A Guide Through the Zoo of Biased SGD

NeurIPS 2023

Collaborators

Grigory Malinovsky
(KAUST, Saudi Arabia)

Igor Sokolov
(KAUST, Saudi Arabia)

Peter Richtárik
(KAUST, Saudi Arabia)

Problem Definition

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

Minimal value of the function

$$\mathbb{E}\left[f(x) - f^*\right] \leq \varepsilon \text{ (convergence in function values)}$$

Find $x \in \mathbb{R}^d$

$$\mathbb{E}\left\|\nabla f(x)\right\|^2 \leq \varepsilon^2 \text{ (gradient norm convergence)}$$

$$\mathbb{E}\left\|x - x^*\right\|^2 \leq \varepsilon \left\|x^0 - x^*\right\|^2 \text{ (iterate convergence)}$$

Minimum of the function

Method Studied: Biased SGD

Algorithm 1 Biased Stochastic Gradient Descent (BiasedSGD)

Input: initial point $x^0 \in \mathbb{R}^d$; learning rate $\gamma > 0$
1: **for** $t = 0, 1, 2, \ldots$ **do**
2: Construct a (possibly biased) estimator $g^t \stackrel{\text{def}}{=} g(x^t)$ of the gradient $\nabla f(x^t)$
3: Compute $x^{t+1} = x^t - \gamma g^t$
4: **end for**

New Assumption: Biased ABC

Biased ABC Assumption. There exist constants $A, B, C, b, c \geq 0$ such that the gradient estimator $g(x)$, for every $x \in \mathbb{R}^d$, satisfies

$$\begin{aligned} \langle \nabla f(x), \mathbb{E}[g(x)] \rangle &\geq b \left\| \nabla f(x) \right\|^2 - c, \\ \mathbb{E}\left[\left\| g(x) \right\|^2 \right] &\leq 2A \left(f(x) - f^* \right) + B \left\| \nabla f(x) \right\|^2 + C. \end{aligned}$$

Motivation for A-term: if $f(x) = \sum_{i=1}^n f_i(x)$ and $\mathbb{E}\left[\|g(x)\|^2\right] = \sum_{i=1}^n q_i \left\| f_i(x) \right\|^2$, $q_i \geq 0$, then $\mathbb{E}\left[\|g(x)\|^2\right]$ can not be bounded solely by $B \left\| \nabla f(x) \right\|^2 + C$.

Diagram of Assumptions

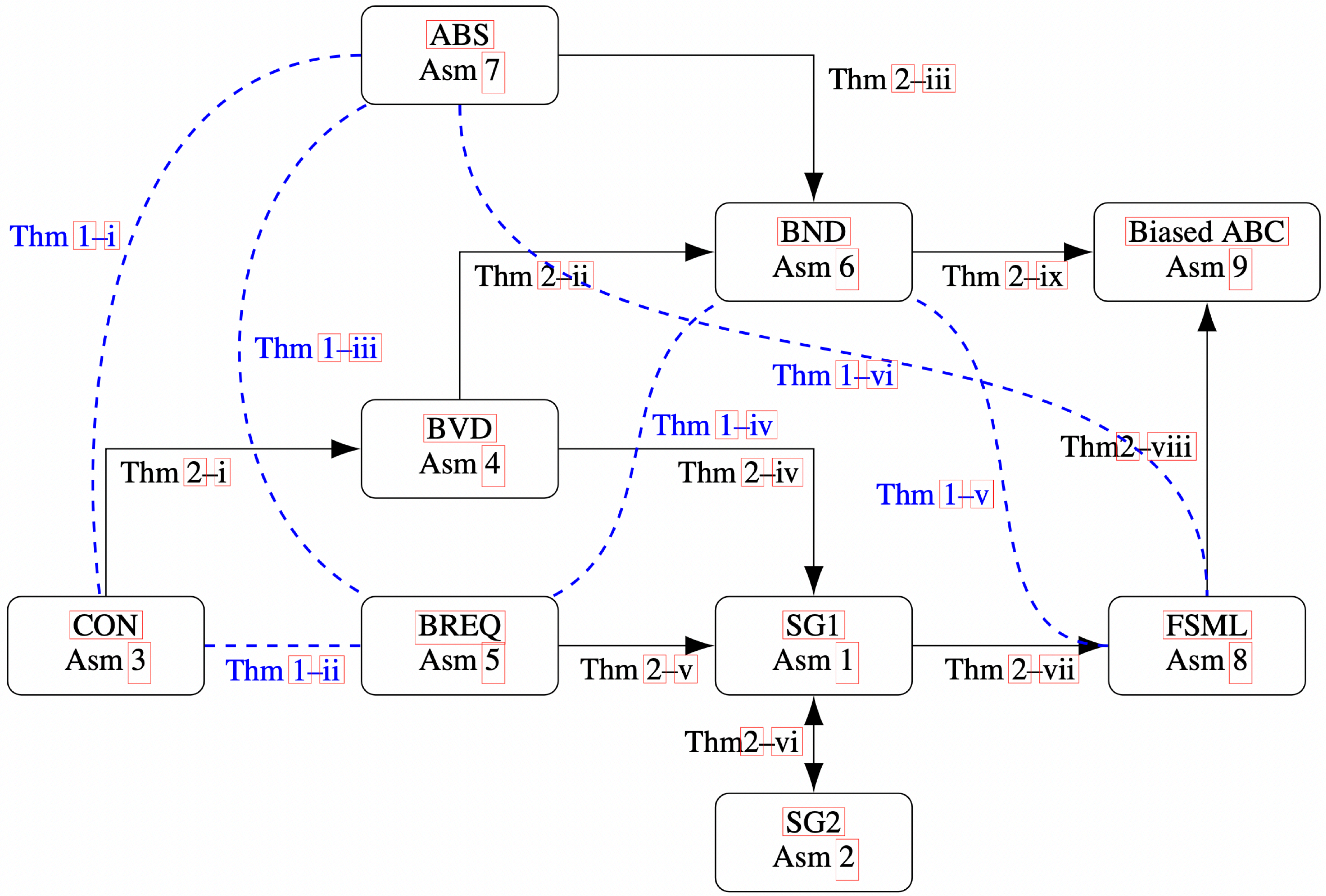

Figure 1: Assumption hierarchy. A single arrow indicates an implication and an absence of a reverse implication. The implications are transitive. A dashed line indicates a mutual abscence of implications. Our newly proposed assumption Biased ABC is the most general one.

Popular Estimators Within Biased ABC Framework

Name of an estimator	Definition	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
Distributed general biased rounding This paper	Def. 2	A_r	B_r	C_r	b_r	c_r
Top-k [Aji and Heafield, 2017; Alistarh et al., 2018]	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	$\frac{Z}{4}$	0	1	0
General biased rounding [Beznosikov et al., 2020]	Def. 8	0	F^2	0	$\frac{G^2}{F}$	0
Natural compression [Horváth et al., 2022]	Def. 9	0	$\frac{9}{8}$	0	1	0
General exponential dithering [Beznosikov et al., 2020]	Def. 10	0	H_a	0	1	0
Natural dithering [Horváth et al., 2022]	Def. 11	0	H_2	0	1	0
Composition of Top-k and exp dithering [Beznosikov et al., 2020]	Def. 12	0	H_a^2	0	$\frac{k}{dH_a}$	0
Gaussian smoothing [Polyak, 1987]	Def. 13	A_{GS}	B_{GS}	C_{GS}	b_{GS}	c_{GS}
Hard-threshold sparsifier [Sahu et al., 2021]	Def. 14	0	1	0	1	$w^2 d$
Scaled integer rounding [Sapio et al., 2021]	Def. 15	0	2	$\frac{2d}{\chi^2}$	$\frac{1}{2}$	$\frac{d}{2\chi^2}$
Biased dithering [Khirirat et al., 2018a]	Def. 16	0	d	0	1	0
Sign compression [Karimireddy et al., 2019]	Def. 17	0	$4 - \frac{2}{d}$	0	$\frac{1}{2d}$	0

Table 8: Summary of the estimators with respective parameters A, B, C, b and c, satisfying our general Biased ABC framework. Constants L_i are from Assumption 13, Δ^* is defined in (26), A_r, B_r, C_r, b_r, c_r are defined in (27)–(31), Z is defined in (32), F and G are defined in (33), H_a is defined in (34), $A_{GS}, B_{GS}, C_{GS}, b_{GS}, c_{GS}$ are defined in (35).

Popular Estimators in Different Frameworks

Name of an estimator \ Assumption	A1	A2	A3	A4	A5	A6	A7	A8	A9
Biased independent sampling [This paper]	✗	✗	✗	✗	✗	✗	✗	✗	✓
Distributed general biased rounding [This paper]	✗	✗	✗	✗	✗	✗	✗	✗	✓
Top-k sparsification [Aji and Heafield, 2017; Alistarh et al., 2018]	✓	✓	✓	✓	✓	✓	✗	✓	✓
Rand-k [Stich et al., 2018]	✓	✓	✗	✓	✗	✓	✗	✓	✓
Biased Random-k [Beznosikov et al., 2020]	✓	✓	✓	✓	✗	✓	✗	✓	✓
Adaptive random sparsification [Beznosikov et al., 2020]	✓	✓	✓	✓	✗	✓	✗	✓	✓
General unbiased rounding [Beznosikov et al., 2020]	✓	✓	✗	✓	✗	✓	✗	✓	✓
General biased rounding [Beznosikov et al., 2020]	✓	✓	✓	✓	✓	✓	✗	✓	✓
Natural compression [Horváth et al., 2022]	✓	✓	✓	✓	✗	✓	✗	✓	✓
General exponential dithering [Beznosikov et al., 2020]	✓	✓	✓	✓	✗	✓	✗	✓	✓
Natural dithering [Horváth et al., 2022]	✓	✓	✓	✓	✗	✓	✗	✓	✓
Composition of Top-k and exp dithering [Beznosikov et al., 2020]	✓	✓	✓	✓	✗	✓	✗	✓	✓
Gaussian smoothing [Polyak, 1987]	✗	✗	✗	✗	✗	✓	✗	✗	✓
Hard-threshold sparsifier [Sahu et al., 2021]	✗	✓	✓	✓	✗	✓	✓	✓	✓
Scaled integer rounding [Sapio et al., 2021]	✓	✓	✗	✓	✓	✓	✓	✓	✓
Biased dithering [Khirirat et al., 2018a]	✓	✓	✗	✗	✓	✗	✗	✓	✓
Sign compression [Karimireddy et al., 2019]	✓	✓	✓	✓	✓	✓	✗	✓	✓

Table 9: Summary on an inclusion of popular estimators into every known framework.

Main results

General Nonconvex case. Let f be L-smooth. Let g be a gradient estimator satisfying BiasedABC assumption. Let $\delta^0 = f(x^0) - f^*$, and choose the stepsize $0 < \gamma \leq \frac{b}{LB}$. Then the iterates $\{x_t\}_{t \geq 0}$ of BiasedSGD satisfy

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

Main results

Convergence under PŁ-condition. Let f be L-smooth and satisfy PŁ-condition with constant $\mu > 0$. Let g be a gradient estimator satisfying BiasedABC assumption. Let $\delta^0 = f(x^0) - f^*$, choose a stepsize

$0 < \gamma \leq \min\{\mu b/(L(A+\mu B)), 1/(\mu b)\}$. Then, for every $T \geq 1$, the iterates $\{x_t\}_{t\geq 0}$ of BiasedSGD satisfy

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

Strongly convex case. Since PŁ-condition is more general than μ-strong convexity assumption, the same result holds for any μ-strongly convex function f. Notice that it implies an iterate convergence since

$$\| x^T - x^* \|^2 \leq \frac{2}{\mu} \mathbb{E}\left[f\left(x^T\right) - f(x^*)\right].$$

Convergence rates comparison

Theorem	Convergence rate	Compared to	Rate we compare to	Match?
Thm 3	$\mathcal{O}\left(\frac{\delta^0 L}{\varepsilon^2}\max\left\{B, \frac{12\delta^0 A}{\varepsilon^2}, \frac{2C}{\varepsilon^2}\right\}\right)$	25-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	$\mathcal{O}\left(\max\left\{\frac{8Q}{\varepsilon^2 q^2}, \frac{4(U+u^2)}{\varepsilon q^2}\right\} L\delta^0\right)$	5-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	$\tilde{\mathcal{O}}\left(\max\left\{\frac{2(M+1)(m+1)}{1-m}, \frac{2(M+1)\varphi^2+\sigma^2}{\epsilon\mu(1-m)+2\varphi^2}\right\} \frac{\kappa}{1-m}\right)$	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 12	$\tilde{\mathcal{O}}\left(\max\left\{2, \frac{L(U+u^2)}{q^2\mu}, \frac{LQ}{\varepsilon\mu^2 q^2}\right\}\right)$	5-Thm 4.6	$\tilde{\mathcal{O}}\left(\max\left\{2, \frac{L\left(U+u^2\right)}{q^2\mu}, \frac{LQ}{\varepsilon\mu^2 q^2}\right\}\right)$	✓
Thm 12	$\tilde{\mathcal{O}}\left(\left(\frac{\beta^2}{\alpha}\right)^2 \frac{L}{\mu}\right)$	4-Thm 12	$\tilde{\mathcal{O}}\left(\frac{\beta^2}{\alpha}\frac{L}{\mu}\right)$	✗
Thm 12	$\tilde{\mathcal{O}}\left(\left(\frac{\beta}{\tau}\right)^2 \frac{L}{\mu}\right)$	4-Thm 13	$\tilde{\mathcal{O}}\left(\frac{\beta}{\tau}\frac{L}{\mu}\right)$	✗
Thm 12	$\tilde{\mathcal{O}}\left(\delta^2 \frac{L}{\mu}\right)$	4-Thm 14	$\tilde{\mathcal{O}}\left(\delta \frac{L}{\mu}\right)$	✗

Table 4: Complexity comparison. We examine whether we can achieve the same convergence rate as obtained under stronger assumptions. In most cases, we ensure the same rate, albeit with inferior multiplicative factors due to the broader scope of the analysis. The notation $\tilde{\mathcal{O}}(\cdot)$ hides a logarithmic factor of $\log \frac{2\delta^0}{\varepsilon}$.