A Stochastic Derivative Free Optimization Method with Momentum

Eduard Gorbunov[1] Adel Bibi[2] Ozan Sener[3] El Houcine Bergou[2,4] Peter Richtárik[1,2]

[1]MIPT, Russia [2]KAUST, Saudi Arabia [3]Intel Labs [4]MaIAGE, INRA, France

1. Introduction

In this paper, we consider the following minimization problem

$$\min_{x\in\mathbb{R}^d} f(x), \quad \text{where} \tag{1}$$

- $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth: $\|\nabla f(x) - \nabla f(y)\|_2 \le L\|x-y\|_2$,
- f is bounded from below by $f(x^*)$ where x^* is a minimizer.

2. Stochastic Three Points Method

Stochastic Three Points method [1] is a new method aimed to solve (1). The key properties of STP are its **simplicity**, **generality** and **practicality**.

Algorithm 1 [1] (STP).

Parameters: some distribution $\mathcal{D}$ over $\mathbb{R}^d$, stepsizes $\{\gamma^k\}_{k\ge 0}$
Initialization: Choose $x_0 \in \mathbb{R}^n$
for $k = 0, 1, 2 \dots$ **do**
  Draw a fresh sample s^k from $\mathcal{D}$
  $x^{k+1} = \arg\min\{f(x^k), f(x^k + \gamma^k s^k), f(x^k - \gamma^k s^k)\}$
end for

3. Key Assumption

Assumption 1. The probability distribution $\mathcal{D}$ on $\mathbb{R}^d$ satisfies the following properties:
1) The quantity $\gamma_{\mathcal{D}} \overset{\text{def}}{=} \mathbf{E}_{s\sim\mathcal{D}}\|s\|_2^2$ is finite.
2) There is a constant $\mu_{\mathcal{D}} > 0$ for a norm $\|\cdot\|_{\mathcal{D}}$ in $\mathbb{R}^d$ such that for all $g \in \mathbb{R}^d$

$$\mathbf{E}_{s\sim\mathcal{D}}|\langle g, s\rangle| \ge \mu_{\mathcal{D}}\|g\|_{\mathcal{D}}.$$

- If $\mathcal{D}$ is the uniform distribution on the unit sphere in $\mathbb{R}^d$, then $\gamma_{\mathcal{D}} = 1$ and $\mathbf{E}_{s\sim\mathcal{D}}|\langle g, s\rangle| \sim \frac{1}{\sqrt{2\pi d}}\|g\|_2$.
- If $\mathcal{D} = N(0, \frac{I}{d})$ then $\gamma_{\mathcal{D}} = 1$ and $\mathbf{E}_{s\sim\mathcal{D}}|\langle g, s\rangle| = \frac{\sqrt{2}}{\sqrt{d\pi}}\|g\|_2$.
- If $\mathcal{D}$ is the uniform distribution on $\{e_1, \dots, e_d\}$, then $\gamma_{\mathcal{D}} = 1$ and $\mathbf{E}_{s\sim\mathcal{D}}|\langle g, s\rangle| = \frac{1}{d}\|g\|_1$.

4. First Ingredient: Momentum Term

Below we introduce Polyak's heavy ball momentum using special technique inspired by virtual iterates analysis from [4].

Algorithm 2 (SMTP).

Parameters: some distribution $\mathcal{D}$ over $\mathbb{R}^d$, stepsizes $\{\gamma^k\}_{k\ge0}$, momentum parameter $\beta \in [0, 1)$
Initialization: Choose $x_0 \in \mathbb{R}^n$
for $k = 0, 1, 2 \dots$ **do**
  Draw a fresh sample s^k from $\mathcal{D}$
  Let $v_+^k = \beta v^{k-1} + s^k$ and $v_-^k = \beta v^{k-1} - s^k$
  Let $x_+^{k+1} = x^k - \gamma^k v_+^k$ and $x_-^{k+1} = x^k - \gamma^k v_-^k$
  Let $z_+^{k+1} = x_+^{k+1} - \frac{\gamma^k\beta}{1-\beta}v_+^k$ and $z_-^{k+1} = x_-^{k+1} - \frac{\gamma^k\beta}{1-\beta}v_-^k$
  Set $z^{k+1} = \arg\min\left\{f(z^k), f(z_+^{k+1}), f(z_-^{k+1})\right\}$
  Set $(x^{k+1}, v^{k+1}) = \begin{cases}(x_+^{k+1}, v_+^{k+1}), & \text{if } z^{k+1} = z_+^{k+1}\\ (x_-^{k+1}, v_-^{k+1}), & \text{if } z^{k+1} = z_-^{k+1}\\ (x^k, v^k), & \text{if } z^{k+1} = z^k\end{cases}$
end for

Key Lemma. Assume that f is L-smooth and $\mathcal{D}$ satisfies Assumption 1. Then for the iterates of SMTP the following inequalities hold:

$$f(z^{k+1}) \le f(z^k) - \frac{\gamma^k}{1-\beta}|\langle\nabla f(z^k), s^k\rangle| + \frac{L(\gamma^k)^2}{2(1-\beta)^2}\|s^k\|_2^2$$

and

$$\mathbf{E}_{s^k\sim\mathcal{D}}\left[f(z^{k+1})\right] \le f(z^k) - \frac{\gamma^k\mu_{\mathcal{D}}}{1-\beta}\|\nabla f(z^k)\|_{\mathcal{D}} + \frac{L(\gamma^k)^2\gamma_{\mathcal{D}}}{2(1-\beta)^2}.$$

Using the lemma above one can get convergence guarantees for SMTP in the similar way as it was done in [1].

5. Second Ingredient: Importance Sampling

Assumption 2 (Coordinate-wise L-smoothness). We assume that the objective f has coordinate-wise Lipschitz gradient, with Lipschitz constants $L_1, \dots, L_d > 0$, i.e.

$$f(x + he_i) \le f(x) + \nabla_i f(x)h + \frac{L_i}{2}h^2, \qquad \forall x \in \mathbb{R}^d, h \in \mathbb{R}.$$

Algorithm 3 (SMTP_IS).

Parameters: stepsize parameters $w_1, \dots, w_n > 0$, probabilities $p_1, \dots, p_n > 0$ summing to 1, momentum parameter $\beta \in [0, 1)$
Initialization: Choose $x_0 \in \mathbb{R}^n$
Set $v^{-1} = 0$ and $z^0 = x^0$
for $k = 0, 1, 2 \dots$ **do**
  Select $i_k = i$ with probability $p_i > 0$
  Choose stepsize γ_i^k proportional to $\frac{1}{w_{i_k}}$
  Let $v_+^k = \beta v^{k-1} + e_{i_k}$ and $v_-^k = \beta v^{k-1} - e_{i_k}$
  Let $x_+^{k+1} = x^k - \gamma_i^k v_+^k$ and $x_-^{k+1} = x^k - \gamma_i^k v_-^k$
  Let $z_+^{k+1} = x_+^{k+1} - \frac{\gamma_i^k\beta}{1-\beta}v_+^k$ and $z_-^{k+1} = x_-^{k+1} - \frac{\gamma_i^k\beta}{1-\beta}v_-^k$
  Set $z^{k+1} = \arg\min\left\{f(z^k), f(z_+^{k+1}), f(z_-^{k+1})\right\}$
  Set $(x^{k+1}, v^{k+1}) = \begin{cases}(x_+^{k+1}, v_+^{k+1}), & \text{if } z^{k+1} = z_+^{k+1}\\ (x_-^{k+1}, v_-^{k+1}), & \text{if } z^{k+1} = z_-^{k+1}\\ (x^k, v^k), & \text{if } z^{k+1} = z^k\end{cases}$
end for

6. Summary of The Convergence Results

Assumptions on f	SMTP Complexity	Importance Sampling	SMTP_IS Complexity
None	${}^{2r_0L\gamma_{\mathcal{D}}}/_{\mu_{\mathcal{D}}^2\epsilon^2}$	$p_i = {}^{L_i}/_{\sum_{i=1}^d L_i}$	${}^{2r_0 d\sum_{i=1}^d L_i}/_{\epsilon^2}$
Convex, $R_0 < \infty$	${}^{L\gamma_{\mathcal{D}}R_0^2\ln(2r_0/\varepsilon)}/_{\mu_{\mathcal{D}}^2\varepsilon}$	$p_i = {}^{L_i}/_{\sum_{i=1}^d L_i}$	${}^{R_0^2\ln(2r_0/\varepsilon)d\sum_{i=1}^d L_i}/_{\epsilon}$
μ-strongly convex	${}^{L\ln(2r_0/\varepsilon)}/_{\mu\mu_{\mathcal{D}}^2}$	$p_i = {}^{L_i}/_{\sum_{i=1}^d L_i}$	${}^{\ln(2r_0/\varepsilon)\sum_{i=1}^d L_i}/_{\mu}$

Note that $r_0 = f(x_0) - f(x_*)$ and that all assumptions listed are in addition to L-smoothness. Complexity means number of iterations in order to guarantee $\mathbf{E}\|\nabla f(\overline{z}^K)\|_{\mathcal{D}} \le \varepsilon$ for the non-convex case, $\mathbf{E}\left[f(z^K) - f(x^*)\right] \le \varepsilon$ for convex and strongly convex cases. $R_0 = \max\left\{\|x - x^*\|_{\mathcal{D}}^* \mid f(x) \le f(x^0)\right\} < +\infty$. We notice that for SMTP_IS $\|\cdot\|_{\mathcal{D}} = \|\cdot\|_1$ and $\|\cdot\|_{\mathcal{D}}^* = \|\cdot\|_\infty$ in non-convex and convex cases and $\|\cdot\|_{\mathcal{D}} = \|\cdot\|_2$ in the strongly convex case.

7. Numerical Experiments

We conduct extensive experiments on challenging non-convex problems on the continuous control task from the MuJoCO suit [3]. In particular, we address the problem of model-free control of a dynamical system.

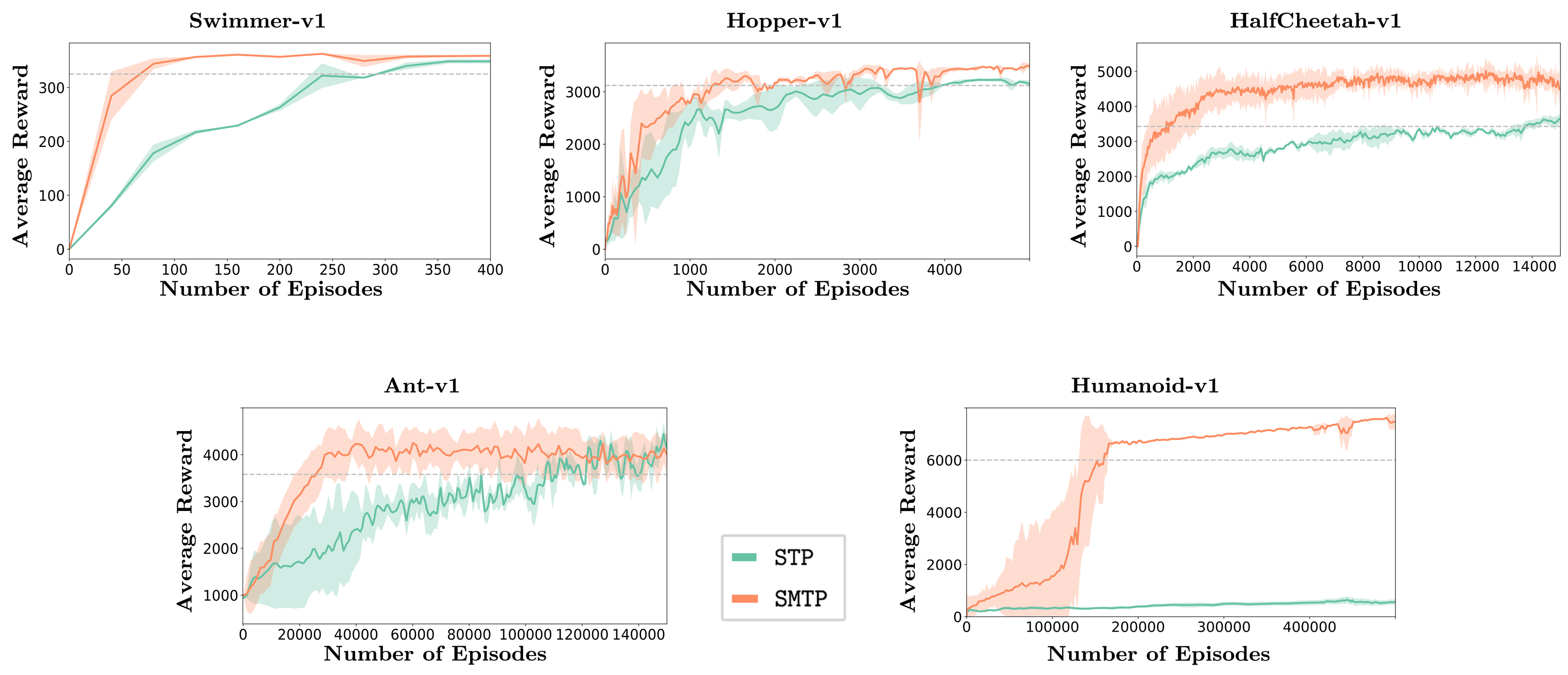

Figure: SMTP is far superior to STP on all 5 different MuJoCo tasks. The horizontal dashed lines are the thresholds used in Table 1 to demonstrate complexity of each method.

Table: For each MuJoCo task, we report the average number of episodes required to achieve a predefined reward threshold. Results for our method is averaged over five random seeds, the rest is copied from [2] (N/A means the method failed to reach the threshold. UNK means the results is unknown since they are not reported in the literature.)

	Threshold	STP	STP_IS	SMTP	SMTP_IS	ARS(V1-t)	ARS(V2-t)	NG-lin	TRPO-nn
Swimmer-v1	325	320	110	80	100	100	427	1450	N/A
Hopper-v1	3120	3970	2400	1264	1408	51840	1973	13920	10000
HalfCheetah-v1	3430	13760	4420	1872	1624	8106	1707	11250	4250
Ant-v1	3580	107220	43860	19890	14420	58133	20800	39240	73500
Humanoid-v1	6000	N/A	530200	161230	207160	N/A	142600	130000	UNK

8. Bibliography

[1] E. H. Bergou, E. Gorbunov, and P. Richtárik. Stochastic three points method for unconstrained smooth minimization. *arXiv preprint arXiv:1902.03591*, 2019.

[2] H. Mania, A. Guy, and B. Recht. Simple random search provides a competitive approach to reinforcement learning. *arXiv preprint arXiv:1803.07055*, 2018.

[3] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 5026–5033. IEEE, 2012.

[4] T. Yang, Q. Lin, and Z. Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. *arXiv preprint arXiv:1604.03257*, 2016.