Università di Genova

DIPARTIMENTO INFORMATICA, BIOINGEGNERIA, ROBOTICA E INGEGNERIA DEI SISTEMI

Computer Science Workshop
PhD program in Computer Science and Systems Engineering

Structured Stochastic Zeroth-order Descent

Marco Rando, Cesare Molinari, Silvia Villa, Lorenzo Rosasco

Zeroth-order Optimization

Solve

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

given $f(x)$ **BUT NOT** $\nabla f(x)$.

Common in

- Adversarial ML [2]
- RL [3]
- Economics [5]
- Robotics [7]

Stochastic Zeroth-order Optimization

Solve

$$\min_{x\in\mathbb{R}^d} f(x) := \mathbb{E}_Z[F(x,Z)] \qquad (1)$$

given $F(x,z)$ **BUT NOT** $\nabla F(x,z)$.

Example: Empirical Risk Minimization

$$f(x) = \frac{1}{n}\sum_{i=0}^{n} F(x, z_i)$$

with F loss function and $(z_i)_{i=0}^n$ data samples.

Algorithm

Structured Stochastic Zeroth-order Descent

For $k = 1, \cdots$, compute

$$x_{k+1} = x_k - \alpha_k \nabla_{(P_k,h_k)} F(x_k, z_k)$$

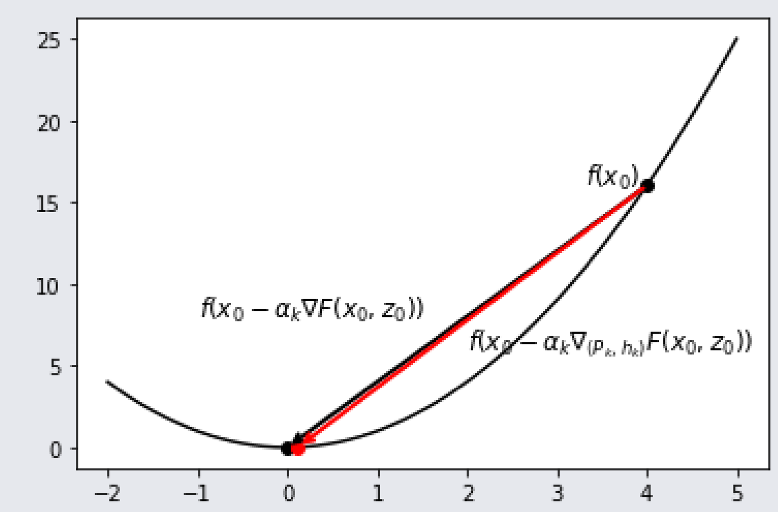
where, for $l \leq d$,

$$\nabla_{(P,h)} F(x,z) = \sum_{i=0}^{l} \frac{F(x + hp^{(i)}) - F(x,z)}{h} p^{(i)}$$

with $P = (p^{(1)}, \cdots, p^{(l)}) \in \mathbb{R}^{d\times l}$ random matrix s.t.

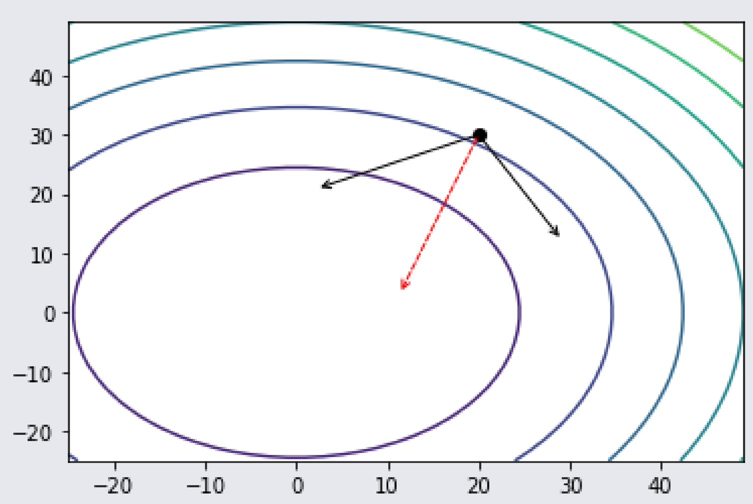
$$\mathbb{E}_P[PP^\intercal] = I \qquad \text{and} \qquad P^\intercal P \overset{\text{a.s.}}{=} \frac{d}{l} I$$

Stochastic Finite-differences



$$\nabla F(x,z) \approx \nabla_{(P,h)} F(x,z)$$

Structured Directions



Orthogonal directions.

Main Results

Main Assumptions

- **Smoothness:** for every z, for every $x_1, x_2 \in \mathbb{R}^d$
$$\|\nabla F(x_1, z) - \nabla F(x_2, z)\|^2 \leq \lambda \|x_1 - x_2\|^2$$
for some $\lambda > 0$.
- **Unbiasedness:** for every $x \in \mathbb{R}^d$
$$\mathbb{E}[\nabla F(x,z)] = \nabla f(x)$$
- **Bounded variance:** there exists $G > 0$ s.t.
$$(\forall x \in \mathbb{R}^d) \quad \mathbb{E}[\|\nabla F(x,z) - \nabla f(x)\|^2] \leq G$$

Convergence rates for convex functions

Let $\alpha_k = \alpha/k^c$ and $h_k = h/k^r$ with $1/2 < c < 1$, $h > 0$ and $\alpha < l/(d\lambda)$. Let $\bar{x}_k$ be the averaged iterate at time $k \in \mathbb{N}$. Then we have that, for every $k \in \mathbb{N}$,

$$\mathbb{E}[f(\bar{x}_k) - f(x^*)] \leq \frac{d}{l}\frac{C'}{k^{1-c}} + o\Big(\frac{1}{k^{1-c}}\Big),$$

where $C' > 0$ is a constant. The number of function evaluations required to obtain an error $\epsilon > 0$, is in

$$\mathcal{O}\bigg(l\Big(\frac{d}{l\epsilon}\Big)^{\frac{1}{1-c}}\bigg).$$

Convergence rates for non-convex functions

Assuming that for some $\gamma > 0$,

$$\forall x \in \mathbb{R}^d, \quad \|\nabla f(x)\|^2 \geq \gamma(f(x) - f^*),$$

let $\alpha_k = \frac{\alpha}{k^c}$ with $1/2 < c \leq 1$ and $h_k = \frac{h}{k^{c/2}}$, with $\alpha < \frac{l}{d\lambda}$ and $h > 0$. Then, there exists a constant $\bar{C} > 0$ s.t.

$$\mathbb{E}[f(x_k) - f(x^*)] \leq \begin{cases} o\big(\frac{1}{k^t}\big), \text{ for every } t < c & \text{if } c < 1 \\ \mathcal{O}\big(\frac{1}{k^\mu}\big) & \text{if } c = 1 \end{cases}$$

with $\mu = \frac{\alpha}{2}\Big(1 - \frac{\lambda\alpha d}{l}\Big)\gamma$. In particular, there is a constant $\tilde{C} > 0$ such that $\mathbb{E}[f_k - f_*] \leq \tilde{C}/k^c$.

Observations

Convex Setting

- The rate approaches the rate of SGD $1/\sqrt{k}$.
- Increasing $l \implies$ better rate.

Non-convex Setting

- **First convergence result** for stochastic zeroth-order methods in this setting.
- Convergence rate close to $1/k$ (SGD rate in strongly convex case).
- Increasing $l \implies$ smaller error in constants.

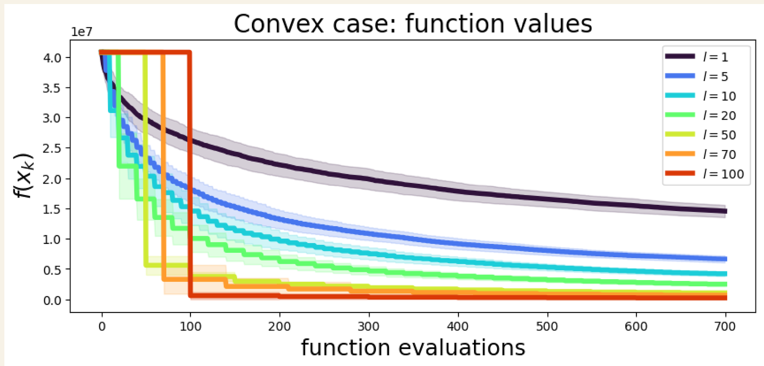
Previous works

Comparison with previous works

- Evolutionary Strategies [1]: few or no theoretical guarantees.
- Bayesian Optimization [6]: cumulative regret scales exponentially with d.
- Direct Search [4]: waste of function evaluations.

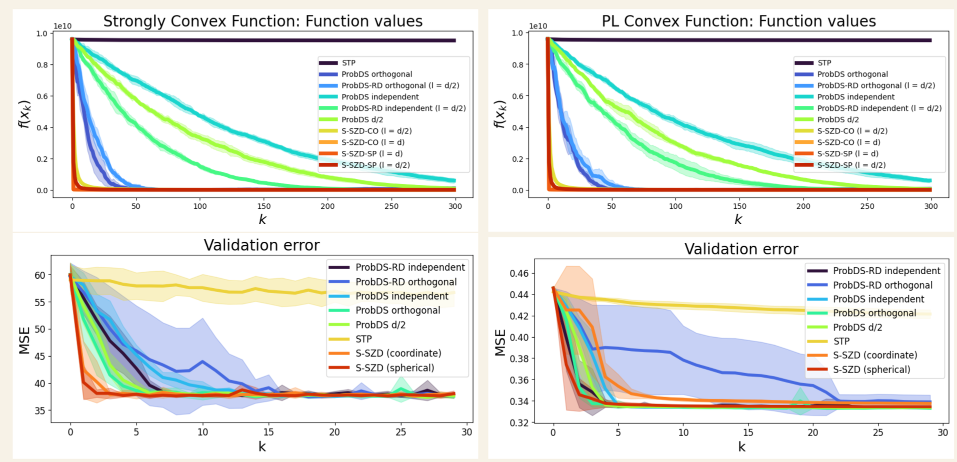
Empirical Results

Impact of number of directions



For large budget, $l > 1$ provides better results w.r.t. $l = 1$.

Comparison with Direct search methods



First row: Our algorithm outperforms state-of-arts direct search (DS) methods in optimizing a strongly convex and a PL convex function.
Second row: Our algorithm achieves lowest validation error faster than DS methods in tuning hyper-parameters of a large-scale kernel methods to solve two regression problems.

Conclusions

- We introduced a new stochastic zeroth order algorithm.
- Convergence rates for convex and non-convex settings.
- Empirical results suggest good performances.

Forthcoming Research

Different research directions

- adaptive strategy for P_k.
- extension for general non-convex functions.

References

[1] Hans-Georg Beyer and Hans-Paul Schwefel. Evolution strategies - a comprehensive introduction. *Natural Computing*, 1:3–52, 03 2002.

[2] HanQin Cai, Daniel McKenzie, Wotao Yin, and Zhenliang Zhang. Zeroth-order regularized optimization (zoro): Approximately sparse gradients and adaptive sampling. *SIAM Journal on Optimization*, 32(2):687–714, 2022.

[3] Krzysztof Choromanski, Mark Rowland, Vikas Sindhwani, Richard Turner, and Adrian Weller. Structured evolution with compact architectures for scalable policy optimization. In *International Conference on Machine Learning*, pages 970–978. PMLR, 2018.

[4] Tamara G. Kolda, Robert Michael Lewis, and Virginia Torczon. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Review*, 45(3):385–482, 2003.

[5] Saeid Qodsi, Reza Tehrani, and Mahdi Bashiri. Portfolio optimization with simulated annealing algorithm. *Financial Research Journal*, 17(1):141–158, 2015.

[6] Marco Rando, Luigi Carratino, Silvia Villa, and Lorenzo Rosasco. Ada-bkb: Scalable gaussian process optimization on continuous domains by adaptive discretization. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7320–7348. PMLR, 28–30 Mar 2022.

[7] Tim Salimans, Jonathan Ho, Xi Chen, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *ArXiv*, abs/1703.03864, 2017.

Acknowledgements

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216, and the Italian Institute of Technology. We gratefully acknowledge the support of NVIDIA Corporation for the donation of the Titan Xp GPUs and the Tesla k40 GPU used for this research. L. R. acknowledges the financial support of the European Research Council (grant SLING 819789), the AFOSR project FA9550-18-1-7009 (European Office of Aerospace Research and Development), the EU H2020-MSCA-RISE project NoMADS - DLV-777826, and the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. S. V. acknowledges the H2020-MSCA-ITN Project Trade-OPT 2019; S. V. Is part of the Indam group “Gruppo Nazionale per l’Analisi Matematica, la Probabilità e le loro applicazioni". C. M. Is part of the Indam group "Gruppo Nazionale per l’Analisi Matematica, la Probabilità e le loro applicazioni".

CONTACTS

Marco Rando
marco.rando@edu.unige.it

Cesare Molinari
molinari@dima.unige.it

Silvia Villa
villa@dima.unige.it

Lorenzo Rosasco
lorenzo.rosasco@unige.it

CSW MaLGa MACHINE LEARNING GENOA CENTER LCSL Dibris DIMA