Swarm-Based Gradient Descent Method for Non-Convex Optimization

Eitan Tadmor [1,*]

in honor of Albert Cohen

Nonlinear Approximation for High-Dimensional Problems
Sorbonne University, July 2 2025

1

* Joint works with J. Lu, A. Zenginoglu, Z. Ding, M. Guerra & Q. Li

Gradient Descent method

- Objective function (or "loss function") $F(\cdot)$; state space $\Omega \subset \mathbb{R}^d$:

$$\text{Find} \quad \underset{\mathbf{x}\in\Omega}{\operatorname{argmin}} F(\mathbf{x})$$

- Gradient Descent (GD) method (Cauchy[2] 1847) $\min\{F(\mathbf{x}) : \nabla F(\mathbf{x}) = 0\}$

  Proceed in gradient direction $\nabla F^n = \nabla F(\mathbf{x}^n)$ discrete time steps $t^{n+1} = t^n + h$

$$\mathbf{x}^{n+1} = \mathbf{x}^n - h\nabla F^n, \qquad \mathbf{x}^n := \mathbf{x}(t^n)$$

- Gradient descent: $\forall\lambda \in (0,1),\ h \leqslant h_\lambda, \quad F(\mathbf{x}^n - h\nabla F^n) \leqslant F(\mathbf{x}^n) - \lambda h|\nabla F^n|^2$
- Different protocols for time stepping $h_\lambda = h(\mathbf{x}^n, \lambda)$:

  Backtracking, Adam ('momentum', Kingma & Ba (2017)), ...

  AEGD ('energy', Liu (2021))...

---

[2] "*I'll restrict myself here to outlining the principles underlying [my method], with the intention to come again over the same subject, in a paper to follow.*"
A. Cauchy. Mèthode gènèrale pour la rèsolution des systémes d'èquations simultanèes. C. R. Acad. Sci. Paris, 25:536-538, 1847.

Swarm-based dynamics

☞ **Agents**: position $\mathbf{x}_i^n = \mathbf{x}_i(t^n) \in \mathbb{R}^d$ with mass $m_i^n = m_i(t^n) \in (0, 1]$

☞ **Communication**: adjust the masses according to relative heights η:

$$\begin{cases} m_i^{n+1} = m_i^n - \eta_i^q m_i^n, & \eta_i := \dfrac{F(\mathbf{x}_i^n) - F_-(t^n)}{F_+(t^n) - F_-(t^n)}, \quad i \neq i_n \\ m_{i_n}^{n+1} = m_{i_n}^n + \displaystyle\sum_{i \neq i_n} \eta_i^q m_i^n, & i_n := \operatorname*{argmin}_i F(\mathbf{x}_i^n) \quad F_-(t^n) = F(\mathbf{x}_{i_n}^n) \end{cases}$$

- Dynamic transfer of mass from high ground to lower ground
- Using η_i^q — the higher q the more tamed mass transfer

☞ **Protocol for time stepping**: dictated by relative masses $\widetilde{m}$:

$$\mathbf{x}_i^{n+1} = \mathbf{x}_i^n - h\big(\mathbf{x}_i^n, \lambda \widetilde{m}_i^{n+1}\big) \nabla F_i^n, \qquad \widetilde{m}_i^{n+1} = \frac{m_i^{n+1}}{m_+^{n+1}}, \quad m_+ := \max_i m_i$$

- Interplay between positions and masses

Time stepping protocol — backtracking

- Protocol for choosing the step size ("learning rate") — backtracking:

$$(\star) \quad \text{For "small enough" } h: \quad F(\mathbf{x}^n - h\nabla F^n) \leqslant F(\mathbf{x}^n) - \lambda h |\nabla F^n|^2, \quad \lambda \in (0,1)$$

Let $h(\mathbf{x}^n, \lambda)$ – the largest for which $(\star)$ holds

· A key aspect: $h(\cdot, \lambda\widetilde{m})$ is decreasing function of the $\widetilde{m} = \dfrac{m}{m_+}$

$$F(\mathbf{x}_i^{n+1}) \leqslant F(\mathbf{x}_i^n) - \lambda \widetilde{m}_i^{n+1} h_i^n |\nabla F_i^n|^2, \quad h_i^n := h(\mathbf{x}_i^n, \lambda \widetilde{m}_i^{n+1})$$

☞ Dynamic distinction — 'leaders' and 'explorers'...

heavy agents, $m_i^{n+1} \sim m_+^{n+1} \ \rightsquigarrow$ small time steps; lead near critical point

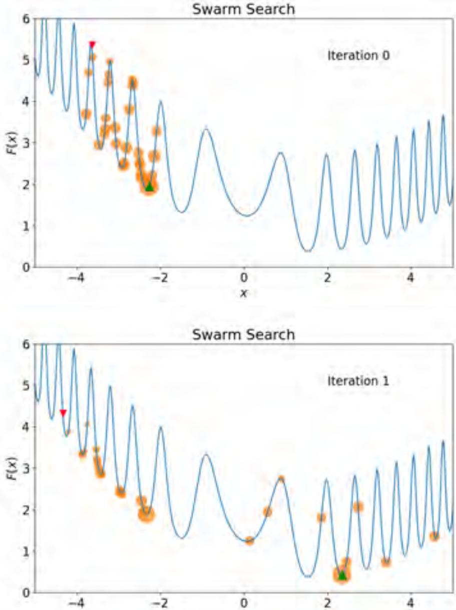
light agents, $m_i^{n+1} \ll m_+^{n+1} \ \rightsquigarrow$ large time steps; explore the landscape

- Each agent sheds an η-fraction of its mass – shifts to current minimizer

- Survival of the fittest: highest agent, $\eta_{\text{max}} = 1$, is eliminated
- Conservation of mass $\sum_i m_i^n = 1$: $\{m_i\}$ viewed as probabilities of agents to identify global minimum

Why communication is important? survival of the fittest

$$F(x) = e^{\sin(2x^2)} + \frac{1}{10}\left(x - \frac{\pi}{2}\right)^2$$



Alignment towards the minimal value

- Consensus Based Optimization (CBO)[3]:
  SDEs with steering towards the weighted min ('Laplace principle')
  · Biologically inspired methods —

ant colony optimization[4]; artificial bee colony optimization[5]; firefly optimization[6]

- Wind-Driven Optimization (WDO)[7]: inspired by modeling pressure and wind
- Particle Swarm Optimization (PSO)[8]: explore the state space with randomized drifts towards the global best position
- Simulated Annealing (SA)[9]: exploration of state space is driven by noise terms that are 'cooled down' as time evolves

[3] Pinnau, Totzeck, Tse, & Martin, A CBO for global optimization, M3AS 27 (2017)
Carrillo, Choi, Totzeck & Tse, An analytical framework for CBO, M3AS 28 (2018)

[4] Mohan & Baskaran, Ant colony optimization Expert Syst. Appl. 39 (2012)

[5] Karaboga et. al., Artificial bee colony (ABC) algorithm, Artif. Intell. Rev. 42 (2014)

[6] Yang, Firefly Algorithms for Multimodal Optimization. SAGA, V. 5792, (2009)

[7] Bayraktar et. al., The Wind Driven Optimization, IEEE Trans. Ant. Prop. 61 (2013)

[8] Kennedy & Eberhart, Particle Swarm Optimization. Proc. IEEE (1995)
Poli, Kennedy & Blackwell, Particle swarm optimization, Swarm Intell. 1 (2007)

[9] Kirkpatrick, Gelatt and Vecchi, Optimization by simulated annealing, Science (1983)
Chen et. al.,Accelerating nonconvex learning via replica exchange diffusion ICLR (2019)

SBGD implementation $\rightsquigarrow$ SBGD$_q$

SET tolerance parameters, *tolm*, *tolmerge* and *tolres*; adjustment parameter $q > 0$
INITIALIZE: # N; positions $\{\mathbf{x}_i^0\}$; masses $m_i^0 = 1/N$: optimal agent, $i_0 = \operatorname{argmin}_i F(\mathbf{x}_i^0)$
for $n = 0, 1, 2, \cdots$ **do**
  Set $F_{\min}^n = F(\mathbf{x}_{i_n}^n)$, $F_{\max}^n = \max_i F(\mathbf{x}_i^n)$
  for $i = 1, \cdots, N$ and $i \neq i_n$ **do**   % Mass transitions
    if $m_i^n < 1/N * tolm$ **then**
      SET $m_i^{n+1} = 0$
      reduce the # of active agents: $N \leftarrow N - 1$
    else $m_i^{n+1} = m_i^n - (\eta_i^n)^q m_i^n$ where $\eta_i^n = \dfrac{F(\mathbf{x}_i^n) - F_{\min}^n}{F_{\max}^n - F_{\min}^n}$
    end if
  end for
  $m_{i_n}^{n+1} = m_{i_n}^n + \sum_{i \neq i_n} (\eta_i^n)^q m_i^n$ % The total mass of the swarm is conserved
  Compute $m_+ = \max_i m_i^{n+1}$
  for $i = 1, \cdots, N$ **do**   % Gradient descent
    Compute relative mass $\widetilde{m}_i^{n+1} := m_i^{n+1}/m_+$
    Compute the step size $h = h(\mathbf{x}_i^n, \lambda \widetilde{m}_i^{n+1})$ according to algorithm 1 (backtracking).
    MARCH: $\mathbf{x}_i^{n+1} = \mathbf{x}_i^n - h \nabla F(\mathbf{x}_i^n)$
  end for
  MERGE agents if their distance $<$ *tolmerge*
  SET the new optimal agent $i_{n+1} = \operatorname{argmin}_i F(\mathbf{x}_i^{n+1})$.
  if $res <$ *tolres* **then** $\mathbf{x}_{min} \leftarrow \mathbf{x}_{i_{n+1}}^{n+1}$ % $res := |\mathbf{x}_{i_{n+1}}^{n+1} - \mathbf{x}_{i_n}^n|_2$
  end if
end for

Quantifying the descent property

- Assume the Lip bound $|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})| \leqslant L|\mathbf{x} - \mathbf{y}|$

$$
\begin{aligned}
F(\mathbf{x}_i^n - h\nabla F_i^n) &\leqslant F(\mathbf{x}_i^n) - h|\nabla F_i^n|^2 + \frac{h^2}{2}L|\nabla F_i^n|^2 \\
&\leqslant F(\mathbf{x}_i^n) - \left(1 - \frac{h}{2}L\right)h|\nabla F_i^n|^2, \qquad \nabla F_i^n := \nabla F(\mathbf{x}_i^n)
\end{aligned}
$$

hence set h_λ : $\left(1 - \frac{1}{2}h_\lambda L\right) \geqslant \lambda\widetilde{m}_i^{n+1}$

$$
\forall h \leqslant h_\lambda : \quad F(\mathbf{x}_i^n - h\nabla F_i^n) \leqslant F(\mathbf{x}_i^n) - \lambda\widetilde{m}_i^{n+1}h|\nabla F_i^n|^2
$$

- Certainly holds for h is small enough : OK if $h_\lambda < \dfrac{2}{L}\left(1 - \lambda\widetilde{m}_i^{n+1}\right)$
- And this does not mean h_λ need to be small ...
- But no access to L ...

Backtracking — implementation

Algorithm 1 Backtracking Line Search

SET the descent parameter $\lambda \in (0,1)$, and shrinkage parameter, $\gamma \in (0,1)$

Compute the relative masses $\widetilde{m}_i^{n+1} = \dfrac{m_i^{n+1}}{m_+^{n+1}}$

INITIALIZE large step size $h = h_0 > \frac{2}{L}$.

while $F\big(\mathbf{x}_i^n - h\nabla F_i^n\big) > F(\mathbf{x}_i^n) - \lambda \widetilde{m}_i^{n+1} h |\nabla F_i^n|^2$ **do** % weighted descent

  $h \leftarrow \gamma h$ % backtracking

end while

SET $h \mapsto h_i^n := h\big(\mathbf{x}_i^n, \lambda \widetilde{m}_i^{n+1}\big) \geqslant \dfrac{2\gamma}{L}(1 - \lambda \widetilde{m}_i^{n+1})$

- Lower-bound: $\dfrac{h_i^n}{\gamma} \geqslant \dfrac{2}{L}\big(1 - \lambda \widetilde{m}_i^{n+1}\big), \quad |\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})| \leqslant L|\mathbf{x} - \mathbf{y}|$

☞ Descent property: $F(\mathbf{x}_i^{n+1}) \leqslant F(\mathbf{x}_i^n) - \dfrac{2\gamma}{L}\big(1 - \lambda \widetilde{m}_i^{n+1}\big)\lambda \widetilde{m}_i^{n+1} |\nabla F_i^n|^2$

· Noll & Rondepierre (2013): backtracking protocol with memory

Descent estimates: $F(\mathbf{x}_i^{n+1}) \leqslant F(\mathbf{x}_i^n) - \dfrac{2\gamma}{L}\big(1-\lambda\widetilde{m}_i^{n+1}\big)\lambda\widetilde{m}_i^{n+1}|\nabla F_i^n|^2$

- Descent property: How small $\widetilde{m}_i^{n+1}$ can get for SBGD minimizers?
- Heaviest agent at $\mathbf{X}_+^n$ w/relative mass $\widetilde{m}=1$; compare w/minimizer $\mathbf{X}_-^n$:

$$\widetilde{m}_{i_n}^{n+1}=1 \ \rightsquigarrow\ F(\mathbf{X}_-^{n+1}) \leqslant F(\mathbf{X}_+^{n+1}) \leqslant F(\mathbf{X}_+^n) - \frac{2\gamma}{L}(1-\lambda)\lambda|\nabla F(\mathbf{X}_+^n)|^2$$

A. Scenario (i): if $F(\mathbf{X}_+^n) \leqslant F(\mathbf{X}_-^n) + \dfrac{\gamma}{L}\big(1-\lambda\big)\lambda|\nabla F(\mathbf{X}_+^n)|^2$ then ...

$$F(\mathbf{X}_-^{n+1}) \leqslant F(\mathbf{X}_+^{n+1}) \leqslant F(\mathbf{X}_+^n) - \frac{2\gamma}{L}(1-\lambda)\lambda|\nabla F_+^n|^2 \leqslant F(\mathbf{X}_-^n) - \frac{\gamma}{L}(1-\lambda)\lambda|\nabla F_+^n|^2$$

B. Scenario (ii): $F(\mathbf{X}_+^n) > F(\mathbf{X}_-^n) + \dfrac{\gamma}{L}\big(1-\lambda\big)\lambda|\nabla F(\mathbf{X}_+^n)|^2$

The minimzier at $\mathbf{X}_-^n$ must be different from the heaviest at $\mathbf{X}_+^n$

hence must gain η-fraction mass from heaviest at $\mathbf{X}_+^n$: $m_-^{n+1} > \eta_+^n m_+^{n+1}$

$$\rightsquigarrow\ \widetilde{m}_-^{n+1} > \eta_+^n \geqslant \frac{1}{M^q}\big(F(\mathbf{X}_+^n) - F(\mathbf{X}_-^n)\big)^q > \frac{\gamma}{ML}(1-\lambda)|\nabla F_+^n|^2$$

$$F(\mathbf{X}_-^{n+1}) \leqslant F(\mathbf{X}_-^n) - \frac{2\gamma}{L}(1-\lambda\widetilde{m}_-^{n+1})\lambda\widetilde{m}_-^{n+1}|\nabla F_-^n|^2$$

$$\leqslant F(\mathbf{X}_-^n) - \Big(\frac{\gamma(1-\lambda)\lambda}{ML}\Big)^q|\nabla F_+^n)|^{2q}\cdot\frac{\gamma}{L}\lambda|\nabla F_-^n|^2$$

☞ telescoping sum A+B ...

Convergence. I (global)

- Recall the mass transfer: $m_i^{n+1} = m_i^n - \eta_i^q m_i^n$, $\eta_i := \frac{F(\mathbf{x}_i^n) - F_{\min}(t^n)}{F_{\max}(t^n) - F_{\min}(t^n)}$

  $\rightsquigarrow F(\mathbf{X}_-^{n+1}) \leqslant F(\mathbf{X}_-^n) - \lambda h_- |\nabla F_-^n|^{2(q+1)}$

Theorem

$$\sum_{n=n_0}^{\infty} \min\left\{|\nabla F(\mathbf{X}_+^n)|, |\nabla F(\mathbf{X}_-^n)|\right\}^{2(q+1)} < C \min_i F(\mathbf{x}_i^0)$$

☞ Limit set of minima of equi-height: $\mathbf{X}_-^{n_\alpha} \stackrel{\alpha\to\infty}{\longrightarrow} \mathbf{X}_\alpha^*$, $\nabla F(\mathbf{X}_-^{n_\alpha}) \stackrel{\alpha\to\infty}{\longrightarrow} 0$

- "Generic" descent property $F(\mathbf{X}_-^{n+1}) \leqslant F(\mathbf{X}_-^n) - \lambda h_- |\nabla F_-^n|^2$

$$\sum_{n=n_0}^{\infty} \min\left\{|\nabla F(\mathbf{X}_+^n)|, |\nabla F(\mathbf{X}_-^n)|\right\}^{2} < C \min_i F(\mathbf{x}_i^0)$$

Convergence. II (local)

- Lojasiewicz bound[10a]: $\exists \mathcal{N}, \beta$ s.t. $\mu|F(\mathbf{x}) - F(\mathbf{x}^*)| \leqslant |\nabla F(\mathbf{x})|^{\beta}, \ \forall \mathbf{x} \in \mathcal{N}(\mathbf{x}^*)$
  - quantifies "Flatness" in terms of "$\beta = \dfrac{m+1}{m} \in (1,2]$" for SA F's[10a]
  - basin of convexity: $m = 1 \ \rightsquigarrow \ \beta = 2$;
- Generic descent property[10b] at $\mathbf{X}_-^{n_\alpha}$: $\quad F_-^{n_\alpha+1} \leqslant F_-^{n_\alpha} - \lambda h_- |\nabla F_-^{n_\alpha}|^2 \ \rightsquigarrow$

$$F_-^{n_\alpha+1} - F_\alpha^* \leqslant F_-^{n_\alpha} - F_\alpha^* - \lambda h_- \big(\mu(F_-^{n_\alpha} - F_\alpha^*)\big)^{2/\beta}, \quad \mathbf{X}_-^{n_\alpha} \in \mathcal{N}(\mathbf{X}_\alpha^*).$$

Theorem

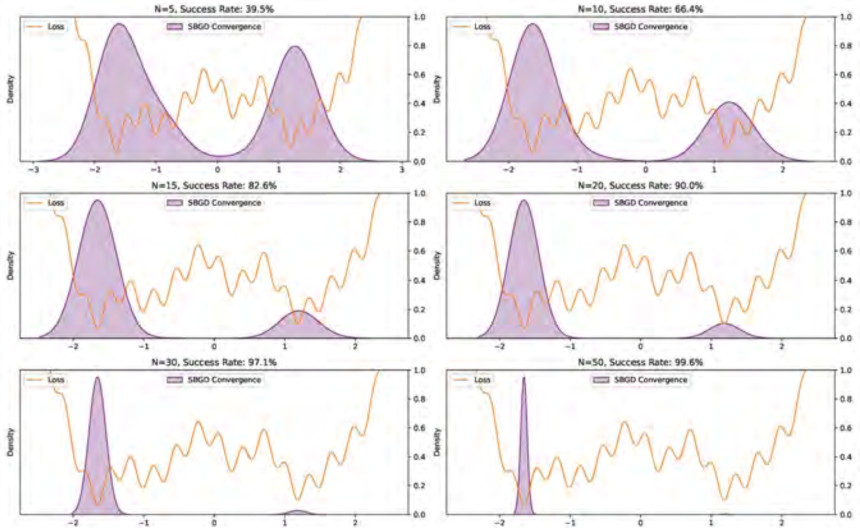
~~Analytic~~ Semi-alegbraic loss function F with minimal flatness $\beta \in (1,2]$.
SBGD minimizers $\{\mathbf{X}_-^n\}_{n \geqslant 0}$. $\exists C = C(\gamma, \lambda, \mu)$, such that

$$F(\mathbf{X}_-^{n_\alpha}) - F(\mathbf{X}_\alpha^*) \begin{cases} \leqslant \Big(1 - \dfrac{2\mu\gamma\lambda(1-\lambda)}{L}\Big)^{n_\alpha} \big(F_{\min}(\mathbf{x}^0) - F(\mathbf{x}^*)\big), & \beta = 2 \\ \lesssim \Big(\dfrac{C}{n_\alpha}\Big)^{\frac{\beta}{2-\beta}}, & \beta \in (1,2) \end{cases}$$

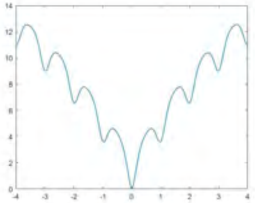
[10a] Lojasiewicz IHES1965; J. Bolte et. al., on Kurdyka-Lojasiewicz ineq. TAMS 2010
[10b] Absil et. al., Convergence of descent iterations for analytic functions, SJOPT2005

Another convincing 1D example (Boffi-Slotine (2020))

$$F(x) = \frac{1}{C}\Big(x^4 - 4x^2 + \frac{1}{5}x + \frac{2}{5}\big(3\sin(20x) - \frac{7}{2}\sin(2\pi x) + \cos(\frac{10ex}{3})\big)\Big)$$



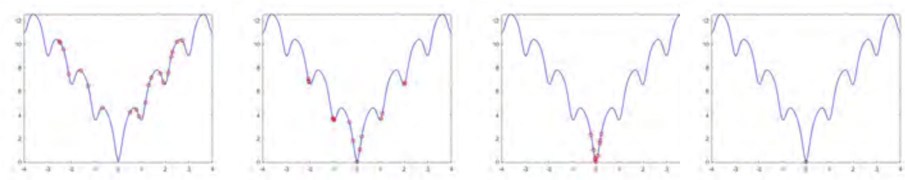
$$F(x) = -20e^{-0.2|x-B|} - e^{\cos(2\pi(x-B))} + 20 + e + C$$



$\mathbf{x}^* = B$		N=10	N=20	N=30	N=10	N=20	N=30
$B = 0$	success rate	100%	100%	100%	100%	100%	100%
	$\mathbb{E}\|x_{SOL} - x^*\|^2$	$8.42e^{-10}$	$8.37e^{-10}$	$3.38e^{-10}$	$8.60e^{-10}$	$1.36e^{-9}$	$1.29e^{-9}$
$B = 5$	success rate	100%	100%	100%	100%	100%	100%
	$\mathbb{E}\|x_{SOL} - x^*\|^2$	$8.41e^{-10}$	$7.58e^{-10}$	$5.01e^{-10}$	$8.51e^{-10}$	$1.25e^{-9}$	$1.21e^{-9}$
$B = 15$	success rate	98.5%	100%	100%	46.5%	75.0%	85.5%
	$\mathbb{E}\|x_{SOL} - x^*\|^2$	$1.41e^{-2}$	$4.69e^{-3}$	$8.27e^{-10}$	$6.44e^{+1}$	$2.26e^{+1}$	$1.14e^{+1}$
$B = 25$	success rate	45.5%	89.0%	98.5%	0%	0%	0%
	$\mathbb{E}\|x_{SOL} - x^*\|^2$	$1.48e^{+2}$	$1.49e^{+1}$	$3.28e^{-1}$	$4.86e^{+2}$	$4.50e^{+2}$	$4.38e^{+2}$

Table: $m = 200$ runs of shifted 1D Ackley function. Left: SBGD. Right: GD($\eta = 0$)(no communication). Backtracking descent parameter $\lambda = 0.2$; shrinkage parameter $\gamma = 0.9$

Dynamics of SBGD agents; 1D Ackley simulation ($B = C = 0$)

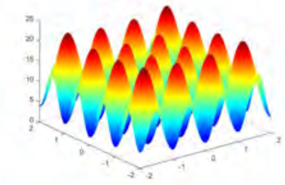


(a) n = 1, N = 20 (b) n = 3, N = 17 (c) n= 8, N = 9 (d) n=15, N = 1

SBGD 1D Ackley

Simulations – two-dimensional

$$F_{\text{Rastrigin}}(x) = \frac{1}{d}\sum_{i=1}^{d}\left\{(\mathbf{x}_B)_i^2 - 10\cos(2\pi(\mathbf{x}_B)_i) + 10\right\} + C$$



N	10	20	30
$SBGD_1$	52.1%	70.0%	75.8%
$SBGD_2$	60.1%	84.3%	91.0%
GD($h\equiv0.004$)	50.5%	70.0%	78.1%
GD($\eta=0$)	51.0%	70.8%	79.3%
Adam(0.8)	29.6%	46.8%	65.5%
Adam(0.2)	40.9%	65.3%	79.4%

N	10	20	30
$SBGD_1$	49.2%	67.0%	72.7%
$SBGD_2$	46.7%	81.9%	89.6%
GD($h\equiv0.004$)	0.0%	0.0%	0.0%
GD($\eta=0$)	2.4%	4.3%	5.9%
Adam(0.8)	31.3%	49.2%	66.9%
Adam(0.2)	0.0%	0.0%	0.0%

Table: $m = 1000$ runs for 2D Rastrigin function, success rates of SBGD compared with GD(h) GD($\eta=0$) and Adam methods (first component).

Left: initial data uniformly generated in $[-3,3]^2$. Right: uniformly generated in $[-3,-1]^2$.

Backtracking parameters: descent parameter $\lambda = 0.8$; shrinkage parameter $\gamma = 0.9$

Simulations – 20-dimensions

$\mathbf{x}_* = B$		N=50	N=100	N=200	N=50	N=100	N=200
$B = 0$	SBGD_2	9.00e-07	2.02e-07	1.43e-07	4.03e-01	4.98e-01	3.91e-01
	GD(η=0)	1.18e-01	6.90e-02	1.88e-02	2.96e-01	3.25e-01	4.35e-01
	Adam(0.5)	3.37e-03	4.03e-03	4.96e-03	3.75e-01	2.86e-01	4.14e-01
$B = 3$	SBGD_2	1.65e-06	1.51e-06	7.96e-07	5.07	3.71	2.92
	GD(η=0)	7.64	6.72	5.67	9.43	9.02	8.57
	Adam(0.5)	2.14e-01	1.22e-01	1.27e-01	9.02	8.63	8.15
$B = 5$	SBGD_2	4.15	1.07	2.36e-01	3.53	2.45	1.29
	GD(η=0)	17.99	17.59	17.11	18.02	17.64	17.18
	Adam(0.5)	11.01	9.27	7.52	17.13	16.73	16.3

Table: $m = 1000$ runs. $\mathbb{E}[\mathbf{x}_{SOL} - \mathbf{x}^*|$ (first component).

Left: 20-dimensional shifted Ackley. Right: for 20-dimensional shifted Rastrigin.

Backtracking parameters: descent parameter $\lambda = 0.2$; shrinkage parameter $\gamma = 0.9$

Multi-D SBRD – "Where No One Has Gone Before"

- The need to enhance "space exploration" in high-D problems
- Time stepping: $\mathbf{x}_i^{n+1} = \mathbf{x}_i^n - h\mathbf{p}_i^n$

☞ Allow general set of directions $\{\mathbf{p}_i^n\}$ such that

$$(*) \qquad \langle \mathbf{p}_i^n, \nabla F(\mathbf{x}_i^n)\rangle \geqslant \frac{1+\widetilde{m}_i^{n+1}}{2}|\nabla F(\mathbf{x}_i^n)|^2 \geqslant \frac{1}{2}|\nabla F(\mathbf{x}_i^n)|^2$$

This secures the "one-half" descent property:

$$F(\mathbf{x}_i^{n+1}) \leqslant F(\mathbf{x}_i^n) - h\langle \mathbf{p}_i^n, \nabla F(\mathbf{x}_i^n)\rangle + \frac{h^2}{2}L|\nabla F(\mathbf{x}_i^n)|^2 \leqslant F(\mathbf{x}_i^n) - \frac{1}{2}(1-Lh)h|\nabla F(\mathbf{x}_i^n)|^2$$

- Heterogeneity of orientations $\rightsquigarrow$ effective exploration of the ambient space
- To secure (*): choose a Random orientation $\mathbf{p}_i^n = |\nabla F(\mathbf{x}_i^n)|\omega_i^n$ such that

$$\left\langle \omega_i^n, \mathbf{q}_i^n \right\rangle = \text{rand}, \quad \text{rand} \in \mathcal{N}\Big(\frac{1+\widetilde{m}_i^{n+1}}{2}, 1\Big), \qquad \mathbf{q}_i^n := \frac{\nabla F(\mathbf{x}_i^n)}{|\nabla F(\mathbf{x}_i^n)|}$$

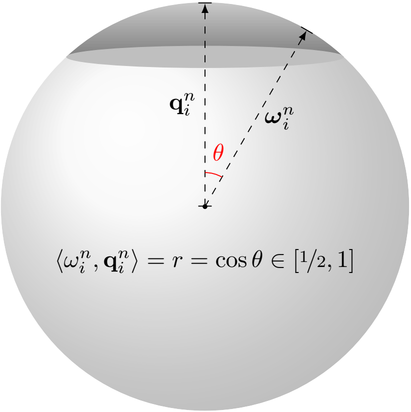


Figure: Spherical cap of $\mathbb{S}^{d-1}$-sphere centered at the gradient orientation $\mathbf{q}_i^n = \dfrac{\nabla F(\mathbf{x}_i^n)}{|\nabla F(\mathbf{x}_i^n)|}$. "Opening": $\cos(\theta) = \dfrac{1 + \lambda \widetilde{m}_i^{n+1}}{2}$ up to 60°

2D Simulations — SBRD 2D Ackley (shift: $\mathbf{x}_B := \mathbf{x} - B\mathbb{1}$)

- $F_{\text{Ackley}}(\mathbf{x}) = -20exp\left\{\dfrac{-0.2}{\sqrt{d}}|\mathbf{x}_B|\right\} - exp\left\{\dfrac{1}{d}\sum_i \cos(2\pi(\mathbf{x}_B)_i)\right\} + 20 + e + C$

SBRD 2D Ackley

Algorithm 2 Random descent direction

if $1 - \widetilde{m}_i^{n+1} \leqslant \mathit{tolrand}$ **then**
  set $\mathbf{p}_i^n = \nabla F(\mathbf{x}_i^n)$ % for heaviest agent use the gradient direction
end if
else set $\mathbf{q}_i^n = \dfrac{\nabla F(\mathbf{x}_i^n)}{|\nabla F(\mathbf{x}_i^n)|}$
Choose random r, $\frac{1}{2}(1 + \widetilde{m}_i^n) < \text{rand} < 1$
  % Random X in $\mathbb{S}^{d-1}$: spherical cap centered at the north pole $\mathbf{z} := (0, \cdots, 0, 1)$
for $i = 1, \cdots, d-1$ **do** % $\dfrac{Y(i)}{|\mathbf{Y}|}$ is random in $\mathbb{S}^{d-2}$ ball ($=\mathbb{S}^{d-1}$ projected to plane)
  $Y(i)$ random $\in \mathcal{N}(0,1)$
end for
for $i = 1, \cdots, d-1$ **do** % from $\mathbb{S}^{d-2}$ to cap 'up' in $\mathbb{S}^{d-1}$
  Set $X(i) = \sqrt{1 - \text{rand}^2}\,\dfrac{Y(i)}{|\mathbf{Y}|}$
end for
  Set $\mathbf{X} = (X(1), \cdots, X(d-1), X(d))$, $X(d) = \text{rand}$
  % From $\mathbf{X}$ to $\boldsymbol{\omega}_i^n$ – random orientation in a spherical cap centered at $\mathbf{q}_i^n$
  % Reflect: $\boldsymbol{\omega}_i^n = \mathbb{P}_i^n \mathbf{X}$, $\mathbb{P}_i^n := \mathbb{I} - 2\dfrac{(\mathbf{q}_i^n - \mathbf{z})(\mathbf{q}_i^n - \mathbf{z})^\top}{|\mathbf{q}_i^n - \mathbf{z}|^2}$
if $1 - \mathbf{q}_i^n(d) \neq 0$ **then**
  Set $\mathbf{v}_i^n = \mathbf{q}_i^n - \mathbf{z}$
  Set $\boldsymbol{\omega}_i^n = \mathbf{X} - 2\dfrac{\langle \mathbf{v}_i^n, \mathbf{X}\rangle}{|\mathbf{v}_i^n|^2}\mathbf{v}_i^n$ % Note the simplification: $|\mathbf{v}_i^n|^2 = 2\big(1 - \mathbf{q}_i^n(d)\big)$
else $\boldsymbol{\omega}_i^n = \mathbf{X}$
end if

Why randomization is important?

d \ N	5	10	25	50	100
4	27.2%	50.4%	91.2%	99.2%	100.0%
8	17.2%	53.0%	100.0%	100.0%	100.0%
12	2.4%	26.0%	96.0%	100.0%	100.0%
16	0.0%	0.0%	0.0%	1.0%	2.4%
20	0.0%	0.0%	0.0%	0.0%	0.0%

d \ N	5	10	25	50	100
4	24.0%	38.6%	77.8%	99.8%	100.0%
8	12.2%	26.6%	60.0%	92.8%	100.0%
12	2.2%	11.6%	56.2%	88.8%	99.6%
16	0.0%	0.4%	26.6%	60.2%	88.0%
20	0.0%	0.0%	0.4%	5.8%	21.6%

Table: Success rates of SBGD (top) vs. SBRD (bottom) for D-dimensional Ackley based on $m = 500$ runs. Backtracking parameters $\lambda = 0.2$ and $\gamma = 0.9$

Mass transfer — dependence on $q \gg 1$

d \ N	10		25		50		100	
q	8	4	8	4	8	4	8	4
Ackley								
14	**4.2%**	3.8%	**66.9%**	60.0%	100.0%	100.0%	100.0%	100.0%
16	0.1%	0.3%	38.4%	38.3%	**99.8%**	95.0%	100.0%	100.0%
18	0.0%	0.0%	14.6%	16.3%	**87.3%**	79.7%	100.0%	99.6%
20	0.0%	0.0%	1.0%	1.4%	**30.7%**	25.1%	**84.7%**	74.5%
Rastrigin								
2	**42.5%**	37.4%	99.0%	98.8%	100.0%	100.0%	100.0%	100.0%
3	6.2%	6.5%	**32.4%**	29.0%	**80.1%**	74.3%	**99.1%**	98.0%
4	0.8%	1.0%	4.7%	4.9%	**14.3%**	11.5%	**35.2%**	30.3%
5	0.2%	0.1%	1.1%	0.9%	**3.0%**	1.7%	**4.8%**	3.7%

Table: Success rates of SBRD vs. SBGD for global optimization of the d-dimensional objective functions with mass transfer parameter $q = 4$ and $q = 8$.

- $F_{\text{Rastrigin}}(\mathbf{x}) = \dfrac{1}{d}|\mathbf{x}_B|^2 - \dfrac{10}{d}\sum_i \cos(2\pi(\mathbf{x}_B)_i) + 10 + C, \quad \mathbf{x}_B := \mathbf{x} - B\mathbb{1}$

SBRD in $d = 1976$[11]

- Training on MNIST:
- MNIST-4-4 Architecture (N = batch size):

Layer type	Output shape	#Params.
1-3 Conv2d-1	[N, 4, 14, 14]	16
MaxPool2d-2	[N, 4, 7, 7]	0
Flatten-3	[N, 196]	0
FullyConn-4	[N, 10]	1960

- Loss function $l(\mathbf{x}, \mathbf{y}) = -\frac{1}{N} \sum_{n=0}^{N-1} \sum_{i} y_i \cdot \log(s(x_i)), \quad s(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$

Trial	Optimizer	N	η	h_0	γ	λ	q
1-8 1	SGD	1	0.1	-	-	-	
2	NSGD (multi-particle)	200	0.1	-	-	-	-
3	NGDBT (multi-backtracking)	200	-	1.0	0.25	0.2	-
4	SBRD	200	-	1.0	0.25	0.2	1.0

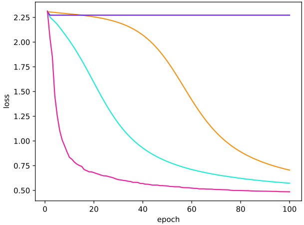


Figure: ——, SGD; ——, NSGD; ——, NGDBT; ——, SBRD

[11] with R. Leonard

SBGD meets Simulated Annealing

- Simulated Annealing (SA) driven by Langevin diffusion:

$$\mathrm{d}\mathbf{x}_i^t = -\nabla F\left(\mathbf{x}_i^t\right)\mathrm{d}t + \sqrt{2\sigma_i^t}\mathrm{d}W_i^t, \qquad i = 1, 2, \ldots, N,$$

⊙ $\{W_i^t\}$ independent Brownian motions, $i = 1, 2, \ldots, N$;

⊙ σ_i^t is the scaled "temperature" or annealing-rate

- Key feature of one-particle SA — protocol for 'cooling down'[12]

  Properly cooling down with annealing rate $\sigma_i^t \propto c/\sqrt{\log t}$

- This is where the masses, $\{m_i^t\}$, come into play:

$$\mathrm{d}m_i^t = -m_i^t\left(F(\mathbf{x}_i^t) - \overline{F}_N^t\right)\mathrm{d}t, \qquad i = 1, 2, \ldots, N,$$

with provisional minimum: $\overline{F}_N^t := \dfrac{\sum_{i=1}^N m_i^t F(\mathbf{x}_i^t)}{\sum_{i=1}^N m_i^t}$

☞ Let the swarm decide how to cool-down $\sigma_i^t = \sigma(m_i^t)\downarrow$

[12] Gidas, ... convergence of the annealing algorithm, JSP (1985)
Geman and Hwang, Diffusions for global optimization. SICOPT (1986)

Why provisional minimum?

- Why provisional minimum? $\overline{F}^t_N := \dfrac{\sum_{i=1}^N m_i^t F(\mathbf{x}_i^t)}{\sum_{i=1}^N m_i^t}$ why not $F(\mathbf{x}^*) = \min\limits_{\mathbf{x}} F(\mathbf{x})$?

☞ Expected that for $t_\epsilon \lesssim \dfrac{1}{\epsilon}$, $N \gtrsim \dfrac{1}{\epsilon^2}$ there holds $\mathbb{E}\left[\overline{F}^t_N - \min\limits_{\mathbf{x}} F(\mathbf{x})\right] \leqslant \epsilon$

- Mean-field: empirical measure $\mu^t_N(\mathbf{x}, m) = \dfrac{1}{N}\sum\limits_{i=1}^N \delta_{\mathbf{x}_i^t}(\mathbf{x}) \otimes \delta_{m_i^t}(m) \ \to \mu_t(\mathbf{x}, m)$:

$$\partial_t \mu = \nabla_{\mathbf{x}} \cdot (\mu \nabla F) + \big(F(\mathbf{x}) - \overline{\mathscr{F}}^t_\mu\big)\partial_m(m\mu) + \sigma(m)\Delta_{\mathbf{x}}\mu \quad \overline{\mathscr{F}}^t_\mu := \frac{\displaystyle\iint mF(\mathbf{x})\mathrm{d}\mu^t(\mathbf{x}, m)}{\displaystyle\iint m\mathrm{d}\mu^t(\mathbf{x}, m)}$$

- Indeed: $\lim\limits_{N\to\infty} \mathbb{E}\left[W_2\left(\mu^t, \mu^t_N\right)\right] = 0, \quad \mathbb{E}\left[\left|\overline{F}^t_N - \overline{\mathscr{F}}^t_\mu\right|\right] < \dfrac{C_t}{\sqrt{N}}$

  and for any $\epsilon > 0$ there exists $t_\epsilon \lesssim \dfrac{1}{\epsilon}$ such that $\overline{\mathscr{F}}^t_\mu < F_* + \epsilon$

Hydrodynamic description

- Density, $\rho = \rho(t, \mathbf{x}) := \int \mu^t(\mathbf{x}, m)\mathrm{d}m$ and Momentum, $\rho \mathrm{v} := \int m\mu^t(\mathbf{x}, m)\mathrm{d}m$
- First two moments of the mean-field

$$\begin{cases} \rho_t + \nabla_{\mathbf{x}} \cdot (\rho \nabla F) = \Delta_{\mathbf{x}} \displaystyle\int \sigma(m)\mu^t(\mathbf{x}, m)\mathrm{d}m, \\ (\rho \mathrm{v})_t + \nabla_{\mathbf{x}} \cdot (\rho \mathrm{v} \nabla F) = \big(\overline{F}(t) - F(\mathbf{x})\big)\rho \mathrm{v} + \Delta_{\mathbf{x}} \displaystyle\int m\sigma(m)\mu^t(\mathbf{x}, m)\mathrm{d}m. \end{cases}$$

Provisional minimum $\overline{F}(t) = \int F(\mathbf{x})\rho \mathrm{v}(t, \mathbf{x})\mathrm{d}\mathbf{x}$ (normalized mass/momentum$= 1$)

☞ Drift-diffusion equation for the velocity v,

$$\rho_t + \nabla_{\mathbf{x}} \cdot (\rho \nabla F) = \Delta_{\mathbf{x}}(\sigma(t, \mathbf{x})\rho(t, \mathbf{x})),$$

$$\mathrm{v}_t + \nabla F \cdot \nabla_{\mathbf{x}} \mathrm{v} = \big(\overline{F}(t) - F(\mathbf{x})\big)\mathrm{v} + \frac{1}{\rho}\Delta_{\mathbf{x}}\big(\sigma(t, \mathbf{x})\rho \mathrm{v}(t, \mathbf{x})\big), \quad \mathbf{x} \in \mathrm{supp}\{\rho(t, \cdot)\}.$$

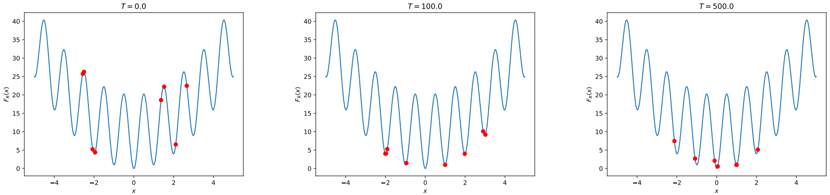
- The decent estimate — expected $\rho(t, \mathbf{x}) \stackrel{t\to\infty}{\longrightarrow} \delta(\mathbf{x} - \mathbf{x}^*)$, $\mathrm{v}(t, \mathbf{x}) \stackrel{t\to\infty}{\longrightarrow} \mathbb{1}(\mathbf{x}^*)$

$$\frac{\mathrm{d}}{\mathrm{d}t}\int (F(\mathbf{x}) - F_*)\rho(t, \mathbf{x})\mathrm{d}\mathbf{x} = -\int |\nabla F(\mathbf{x})|^2\rho(t, \mathbf{x})\mathrm{d}\mathbf{x} + \int \Delta_{\mathbf{x}} F(\mathbf{x})\sigma(t, \mathbf{x})\rho(t, \mathbf{x})\mathrm{d}\mathbf{x},$$

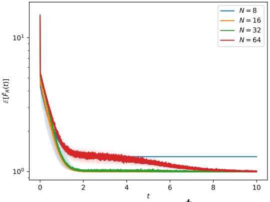
Communication — $\sigma(\cdot) \geqslant 0$ is the key; observe: $\sigma(t, \mathbf{x})$ should vanish as $\mathbf{x} \to \mathbf{x}^*$

$$\sigma(m) = \begin{cases} 2exp(m/(m-1/8)) & m < 1/8 \\ 0, & m \geqslant 1/8 \end{cases} ; \quad N = 8, \ h = 10^{-4}, \ T = 500$$

- Particles eventually find the global minimum after a long time



Convergence of $\overline{F}_N^t$ in finding the global minimum is significantly faster



Mean-field convergence: 20 trajectories of $\overline{F}_N^t$; bigger $N \mapsto$ a lower value of F_*

THANK YOU, Albert