A Stochastic Derivative-Free Optimization Method with Importance Sampling: Theory and Learning to Control

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

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

We consider the following minimization problem:

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

Assumption 1 *The objective function f has coordinatewise Lipschitz gradient, with Lispchitz constants $L_1,\dots,L_n>0$. Moreover, f is bounded from below by $f(x_*)\in\mathbb{R}$. That is, f satisfies*

$$f(x_*) \le f(x+te_i) \le f(x) + \nabla_i f(x) t + \frac{L_i}{2}t^2$$

for all $x\in\mathbb{R}^n$ and $t\in\mathbb{R}$, where $\nabla_i f(x)$ is the ith partial derivative of f at x.

Algorithm 1 Stochastic Three Points Method with Importance Sampling ($\mathtt{STP_{IS}}$)

Initialization

Choose initial iterate $x_0\in\mathbb{R}^n$, stepsize parameters $v_1,\dots,v_n>0$ and probabilities $p_1,\dots,p_n>0$ summing up to 1.

For $k=0,1,2,\dots$

1. Select $i_k=i$ with probability $p_i>0$
2. Choose stepsize α_{i_k} proportional to $1/v_{i_k}$
3. Let $x_+ = x_k+\alpha_k e_{i_k}$ and $x_- = x_k - \alpha_k e_{i_k}$
4. $x_{k+1} = \arg\min\{f(x_k), f(x_+), f(x_-)\}$

Assumptions on f	Uniform Sampling Complexity	Importance Sampling	Importance Sampling Complexity [NEW]
None	$\frac{4\sqrt{2}r_0 n^2 L}{\epsilon^2}$	$p_i=\frac{\sqrt{L_i}}{\sum_{i=1}^n\sqrt{L_i}}$	$\frac{4\sqrt{2}r_0\left(\sum_{i=1}^n\sqrt{L_i}\right)^2}{\epsilon^2}$
None	$\frac{4\sqrt{2}r_0 n^2 L}{\epsilon^2}$	$p_i=\frac{L_i}{\sum_{i=1}^n L_i}$	$\frac{4\sqrt{2}r_0 n\left(\sum_{i=1}^n L_i\right)}{\epsilon^2}$
Convex, $R_0<\infty$	$8R_0^2n^2L\left(\frac{1}{\epsilon}-\frac{1}{r_0}\right)$	$p_i=\frac{L_i}{\sum_{i=1}^n L_i}$	$8R_0^2 n\sum_{i=1}^n L_i\left(\frac{1}{\epsilon}-\frac{1}{r_0}\right)$
λ-strongly convex	$\frac{nL}{\lambda}\log\left(\frac{r_0}{\epsilon}\right)$	$p_i=\frac{L_i}{\sum_{i=1}^n L_i}$	$\frac{\sum_{i=1}^n L_i}{\lambda}\log\left(\frac{r_0}{\epsilon}\right)$

Theorem 1 *Let Assumption 1 be satisfied. Choose $\alpha_{i_k}=\frac{\epsilon}{nv_{i_k}}$ where $\sum_{i=1}^n\frac{p_iL_i}{v_i^2}<2n\left(\min_i\frac{p_i}{v_i}\right)$. If*

$$K\ge\frac{2n(f(x_0)-f(x_*))}{\left(\min_i\frac{p_i}{v_i}\right)\left(1-\frac{\sum_{i=1}^n\frac{p_iL_i}{v_i^2}}{2n\left(\min_i\frac{p_i}{v_i}\right)}\right)\epsilon^2},\ \text{then}\ \min_{k=0,1,\dots,K}\mathbb{E}\left[\|\nabla f(x_k)\|_1\right]\le\epsilon.$$

Theorem 2 *Let Assumption 1 be satisfied. Choose $\alpha_{i_k}=\frac{\alpha_0}{v_{i_k}\sqrt{k+1}}$ where $\alpha_0>0$. If*

$$K\ge\frac{2\left(\frac{\sqrt{2}(f(x_0)-f(x_*))}{\alpha_0}+\frac{\alpha_0\sum_{i=1}^n\frac{p_iL_i}{v_i^2}}{2}\right)^2}{\left(\min_i\frac{p_i}{v_i}\right)^2\epsilon^2},\ \text{then}\ \min_{k=0,1,\dots,K}\mathbb{E}\left[\|\nabla f(x_k)\|_1\right]\le\epsilon.$$

Theorem 3 *Let Assumption 1 hold and be satisfied. Choose $\alpha_{i_k}=\frac{|f(x_k+te_{i_k})-f(x_k)|}{tv_{i_k}}$ and sufficiently small t. If*

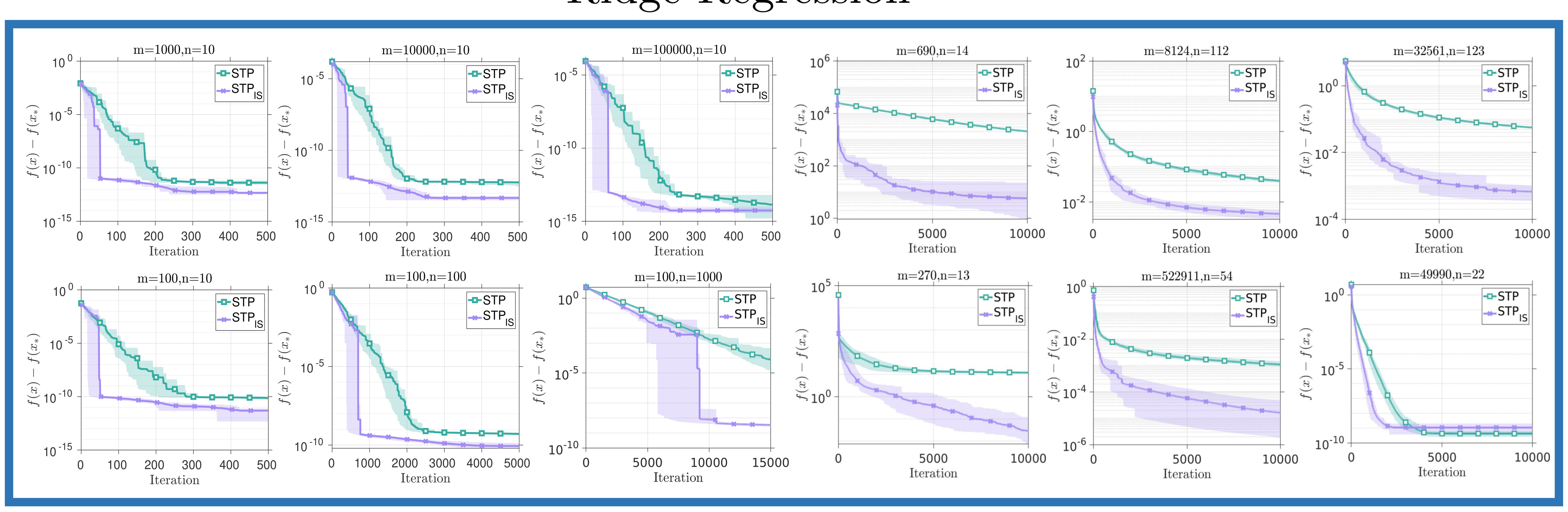
$$k\ge\frac{8R_0^2n}{\min_i\frac{p_i}{v_i}}\left(\frac{1}{\epsilon}-\frac{1}{r_0}\right),\ \text{then}\ \mathbb{E}[f(x_k)]-f(x_*)\le\epsilon.$$

Theorem 4 *Let Assumption 1 hold for λ-strongly convex f. Choose $\alpha_{i_k}=\frac{|f(x_k+te_{i_k})-f(x_k)|}{tv_{i_k}}$ and a sufficiently small t. If*

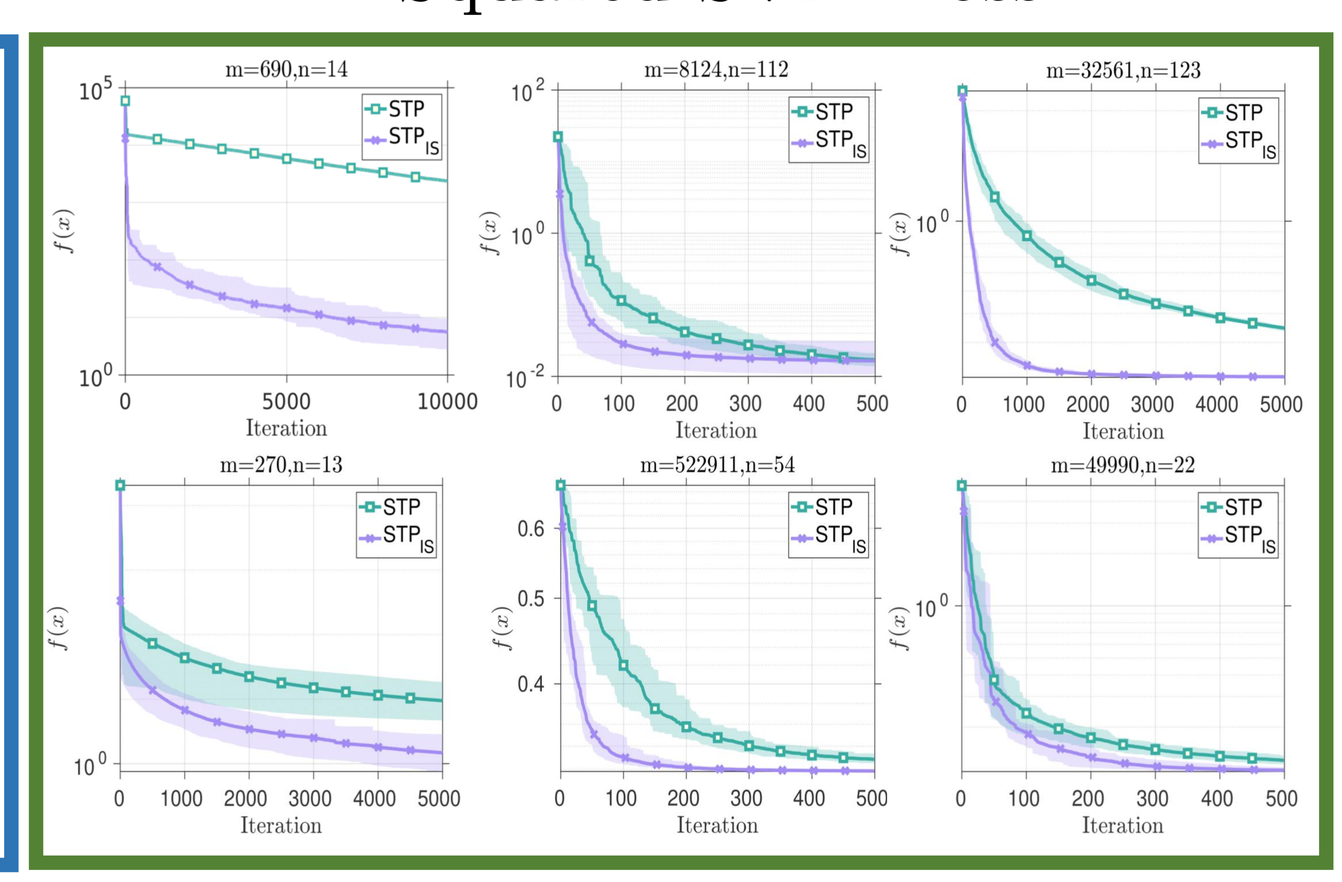
$$k\ge\frac{\max_i\frac{v_i}{p_i}}{\lambda}\log\left(\frac{2\left(f(x_0)-f(x_*)\right)}{\epsilon}\right),\ \text{then}\ \mathbb{E}[f(x_k)]-f(x_*)\le\epsilon.$$

$r_0=f(x_0)-f(x_*)$ $\qquad$ $R_0\stackrel{\text{def}}{=}\max_x\{\|x-x_*\|_\infty\ :f(x)\le f(x_0)\}<\infty$

Ridge Regression

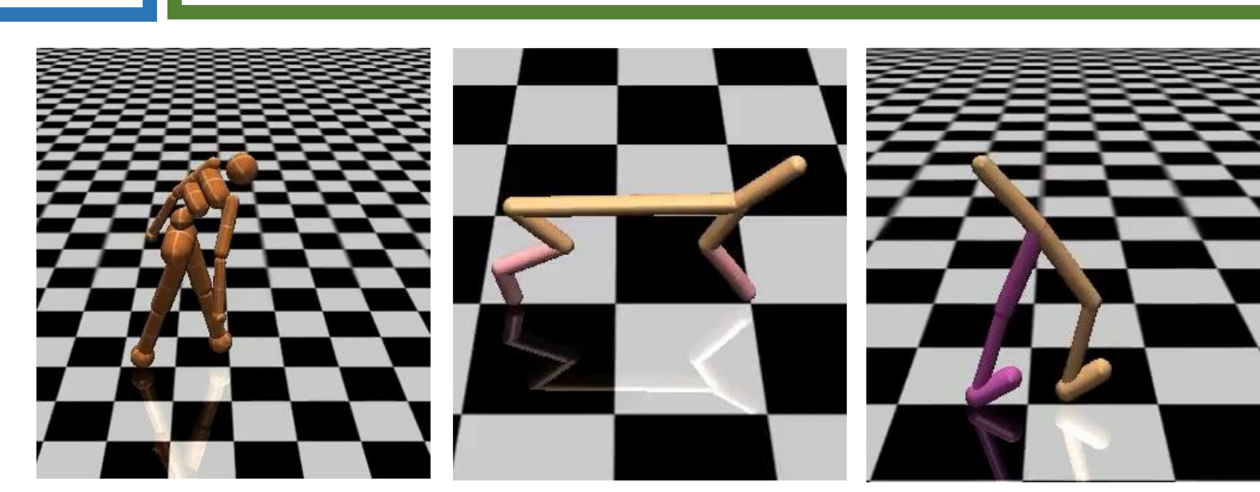


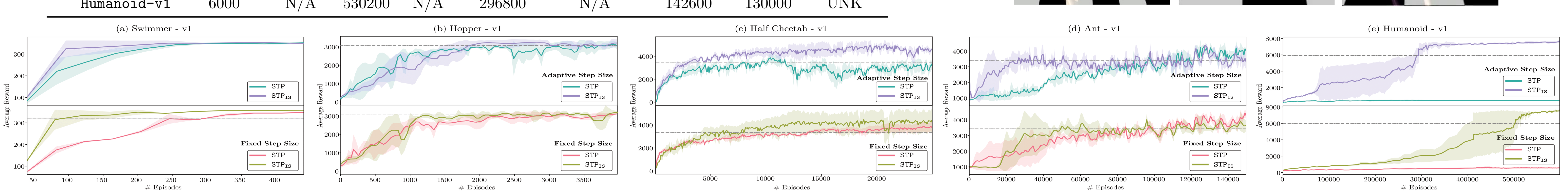
Squared SVM Loss



Continuous Control Experiments:

	Threshold	Fixed Step Size STP	Fixed Step Size $\mathtt{STP_{IS}}$	Adaptive Step Size STP	Adaptive Step Size $\mathtt{STP_{IS}}$	ARS(V1-t)	ARS(V2-t)	NG-lin	TRPO-nn
Swimmer-v1	325	320	110	200	90	100	427	1450	N/A
Hopper-v1	3120	3970	2400	3720	1870	51840	1973	13920	10000
HalfCheetah-v1	3430	13760	4420	5040	2710	8106	1707	11250	4250
Ant-v1	3580	107220	43860	96980	26480	58133	20800	39240	73500
Humanoid-v1	6000	N/A	530200	N/A	296800	N/A	142600	130000	UNK





(a) Swimmer - v1 (b) Hopper - v1 (c) Half Cheetah - v1 (d) Ant - v1 (e) Humanoid - v1

References. "Stochastic Three Points Method for Unconstrained Smooth Minimization". El Houcine Bergou, Eduard Gorbunov, Peter Richtárik.

Acknowledgments. This work was partially supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research.