Proximal-based adaptive simulated annealing for global optimization

ICASSP 2022

Emilie Chouzenoux[†], Víctor Elvira[*‡], and Thomas Guilmeau[†]

† Université Paris-Saclay, CentraleSupélec, Inria, CVN, Gif-sur-Yvette, France
∗ School of Mathematics, University of Edinburgh, United Kingdom
‡ The Alan Turing Institute, United Kingdom

Non-convex problems are hard...

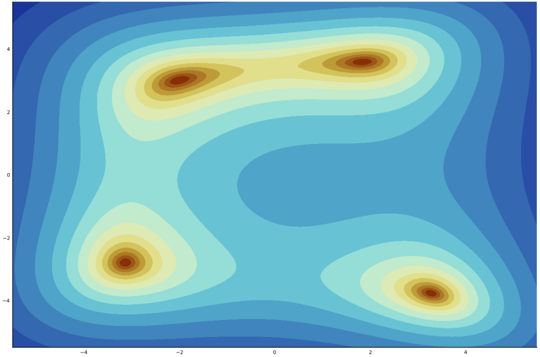


Figure: Level sets of a non-convex function in $\mathbb{R}^2$

Non-convex problems possibly have

- several global minimizers,
- local minima,
- saddle points,
- a combination of the three...

And it is hard to avoid them

Many data science tasks can be formulated as optimization problems

$$\text{Find } x \in \mathcal{X} \text{ s.t. } f(x) = f_* = \min_{x \in \mathcal{X}} f(x).$$

Not all of these problems are convex. It can be because of:

- Sparsity penalty[1],
- Low-rank prior[2],
- Non-linear inverse problems[3]...

[1] A. Marmin et al. "Sparse signal reconstruction for nonlinear models via piecewise rational optimization". In: *Signal Processing* 179 (2021), 107835:1–107835:13.

[2] Y. Chi, Y. M. Lu, and Y. Chen. "Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview". In: *IEEE Transactions on Signal Processing* 62.20 (2019), pp. 5239–5269.

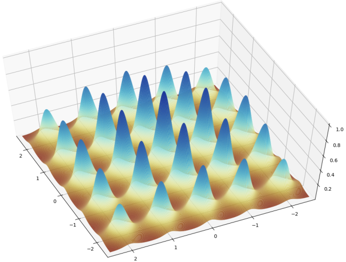
[3] T. Bonesky, D. Lorenz, and P. Maas. "A generalized conditional gradient method for nonlinear operator equations with sparsity constraints". In: *Inverse Problems* 23.5 (2007).

Global minimizers and Boltzmann distributions

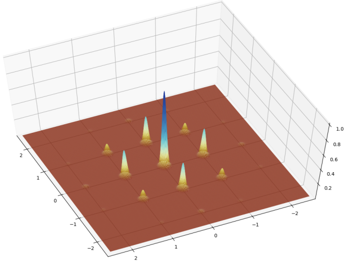
For global optimization, we are interested in exploring two types of sets

- $S_* = \{x \in \mathcal{X}, f(x) = f_*\}$,
- $S_\varepsilon = \{x \in \mathcal{X}, f(x) \leq f_* + \varepsilon\}$.

The Boltzmann distributions π_T concentrate on those sets as the parameter T goes to 0.



Temperatures	**Boltzmann distributions**
T	$\pi_T(x) = \exp\left(-\frac{1}{T} f(x) - B(T)\right)$
$\downarrow$	$\downarrow$
0	$\delta_{S_*}(x)$



$B(T)$ is the log-partition function of π_T.

Approaching the Boltzmann distributions

Boltzmann distributions are intractable:

- Their normalization constants $\int \exp\left(-\frac{1}{T}f(x)\right) dx$ are unknown,
- Generating samples $x \sim \pi_T$ is hard,
- Lower values values of T make it even more challenging!

SA in a nutshell

SA algorithms track a sequence of intractable Boltzmann distributions $\{\pi_{T_k}\}_k$ with $T_k \to 0$ by constructing a sequence of tractable proposal distributions $\{q_k\}_k$.

Most of the times, proposals are constructed by iterating Markov kernels $q_{k+1} = q_k P_k$. In this work, we focus on parametric proposals $q_k = q_{\theta_k}$.

Standard simulated annealing[5] (SA)

Consider the Metropolis-Hastings kernel P_k such that $\pi_{T_k} = \pi_{T_k} P_k$.
Applying this kernel[4] brings us closer to π_{T_k}: qP_k is closer to π_{T_k} than q.
Simulated annealing consists in tracking π_{T_k} with $q_k = q_0 P_1 \cdots P_{k-1} P_k$.

Convergence of SA

If $T_k = \frac{C(f)}{\log(k+1)}$, then $||\pi_{T_k} - q_k||_{TV} \longrightarrow 0$.

- ***TV* convergence implies convergence to the set S_ε for any $\varepsilon > 0$,**
- **The logarithmic schedule is often considered too slow.**

[4] G. O. Roberts and J. S. Rosenthal. "General State Space Markov Chain and MCMC algorithms". In: *Probability Surveys* 1 (2004), pp. 20–71.

[5] H. Haario and E. Saksman. "Simulated Annealing Process in General State Space". In: *Advances in Applied Probability* 23.4 (1991), pp. 866–893.

Parametric proposals

$\mathcal{Q} = \{q_\theta,\ \theta \in \Theta\}$ is an exponential family if

$$q_\theta(x) = \exp\left(\langle \theta, \Gamma(x) \rangle - A(\theta)\right), \quad \forall x \in \mathcal{X}, \theta \in \Theta, \tag{1}$$

with A being the log-partition function.
Many classes of distributions are exponential families:

- Gaussian distributions, with $\theta = (\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1})^\top$ and $\Gamma(x) = (x, xx^\top)^\top$,
- Boltzmann distributions, with parameter $\theta = \frac{1}{T}$ and $\Gamma(x) = -f(x)$...

Moment-matching optimality conditions

$$\theta^* = \underset{\theta \in \Theta}{\arg\min}\, KL(\pi, q_\theta) \iff q_{\theta^*}(\Gamma) = \pi(\Gamma). \tag{2}$$

Model annealing random search[6] (MARS)

The MARS algorithm relies on this framework of successive minimizations:

1. Design an intermediate target $\hat{\pi}_{k+1} = \alpha_{k+1}\pi_{T_{k+1}} + (1-\alpha_k)q_{\theta_k}$
2. Approach $q_{\theta_{k+1}}(\Gamma) = \hat{\pi}_{k+1}(\Gamma)$ with importance sampling (N_k samples).

MARS convergence guarantees

The convergence $q_{\theta_k}(\Gamma) \to \delta_{S_*}(\Gamma)$ is guaranteed if $\lambda_k = k^{-\gamma}$, $\sum_k \alpha_k = +\infty$, $\sum_k \alpha_k^2 > +\infty$ and either $T_k = \frac{T_0}{\log(k+1)}$ and $N_k = N_0 k^\beta$ or $T_k = \frac{T_0}{1+ck}$ and $N_k = N_0\beta^k$.

- **Logarithmic cooling schedule, with polynomially increasing number of samples,**
- **Linear cooling schedule, with exponentially increasing number of samples.**

[6] J. Hu and P. Hu. "Annealing adaptive search, cross-entropy, and stochastic approximation in global optimization". In: *Naval Research Logistics* 58.5 (2011).

Proposed adaptive cooling schedule

What is expected from an adaptive cooling schedule?

- If the proposal is a good fit, the schedule should speed up. Else, it should slow down.
- Temperature decrease should be promoted but stopping at $T > 0$ must be possible.

Boltzmann distributions are exponential, so why not adapt T as well as θ ?

Variational formulation

Our approach is to solve

$$\underset{T>0,\, \theta\in\Theta}{\text{minimize}}\, KL(\pi_T, q_\theta) + \lambda R(T), \tag{3}$$

with an alternating Bregman proximal algorithm.

Alternating proximal simulated annealing (APSA)

We propose to minimize the quantity $F_\lambda : (T, \theta) \longmapsto KL(\pi_T, q_\theta) + \lambda R(T)$ by alternating Bregman proximal steps[7] that reads like

$$\theta_k = \overleftarrow{\text{prox}}^A_{\rho^{-1} F_\lambda(T_k, \cdot)}(\theta_{k-1}) = \underset{\theta \in \Theta}{\arg\min}\,(KL(\pi_{T_k}, q_\theta) + \lambda R(T_k)) + \rho KL(q_{\theta_{k-1}}, q_\theta), \quad (4)$$

$$T_{k+1} = \overrightarrow{\text{prox}}^B_{\rho^{-1} F_\lambda(\cdot, \theta_k)}(T_k) = \underset{T > 0}{\arg\min}\,(KL(\pi_T, q_{\theta_k}) + \lambda R(T)) + \rho KL(\pi_T, \pi_{T_k}). \quad (5)$$

A decrease property

For every $k \in \mathbb{N}$, we have

$$KL(\pi_{T_{k+1}}, q_{\theta_{k+1}}) + \lambda R(T_{k+1}) \leq KL(\pi_{T_k}, q_{\theta_k}) + \lambda R(T_k). \quad (6)$$

[7] H. Bauschke, P. Combettes, and D. Noll. "Joint minimization with alternating Bregman proximity operators". In: *Pacific Journal of Optimization* 2 (2006).

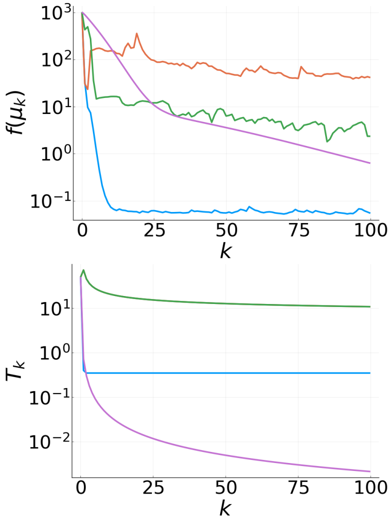
Numerical experiments

We use the Rosenbrock function in $\mathbb{R}^2$ as a benchmark to compare

- MARS (orange),
- mFSA (purple),
- SMC-SA (green),
- APSA (blue).

We used Gaussian proposals indexed by (μ, Σ) for MARS and APSA. For mFSA and SMC-SA, we used $f(\mu_k) = \frac{1}{N_k}\sum_{i=1}^{N_k} f(x_k^i)$.

- **APSA finds the best values of μ,**
- **The cooling stops before reaching 0.**



Conclusion

What we saw in this talk:

- **We proposed a variational formulation of adaptive simulated annealing,**
- **The resulting scheme alternatively adapts a parametric proposal and a temperature,**
- **It is able to reach good values of the objective very fast,**
- **But further understanding of its convergence is still needed!**

Thank you for your attention!

References

Akyildiz, O. and J. Míguez. "Convergence rates for optimised adaptive importance samplers". In: *Statistic and Computing* 31.12 (2021).

Andrieu, C., L. A. Breyer, and A. Doucet. "Convergence of simulated annealing using Foster-Lyapunov criteria". In: *Journal of Applied Probability* 38.4 (2001), pp. 975–994.

Bauschke, H., P. Combettes, and D. Noll. "Joint minimization with alternating Bregman proximity operators". In: *Pacific Journal of Optimization* 2 (2006).

Bezanson, J. et al. "Julia: A Fresh Approach to Numerical Computing". In: *SIAM Review* 59.1 (2017), pp. 65–98.

Bonesky, T., D. Lorenz, and P. Maas. "A generalized conditional gradient method for nonlinear operator equations with sparsity constraints". In: *Inverse Problems* 23.5 (2007).

Černý, V. "Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm". In: *Journal of Optimization Theory and Applications* 45.1 (1985), pp. 41–51.

Chan, T. F., S. Esedoglu, and M. Nikolova. "Algorithms for Finding Global Minimizers of Image Segmentation and Denoising Models". In: *SIAM Journal of Applied Mathematics* 66.5 (2006), pp. 1632–1648.

Chi, Y., Y. M. Lu, and Y. Chen. "Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview". In: *IEEE Transactions on Signal Processing* 62.20 (2019), pp. 5239–5269.

Chopin, N. and O. Papaspilopoulos. *An Introduction to Sequential Monte Carlo*. Springer, 2020.

Dekkers, A. and E. Aarts. "Global optimization and simulated annealing". In: *Mathematical Programming* 50.3 (1991), pp. 367–393.

Gielis, G. and C. Maes. "A simple approach to time-inhomogeneous dynamics and applications to (fast) simulated annealing". In: *Journal of Physics A: Mathematical and General* 32.29 (1999), pp. 5389–5407.

References (cont.)

Guilmeau, T., E. Chouzenoux, and V. Elvira. “Simulated Annealing: a Review and a New Scheme”. In: *2021 IEEE Statistical Signal Processing Workshop (SSP)*. 2021, pp. 101–105.

Haario, H. and E. Saksman. “Simulated Annealing Process in General State Space”. In: *Advances in Applied Probability* 23.4 (1991), pp. 866–893.

Haeffele, B. D. and R. Vidal. “Global Optimality in Neural Network Training”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 4390–4398.

Hu, J. and P. Hu. “Annealing adaptive search, cross-entropy, and stochastic approximation in global optimization”. In: *Naval Research Logistics* 58.5 (2011).

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. “Optimization by Simulated Annealing”. In: *Science* 220.4598 (1983), pp. 671–680.

Marmin, A. et al. “Sparse signal reconstruction for nonlinear models via piecewise rational optimization”. In: *Signal Processing* 179 (2021), 107835:1–107835:13.

Onbaşoğlu, E. and L. Özdamar. “Parallel Simulated Annealing Algorithms in Global Optimization”. In: *Journal Of Global Optimization* 19.1 (2001), pp. 27–50.

Roberts, G. O. and J. S. Rosenthal. “General State Space Markov Chain and MCMC algorithms”. In: *Probability Surveys* 1 (2004), pp. 20–71.

Rubenthaler, S., T. Rydén, and M. Wiktorsson. “Fast simulated annealing in $\mathbb{R}^d$ with an application to maximum likelihood estimation in state-space models”. In: *Stochastic Processes and their Applications* 119.6 (2009), pp. 1912–1931.

References (cont.)

Zhou, E. and X. Chen. "Sequential Monte Carlo simulated annealing". In: *Journal of Global Optimization* 55 (2013), pp. 101–124.