Finding global minima of non-convex functions via swarm-based gradient descent (SBGD)

Moritz Danzebrink and Janina Tikko

Work group of Prof. Dr. Angela Kunoth, University of Cologne

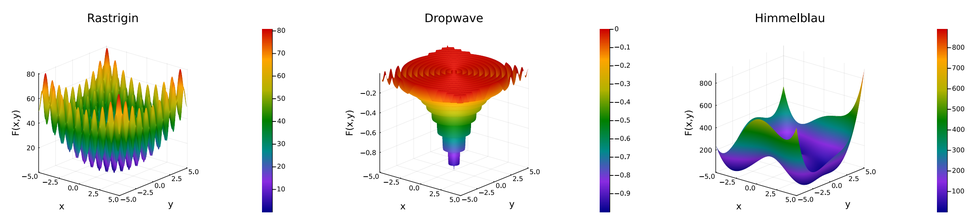
Problem

Consider a function $F : \Omega \subset \mathbb{R}^d \to \mathbb{R}$, not necessarily convex. Find a global minimum of F defined as

$$\underset{\mathbf{x}\in\Omega\subset\mathbb{R}^d}{\operatorname{argmin}} \; F(\mathbf{x})$$

Possible challenges for standard gradient descent methods:

- High frequency of local minima
- Varying amplitude between maxima and minima
- Different kinds of minima like valleys or basins
- Flat surface that gives little to no information to the minimizer



https://infinity77.net/global_optimization/test_functions.html

Motivation: Modifying gradient descent by applying swarm-dynamics

Swarm-based approach

Combine swarm-behavior according to Cucker and Smale [1] with gradient descent

Agents: Consider $N \in \mathbb{N}$ agents from $\mathbb{R}^d \times (0, 1]$

- Each agent is characterized by a position $\mathbf{x}_i(t) \in \mathbb{R}^d$ and a mass $m_i(t) \in (0, 1]$
- The total mass of all agents is constant at all times

$$\sum_{i=1}^{N} m_i(t) = 1$$

I. Communication: between agents is realized by mass transition

- Define the relative height of an agent as

$$\eta_i(t) := \frac{F(\mathbf{x}_i(t)) - F_{\min}(t)}{F_{\max}(t) - F_{\min}(t)} \geq 0, \tag{1}$$

with $F_{\max}(t)$ the maximum height and $F_{\min}(t)$ the minimum height of the swarm at time t

- The mass transition is defined as

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} m_i(t) = -\eta_i(t)^p m_i(t) & i \neq i^*(t) \\ m_i(t) = 1 - \displaystyle\sum_{j\neq i^*} m_j(t) & i = i^*(t)\,, \end{cases} \tag{2}$$

with $p > 0$ and where $i^*(t)$ denotes the index of the current minimizer

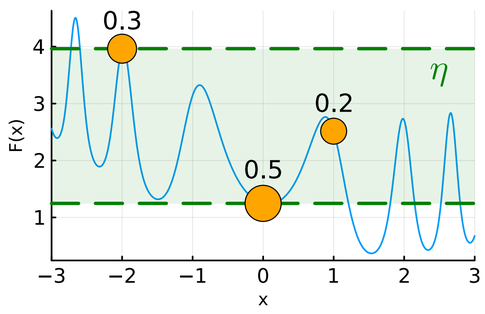


Figure 1: Before communication

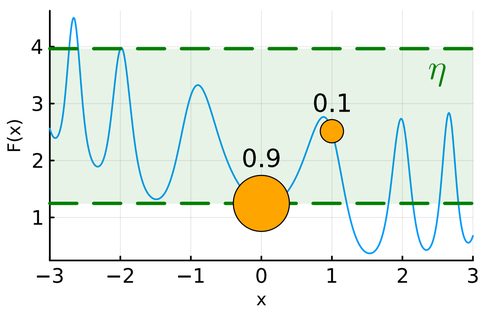


Figure 2: After communication

⇒ Only current minimizer receives mass from other agents
⇒ Highest placed agent is naturally eliminated

II. Stepsize protocol: Positions are adjusted by a stepsize h_i in direction of the negative gradient

- The stepsize h_i depends on the current position $\mathbf{x}_i(t)$ and the relative mass of the agent:

$$\tilde{m}_i(t) := \frac{m_i(t)}{m_+(t)},$$

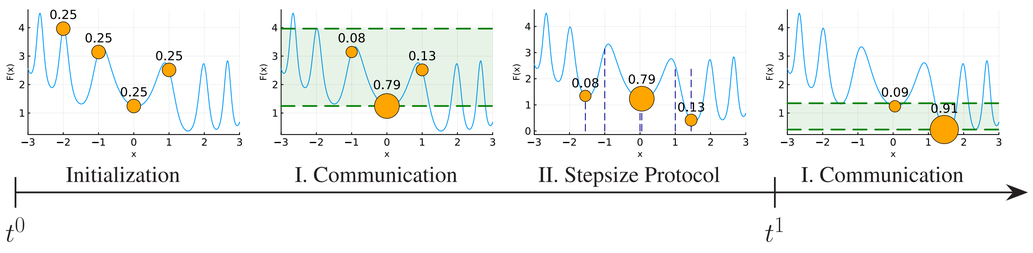
with $m_+(t) := \max_{i=1,\dots,N} m_i(t)$

- The relative mass $\tilde{m}_i$ can alternatively be understood as the probability of the agent to be a minimizer

⇒ h_i is a decreasing function of the relative mass, which assigns small stepsizes to heavier agents and larger stepsizes to lighter agents

I. + II. ⇝ Time discretization

- At the beginning set $N \in \mathbb{N}$ agents to random positions $\{\mathbf{x}_i^0\}$ with equal mass $\{m_i^0 = \frac{1}{N}\}$
- A timestep is defined by $t^{n+1} = t^n + \Delta t$ with $\Delta t = 1$
- Repeat until one agent remains: I. Communication, II. Stepsize protocol



Role of the relative mass

To obtain a stepsize h, we use the Backtracking line-search method:

- To guarantee descent on the function, the Wolfe condition [6] is applied

$$F(\mathbf{x}^{n+1}(h)) \leq F(\mathbf{x}^n) - \lambda h |\nabla F(\mathbf{x}^n)|^2, \quad \lambda \in (0, 1) \tag{3}$$

- Start with h large enough so that

$$F(\mathbf{x}^n - h\nabla F(\mathbf{x}^n)) > F(\mathbf{x}^n) - \lambda h |\nabla F(\mathbf{x}^n)|^2$$

- Successively reduce step size with shrinking factor $\gamma > 0$ until (3) is reached for $h = h(\mathbf{x}^n, \lambda)$

The relative mass of the individual agents is used to control the stepsize according to the stepsize protocol. For $i = 1, ..., N$ we thus obtain the step size

$$h_i^n = h(\mathbf{x}_i^n, \lambda(\tilde{m}^{n+1})^q) \tag{4}$$

The parameter $q > 0$ determines the influence of the relative mass.

Optimization approach of q [2]

- q has impact on quality of the solution and runtime of the algorithm
- Runtime is mostly influenced by the number of backtracking iterations
- As shown in [3] with certain thresholds q can be chosen from $(0, 76]$ for implementation in Julia
- However, the interesting changes happen for $q \in (0, 4]$

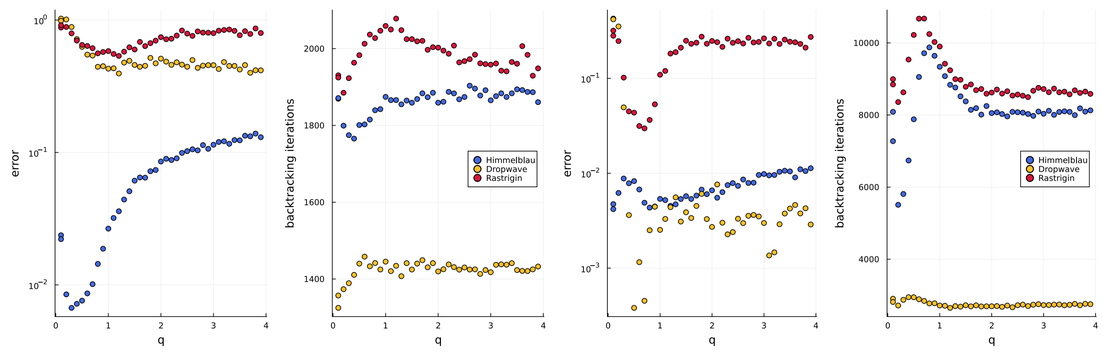


Figure 3: $N = 10$

Figure 4: $N = 50$

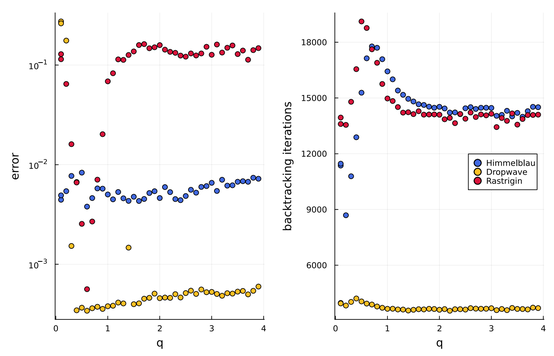


Figure 5: $N = 100$

$q\backslash N$	10	50	100	Function
0.2	64.2%	68.4%	68.9%	Himmelblau
0.5	68.4%	60.0%	58.3%	
1	46.0%	75.9%	76.8%	
0.2	0.17%	18.1%	36.6%	Dropwave
0.5	16.1%	99.9%	100%	
1	28.2%	98.3%	99.8%	
0.2	8%	57.5%	74.3%	Rastrigin
0.5	16.9%	92.8%	99.9%	
1	19.6%	79.2%	86.6%	

Table 1: Success rate of 1000 simulations with SBGD on initial data in $[-5, 5]^2$

⇒ In general more agents seem to be better, and $q = 0.5$ is a good choice
⇒ There is a tradeoff between accuracy and runtime

SBGD compared to other methods [4]

Success rates of 1000 independent simulations of $\text{SBGD}_{p,q}$ for 2D Rastrigin compared with

- Gradient Descent with constant stepsize $h = 0.004$ (GD(0.004))
- Backtracking Gradient Descent, meaning no communication between agents (GD(BT))
- Adam's method with initial stepsize $h_0 = 0.2$ (Adam(0.2)) and $h_0 = 0.8$ (Adam(0.8))

N	5	10	15	20	30
$\text{SBGD}_{1,1}$	34.4%	52.1%	62.6%	70.0%	75.8%
$\text{SBGD}_{2,1}$	34.5%	60.1%	75.3%	84.3%	91.0%
GD(0.004)	36.3%	50.5%	60.0%	70.0%	78.1%
GD(BT)	35.0%	51.0%	62.0%	70.8%	79.3%
Adam(0.8)	23.7%	29.6%	39.1%	46.8%	65.5%
Adam(0.2)	32.1%	40.9%	55.9%	65.3%	79.4%

Table 2: Initial data in $[-3, 3]^2$

N	5	10	15	20	30
$\text{SBGD}_{1,1}$	17.0%	49.2%	61.7%	67.0%	72.7%
$\text{SBGD}_{2,1}$	14.2%	46.7%	68.4%	81.9%	89.6%
GD(0.004)	0.0%	0.0%	0.0%	0.0%	0.0%
GD(BT)	1.8%	2.4%	3.4%	4.3%	5.9%
Adam(0.8)	24.5%	31.3%	41.4%	49.2%	66.9%
Adam(0.2)	0.0%	0.0%	0.0%	0.0%	0.0%

Table 3: Initial data in $[-3, -1]^2$

Conclusion: Communication is key!

References

[1] F. Cucker, S. Smale, *Emergent Behavior in Flocks*, IEEE Transactions on automatic control, Volume 52, Pages 852-862, 2007, doi:10.1109/TAC.2007.895842.

[2] M. Danzebrink, J. Tikko, *Finding global minima of non-convex functions via Swarm-Based methods*, in preparation.

[3] M. Danzebrink, *Optimization approach for Swarm-Based Gradient Descent in multiple arguments*, Bachelor Thesis, Department of Mathematics and Computer Science, Division of Mathematics, University of Cologne, November 2024.

[4] L. Jingcheng, E. Tadmor, A. Zenginoğlu, *Swarm-based gradient descent method for non-convex optimization*, Communications of the American Mathematical Society, Volume 4, Pages 787-822, 2024, doi:10.1090/cams/42.

[5] J. Tikko, *Introduction: Swarm-based gradient descent for non convex optimization*, 2024, https://arxiv.org/abs/2404.00005.

[6] P. Wolfe, *Convergence conditions for ascent methods*, Siam Review, Volume 11, Pages 226-235, 1969, doi:10.1137/1013035.

Janina's work was supported by Hypatia.Science, an initiative for the promotion of young female scientists at the Department of Mathematics and Computer Science of the University of Cologne.