PROXIMAL-BASED ADAPTIVE SIMULATED ANNEALING FOR GLOBAL OPTIMIZATION

Thomas Guilmeau†*, Emilie Chouzenoux*†*, and Víctor Elvira*∗‡

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

ABSTRACT

Simulated annealing (SA) is a widely used approach to solve global optimization problems in signal processing. The initial non-convex problem is recast as the exploration of a sequence of Boltzmann probability distributions, which are increasingly harder to sample from. They are parametrized by a temperature that is iteratively decreased, following the so-called cooling schedule. Convergence results of SA methods usually require the cooling schedule to be set a priori with slow decay. In this work, we introduce a new SA approach that selects the cooling schedule on the fly. To do so, each Boltzmann distribution is approximated by a proposal density, which is also sequentially adapted. Starting from a variational formulation of the problem of joint temperature and proposal adaptation, we derive an alternating Bregman proximal algorithm to minimize the resulting cost, obtaining the sequence of Boltzmann distributions and proposals. Numerical experiments in an idealized setting illustrate the potential of our method compared with state-of-the-art SA algorithms.

***Index Terms*—** Global optimization, simulated annealing, adaptive cooling schedule, Kullback-Leibler divergence, alternating minimization.

1. INTRODUCTION

Many problems ranging from signal processing to machine learning require to find the global minimizer of non-convex functions. Non-convexity can arise due to sparsity-inducing regularizers [1], non-linear activation functions [2], or discrete domain constraints [3], among other factors. Standard optimization methods converge at best to local minima which may be of limited interest, so specific space exploration strategies are needed to reach global optimality. For instance, branch & bound methods rely on subspaces that are retained or discarded [4]. In particle swarm optimization methods, interacting particles are simulated for efficient exploration [5].

In this work, we focus on the class of SA algorithms, which explore the space by sampling from the Boltzmann distributions associated to the objective function. These distributions, indexed by a temperature T, have their modes located at the global minimizers of the objective. When T is high, they are flat, hence easy to explore by sampling. In contrast, they become increasingly concentrated around their modes, hence more informative, as T decays. By following a cooling schedule, i.e. a sequence of temperatures $\{T_k\}_{k\in\mathbb{N}}$ going to zero, the optimization problem is thus recast as a sequence of increasingly harder sampling problems [6]. This progressive strategy allows to benefit from the early iterations at high T to sample from the last harder distributions. Samples approaching the Boltzmann distributions can be generated by iterating Markov kernels [7], or with parametric proposal distributions [8]. In this work, we adopt the latter approach.

SA algorithms performance requires designing good proposals and cooling schedule. On the one hand, the construction of parametric proposals for sampling tasks has been well-explored. A typical approach is to minimize a divergence between the proposal and the target [9, 8, 10, 11]. On the other hand, the adaptation of cooling schedules has been underlooked in the literature. In fact, many convergence results for SA require a logarithmic cooling schedule [7, 8, 12]. Recent convergence results account for faster schedules [13]. However, fixed cooling schedules backed by theoretical guarantees often involve constants which are either intractable [7] or empirically chosen [8, 12]. One exception is [14], where temperatures are set by solving non-linear optimal control problems.

The main contribution of this work is a novel SA framework integrating an iterative alternating approach for both temperature and parametric proposal on-the-fly adaptation. We formulate the adaptation problem as the minimization of a loss function involving the Kullback-Leibler (KL) divergence between the Boltzmann distribution and the proposal, as well as a regularization term promoting temperature decay. We introduce an alternating Bregman proximal algorithm [15] to solve the resulting optimization problem. Suitable sampling strategies are considered to solve the inner problems.

The paper is structured as follows. In Section 2, we present the global optimization problem at hand, and introduce the necessary background on SA and Bregman proximal tools. In Section 3, we describe our variational formulation for adaptive SA as well as an alternating proximal approach to resolve it. We conduct some numerical experiments in Section 4, before concluding in Section 5.

T.G. and E.C. acknowledge support from the ERC Starting Grant MAJORIS ERC-2019-STG-850925. V.E. acknowledges support from the ANR of France under PISCES (ANR-17-CE40-0031-01) project.

2. BACKGROUND

2.1. Problem statement and notation

In this work, we are interested in finding f_* such that

$$f_* = \min_{x \in \mathcal{X}} f(x), \tag{1}$$

where $\mathcal{X} \subset \mathbb{R}^d$ is the search space, and $f : \mathcal{X} \to \mathbb{R}$ is the objective function. We assume that $\mathcal{X}$ and f are such that Problem (1) is well-defined. In the remainder, the euclidean scalar product is denoted by $\langle \cdot, \cdot \rangle$ and $|| \cdot ||$ is the euclidean norm. The Borel algebra of $\mathcal{X}$ is denoted by $\mathcal{B}(\mathcal{X})$. $\mathcal{M}(\mathcal{X})$ is the set of probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. For a given $p \in \mathcal{M}(\mathcal{X})$, a measurable function h and a set $S \subset \mathcal{X}$, we will write $p(h) = \int_{\mathcal{X}} h(x)p(dx)$ and $p(S) = \int_S p(dx)$. We refer to [16] for further details on measure theory concepts.

2.2. Boltzmann distributions and SA

SA algorithms rely on the Boltzmann distributions, denoted by π_T for $T > 0$. They relate the probability of $x \in \mathcal{X}$, its energy $f(x)$ and the temperature T through

$$\pi_T(x) = \exp\left(-\frac{1}{T} f(x) - B(T)\right), \tag{2}$$

where B is the log-partition function of π_T, that is

$$B(T) = \log\left(\int \exp\left(-\frac{1}{T} f(x)\right) dx\right), \quad \forall T > 0. \tag{3}$$

Boltzmann distributions are interesting from a global optimization perspective, as their mass concentrates on the minimizers of f as $T \to 0$ [7, Eq. (5.13)], i.e,

$$\lim_{T \to 0} \pi_T(S_\epsilon) = 1, \quad \forall \epsilon > 0, \tag{4}$$

where $S_\epsilon = \{x \in \mathcal{X},\ f(x) \leq f_* + \epsilon\}$ for any $\epsilon > 0$.

However, Boltzmann distributions are usually intractable. SA algorithms amount to exploring a sequence $\{\pi_{T_k}\}_{k \in \mathbb{N}}$ associated to a cooling schedule $\{T_k\}_{k \in \mathbb{N}}$, with $T_k \to 0$. For each $k \in \mathbb{N}$, a proposal q_k is used to generate samples approximating π_{T_k}. Proposals are chosen such that q_k gets closer to π_{T_k} as $k \to \infty$. They can be constructed from Markov kernels $\{P_k\}_{k \in \mathbb{N}}$, i.e, $q_k = q_{k-1}P_k$ [7]. Alternatively, they can be of the form q_{θ_k}, parametrized by some θ_k [8].

2.3. Kullback-Leibler divergence and proximal operators

Let us now introduce two mathematical tools that will be at the core of our proposed adaptive SA algorithm. First, the KL divergence is widely used to measure the discrepancy between probability distributions and thus to fit proposals to targets [8, 9]. It reads

$$KL(\pi, q) = \int \log\left(\frac{\pi(x)}{q(x)}\right) \pi(dx), \quad \forall \pi, q \in \mathcal{M}(\mathcal{X}). \tag{5}$$

We then introduce the concept of Bregman proximity operator. Given a Bregman divergence d_ψ with associated function ψ [17], and a convex lower-semicontinuous function h, we define, following [15], the two Bregman proximity operators

$$\overleftarrow{\text{prox}}_h^\psi(\theta) := \arg\min_{\theta' \in \Theta} h(\theta') + d_\psi(\theta', \theta), \tag{6}$$

$$\overrightarrow{\text{prox}}_h^\psi(\theta) := \arg\min_{\theta' \in \Theta} h(\theta') + d_\psi(\theta, \theta'). \tag{7}$$

Well-posedness of the above definition can be ensured by [15, Lemma 2.1, Proposition 3.5]. It is worth noting that, when $\psi(\cdot) = ||\cdot||^2$, the euclidean distance and standard proximity operator are recovered. In this work, we will instead focus on another choice for ψ, so that d_ψ reads as the KL divergence between two distributions of interest.

3. PROPOSED ANNEALING FRAMEWORK

3.1. Proposed formulation

We first propose to use parametric proposal distributions from an exponential family, that is with density

$$q_\theta(x) = \exp\left(\langle \theta, \Gamma(x) \rangle - A(\theta)\right). \tag{8}$$

The above is parametrized by $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$. Exponential families are a standard choice for proposal distributions in sampling [11, 8]. The choice of sufficient statistics $\Gamma : \mathcal{X} \to \mathbb{R}^{d_\theta}$ determines the family. If $\Gamma(x) = (x, xx^\top)^\top$, the Gaussian family is recovered, while the Boltzmann family is a particular case with $\Gamma(x) = -f(x)$. Eq. (8) also involves the log-partition function

$$A(\theta) = \log\left(\int \exp\left(\langle \theta, \Gamma(x) \rangle\right) dx\right), \quad \forall \theta \in \Theta. \tag{9}$$

We aim at constructing efficiently $\{\theta_k, T_k\}_{k \in \mathbb{N}}$ to minimize f. To do so, we introduce the loss function

$$F_\lambda(T, \theta) = KL(\pi_T, q_\theta) + \lambda R(T), \quad \forall T > 0, \theta \in \Theta. \tag{10}$$

Hereabove, the first term measures the discrepancy between the proposal and the sought Boltzmann distribution. Moreover, $R : \mathbb{R}_+ \to \mathbb{R}_+$ is a regularization term, weighted by some $\lambda > 0$, that promotes the decay of $\{T_k\}_{k \in \mathbb{N}}$. In practice, R must be increasing and null at $T = 0$. The minimization of F_λ hence promotes a good fitting of the Boltzmann distributions by the proposals and a low final temperature.

3.2. Proposed alternating Bregman proximal algorithm

Let us now present our approach for minimizing the proposed loss function F_λ. We introduce the shorter notations

$$H(T) = \int \log(\pi_T(x))\pi_T(x)dx, \quad \forall T > 0, \tag{11}$$

$$(\pi_T(\Gamma))_i = \int (\Gamma(x))_i \, \pi_T(x)dx, \quad \forall T > 0, \forall i \in \{1, ..., d_\theta\}. \tag{12}$$

We can thus rewrite in a more explicit way

$$F_\lambda(T,\theta) = H(T) + \lambda R(T) + \langle \theta, \pi_T(\Gamma)\rangle + A(\theta), \quad \forall T>0, \theta\in\Theta. \quad (13)$$

We see in (13) that T and θ are coupled only in one term of the loss function, with a linear dependency on θ, thus motivating an alternating procedure. We opt for proximal alternating methods, where one alternatively computes the proximity operator of the loss function with respect to each of the variables. This leads to more stable convergence behavior [18, 19] than the standard Gauss-Seidel technique, especially in the challenging non-convex setting.

We now choose the metric in which the proximity operator is computed. The KL divergence appears in F_λ, so it is natural to use it in our proximal steps. Since both Boltzmann and the parametric proposal distributions are exponential, we can link the Bregman and the KL divergences as follows [17]

$$d_B(T',T) = KL(\pi_T, \pi_{T'}), \quad \forall T,T'>0, \quad (14)$$
$$d_A(\theta,\theta') = KL(q_{\theta'}, q_\theta), \quad \forall \theta,\theta'\in\Theta, \quad (15)$$

where B and A are defined in Eq. (3) and (9), respectively.

These choices lead to Alg. 1, with two positive hyperparameters, λ and ρ. They control respectively the trade-off between adaptation and low-temperature, and the inertia between iterates.

Algorithm 1: Alternating proximal SA (APSA)

Initialization with $T_0>0$, $\theta_0\in\Theta$.
for $k=1,2,\ldots$ **do**
  Temperature adaptation:
  $$T_{k+1} = \overrightarrow{\text{prox}}^{B}_{\rho^{-1}F_\lambda(\cdot,\theta_k)}(T_k). \quad (16)$$
  Proposal adaptation:
  $$\theta_{k+1} = \overleftarrow{\text{prox}}^{A}_{\rho^{-1}F_\lambda(T_{k+1},\cdot)}(\theta_k). \quad (17)$$
end

As an alternating proximal algorithm, Alg. 1 enjoys a monotonicity property akin to [15, Proposition 4.1]. Namely,

$$F_\lambda(T_k,\theta_k) \geq F_\lambda(T_{k+1},\theta_{k+1}), \quad \forall k\in\mathbb{N}. \quad (18)$$

Since the loss takes non-negative values, we can deduce convergence of $\{F_\lambda(T_k,\theta_k)\}_{k\in\mathbb{N}}$ to some non-negative value.

3.3. Temperature adaptation

We discuss the practical resolution of the inner Problem (16). Eq. (16) is a scalar but non-convex minimization problem. We propose to approximate its solution by an intensive grid search, restricting the values of T to a grid of the form $\{T^{(i)}\}_{1\leq i\leq N_T}$, with $T^{(i)} = ih$, with $h = \frac{T_{N_T}}{N_T}$. This strategy, though basic, presents the great advantage of allowing the precomputation of the values $H(T^{(i)})$, $B(T^{(i)})$, $\pi_{T^{(i)}}(f)$, and $\pi_{T^{(i)}}(\Gamma)$ for every $1\leq i\leq N_T$. From these, we can compute $F_\lambda(T^{(i)},\theta)$ using Eq. (13) and $KL(\pi_{T^{(i)}},\pi_{T^{(j)}}) = H(T^{(i)}) + \frac{1}{T^{(j)}}\pi_{T^{(i)}}(f) + B(T^{(j)})$, for every $\theta\in\Theta$ and $i,j\in\{1,\ldots,N_T\}$.

Although these computations are not realistic in high dimensions, they allow to solve Eq. (16) with high precision, thus yielding a proof-of-concept implementation of Alg. 1.

3.4. Proposal adaptation

We focus now on Eq. (17), which reads as the minimization of $\theta \longmapsto KL(\pi_{T_{k+1}}, q_\theta) + \rho KL(q_{\theta_k}, q_\theta)$ on Θ. Since we are manipulating exponential proposals, we can rewrite

$$KL(\pi_T, q_\theta) = H(T) - \langle \theta, \pi_T(\Gamma)\rangle + A(\theta), \quad \forall T>0,\ \theta\in\Theta. \quad (19)$$

Function A is convex [11], analytic on Θ [20, Theorem 2.2], and its gradient is $\nabla A(\theta) = q_\theta(\Gamma)$. Therefore, solving (17) is equivalent to solving $-\pi_{T_k}(\Gamma) + q_\theta(\Gamma) + \rho(-q_{\theta_k}(\Gamma) + q_{\theta_{k+1}}(\Gamma)) = 0$. Thus, Eq. (17) admits the following explicit solution:

$$q_{\theta_{k+1}}(\Gamma) = \frac{1}{1+\rho}\pi_{T_{k+1}}(\Gamma) + \frac{\rho}{1+\rho}q_{\theta_k}(\Gamma). \quad (20)$$

In this update, only $q_{\theta_k}(\Gamma)$ is accessible. We propose to approximate $\pi_{T_{k+1}}(\Gamma)$ using importance sampling [21], as it was done in [8]. Consider N samples $\{x_n\}_{1\leq n\leq N}$ drawn from q_{θ_k}, then $\pi_{T_{k+1}}(\Gamma) \approx \sum_{n=1}^N \overline{w}_n\Gamma(x_n)$, with normalized weights $\overline{w}_n$ obtained from $w_n = \frac{\pi_{T_{k+1}}(x_n)}{q_{\theta_k}(x_n)}$, amounting to N evaluations of f per iteration.

Note that the precomputed values $\pi_{T^{(i)}}(\Gamma)$ could also be used at this stage. However, we consider that we can gain better insights on realistic implementations of Alg. 1 by making use of a sampling step to evaluate (20).

3.5. Discussion

At each iteration, the proposed Alg. 1 fits an exponential proposal to the current Boltzmann distribution. This proposal adaptation strategy is rather common when the target is fixed, as in [11]. In the context of SA, our approach stems from the MARS framework of [8]. The novelty here is the introduction of the temperature adaptation step and the use of Bregamn proximity steps. We can also cite the cross-entropy method [9] for global optimization, which also adapts parametric proposals by KL divergence minimization. In this method however, the targets are constructed by truncating the proposals to keep the areas with the best values of f.

In a broader context, expectation-minimization algorithms for statistical inference can also be reformulated using alternating divergence minimization [22, 23].

4. NUMERICAL EXPERIMENTS

4.1. Considered examples and setting

We consider two benchmark problems in $\mathbb{R}^2$, whose minimum is $f_* = 0$. Problem (P_1) relies on the ill-conditioned Rosenbrock function with a unique minimizer at $x_* = (1,1)^T$ located in a large banana-shaped valley. Problem (P_2) aims at minimizing the highly multimodal Rastrigin function, minimized at $x_* = 0$. Their respective objective functions are

$$f_1(x) = 5(x_2 - x_1^2)^2 + (1 - x_1)^2, \quad \forall x \in \mathbb{R}^2, \tag{21}$$

$$f_2(x) = 2 + \sum_{i=1}^{2} x_i^2 - \cos(2\pi x_i), \quad \forall x \in \mathbb{R}^2. \tag{22}$$

We use the regularization function $R(T) = T^2$, and Gaussian proposals with parameters μ and Σ. We evaluate $\{f(\mu_k)\}_{k\in\mathbb{N}}$ to outline the performance of the SA algorithms. We also record $\{T_k\}_{k\in\mathbb{N}}$ to understand the temperature adaptation behavior of Alg. 1. We set $N_T = 1000$, $T_{N_T} = 50$, and $N = 10^5$. Precomputations were done in Julia [24] with the numerical integration package `hcubature`. Algorithms were initialized with $T_0 = T_{N_T}$, $\Sigma_0 = 10\,\mathrm{Id}$ and random $\mu_0 \in \mathbb{R}^2$.

4.2. Influence of APSA hyper-parameters

We display in Fig. 1 the evolution of $\{f(\mu_k)\}_{k\in\mathbb{N}}$ and $\{T_k\}_{k\in\mathbb{N}}$ for different λ and ρ, on both problems.

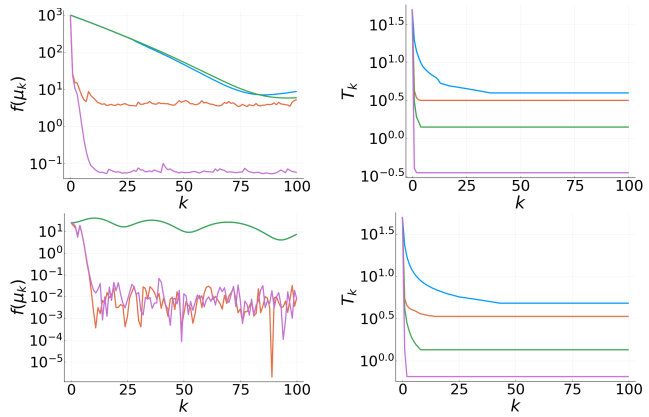

Fig. 1. Influence of the parameters of APSA for solving (P_1) (top) and (P_2) (bottom): $\lambda = 0.05, \rho = 100.0$ (blue), $\lambda = 0.05, \rho = 1.0$ (red), $\lambda = 5.0, \rho = 100.0$ (green), $\lambda = 5.0, \rho = 1.0$ (purple).

In both examples, high values of ρ seem to slow the iterates down, while high values of λ encourage low values of T. This influence is clear if we inspect $\{T_k\}_{k\in\mathbb{N}}$. Regarding $\{f(\mu_k)\}_{k\in\mathbb{N}}$, we can see that the two plots with $\rho = 100.0$ are very close. It is also the case on (P_2) for $\rho = 1.0$. The role of λ is important as can be seen on Fig. 1 (top-left), where APSA reaches small values for only one combination of λ, ρ.

4.3. Comparison with other algorithms

We now compare our approach to three other SA algorithms with fixed schedules, among which MARS is the only one using parametric proposals: the SMCSA algorithm [12], where N particles interact through weighting and resampling, with logarithmic cooling schedule $T_k = \frac{T_0}{\log(k+1)}$ for $k \geq 1$, the MARS algorithm of [8], which uses Gaussian proposals with logarithmic cooling schedule, and the multi-start fast SA (mFSA), which consists in running N parallel fast SA algorithms [13] with schedule $T_k = \frac{1}{(k+1)\log(k+1)}$ for $k \geq 1$.

Among the parameters tested in Fig. 1, we retain $\lambda = 5$, $\rho = 1$. For SMCSA and mFSA, the values $\{f(\mu_k)\}_{k\in\mathbb{N}}$ are the averages of the sampled objective values at iteration k.

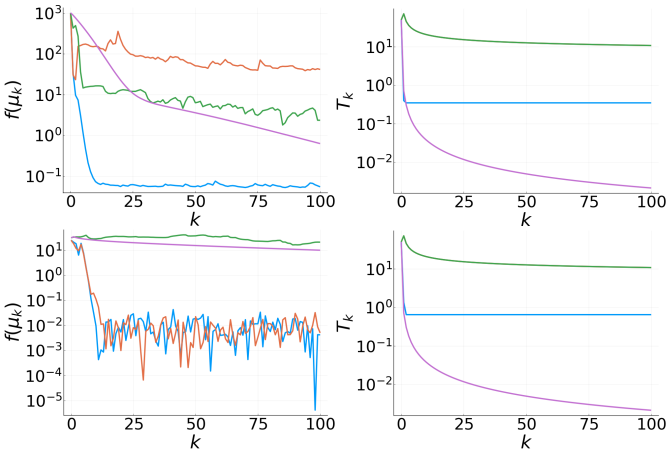

Fig. 2. One run of the algorithms on (P_1) (top) and (P_2) (bottom): MARS (red), APSA (blue), mFSA (purple) and SMCSA (green).

On (P_1), the APSA algorithm shows the best performance, with a very fast convergence. On (P_2), the performances of APSA and MARS are indistinguishable, and are clearly better than those of mFSA and SMCSA.

The performance of APSA are up-to-par with the other tested algorithms. Surprisingly, the adapted temperatures do not reach $T = 0$. This may indicate that reaching a low final temperature is enough to globally minimize f. Moreover, the final temperature of APSA is actually reached very fast, showing that adaptative schedules can be faster.

5. CONCLUSION

In this work, we have proposed a variational formulation associated to an alternating proximal framework for the design of a simulated annealing algorithm with adaptive cooling schedule. This is in stark contrast with existing methods, which need a fixed cooling schedule, often too slow. As a proof of concept, we matched state-of-the-art SA algorithms performance with an idealized implementation. We plan on leveraging these promising insights to design a more practical adaptive SA algorithm with sound convergence analysis.

6. REFERENCES

[1] A. Marmin, M. Castella, J.-C. Pesquet, and L. Duval, "Sparse signal reconstruction for nonlinear models via piecewise rational optimization," *Signal Processing*, vol. 179, pp. 107835:1–107835:13, 2021.

[2] B. D. Haeffele and R. Vidal, "Global optimality in neural network training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, 2017, pp. 4390–4398.

[3] T. F. Chan, S. Esedoglu, and M. Nikolova, "Algorithms for finding global minimizers of image segmentation and denoising models," *SIAM Journal of Applied Mathematics*, vol. 66, no. 5, pp. 1632–1648, 2006.

[4] J. M. Fowkes, N. I. M. Gould, and C. L. Farmer, "A branch and bound algorithm for the global optimization of hessian lipschitz continuous functions," *Journal of Global Optimization*, vol. 56, pp. 1791–1815, 2013.

[5] M. R. Bonyadi and Z. Michalewicz, "Particle swarm optimization for single objective continuous space problems: a review," *Evolutionary Computation*, vol. 25, no. 1, pp. 1–54, 2017.

[6] P. Del Moral, A. Doucet, and A. Jasra, "Sequential Monte Carlo samplers," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 68, no. 3, pp. 411–436, 2006.

[7] H. Haario and E. Saksman, "Simulated annealing process in general state space," *Advances in Applied Probability*, vol. 23, no. 4, pp. 866–893, 1991.

[8] J. Hu and P. Hu, "Annealing adaptive search, cross-entropy, and stochastic approximation in global optimization," *Naval Research Logistics*, vol. 58, no. 5, pp. 457–477, 2011.

[9] D. Kroese, S. Porotsky, and R. Rubinstein, "The cross-entropy method for continuous multi-extremal optimization," *Methodology and Computing in Applied Probability*, vol. 8, pp. 383–407, 2006.

[10] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Míguez, and P. M. Djuric, "Adaptive importance sampling: The past, the present, and the future," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 60–79, 2017.

[11] O. Akyildiz and J. Míguez, "Convergence rates for optimised adaptive importance samplers," *Statistics and Computing*, vol. 31, no. 12, 2021.

[12] E. Zhou and X. Chen, "Sequential Monte Carlo simulated annealing," *Journal of Global Optimization*, vol. 55, pp. 101–124, 2013.

[13] S. Rubenthaler, T. Rydén, and M. Wiktorsson, "Fast simulated annealing in $\mathbb{R}^d$ with an application to maximum likelihood estimation in state-space models," *Stochastic Processes and their Applications*, vol. 119, no. 6, pp. 1912–1931, 2009.

[14] O. Molvalioglu and Z. Zabinsky, "Meta-control of an interacting-particle algorithm for global optimization," *Non-linear Analysis : Hybrid Systems*, vol. 4, pp. 659–671, 2010.

[15] H. Bauschke, P. L. Combettes, and D. Noll, "Joint minimization with alternating Bregman proximity operators," *Pacific Journal of Optimization*, vol. 2, 2006.

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

[17] F. Nielsen and R. Nock, "Entropies and cross-entropies of exponential families," in *Proceedings of 17th IEEE International Conference on Image Processing (ICIP 2010)*, 2010, pp. 3621–3624.

[18] J. Bolte, P.L. Combettes, and J.-C. Pesquet, "Alternating proximal algorithm for blind image recovery," in *Proceedings of 17th IEEE International Conference on Image Processing (ICIP 2010)*, 2010, pp. 1673–1676.

[19] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization methods for non-convex problems: an approach based on the Kurdyka-Lojasiewicz inequality," *Mathematics of Operations Research*, vol. 35, no. 2, pp. 438–457, 2010.

[20] L. D. Brown, *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*, Institute of Mathematical Statistics, 1986.

[21] V. Elvira and L. Martino, "Advances in importance sampling," *Wiley StatsRef: Statistics Reference Online*, pp. 1–22, 2021.

[22] S. Amari, "Information geometry of the EM and em algorithms for neural networks," *Neural Networks*, vol. 8, no. 9, pp. 1379–1408, 1995.

[23] Y. Fujimoto and N . Murata, "A modified EM algorithm for mixture models based on Bregman divergence," *Annals of the Institute of Statistical Mathematics*, vol. 59, pp. 3–25, 2007.

[24] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A fresh approach to numerical computing," *SIAM Review*, vol. 59, pp. 65–98, 2017.