的实际性能，包括回溯线搜索、小批量处理，以及一种用于避免分析中保守假设所带来影响的启发式方法。

对于给定的 f： $\mathbb{R}^d \to \mathbb{R}$，光滑但非凸的情形，我们考虑以下最小化问题：

$$\min_{w\in\mathbb{R}^d} f(w).$$

我们寻求如下定义的近似二阶解。

定义 1.1 ((ϵ_g, ϵ_H)-二阶驻点)。 对于给定的正值 ϵ_g 和 ϵ_H，我们称 w 为一个（ϵ_g, ϵ_H）-近似二阶解（简称为（ϵ_g, ϵ_H）-二阶驻点），如果

$$\|\nabla f(w)\| \le \epsilon_g, \quad \lambda_{\min}\left(\nabla^2 f(w)\right) \ge -\epsilon_H. \tag{1}$$

我们主要关注 $\epsilon_g = \alpha$ 和 $\epsilon_H = \sqrt{M\alpha}$的情形，其中 $\alpha > 0$。也就是说，我们寻求一个 $(\alpha, \sqrt{M\alpha})$-二阶驻点，其中 M是 $\nabla^2 f$的利普希茨常数。

我们将重点关注经验风险最小化（ERM）。

定义 1.2 （ERM）。 给定数据集 $D = \{x_1, \ldots, x_n\}$和损失函数 $l(w, x)$，我们寻求使经验风险最小化的参数 $w \in \mathbb{R}^d$

$$f(w) = L(w, D) := \frac{1}{n}\sum_{i=1}^{n} l\left(w, x_i\right). \tag{2}$$

经验风险最小化是机器学习中的经典问题，已得到广泛研究；例如，参见 Shalev-Shwartz 和 Ben-David [2014]。本文介绍用于求解经验风险最小化的差分隐私技术。

*威斯康星大学麦迪逊分校工业与系统工程系，威斯康星州 53706，美国（电子邮件：changyu.gao@wisc.edu）

†威斯康星大学麦迪逊分校计算机科学系，威斯康星州 53706，美国（电子邮件：swright@cs.wisc.edu）

以往关于经验风险最小化和优化的差分隐私算法研究主要聚焦于凸损失函数。近期针对非凸经验风险最小化的差分隐私算法研究 [Wang 等人，2018a，Wang 和 Xu，2019，Zhang 等人， 2017] 旨在找到一个近似驻点，该点仅满足条件 (1) 中的第一个条件。Wang 和 Xu [2021] 提出了一种信赖域类算法（DP-TR），在特定的 ϵ_g和 ϵ_H取值下，可为经验风险最小化问题提供满足条件 (1) 中两个条件的近似二阶解。该方法要求在每次迭代中精确求解信赖域子问题，并将信赖域半径固定为一个较小值，类似于线搜索方法中的“短步”。更早的一项工作 [Wang 等人， 2019] 提出了 DP-GD 算法，该算法沿含噪梯度方向采取短步，然后从生成的所有迭代点中筛选出一个满足二阶必要条件的点。我们的方法达到了与 DP-GD 相同的样本复杂度界：对于 ρ-zCDP，该界为 $O\left(\frac{\sqrt{d}}{\alpha^2\sqrt{\rho}}\right)$；对于 (ε,δ)-DP，该界为 $O\left(\frac{\sqrt{d}\ln(1/\delta)}{\alpha^2\varepsilon}\right)$，用于求得 $(\alpha,\sqrt{M\alpha})$-二阶驻点；其迭代复杂度为 $O(\alpha^{-2})$。我们的贡献可概括如下。

- 我们的算法简单易懂，基于一种用于求取近似二阶解的简单（非隐私）线搜索算法。只有在单独使用一阶（梯度）信息无法取得足够进展时，才会评估二阶信息（带噪海森矩阵）。相比之下，DP- GD 仅在检查二阶近似条件时使用（带噪）海森矩阵，而 DP-TR 则要求每次迭代都计算带噪海森矩阵。

- 我们的算法实用且快速。DP-TR 的最坏情况样本复杂度上界略优于我们的方法，具体取决于 $O(\alpha^{-7/4})$，而非 $O(\alpha^{-2})$。不过，我们的方法采用线搜索，因此能够适应函数的局部几何，从而在实际应用中的表现优于最坏情况界所预示的水平。相比之下，DP-TR 使用一个较小的信赖域，其大小依据函数的最坏情况全局性质确定，并且要求精确求解其信赖域子问题。这一操作相对昂贵；当梯度足够大、无需利用海森矩阵即可取得有效进展时，这种做法也并不理想。实验表明，与 DP-TR 相比，我们的算法所需迭代次数更少，且平均每次迭代的计算量更小，因此在实践中运行速度显著更快。无论如何，DP-TR 的小批量 版本的样本复杂度为 $O(\alpha^{-2})$，与我们算法的小批量版本的样本复杂度相当。

- 我们采用线搜索和小批量处理来加速算法。Chen and Lee 2020] 提出了差分隐私线搜索算法。我们采用了与他们的工作相同的稀疏向量技术，但对用于检查充分下降条件的查询敏感度进行了更精确的分析。此外，我们还严格分析了以高概率保证的函数下降量。

- 为了补充最坏情况分析，我们提出了一种启发式方法，在保留分析所提供保证的同时，实现更快的收敛。

本文其余部分安排如下。第2节回顾差分隐私的基本定义和性质，并概述我们对待优化函数 f 的假设。第3节介绍我们的算法及其分析。第3.1节讨论该算法的基本短步版本，第3.2节介绍其向实用线搜索方法的扩展。第3.3节介绍该算法的小批量改编。第4节展示实验结果，并说明我们算法的有效性。

2 预备知识

我们的分析使用了几种差分隐私的变体，包括 (ε,δ)-差分隐私 [Dwork 和 Roth、 2014], (α,ϵ)-Rényi 差分隐私[Mironov、 2017], 以及 zCDP [Bun and Steinke 2016]。我们在附录 A 中回顾了它们的定义和性质。

我们对目标函数 f的光滑性作出如下假设 .

假设 2.1. 我们假设 f 有下界 $\underline{f}$。进一步假设 f 是 G-光滑的，且具有 M-利普希茨连续海森矩阵，也就是说，对所有 $w_1, w_2 \in \mathrm{dom}(f)$，都有

$$\|\nabla f(w_1) - \nabla f(w_2)\| \le G\|w_1 - w_2\|,$$
$$\|\nabla^2 f(w_1) - \nabla^2 f(w_2)\| \le M\|w_1 - w_2\|,$$

其中，$\|\cdot\|$ 分别表示向量 2-范数和矩阵 2-范数。本文通篇使用这一记号。

对于 f 的经验风险最小化版本（见定义 1.2），我们还需要作出额外假设。

假设 2.2. 对于经验风险最小化设置 (2)，我们假设损失函数 $l(w,x)$ 是 G-光滑的，且关于 w 具有 M-利普希茨连续海森矩阵。因此，$L(w,D)$（n 个样本的平均损失）也同样是 G-光滑的，且关于 w 具有 M-利普希茨连续海森矩阵。此外，我们假设 $l(w,x)$ 的函数值、梯度和海森矩阵均有界。也就是说，存在常数 B、B_g 和 B_H，使得对于任意 w, x，都有

$$0 \le l(w,x) \le B,\ \|\nabla_w l(w,x)\| \le B_g,\ \|\nabla_w^2 l(w,x)\| \le B_H.$$

因此，$L(w,D)$ 和 $\nabla L(w,D)$ 的 l^2 敏感度分别有 B/n 和 $2B_g/n$ 上界。我们有

$$\begin{aligned}&\|\nabla^2 L(w,D) - \nabla^2 L(w,D')\|_F \\ &\le \sqrt{d}\,\|\nabla^2 L(w,D) - \nabla^2 L(w,D')\| \le \frac{2B_H\sqrt{d}}{n}.\end{aligned}$$

为简化记号，我们定义 $g(w) := \nabla f(w)$ 和 $H(w) := \nabla^2 f(w)$。根据 ℓ_2-敏感度的定义 (21)，f、g 和 H 的敏感度为

$$\Delta_f = \frac{B}{n}, \quad \Delta_g = 2\frac{B_g}{n}, \quad \Delta_H = \frac{2B_H\sqrt{d}}{n}. \tag{3}$$

3 算法及其性质

我们的算法起点是 Wright 和 Recht [2022, 第 3 章6]中描述的基础算法，该算法保证收敛到满足近似二阶条件的点。为简单起见，我们使用以下记号来描述和分析该方法:

$$\begin{aligned}f_k &:= f(w_k),\\ g_k &:= g(w_k) = \nabla f(w_k),\\ H_k &:= H(w_k) = \nabla^2 f(w_k).\end{aligned} \tag{4}$$

我们采用高斯机制对梯度和海森矩阵添加扰动，并记为

$$\tilde{g}_k = g_k + \varepsilon_k, \quad \tilde{H}_k = H_k + E_k,$$

其中，对于某个选定的参数 σ_g，$\varepsilon_k \sim \mathcal{N}(0, \Delta_g^2\sigma_g^2 I_d)$，且 E_k是一个对称矩阵，其对角线及对角线以上的每个元素均独立同分布于 $\mathcal{N}\left(0, \Delta_H^2\sigma_H^2\right)$，其中 σ_H为某个选定的值。令 $\tilde{\lambda}_k$表示 H_k 的最小~特征值，其对应的特征向量为 $\tilde{p}_k$，其符号和范数选取为满足

$$\|\tilde{p}_k\| = 1 \quad \text{and} \quad (\tilde{p}_k)^T\tilde{g}_k \le 0. \tag{5}$$

算法 1 给出了我们优化算法的一般形式。我们将讨论两种策略——“短步”策略和基于回溯线搜索的策略——以选择沿方向 $\gamma_{k,\,g}$ 和 $\gamma_{k,\,H}$所采取的步长

的步长，分别沿方向 $\tilde{g}_k$ 和 $\tilde{p}_k$进行。对于每种变体，我们定义一个量 MIN_DEC，表示最小下降量，并将其与 f 的一个特定下界结合起来，定义所需迭代次数的上界T。每次迭代中，如果 $\|\tilde{g}_k\| > \epsilon_g$，我们就沿扰动梯度方向 $\tilde{g}_k$ 的反方向迈出一步。否则，我们检查扰动海森矩阵 $\tilde{H}_k$的最小特征值 λ_k。如果 $\lambda_k < -\epsilon_H$，我们就沿方向 $\tilde{p}_k$迈出一步。在其余情况下，我们有 $\|\tilde{g}_k\| \le \epsilon_g$ 和$\lambda_k \ge -\epsilon_H$，因此近似二阶条件得到满足（误差至多为某些常数因子），并将当前迭代点 w_k 作为二阶驻点解输出。

算法 1 具有二阶保证的差分隐私优化（一般形式）

给定： 每次迭代的最小下降量 MIN_DEC、容差 ϵ_g、 ϵ_H、噪声参数 σ_f、 σ_g、 σ_H

初始化 w_0并采样 $z \sim \mathcal{N}(0, \Delta_f^2\sigma_f^2)$，按如下方式计算所需迭代次数的上界

$$T = \left\lceil \frac{f(w_0) + |z| - \underline{f}}{\text{MIN_DEC}} \right\rceil \tag{6}$$

使用 T 计算 σ_g 和 σ_H （详见定理）

对于 $k = 0, 1, \ldots, T-1$ **执行**

采样 $\varepsilon_k \sim \mathcal{N}\left(0,\ \Delta_g^2\sigma_g^2 I_d\right)$

计算扰动梯度 $\tilde{g}_k = g_k + \varepsilon_k$

如果 $\|\tilde{g}_k\| > \epsilon_g$**，则**

选择步长 $\gamma_{k,\ g}$ 并设置 $w_{k+1} \leftarrow w_k - \gamma_{k,\ g}\tilde{g}_k$ ▷ Gradient step

else

采样 E_k，使得 E_k是一个 $d \times d$ 对称矩阵，其中对角线及其上方的每个元素与 $\mathcal{N}(0,\ \Delta_H^2\sigma_H^2)$ 独立同分布

计算扰动海森矩阵 $\tilde{H}_k = H_k + E_k$

计算 的最小特征值 $\tilde{H}_k$ 及其对应的特征向量（ $\tilde{\lambda}_k$、 $\tilde{p}_k$）满足（5）

if $\tilde{\lambda}_k \le -\epsilon_H$ **然后**

选择步长 $\gamma_{k,\ H}$ 并设置 $w_{k+1} \leftarrow w_k + \gamma_{k,\ H}\tilde{p}_k$ ▷ 负曲率步

else

返回 w_k

结束 if

结束 if

结束 for

σ_f、 σ_g、 σ_H 分别决定为函数、梯度和海森矩阵计算添加的噪声量，目的是通过高斯机制保护隐私。我们可以为整个算法设定一个隐私级别（例如，ρ 在 ρ-零集中差分隐私下），求出所用算法 1 的变体所需迭代次数的上界，然后选择 σ_f、 σ_g和σ_H 以确保达到该隐私级别。反过来，我们也可以为 σ_f、 σ_g和 σ_H 选择正值，再确定这种选择能够确保的隐私级别。随着算法的推进，我们可以跟踪隐私泄露，从而有可能采用自适应方案来选择噪声方差。

3.1 短步

在该算法的短步版本中，我们按如下方式选择步长

$$\gamma_{k,g} \equiv \frac{1}{G}, \quad \gamma_{k,H} \equiv \frac{2|\tilde{\lambda}_k|}{M}. \tag{7}$$

MIN_DEC 以及噪声参数 σ_f、 σ_g 和 σ_H 的选择将在以下结果中讨论。

首先，我们讨论隐私保证及其与噪声方差和迭代次数的关系。

定理 3.1。 给定噪声参数 σ_f、 σ_g、 σ_H 。假设算法 *1* 的一次运行进行了 k_g 次梯度步和 k_H 次负曲率步。那么该次运行满足 ρ-z 差分隐私，其中 $\rho = \frac{1}{2}\left(\frac{1}{\sigma_f^2} + \frac{k_g+k_H}{\sigma_g^2} + \frac{k_H}{\sigma_H^2}\right)$.

回顾一下， T 是式 (6) 中定义的最大迭代次数。令 $\bar{\rho} = \frac{1}{2}\left(\frac{1}{\sigma_f^2} + \frac{T}{\sigma_g^2} + \frac{T}{\sigma_H^2}\right)$. 。我们总有 $\bar{\rho} \geq \rho$，因此该算法满足 $\bar{\rho}$-z 差分隐私。反过来，对于给定的 $\rho > 0$ 和 $c_f \in (0,1)$，我们可以选择 $\rho_f = c_f\rho$，并且

$$\sigma_f^2 = \frac{1}{2\rho_f}, \quad \sigma_g^2 = \sigma_H^2 = \frac{T}{(1-c_f)\rho}. \tag{8}$$

以确保该算法满足 ρ-z 差分隐私。

证明。 该证明直接源于高斯机制的 zCDP 保证，并结合 zCDP 的后处理和组合。

□

备注 3.2。 在我们的算法中，实际噪声会按式 (3) 中定义的相应敏感度 Δ 进行缩放。后续算法也采用相同做法。实际中，我们预计大多数步骤都是梯度步，因此 $\bar{\rho}$ 会高估实际的隐私级别 ρ。因此在实践中，我们选择噪声方差时可以比最坏情况理论所建议的更大胆。我们将在第 3.4 节讨论一种两阶段方法。

现在我们讨论算法 1 的输出保证。首先，我们估计每个短步中的 MIN_DEC。

引理 3.3。 采用短步步长选择(7)时，若噪声满足以下条件：存在正数常量 c、 c_1 和 c_2，使得 $c_1 < \frac{1}{2}$ 和 $c_2 + c < \frac{1}{3}$，并且

$$\|\varepsilon_k\| \leq \min\left(c_1\epsilon_g, \frac{c_2}{M}\epsilon_H^2\right), \tag{9a}$$

$$\|E_k\| \leq c\,\epsilon_H, \tag{9b}$$

那么，每一步的下降量至少为

$$\text{MIN_DEC} = \min\left(\frac{1-2c_1}{2G}\epsilon_g^2,\, 2\left(\frac{1}{3} - c_2 - c\right)\frac{\epsilon_H^3}{M^2}\right). \tag{10}$$

真实梯度和海森矩阵的真实最小特征值满足以下条件，

$$\|g_k\| \leq (1+c_1)\,\|\tilde{g}_k\|, \quad \lambda_k > -(1+c)|\tilde{\lambda}_k|. \tag{11}$$

备注 3.4. (9) 中的常数 c、 c_1 和 c_2 控制含噪梯度和海森矩阵估计的精度。选择较小的容差 ϵ_g 和 ϵ_H 时，MIN_DEC 会更小。较小的容差也会得到更精确的解，但要满足条件 (9)，我们需要更小的噪声方差。这一要求意味着经验风险最小化问题所需的样本量 n 更大，正如我们将在定理 3.6 中看到的那样。

推论 3.5. 假设每次迭代中的噪声都满足(9)，则该算法的短步版本（使用(7)）将输出一个$((1+c_1)\epsilon_g, (1+c)\,(10)/\epsilon_H)$-二阶驻点。

基于以上结果，我们现在分析固定步长算法在经验风险最小化设置下的保证。

定理 3.6（短步算法的样本复杂度）。 考虑经验风险最小化设置。假设样本数 n 满足 $n \geq n_{\min}$，其中

$$n_{\min} := \max\left(\frac{\sqrt{2d}B_g\sigma_g\log\frac{T}{\zeta}}{\min\left(c_1\epsilon_g,\, \frac{c_2}{M}\epsilon_H^2\right)}, \frac{C\sqrt{d}B_H\sigma_H\log\frac{T}{\zeta}}{c\,\epsilon_H}\right).$$

以至少 $\{(1-\frac{\zeta}{T})(1-C\exp(-C_1Cd))\}^T$ 的概率，其中 C 和 C_1 是引理 *3.8* 中的普适常数，算法的短步版本（使用 (7)、(10)）的输出是一个 $((1+c_1)\epsilon_g, (1+c)\epsilon_H)$-二阶驻点。

在选择 σ 的取值如 (8) 所示并忽略对数项和常数后，$n_{\min}$ 对 (ϵ_g, ϵ_H)、ρ 和 d 的渐近依赖关系为

$$n_{\min} = \frac{\sqrt{d}}{\sqrt{\rho}} \tilde{O}\left(\max\left(\epsilon_g^{-2}, \epsilon_g^{-1}\epsilon_H^{-2}, \epsilon_H^{-7/2}\right)\right). \tag{12}$$

当 $(\epsilon_g, \epsilon_H) = (\alpha, \sqrt{M\alpha})$ 时，依赖关系简化为 $\frac{\sqrt{d}}{\sqrt{\rho}}\tilde{O}(\alpha^{-2})$。

在证明定理 3.6 之前，我们先介绍两个集中结果。

引理 3.7（高斯集中，[Vershynin, 2018]）。对于 $x \sim \mathcal{N}(0, \sigma^2 I_d)$，以至少 $1-\eta$ 的概率，对任意 $1 > \eta > 0$，有

$$\|x\| \le \sqrt{2d}\sigma \log\frac{1}{\eta}.$$

引理 3.8（Wigner 系综的上尾估计，[Tao，2012，第 110]页）。设 $M = (m_{ij})_{1\le i,\ j\le d}$ 为一个$d\times d$ 随机对称矩阵。假设 M 的系数 m_{ij} 对于 $j \ge i$ 相互独立、均值为零，并且具有一致的次高斯尾部。存在普适常数 C，$C_1 > 0$，使得对于所有 $A \ge C$，有

$$\mathbf{P}\left(\|M\| > A\sqrt{d}\right) \le C\exp(-C_1 Ad).$$

证明。（定理 3.6）根据集中结果，在迭代 k 中，以至少 $(1-\frac{\zeta}{T})(1-C\exp(-C_1Cd))$ 的概率，我们有

$$\|\varepsilon_k\| \le \sqrt{2d}\Delta_g\sigma_g \log\frac{T}{\zeta}, \tag{13a}$$

$$\|E_k\| \le C\sqrt{d}\Delta_H\sigma_H. \tag{13b}$$

我们需要找到一个关于 n 的条件，以确保右侧各项小于式 (9) 的右侧各项。将式 (3) 中的 Δ_g 和 Δ_H 代入，并通过重新排列各项求解 $n_{\min}$。如果集中结果对所有迭代都成立，则由推论 3.5 可得结论。

现在来计算成功概率。对于每次迭代，集中结果成立的概率至少为 $(1-\frac{\zeta}{T})(1-C\exp(-C_1Cd))$（如果不计算扰动海森矩阵，该概率会更高，至少为 $1-\frac{\zeta}{T}$）。利用条件概率，在给定迭代次数 τ 的条件下，总体成功概率为$\{(1-\frac{\zeta}{T})(1-C\exp(-C_1Cd))\}^\tau$。由于 $\tau \le T$，总体成功概率至少为 $\{(1-\frac{\zeta}{T})(1-C\exp(-C_1Cd))\}^T$。

对于第二部分，回顾式 (8)，可知 $\sigma_g = \sigma_H = \sqrt{T}/\sqrt{(1-c_f)\rho_\circ}$ 结合式 (10) 以及我们在式 (6) 中对 T 的选择，可得 $\sqrt{T} = O(\max(\epsilon_g^{-1}, \epsilon_H^{-3/2}))$。将 σ_g 和 σ_H 代入，即可得到 $n_{\min}$ 的渐近界。

□

备注 3.9。当条件 (9) 不成立时，算法可能无法收敛到一个 $((1+c_1)\epsilon_g, (1+c)\epsilon_H)$-二阶驻点。首先，扰动梯度和海森矩阵中的噪声可能过大，以致该步不是下降方向。其次，由于噪声，在检查近似二阶条件时，我们可能过早终止，或未能及时终止。若我们过早终止且噪声并不过大，该解仍可能是可接受的，因为含噪评估满足终止条件。

3.2 线搜索算法

我们可以使用回溯进行线搜索，而不是采用保守的固定步长。回溯线搜索需要一个初始值 γ_0、步长的下降参数 $\beta \in (0,1)$，以及决定所需下降量的常数$c_g \in (0, 1-c_1)$、$c_H \in (0, 1-c-\sqrt{\frac{8}{3}c_2})$。每次线搜索都会依次尝试

值 γ_0、 $\beta\gamma_0$、 $\beta^2\gamma_0$，.....，直到找到满足充分下降条件的值。对于梯度步，该条件为

$$f(w-\gamma\tilde{g}) \le f(w) - c_g\gamma\|\tilde{g}\|^2, \tag{SD1}$$

而对于负曲率步，则是

$$f(w+\gamma\tilde{p}) \le f(w) - \frac{1}{2}c_H\gamma^2|\tilde{\lambda}|. \tag{SD2}$$

为使线搜索满足差分隐私，我们采用 [Dwork 和 Roth2014]提出的稀疏向量技术（SVT）。我们根据（SD1）和（SD2）定义查询:

$$q_g(\gamma, w) = f(w) - f(w-\gamma\tilde{g}) - c_g\gamma\|\tilde{g}\|^2, \tag{14a}$$

$$q_H(\gamma, w) = f(w) - f(w+\gamma\tilde{p}) - \frac{1}{2}c_H\gamma^2|\tilde{\lambda}|, \tag{14b}$$

其非负性分别对应于各项充分下降条件。

算法 2 给出了使用 SVT 的差分隐私线搜索算法，该算法改编自 [Dwork 和 Roth 2014]的 AboveThreshold 算法。

通过满足充分下降条件，我们力求使函数值比短步算法获得更显著的改善。作为回退策略，如果线搜索失败，我们采用与短步算法相近的步长（仅相差一个常数因子），从而获得与短步情形相近的下降量。我们在算法 3 中给出了结合线搜索的完整算法。在该算法中，我们计算回退步长 $\bar{\gamma}$，并将其乘以乘数 b（$b>1$），作为线搜索的初始步长 $b\bar{\gamma}$。据此计算查询敏感度 Δ_q，并调用私有线搜索子程序，以找到满足充分下降条件的步长 γ。

算法 2 使用稀疏向量技术的私有回溯线搜索

给定： 查询 q 及其敏感度 Δ_q、初始步长乘数 b、回退步长 $\bar{\gamma}$、下降参数β、隐私参数 λ

函数 差分隐私线搜索$(q, \Delta_q, \gamma^{\text{init}}, \bar{\gamma}, \beta, \lambda)$
初始化 $\gamma \leftarrow \gamma^{\text{init}}$。
采样 $\xi \sim$ 拉普拉斯分布 $(2\lambda\Delta_q)$
对于 $i = 1, 2, \ldots, i_{\max} = \lfloor \log_\beta \frac{\bar{\gamma}^{\text{初始化}}}{\gamma} \rfloor + 1$ **执行**
采样 $\nu_i \sim$ 拉普拉斯分布（$4\lambda\Delta_q$）
计算 $q_i = q$（γ）和 $\tilde{q}_i = q_i + \nu_i$
如果 $\tilde{q}_i \ge \xi$ **则**
停止并输出 γ
else
$\gamma \leftarrow \beta\gamma$
结束如果
结束循环
停止并输出 $\bar{\gamma}$。

我们有以下隐私保证。

定理 3.10。 假设给定了 σ_f、 σ_g、 σ_H 和 λ。假设线搜索算法的一次实际运行进行了 k_g 次梯度步和 k_H 次负曲率步。该次运行是 ρ-*zCDP*，其中 $\rho = \frac{1}{2}\left(\frac{1}{\sigma_f^2} + \frac{k_g+k_H}{\sigma_g^2} + \frac{k_H}{\sigma_H^2} + \frac{k_g+k_H}{\lambda^2}\right)$.

回想一下， T是式(6)中定义的最大迭代次数。令 $\bar{\rho} = \frac{1}{2}\left(\frac{1}{\sigma_f^2} + \frac{T}{\sigma_g^2} + \frac{T}{\sigma_H^2} + \frac{T}{\lambda^2}\right)$. 。我们总有 $\bar{\rho} \ge \rho$，因此该算法是 $\bar{\rho}$-z差分隐私。反过来，对于给定的 $\rho > 0$ 和 $\rho_f \in (0, \rho)$，我们可以选择

$$\sigma_f^2 = \frac{1}{2\rho_f}, \quad \sigma_g^2 = \sigma_H^2 = \lambda^2 = \frac{3T}{2(\rho - \rho_f)}, \tag{15}$$

算法 3差分隐私 O　具有二阶保证和回溯线搜索的优化算法

给定： 噪声上界参数 c_1, c_2, c, 充分下降参数 c_g, c_H 和初始步长乘数 b_g, b_H, 线搜索下降参数 β_g, β_H, 容差 ϵ_g和 ϵ_H, 噪声参数 σ_f,σ_g, σ_H 和 λ_{SVT}

初始化 w_0, 采样 $z \sim \mathcal{N}(0, \Delta_f^2\sigma_f^2)$ 并根据 (18) 计算 MIN_DEC

计算所需迭代次数的上界 $T = \left\lceil \frac{f(w_0)+|z|-\underline{f}}{\text{MIN_DEC}} \right\rceil$

$\bar{\gamma}_g \leftarrow 2(1-c_1-c_g)/G$

对于 $k = 0, 1, \ldots, T-1$ **执行**

采样 $\varepsilon_k \sim \mathcal{N}\left(0, \Delta_g^2\sigma_g^2 I_d\right)$

计算扰动梯度 $\tilde{g}_k = g_k + \varepsilon_k$

如果 $\|\tilde{g}_k\| > \epsilon_g$ **则**

定义 $q_{k,g}(\gamma) = f(w_k) - f(w_k - \gamma\tilde{g}_k) - c_g\gamma\|\tilde{g}_k\|^2$

$\gamma_{k,g}^{\text{init}} \leftarrow b_g\bar{\gamma}_g$, $\Delta_{q_{k,g}} \leftarrow \frac{2}{n}\gamma_{k,g}^{\text{init}} B_g\|\tilde{g}_k\|$ ▷ 线搜索查询敏感度

$\gamma_{k,g} \leftarrow$ 差分隐私线搜索$(q_{k,g}, \Delta_{q_{k,g}}, \gamma_{k,g}^{\text{init}}, \bar{\gamma}_g, \beta_g, \lambda_{SVT})$ ▷ 回溯线搜索

$w_{k+1} \leftarrow w_k - \gamma_{k,g}\tilde{g}_k$ ▷ 梯度步

else

采样 E_k, 使得 E_k 是一个 $d \times d$ 对称矩阵，其对角线及对角线以上的每个元素均独立同分布为 $\mathcal{N}(0, \Delta_H^2\sigma_H^2)$

计算扰动海森矩阵 $\tilde{H}_k = H_k + E_k$

计算.....的最小特征值 $\tilde{H}_k$ 以及相应的特征向量（$\tilde{\lambda}_k, \tilde{p}_k$）满足（5）

if $\tilde{\lambda}_k < -\epsilon_H$ **则**

定义 $q_{k,H}(\gamma) = f(w_k) - f(w_k + \gamma\tilde{p}_k) - \frac{1}{2}c_H\gamma^2|\tilde{\lambda}_k|$

$\bar{\gamma}_{k,H} \leftarrow t_2|\tilde{\lambda}_k|/M$, $\gamma_{k,H}^{\text{init}} \leftarrow b_H\bar{\gamma}_{k,H}$, $\Delta_{q_{k,H}} \leftarrow \frac{2}{n}\gamma_{k,H}^{\text{init}} B_g$

$\gamma_{k,H} \leftarrow$ 差分隐私线搜索$(q_H, \Delta_{q_{k_H}}, \gamma_{k,H}^{\text{init}}, \bar{\gamma}_{k,H}, \beta_H, \lambda_{SVT})$ ▷ 回溯线搜索

$w_{k+1} \leftarrow w_k + \gamma_{k,H}\tilde{p}_k$ ▷ 负曲率步

else

返回 w_k

结束 if

结束 if

结束 for

以确保算法是 ρ-z*CDP*。

证明。 我们知道稀疏向量技术是 $(1/\lambda)$-差分隐私。因此，它满足 $(1/(2\lambda^2))$-zCDP。该结果可直接由高斯机制的z CDP 保证，以及对 zCDP 的后处理和组合得出。 □

现在我们讨论算法输出的保证。我们首先推导充分下降的必要条件。

引理 3.11。 假设与之前相同的有界噪声条件(9) 。选取充分下降系数 $c_g \in (0, 1-c_1)$ $c_H \in (0, 1-c-\sqrt{\frac{8}{3}c_2})$，令 $\bar{\gamma}_g = 2(1-c_1-c_g)/G$和 $\bar{\gamma}_H = t_2|\tilde{\lambda}|/M$如算法 *3* 中所定义，则当 $\gamma \le \bar{\gamma}_g$ 和$\gamma \in [(t_1/t_2)\bar{\gamma}_H, \bar{\gamma}_H]$时，充分下降条件(SD1) 和(SD2) 分别成立，其中 $0 < t_1 < t_2$是以下二次方程的解（根据我们对 c, c_2, c_H的选择，存在实数解），

$$r(t) := -\frac{1}{6}t^2 + \frac{1}{2}(1-c-c_H)t - c_2 = 0,$$

具体而言，我们有

$$t_1, t_2 = \frac{3}{2}(1-c-c_H) \pm 3\sqrt{\frac{1}{4}(1-c-c_H)^2 - \frac{2}{3}c_2}. \tag{16}$$

特别地，我们有 $q_g(\bar{\gamma}_g) \geq 0$，以及 $q_H(\bar{\gamma}_H) \geq 0$。

现在我们推导每次迭代的最小下降量。

引理 3.12。使用 *DP* 线搜索算法 *3*，并假设与之前相同的有界噪声条件(9)。选取充分下降系数 $c_g \in (0, 1-c_1)$ $c_H \in (0, 1-c-\sqrt{\frac{8}{3}c_2})$，并按之前的方式定义 $\bar{\gamma}_g$ 和 $\bar{\gamma}_H$。选择初始步长乘数 $b_g, b_H > 1$ 和下降参数 $\beta_g \in (0,1)$ $\beta_H \in (t_1/t_2, 1)$。令 $i_{\max} = \lfloor \log^{\beta} \max(b_g, b_H) \rfloor + 1$。若 n 至少为

$$16\lambda \left(\log i_{\max} + \log \frac{T}{\xi} \right) B_g \max \left(\frac{2b_g}{c_g \epsilon_g}, \frac{4b_H M}{t_2 c_H \epsilon_H^2} \right), \tag{17}$$

以至少 $1-\xi/T$ 的概率，单步下降量至少为

$$\text{MIN_DEC} = \min \left(\frac{1}{G}(1-c_1-c_g)c_g\epsilon_g^2, \frac{1}{4}c_H t_2^2 \frac{\epsilon_H^3}{M^2} \right). \tag{18}$$

基于上述结果，我们现在可以分析经验风险最小化设置下线搜索算法的保证。

定理 3.13（线搜索算法的样本复杂度）。假设满足与前一个引理相同的条件，则以至少 $\{(1-\frac{\zeta}{T})(1-C\exp(-C_1 C d))(1-\xi/T)\}^T$ 的概率，若样本数量 n 满足 $n \geq n_{\min}$，其中

$$n_{\min} := \max \left(\frac{\sqrt{2d} B_g \sigma_g \log \frac{T}{\zeta}}{\min\left(c_1\epsilon_g, \frac{c_2}{M}\epsilon_H^2\right)}, \frac{C\sqrt{d} B_H \sigma_H \log \frac{T}{\zeta}}{c\,\epsilon_H}, 16\lambda \left(\log i_{\max} + \log \frac{T}{\xi} \right) B_g \max \left(\frac{2b_g}{c_g\epsilon_g}, \frac{4b_H M}{t_2 c_H \epsilon_H^2} \right) \right). \tag{19}$$

该算法的输出是一个$((1+c_1)\epsilon_g, (1+c)\epsilon_H)$-二阶驻点。选择 σ 和 λ于(15)中，忽略对数项和常数后，$n_{\min}$ 对(ϵ_g, ϵ_H)及 ρ 的渐近依赖关系为

$$n_{\min} = \frac{\sqrt{d}}{\sqrt{\rho}} \tilde{O}\left(\max\left(\epsilon_g^{-2}, \epsilon_g^{-1}\epsilon_H^{-2}, \epsilon_H^{-7/2} \right) \right). \tag{20}$$

当 $(\epsilon_g, \epsilon_H) = (\alpha, \sqrt{M\alpha})$ 时，依赖关系简化为 $\frac{\sqrt{d}}{\sqrt{\rho}} \tilde{O}(\alpha^{-2})$。

证明。证明与定理 3.6 类似，使用引理 18 即可。由于 SVT 线搜索步骤，我们的成功概率中多了一项。对于 $n_{\min}$ 的渐近界，我们注意到 MIN_DEC (18)、T、σ_g 和 σ_H 与短步算法中的对应项相同，只差一个常数。对 n 的额外要求 (17) 为 $O\left(\lambda \log T \max(\epsilon_g^{-1}, \epsilon_H^{-2}) \right)$。由于我们选择了 $\lambda = \sigma_g$，它与 (19) 中 max 表达式的第一项同阶。因此，$n_{\min}$ 的渐近界与短步算法相同。

□

3.3 小批量处理

小批量处理是加快算法速度的一种有效方法。通过从完整数据集中抽取数据点子集，我们可以计算该小批量上的平均风险，并修改短步算法，以在该子集上计算梯度和海森矩阵。

形式上，在每次迭代 k中，我们从 D 中无放回地抽取 m 个数据点，形成小批量 S_k。此时，目标函数为集合 S_k上的平均风险，即:

$$f_{S_k} := \frac{1}{m} \sum_{i \in S_k} \ell(w_k, x_i).$$

我们证明，对于（ϵ_g, ϵ_H）=（$\alpha, \sqrt{M\alpha}$）的（ε, δ）-差分隐私，算法的小批量版本的样本复杂度仍为 $\tilde{O}(\sqrt{d \ln(1/\delta)}\varepsilon\alpha^2)$，与 DP-TR 小批量版本的样本复杂度相匹配。详情见附录 B。

3.4 讨论：两阶段策略与特征值计算

为了加快算法速度，我们提出了一种两阶段策略，以解决式（6）中 T 的悲观估计问题；该估计基于最坏情况分析得出的 MIN_DEC。

第一阶段包括使用一部分隐私预算（例如 3/4）来尝试使用较小的 T 值。使用较小的 T 会产生更少的噪声，并有可能加快收敛。如果我们无法找到理想的解，就会进入第二阶段。在此阶段，我们利用剩余的隐私预算，回退到原始方法来估计 T。使用上一次迭代的结果作为热启动，有助于提高该阶段的效率。

此外，我们可以使用 Lanczos 方法来寻找最小特征值和特征向量的近似值，以替代直接特征值计算。这种替代方法会带来略有不同的分析。相关讨论请参见附录 C。

4 实验

我们开展数值实验，以展示差分隐私优化算法的性能，并遵循与 [Wang 和 Xu， 2021] 类似的实验方案。我们使用 Covertype 数据集，并执行必要的数据预处理。有关该数据集及其他实验的详细信息，请参见附录 E。

令 x_i 为特征向量， $y_i \in \{-1, +1\}$ 为二元标签。我们研究非凸经验风险最小化损失 [1]:

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i \left\langle x_i, w \right\rangle\right)\right) + \lambda r(w),$$

其中 $r\left(w\right) = \sum_{i=1}^{p} \frac{w_i^2}{1+w_i^2}$ 是非凸正则化项。在我们的实验中，我们选择 $\lambda = 10^{-3}$。

我们将自己的算法与 DP-TR 进行比较。 [2] 为确保结果比较的一致性，我们修改了 DP-TR 方法，加入了对近似二阶条件的显式检查，类似于我们的算法所采用的方法。这样修改后，当这些条件满足时，DP-TR 即可终止，

我们在两种设置下进行实验：

1. 求得一个宽松解， $\epsilon_g = 0.060$ 和 $\epsilon_H \approx 0.245$。在这种设置下，我们对二阶驻点的要求较宽松。相较于所需的样本复杂度，这对应较大的样本量 n。
2. 求得一个紧致解： $\epsilon_g = 0.030$和 $\epsilon_H \approx 0.173$。在这种设置下，我们对二阶驻点的要求较严格。相较于所需的样本复杂度，我们使用较小的样本量 n。

对于每种设置，我们选择不同的隐私预算水平 ε，并使用五个不同的随机种子运行每种配置。为便于比较，我们在必要时将差分隐私方案转换为 (ε, δ)-DP。汇总结果见下表。每个表项报告五次运行所得数值的均值 $\pm$ 标准差。如果五次运行中有任何一次未能找到解，或者找到了解但因噪声而未能终止，我们将运行时间 [3] 报告为 $\times$。

表中使用以下方法缩写：TR 表示 DP-TR，OPT 表示我们提出的算法，2OPT 表示这些算法的两阶段变体，OPT-LS 表示我们提出的采用线搜索的算法，而带有“-B”的算法采用小批量处理。

实验结果表明，在隐私预算较高时 $\varepsilon = 0.6$,1.0，为求得宽松解，我们的短步算法 OPT 优于 TR，运行时间大幅缩短，最终损失也更低。在隐私预算较低时 $\varepsilon = 0.2$，尽管 OPT 可能无法成功终止，但我们发现其最终损失甚至低于 TR。原因如下：由于对下降量进行了保守估计，每次迭代的

[1]经检查，只要特征向量 x_i' 有界，损失函数就具有 Lipschitz 梯度和海森矩阵。[2]我们尝试实现 DP-GD，但使用 DP-GD 论文中描述的算法参数无法得到实际可用的结果。[3]此处的运行时间以由 Python 函数 time.perf_counter() 确定的单位表示

表1：Covertype：寻找宽松解，（ϵ_g，ϵ_H）=(0.060, 0.245)

方法	$\varepsilon = 0.2$		$\varepsilon = 0.6$		$\varepsilon = 1.0$	
	最终损失	运行时间	loss	运行时间	loss	运行时间
TR	0.729 ± 0.028	10.1 ± 9.9	0.729 ± 0.026	8.3 ± 8.6	0.729 ± 0.026	9.5 ± 9.1
TR-B	0.729 ± 0.029	2.2 ± 2.0	0.728 ± 0.027	2.2 ± 2.4	0.729 ± 0.028	2.5 ± 2.4
OPT	0.581 ± 0.057	$\times$	0.712 ± 0.018	0.6 ± 0.2	0.712 ± 0.017	0.5 ± 0.2
OPT-B	0.712 ± 0.018	3.1 ± 2.9	0.712 ± 0.018	3.2 ± 3.0	0.712 ± 0.018	2.9 ± 2.9
OPT-LS	0.577 ± 0.032	$\times$	0.687 ± 0.028	$\mathbf{0.4 \pm 0.1}$	0.699 ± 0.018	$\mathbf{0.4 \pm 0.1}$
2OPT	0.626 ± 0.078	$\times$	0.712 ± 0.017	0.6 ± 0.2	0.712 ± 0.018	0.6 ± 0.2
2OPT-B	0.712 ± 0.018	1.4 ± 0.3	0.712 ± 0.018	1.4 ± 0.4	0.712 ± 0.018	2.0 ± 1.7
2OPT-LS	0.699 ± 0.018	$\mathbf{0.5 \pm 0.2}$	0.699 ± 0.018	0.5 ± 0.2	0.699 ± 0.018	0.5 ± 0.2

表 2：Covertype：寻找紧致解：(ϵ_g, ϵ_H) =(0.030, 0.173)

方法	$\varepsilon = 0.2$		$\varepsilon = 0.6$		$\varepsilon = 1.0$	
	最终损失	运行时间	loss	运行时间	loss	运行时间
TR	0.516 ± 0.005	$\times$	0.607 ± 0.007	99.6 ± 32.2	0.607 ± 0.005	90.8 ± 21.6
TR-B	0.517 ± 0.005	$\times$	0.603 ± 0.005	32.6 ± 7.9	0.607 ± 0.003	33.4 ± 14.4
OPT	0.506 ± 0.001	$\times$	0.535 ± 0.015	$\times$	0.592 ± 0.003	1.8 ± 0.5
OPT-B	0.597 ± 0.003	$\mathbf{1.3 \pm 0.3}$	0.597 ± 0.003	1.3 ± 0.2	0.597 ± 0.003	1.4 ± 0.3
OPT-LS	0.525 ± 0.009	$\times$	0.527 ± 0.009	$\times$	0.549 ± 0.006	$\times$
2OPT	0.502 ± 0.001	$\times$	0.513 ± 0.003	$\times$	0.519 ± 0.003	$\times$
2OPT-B	0.597 ± 0.003	2.1 ± 0.4	0.597 ± 0.003	2.3 ± 0.5	0.597 ± 0.003	2.3 ± 0.6
2OPT-LS	0.577 ± 0.008	2.1 ± 1.0	0.591 ± 0.001	$\mathbf{0.6 \pm 0.1}$	0.591 ± 0.001	$\mathbf{0.8 \pm 0.2}$

隐私预算较低，因此噪声会使我们无法足够准确地检验 2S 条件。在实践中，我们可以提前停止；即使未能满足终止条件，所得解仍然可以接受。也可以采用启发式方法，额外消耗隐私预算来检验 2S 条件。

线搜索和小批量处理提升了短步算法的表现，尤其是与我们的两阶段策略结合时。我们指出，与 OPT 类似，OPT-LS 的理论最小下降量更加保守。两阶段策略采用激进估计的下降量，与线搜索相辅相成。我们观察到，2OPT-LS 在所有隐私预算水平和两种设置下都表现稳定。最后，我们指出，海森矩阵评估次数极少。详情请参见附录 E.3。

5 结论

我们基于一种用于寻找光滑非凸函数近似二阶最优点的基础算法，开发了简单的差分隐私优化算法。在非凸经验风险最小化问题上，所提出的算法基于带噪海森矩阵执行带噪梯度步或负曲率步。为了获得比保守的短步长方法更实用的方法，我们采用线搜索、小批量处理和两阶段策略。我们使用 zCDP（小批量处理时使用 Rényi 差分隐私）跟踪隐私泄露。我们的工作达到了与 DP-GD 相同的样本复杂度，但分析简单得多。尽管 DP-TR 的样本复杂度更优，其小批量版本的复杂度与我们的方法相同。我们的算法在运行时间方面显著优于 DP-TR。结合线搜索的 2OPT-LS

和两阶段策略，在数值实验中始终优于 DP-TR .

参考文献

B. Balle、G. Barthe 和 M. Gaboardi。通过子采样实现的隐私放大：通过耦合与散度进行精确分析，2018年11月。M. Bun 和 T. Steinke。集中差分隐私：简化、扩展与下界。*arXiv:1605.02065 [cs, math]*，2016年5月。Y. Carmon、J. C. Duchi、O. Hinder 和 A. Sidford。非凸优化的加速方法，2017年2月。C. Chen 和 J. Lee。差分隐私优化的随机自适应线搜索。*arXiv:2008.07978 [cs, stat]*,2014年8月。C. Dwork 和 A. Roth。差分隐私的算法基础。《理论计算机科学基础与趋势》，9(3–4):211–407，2014 年 8 月。ISSN 1551-305X。doi: 10.1561/0400000042。J. M. Kohler 和 A. Lucchi。非凸优化的子采样三次正则化。*arXiv:1705.05933 [cs, math, stat]*，2017年7月。I. Mironov。Rényi差分隐私。*2017*年*IEEE*第*30*届计算机安全基础研讨会（*CSF*），第263–275页，2017年8月。doi: 10.1109/CSF.2017.11。S. Shalev-Shwartz 和 S. Ben-David。理解机器学习：从理论到算法。剑桥大学出版社，剑桥，2014年。ISBN 978-1-107-05713-5。doi: 10.1017/CBO9781107298019。T. Tao。随机矩阵理论专题。美国数学会，2012年3月。ISBN 978-0-8218-7430-1。R. Vershynin。高维概率：数据科学中的应用导论。剑桥统计与概率数学丛书。剑桥大学出版社，剑桥，2018 年。ISBN 978-1-108-41519-4。doi: 10.1017/9781108231596。D. Wang 和 J. Xu。具有光滑非凸损失函数的差分隐私经验风险最小化：非平稳视角。*AAAI*人工智能大会论文集，33(01): 1182–1189，2019年7月。ISSN 2374-3468。doi: 10.1609/aaai.v33i01.33011182。D. Wang 和 J. Xu。通过DP信赖域方法以保护隐私且可扩展地逃离经验风险的鞍点。在 F. Hutter、K. Kersting、J. Lijffijt 和 I. Valera（编）所著的数据库中的机器学习与知识发现中，《计算机科学讲义》，第90–106页，尚，2021年。Springer国际出版。ISBN 978-3-030-67664-3。doi: 10.1007/978-3-030-67664-3_6。D. Wang、M. Ye 和 J. Xu。重新审视差分隐私经验风险最小化：更快且更通用。*arXiv:1802.05251 [cs, stat]*,2018年2月。D. Wang、C. Chen 和 J. Xu。具有非凸损失函数的差分隐私经验风险最小化。在第*36*届国际机器学习大会论文集中，第6526–6535页。PMLR，2019年5月。Y.-X. Wang、B. Balle 和 S. Kasiviswanathan。子采样Rényi差分隐私与解析矩会计器。*arXiv:1808.00087 [cs, stat]*, 2018年12月。S. J. Wright 和 B. Recht。数据分析优化。剑桥大学出版社，剑桥，2022。ISBN 978-1-316-51898-4。doi: 10.1017/9781009004282。

J. Zhang、K. Zheng、W. Mou 和 L. Wang。《平滑目标的高效私有经验风险最小化》。载于第二十六届国际人工智能联合会议论文集，第 3922–3928 页，澳大利亚墨尔本，2017年8月。国际人工智能联合会议组织。ISBN 978-0- 9992411-0-3。doi: 10.24963/ijcai.2017/548。

差分隐私简要回顾

定义 A.1 ((ε,δ)-差分隐私 [Dwork 和 Roth, 2014])。 若随机化算法 $\mathcal{A}$ 满足以下条件，则称其为 (ε,δ)-差分隐私：对于所有相邻数据集 D, D'，以及 $\mathcal{A}$的输出空间中的所有事件 S，均有：

$$\Pr\left(\mathcal{A}(D)\in S\right)\leq e^{\varepsilon}\Pr\left(\mathcal{A}(D')\in S\right)+\delta.$$

当 $\delta=0$时，我们称 $\mathcal{A}$ 满足 ε-差分隐私。在经验风险最小化设置中，如果两个数据集仅有一个数据点不同，我们就称 D' 是 D 的相邻数据集；也就是说，将 D 中的某个数据点 x_k 替换为 x'_k，即可得到数据集 D'。

Mironov 提出了 Rényi 差分隐私（RDP），作为对原始差分隐私的放宽。

定义 A.2 （Rényi 散度）。 对于定义在 $\mathcal{R}$上的两个概率分布 P 和 Q，阶数为 $\alpha>1$ 的 Rényi 散度为

$$D_\alpha(P\|Q)=\frac{1}{\alpha-1}\log \mathrm{E}_{w\sim Q}\left(\frac{P(w)}{Q(w)}\right)^{\alpha}.$$

定义 A.3 ((α,ϵ)-Rényi 差分隐私 [Mironov, 2017])。 若随机化算法 $\mathcal{M}:\mathcal{D}\to\mathcal{R}$ 满足以下条件，则称其为 (α,ϵ)-Rényi 差分隐私：对于所有相邻数据集对 D, D'，均有

$$D_\alpha(\mathcal{M}(D)\|\mathcal{M}(D'))\leq\epsilon.$$

另一种差分隐私概念是零集中差分隐私（zCDP），它要求所有阶数的散度都满足线性界。

定义 A.4 （zCDP, [Bun and Steinke, 2016]）。 若对所有相邻数据集对 D, D' 以及所有 $\alpha\in(1,\infty)$，下式均成立，则随机化算法 $\mathcal{M}:\mathcal{D}\to\mathcal{R}$ 满足（ξ,ρ）-zCDP：

$$D_\alpha(\mathcal{M}(D)\|\mathcal{M}(D'))\leq\xi+\rho\alpha.$$

等价地，若对所有 $\alpha\in(1,\infty)$，$\mathcal{M}$ 都满足（$\alpha,\xi+\rho\alpha$）-RDP，则随机化算法 $\mathcal{M}$ 满足 ρ-zCDP。若此定义对 $\xi=0$ 成立，则我们改用术语 ρ-zCDP。

RDP 和 zCDP 有一些共同性质（ [Mironov，2017；Bun and Steinke, 2016]）。

命题 A.5 (Rényi 差分隐私的组合)。 假设 $\mathcal{M}_1:\mathcal{D}\to\mathcal{R}_1$ 满足 (α,ϵ_1)-*RDP*，且 $M_2:\mathcal{R}_1\times\mathcal{D}\to\mathcal{R}_2$ 满足 (α,ϵ_2)-*RDP*。那么，定义为 (X,Y) 的机制，其中 $X\sim\mathcal{M}_1(D)$ 且 $Y\sim\mathcal{M}_2(X,D)$，满足$(\alpha,\epsilon_1+\epsilon_2)$-*RDP*。

命题 A.6 （ zCDP 的组合）。假设 $\mathcal{M}_1:\mathcal{D}\to\mathcal{R}_1$ 是 ρ_1-*zCDP*，且 $M_2:\mathcal{R}_1\times\mathcal{D}\to\mathcal{R}_2$是 ρ_2-*zCDP*。那么，定义为 （X， Y）的机制，其中 $X\sim\mathcal{M}_1$（D）且 $Y\sim\mathcal{M}_2$（X， D），是 （$\rho_1+\rho_2$）-*zCDP*。

命题 *A.7*（后处理下的保持性）。考虑映射 $\mathcal{M}:\mathcal{D}\to\mathcal{R}$ 和 $g:\mathcal{R}\to\mathcal{R}'$。根据数据处理不等式的类似结论可得 $D_\alpha(P\|Q)\geq D_\alpha(g(P)\|g(Q))$。这表明，如果 $\mathcal{M}(\cdot)$ 满足 (α,ϵ)-*Rényi* 差分隐私，那么 $g(\mathcal{M}(\cdot))$ 也满足。类似地，如果 $\mathcal{M}(\cdot)$ 满足 ρ-*zCDP*，那么 $g(\mathcal{M}(\cdot))$ 也满足。

我们可以轻松地将一种差分隐私定义转换为另一种。

命题 A.8 （Rényi 差分隐私到 （ε， δ）-差分隐私）。 **</style}如果 是一个**（α， ϵ）-*Rényi* 差分隐私机制，那么对于任意 $0<\delta<1$，它都是 $\left(\epsilon+\log_{\alpha-1}^{1/\delta},\ \delta\right)$-差分隐私。

命题 A.9 （ϵ-差分隐私到 zCDP）。 **</style}如果 是一个 -差分隐私机制，那么它也是**（$\frac{1}{2}\varepsilon^2$）-*zCDP*。

命题 A.10 （zCDP 到 （ε， δ）-差分隐私）。 **</style}假设：** $\mathcal{D}\to\mathcal{R}$ 是（ξ， ρ）-*zCDP*。那么对于所有 $\delta>0$，M也满足（ε， δ）-差分隐私，并且

$$\varepsilon=\xi+\rho+\sqrt{4\rho\log(1/\delta)}.$$

因此，对于给定的 ε和 δ，要实现(ε,δ)-差分隐私保证，只需满足以下(ξ,ρ)-*zCDP*：

$$\rho=(\sqrt{\varepsilon-\xi+\log(1/\delta)}-\sqrt{\log(1/\delta)})^2\approx\frac{(\varepsilon-\xi)^2}{4\log(1/\delta)}.$$

实现差分隐私的一种常见方法是向输出添加高斯噪声。

命题 A.11 (高斯机制)。 给定任意函数 h： $\mathcal{X}^n \to \mathbb{R}^d$，高斯机制定义如下：

$$\mathbf{G}_\sigma h(D) = h(D) + N(0, \Delta_h^2 \sigma^2 I_d),$$

其中 Δ_h 表示函数 h 的 ℓ_2-敏感度，定义为

$$\Delta_h = \sup_{D \sim D'} \|h(D) - h(D')\|. \tag{21}$$

对于所有 $\alpha \geq 1$ ，高斯机制 $\mathbf{G}_\sigma h$ 满足 $(\alpha, \alpha/(2\sigma^2))$-*Rényi* 差分隐私，因此也满足 $(1/2\sigma^2)$-*zCDP*。

关于 (ε, δ)-差分隐私的结果推迟至附录 B.2。

B 小批量算法分析

对于算法的线搜索版本，若要使用小批量损失检查充分下降条件，则需要额外的假设。为简单起见，本节仅考虑该算法的简化版本。

B.1 Rényi 差分隐私分析

如正文所述，在每次迭代 k 中，我们从 D 中无放回抽取 m 个数据点，构成小批量 S_k。我们计算集合 S_k 上的平均风险，即

$$f_{S_k} := \frac{1}{m} \sum_{i \in S_k} \ell(w_k, x_i).$$

我们以类似的方式在大小为 m 的小批量 S_k 上计算梯度和海森矩阵，分别记为 g_{S_k}、H_{S_k}，并令 $\tilde{g}_{S_k}$、H_{S_k} 分别表示它们受扰动后的版本。算法的其他部分保持不变。如式 (3) 所述，敏感度 Δ_f、Δ_g、Δ_H 将相应缩放，即将其分母中的 n 替换为小批量大小 m。本节中，我们用 $s = m/n$ 表示采样比例。

令 g_k 和 H_k 分别表示在完整数据集 D 上计算的梯度和海森矩阵。我们可以将它们的含噪近似的偏差分解如下，

$$\|\tilde{g}_{S_k} - g_k\| \leq \|\tilde{g}_{S_k} - g_{S_k}\| + \|g_{S_k} - g_k\|, \tag{22a}$$

$$\left\|\tilde{H}_{S_k} - H_k\right\| \leq \left\|\tilde{H}_{S_k} - H_{S_k}\right\| + \|H_{S_k} - H_k\|, \tag{22b}$$

其中，界中的第一项源于添加的高斯噪声，第二项则源于子采样。我们将以高概率分别界定这两项。

W我们有以下来自 [Kohler 和 Lucchi， 2017]的子采样集中性结果

引理 B.1 （梯度偏差界）。 我们以至少为 $1 - \eta$ 的概率有

$$\|g_{S_k} - g_k\| \leq 4\sqrt{2} B_g \sqrt{\frac{\log(2d/\eta) + 1/4}{|S_k|}}.$$

引理 B.2 (海森矩阵偏差界)。 我们以至少 $1 - \eta$ *th* 的概率有 *at*

$$\|H_{S_k} - H_k\| \leq 4 B_H \sqrt{\frac{\log(2d/\eta)}{|S_k|}}.$$

对于子采样误差 (22)，我们使用前文所述的高斯集中结果 13 来界定第一项，并使用上文给出的子采样结果来界定第二项。对于第 k 次迭代，以至少 $(1-\frac{\zeta}{T})(1-C\exp(-C_1Cd))(1-\eta/T)^2$, 的概率，我们有

$$\begin{aligned}\|\tilde{g}_{S_k}-g_k\| &\le \sqrt{2d}\eta_g\sigma_g\log\frac{T}{\zeta}+4\sqrt{2}B_g\sqrt{\frac{\log(2dT/\eta)+1/4}{|S_k|}},\\ \left\|\tilde{H}_{S_k}-H_k\right\| &\le C\sqrt{d}\eta_H\sigma_H+4B_H\sqrt{\frac{\log(2dT/\eta)}{|S_k|}}.\end{aligned} \tag{23}$$

只需要求上述界右侧的每一项都不超过式 (9) 右侧对应项的 $1/2$，这样我们就可以采用类似的分析。对于子采样造成的偏差，我们需要

$$4\sqrt{2}B_g\sqrt{\frac{\log(2dT/\eta)+1/4}{|S_k|}} \le \frac{1}{2}\min\left(c_1\epsilon_g,\frac{c_2}{M}\epsilon_H^2\right), \tag{24a}$$

$$4B_H\sqrt{\frac{\log(2dT/\eta)}{|S_k|}} \le \frac{1}{2}c\,\epsilon_H. \tag{24b}$$

重新整理各项，可得关于小批量大小的条件：

$$|S_k| \ge \max\left(64B_g^2(\log(2dT/\eta)+1/4)\max\left(c_1^{-2}\epsilon_g^{-2},\frac{M^2}{c_2^2}\epsilon_H^{-4}\right),32B_H^2\log(2dT/\eta)c^{-2}\epsilon_H^{-2}\right). \tag{25}$$

根据与全批量情形相同的分析（参见定理 3.13），以下收敛结果显然成立。

定理 B.3. 以至少 $\{(1-\zeta/T)(1-C\exp(-C_1Cd))(1-\eta/T)^2\}^T$ 的概率，假设样本数量 n 满足 $n\ge n_{\min}$，其中

$$\begin{aligned}n_{\min} := \max\Bigg(&\frac{2\sqrt{2d}B_g\sigma_g\log\frac{T}{\zeta}}{\min\left(c_1\epsilon_g,\frac{c_2}{M}\epsilon_H^2\right)},\frac{2C\sqrt{d}B_H\sigma_H\log\frac{T}{\zeta}}{c\,\epsilon_H},\\ &s^{-1}64B_g^2(\log(2dT/\eta)+1/4)\max\left(c_1^{-2}\epsilon_g^{-2},\frac{M^2}{c_2^2}\epsilon_H^{-4}\right),s^{-1}32B_H^2\log(2dT/\eta)c^{-2}\epsilon_H^{-2}\Bigg).\end{aligned} \tag{26}$$

小批量短步算法的输出是一个$((1+c_1)\epsilon_g,(1+c)\epsilon_H)$-二阶驻点。取(8)中的 σ，忽略对数项和常数后，$n_{\min}$ 对(ϵ_g, ϵ_H) 及 ρ 的渐近依赖关系为

$$n_{\min}=\frac{\sqrt{d}}{\sqrt{\rho}}\tilde{O}\left(\max\left(\epsilon_g^{-2},\epsilon_g^{-1}\epsilon_H^{-2},\epsilon_H^{-7/2}\right)\right). \tag{27}$$

现在我们来讨论隐私保证。使用 z-CDP 无法处理子采样问题，但在 RDP 框架下，Wang 等人 [2018b] 给出了针对子采样的广义分析：

定理 B.4 (子采样机制的 Rényi 差分隐私)。给定一个包含 n 个从定义域 $\mathcal{X}$ 中抽取的点的数据集，以及一个（随机化）机制 $\mathcal{M}$，该机制针对 $m\le n$ 接收来自 $\mathcal{X}^m$ 的输入。定义随机化算法$\mathcal{M}\circ$子采样如下：（*1*）子采样：从数据集中无放回地子采样 *m* 个数据点（采样比例为 $s=m/n$）；（*2*）应用 $\mathcal{M}$：一个以子采样后的数据集为输入的随机化算法。对于所有整数 $\alpha\ge 2$，如果 $\mathcal{M}$ 是 $(\alpha,\epsilon(\alpha))$-***Rényi*** 差分隐私，那么这个新的随机化算法 $\mathcal{M}\circ$ 子采样

满足 $(\alpha,\ \epsilon'(\alpha))$-*Rényi* 差分隐私，其中，

$$\epsilon'(\alpha) \le \frac{1}{\alpha-1}\log\left(1+s^2\binom{\alpha}{2}\min\left\{4\left(e^{\epsilon(2)}-1\right), e^{\epsilon(2)}\min\left\{2,\left(e^{\epsilon(\infty)}-1\right)^2\right\}\right\} + \sum_{j=3}^{\alpha} s^j\binom{\alpha}{j}e^{(j-1)\epsilon(j)}\min\left\{2,\left(e^{\epsilon(\infty)}-1\right)^j\right\}\right). \quad (28)$$

对于高斯机制，我们有

$$\epsilon(\alpha) = \frac{\alpha}{2\sigma^2},$$

因此 $\epsilon_{\mathcal{M}(\infty)} = \infty$，界可简化为

$$\epsilon'(\alpha) \le \frac{1}{\alpha-1}\log\left(1+s^2\binom{\alpha}{2}\min\left\{4\left(e^{1/\sigma^2}-1\right), 2e^{1/\sigma^2}\right\} + \sum_{j=3}^{\alpha} 2s^j\binom{\alpha}{j}e^{(j-1)j/(2\sigma^2)}\right) =: \epsilon'_{\mathcal{N}}(\alpha;\sigma,s),$$

我们将其记作 $\epsilon'_{\mathcal{N}}(\alpha;\ \sigma, s)$。当 s 较小且 σ 较大时，可以忽略高阶项，并将右侧写为

$$\epsilon'_{\mathcal{N}}(\alpha;\sigma,s) \approx \frac{1}{\alpha-1}\left(s^2\frac{\alpha(\alpha-1)}{2}\cdot 4\frac{1}{\sigma^2}\right) = 2s^2\frac{\alpha}{\sigma^2}, \quad (29)$$

其中，对于较小的 t，我们采用近似 $e^t \approx 1+t$。

定理 B.5. 考虑使用子采样的算法短步版本。给定σ_f，σ_g，σ_H, λ以及采样比例 s。假设子采样算法的一次实际运行进行了 k_g次梯度步和 k_H 次负曲率步。该次运行满足 $(\alpha,\ \epsilon(\alpha))$-*Rényi* 差分隐私，其中

$$\epsilon(\alpha) = \frac{\alpha}{2\sigma_f^2} + k_g\epsilon'_{\mathcal{N}}(\alpha;\sigma_g,s) + k_H\epsilon'_{\mathcal{N}}(\alpha;\sigma_{g,H},s),$$

其中$1/\sigma_{g,\ H}^2 = 1/\sigma_g^2 + 1/\sigma_H^2$。令

$$\bar{\epsilon}(\alpha) = \frac{\alpha}{2\sigma_f^2} + T\epsilon'_{\mathcal{N}}(\alpha;\sigma_g,s) + T\epsilon'_{\mathcal{N}}(\alpha;\sigma_{g,H},s). \quad (30)$$

我们始终有 $\bar{\epsilon}(\alpha) \ge \epsilon(\alpha)$，因此该算法满足 $(\alpha,\bar{\epsilon}(\alpha))$-*Rényi* 差分隐私。

证明。 证明可直接由上述 RDP 采样定理和 Rényi 差分隐私的组合得出。 □

鉴于子采样隐私保证 $\epsilon'_{\mathcal{N}}(\alpha;\ \sigma, s)$ 的复杂性，我们无法给出设置参数 $\sigma_f, \sigma_g, \sigma_H$ 的显式公式。不过，给定 (ε,δ)-差分隐私预算，我们可以优化参数以满足隐私保证。回顾从 $(\alpha,\epsilon(\alpha))$-Rényi 差分隐私到 $(\varepsilon_{DP},\delta_{DP})$-差分隐私的转换，给定 δ，我们求解

$$\varepsilon_{DP}(\epsilon(\cdot)) = \min_{\alpha}\left(\epsilon(\alpha) + \frac{\log 1/\delta_{DP}}{\alpha-1}\right).$$

因此，我们 可以优化参数 $\sigma_f, \sigma_g, \sigma_H$，使下列目标函数最小化

$$\max\left(\bar{\varepsilon}_{DP} - \varepsilon_{DP}(\bar{\epsilon}(\cdot)), 0\right), \quad (31)$$

其中 $\bar{\varepsilon}_{DP}$ i 是目标隐私预算，我们将 $\bar{\epsilon}(\cdot)$ 替换为其对应版本（30） .

使用(ε, δ)-差分隐私的样本复杂度

在 (ε, δ)-差分隐私方案下，子采样更容易处理，我们将推导出一个样本复杂度界。首先，我们介绍 (ε, δ)-差分隐私中的若干有用结果。

命题 *B.6*（(ε, δ)-差分隐私的组合）。假设 $\mathcal{M}_1 : \mathcal{D} \to \mathcal{R}_1$ 是 $(\varepsilon_1, \delta_1)$-差分隐私，且 $M_2 : \mathcal{R}_1 \times \mathcal{D} \to \mathcal{R}_2$ 是 $(\varepsilon_2, \delta_2)$-差分隐私。那么，定义为 (X, Y) 的机制，其中 $X \sim \mathcal{M}_1(D)$ 和 $Y \sim \mathcal{M}_2(X, D)$，是 $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-差分隐私。

定义 B.7（(ε, δ)-差分隐私的高斯机制）。给定任意函数 $h : \mathcal{X}^n \to \mathbb{R}^d$，高斯机制定义如下：

$$\mathbf{G}_\sigma h(D) = h(D) + N(0, \Delta_h^2 \sigma^2 I_d),$$

其中，Δ_h 是函数 h 和 $\sigma \geq \frac{\sqrt{2\ln(1.25/\delta)}\Delta_h}{\epsilon}$ 的 ℓ_2-敏感度。于是，高斯机制 $\mathrm{G}_\sigma h$ 满足 (ϵ, δ)-差分隐私。

定理 B.8（通过子采样实现隐私放大 [Balle 等人， 2018]）。。给定一个包含 n 个点的数据集，这些点从定义域 $\mathcal{X}$ 中抽取；再给定一个（随机化）机制 $\mathcal{M}$，它对来自 $\mathcal{X}^m$ 的输入进行 $m \leq n$，将随机化算法 $\mathcal{M} \circ$ 子采样 定义如下：*(1)* 子采样：从数据集中无放回地子采样 m 个数据点（采样参数为 $s = m/n$）；*(2)* 应用 $\mathcal{M}$：将子采样后的数据集作为输入的随机化算法。若 $\mathcal{M}$ 是(ε, δ)-差分隐私，则 $\mathcal{M} \circ$子采样 是(ε', δ')-差分隐私，其中$\varepsilon' = \log(1 + s(e^\varepsilon - 1)) \leq s(e^\varepsilon - 1))$ 且 $\delta' = s\delta$。

定理 B.9 (高级组合)。对于所有 ε_0， δ_0， $\delta_0' \geq 0$，$(\varepsilon_0, \delta_0)$-差分隐私机制类在进行 k 重自适应组合后满足$(\varepsilon, k\delta_0 + \delta_0')$-差分隐私，条件如下：

$$\varepsilon = \sqrt{2k \ln(1/\delta_0')}\varepsilon + k\varepsilon(e^\varepsilon - 1).$$

作为推论，对于 $0 < \varepsilon < 1$，只需选择 $\epsilon_0 = \frac{\varepsilon}{2\sqrt{2k\log(1/\delta_0')}}$ 即可确保组合满足$(\varepsilon, k\delta_0 + \delta_0')$-差分隐私。具体而言，我们还可以选择 $\delta_0' = \delta/2$ 和 $\delta_0 = \delta/(2k)$，使其满足(ε, δ)-差分隐私。

对于使用无放回抽样的算法，我们有以下隐私保证：

定理 *B.10*。考虑使用子采样的算法的短步版本。给定隐私参数 $\varepsilon, \delta, \varepsilon_f, \delta_f \in (0, 1)$，满足 $\varepsilon_f < \varepsilon$ 和 $\delta_f < \delta$，以及子采样参数 s。令 $\varepsilon_0 = (\varepsilon - \varepsilon_f)/(8s\sqrt{2T\ln(2/(\delta - \delta_f))})$ 和 $\delta_0 = (\delta - \delta_f)/(4sT)$，其中 T按式 *(6)* 中的方法估算。 $\sigma_f = \sqrt{2\ln(1.25/\delta_f)}/\varepsilon_f$，$\sigma_g = \sigma_H = \frac{\sqrt{2\ln(1.25/\delta_0)}}{\epsilon_0}$。该算法满足$(\varepsilon, \delta)$-差分隐私。

证明。根据高斯机制，估计 T 的步骤满足 $(\varepsilon_f, \delta_f)$-差分隐私。只需证明其余步骤满足 $(\varepsilon - \varepsilon_f, \delta - \delta_f)$-差分隐私。根据高级组合，我们只需证明每次迭代满足 $(4s\varepsilon_0, 2s\delta_0)$-差分隐私。

考虑一次不使用子采样的迭代。根据高斯机制和稀疏向量技术的使用方式，我们知道，计算扰动梯度步骤和扰动海森矩阵步骤都满足 $(\varepsilon_0, \delta_0)$-差分隐私。根据组合，我们知道整个迭代满足 $(2\varepsilon_0, 2\delta_0)$-差分隐私。应用隐私放大定理 B.8，我们知道，使用子采样的每次迭代满足 $(4s\varepsilon_0, 2s\delta_0)$-差分隐私。 □

现在讨论样本复杂度。

定理 B.11. 对于 $c_f \in (0, 1)$，设置 $\varepsilon_f = c_f\varepsilon$ 和 $\delta_f = c_f\delta$，在定理 *B.3* 中选择参数 σ_g, σ_H 时，定理 *B.3* 中 $n_{\min}$ 对 (ϵ_g, ϵ_H)、(ε, δ)、 的渐近依赖关系为

$$n_{\min} = \tilde{O}\left(\frac{\sqrt{d}\ln(1/\delta)}{\varepsilon}\max\left(\epsilon_g^{-2}, \epsilon_g^{-1}\epsilon_H^{-2}, \epsilon_H^{-4}\right)\right). \tag{32}$$

当 $(\epsilon_g, \epsilon_H) = (\alpha, \sqrt{M\alpha})$ 时，这种依赖关系简化为 $\tilde{O}(d/\ln(1/\delta)\varepsilon\alpha^2)$，与通过 $\sqrt{\rho} = O(\frac{\varepsilon}{\sqrt{\ln(1/\delta)}})$，利用命题 *A.10* 将 ρ-*zCDP* 转换为(ε, δ)-差分隐私后，全批量版本算法的结果一致（相差一个 $\sqrt{\ln(1/\delta)}$ 因子）。

证明。与之前一样，我们有 $\sqrt{T} = O(\max(\epsilon_g^{-1}, \epsilon_H^{-3/2}))$。化简后，$\sigma_g$ 和 σ_H 的阶为

$$\frac{s}{\varepsilon}\sqrt{T}\sqrt{\ln(2sT/\delta)\ln(1/\delta)} = \tilde{O}\left(\frac{s}{\varepsilon}\max(\epsilon_g^{-1}, \epsilon_H^{-3/2})\right).$$

$n_{\min}$ 的渐近依赖关系可通过将 σ_g 和 σ_H 代入式（26）得出。 □

C 使用 Lanczos 方法计算最小特征值

在我们的算法中，需要计算扰动海森矩阵的最小特征值。使用随机化 Lanczos 算法可以高效地完成这一计算。根据 [Carmon 等人 2017]，我们有以下结果

引理 C.1。假设使用 *Lanczos* 方法估计 H 的最小特征值，初始向量是在单位球面上均匀随机生成的向量，其中 $\|H\| \leq M$。对于任意 $\delta \in [0, 1)$，该方法能以不超过下述次数，将 H 的最小特征值估计到绝对精度 $\epsilon/2$，并同时得到一个对应方向 v

$$\min\left\{d, 1 + \left\lceil \frac{1}{2}\ln\left(2.75d/\delta^2\right)\sqrt{\frac{M}{\epsilon}}\right\rceil\right\} \ iterations$$

以至少 $1-\delta$ 的概率。

~ ~在我们的算法中使用 Lanczos 方法时，只要满足 $\lambda \leq -\epsilon_H/2$，我们就输出 $\lambda_{\min}(H)$ 的一个估计值 λ，以及相应的特征向量。如果 $\lambda > -\epsilon_H/2$，我们就判定 $\lambda_{\min}(H) \geq -\epsilon_H$，错误概率至多为 δ_L。将 ϵ_H 替换为 $\epsilon_H/2$，并将 Lanczos 算法的成功概率 $1-\delta_L$ 乘入每次迭代的成功概率之积后，我们的分析和收敛结果仍然成立。

请注意，Lanczos 方法只需要计算 Hessian-向量乘积，其计算成本取决于目标函数的具体形式。对于线性经验风险最小化，每一项的形式为 $l(w^T x_i)$，因此海森矩阵是秩一项的加权和。因此，使用真实 Hessian 矩阵计算 Hessian-向量乘积需要 $O(nd)$ 次运算，而将添加的噪声矩阵与向量相乘则额外需要 $O(d^2)$ 次运算。因此，随机化 Lanczos 方法的总成本取决于所需精度，为上述迭代次数上界的 $O(nd + d^2)$ 倍。相比之下，完整计算一次 Hessian 矩阵的成本可能达到 $O(nd^2)$，而进行特征值分解则需要 $O(d^3)$。[4]

D 缺失的证明

D.1 引理 3.3 的证明

引理 D.1. 采用短步步长选择(7)时，若噪声满足以下条件，其中 c、c_1 和 c_2 为正数常量，且 $c_1 < \frac{1}{2}$ 和 $c_2 + c < \frac{1}{3}$，

$$\|\varepsilon_k\| \leq \min\left(c_1\epsilon_g, \frac{c_2}{M}\epsilon_H^2\right), \tag{33a}$$

$$\|E_k\| \leq c\,\epsilon_H, \tag{33b}$$

则每一步的下降量至少为

$$\text{MIN_DEC} = \min\left(\frac{1-2c_1}{2G}\epsilon_g^2,\ 2\left(\frac{1}{3} - c_2 - c\right)\frac{\epsilon_H^3}{M^2}\right). \tag{34}$$

[4]我们仍 ev 在我们的实现中计算含噪海森矩阵，因为 PyTorch 尚未对此提供优化支持 .

真实梯度和海森矩阵的真实最小特征值满足以下条件 ，

$$\|g_k\| \leq (1+c_1)\|\tilde{g}_k\|, \quad \lambda_k > -(1+c)|\tilde{\lambda}_k|. \tag{35}$$

证明。 我们将使用 以下两个标准界，它们可由关于 f 的光滑性假设推出 ：

$$f(w+p) \leq f(w) + \nabla f(w)^\top p + \frac{G}{2}\|p\|^2, \tag{36}$$

$$f(w+p) \leq f(w) + \nabla f(w)^T p + \frac{1}{2}p^T\nabla^2 f(w)p + \frac{1}{6}M\|p\|^3. \tag{37}$$

为简便起见，在下面的分析中我们省略迭代编号 k 。

对于梯度步，我们有 $\|\tilde{g}\| > \epsilon_g$。 记 $g = \tilde{g} - \varepsilon$。 利用 $\|\varepsilon\| \leq c_1\epsilon_g < c_1\|\tilde{g}\|$，由 (36) 可得

$$\begin{aligned}
f(w-\gamma_g\tilde{g}) &\leq f - \gamma_g(\tilde{g}-\varepsilon)^T\tilde{g} + \frac{G}{2}\gamma_g^2\|\tilde{g}\|^2 \\
&\leq f - \frac{1}{G}(\tilde{g}-\varepsilon)^T\tilde{g} + \frac{1}{2G}\|\tilde{g}\|^2 \\
&\leq f - \frac{1}{2G}\|\tilde{g}\|^2 + \frac{1}{G}\|\varepsilon\|\,\|\tilde{g}\| \\
&\leq f - \frac{1}{2G}\|\tilde{g}\|^2 + \frac{1}{G}c_1\|\tilde{g}\|^2 \\
&= f - \frac{1}{2G}(1-2c_1)\|\tilde{g}\|^2 \\
&\leq f - \frac{1}{2G}(1-2c_1)\varepsilon_g^2,
\end{aligned}$$

而真实梯度满足

$$\|g\| \leq \|\tilde{g}\| + \|\varepsilon\| \leq (1+c_1)\|\tilde{g}\|.$$

采取负曲率步时，我们有 $\tilde{\lambda} < -\epsilon_H$。根据假设，有 $\|\varepsilon\| \leq \frac{c_2}{M}\epsilon_H^2 < \frac{c_2}{M}|\tilde{\lambda}|^2$
~ ~和 $\|E\| \leq c\,\epsilon_H < c|\lambda|$。回顾 $\tilde{p}$ 的定义 (5)，并记 $g = \tilde{g} - \varepsilon$、$H = H - E$。由 (37)，我们有

$$\begin{aligned}
f(w+\gamma_H\tilde{p}) &\leq f + \gamma_H g^T\tilde{p} + \frac{1}{2}\gamma_H^2\tilde{p}^T H\tilde{p} + \frac{1}{6}M\gamma_H^3\|\tilde{p}\|^3 \\
&= f + \gamma_H\tilde{g}^T\tilde{p} + \frac{1}{2}\gamma_H^2\tilde{p}^T\tilde{H}\tilde{p} + \frac{1}{6}M\gamma_H^3\|\tilde{p}\|^3 - \gamma_H\varepsilon^T\tilde{p} - \frac{1}{2}\gamma_H^2\tilde{p}^T E\tilde{p} \\
&\leq f + \frac{1}{2}\left(\frac{2|\tilde{\lambda}|}{M}\right)^2(-|\tilde{\lambda}|) + \frac{1}{6}M\left(\frac{2|\tilde{\lambda}|}{M}\right)^3 - \frac{2|\tilde{\lambda}|}{M}\varepsilon^T\tilde{p} - \frac{1}{2}\left(\frac{2|\tilde{\lambda}|}{M}\right)^2\tilde{p}^T E\tilde{p} \\
&\leq f - \frac{2}{3}\frac{|\tilde{\lambda}|^3}{M^2} + \frac{2|\tilde{\lambda}|}{M}\|\varepsilon\| + \frac{2|\tilde{\lambda}|^2}{M^2}\|E\| \\
&\leq f - \left(\frac{2}{3} - 2c_2 - 2c\right)\frac{|\tilde{\lambda}|^3}{M^2} \\
&\leq f - \left(\frac{2}{3} - 2c_2 - 2c\right)\frac{\epsilon_H^3}{M^2},
\end{aligned}$$

前提是 $c_2 + c < 1/3$。令 λ 表示 H 的最小特征值。由韦尔不等式可得

$$|\tilde{\lambda} - \lambda| \leq \|E\| \leq c|\tilde{\lambda}|,$$

因此，

$$\lambda > \tilde{\lambda} - c|\tilde{\lambda}| \geq -(1+c)|\tilde{\lambda}|.$$

□

D.2 推论 3.5 的证明

推论 D.2. 假设噪声在每次迭代中都满足(9)，则算法的短步版本（使用(7)）将输出一个$((1+c_1)\epsilon_g, (1+c)$ (10)$)$ $\epsilon_H)$-二阶驻点。

证明。 根据刚刚推导出的最小下降量 (10)，可知算法将在 T^* 次迭代后终止，其中

$$T^* = \frac{f(w_0) - f^*}{\text{MIN_DEC}}.$$

我们在 (6) 中对 T 的选择是 T^* 的上界，因此算法将在 T 次迭代内停止。当算法在第 ˜ 次迭代 k 时停止，我们有 $\|\tilde{g}_k\| \le \epsilon_g$ 和 $\lambda_k \ge -\epsilon_H$。由 (11) 可知，输出是一个 $((1+c_1)\epsilon_g, (1+c)\epsilon_H)$-二阶驻点。

□

D.3 引理 3.11 的证明

引理 D.3. 假设与之前相同的有界噪声条件 (9) 。取充分下降系数 $c_g \in (0, 1-c_1)$ $c_H \in (0, 1-c-\sqrt{\frac{8}{3}c_2})$，令 $\bar{\gamma}_g = 2(1-c_1-c_g)/G$ 和 $\bar{\gamma}_H = t_2|\tilde{\lambda}|/M$ 如算法 *3* 中所定义，则当 $\gamma \le \bar{\gamma}_g$ 和 $\gamma \in [(t_1/t_2)\bar{\gamma}_H, \bar{\gamma}_H]$ 时，充分下降条件 (SD1) 和 (SD2) 分别成立。其中 $0 < t_1 < t_2$ 是以下二次方程的解（根据我们对 c, c_2, c_H 的选择，存在实数解），

$$r(t) := -\frac{1}{6}t^2 + \frac{1}{2}(1-c-c_H)t - c_2 = 0,$$

具体而言，我们有

$$t_1, t_2 = \frac{3}{2}(1-c-c_H) \pm 3\sqrt{\frac{1}{4}(1-c-c_H)^2 - \frac{2}{3}c_2}. \tag{38}$$

特别地，我们有 $q_g(\bar{\gamma}_g) \ge 0$ 且 $q_H(\bar{\gamma}_H) \ge 0$。

证明。 分析与 D.1 类似。为简便起见，我们再次省略迭代下标 k。

对于梯度步，我们有 $\|\tilde{g}\| > \epsilon_g$。记 $g = \tilde{g} - \varepsilon$。利用 $\|\varepsilon\| \le c_1\epsilon_g < c_1\|\tilde{g}\|$，由 (36) 可得

$$\begin{aligned} f(w-\gamma\tilde{g}) &\le f(w) - \gamma(\tilde{g}-\varepsilon)^\top\tilde{g} + \frac{G}{2}\gamma^2\|\tilde{g}\|^2 \\ &\le f(w) - \left(\gamma - \frac{G}{2}\gamma^2\right)\|\tilde{g}\|^2 + \gamma\|\tilde{g}\|\|\varepsilon\| \\ &\le f(w) - \gamma\left(1 - \frac{G}{2}\gamma - c_1\right)\|\tilde{g}\|^2, \end{aligned}$$

根据 $\bar{\gamma}_g$ 的定义，当 $\gamma \le \bar{\gamma}_g$ 且.....时，(SD1) 成立

采取负曲率步时，我们有 $\|\varepsilon\| \le \frac{c_2}{M}\epsilon_H^2 < \frac{c_2}{M}|\tilde{\lambda}|^2$˜ ˜和 $\|E\| \le c\,\epsilon_H < c|\lambda|$。回顾 $\tilde{p}$ 的定义 (5)，并记作 $g = \tilde{g} - \varepsilon$, $H = H - E$。由 (37) 可知，对于 $\gamma > 0$,

$$\begin{aligned} f(w+\gamma\tilde{p}) &\le f(w) + \gamma\tilde{g}^T\tilde{p} + \frac{1}{2}\gamma^2\tilde{p}^T H\tilde{p} + \frac{1}{6}M\gamma^3\|\tilde{p}\|^3 - \gamma\varepsilon^T\tilde{p} - \frac{1}{2}\gamma^2\tilde{p}^T E\tilde{p} \\ &\le f(w) - \frac{1}{2}\gamma^2|\tilde{\lambda}| + \frac{1}{6}M\gamma^3 + \gamma\|\varepsilon\| + \frac{1}{2}\gamma^2\|E\| \\ &\le f(w) - \underbrace{\left(\frac{1}{2}\gamma^2(1-c)|\tilde{\lambda}| - \gamma\frac{c_2}{M}|\tilde{\lambda}|^2 - \frac{1}{6}M\gamma^3\right)}_{g(\gamma)}. \end{aligned}$$

通过重新参数化 $\gamma = \frac{t|\lambda|}{M}$，我们得到

$$g(\gamma) - \frac{1}{2}c_H\gamma^2|\tilde{\lambda}| = \left(-\frac{1}{6}t^3 + \frac{1}{2}(1-c-c_H)t^2 - c_2 t\right)\frac{|\tilde{\lambda}|^3}{M^2} = t\cdot r(t)\frac{|\tilde{\lambda}|^3}{M^2}.$$

注意，当 $g(\gamma) \geq \frac{1}{2}c_H\gamma^2|\tilde{\lambda}|$ 时，(SD2) 成立。该结果由以下事实得出：对于 $t \in [t_1, t_2]$，$r(t) \geq 0$。

□

D.4 引理 3.12 的证明

引理 D.4. 使用 *DP* 线搜索算法 *3*，并假设与前文相同的有界噪声条件(9) 。取充分下降系数 $c_g \in (0, 1-c_1)$ $c_H \in (0, 1-c-\sqrt{\frac{8}{3}c_2})$，并按前文定义 $\bar{\gamma}_g$ 和 $\bar{\gamma}_H$。选择初始步长乘数 $b_g, b_H > 1$，以及下降参数 $\beta_g \in (0,1)$ $\beta_H \in (t_1/t_2, 1)$。令 $i_{\max} = \lfloor \log^{\beta} \max(b_g, b_H) \rfloor + 1$。若 n 满足以下条件：

$$n \geq 16\lambda\left(\log i_{\max} + \log(T/\xi)\right)\max\left(2b_g\frac{B_g}{c_g\epsilon_g}, 4b_H\frac{B_g M}{t_2 c_H \epsilon_H^2}\right), \tag{39}$$

以至少 $1-\xi/T$ 的概率，单步的下降量至少为

$$\text{MIN_DEC} = \min\left(\frac{1}{G}(1-c_1-c_g)c_g\epsilon_g^2, \frac{1}{4}c_H t_2^2\frac{\epsilon_H^3}{M^2}\right). \tag{40}$$

证明。 对于经验风险最小化设置中的损失函数 l，根据中值定理，存在$t \in$ (0,1)，使得

$$|l(w,x) - l(w-\gamma\tilde{g}, x)| = |\gamma\tilde{g}^T\nabla_w l(w - t\gamma\tilde{g}, x)| \leq \gamma B_g\|\tilde{g}\|.$$

我们首先分析 q_g 和 q_H 的敏感度。根据 (14a) 中 $q_g(\gamma, w)$ 的定义以及上面的界

$$\begin{aligned}
|q_g(\gamma, w, D) - q_g(\gamma, w, D')| &= \frac{1}{n}|l(w, x_k) - l(w-\gamma\tilde{g}, x_k) - l(w, x_k') - l(w-\gamma\tilde{g}, x_k')| \\
&\leq \frac{1}{n}|l(w, x_k) - l(w-\gamma\tilde{g}, x_k)| + |l(w, x_k') + l(w-\gamma\tilde{g}, x_k')| \\
&\leq \frac{2}{n}\gamma B_g\|\tilde{g}\|.
\end{aligned}$$

我们有 $\gamma \leq \gamma_g^{\text{init}} = b_g\bar{\gamma}_g$。因此，对于 q_g 的敏感度，有

$$\Delta_{q_g} = \frac{2}{n}\gamma_g^{\text{init}} B_g\|\tilde{g}\| = \frac{2}{n}b_g\bar{\gamma}_g B_g\|\tilde{g}\|. \tag{41}$$

对于 q_H，使用定义 (14b)，并采用与上文类似的论证，我们得到敏感度的以下上界：

$$\Delta_{q_H} = \frac{2}{n}\gamma_H^{\text{init}} B_g\|p\| = \frac{2}{n}b_H\bar{\gamma}_H B_g. \tag{42}$$

根据稀疏向量技术的性质（参见 Dwork 和 Roth [2014, 定理 3.24]），我们知道，以至少 $1-\xi/T$ 的概率，线搜索将输出一个 γ，使得

$$q(\gamma) \geq -t\Delta_q, \quad \text{where } t = 8\lambda\left(\log i_{\max} + \log(T/\xi)\right).$$

这里， $i_{\max}$ 是差分隐私线搜索的最大迭代次数；对于梯度步和负曲率步，该值分别应为 $\lfloor \log_\beta b_g \rfloor + 1$ 和$\lfloor \log_\beta b_H \rfloor + 1$。我们取两者中的最大值作为 $i_{\max}$，这样界限在两种情况下都成立。（如果最终步长采用回退值，那么最终

输出 $\bar{\gamma}$ 将自动满足上述条件，因为根据定义，我们有 $q(\bar{\gamma}) \geq 0$。）使用 t重写关于 n 的假设，

$$2tb_g\frac{B_g}{n} \leq \frac{1}{2}c_g\epsilon_g, \quad 2tb_H\frac{B_g}{n} \leq \frac{1}{4}c_H\epsilon_H \cdot \frac{t_2\epsilon_H}{M}.$$

结合 (41) 和 (42) 可得

$$t\Delta_{q_g} = 2tb_g\frac{B_g}{n}\bar{\gamma}_g\|\tilde{g}\| \leq \frac{1}{2}c_g\epsilon_g\bar{\gamma}_g\|\tilde{g}\| \leq \frac{1}{2}c_g\bar{\gamma}_g\|\tilde{g}\|^2,$$
$$t\Delta_{q_H} = 2tb_H\frac{B_g}{n}\bar{\gamma}_H \leq \frac{1}{4}c_H\epsilon_H\bar{\gamma}_H \cdot \frac{t_2\epsilon_H}{M} \leq \frac{1}{4}c_H\epsilon_H\bar{\gamma}_H^2,$$

˜这里分别使用了梯度步和负曲率步的前提条件 $\|\tilde{g}\| > \epsilon_g$ 和 $\lambda < -\epsilon_H$、-，以及根据定义有 $\bar{\gamma}_H = \frac{t_2|\lambda|}{M} \geq \frac{t_2\epsilon_H}{M}$这一事实。对于输出 γ，我们总有 $\gamma \geq \bar{\gamma}$，因为 γ最多下降 $\lfloor \log_\beta \max b \rfloor + 1$ 次。因此，下降量至少为

$$c_g\gamma_g\|\tilde{g}\|^2 - t\Delta_{q_g} \geq \frac{1}{2}c_g\bar{\gamma}_g\|\tilde{g}\|^2 \geq \frac{1}{2}c_g\bar{\gamma}_g\epsilon_g^2 = \frac{1}{G}(1-c_1-c_g)c_g\epsilon_g^2,$$
$$\frac{1}{2}c_H\gamma_H^2\|\tilde{\lambda}\| - t\Delta_{q_H} \geq \frac{1}{4}c_H\epsilon_H\bar{\gamma}_H^2 \geq \frac{1}{4}c_Ht_2^2\frac{\epsilon_H^3}{M^2},$$

分别适用于梯度步和负曲率步。对右侧两个量取最小值即可得到该结果。

□

E 实验设置及其他实验

算法使用 PyTorch 实现。对于小批量算法中使用的 Rényi 差分隐私隐私核算，我们使用 autodp 软件包[5]。所有实验均在配备 36 核 Intel Xeon Gold 6254 3.1GHz CPU 的集群上进行，每次运行使用 8 个 CPU 核心。

E.1 数据集

Covertype 数据集[6] 包含 $n = 581012$ 个数据点。每个数据点的形式为 (x, y)，其中 x 是一个 54 维特征向量（前 10 列为数值维度，第 11–14 列是荒野区域独热向量，最后 40 列是土壤类型独热向量）， y 是标签，取值为$\{1, 2, \ldots, 7\}$中的一个。

对于预处理，我们对前 10 个数值列进行归一化，并仅保留满足$y = 1$、2 的样本。该受限集合中剩余的样本数为 $n = 495141$。我们将 $y = 2$ 重新编码为 $y = -1$，使得 $y \in \{-1, 1\}$。

IJCNN 数据集[7] 包含 $n = 4999$ 个数据点。每个数据点由 (x, y) 组成，其中 x 是一个 22 维特征向量（前 10 列为独热向量，第 11–22 列为数值列）， y 是二元标签。对于预处理，我们对数据进行归一化。

下面我们使用 IJCNN 数据集重复相同的实验。

E.2 使用第4节中的损失进行 IJCNN 实验

我们指出，对于 $\varepsilon = 0.2$ 下的 2OPT-LS，由于存在数值问题，我们使用的 autodp 软件包无法处理隐私预算非常低时的子采样，因此无法获得结果（标记为 NA）。

[5]开源代码库：https://github.com/yuxiangw/autodp[6]数据来源：UCI机器学习库 https://archive.ics.uci.edu/ml/datasets/covertype[7]数据来源：LIBSVM数据仓库 https://www.openml.org/search?type=data&sort=runs&id=1575&status=active

表3：IJCNN：寻找一个宽松解，$(\epsilon_g, \epsilon_H) = (0.040, 0.200)$

方法	$\varepsilon = 0.2$		$\varepsilon = 0.6$		$\varepsilon = 1.0$	
	最终损失	运行时间	loss	运行时间	loss	运行时间
TR	0.621 ± 0.009	$\times$	0.71 ± 0.025	1.3 ± 1.5	0.718 ± 0.019	0.8 ± 1.1
TR-B	0.622 ± 0.009	$\times$	0.718 ± 0.046	0.4 ± 0.4	0.72 ± 0.043	1.3 ± 1.6
OPT	0.603 ± 0.012	$\times$	0.644 ± 0.014	$\times$	0.702 ± 0.015	0.1 ± 0.0
OPT-B	0.71 ± 0.022	2.7 ± 2.4	0.71 ± 0.022	2.6 ± 2.5	0.71 ± 0.022	2.8 ± 2.3
OPT-LS	0.671 ± 0.03	$\times$	0.593 ± 0.02	$\times$	0.658 ± 0.019	$\times$
2OPT	0.631 ± 0.052	$\times$	0.679 ± 0.048	$\times$	0.676 ± 0.056	$\times$
2OPT-B	0.71 ± 0.022	1.0 ± 0.3	0.71 ± 0.022	1.0 ± 0.3	0.71 ± 0.022	1.1 ± 0.3
2OPT-LS	NA	NA	0.696 ± 0.013	0.045 ± 0.005	0.693 ± 0.013	0.047 ± 0.005

表 4：IJCNN：寻找紧致解：$(\epsilon_g, \epsilon_H) = (0.020, 0.141)$

方法	$\varepsilon = 0.2$		$\varepsilon = 0.6$		$\varepsilon = 1.0$	
	最终损失	运行时间	loss	运行时间	loss	运行时间
TR	0.625 ± 0.008	$\times$	0.59 ± 0.009	$\times$	0.57 ± 0.009	$\times$
TR-B	0.625 ± 0.009	$\times$	0.591 ± 0.009	$\times$	0.574 ± 0.009	$\times$
OPT	0.643 ± 0.011	$\times$	0.583 ± 0.019	$\times$	0.515 ± 0.039	$\times$
OPT-B	0.667 ± 0.013	1.1 ± 0.3	0.667 ± 0.013	0.9 ± 0.1	0.667 ± 0.013	1.0 ± 0.1
OPT-LS	0.835 ± 0.069	$\times$	0.635 ± 0.022	$\times$	0.598 ± 0.024	$\times$
2OPT	0.537 ± 0.027	$\times$	0.522 ± 0.074	$\times$	0.552 ± 0.091	$\times$
2OPT-B	0.666 ± 0.014	1.1 ± 0.4	0.666 ± 0.014	1.3 ± 0.6	0.666 ± 0.014	1.1 ± 0.3
2OPT-LS	NA	NA	0.623 ± 0.06	$\times$	0.649 ± 0.004	0.1 ± 0.1

E.3 额外实验

此外，我们还考虑逻辑损失

$$\nabla l(w) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + \exp\left(-y_i \left\langle x_i, w \right\rangle\right)} + \frac{\lambda}{2} \|w\|^2,$$

并使用 $\lambda = 10^{-3}$ 在上述数据集上重复实验。我们可以验证，只要特征向量 x_i 有界，所选的两种损失就具有 Lipschitz 梯度和海森矩阵。

在这组实验中，我们得到解 $(\epsilon_g, \epsilon_H) = (0.040, 0.200)$ 和 $(\epsilon_g, \epsilon_H) = (0.020, 0.141)$。我们还展示了带噪海森矩阵评估次数的汇总结果。值得注意的是，我们的算法所需的带噪海森矩阵评估次数非常少，而 DP-TR 每次迭代都需要计算带噪海森矩阵。

E.3.1 使用逻辑损失的 Covertype 实验

表 5：Covertype（逻辑损失）：寻找宽松解，$(\epsilon_g, \epsilon_H) = (0.040, 0.200)$

方法	$\varepsilon = 0.2$		$\varepsilon = 0.6$		$\varepsilon = 1.0$	
	最终损失	运行时间	loss	运行时间	loss	运行时间
TR	0.425 ± 0.009	$\times$	0.388 ± 0.001	$\times$	0.381 ± 0.001	$\times$
TR-B	0.425 ± 0.009	$\times$	0.388 ± 0.002	$\times$	0.382 ± 0.001	$\times$
OPT	0.442 ± 0.006	$\times$	0.539 ± 0.022	0.3 ± 0.1	0.539 ± 0.022	0.4 ± 0.1
OPT-B	0.539 ± 0.022	11.1 ± 1.7	0.539 ± 0.022	10.6 ± 0.4	0.539 ± 0.022	10.6 ± 0.4
OPT-LS	0.385 ± 0.002	$\times$	0.455 ± 0.014	$\times$	0.539 ± 0.022	0.4 ± 0.1
2OPT	0.539 ± 0.022	0.3 ± 0.1	0.539 ± 0.022	0.3 ± 0.1	0.539 ± 0.022	0.4 ± 0.1
2OPT-B	0.539 ± 0.022	1.0 ± 0.0	0.539 ± 0.022	1.0 ± 0.1	0.539 ± 0.022	1.0 ± 0.1
2OPT-LS	0.539 ± 0.022	0.4 ± 0.1	0.539 ± 0.022	0.3 ± 0.1	0.539 ± 0.022	0.4 ± 0.1

表 6：Covertype 海森矩阵评估次数（逻辑损失）：寻找一个宽松解，$(\epsilon_g, \epsilon_H) = (0.040, 0.200)$

方法	$\varepsilon = 0.2$		$\varepsilon = 0.6$		$\varepsilon = 1.0$	
	海森矩阵评估次数	运行时间	海森矩阵评估次数	运行时间	海森矩阵评估次数	运行时间
TR	375.0 ± 0.0	$\times$	375.0 ± 0.0	$\times$	375.0 ± 0.0	$\times$
TR-B	375.0 ± 0.0	$\times$	375.0 ± 0.0	$\times$	375.0 ± 0.0	$\times$
OPT	190.6 ± 32.3	$\times$	1.0 ± 0.0	0.3 ± 0.1	1.0 ± 0.0	0.4 ± 0.1
OPT-B	1.0 ± 0.0	11.1 ± 1.7	1.0 ± 0.0	10.6 ± 0.4	1.0 ± 0.0	10.6 ± 0.4
OPT-LS	0.0 ± 0.0	$\times$	282.4 ± 174.252	$\times$	1.0 ± 0.0	0.4 ± 0.1
2OPT	1.0 ± 0.0	0.3 ± 0.1	1.0 ± 0.0	0.3 ± 0.1	1.0 ± 0.0	0.4 ± 0.1
2OPT-B	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.1	1.0 ± 0.0	1.0 ± 0.1
2OPT-LS	1.0 ± 0.0	0.4 ± 0.1	1.0 ± 0.0	0.3 ± 0.1	1.0 ± 0.0	0.4 ± 0.1

表 7：Covertype（逻辑损失）：寻找紧致解：$(\epsilon_g, \epsilon_H) = (0.020, 0.141)$

方法	$\varepsilon = 0.2$		$\varepsilon = 0.6$		$\varepsilon = 1.0$	
	最终损失	运行时间	loss	运行时间	loss	运行时间
TR	0.408 ± 0.004	$\times$	0.381 ± 0.0	$\times$	0.378 ± 0.0	$\times$
TR-B	0.408 ± 0.004	$\times$	0.381 ± 0.0	$\times$	0.378 ± 0.0	$\times$
OPT	0.379 ± 0.001	$\times$	0.38 ± 0.001	$\times$	0.39 ± 0.002	$\times$
OPT-B	0.454 ± 0.004	1.3 ± 0.2	0.454 ± 0.004	1.4 ± 0.4	0.454 ± 0.004	1.5 ± 0.6
OPT-LS	0.41 ± 0.005	$\times$	0.381 ± 0.001	$\times$	0.378 ± 0.0	$\times$
2OPT	0.386 ± 0.006	$\times$	0.378 ± 0.0	$\times$	0.377 ± 0.0	$\times$
2OPT-B	0.454 ± 0.004	2.0 ± 0.2	0.454 ± 0.004	2.0 ± 0.3	0.454 ± 0.004	1.8 ± 0.2
2OPT-LS	0.441 ± 0.007	$\times$	0.447 ± 0.009	0.6 ± 0.2	0.447 ± 0.008	0.7 ± 0.2

E.3.2 使用逻辑损失的 IJCNN 实验

表 8: Covertype 海森矩阵评估次数（逻辑损失）：寻找紧致解：$(\epsilon_g, \epsilon_H) = (0.020, 0.141)$

方法	$\varepsilon = 0.2$		$\varepsilon = 0.6$		$\varepsilon = 1.0$	
	海森矩阵评估次数	运行时间	海森矩阵评估次数	运行时间	海森矩阵评估次数	运行时间
TR	1061.0 ± 0.0	$\times$	1061.0 ± 0.0	$\times$	1061.0 ± 0.0	$\times$
TR-B	1061.0 ± 0.0	$\times$	1061.0 ± 0.0	$\times$	1061.0 ± 0.0	$\times$
OPT	0.0 ± 0.0	$\times$	58.6 ± 24.765	$\times$	372.6 ± 141.077	$\times$
OPT-B	1.0 ± 0.0	1.3 ± 0.2	1.0 ± 0.0	1.4 ± 0.4	1.0 ± 0.0	1.5 ± 0.6
OPT-LS	0.0 ± 0.0	$\times$	0.0 ± 0.0	$\times$	1.0 ± 1.0	$\times$
2OPT	0.0 ± 0.0	$\times$	0.0 ± 0.0	$\times$	0.2 ± 0.447	$\times$
2OPT-B	1.0 ± 0.0	2.0 ± 0.2	1.0 ± 0.0	2.0 ± 0.3	1.0 ± 0.0	1.8 ± 0.2
2OPT-LS	33.0 ± 6.745	$\times$	1.0 ± 0.0	0.6 ± 0.2	1.0 ± 0.0	0.7 ± 0.2

表 9：IJCNN（逻辑损失）：寻找宽松解，$(\epsilon_g, \epsilon_H) = (0.040, 0.200)$

方法	$\varepsilon = 0.2$		$\varepsilon = 0.6$		$\varepsilon = 1.0$	
	最终损失	运行时间	loss	运行时间	loss	运行时间
TR	0.477 ± 0.01	$\times$	0.454 ± 0.006	$\times$	0.446 ± 0.004	$\times$
TR-B	0.477 ± 0.01	$\times$	0.454 ± 0.006	$\times$	0.446 ± 0.004	$\times$
OPT	0.46 ± 0.01	$\times$	0.439 ± 0.002	$\times$	0.463 ± 0.005	$\times$
OPT-B	0.501 ± 0.008	8.6 ± 0.2	0.501 ± 0.008	11.0 ± 4.9	0.501 ± 0.008	9.0 ± 1.0
OPT-LS	0.607 ± 0.076	$\times$	0.473 ± 0.012	$\times$	0.454 ± 0.004	$\times$
2OPT	0.493 ± 0.022	$\times$	0.501 ± 0.008	$0.0^8 \pm 0.0$	0.501 ± 0.008	0.0 ± 0.0
2OPT-B	0.501 ± 0.008	1.0 ± 0.0	0.501 ± 0.008	1.4 ± 0.9	0.501 ± 0.008	1.1 ± 0.3
2OPT-LS	1.002 ± 0.321	$\times$	0.501 ± 0.008	0.0 ± 0.0	0.501 ± 0.008	0.0 ± 0.0

Table 10：IJCNN 海森矩阵评估次数（逻辑损失）：寻找一个宽松解，$(\epsilon_g, \epsilon_H) = (0.040, 0.20 \quad 0)$

方法	$\varepsilon = 0.2$		$\varepsilon = 0.6$		$\varepsilon = 1.0$	
	海森矩阵评估次数	运行时间	海森矩阵评估次数	运行时间	海森矩阵评估次数	运行时间
TR	375.0 ± 0.0	$\times$	375.0 ± 0.0	$\times$	375.0 ± 0.0	$\times$
TR-B	375.0 ± 0.0	$\times$	375.0 ± 0.0	$\times$	375.0 ± 0.0	$\times$
OPT	0.0 ± 0.0	$\times$	0.5 ± 0.577	$\times$	85.25 ± 47.647	$\times$
OPT-B	1.0 ± 0.0	8.6 ± 0.2	1.0 ± 0.0	11.0 ± 4.9	1.0 ± 0.0	9.0 ± 1.0
OPT-LS	0.0 ± 0.0	$\times$	0.0 ± 0.0	$\times$	0.25 ± 0.5	$\times$
2OPT	6.5 ± 1.732	$\times$	1.0 ± 0.0	0.0 ± 0.0	1.0 ± 0.0	0.0 ± 0.0
2OPT-B	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.4 ± 0.9	1.0 ± 0.0	1.1 ± 0.3
2OPT-LS	0.25 ± 0.5	$\times$	1.0 ± 0.0	0.0 ± 0.0	1.0 ± 0.0	0.0 ± 0.0

表11：IJCNN（逻辑损失）：寻找紧致解：$(\epsilon_g, \epsilon_H) = (0.020, 0.141)$

方法	$\varepsilon = 0.2$		$\varepsilon = 0.6$		$\varepsilon = 1.0$	
	最终损失	运行时间	loss	运行时间	loss	运行时间
TR	0.478 ± 0.01	$\times$	0.454 ± 0.005	$\times$	0.446 ± 0.003	$\times$
TR-B	0.479 ± 0.01	$\times$	0.454 ± 0.005	$\times$	0.446 ± 0.003	$\times$
OPT	0.484 ± 0.012	$\times$	0.447 ± 0.007	$\times$	0.44 ± 0.003	$\times$
OPT-B	0.501 ± 0.007	7.9 ± 6.3	0.501 ± 0.007	7.8 ± 6.2	0.501 ± 0.007	9.2 ± 7.4
OPT-LS	0.761 ± 0.066	$\times$	0.553 ± 0.037	$\times$	0.492 ± 0.016	$\times$
2OPT	0.527 ± 0.024	$\times$	0.462 ± 0.013	$\times$	0.501 ± 0.007	0.0 ± 0.0
2OPT-B	0.502 ± 0.007	1.5 ± 0.4	0.502 ± 0.007	1.3 ± 0.1	0.502 ± 0.007	1.6 ± 0.3
2OPT-LS	3.504 ± 0.98	$\times$	0.798 ± 0.081	$\times$	0.459 ± 0.01	0.4 ± 0.4

Tab表 12：IJCNN 逻辑损失的海森矩阵评估次数：寻找紧致解：$(\epsilon_g, \epsilon_H) = (0.020, 0.14 \quad 1)$

方法	$\varepsilon = 0.2$		$\varepsilon = 0.6$		$\varepsilon = 1.0$	
	海森矩阵评估次数	运行时间	海森矩阵评估次数	运行时间	海森矩阵评估次数	运行时间
TR	1061.0 ± 0.0	$\times$	1061.0 ± 0.0	$\times$	1061.0 ± 0.0	$\times$
TR-B	1061.0 ± 0.0	$\times$	1061.0 ± 0.0	$\times$	1061.0 ± 0.0	$\times$
OPT	0.0 ± 0.0	$\times$	0.0 ± 0.0	$\times$	0.0 ± 0.0	$\times$
OPT-B	1.0 ± 0.0	7.9 ± 6.3	1.0 ± 0.0	7.8 ± 6.2	1.0 ± 0.0	9.2 ± 7.4
OPT-LS	0.0 ± 0.0	$\times$	0.0 ± 0.0	$\times$	0.0 ± 0.0	$\times$
2OPT	0.0 ± 0.0	$\times$	3.0 ± 1.581	$\times$	1.0 ± 0.0	0.0 ± 0.0
2OPT-B	1.0 ± 0.0	1.5 ± 0.4	1.0 ± 0.0	1.3 ± 0.1	1.0 ± 0.0	1.6 ± 0.3
2OPT-LS	0.0 ± 0.0	$\times$	0.4 ± 0.548	$\times$	4.8 ± 4.97	0.4 ± 0.4