Privacy for Free: Communication-Efficient Learning with Differential Privacy Using Sketches

0. Abstract

1. 2 concerns in distributed learning:

- Communication reduction
- Privacy preservation

2. 证明了：

Count Sketch 具有 differential privacy 属性

3. 提出了；

DiffSketch，通过Sketch技术，压缩传输信息，同时实现了 Communication efficiency and privacy.

4. 实验了：

- DiffSketch可以提供强DP保证，且能够在不损失 accuracy 的情况下，减小20-50倍的通信开销
- 与分开处理privacy 和 communication 的情况相比，DiffSketch在提供相同水平 privacy 和 communication情况下，accuracy 提升了5%-50%

1. Introduction

Prior efforts: 分开考虑 privacy and communication.

1. 减小传输messages 的大小

- sparsification
- quantization
- subsampling

2. privacy-preserving

- add noise DP
- cryptographic eg. SMC

但是这些方法是要么能减小communication，要么能保护隐私

直接组合这些方法会降低accuracy

Sketching algorithms(sketches) 联合考虑了 communication and privacy.

Sketches 使用独立的 hash functions在有界errors 的情况下压缩输入的数据，同时实现了space and accuracy 的 trade-off.

通过在独立的hash functions 随机化数据的情况下，sketches 可以提供privacy---> DP

Sketches 可以被用来实现 Local differential privacy

DiffSketch，由 Count Sketch 得出的 framework.

实验验证了 DiffSketch 在 distributed SGD and FedAvg，在固定 privacy and communication 水平的情况下，DiffSketch可以在 test accuracy 上提升5%-50%。

Contributions

1. 证明了Count Sketch 具有 differential privacy.
2. 设计了 DiffSketch，基于 Count Sketch，能够保证 communication-efficient and differentially-private.
3. 实验证明了 DiffSketch 提供了提供了更强的 privacy and communication reduction.

2. Related Work and Background

Communication-efficient learning

1. Compression methods.
2. Reduce the total number of communication rounds.

Privacy in distributed learning

1. statistical

通过添加 random perturbations(Gaussian or Laplacian distribution) 来提供 differential privacy.

根据 server 是否可信，分为 local and global.

2. cryptography

secure multiparty computation (SMC), 多个参与方在不知道任何一方的 input information 情况下，协同计算一个 function

缺点：显著的 communication and computation overheads.

Connections between communication and privacy

现有的技术只能保护有限的信息，例如：原始数据的协方差

DiffSketch 实现了在DP 和 convergence 的 accuracy.

cpSGD 实现了 private and communication-efficient (但是是分开处理的)

Sketches

Sketching 算法提供了对于不同的统计数据的大致估计，例如计算数据集的不同的 count or average.

sketching 被用于 large-scale machine learning，通过 compressing model updates 或者减小 training 过程中的 memory.

recent efforts

- 使用 random noise、random sampling or other randomization 来提供 differential privacy.
- 但是这是在 sketches 技术上，添加了其他的机制

Our analysis

sketches 有固有的 differential privacy

3. Differential Privacy of Sketches

3.1 Preliminaries

Count Sketch 用于压缩 real-value vectors.

使用 t 个独立的 hash functions, 将 n 长向量中共的每个 element map 到 t 个 distinct bins，这可能导致 collisions, sketch 的目标也就是在 bounded errors 上近似 the true value.

Sketch 方法具有线性性，可用来 compress messages and merge sketched messages：

$$S(g_1 + g_2) = S(g_1) + S(g_2)$$

Count Sketch 主要包括两个 operations:

1. Encoding g in to a sketch table $S_{t\times k}(g)$

- 使用 hash functions 将 n 长 vector map 到 k 个位置

  $\{h_j, 1 \le j \le t : [n] -> [k]\}$

- 同时使用 2-wise independent sign hash functions

  $\{sign_j, 1 \leq j \leq t : [n] -> \{+1, -1\}\}$

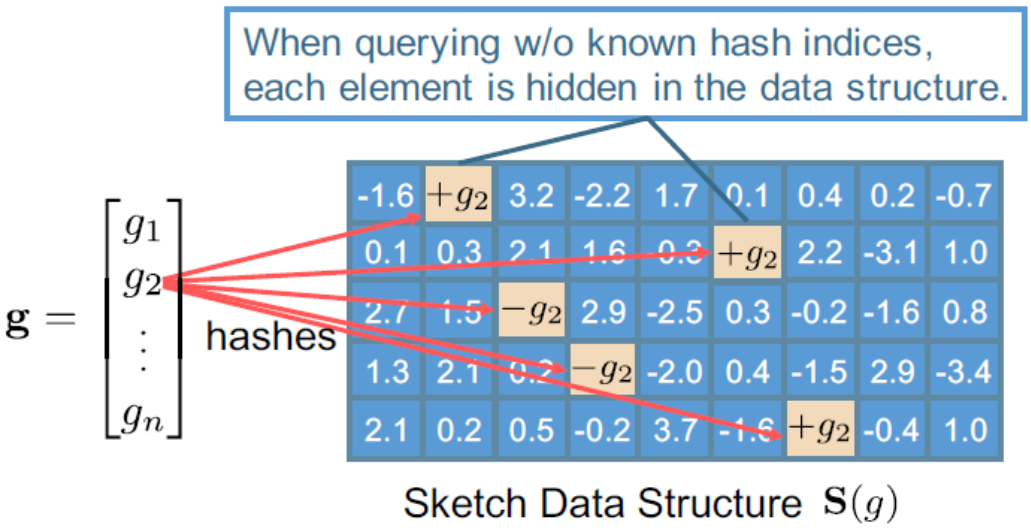
  将每个g_i 隐射到table 的 t 个不同的 bins

2. Querying $S_{t\times k}(g)$

   得到的是 g 的一个 estimation $\tilde{g}$

   为了查询 g_i，通过 $h_j(i)$的indexes，查询 t个近似值的中位数

   从 independent hash functions 的 randomization 提供了 differential privacy.



Definition 1(ϵ-differential privacy)

> input data D_1, D_2
>
> output S
>
> randomized mechanism M

$$Pr[\mathcal{M}(D_1) \in S] \leq e^{\varepsilon} \cdot Pr[\mathcal{M}(D_2) \in S].$$

ϵ越小，两个输出的分布就越接近

local privacy model

global privacy model

3.2 Differential Privacy Analysis

量化输入数据分布假设：

（1）每个 element 都是 bounded.

（2）inputs 是从 Gaussian distribution 中提取

Assumption 1 (input vector distribution)

n 长向量，input vector D，每个 element $d_i \sim N(0, \sigma^2)$，并且被常数 α bounded.

Lemma 1 (Estimation error of Count Sketch)

对于 Count Sketch，t arrays of k bins, input element $d_i \in D$, query Q，概率 p ≥ $1 - \delta$

$$|\mathcal{Q}(\mathcal{M}(d_i)) - d_i| \leq \mu \|D\|_2,$$

where $k = O\left(\frac{e}{\mu^2}\right)$, *and* $t = O\left(\ln(\frac{1}{\delta})\right)$.

估计值与原始值非常接近

Theorem 2 (ϵ-differential privacy of Count Sketch)

Theorem 2 (ε-differential privacy of Count Sketch). *For a sketching algorithm $\mathcal{M}$ using Count Sketch $S_{t\times k}$ with t arrays of k bins, for any input vector D with length n satisfying Assumption 1, $\mathcal{M}$ achieves $t \cdot \ln\left(1 + \frac{\beta\alpha^2 k(k-1)}{\sigma^2(n-2)}\left(1 + \ln(n-k)\right)\right)$-differential privacy with high probability, where β is a positive constant satisfying $\frac{\alpha^2 k(k-1)}{\sigma^2(n-2)}\left(1 + \ln(n-k)\right) \leq \frac{1}{2} - \frac{1}{\beta}$.*

accuracy 、communication、privacy 之间的 trade-off.

例如：使用更小的 sketching table 去压缩 input vector，压缩更激进，M 机制会有更小的隐私预算 ϵ，损耗了 accuracy.

4. Proposed Framework: DiffSketch

4.1 DiffSketch: A Framework for Distributed Learning

DiffSketch:

1.**Aggregation and Query steps**

computing gradients parallel

aggregate local information

2.**Compression and Validation**

步骤：

（1）Compression

entity k 使用 sketches , compress the local information v, to $S(v_k)$

（2）Validation

local worker verify 是否 sketching 后满足 ϵ-differential privacy。不满足的时候，需要添加额外noise

（3）Aggregation

central server 聚合 local information,并将聚合后的内容发送回去

（4）Query

每个 local entity 可以从聚合的 sketch 中大致恢复出 updated global information.

DiffSketch 允许每个 worker在本地 sketch the vectors, 因此可以达到与LDP 类似的 privacy level.

Algorithm 2 Proposed framework: `DiffSketch`.

1: **Input:** T, $v_1, \cdots, v_m$, ε
2: **for** $t = 0, \cdots, T-1$ **do**
3: **Compression:** Each entity k compresses v_k to obtain $\boldsymbol{S}(v_k)$ based on Algorithm 1
4: **Validation:** Each entity validates if it satisfies ε-differential privacy for a given ε. If not, add appropriate Laplacian or Gaussian noise to $\boldsymbol{S}(v_k)$
5: **Aggregation:** The server aggregates local information to obtain $\boldsymbol{S}(v) = \frac{1}{m}(\boldsymbol{S}_1 + \cdots + \boldsymbol{S}_m)$
6: **Query:** Each entity queries $\boldsymbol{S}(v)$ for the mean $\tilde{v}$ based on Algorithm 1
7: **end for**

Advantages:

(1) 可证明的 accuracy and differential privacy

(2) hashing-based computations 在 Compression and Query 是 轻量级的操作

4.2 DiffSketch for Distributed SGD

mini-batch SGD

$$\min_{w} F(w) = \sum_{k=1}^{m} p_k F_k(w) ,$$

> F_k 是worker k 的 local loss

> m 是workers 的数量
>
> p_k 是 worker k 的权重 weight

server 无法得知 raw gradients, 因为已经被压缩并且 mask 为小的 sketching tables.

local worker 可以以较高的 accuracy 恢复出 the merged gradients.

Algorithm 3 `DiffSketch` with distributed SGD.

Input: T, η, w^0, ε
for $t = 0, \cdots, T-1$ **do**
  if $t > 0$ **then**
    Server sends the sketched global gradient $\boldsymbol{S}(g^t)$ to all workers
    Each worker queries $\boldsymbol{S}(g^t)$ for $\tilde{g}^t$
    Each worker updates: $w^t = w^{t-1} - \eta \tilde{g}^t$
  end if
  Each worker k runs (mini-batch) SGD on w^t to obtain local gradients g_k^{t+1}
  Each worker sketches the gradients locally to obtain $\boldsymbol{S}(g_k^{t+1})$
  Each worker adds additional Laplacian noise to $\boldsymbol{S}(g_k^{t+1})$ if not satisfying ε-differential privacy
  Each worker sends $\boldsymbol{S}(g_k^{t+1})$ to the server
  Server aggregates the model updates: $\boldsymbol{S}(g^{t+1}) = \frac{1}{m}\sum_{k=1}^{m} \boldsymbol{S}(g_k^{t+1})$
end for

gradient.

实验证明：当 compression ratio 达到 50× 时候， accuracy reduction是非常微弱的

Convergence

依赖于

（1）the bounded estimation error of Count Sketch

（2）unbiasedness of Count Sketch

满足：$E[Q(M(d_i))] = d_i$

以上Lemma确保了 unbiased and uniformly bounded.

DiffSketch 与 distributed SGD 有相同的收敛率

hash functions 数量 t 越大，以及 bins 数量 k 越大，我们压缩的越少，得到一个更紧的 convergence bound，更小的 recovery error.

4.3 DiffSketch for Federated Learning

(1) randomly samples a subset of devices.

(2) perform E epochs of local updates.

(3) apply Count Sketch to compress the updates.

(4) average the updates centrally.

Algorithm 4 `DiffSketch` in federated learning.

Input: K, T, η, E, w^0, p_k, ε
for $t = 0, \cdots, T-1$ **do**
  Server samples a subset S_t of K devices (each device is chosen with probability p_k)
  if $t > 0$ **then**
    Server sends the sketched global model $\boldsymbol{S}(\Delta w^t)$ to all chosen devices
    Each device k queries $\boldsymbol{S}(\Delta w^t)$ for $\Delta \tilde{w}^t$
    Each device k updates: $w^t = w^{t-1} + \Delta w^t$
  end if
  Each device k updates w^t for E epochs of SGD on F_k with step-size η to obtain Δw_k^{t+1}
  Each device k sketches the updates locally to obtain $\boldsymbol{S}(\Delta w_k^{t+1})$
  Each device k adds additional Laplacian noise to $\boldsymbol{S}(\Delta w_k^{t+1})$ if not satisfying ε-differential privacy
  Each device k sends $\boldsymbol{S}(\Delta w_k^{t+1})$ to the server
  Server aggregates the model updates: $\boldsymbol{S}(\Delta w^{t+1}) = \frac{1}{K} \sum_{k \in S_t} \boldsymbol{S}(\Delta w_k^{t+1})$
end for

updates

实验证明：DiffSketch 可以压缩通信20倍，提供了强的 local privacy guarantees.

FedAvg是一个启发式的方法，因此即使 robust 性能，但是可能也不会收敛。

5. Evaluation

5.1 Simulation Setups

Datasets

(1)distributed SGD

MNIST dataset for image classification

(2)federated learning

Shakespeare dataset

Implementation

local worker 将 local gradients 与 query results 进行比较

增大 input vector 的 size，可以压缩更多的 values.

5.2 Comparison with Baselines

cpSGD 是最新的旨在优化 communication and privacy （但是是分布处理的）

1、quantizes the gradients.

2、add Binomial noise 提供 differential privacy.

如下图所示， DiffSketch 比 cpSGD test accuracy 提高了 50%

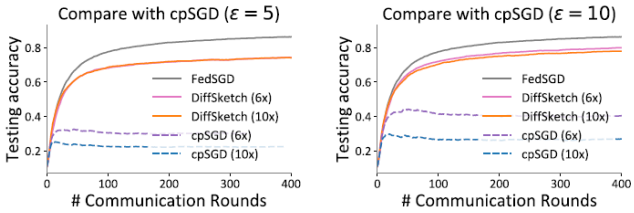


Figure 2: `DiffSketch` compared with `cpSGD`. Under the same privacy guarantees ($\varepsilon = 5$ or 10) and the same communication compression ratio, `DiffSketch` achieves significantly higher accuracy. If we continue to decrease ε to improve the privacy, `cpSGD` will perform even worse.

Other baselines

(1) 直接给 raw gradients 添加 Laplacian noise

(2) 把 sketches 看作 plain-text, 所有 sketches 结束后，添加 Laplacian noise.

与其他的 baselines 相比

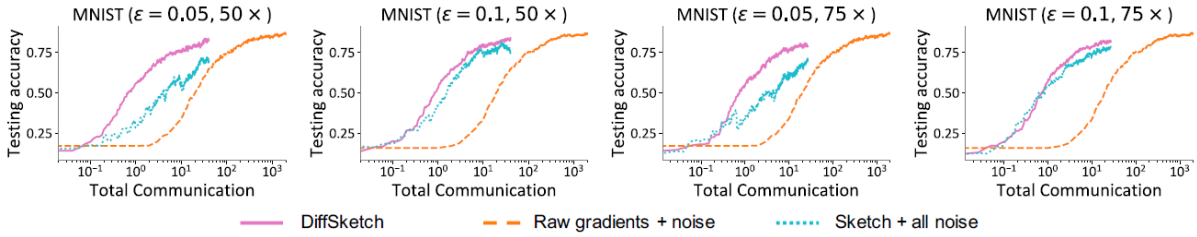


Figure 3: `DiffSketch` compared with other baselines in the distributed SGD setting. We show the test accuracy versus total communication in log scale. The total amount of communication is normalized

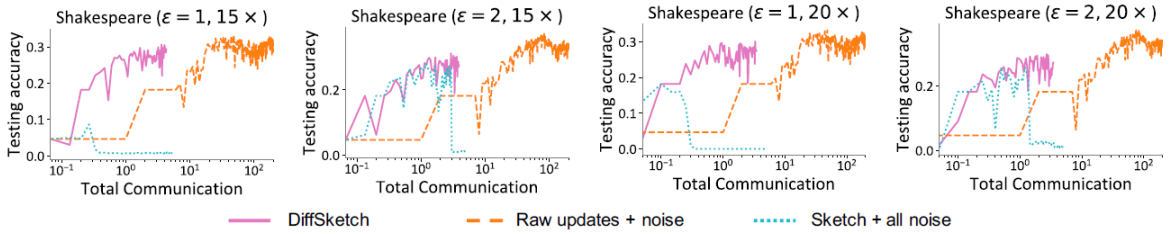


Figure 4: `DiffSketch` compared with other baselines in federated learning. Again, we note here that the total

DiffSketch 收敛更快，而且通信量也显著减小

DiffSketch 更加准确

为了保证更强的隐私性，baselines 需要在 sketches 的基础上添加更多的noise，导致模型性能下降。而DiffSketch 只需要添加小部分的noise。

5.3 Trade-offs in DiffSketch

Accuracy vs. compression

随着 compression ratio 的增大，收敛更快，但是也带来了更多的 accuracy 的 loss

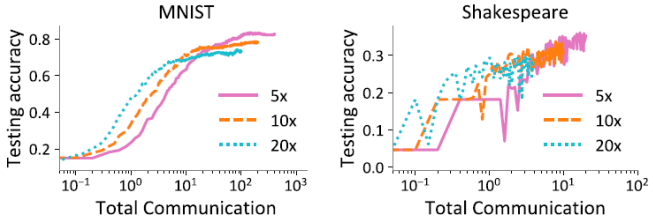


Figure 5: As the compression ratio becomes higher, `DiffSketch` converges faster, but with potentially lower accuracies.

Privacy vs. communication

compression 越大， more accuracy

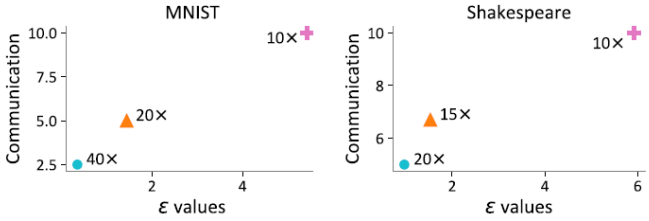


Figure 6: As we compress more, we can reduce communication and obtain stronger privacy guarantees with smaller ε values.

6. Conclusion

- 证明了 Count Sketch 有固有的DP 属性
- 设计 DiffSketch framework
- 将 DiffSketch 应用于经典的 distributed SGD and federated learning.
- 实验证明 DiffSketch 能提升 5% - 50% 的 准确率
- future work：是否有机制能够同样提供同时提供 privacy and communication.