Differentially Private Optimization with Sparse Gradients

Badih Ghazi[1], Cristóbal Guzmán[1,2], Pritish Kamath[1], Ravi Kumar[1], Pasin Manurangsi[1]

[1] Google Research, [2] Pontificia Universidad Católica de Chile

Motivation

- Gradient sparsity arises naturally in ML
- The fundamental question of how gradient (or data) sparsity influences excess risk rates in DP learning has been scarcely studied [ZMH:2021]

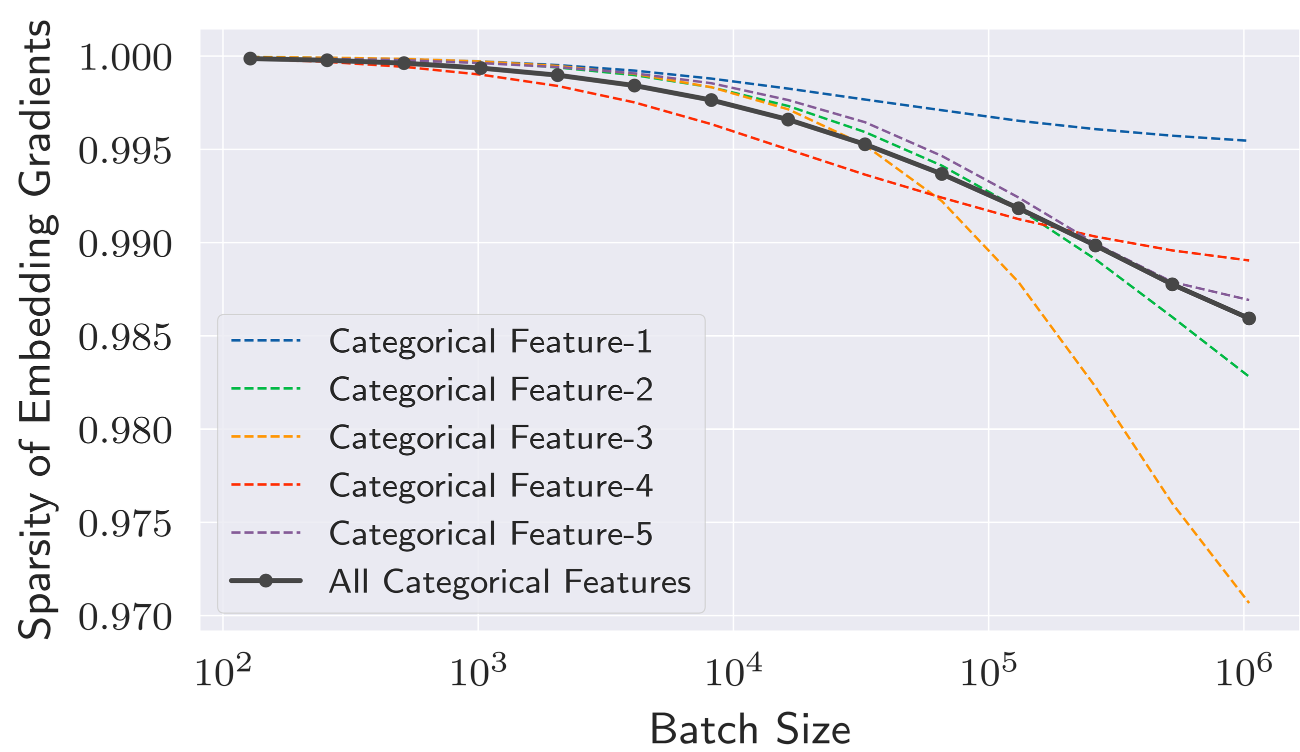

Figure 1: Gradient sparsity for embedding features in ads model [GHKKMSZ:2022]

Setting

1. Feasible set $\mathsf{X} \subseteq \mathbb{R}^d$ is closed, convex
2. Distribution D supported on $\mathcal{Z}$, and dataset $S \sim \mathsf{D}^n$
3. $f(\cdot; z)$ convex (if not, we aim at stationary points) and L-Lipschitz and/or H-smooth wrt $\|\cdot\|_2$
4. **Gradient Sparsity:** $\sup_{x,z} \|\nabla f(x,z)\|_0 \le s$

Stochastic Optimization (SO)

Objective: $\min_{x \in \mathsf{X}} F_{\mathsf{D}}(x) := \mathbb{E}_{z\sim\mathsf{D}}[f(x,z)]$

Algorithm $\mathcal{A} : \mathcal{Z}^n \mapsto \mathsf{X}$ is α-accurate for (SO) if

$$\mathbb{E}_{\mathcal{A},S}[F_{\mathsf{D}}(\mathcal{A}(S))] - \min_{x\in\mathsf{X}} F_{\mathsf{D}}(x) \le \alpha$$

(ϵ, δ)-Differential Privacy

For neighbouring datasets S, S'

$$\mathbb{P}(\mathcal{A}(S) \in E) \le e^{\varepsilon}\mathbb{P}(\mathcal{A}(S') \in E) + \delta$$

Our Results

Setting	Upper bound	Lower bound
ε-DP	$1 \wedge \sqrt{\frac{s \ln d}{\varepsilon n}} \wedge \frac{\sqrt{sd}}{\varepsilon n}$	$1 \wedge \sqrt{\frac{s \ln(d/(\varepsilon n))}{\varepsilon n}} \wedge \frac{\sqrt{sd}}{\varepsilon n}$
(ε,δ)-DP	$1 \wedge \frac{(s \ln(d/s) \ln(1/\delta))^{1/4}}{\sqrt{\varepsilon n}} \wedge \frac{\sqrt{d \ln(1/\delta)}}{\varepsilon n}$	$1 \wedge \frac{(s \ln(1/\delta))^{1/4}}{\sqrt{\varepsilon n}} \wedge \frac{\sqrt{d \ln(1/\delta)}}{\varepsilon n}$

Table 1: Upper/lower bounds for DP ℓ_2-mean estimation. New regimes in red.

Setting	Guarantee	New Upper bound (sparse)	Upper bound (non-sparse)
(ε,δ)-DP	Cvx. ERM	$\frac{(s \ln(d) \ln(1/\delta))^{1/4}}{\sqrt{\varepsilon n}} \wedge \mathcal{R}_{\varepsilon,\delta}$	$\mathcal{R}_{\varepsilon,\delta}$
	SCO	$\frac{(s \ln(d) \ln(1/\delta))^{1/4}}{\sqrt{\varepsilon n}} \wedge \mathcal{R}_{\varepsilon,\delta} + \frac{1}{\sqrt{n}}$	$\mathcal{R}_{\varepsilon,\delta} + \frac{1}{\sqrt{n}}$
ε-DP	Cvx. ERM	$\left(\frac{s \ln(d)}{\varepsilon n}\right)^{1/3} \wedge \mathcal{R}_{\varepsilon}$	$\mathcal{R}_{\varepsilon}$
	SCO	$\left(\frac{s \ln(d)}{\varepsilon n}\right)^{1/3} \wedge \mathcal{R}_{\varepsilon} + \frac{1}{\sqrt{n}}$	$\mathcal{R}_{\varepsilon} + \frac{1}{\sqrt{n}}$
(ε,δ)-DP	Emp. Grad. Norm	$\frac{(s \ln(d/s) \ln^3(1/\delta))^{1/8}}{(\varepsilon n)^{1/4}} \wedge (\mathcal{R}_{\varepsilon,\delta})^{2/3}$	$(\mathcal{R}_{\varepsilon,\delta})^{2/3}$

Table 2: Rates for DP optimization with sparse gradients. Polylog(n) factors omitted. Above $\mathcal{R}_{\varepsilon,\delta} = \sqrt{d \ln(1/\delta)}/[\varepsilon n]$ and $\mathcal{R}_{\varepsilon} = d/[\varepsilon n]$. New regimes in red.

Upper Bounds for Mean Estimation

Algorithm: Projection Mechanism [NTZ:2013]

Input: $\bar{z}(S) = \frac{1}{n}\sum_{i=1}^n z_i$

Output: $\widehat{z} = \arg\min\{\|z - [\bar{z}(S) + \xi]\|_2 : z \in \mathcal{B}_1^d(0, L\sqrt{s})\}$ where

$$\xi \sim \begin{cases} \mathsf{Lap}(\sigma)^{\otimes d} \text{ with } \sigma = \frac{2L\sqrt{s}}{n\varepsilon} & \text{if } \delta = 0 \\ \mathcal{N}(0, \sigma^2 I) \text{ with } \sigma^2 = \frac{8L^2 \ln(1.25/\delta)}{(n\varepsilon)^2} & \text{if } \delta > 0 \end{cases}$$

If $\delta > 0$, a tighter bound can be obtained by (noisy) compressed sensing.

Algorithm: Gaussian ℓ_1-Recovery

Input: $\bar{z}(S) = \frac{1}{n}\sum_{i=1}^n z_i$

$m = n\varepsilon\sqrt{\frac{s\ln(d/s)}{\ln(1/\delta)}}$, $\sigma^2 = \frac{18L^2\ln(2.5/\delta)}{(n\varepsilon)^2}$, $A \sim (\mathcal{N}(0, 1/m))^{\otimes m\times d}$, $b = A\bar{z}(S) + \xi$

Output: $\tilde{z} = \arg\min\{\|z\|_1 : Az = b\}$

Algorithms are nearly optimal, evidenced by lower bounds obtained by a novel block-diagonal construction whose blocks contain dense-case hard datasets.

Upper Bounds for DP-SO

- We introduce a novel output perturbation with ℓ_∞-projection that is provably nearly optimal for DP-ERM/SCO in high-dimension.

Algorithm: Output Perturbation with ℓ_∞-Projection

Input: Dataset $S = (z_1, \ldots, z_n)$, regularization param. $\lambda > 0$

$x_\lambda^*(S) = \arg\min\{\frac{1}{n}\sum_{i=1}^n f(x; z_i) + \frac{\lambda}{2}\|x\|_2^2\}$

$\tilde{x} = x_\lambda^*(S) + \xi$ where

$$\xi \sim \begin{cases} \mathsf{Lap}(\sigma)^{\otimes d} \text{ with } \sigma = \frac{2\sqrt{2s}L}{\lambda n\varepsilon}\left(\frac{2H}{\lambda} + 1\right) & \text{if } \delta = 0 \\ \mathcal{N}(0, \sigma^2 I) \text{ with } \sigma^2 = \frac{8L^2\ln(1.25/\delta)}{(\lambda n\varepsilon)^2} & \text{if } \delta > 0 \end{cases}$$

Output: $\widehat{x} = \arg\min \|x - \tilde{x}\|_\infty$

- We introduce a bias-reduced DP-SGD method for convex and nonconvex losses, following the debiasing approach in [BG:2015]
- SGD analysis depends on a randomly increasing privacy budget [WRRW:2023] $\Rightarrow$ algorithm runs for a (random) stopping time
- Bias-reduction introduces heavy-tailed stochastic oracles whose convergence can only be guaranteed with constant success probability. DP boosting approaches imply high-probability guarantees [LT:2019]

Subsampled Bias-Reduced Gradient Estimator

Input: $S = (z_1, \ldots, z_n)$, $x \in \mathcal{X}$

Sample $N \in [\log(n) - 1]$ with $\mathbb{P}[N = k] \propto 2^{-k} =: p_k$

$B \sim \mathrm{Unif}(\binom{n}{2^{N+1}})$; O, E equipartition of B; and $i \sim \mathrm{Unif}([n])$

Apply DP-Mean Estimation

$[\nabla F_B, \nabla F_O, \nabla F_E, \nabla f(\cdot, z_i)](x) \mapsto [G_B^+(x), G_O^-(x), G_E^-(x), G_i(x)]$

Output: $\mathcal{G}(x) = \frac{1}{p_N}\left[G_B^+(x) - \frac{1}{2}\left(G_O^-(x) + G_E^-(x)\right)\right] + G_i(x)$

References

ZMH:2021: Zhang, Mironov, Hejazinia, "Wide Network Learning with Differential Privacy" *arXiv:2103.01294*

GHKKMSZ:2022: Ghazi, Huang, Kamath, Kumar, Manurangsi, Sinha, Zhang. "Sparsity-preserving differentially private training of large embedding models", *NeurIPS 2022*

NTZ:2013: Nikolov, Talwar, Zhang. "The geometry of differential privacy: the sparse and approximate cases", *STOC 2013*

BG:2015: Blanchet, Glynn. "Unbiased Monte Carlo for opt. and functions of expectations via multi-level randomization", *WSC 2015*

WRRW:2023: Whitehouse, Ramdas, Rogers, Wu. "Fully-adaptive composition in differential privacy", *ICML 2023*

LT:2019: Liu, Talwar. "Private selection from private candidates", *STOC 2019*