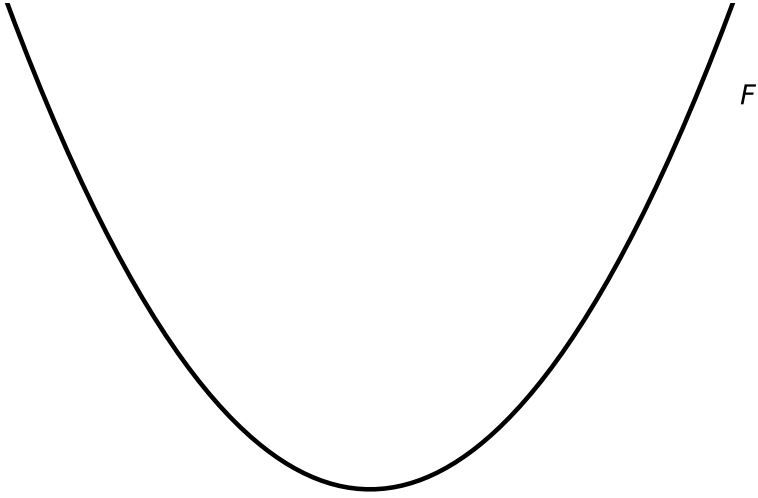
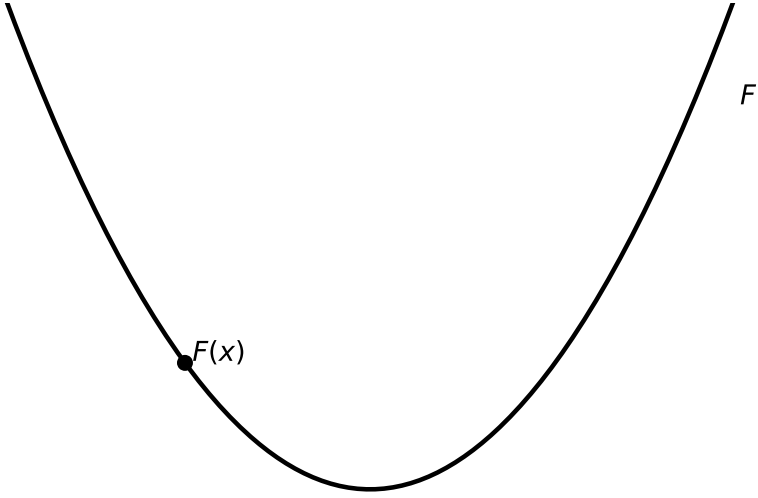
UniGe | MaLGa

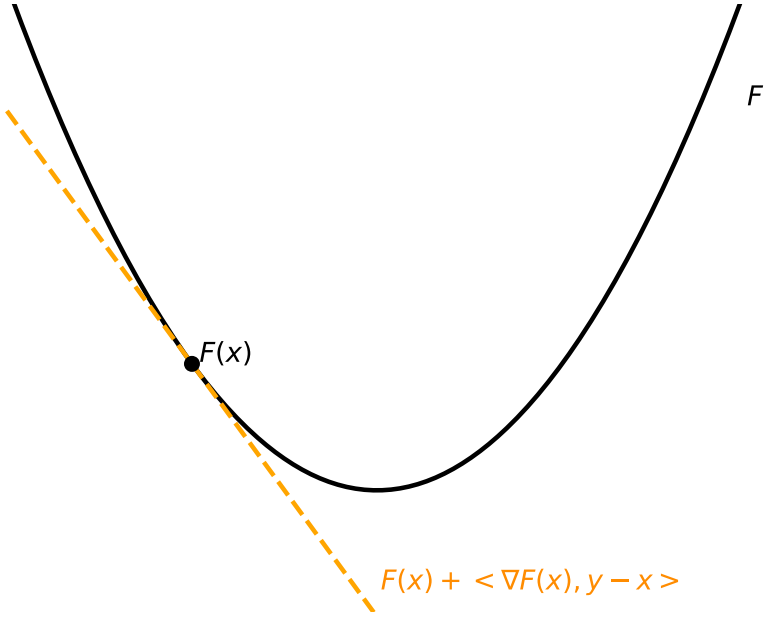
Recent Developments in Zeroth-order Optimization Algorithms
NP-Twins 2024

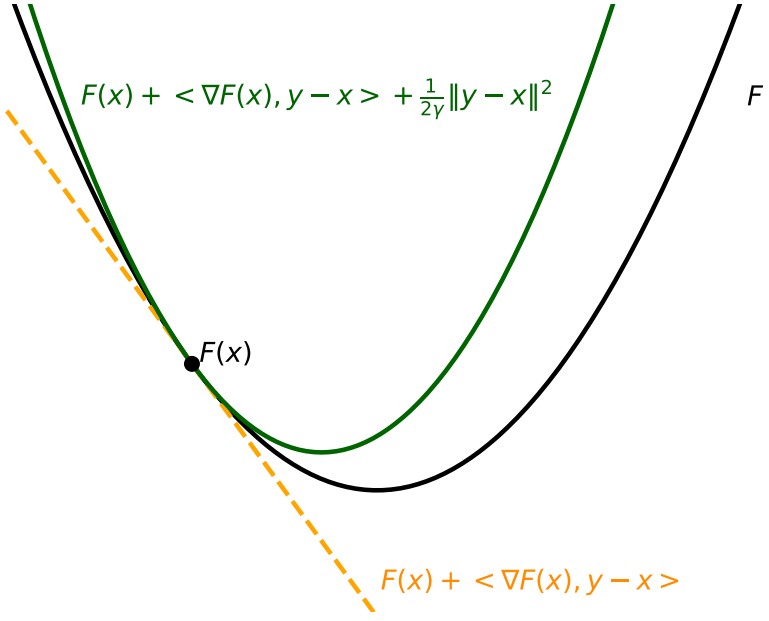
Marco Rando, Cesare Molinari, Lorenzo Rosasco and Silvia Villa

MaLGa - DIBRIS, University of Genoa


F


F
F(x)


F
F(x)
F(x) + < ∇F(x), y − x >


$F(x) + <\nabla F(x), y - x> + \frac{1}{2\gamma}\|y - x\|^2$
F
$F(x)$
$F(x) + <\nabla F(x), y - x>$

Gradient Descent

At every iteration $k \in \mathbb{N}$,

$$
\begin{aligned}
x_{k+1} &= \arg\min_{y} \left\{ F(x_k) + \langle \nabla F(x_k), y - x_k \rangle + \frac{1}{2\gamma} \|y - x_k\|^2 \right\} \\
&= x_k - \gamma \nabla F(x_k).
\end{aligned}
$$

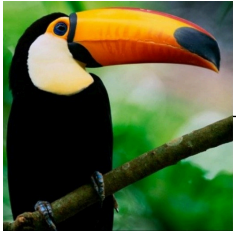
Gradient Descent

At every iteration $k \in \mathbb{N}$,

$$x_{k+1} = x_k - \gamma \nabla F(x_k).$$

BUT ∇F must be available!

A motivating example



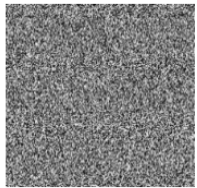
f

toucan (97 %)

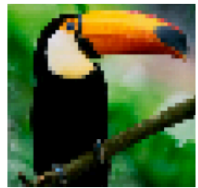
A motivating example

toucan (97%)

x



v s.t. $||v|| \leq \varepsilon$



cat (95%)

$x + v$

The optimization perspective

$$v^* \in \arg\max_{v \in \mathbb{R}^d} F(v) := L(f(x+v), y) \qquad \text{subject to} \qquad \|v\| \leq \varepsilon.$$

Where

- $(x, y) \in X \times Y$ input-output
- ε maximum pertubation threshold
- $f : X \to Y$ classifier
- $L : Y \times Y \to \mathbb{R}$ loss function

The optimization perspective

$$v^* \in \arg\max_{v \in \mathbb{R}^d} F(v) := L(f(x+v), y) \qquad \text{subject to} \qquad \|v\| \leq \varepsilon.$$

- **Black-box optimization**: Gradients not available.

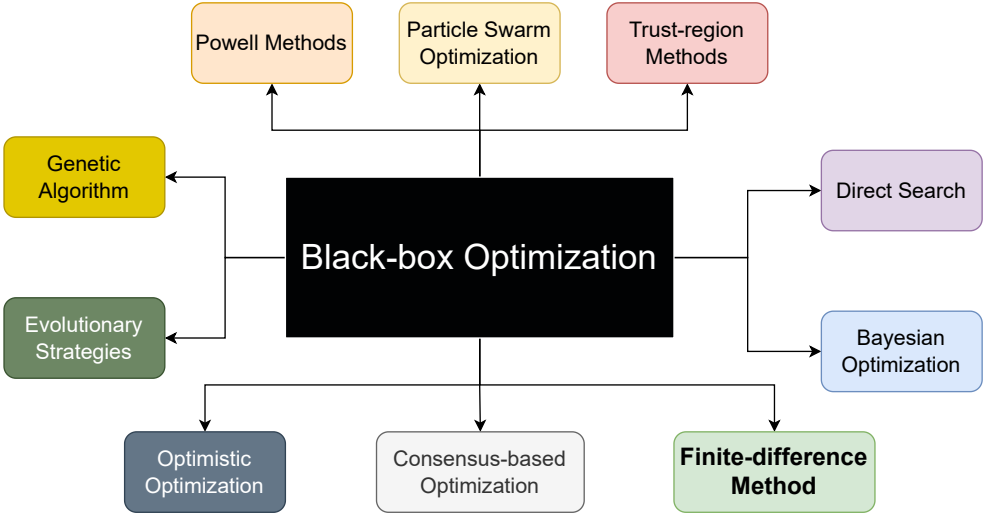
The optimization perspective

$$v^* \in \arg\max_{v \in \mathbb{R}^d} F(v) := L(f(x+v), y) \qquad \text{subject to} \qquad \|v\| \leq \varepsilon.$$

- **Black-box optimization**: Gradients not available.

Other examples

- Robotics
- (Hyper)-parameter tuning
- Reinforcement Learning


Powell Methods
Particle Swarm Optimization
Trust-region Methods
Genetic Algorithm
Black-box Optimization
Direct Search
Evolutionary Strategies
Bayesian Optimization
Optimistic Optimization
Consensus-based Optimization
Finite-difference Method

Finite-Difference Method

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

We can use F

$x_{k+1} = x_k - \gamma_k \nabla$ ~~$F(x_k)$~~

$g_k(x_k) \approx \nabla F(x_k)$

$x_{k+1} = x_k - \gamma_k g_k(x_k)$

FD: Finite-Difference Method

$$\nabla F(x) = \sum_{i=1}^{d} \lim_{h \to 0} \frac{F(x + he_i) - F(x)}{h} e_i \approx \sum_{i=1}^{d} \frac{F(x + he_i) - F(x)}{h} e_i =: g(x)$$

Function evaluation cost

$$x_{k+1} = x_k - \gamma_k \sum_{i=1}^{d} \frac{F(x_k + h_k e_i) - F(x_k)}{h_k} e_i$$

The iteration is expensive: $d + 1$ function evaluations

Random Finite-Difference Method

At every iteration $k \in \mathbb{N}$, sample $(v_k^{(i)})_{i=1}^{\ell}$ with $\ell \leq d$ and compute

$$x_{k+1} = x_k - \gamma_k \sum_{i=1}^{\ell} \frac{F(x_k + h_k v_k^{(i)}) - F(x)}{h_k} v_k^{(i)} \qquad \text{with} \qquad \ell \leq d$$

- random directions $(v_k^{(i)})_{i=1}^{\ell}$

[4] Duchi, J. C., Jordan, M. I., Wainwright, M. J., & Wibisono, A. (2015). Optimal rates for zero-order convex optimization: The power of two function evaluations. IEEE Transactions on Information Theory, 61(5), 2788-2806.
[8] Nesterov, Y., & Spokoiny, V. (2017). Random gradient-free minimization of convex functions. Foundations of Computational Mathematics, 17(2), 527-566.
[5] Ghadimi, S., & Lan, G. (2013). Stochastic first-and zeroth-order methods for nonconvex stochastic programming. SIAM journal on optimization, 23(4), 2341-2368.

Random Finite-Difference Method

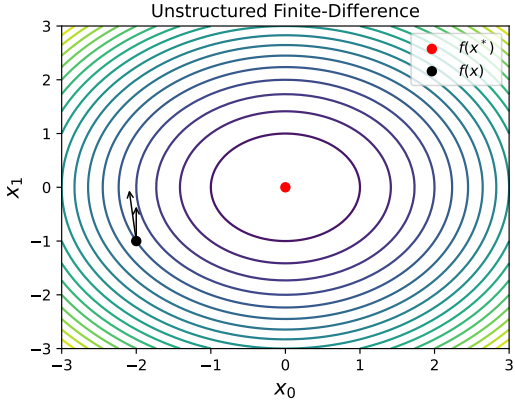
At every iteration $k \in \mathbb{N}$, sample $(v_k^{(i)})_{i=1}^{\ell}$ with $\ell \leq d$ and compute

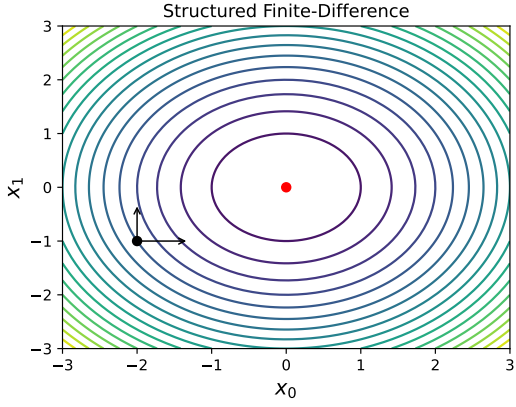
$$x_{k+1} = x_k - \gamma_k \sum_{i=1}^{\ell} \frac{F(x_k + h_k v_k^{(i)}) - F(x)}{h_k} v_k^{(i)} \qquad \text{with} \qquad \ell \leq d$$

- random directions $(v_k^{(i)})_{i=1}^{\ell}$
  - **unstructured directions:** sampled i.i.d. from some distribution
  - **structured directions:** orthogonal, let V be a random matrix with columns $v^{(i)}$

$$\mathbb{E}[VV^\intercal] = I \qquad \text{and} \qquad V^\intercal V \overset{a.s.}{=} \frac{d}{\ell} I$$

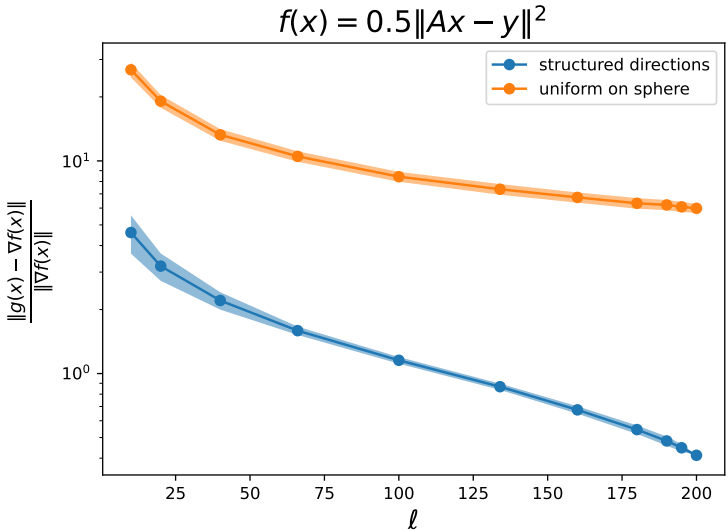
[4] Duchi, J. C., Jordan, M. I., Wainwright, M. J., & Wibisono, A. (2015). Optimal rates for zero-order convex optimization: The power of two function evaluations. IEEE Transactions on Information Theory, 61(5), 2788-2806.
[8] Nesterov, Y., & Spokoiny, V. (2017). Random gradient-free minimization of convex functions. Foundations of Computational Mathematics, 17(2), 527-566.
[5] Ghadimi, S., & Lan, G. (2013). Stochastic first-and zeroth-order methods for nonconvex stochastic programming. SIAM journal on optimization, 23(4), 2341-2368.


Unstructured Finite-Difference
$f(x^*)$
$f(x)$
x_0
x_1


Structured Finite-Difference
x_0
x_1

Gradient Accuracy Comparison

Let $d = 200$ and $A \in \mathbb{R}^{d \times d}$,



Structured and Unstructured Directions

- **Higher approximation accuracy than unstructured methods** [1].
- **Better empirical results** [3].
- **Few theoretical results available** [6, 1, 7, 12]
  - **No non-smooth analysis was provided**.

[1] Berahas, A. S., Cao, L., Choromanski, K., & Scheinberg, K. (2022). A theoretical and empirical comparison of gradient approximations in derivative-free optimization. Foundations of Computational Mathematics, 22(2), 507-560.
[3] K. Choromanski, M. Rowland, V. Sindhwani, R. Turner, and A. Weller. Structured evolution with compact architectures for scalable policy optimization. In Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 970–978. PMLR, 10–15 Jul 2018.
[6] Grapiglia, G. N. (2024). Worst-case evaluation complexity of a derivative-free quadratic regularization method. Optimization Letters, 18(1), 195-213.
[7] Kozak, D., Molinari, C., Rosasco, L., Tenorio, L., & Villa, S. (2023). Zeroth-order optimization with orthogonal random directions. Mathematical Programming, 199(1), 1179-1219.
[12] Wang, T., & Feng, Y. (2024). Convergence Rates of Zeroth Order Gradient Descent for Łojasiewicz Functions. INFORMS Journal on Computing.
[13] Zhang, C., Benz, P., Lin, C., Karjauv, A., Wu, J., & Kweon, I. S. (8 2021). A Survey on Universal Adversarial Attack. In Z.-H. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21 (pp. 4687–4694). doi:10.24963/ijcai.2021/635.
[14] Zhang, S., Shen, T., Sun, H., Dong, Y., Xie, D., & Zhang, H. (2022). Zeroth-order stochastic coordinate methods for decentralized non-convex optimization. arXiv preprint arXiv:2204.04743.

Non-smooth Optimization

$$x^* \in \arg\min_{x \in \mathbb{R}^d} F(x)$$

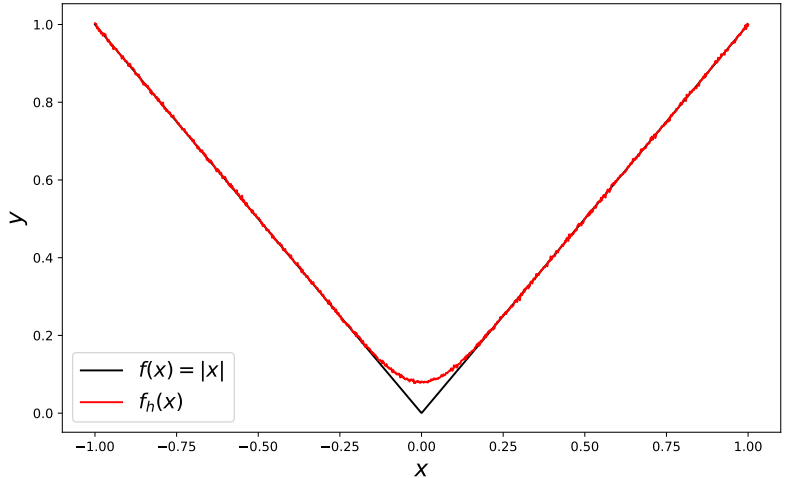
- **BUT:** F non-smooth

Smoothing

$$F_h(x) := \mathbb{E}_{u \sim \mathbb{B}^d}[F(x + hu)]$$

Smoothing

$$F_h(x) := \mathbb{E}_{u \sim \mathbb{B}^d}[F(x + hu)]$$



Smoothing

$$F_h(x) := \mathbb{E}_{u \sim \mathbb{B}^d}[F(x + hu)]$$

- F_h is differentiable [2].
- F convex $\implies$ F_h convex and $F(x) \leq F_h(x)$
- F L-Lipschitz $\implies$ F_h L-Lipschitz and $F_h(x) \leq F(x) + Lh$
- F L-**Lipschitz** $\implies$ F_h $L\sqrt{d}/h$-**Smooth!**

[2] D. P. Bertsekas. Stochastic optimization problems with nondifferentiable cost functionals. Journal of Optimization Theory and Applications, 12(2):218–231, Aug 1973.

J. C. Duchi, P. L. Bartlett, L. Peter, and M. J. Wainwright. Randomized smoothing for stochastic optimization. SIAM Journal on Optimization, 22(2):674–701, 2012

X. Gao, B. Jiang, and S. Zhang. On the information-adaptive variants of the admm: An iteration complexity perspective. Journal of Scientific Computing, 76(1):327–363, Jul 2018

T. Lin, Z. Zheng, and M. Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems, volume 35, pages 26160–26175. Curran Associates, Inc., 2022.

F. Yousefian, A. Nedi ´c, and U. V. Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. Automatica, 48(1):56–67, 2012.

Smoothing Lemma for Structured Approximation

Structured Central Finite-difference
Sample G uniformly from $O(d)$,

$$g(x) := \frac{d}{\ell} \sum_{i=1}^{\ell} \frac{F(x + hGe_i) - F(x - hGe_i)}{2h} Ge_i.$$

Smoothing Lemma [10]
Then, let $F_h(x) := \mathbb{E}_{u \sim \mathbb{B}^d}[F(x + hu)]$,

$$\mathbb{E}_G[g(x)] = \nabla F_h(x).$$

[10] Rando, M., Molinari, C., Rosasco, L., & Villa, S. (2023). An Optimal Structured Zeroth-order Algorithm for Non-smooth Optimization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), Advances in Neural Information Processing Systems (Vol. 36, pp. 36738–36767).

Algorithm

At every iteration $k \in \mathbb{N}$, sample G_k uniformly from $O(d)$ and compute

$$x_{k+1} = x_k - \gamma_k \underbrace{\frac{d}{\ell} \sum_{i=1}^{\ell} \frac{F(x_k + h_k G_k e_i) - F(x_k - h_k G_k e_i)}{2h_k} G_k e_i}_{=: g_k(x_k)}$$

Convergence Rate in Convex Non-smooth setting

Theorem 1 [10] If F is L-Lipschitz continuous, choosing $\gamma_k = \gamma/\sqrt{(k+1)}$ and $h_k = h/\sqrt{(k+1)}$

$$\mathbb{E}[F(\bar{x}_k) - F(x^*)] \leq \frac{C}{\gamma\sqrt{k}} + o\left(\frac{1}{\sqrt{k}}\right)$$

with $C > 0$.

Moreover, the number of function evaluations required to obtain an error $\varepsilon \in (0, 1)$ is $\mathcal{O}(d\epsilon^{-2})$

[10] Rando, M., Molinari, C., Rosasco, L., & Villa, S. (2024). An optimal structured zeroth-order algorithm for non-smooth optimization. Advances in Neural Information Processing Systems, 36.

Convergence Rate in Non-Convex Non-smooth setting

Theorem 2 [10] If F is L-Lipschitz continuous, choosing $\gamma_k = \gamma/\sqrt{(k+1)}$ and $h_k = h$

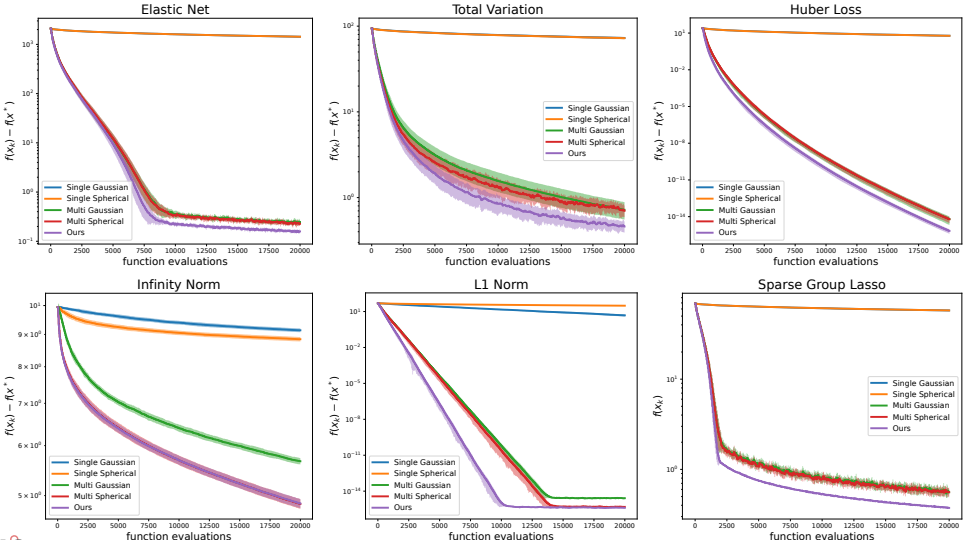
$$\frac{\sum_{i=0}^{k} \gamma_i \mathbb{E}[\|\nabla F_h(x_i)\|^2]}{\sum_{i=0}^{k} \gamma_i} \leq C \frac{F_h(x_0) - \min F}{\gamma\sqrt{k}} + o\left(\frac{1}{\sqrt{k}}\right)$$

with $C > 0$.

Moreover, the number of function evaluations required to obtain an error $\varepsilon \in (0, 1)$ is $\mathcal{O}(d\sqrt{d}h^{-1}\epsilon^{-2})$

[10] Rando, M., Molinari, C., Rosasco, L., & Villa, S. (2024). An optimal structured zeroth-order algorithm for non-smooth optimization. Advances in Neural Information Processing Systems, 36.

Numerical Experiments



Conclusions

Summing up

- A structured finite-difference method for non-smooth optimization.
  - Theory + experiments

Conclusions

Summing up

- A structured finite-difference method for non-smooth optimization.
  - Theory + experiments

Results

- Smoothing Lemma
- Convergence rates
- Convergence of the iterates
- Empirical Experiments

Research Directions

- Extensions
  - Variance Reduction (coming soon on arXiv.)
- Non-convex setting
- Analysis in stochastic non-smooth setting
  - Stochastic smooth setting [11]
- Applications [9, 3]

[9] Rando, M., Demetrio, L., Rosasco, L., & Roli, F. (2024). A New Formulation for Zeroth-Order Optimization of Adversarial EXEmples in Malware Detection. Submitted to TIFS.
[11] Rando, M., Molinari, C., Villa, S., & Rosasco, L. (2024). Stochastic zeroth order descent with structured directions. Computational Optimization and Applications, 1-37.
[3] Choromanski, K., Rowland, M., Sindhwani, V., Turner, R., & Weller, A. (2018, July). Structured evolution with compact architectures for scalable policy optimization. In International Conference on Machine Learning (pp. 970-978). PMLR.

Thank you for your Attention! :)

References I

Albert S. Berahas, Liyuan Cao, Krzysztof Choromanski, and Katya Scheinberg.
A theoretical and empirical comparison of gradient approximations in derivative-free optimization.
Foundations of Computational Mathematics, 22(2):507–560, Apr 2022.

Dimitri P Bertsekas.
Stochastic optimization problems with nondifferentiable cost functionals.
Journal of Optimization Theory and Applications, 12(2):218–231, 1973.

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.

References II

John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono.
Optimal rates for zero-order convex optimization: The power of two function evaluations.
IEEE Transactions on Information Theory, 61(5):2788–2806, 2015.

Saeed Ghadimi and Guanghui Lan.
Stochastic first-and zeroth-order methods for nonconvex stochastic programming.
SIAM journal on optimization, 23(4):2341–2368, 2013.

Geovani Nunes Grapiglia.
Worst-case evaluation complexity of a derivative-free quadratic regularization method.
Optimization Letters, 18(1):195–213, 2024.

References III

- David Kozak, Cesare Molinari, Lorenzo Rosasco, Luis Tenorio, and Silvia Villa.
  Zeroth-order optimization with orthogonal random directions.
  Mathematical Programming, 199(1):1179–1219, 2023.
- Yurii Nesterov and Vladimir Spokoiny.
  Random gradient-free minimization of convex functions.
  Foundations of Computational Mathematics, 17(2):527–566, 2017.
- Marco Rando, Luca Demetrio, Lorenzo Rosasco, and Fabio Roli.
  A new formulation for zeroth-order optimization of adversarial exemples in malware detection.
  arXiv preprint arXiv:2405.14519, 2024.

References IV

Marco Rando, Cesare Molinari, Lorenzo Rosasco, and Silvia Villa.
An optimal structured zeroth-order algorithm for non-smooth optimization.
In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 36738–36767. Curran Associates, Inc., 2023.

Marco Rando, Cesare Molinari, Silvia Villa, and Lorenzo Rosasco.
Stochastic zeroth order descent with structured directions.
Computational Optimization and Applications, pages 1–37, 2024.

Tianyu Wang and Yasong Feng.
Convergence rates of zeroth order gradient descent for łojasiewicz functions.
INFORMS Journal on Computing, 2024.

References V

Chaoning Zhang, Philipp Benz, Chenguo Lin, Adil Karjauv, Jing Wu, and In So Kweon.
A survey on universal adversarial attack.
In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4687–4694. International Joint Conferences on Artificial Intelligence Organization, 8 2021.
Survey Track.

Shengjun Zhang, Tan Shen, Hongwei Sun, Yunlong Dong, Dong Xie, and Heng Zhang.
Zeroth-order stochastic coordinate methods for decentralized non-convex optimization.
arXiv preprint arXiv:2204.04743, 2022.