Differentially Private Random Block Coordinate Descent

Artavazd Maranjyan, Abdurakhmon Sadiev, Peter Richtárik

NEURAL INFORMATION PROCESSING SYSTEMS

King Abdullah University of Science and Technology
KAUST
Saudi Arabia

Problem Formulation

$$w^\star \in \arg\min_{w\in\mathbb{R}^d} \left\{ f(w) := \frac{1}{n}\sum_{i=1}^{n} \ell(w;\zeta_i) \right\}.$$

$\ell(w;\zeta_i) : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$ is the loss function for a sample ζ_i, and $D = (\zeta_1, \dots, \zeta_n)$ is a dataset of n samples drawn from the universe $\mathcal{X}$.

Sketches

Given a random set $S \sim \mathcal{S}$, define

$$p_j := \mathrm{Prob}(j \in S), \quad j \in [d].$$

We also denote $\mathbf{P} = \mathrm{Diag}\,(p_1, \dots, p_d)$.

Definition 0 (Unbiased diagonal sketch). For a given random set $S \sim \mathcal{S}$ we define a random diagonal matrix (sketch) $\mathbf{C} = \mathbf{C}(S) \in \mathbb{R}^{d\times d}$ via

$$\mathbf{C} = \mathrm{Diag}\,(\mathrm{c}_1, \dots, \mathrm{c}_d)\,, \quad \mathrm{c}_j = \begin{cases} \frac{1}{p_j}, & \text{if } j \in S, \\ 0, & \text{otherwise.} \end{cases}$$

Equivalently, we can write

$$\mathbf{C} = \mathbf{I}_S \mathbf{P}^{-1},$$

where $\mathbf{I}_S = \mathrm{Diag}\,(\delta_1, \delta_2, \dots, \delta_d)$ is a diagonal matrix with

$$\delta_i = \begin{cases} 1, & \text{if } \quad i \in S, \\ 0, & \text{if } \quad i \notin S. \end{cases}$$

Assumptions

Assumption 1. Let $S \sim \mathcal{S}$ be *nonvacuous*, i.e., $P(S = \emptyset) = 0$, and *proper*, meaning that $p_j > 0$ for all $j \in [d]$.

Assumption 2 (Component smoothness). Function $f : \mathbb{R}^d \to \mathbb{R}$ is $\mathbf{M}$-component-smooth for $M_1, \dots, M_d > 0$. That is, for all $v, w \in \mathbb{R}^d$,

$$f(w) \le f(v) + \langle \nabla f(v), w - v\rangle + \frac{1}{2}\|w - v\|^2_{\mathbf{M}}.$$

Algorithm 1 DP-SkGD

Input: Initial point $w^0 \in \mathbb{R}^d$, step sizes $\mathbf{\Gamma} = \mathrm{Diag}\,(\gamma_1, \dots, \gamma_d)$, number of iterations T, number of inner loops K, probability distribution $\mathcal{S}$ over the subsets of $[d]$, noise scales σ_U for $U \in \mathrm{Range}\,(\mathcal{S})$
for $t = 0, \dots, T-1$ **do**
  Set $\theta^0 = w^t$
  for $k = 0, \dots, K-1$ **do**
    Sample a subset $S \sim \mathcal{S}$ and let $\mathbf{C} = \mathbf{C}(S)$
    Draw $\eta \sim \mathcal{N}\,(0, \sigma_S \mathbf{I})$
    $\theta^{k+1} = \theta^k - \mathbf{\Gamma}\mathbf{C}\left(\nabla f(\theta^k) + \eta\right)$
  end for
  $w^{t+1} = \frac{1}{K}\sum_{k=1}^{K}\theta^k$
end for

Assumption 3 (Component Lipschitzness). Let $\mathcal{S}$ be a probability distribution over the 2^d subsets of $[d]$. Function $\ell(\cdot;\zeta) : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$ is $L_{\mathcal{S}}$-component-Lipschitz with $L_U > 0$ for all $U \in \mathrm{Range}\,(\mathcal{S})$, for all $\zeta \in \mathcal{X}$. This means that for all $v, w \in \mathbb{R}^d$, we have:

$$|\ell\,(w + \mathbf{I}_U v;\zeta) - \ell\,(w;\zeta)| \le L_U \|\mathbf{I}_U v\|.$$

Block Coordinates

Consider a partition of $[d]$ into b nonempty blocks, denoted as $A_1, \dots, A_b$. Let $S = A_j$ with probability $q_j > 0$, where $\sum_j q_j = 1$. For each $i \in [n]$, let $B(i)$ indicate which block i belongs to. In other words, $i \in A_j$ iff $B(i) = j$. Then $p_i := \mathrm{Prob}(i \in S) = q_{B(i)}$. We call the resulting method DP-SkGD-BS. We define

$$L_{\{A_1,\dots,A_b\}} := \sum_{i=1}^{d} L_{B(i)} e_i.$$

$$R_{\mathbf{MP}^{-1}} = \left\|w^0 - w^\star\right\|^2_{\mathbf{MP}^{-1}}.$$

Importance Sampling

$$q_i = \frac{\max_{j\in A_i} M_j}{\sum_{i=1}^{b}\max_{j\in A_i} M_j}.$$

Table 1: Utility guarantees for DP-SkGD-BS with varying values of b and different sampling strategies, along with DP-CD, DP-SGD, and DP-SVRG.

	Convex	Strongly-convex
DP-SkGD-BS (this paper)	$\mathcal{O}_*\left(\left\|L_{\{A_1,\dots,A_b\}}\right\|_{\mathbf{M}^{-1}} R_{\mathbf{MP}^{-1}}\right)$	$\widetilde{\mathcal{O}}_-\left(\left\|L_{\{A_1,\dots,A_b\}}\right\|^2_{\mathbf{M}^{-1}} \frac{1}{\mu}\max_{i\in[d]}\left\{\frac{M_i}{p_i}\right\}\right)$
Uniform Sampling	$\mathcal{O}_*\left(\left\|L_{\{A_1,\dots,A_b\}}\right\|_{\mathbf{M}^{-1}} R_{\mathbf{M}}\sqrt{b}\right)$	$\widetilde{\mathcal{O}}_-\left(\left\|L_{\{A_1,\dots,A_b\}}\right\|^2_{\mathbf{M}^{-1}} \frac{1}{\mu} M_{\max} b\right)$
Importance Sampling	$\mathcal{O}_*\left(\left\|L_{\{A_1,\dots,A_b\}}\right\|_{\mathbf{M}^{-1}} R_{\mathbf{I}}\sqrt{\sum_{i=1}^{b}\max_{j\in A_i} M_j}\right)$	$\widetilde{\mathcal{O}}_-\left(\left\|L_{\{A_1,\dots,A_b\}}\right\|_{\mathbf{M}^{-1}} \frac{1}{\mu}\sum_{i=1}^{b}\max_{j\in A_i} M_j\right)$
DP-SkGD-BS (this paper) $b = d$	$\mathcal{O}_*\left(\left\|L_{\{1,\dots,d\}}\right\|_{\mathbf{M}^{-1}} R_{\mathbf{MP}^{-1}}\right)$	$\widetilde{\mathcal{O}}_-\left(\left\|L_{\{1,\dots,d\}}\right\|^2_{\mathbf{M}^{-1}} \frac{1}{\mu}\max_{i\in[d]}\left\{\frac{M_i}{p_i}\right\}\right)$
Uniform Sampling	$\mathcal{O}_*\left(\left\|L_{\{1,\dots,d\}}\right\|_{\mathbf{M}^{-1}} R_{\mathbf{M}}\sqrt{d}\right)$	$\widetilde{\mathcal{O}}_-\left(\left\|L_{\{1,\dots,d\}}\right\|^2_{\mathbf{M}^{-1}} \frac{1}{\mu} M_{\max} d\right)$
Importance Sampling	$\mathcal{O}_*\left(\left\|L_{\{1,\dots,d\}}\right\|_{\mathbf{M}^{-1}} R_{\mathbf{I}}\sqrt{\mathrm{Tr}\,(\mathbf{M})}\right)$	$\widetilde{\mathcal{O}}_-\left(\left\|L_{\{1,\dots,d\}}\right\|^2_{\mathbf{M}^{-1}} \frac{1}{\mu}\,\mathrm{Tr}\,(\mathbf{M})\right)$
DP-CD (Mangold et al., 2022)	$\mathcal{O}_*\left(\left\|L_{\{1,\dots,d\}}\right\|_{\mathbf{M}^{-1}} R_{\mathbf{M}}\sqrt{d}\right)$	$\widetilde{\mathcal{O}}_-\left(\left\|L_{\{1,\dots,d\}}\right\|^2_{M^{-1}} \frac{1}{\mu_{\mathbf{M}}} d\right)$
DP-SkGD-BS (this paper) $b = 1$	$\mathcal{O}_*\left(L\sqrt{\mathrm{Tr}\,(\mathbf{M}^{-1})} R_{\mathbf{M}}\right)$	$\widetilde{\mathcal{O}}_-\left(L^2\,\mathrm{Tr}\,(\mathbf{M}^{-1})\,\frac{1}{\mu} M_{\max}\right)$
DP-SGD (Bassily et al., 2014) DP-SVRG (Wang et al., 2017)	$\mathcal{O}_*\left(L\sqrt{d} R_{\mathbf{I}}\right)$	$\widetilde{\mathcal{O}}_-\left(L^2 \frac{1}{\mu_{\mathbf{I}}} d\right)$

We use the notation $\mathcal{O}_*$ to suppress the common term $\frac{\sqrt{\log(1/\delta)}}{n\epsilon}$, which appears consistently across all rates. Similarly, we denote $\mathcal{O}_-$ to suppress the term $\frac{\log(1/\delta)}{n^2\epsilon^2}$, as it is also consistent across all rates.

Our method gains an advantage over DP-CD due to the use of importance sampling.

To illustrate, consider the case where $b = d$ (i.e., single coordinate sampling). Assume that $M_1 \gg M_j$ for all $j \ne 1$, and similarly, $|w_1^0 - w_1^\star| \gg |w_j^0 - w_j^\star|$ for all $j \ne 1$. Moreover, suppose $M_1|w_1^0 - w_1^\star| \gg M_j|w_j^0 - w_j^\star|$. Then, in the convex case, we get

$$R_{\mathbf{I}}\sqrt{\mathrm{Tr}\,(\mathbf{M})} \approx \sqrt{M_1|w_1^0 - w_1^\star|} \approx R_{\mathbf{M}}.$$

Thus, DP-SkGD-BS with importance sampling can be up to $\sqrt{d}$ times faster.

References

Mangold, P., Bellet, A., Salmon, J., and Tommasi, M. (2022). Differentially private coordinate descent for composite empirical risk minimization. Proceedings of the 39th International Conference on Machine Learning, volume 162.

Bassily, R., Smith, A., and Thakurta, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. In 2014 IEEE 55th annual symposium on foundations of computer science, pages 464–473. IEEE.

Wang, D., Ye, M., and Xu, J. (2017). Differentially private empirical risk minimization revisited: Faster and more general. Advances in Neural Information Processing Systems, 30.