A Guide Through the Zoo of Biased SGD

Yury Demidovich Grigory Malinovsky Igor Sokolov Peter Richtárik

King Abdullah University of Science and Technology (KAUST)

The Problem

We study convergence properties and worst-case complexity bounds of stochastic gradient descent with a *biased* gradient estimator (BiasedSGD; see Algorithm 1) for solving general optimization problems of the form

$$\min_{x\in\mathbb{R}^d} f(x),$$

where the function $f:\mathbb{R}^d\to\mathbb{R}$ is possibly nonconvex, satisfies several smoothness and regularity conditions: f is differentiable, L-smooth (i.e., $\|\nabla f(x)-\nabla f(y)\|\le L\|x-y\|$ for all $x,y\in\mathbb{R}^d$), and bounded from below by $f^*\in\mathbb{R}$. Given an error tolerance $\varepsilon>0$, we seek a random vector $x\in\mathbb{R}^d$ such that one of the following inequalities holds:

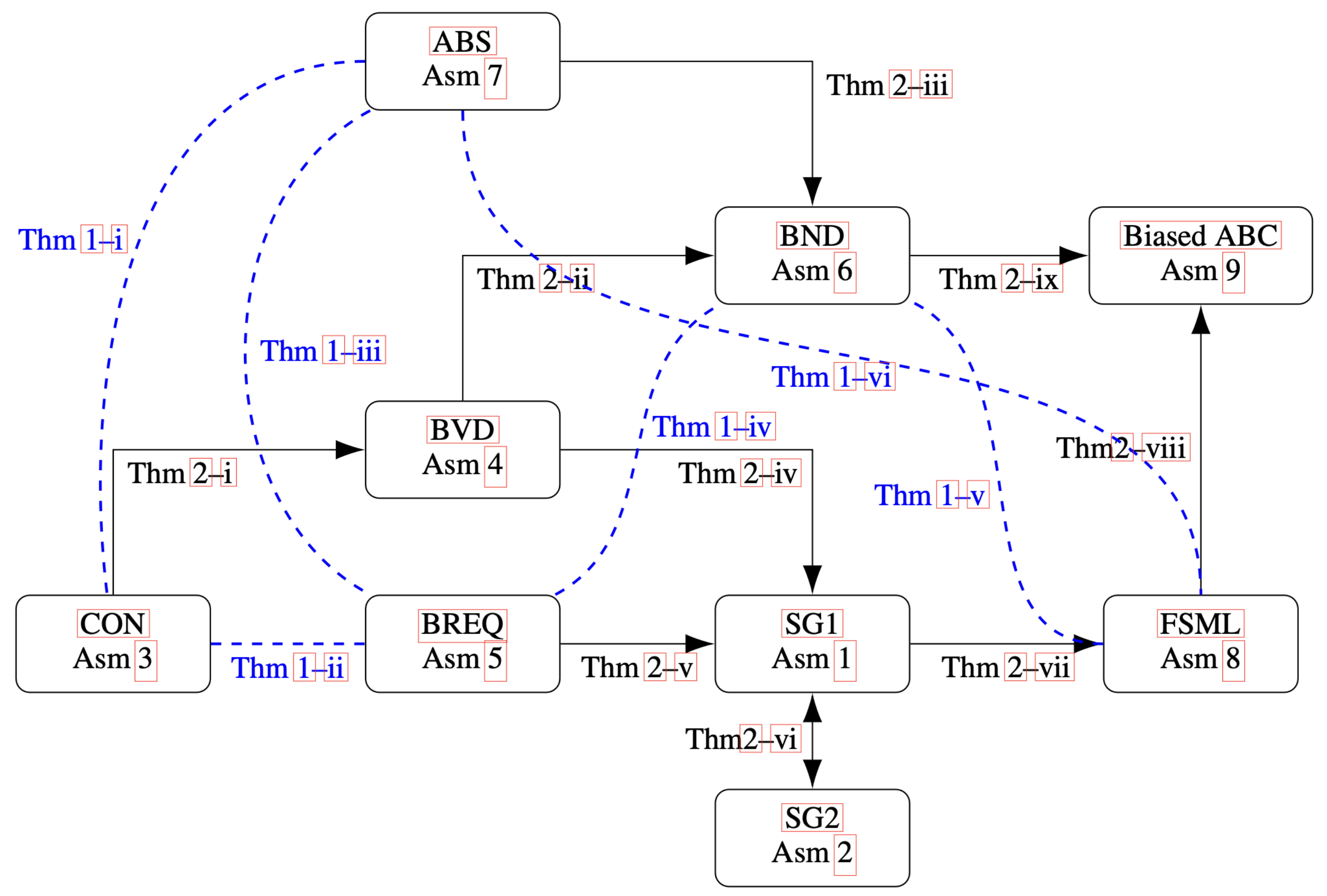
i) $\mathbb{E}\left[f(x)-f^*\right]\le\varepsilon$; **ii)** $\mathbb{E}\|x-x^*\|^2\le\varepsilon\left\|x^0-x^*\right\|^2$; **iii)** $\mathbb{E}\|\nabla f(x)\|^2\le\varepsilon^2$.

Algorithm 1: BiasedSGD

```
Parameters: Stepsize $\gamma > 0$, initial iterate $x^0 \in \mathbb{R}^d$
for $t = 0, 1, \dots$ do
    Construct a (possibly biased) estimator $g^t := g(x^t)$ of the gradient $\nabla f(x^t)$
    Compute $x^{t+1} = x^t - \gamma g^t$
end
```

A Zoo of Assumptions

There exists a Zoo of assumptions on the stochastic gradient estimators in works on BiasedSGD. In the diagram we present the assumptions from the literature and establish transitive connections between them.



ABS Asm 7

BND Asm 6

Biased ABC Asm 9

BVD Asm 4

CON Asm 3

BREQ Asm 5

SG1 Asm 1

FSML Asm 8

SG2 Asm 2

Thm 2–iii

Thm 1–i

Thm 2–ii

Thm 2–ix

Thm 1–iii

Thm 1–vi

Thm 1–iv

Thm2–viii

Thm 2–i

Thm 2–iv

Thm 1–v

Thm 1–ii

Thm 2–v

Thm 2–vii

Thm2–vi

Our work is motivated by the need of a more accurate and informative analysis of BiasedSGD in the strongly convex and nonconvex settings, which are problems of key importance in optimization and deep learning. Our results are generic and cover both subsampling and compressive estimators, among others. We generalize the existing conditions on the first moment and combine them with ABC-assumption to develop our Biased ABC framework.

New Assumption: Biased ABC

$\exists\, A,B,C,b,c\ge 0$ s.t. $\forall\, x\in\mathbb{R}^d$ the gradient estimator $g(x)$ satisfies

$$\langle\nabla f(x),\mathbb{E}[g(x)]\rangle\ge b\,\|\nabla f(x)\|^2-c, \tag{1}$$

$$\mathbb{E}\left[\|g(x)\|^2\right]\le 2A\left(f(x)-f^*\right)+B\,\|\nabla f(x)\|^2+C. \tag{2}$$

Khaled and Richtárik [4] proposed (2) in the unbiased case. $A\left(f(x)-f^*\right)$ in (2) emerges when $f(x)=\sum_{i=1}^n f_i(x)$ and we bound the expression $\sum_{i=1}^n q_i\,\|\nabla f_i(x)\|^2$, $q_i\ge 0$ (typical second moment bound for estimators based on sampling): it can not be confined solely by $B\,\|\nabla f(x)\|^2$, nor by a constant C, yet smoothness suffices to bound this by $A\left(f(x)-f^*\right)$. Weaker versions of (1) were proposed in Bottou et al. [3] and in Beznosikov et al. [2].

Our first theorem, described informally below, provides required counterexamples of problems and estimators for the diagram.

Theorem 1

The assumptions connected by dashed lines in the diagram are mutually non-implicative.

Our second theorem, described informally below, states that our new Biased ABC assumption is the least restrictive of the assumptions in the literature.

Theorem 2

Biased ABC Assumption is the weakest among the assumptions in the literature.

In fact, conditions (1) and (2) are the least restrictive individually.

We summarize known assumptions on biased stochastic gradients. Estimators satisfying any of them, belong to our general Biased ABC framework with parameters A, B, C, b and c provided in this table.

Assumption	A	B	C	b	c
Asm 1 (SG1) [Beznosikov et al., 2020]	0	β^2	0	$\frac{\alpha}{\beta}$	0
Asm 2 (SG2) [Beznosikov et al., 2020]	0	β^2	0	τ	0
Asm 3 (CON) [Beznosikov et al., 2020]	0	$2\left(2-\frac{1}{\delta}\right)$	0	$\frac{1}{2\delta}$	0
Asm 4 (BVD) [Condat et al., 2022]	0	$2(1+\xi+\eta)$	0	$\frac{1-\eta}{2}$	0
Asm 5 (BREQ) [Khirirat et al., 2018b]	0	ζ	0	ρ	0
Asm 6 (BND) [Ajalloeian and Stich, 2020]	0	$2(M+1)(m+1)$	$2(M+1)\varphi^2+\sigma^2$	$\frac{1-m}{2}$	$\frac{\varphi^2}{2}$
Asm 7 (ABS) [Sahu et al., 2021]	0	2	$2\Delta^2$	$\frac{1}{2}$	$\frac{\Delta^2}{2}$
Asm 8 (FSML) [Bottou et al., 2018]	0	$U+u^2$	Q	q	0

Note that the constants are too pessimistic: given the estimator satisfying one of these assumptions, direct computation of constants in Biased ABC framework for it might lead to much more accurate results.

Convergence Analysis

We show the convergence of BiasedSGD under Biased ABC assumption in nonconvex case.

Theorem 3

Let $\delta^0:=f(x^0)-f^*$, and choose the stepsize such that $0<\gamma\le\frac{b}{LB}$. Then the iterates $\{x^t\}_{t\ge0}$ of BiasedSGD (Algorithm (1)) satisfy

$$\min_{0\le t\le T-1}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right]\le\frac{2\left(1+LA\gamma^2\right)^T}{b\gamma T}\delta^0+\frac{LC\gamma}{b}+\frac{c}{b}.$$

One of the popular generalizations of strong convexity in the literature is the **Polyak–Łojasiewicz assumption:** $\exists\,\mu>0$ s.t. $\forall x\in\mathbb{R}^d:\|\nabla f(x)\|^2\ge 2\mu\left(f(x)-f^*\right)$.

Theorem 4

Suppose PŁ-condition holds. Let $\delta^0:=f(x^0)-f^*$ and choose a stepsize such that $0<\gamma<\min\left\{\frac{\mu b}{L(A+\mu B)},\frac{1}{\mu b}\right\}$. For every $T\ge1$, we have

$$\mathbb{E}\left[f(x^T)-f^*\right]\le\left(1-\gamma\mu b\right)^T\delta^0+\frac{LC\gamma}{2\mu b}+\frac{c}{\mu b}.$$

The function f is μ-strongly convex, if there exists $\mu\ge0$ such that

$$f(x)+\langle\nabla f(x),y-x\rangle+\frac{\mu}{2}\|y-x\|^2\le f(y)\quad\forall x,y\in\mathbb{R}^d.$$

Since PŁ-condition is more general than μ-strong convexity, we can apply Theorem 4 to strongly convex functions.

We examine whether we achieve same rates as obtained under stronger assumptions.

Theorem	Convergence rate	Compared to	Rate we compare to	Match?
Thm 3	$\mathcal{O}\left(\frac{\delta^0 L}{\varepsilon^2}\max\left\{B,\frac{12\delta^0 A}{\varepsilon^2},\frac{2C}{\varepsilon^2}\right\}\right)$	[4]-Thm 2	$\mathcal{O}\left(\frac{\delta^0 L}{\varepsilon^2}\max\left\{B,\frac{12\delta^0 A}{\varepsilon^2},\frac{2C}{\varepsilon^2}\right\}\right)$	✓
Thm 3	$\mathcal{O}\left(\max\left\{\frac{8(M+1)(m+1)}{(1-m)^2\varepsilon},\frac{16(M+1)\varphi^2+2\sigma^2}{(1-m)^2\varepsilon^2}\right\}L\delta^0\right)$	[1]-Thm 4	$\mathcal{O}\left(\max\left\{\frac{M+1}{(1-m)\varepsilon},\frac{2\sigma^2}{(1-m)^2\varepsilon^2}\right\}L\delta^0\right)$	✗
Thm 3	$\mathcal{O}\left(\max\left\{\frac{8Q}{\varepsilon^2 q^2},\frac{4(U+u^2)}{\varepsilon q^2}\right\}L\delta^0\right)$	[3]-Thm 4.8	$\mathcal{O}\left(\max\left\{\frac{8Q}{\varepsilon^2 q^2},\frac{4(U+u^2)}{\varepsilon q^2}\right\}L\delta^0\right)$	✓
Thm 4 PŁ	$\tilde{\mathcal{O}}\left(\max\left\{\frac{2(M+1)(m+1)}{1-m},\frac{2(M+1)\varphi^2+\sigma^2}{\varepsilon\mu(1-m)+2\varphi^2}\right\}\frac{\kappa}{1-m}\right)$	[1]-Thm 6	$\tilde{\mathcal{O}}\left(\max\left\{(M+1),\frac{\sigma^2}{\varepsilon\mu(1-m)+\varphi^2}\right\}\frac{\kappa}{1-m}\right)$	✗
Thm 4 μ-cvx	$\tilde{\mathcal{O}}\left(\max\left\{2,\frac{L(U+u^2)}{q^2\mu},\frac{LQ}{\varepsilon\mu^2q^2}\right\}\right)$	[3]-Thm 4.6	$\tilde{\mathcal{O}}\left(\max\left\{2,\frac{L\left(U+u^2\right)}{q^2\mu},\frac{LQ}{\varepsilon\mu^2q^2}\right\}\right)$	✓
Thm 4 μ-cvx	$\tilde{\mathcal{O}}\left(\left(\frac{\beta^2}{\alpha}\right)^2\frac{L}{\mu}\right)$	[2]-Thm 12	$\tilde{\mathcal{O}}\left(\frac{\beta^2}{\alpha}\frac{L}{\mu}\right)$	✗
Thm 4 μ-cvx	$\tilde{\mathcal{O}}\left(\left(\frac{\beta}{\tau}\right)^2\frac{L}{\mu}\right)$	[2]-Thm 13	$\tilde{\mathcal{O}}\left(\frac{\beta}{\tau}\frac{L}{\mu}\right)$	✗
Thm 4 μ-cvx	$\tilde{\mathcal{O}}\left(\delta^2\frac{L}{\mu}\right)$	[2]-Thm 14	$\tilde{\mathcal{O}}\left(\delta\frac{L}{\mu}\right)$	✗

In most cases, we ensure the same rate, albeit with inferior multiplicative factors due to the broader scope of the analysis. The notation $\tilde{\mathcal{O}}\left(\cdot\right)$ hides a logarithmic factor of $\log\frac{2\delta^0}{\varepsilon}$. We recover the optimal rates in the unbiased case and prove they are optimal in the biased case.

Popular Estimators Within Biased ABC Framework

We introduce a new *Biased independent sampling* estimator. Let $0<p_i\le1$, $S_i=\{i\}$ with probability p_i or $\varnothing$ otherwise, $i\in[n]$, $\sum_{i=1}^n p_i\in(0,n]$. Let $S:=\bigcup_{i=1}^n S_i$. Let $\mathbb{I}_i=1$ if $i\in S$ and $\mathbb{I}_i=0$ otherwise. Define $g(x)=\frac{1}{|S|}\sum_{i=1}^n\mathbb{I}_i\nabla f_i(x)$. Practical use:

- If there is no access to the entire dataset, a *fixed batch strategy* can be employed. This strategy involves sampling a single batch S at step 0 and subsequently.
- $\forall i\in[n]$, an oracle decides with an unknown probability p_i whether to provide the information of ∇f_i at the iteration t or not.

We provide a description of popular gradient estimators in terms of the Biased ABC framework. Unlike the existing assumptions, which implicitly assume that the bias comes from either perturbation or compression, Biased ABC also holds in settings such as subsampling.

Estimator	Def	A	B	C	b	c
Biased independent sampling [This paper]	Def. 1	$\frac{\max_i\{L_i\}}{\min_i p_i}$	0	$2A\Delta^*+s^2$	$\min_i\{p_i\}$	0
Top-k [Aji and Heafield, 2017]	Def. 3	0	1	0	$\frac{k}{d}$	0
Rand-k Stich et al. [2018]	Def. 4	0	$\frac{d}{k}$	0	1	0
Biased Rand-k [Beznosikov et al., 2020]	Def. 5	0	$\frac{k}{d}$	0	$\frac{k}{d}$	0
Adaptive random sparsification [Beznosikov et al., 2020]	Def. 6	0	1	0	$\frac{1}{d}$	0
General unbiased rounding [Beznosikov et al., 2020]	Def. 7	0	$\sup_{k\in\mathbb{Z}}\frac{a_k^2+a_{k+1}^2}{4a_k a_{k+1}}+\frac{1}{2}$	0	1	0
Natural compression [Horváth et al., 2022]	Def. 9	0	$\frac{9}{8}$	0	1	0
Scaled integer rounding [Sapio et al., 2021]	Def. 15	0	2	$\frac{2d}{\chi^2}$	$\frac{1}{2}$	$\frac{d}{2\chi^2}$

We list several popular estimators and indicate which of the assumptions in the literature they satisfy. Only Assumption 9 (Biased ABC) encompasses them all.

Estimator \ Assumption	A1	A2	A3	A4	A5	A6	A7	A8	A9
Biased independent sampling [This paper]	✗	✗	✗	✗	✗	✗	✗	✗	✓
Top-k sparsification [Aji and Heafield, 2017]	✓	✓	✓	✓	✓	✓	✗	✓	✓
Rand-k [Stich et al., 2018]	✓	✓	✗	✓	✗	✓	✗	✓	✓
Biased Rand-k [Beznosikov et al., 2020]	✓	✓	✓	✓	✗	✓	✗	✓	✓
Adaptive random sparsification [Beznosikov et al., 2020]	✓	✓	✓	✓	✗	✓	✗	✓	✓
General unbiased rounding [Beznosikov et al., 2020]	✓	✓	✗	✓	✗	✓	✗	✓	✓
Natural compression [Horváth et al., 2022]	✓	✓	✓	✓	✗	✓	✗	✓	✓
Scaled integer rounding [Sapio et al., 2021]	✓	✓	✗	✓	✓	✓	✓	✓	✓

References

[1] Ahmad Ajalloeian and Sebastian U Stich. Analysis of SGD with biased gradient estimators. arXiv preprint arXiv:2008.00051, 2020.

[2] Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. arXiv preprint arXiv:2002.12410, 2020.

[3] Léon Bottou, Frank Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. SIAM Review, 60(2):223–311, 2018.

[4] Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. Transactions on Machine Learning Research, 2023.