Relaxed GKBO method for global optimization of non-convex high dimensional functions

Giacomo Albi [1] Federica Ferrarese [1,2] Claudia Totzeck [3]

[1] University of Verona, Italy

[2] University of Trento, Italy

[3] Bergische Universität Wuppertal, Germany

Final workshop – PRIN 2017

Catania, February 21st 2023

Numerical optimization methods[1]

- Gradient-based methods[2]
  - Newton,
  - Quasi-Newton,
  - Stochastic gradient descent.
- Gradient-free methods[3,4,5]
  - Simulated annealing,
  - Particle swarm optimization,
  - Genetic algorithm.

Applications

- Training of ANNs, improve Machine Learning algorithms.
- Efficient solution of large-scale optimization problems.

---

[1] C. Totzeck. Trends in Consensus-Based Optimization, 2021

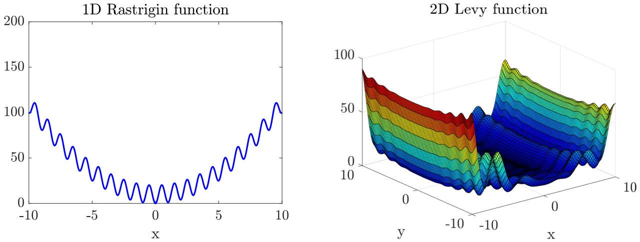
[2] L. Bottou, F. E. Curtis, J. Nocedal. Optimization Methods for Large-Scale Machine Learning, 2018.

[3] M. Fornasier, H. Huang, L. Pareschi, P. Sünnen. Consensus-based optimization on hypersurfaces: well-posedness and mean-field limit, 2020

[4] S. Grassi, L. Pareschi. From particle swarm optimization to consensus based optimization, 2020

[5] D. Kalise,A. Sharma, M. V. Tretyakov. Consensus based optimization via jump-diffusion stochastic differential equations,2022

Aim: study of efficient numerical methods for global optimization of non-convex high dimensional functions.



Idea: combine Consensus based optimization method[6,7] following a Kinetic approach[8] and Continuos Genetic algorithm[9]

[6] R. Pinnau, C. Totzeck, O. Tse, S. Martin. A consensus-based model for global optimization and its mean-field limit, 2016

[7] J. A. Carrillo, Y-P Choi, C. Totzeck, O. Tse An analytical framework for a consensus-based global optimization method, 2018

[8] A. Benfenati, G. Borghi, L. Pareschi. Binary interaction methods for high dimensional global optimization and machine learning, 2022

[9] C. F. M. Toledo, L. Oliveira, P. M. França. Global optimization using a genetic algorithm with hierarchically structured population, 2014

- Kinetic based optimization (KBO) methods are inspired from the study of opinion dynamics and consensus formation.
- Each agent in position x is subjected to a drift and a random perturbation. Its post interaction position is

$$x' = x + \nu_F(\hat{x}(t) - x) + \sigma_F D(x)\xi, \tag{1}$$

where

- σ_F, ν_F are positive parameters and ξ a normally distributed random number,
- $D(x)$ is the diffusion matrix defined to be either

$$\begin{aligned} D(x) &= |\hat{x}(t) - x| Id_d, && \text{isotropic} \\ D(x) &= diag\{(\hat{x}(t) - x)_1, \dots (\hat{x}(t) - x)_d\}, && \text{anisotropic} \end{aligned} \tag{2}$$

- the term $\hat{x}(t)$ represent the estimate of the position of the global minimizer and it is computed as

$$\hat{x}(t) = \frac{\int_{\mathbb{R}^d} x e^{-\alpha \mathcal{E}(x)} g(x,t) dx}{\int_{\mathbb{R}^d} e^{-\alpha \mathcal{E}(x)} g(x,t) dx}, \tag{3}$$

for any probability g, where $\mathcal{E}$ is the cost function, according to Laplace's principle.

- The Continuos Genetic Algorithm (GA) is biologically inspired and its idea arises from natural selection process that mimics biological evolution.
- It selects individuals from the current population and uses them as parents to produce the children for the next generation:
  - Parents (leaders) are chosen to be the ones with best position on the cost function and do not modify their position.
  - Children (followers) are subjected to crossover

$$\begin{aligned} &\text{with rate} \quad \nu_F \quad \Rightarrow \quad x' = x_*, \\ &\text{with rate} \quad 1-\nu_F \quad \Rightarrow \quad x' = x, \end{aligned} \tag{4}$$

and mutations of the type

$$\begin{aligned} &x' = x + \sigma_F \xi, && \text{standard GA} \\ &x' = x + \sigma_F D(x) \xi, && \text{modified GA} \end{aligned} \tag{5}$$

where x, x_* denotes the pre-interaction positions of a follower and a leader respectively, x' denotes the post-interaction position of a follower, ν_F, σ_F are positive parameters, ξ is a normal distributed random number and $D(x)$ a diffusion matrix.

Relaxed Genetic kinetic based algorithm

- The relaxed GKBO method[10]
- To each agent with associate a position x and a label $\lambda \in \{0, 1\}$:
  - if $\lambda = 0$ then the agent is in the followers status;
  - if $\lambda = 1$ then the agent is in the leaders status.
- Each pair of agents (x, λ), (x_*, λ_*) interact toward the following binary interactions rules

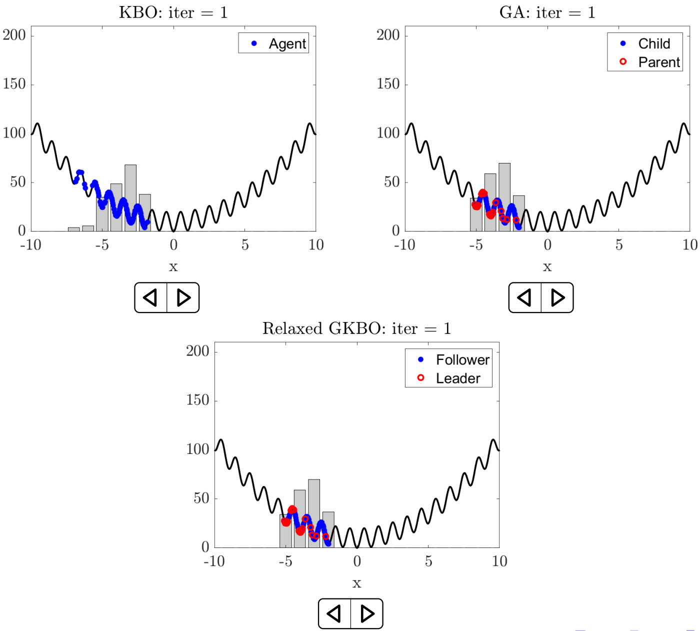
$$\begin{cases} x' = x + (\nu_F(x_* - x) + \sigma_F D(x)\xi)(1-\lambda)\lambda_* + \frac{\nu_L}{\beta}(\hat{x}(t) - x)\lambda, \\ x'_* = x_*, \end{cases} \tag{6}$$

  with
  - ν_L, β, σ_F, ν_F are positive parameters,
  - $D(x)$ is the diffusion matrix defined as in (2),
  - ξ a normally distributed random number,
  - the term $\hat{x}(t)$ represent the estimate of the position of the global minimizer.

  combines the ideas of the KBO and the standard genetic algorithm.

[10] G. Albi, F.F. and C. Totzeck. Relaxed Genetic Kinetic based optimization methods, in preparation.

Rastrigin function: videos



- The evolution of the density function

$$f = f(x, \lambda, t), \qquad f : \mathbb{R}^d \times \{0, 1\} \times \mathbb{R}_+ \to \mathbb{R}_+ \tag{7}$$

is described by the integro-differential equation of Boltzmann type that in weak form reads

$$\frac{\partial}{\partial t} \int_{\mathbb{R}^d} f_\lambda(x, t)\phi(x)dx - \int_{\mathbb{R}^d} \mathcal{T}[f_\lambda](x, t)\phi(x)dx = \eta \sum_{\lambda_* \in \{0,1\}} \left\langle \int_{\mathbb{R}^{2d}} \Big[\phi(x') - \phi(x)\Big] f_\lambda(x, t) f_{\lambda_*}(x_*, t)dxdy \right\rangle, \tag{8}$$

for any test function $\phi \in \mathcal{C}^\infty(\mathbb{R}^d)$, $\eta > 0$ where for simplicity we write $f_\lambda(x, t) = f(x, \lambda, t)$, and where $\mathcal{T}[f_\lambda](x, t)$ describes the leaders-followers change of label.

- Introduce the scaling parameter $\varepsilon > 0$ and consider the scaling

$$\nu_F \to \frac{\nu_F}{\rho_1}\varepsilon, \qquad \nu_L \to \frac{\nu_L}{\rho_1}\varepsilon, \qquad \sigma_F \to \frac{\sigma_F}{\rho_1}\sqrt{\varepsilon}, \qquad \eta \to \frac{1}{\varepsilon}. \tag{9}$$

  where ρ_1 denotes the leaders mass.
- Consider a Taylor expansion of the test function $\phi(x')$ centred in x.
- Plug it in to equation (8) and integrate by parts to get the equations that in strong form read

$$\begin{aligned}\frac{\partial}{\partial t} f_0(x,t) - \mathcal{T}[f_0](x,t) &= \frac{\sigma_F^2}{2}\Delta_x\left[D^2(x)f_0(x,t)\right] + \nu_F \nabla_x \cdot \left[\left(\frac{m_1(t)}{\rho_1} - x\right) f_0(x,t)\right],\\ \frac{\partial}{\partial t} f_1(x,t) - \mathcal{T}[f_1](x,t) &= \frac{\nu_L}{\beta\rho_1}\nabla_x \cdot \left[\left(\hat{x}(t) - x\right) f_1(x,t)\right],\end{aligned} \tag{10}$$

  where $m_1(t)$ denotes the leaders mean at time t.
- Rigorous derivation of the grazing collision limit is done in [11].

[11] L. Pareschi and G. Toscani. Interacting multiagent systems: kinetic equations and Monte Carlo methods, 2013.

The transition operators acts as follows

$$\begin{aligned}\mathcal{T}[f](x,0,t) =& \pi_{L\to F}(x,\lambda;f)f(x,1,t) - \pi_{F\to L}(x,\lambda;f)f(x,0,t),\\ \mathcal{T}[f](x,1,t) =& \pi_{F\to L}(x,\lambda;f)f(x,0,t) - \pi_{L\to F}(x,\lambda;f)f(x,1,t),\end{aligned} \tag{11}$$

where $\pi_{F\to L}(\cdot)$ and $\pi_{L\to F}(\cdot)$ are certain transition rates. Leaders can emerge:

- randomly with constant rates

$$\pi_{L\to F} = q_{LF}, \qquad \pi_{F\to L} = q_{FL},$$

  with $q_{LF}, q_{FL} > 0$;
- according to a weighted strategy: associate to each agent a weight $\omega(x,t)$ dependent on their position on the cost function, then

$$\pi_{L\to F} = \begin{cases} 1, & \text{if } \omega(x,t) > \bar{\omega},\\ 0, & \text{if } \omega(x,t) \le \bar{\omega},\end{cases} \qquad \pi_{F\to L} = \begin{cases} 0, & \text{if } \omega(x,t) \ge \bar{\omega},\\ 1, & \text{if } \omega(x,t) < \bar{\omega},\end{cases}$$

  where $\bar{\omega}$ is a certain threshold;
- according to a mixed strategy combing the two.

- Define
$$m(t) = m_0(t) + m_1(t), \qquad V(t) = v_0(t) + v_1(t), \tag{12}$$
where
$$m_0(t) = \int_{\mathbb{R}^d} x\, f_0(x,t)dx, \qquad m_1(t) = \int_{\mathbb{R}^d} x\, f_1(x,t)dx,$$
$$v_0(t) = \int_{\mathbb{R}^d} \left|x - \frac{m_0}{\rho_0}\right|^2 f_0(x,t)dx, \qquad v_1(t) = \int_{\mathbb{R}^d} \left|x - \frac{m_1}{\rho_1}\right|^2 f_1(x,t)dx. \tag{13}$$
to be the mean and variance of $f_\lambda(x,t)$ for $\lambda \in \{0,1\}$.
- Following the idea in[12,13] we state the following Propositions, showing the decay of the variance and proving the convergence to the global minimum.

[12] A. Benfenati, G. Borghi, L. Pareschi. Binary interaction methods for high dimensional global optimization and machine learning, 2022

[13] J. A. Carrillo, Y-P Choi, C. Totzeck, O. Tse An analytical framework for a consensus-based global optimization method, 2018

Proposition (G.Albi, F.F., C.Totzeck)

Suppose to be in the stationary state and that the transition rates are constant. Assume the cost function $\mathcal{E}(x)$ positive and for all $x \in \mathbb{R}^d$

$$\underline{\mathcal{E}} := \inf_x \mathcal{E}(x) \leq \mathcal{E}(x) \leq \sup_x \mathcal{E}(x) := \overline{\mathcal{E}}. \tag{14}$$

Thus, if

$$\frac{\nu_L}{\beta} = \nu_F, \qquad \nu_F > \max\left\{ \frac{k\sigma_F^2 e^{\alpha(\bar{\mathcal{E}} - \underline{\mathcal{E}})}}{2}, \frac{\rho_1}{2} \right\} \tag{15}$$

with $k = d$ in the case of isotropic diffusion and $k = 1$ in the case of anisotropic diffusion, then

$$V(t) \to 0, \qquad \textit{for } \; t \to \infty.$$

- By direct computation, show that

$$\frac{dV(t)}{dt} \leq \mathcal{C}_v V(t) + \mathcal{C}_m \left(\frac{m_0(t)}{\rho_0} - \frac{m_1(t)}{\rho_1} \right)^2, \tag{16}$$

  for some constant $\mathcal{C}_v, \mathcal{C}_m > 0$.

- Show that

$$\left(\frac{m_0(t)}{\rho_0} - \frac{m_1(t)}{\rho_1} \right)^2 \to 0, \qquad \text{for } \; t \to \infty,$$

  and in particular that it is bounded from above by a certain constant $\bar{\mathcal{C}}$.

- Apply Grönwall lemma in (16) to show

$$V(t) \leq \left(V(0) + \mathcal{C}_m \, \bar{\mathcal{C}} \; t \right) e^{-\mathcal{C}_v t}.$$

- Take the limit $t \to \infty$ to get $V(t) \to 0$.

Convergence to global minimum

Proposition (G.Albi, F.F., C.Totzeck)

Suppose the same assumptions of Proposition 1 hold. Furthermore, assume the cost function $\mathcal{E} \in C^2(\mathbb{R}^d)$ *and that* $\exists c_1, c_2 > 0$ *s.t.*

$$\sup_{y \in \mathbb{R}^2} |\nabla \mathcal{E}(y)| \leq c_1, \qquad \sup_{y \in \mathbb{R}^2} |\Delta \mathcal{E}(y)| \leq c_2. \tag{17}$$

Choose the parameters s.t.

$$\frac{\mu}{M_\alpha^2(0)} \leq \frac{3}{4}, \tag{18}$$

with μ *a certain constant dependent on the parameters, and*

$$M_\alpha(t) = \int_{\mathbb{R}^d} e^{-\alpha \mathcal{E}(x)} g(x) dx. \tag{19}$$

Then $\exists\, \tilde{x} \in \mathbb{R}^d$ *s.t.* $m(t) \to \tilde{x}$ *as* $t \to \infty$ *and*

$$\mathcal{E}(\tilde{x}) = \underline{\mathcal{E}}. \tag{20}$$

Algorithm (Relaxed GKBO)

1. *Given* N_s *samples* (x_i^0, λ_i^0) *from the initial distribution* $f_\lambda(x, 0)$.
2. *Compute* $\hat{x}^0$ *as in equation* (3).
3. `while` $n < N_t$ `and` $j < j_{stall}$
   1. `for` $i = 1$ `to` N_s
      - *Select randomly a leader with position* y_k^n, $k \neq i$.
      - *Compute the positions change* x_i^{n+1} *as*

$$x_i^{n+1} = x_i^n + \nu_F h\left(y_k^n - x_i^n\right) + \sigma_F \sqrt{h} D\xi\left(1 - \lambda_i^n\right) + h\frac{\nu_L}{\beta}(\hat{x}^n - x_i^n)\lambda_i^n. \quad (21)$$

      - `if` $\lambda_i^n = 0$, *with probability* $h\,\pi_{F\to L}$ *agents* i *becomes a leader:* $\lambda_i^{n+1} = 1$.
      - `if` $\lambda_i^n = 1$, *with probability* $h\,\pi_{L\to F}$ *agents* i *becomes a follower:* $\lambda_i^{n+1} = 0$.

      `end for`
   2. *Compute* $\hat{x}^{n+1}$ *as in equation* (3).
   3. `if` $\|\hat{x}^{n+1} - \hat{x}^n\|_\infty \leq \delta_{stall}$

      $j \leftarrow j + 1$

      `end if`

   `end while`

[14] K. Nanbu. Direct simulation scheme derived from the Boltzmann equation. i. monocomponent gases, 1980.

- Consider the Rastrigin function as benchmark function.
- Fix $\nu_F = 1$, $\nu_L = 4$, $\beta = 0.4$.
- Run $M = 20$ simulations and consider one successful if

$$\|\hat{x}(t) - \bar{x}\|_\infty \leq 0.25,$$

with $\bar{x}$ defined to be the position of the global minimizer.

Leaders dependence. In the Table the iterations number (success rate) for the GKBO algorithm as the leaders mass at the equilibrium ρ_1^∞ varies for $\sigma_F = 4$ and $d = 20$.

	GKBO random	GKBO $\bar{p} = 0.1$	GKBO weighted
$\rho_1^\infty = 0.05$	1927 (0.15)	1488 (1)	3029 (1)
$\rho_1^\infty = 0.1$	1496 (1)	1578 (1)	2938 (1)
$\rho_1^\infty = 0.15$	1862 (1)	2181 (1)	3014 (1)
$\rho_1^\infty = 0.2$	–	4749 (0.6)	3073 (1)

The success rate and iterations number for the KBO algorithm are 1 and 7000 respectively.

Success rate as both the diffusion parameter and the dimension vary.

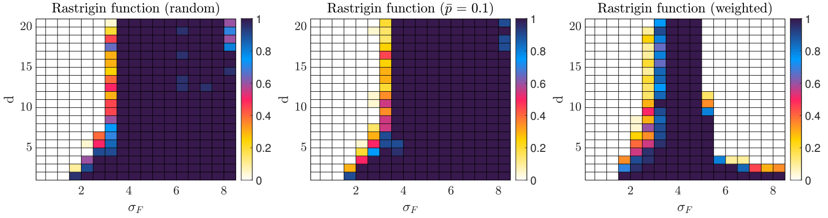


Figure: Success rate as σ_F and d vary with dynamics simulated with the GKBO method. On the left, random leaders generation. In the centre, mixed strategy with $\bar{p} = 0.1$. On the right, weighted leaders generation.

Success rate and iterations number as the diffusion parameters varies.

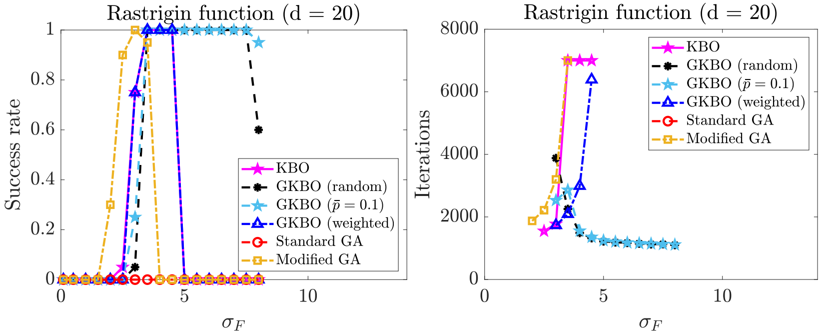


Figure: Success rate and iterations number as σ_F varies for $d = 20$ with dynamics simulated with the GKBO, KBO, GA methods.

Success rate and iterations number as the dimension varies.

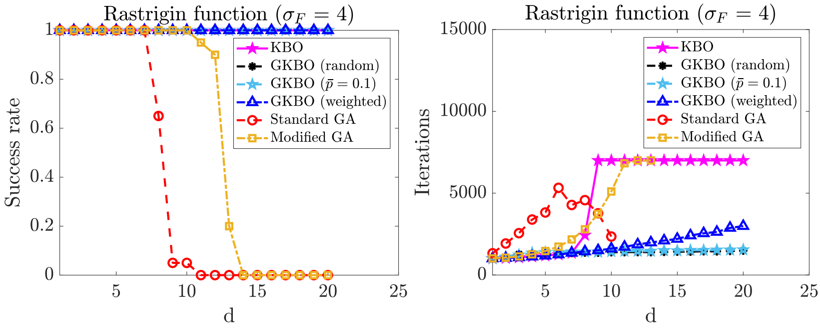


Figure: Success rate and iterations number as d varies for $\sigma_F = 4$ with dynamics simulated with the GKBO, KBO, GA methods.

Accuracy of the KBO and the GKBO algorithms.

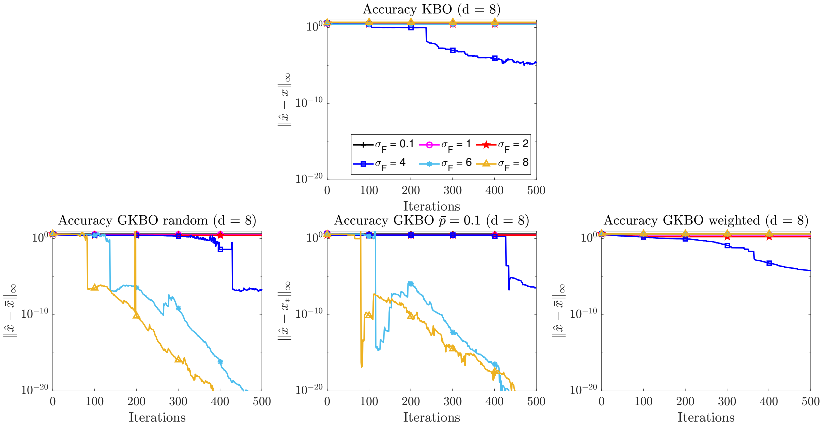


Figure: Accuracy of the KBO (first row) and GKBO (second row) algorithms as σ_F varies for $d = 8$ with dynamics simulated with the GKBO, KBO, GA methods.

Success rate and iterations number for different benchmark functions[15]

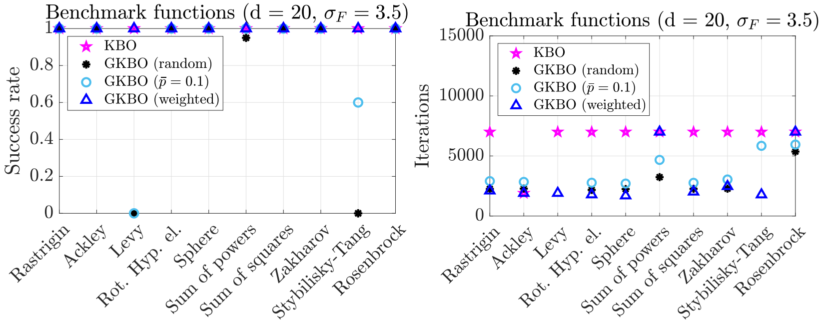


Figure: Success rate and iterations number for $d = 20$ and $\sigma_F = 3.5$ with dynamics simulated with the GKBO method and with the KBO method.

[15] M. Jamil and X.-S. Yang. A literature survey of benchmark functions for global optimisation problems, 2013

- We have introduced an efficient numerical methods for global optimization of non-convex high dimensional functions gluing together the ideas of the KBO and the GA algorithms.
- Results about convergence to the global minimizer are still valid.
- By introducing leaders, it is possible to improve the success rate and to reduce the iterations number of the considered algorithms.
- Future plan: extended this algorithm to localized versions useful to minimize functions with multiple global minima.

- We have introduced an efficient numerical methods for global optimization of non-convex high dimensional functions gluing together the ideas of the KBO and the GA algorithms.
- Results about convergence to the global minimizer are still valid.
- By introducing leaders, it is possible to improve the success rate and to reduce the iterations number of the considered algorithms.
- Future plan: extended this algorithm to localized versions useful to minimize functions with multiple global minima.

Thank you for your attention!