UniGe | MaLGa

An Optimal Structured Zeroth-order Method for Non-smooth optimization

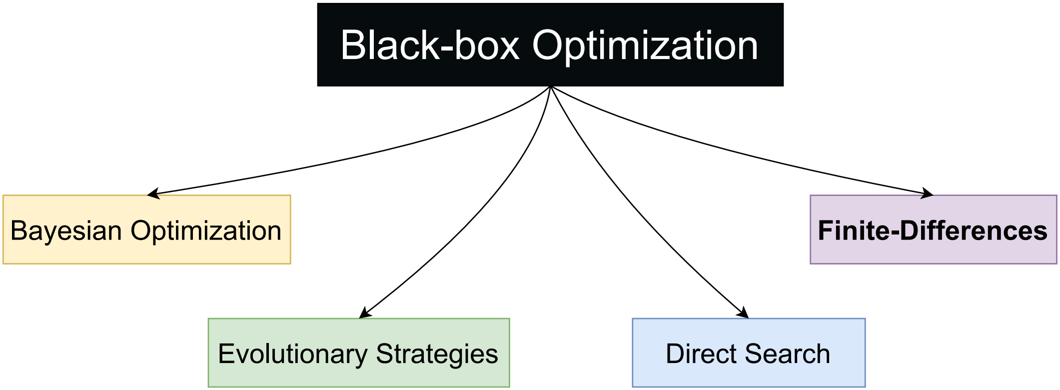
Marco Rando , Cesare Molinari, Lorenzo Rosasco, Silvia Villa
MaLGa- Machine learning Genova Center
Università di Genova

Black-box optimization problem

- **No explicit formulation of f.**
- **Gradient is not available.**
- **(perturbed) function values are (generally) available.**

GOAL

$$x^* \in \arg\min_{x \in X} f(x)$$


Black-box Optimization
Bayesian Optimization
Evolutionary Strategies
Direct Search
Finite-Differences

Finite-difference methods

Gradient Descent

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

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

Zeroth-order "Descent"

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

UniGe | MaLGa

Gradient Surrogate

$$g(x) := \frac{d}{\ell} \sum_{i=1}^{\ell} \frac{f(x + hv^{(i)}) - f(x - hv^{(i)})}{2h} v^{(i)}.$$

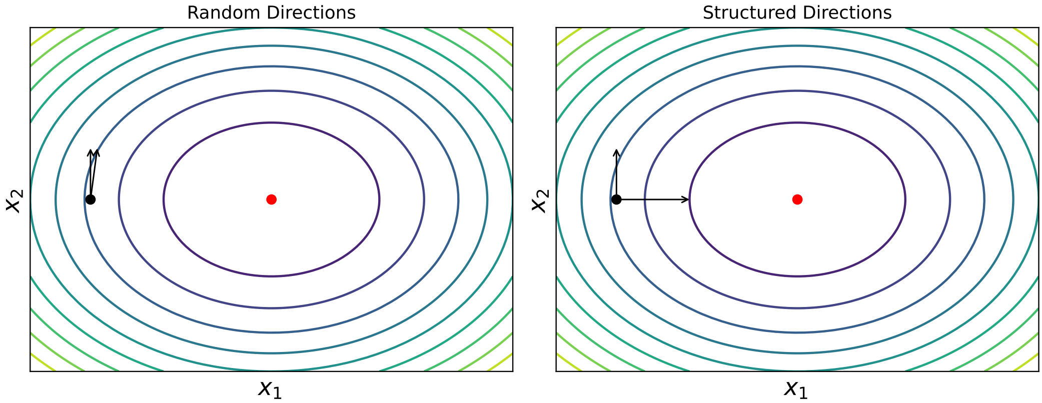
- **Directions**.
- **Number of directions**.
- **Discretization parameter**.

UniGe | MaLGa

$$g(x) := \frac{d}{l} \sum_{i=1}^{l} \frac{f\left(x + hv^{(i)}\right) - f\left(x - hv^{(i)}\right)}{2h} v^{(i)}$$

Random Directions

Structured Directions



UniGe | MaLGa

Random vs Structured approximations

Random Directions

- Simple.
- Higher number of directions than structured methods to achieve similar gradient accuracy Berahas et al. (2022).
- Many applications e.g, Cai et al. (2021); Salimans et al. (2017); Mania et al. (2018)

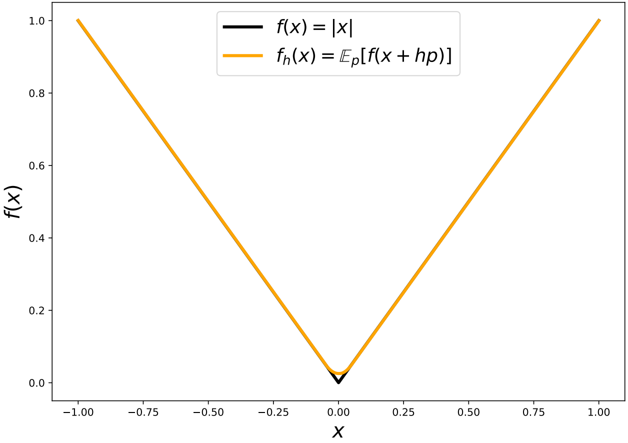
Structured Directions

- Better exploration than random methods.
- Analysis is very limited e.g, **no non-smooth analysis**.
- Actually, few applications e.g, Choromanski et al. (2018).

Goal: non-smooth analysis for structured finite-difference method.

Non-smooth Setting

$$f_h(x) := \mathbb{E}_{u \in \mathbb{B}^d}[f(x + hu)]$$



Smoothing

Let

$$f_h(x) := \mathbb{E}_u[f(x+hu)]$$

- f_h is differentiable (Bertsekas, 1973).
- if f is L-Lipschitz continuous, f_h is smooth!
- if f convex and L-Lipschitz,

$$(\forall x \in \mathbb{R}^d) \qquad f(x) \leq f_h(x) \leq f(x) + Lh$$

- if f convex and L-smooth,

$$(\forall x \in \mathbb{R}^d) \qquad f(x) \leq f_h(x) \leq f(x) + \frac{Lh^2}{2}$$

Smoothing Lemma for Structured Surrogates

Define $f_h(x) := \mathbb{E}_{u \in \mathbb{B}^d}[f(x + hu)]$. Then, for every $G \in O(d)$, define

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

Then,

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

Algorithm

For $k = 1, \cdots,$

sample G_k from $O(d)$

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

UniGe | MaLGa

Main Results

In convex Lipschitz non-smooth setting

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

Complexity in function evaluations is $\mathcal{O}(d\varepsilon^{-2})$

Main Results

In non-convex non-smooth Lipschitz setting

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

Complexity in function evaluations is $\mathcal{O}(d\sqrt{d}h^{-1}\varepsilon^{-2})$

UniGe | MaLGa

Main Results

Convex Setting

$$\mathbb{E}[f(\bar{x}_k) - f(x^*)] \leq \frac{d}{\ell}\frac{C}{k}.$$

Complexity in function evaluations is $\mathcal{O}(d\varepsilon^{-1})$.

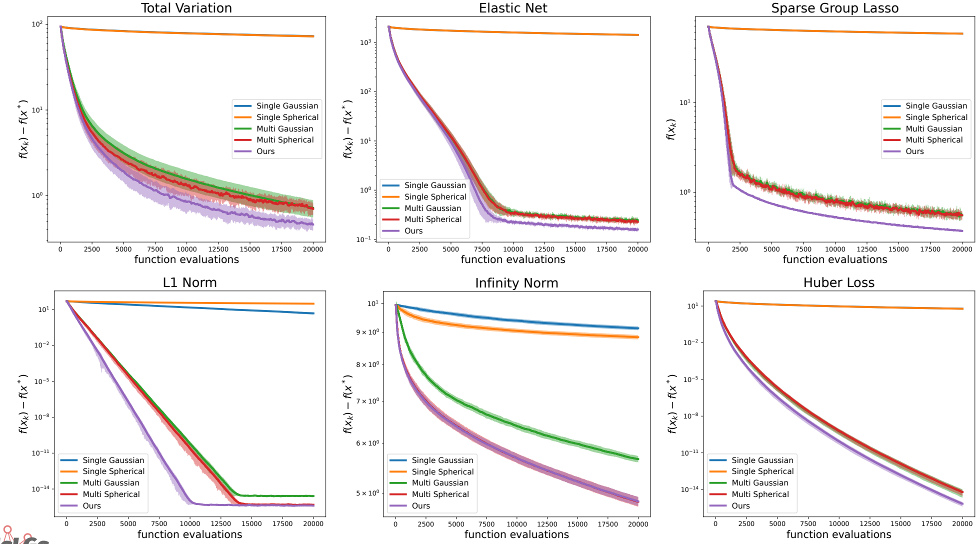
Non-convex setting

let $\Delta := \left(\frac{1}{2} - \frac{L_1 d}{\ell}\bar{\alpha}\right)$ with $\alpha_k \leq \bar{\alpha} < \ell/(2dL)$

$$\frac{\sum_{i=0}^{k}(\gamma_i \mathbb{E}[\|\nabla f(x_i)\|^2])}{\left(\sum_{i=0}^{k}\gamma_i\right)} \leq \left[\frac{f(x_0) - \min f}{\Delta\alpha} + \frac{C_1 d^2 h^2}{\Delta} + \frac{C_2 \alpha h^2 d^2}{\Delta \ell}\right] \cdot \frac{1}{k}$$

Complexity in function evaluations is $\mathcal{O}(d\varepsilon^{-1})$ with $h = \mathcal{O}(1/d)$.

Numerical Experiments



Conclusions

- Smoothing Lemma for structured surrogates.
- Analysis in non-smooth convex setting.
- Analysis in non-smooth non-convex setting.
- Analysis in smooth setting.
- Numerical experiments.

Thank you for your Attention! :)

References I

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

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

Cai, H., Lou, Y., McKenzie, D., and Yin, W. (2021). A zeroth-order block coordinate descent algorithm for huge-scale black-box optimization. In *International Conference on Machine Learning*, pages 1193–1203. PMLR.

Choromanski, K., Rowland, M., Sindhwani, V., Turner, R., and Weller, A. (2018). Structured evolution with compact architectures for scalable policy optimization. In *International Conference on Machine Learning*, pages 970–978. PMLR.

Crama, Y. and Schyns, M. (2003). Simulated annealing for complex portfolio selection problems. *European Journal of operational research*, 150(3):546–571.

References II

Duchi, J. C., Jordan, M. I., Wainwright, M. J., and Wibisono, A. (2014). Optimal rates for zero-order convex optimization: the power of two function evaluations.

Kozak, D., Molinari, C., Rosasco, L., Tenorio, L., and Villa, S. (2023). Zeroth-order optimization with orthogonal random directions. *Mathematical Programming*, 199(1):1179–1219.

Mania, H., Guy, A., and Recht, B. (2018). Simple random search of static linear policies is competitive for reinforcement learning. *Advances in Neural Information Processing Systems*, 31.

Mester, D. and Bräysy, O. (2007). Active-guided evolution strategies for large-scale capacitated vehicle routing problems. *Computers & operations research*, 34(10):2964–2975.

Pannetier, J., Bassas-Alsina, J., Rodriguez-Carvajal, J., and Caignaert, V. (1990). Prediction of crystal structures from crystal chemistry rules by simulated annealing. *Nature*, 346(6282):343–345.

References III

Salimans, T., Ho, J., Chen, X., Sidor, S., and Sutskever, I. (2017). Evolution strategies as a scalable alternative to reinforcement learning.

Van Batenburg, F., Gultyaev, A. P., and Pleij, C. W. (1995). An apl-programmed genetic algorithm for the prediction of rna secondary structure. *Journal of theoretical Biology*, 174(3):269–280.

White, C., Neiswanger, W., and Savani, Y. (2021). Bananas: Bayesian optimization with neural architectures for neural architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10293–10301.

Zalkind, D. S., Dall'Anese, E., and Pao, L. Y. (2020). Automatic controller tuning using a zeroth-order optimization algorithm. *Wind Energy Science*, 5(4):1579–1600.

Approximating the gradient

$$\nabla f(x) = \sum_{i=1}^{d} \lim_{h \to 0} \frac{f(x + he_i) - f(x)}{h} e_i.$$

Problem: we cannot compute the $\lim$.

Approximating the gradient

Fix an $h > 0$,

$$\nabla f(x) \approx \sum_{i=1}^{d} \frac{f(x + he_i) - f(x)}{h} e_i.$$

Problem: it can be expensive to evaluate.

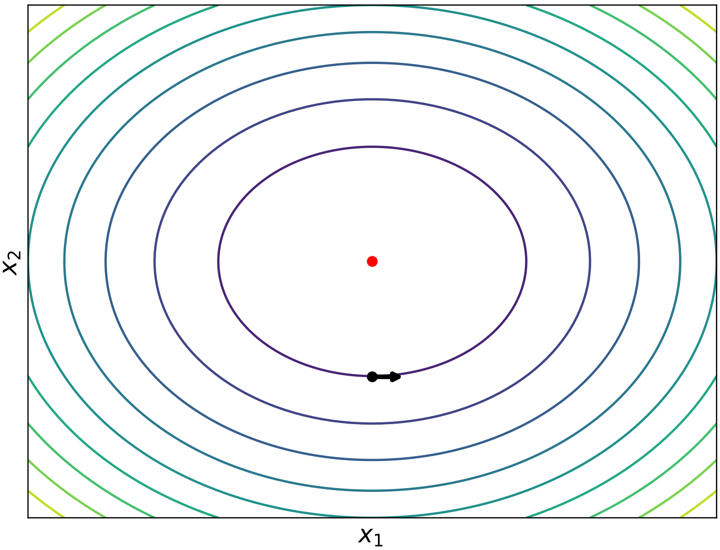
Approximating the gradient

Fix an $h > 0$ and $0 < \ell \leq d$,

$$\nabla f(x) \approx \sum_{i=1}^{\ell} \frac{f(x + he_i) - f(x)}{h} e_i.$$

Problem: some directions will be never explored.

Approximating the gradient



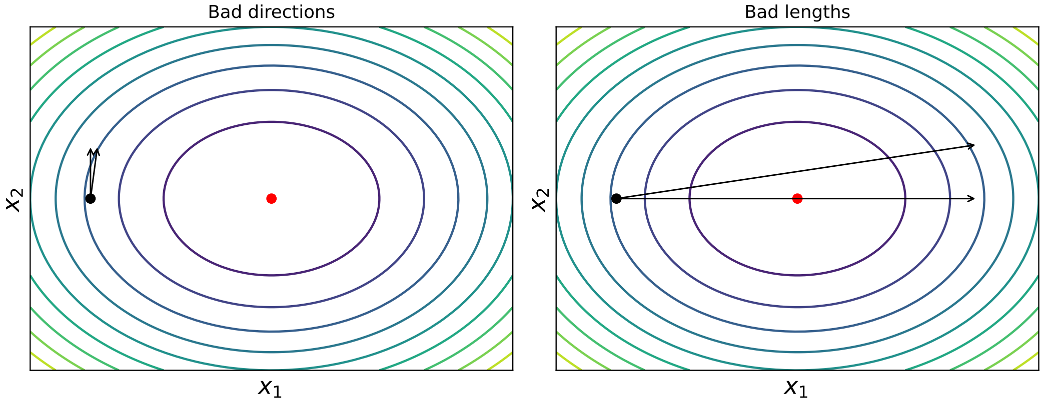
Approximating the gradient

Fix an $h > 0$, $0 < \ell \leq d$ and let $(p^{(i)})_{i=1}^{\ell}$ be random directions,

$$\nabla f(x) \approx \sum_{i=1}^{\ell} \frac{f(x + hp^{(i)}) - f(x)}{h} p^{(i)}.$$

Problem: no control on the directions.

Approximating the gradient

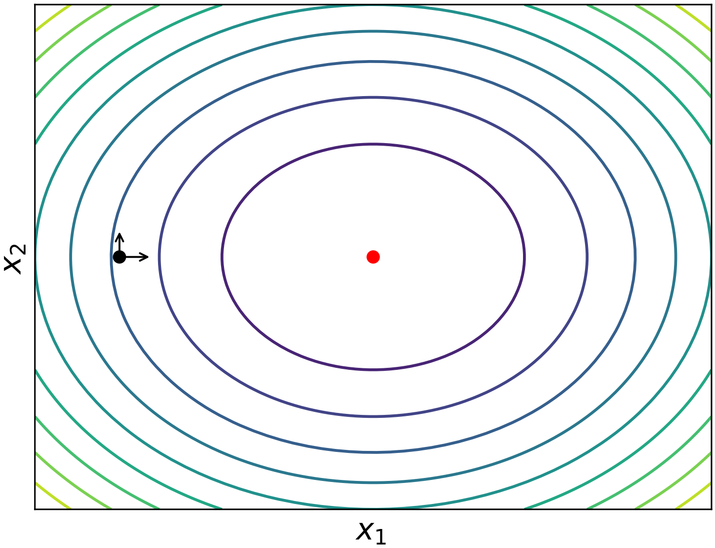


Approximating the gradient

Fix an $h > 0$, $0 < \ell \leq d$ and let $(p^{(i)})_{i=1}^{\ell}$ be random orthogonal directions,

$$\nabla f(x) \approx \sum_{i=1}^{\ell} \frac{f(x + hp^{(i)}) - f(x)}{h} p^{(i)} =: g(x).$$

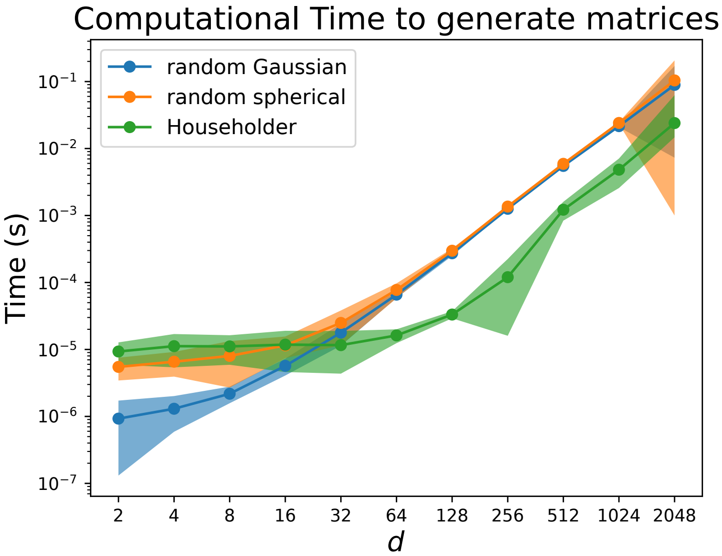
Approximating the gradient



Approximating the gradient

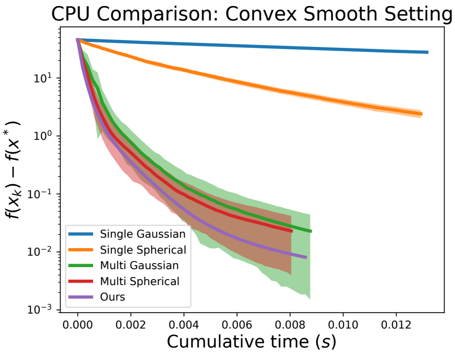
$$x_{k+1} = x_k - \gamma_k \sum_{i=1}^{\ell} \frac{f(x + h_k p_k^{(i)}) - f(x)}{h_k} p_k^{(i)}$$

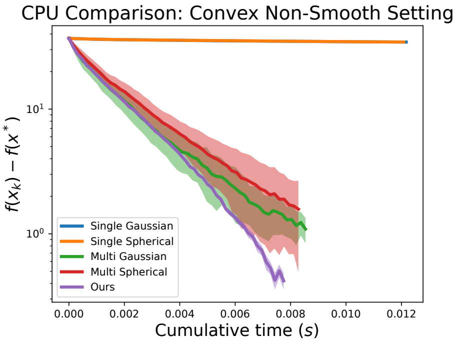
Time-cost comparison



UniGe | MaLGa

Computational Cost of Orthogonal matrices





UniGe | MaLGa