ADAPTIVE SIMULATED ANNEALING THROUGH ALTERNATING RÉNYI DIVERGENCE MINIMIZATION

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 is a popular approach to solve non-convex and black-box optimization problems. It consists in running a non-homogeneous Markov chain to sample from a sequence of Boltzmann probability distributions. This sequence is controlled by a cooling schedule, which governs the concentration of the mass of the Boltzmann distributions around the global minimizers. However, convergence is often slow, difficult to assess, and requires a fixed cooling schedule. We propose here a new simulated annealing algorithm with adaptive cooling schedule, which draws samples from variational approximations of the Boltzmann distributions. Our approach is theoretically sound and relies on an alternating Bregman proximal-gradient scheme minimizing a regularized Rényi divergence. Numerical experiments illustrate the performance of the method.

Index Terms— Adaptive simulated annealing, alternating Bregman proximal-gradient algorithm, Rényi divergence.

1. INTRODUCTION

In this work, we focus on the black-box minimization of a non-convex cost. In this setting, only the cost function can be evaluated, without access to any extra information (e.g. gradients). This can occur while optimizing hyper-parameters [1] or when dealing with categorical variables [2]. Finding the global minimum in such problems requires schemes able to explore the space and avoid local minima. Many available methods are reviewed in [3], including the natural evolution strategies (NES) [4] or the cross-entropy (CE) method [5].

Among these methods, simulated annealing (SA) [6] relies on Boltzmann distributions, which are probability densities constructed from the cost function and parametrized by a temperature. Low-temperature Boltzmann densities are concentrated around the global minimizers, but it is hard to sample from them. Therefore, SA schemes use decreasing temperatures to generate samples approximating these Boltzmann distributions, and thus approaching the global minimizers.

The cooling schedule (i.e., the sequence of temperatures) must be fixed in advance and decrease slowly enough to reach convergence [6]. In particular, the cooling schedule cannot be adapted to the quality of the samples generated in past iterations. Many adaptive SA algorithms have thus been proposed [7, 8] to circumvent these issues, but mostly without the theoretical guarantees of the original SA scheme [6].

Most SA methods rely on Markov chain Monte Carlo schemes to draw samples, which allows to derive asymptotic convergence results, but can make convergence slow and difficult to assess [9]. Sampling from parametric densities, as in the NES [4] and CE [5] algorithms, provides an interesting perspective for SA schemes, as demonstrated in [10].

Our recent APSA algorithm [11] provides a methodology to adapt the cooling schedule in SA while generating parametric approximations of the Boltzmann distributions. To do so, it reformulates the problem into the minimization of the sum of the Kullback-Leibler (KL) divergence between the Boltzmannn distribution and its parametric approximation, and a regularizer term. This objective function is then minimized using an alternating Bregman proximal algorithm, relying on numerical integration to adapt the temperature.

In this work, we propose a novel SA algorithm that leverages the core ideas of [11], with a new choice of divergence, regularizer, and optimization procedure. These novel features lead to an improved algorithm which (i) is more efficient, (ii) exploits available information about the global minimum of the objective, (iii) admits theoretical guarantees, and (iv) shows good empirical performance. Specifically, we propose to use a Rényi divergence instead of the KL divergence, and we introduce a novel and more principled regularizer. The parametric distributions are chosen in an exponential family. We propose an alternating scheme involving Bregman proximal gradient steps. This allows to use the geometry of the considered families [12]. We show in two benchmark cost functions that the resulting method exhibits good performance compared to existing black-box algorithms.

T.G. and E.C. acknowledge support from the ERC Starting Grant MAJORIS ERC-2019-STG-850925. The work of V. E. is supported by the *Agence Nationale de la Recherche* of France under PISCES (ANR-17-CE40-0031-01), the Leverhulme Research Fellowship (RF-2021-593), and by ARL/ARO under grant W911NF-22-1-0235.

The rest of the paper is organized as follows. Background is provided in Section 2. We detail our method in Section 3. Numerical experiments are presented in Section 4, before concluding in Section 5.

2. PRELIMINARIES

2.1. Optimization problem and notation

We consider the following optimization problem

$$\text{Find } x_* \in \mathcal{X} \text{ s.t. } f(x_*) = f_* = \min_{x \in \mathcal{X}} f(x). \tag{1}$$

Hereabove, we consider a space $\mathcal{X} \subset \mathbb{R}^d$ and a cost function $f : \mathcal{X} \to \mathbb{R}$. In the following, $\mathcal{H}$ denotes a finite-dimensional Hilbert space with scalar product $\langle \cdot, \cdot \rangle$. $\mathcal{B}(\mathcal{X})$ denotes the Borel algebra of $\mathcal{X}$ and given a measure ν on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, we denote by $\mathcal{P}(\mathcal{X}, \nu)$ the set of probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ which admits a density with respect to ν. Given a measurable function h and $p \in \mathcal{P}(\mathcal{X}, \nu)$, we write $p(h) = \int h(x)p(x)\nu(dx)$. For a function h, we denote its conjugate function by h^*.

2.2. Exponential and Boltzmann families

Definition 1. *The exponential family with sufficient statistics* Γ *is the family* $\mathcal{Q} = \{q_\theta \in \mathcal{P}(\mathcal{X}, \nu),\ \theta \in \Theta\}$ *such that*

$$q_\theta(x) = \exp\left(\langle \theta, \Gamma(x) \rangle - A(\theta)\right),\ \forall x \in \mathcal{X}, \tag{2}$$

with A *being the log-partition function, such that* $\Theta = \operatorname{dom} A \subset \mathcal{H}$*, and which reads:*

$$A(\theta) = \log\left(\int \exp\left(\langle \theta, \Gamma(x) \rangle\right) \nu(dx)\right),\ \forall \theta \in \Theta. \tag{3}$$

Example 1. *Gaussian distributions with mean* μ *and covariance* Σ *form an exponential family [13], with* ν *being the Lebesgue measure and parameters* $\theta = (\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1})^\top$.

The Boltzmann distributions, denoted by π_β for $\beta > 0$, are central in SA schemes. π_β gives the probability of a state $x \in \mathcal{X}$ depending on $f(x)$ and the inverse temperature β.

Definition 2. *The family* $\mathcal{B}$ *of Boltzmann distributions associated with* f *is the set of densities defined for* $\beta > 0$ *by*

$$\pi_\beta(x) = \exp\left(-\beta f(x) - B(\beta)\right),\ \forall x \in \mathcal{X}, \tag{4}$$

where B *is the log-partition function of* π_β*:*

$$B(\beta) = \log\left(\int \exp\left(-\beta f(x)\right) \nu(dx)\right),\ \forall \beta > 0. \tag{5}$$

Remark 1. *The Boltzmann family* $\mathcal{B}$ *forms an exponential family, with sufficient statistics* $-f$ *and parameter* β.

The mass of the Boltzmann distributions concentrate on the global minimizers of f as $\beta \to +\infty$. Indeed, [6, Eq. (5.13)] implies that

$$\lim_{\beta \to +\infty} \pi_\beta(f) = f_*. \tag{6}$$

2.3. Rényi and KL divergences

Rényi and KL divergences are widely used as discrepancy measures between probability distributions.

Definition 3. *The* Rényi divergence *with parameter* $\alpha \in (0,1) \cup (1, +\infty)$ *and the* KL divergence *between* p_1 *and* p_2 *in* $\mathcal{P}(\mathcal{X}, \nu)$ *are respectively defined by*

$$RD_\alpha(p_1, p_2) = \frac{1}{\alpha - 1} \log\left(\int p_1(x)^\alpha p_2(x)^{1-\alpha} \nu(dx)\right),$$
$$KL(p_1, p_2) = \int \log\left(\frac{p_1(x)}{p_2(x)}\right) p_1(x)\nu(dx).$$

If one of the above is not well-defined, its value is $+\infty$.

The KL divergence is a limiting case of Rényi divergence [14], since $\lim_{\alpha \to 1,\, \alpha \leq 1} RD_\alpha(p_1, p_2) = KL(p_1, p_2)$.

3. PROPOSED APPROACH

3.1. A Rényi divergence minimization problem

In order to build our SA method, we propose to approximate the Boltzmann distributions from $\mathcal{B}$ by parametric distributions from an exponential family $\mathcal{Q}$. Therefore, we search for parameters β, θ such that q_θ and π_β are close and β is sufficiently large. To achieve this goal, we consider the minimization of the following function:

$$J_\alpha(\beta, \theta) = d_\alpha(\beta, \theta) + r(\beta),\ \forall \beta > 0,\ \theta \in \Theta, \tag{7}$$

where $\alpha \in (0,1)$, $d_\alpha(\beta, \theta) = RD_\alpha(\pi_\beta, q_\theta)$. Contrary to the Rényi divergence above, the KL divergence lacks the near-symmetry property [14, Proposition 2], which complicated greatly the temperature adaptation step in [11]. We furthermore introduce the regularizer

$$r(\beta) = B(\beta) + \beta f_\epsilon,\ \forall \beta > 0, \tag{8}$$

where $f_\epsilon = f_* + \epsilon$ for any small value $\epsilon > 0$. r is convex and differentiable on $\mathbb{R}_{++}$. It is lower bounded as it is equal, up to an additive constant, to $KL(\pi_{\beta_\epsilon}, \pi_\beta)$ where $\pi_{\beta_\epsilon}(f) = f_\epsilon$. It promotes the concentration of π_β around the global minimizers of f since $\nabla r(\beta) = 0$ if and only if $\pi_\beta(f) = f_\epsilon$.

If a possible value of f_ϵ is known in advance, it can be exploited directly. This is the case for instance in overparametrized neural networks, where $f_* = 0$ [15]. However, we will write our algorithm without this a priori knowledge.

3.2. Proposed algorithm

We propose to minimize J_α by using an *alternating Bregman proximal gradient algorithm* where the Bregman divergences are induced by the log-partition functions A and B. These are equivalent to the KL divergences [12]:

$$d_A(\theta, \theta') = KL(q_{\theta'}, q_\theta),\ \forall \theta, \theta' \in \Theta,$$
$$d_B(\beta, \beta') = KL(\pi_{\beta'}, \pi_\beta),\ \forall \beta, \beta' > 0.$$

Such choices allow to decouple the updates of β and θ, so that each update reduces to a variational inference update. Indeed, in this framework, Boltzmann distributions are approached by a variational approximation from $\mathcal{Q}$, and proposals from $\mathcal{Q}$ are also approximated by a Boltzmann distribution. Our choices also allow the use of the underlying geometries of $\mathcal{Q}$ and $\mathcal{B}$, which often leads to more efficient schemes [16, 17]. Finally, the available studies of [18, 19, 20] give our method a solid theoretical background.

We are now ready to present our approach, whose iteration $k \in \mathbb{N}$ consists in the following three steps with step-sizes $\eta_k, \tau_k \in (0,1]$, and $\widetilde{\eta}_k = \frac{1-\alpha}{\alpha}\eta_k$ with $\alpha \in (0,1)$. First, we perform a Bregman gradient descent step on $d_\alpha(\cdot, \theta_{k-1})$ within the Bregman divergence d_B with step-size $\widetilde{\eta}_k$:

$$\beta_{k-\frac{1}{2}} = \nabla B^*(\nabla B(\beta_{k-1}) - \widetilde{\eta}_k \nabla_\beta d_\alpha(\beta_{k-1}, \theta_{k-1})). \tag{9}$$

Second, we perform a Bregman proximal step on r within the divergence d_B and with step-size $\widetilde{\eta}_k$:

$$\beta_k = \underset{\beta>0}{\arg\min}\, r(\beta) + \frac{1}{\widetilde{\eta}_k} d_B(\beta, \beta_{k-\frac{1}{2}}). \tag{10}$$

Third, we perform a Bregman gradient descent step on $d_\alpha(\beta_k, \cdot)$ within the divergence d_A and with step-size τ_k:

$$\theta_k = \nabla A^*(\nabla A(\theta_{k-1}) - \tau_k \nabla_\theta d_\alpha(\beta_k, \theta_{k-1})). \tag{11}$$

We can compute for any $\beta > 0$, $\theta \in \Theta$,

$$\nabla_\beta d_\alpha(\beta, \theta) = \frac{\alpha}{1-\alpha}(\pi_\beta(-f) - \pi^{(\alpha)}_{\beta,\theta}(-f)), \tag{12}$$

$$\nabla_\theta d_\alpha(\beta, \theta) = q_\theta(\Gamma) - \pi^{(\alpha)}_{\beta,\theta}(\Gamma), \tag{13}$$

where $\pi^{(\alpha)}_{\beta,\theta} \in \mathcal{P}(\mathcal{X}, \nu)$ and has a density defined by

$$\pi^{(\alpha)}_{\beta,\theta}(x) \propto \left(\frac{\exp(-f(x))^\beta}{q_\theta(x)}\right)^\alpha q_\theta(x), \, \forall x \in \mathcal{X}. \tag{14}$$

Then, using results from [12, 19, 13], we can show that Eq. (9)-(11) are equivalent for every $k \in \mathbb{N}$ to

$$\pi_{\beta_{k-\frac{1}{2}}}(f) = (1-\eta_k)\,\pi_{\beta_{k-1}}(f) + \eta_k \pi^{(\alpha)}_{\beta_{k-1},\theta_{k-1}}(f), \tag{15}$$

$$\pi_{\beta_k}(f) = \frac{1}{1+\widetilde{\eta}_k}\pi_{\beta_{k-\frac{1}{2}}}(f) + \frac{\widetilde{\eta}_k}{1+\widetilde{\eta}_k} f_\epsilon, \tag{16}$$

$$q_{\theta_k}(\Gamma) = (1-\tau_k) q_{\theta_{k-1}}(\Gamma) + \tau_k \pi^{(\alpha)}_{\beta_k,\theta_{k-1}}(\Gamma). \tag{17}$$

Let us now state our convergence result for the proposed iterative scheme, using denominations from [13].

Proposition 1. *Let $\epsilon > 0$. Assume that $\mathcal{B}$ and $\mathcal{Q}$ are regular and full, that $\tau_k, \eta_k \in (0,1]$ for all $k \in \mathbb{N}$, and that the updates are well-defined. Then, the sequence $\{J_\alpha(\beta_k, \theta_k)\}_{k\in\mathbb{N}}$ is decreasing. If further, $\mathcal{B} \subset \mathcal{Q}$, then a fixed point of the algorithm (β_*, θ_*) is such that $\pi_{\beta_*}(f) = q_{\theta_*}(f) = f_\epsilon$.*

Proposition 1 establishes a monotonic decrease of the objective J_α. The second part of the result applies in particular when f is quadratic and $\mathcal{Q}$ is the Gaussian family.

3.3. Our Rényi-based adaptive SA (RASA) scheme

We now discuss the practical implementation of our algorithm for black-box optimization, meaning that only evaluations of f are available. We propose to approximate $\pi^{(\alpha)}_{\beta,\theta}$ and π_β using adaptive importance sampling [21]. To this end, consider $\theta \in \Theta$ and denote by X_θ a set of N samples from q_θ. We define for $\beta > 0$, $\alpha > 0$ the empirical distributions

$$\pi^{(\beta,\alpha)}_{X_\theta} \propto \sum_{x\in X_\theta} \left(\frac{\exp(-f(x))^\beta}{q_\theta(x)}\right)^\alpha \delta_x. \tag{18}$$

The distribution above approximates $\pi^{(\alpha)}_{\beta,\theta}$, while $\pi^{(\beta,1)}_{X_\theta}$ approximates π_β. This yields Algorithm 1, which approximates the idealized iterative algorithm described by Eq. (15)-(17), with a precision depending on the sample size N.

```
Algorithm 1: RASA algorithm
input: Choose $\beta_0 > 0$, $\theta_0 \in \Theta$, $N \in \mathbb{N}$, and $\{\tau_k, \eta_k\}_{k\in\mathbb{N}}$
       in $(0,1]$. Set $\widetilde{\eta}_k = \frac{1-\alpha}{\alpha}\eta_k$ for every $k \in \mathbb{N}$.
for $k = 1, 2, \dots$ do
    Sample $N$ points $X_{\theta_{k-1}} = \{x_n\}_{n=1}^N$ from $q_{\theta_{k-1}}$.
    Set $f_k = \min\{f(x),\, x \in X_{\theta_l},\, 0 \le l \le k-1\}$.
    If $k = 1$, set $\pi^{(\beta_{k-1},1)}_{X_{\theta_{k-2}}}(f) = \pi^{(\beta_{k-1},1)}_{X_{\theta_{k-1}}}(f)$.
    Compute the following empirical moment:
        $\pi_{\beta_{k-\frac{1}{2}}}(f) = (1-\eta_k)\,\pi^{(\beta_{k-1},1)}_{X_{\theta_{k-2}}}(f) + \eta_k \pi^{(\beta_{k-1},\alpha)}_{X_{\theta_{k-1}}}(f).$
    Set $\beta_k$ such that
        $\pi^{(\beta_k,1)}_{X_{\theta_{k-1}}}(f) = \frac{1}{1+\widetilde{\eta}_k}\pi_{\beta_{k-\frac{1}{2}}}(f) + \frac{\widetilde{\eta}_k}{1+\widetilde{\eta}_k} f_k.$
    Adapt the proposal by
        $q_{\theta_k}(\Gamma) = (1-\tau_k) q_{\theta_{k-1}}(\Gamma) + \tau_k \pi^{(\beta_k,\alpha)}_{X_{\theta_{k-1}}}(\Gamma).$
```

Note that $\beta \longmapsto \pi^{(\beta,1)}_{X_{\theta_{k-1}}}(f)$ is a decreasing continuous function. Thus, β_k can be found by a dichotomy search.

3.4. Discussion

In [11], numerical integration was used for temperature adaptation. In contrast, using the Rényi divergence and our novel regularizer simplifies the updates of β which are now convex combinations of integrals, like the updates of θ.

The adaptation step of θ of the MARS algorithm [10] is recovered in Algorithm 1 with $\alpha = 1$. However, a fixed cooling schedule was used in [10], while ours is fully adaptive.

Non-linear importance weights arise in Algorithm 1 from the densities $\pi^{(\alpha)}_{\beta,\theta}$ and the approximations $\pi^{(\beta,\alpha)}_{X_\theta}$. Such weights are known to prevent weight degeneracy issues [22].

4. NUMERICAL SIMULATIONS

4.1. Test problems and compared algorithms

We use two benchmarks with minimum value f_* reached at a unique x_*. Both are defined for any $x \in \mathbb{R}^d$. We consider a banana-shaped Rosenbrock function and a multimodal Rastrigin function, defined respectively by

$$f(x - x_*) = \sum_{i=1}^{d-1} \left(10(x_{i+1} - 1 - (x_i - 1)^2)^2 + x_i^2\right) + f_*,$$

$$f(x - x_*) = 4d + \sum_{i=1}^{d} \left(0.4x_i^2 - 4\cos(2\pi x_i)\right) + f_*.$$

The minimum values f_* are sampled uniformly in $[-1, 1]$, and the two functions are translated so that they reach the value f_* at x_* uniformly sampled in $[-1, 1]^d$.

We compare our RASA algorithm with similar black-box optimization schemes updating approximating densities using sampling. Namely, we consider the APSA scheme of [11], the MARS algorithm [10], and the CE algorithm [5].

All the algorithms are run with step-size $\tau_k = \frac{0.5}{k+1}$ for the parameters updates, and step-size $\eta_k = 0.9$ for the eventual temperature updates. The CE algorithm is run using the proportion $\rho = 0.5$. We simulated $N = 100$ samples at each iteration $k \in \mathbb{N}$ from Gaussian proposals with means μ_k. The algorithms are initialized with mean μ_0 uniformly sampled in $[-5, 5]^d$, covariance $\Sigma_0 = 10I$, and initial temperature $\beta_0 = 0.1$. At each iteration, the dichotomy search in RASA is performed in the interval $[0.1\beta_{k-1}, 1.5\beta_{k-1}]$.

4.2. Comparison of the cooling schedules

We first compare the cooling schedules of the APSA, MARS, and RASA algorithms in dimension $d = 2$. Indeed, the APSA scheme requires numerical integration steps which become intractable in higher dimensions.

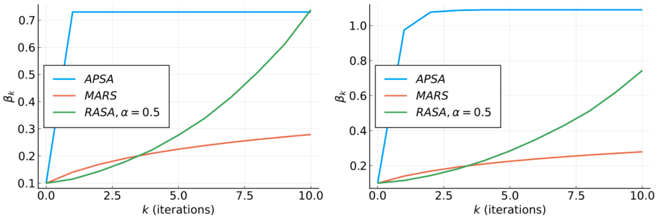

Fig. 1. Plots of the quantities β_k averaged over 500 runs for the Rastrigin (left) and Rosenbrock (right) functions in dimension $d = 2$.

Figure 1 shows that the three methods have very different cooling schedules. As observed in [11], the cooling schedule of the APSA algorithm reaches a high stationary value very fast. The MARS algorithm follows a logarithmic cooling schedule, which is the slowest of the three. In contrast, the cooling schedule of the APSA algorithm increases in a convex fashion, much faster than the logarithmic cooling schedule and without the stagnation of the APSA schedule.

4.3. Simulation results in high dimension

We now compare our RASA algorithm with the MARS and CE algorithms for $d = 50$, to see how our adaptive cooling schedule translates in terms of optimization performance. Note that the APSA method is intractable in this setting.

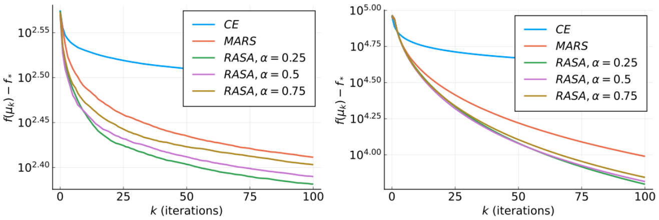

Fig. 2. Plots of the quantity $f(\mu_k) - f_*$ averaged over 500 runs for the Rastrigin (left) and Rosenbrock (right) functions in dimension $d = 50$.

We can see in Fig. 2 that RASA outperforms the MARS and CE algorithms on both benchmarks. This demonstrates the positive impact on optimization performance of our adaptive cooling schedule. Note that contrary to the CE algorithm, which only uses a ranking of the samples, the RASA and MARS algorithms use each sample's cost function value.

We can also notice that the lowest value of α yields the best performance. Indeed, transforming the weights as in Eq. (18) reduces weight degeneracy [22], which may lower the approximation error arising from sampling with a limited number of points in high dimension. This shows the interest of the additional parameter α, coming from the use of a Rényi divergence instead of a KL divergence as in [10].

5. CONCLUSION

In this work, we presented a novel SA algorithm for black-box optimization. It approaches Boltzmann distributions by variational approximations and treats the Boltzmann distributions as variational approximations themselves. This creates an adaptive cooling schedule while avoiding some drawbacks of the commonly used Markov chains schemes. This procedure is interpreted as an alternating Bregman proximal-gradient algorithm, allowing to gain novel theoretical insights. Numerical experiments illustrate the good behavior of the resulting method. This work also opens potential research tracks. On a theoretical level, the links between our practical implementation and its exact counterpart remain to be established precisely. The use of more complex approximating distributions, such as mixtures, can also be investigated, in particular in relation to multi-modal objectives.

6. REFERENCES

[1] S. Watanabe and J. Le Roux, "Black box optimization for automatic speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014))*, Firenze, Italy, 4-9 May 2014, pp. 3256–3260.

[2] M. A. Abramson, T. J. Asaki, J. E. Dennis, K. R. O'Reilly, and R. L. Pingel, "Quantitative object reconstruction using Abel transform x-ray tomography and mixed variable optimization," *SIAM Journal on Imaging Sciences*, vol. 1, no. 3, pp. 322–342, 2008.

[3] L. M. Rios and N. V. Sahinidis, "Derivative-free optimization: a review of algorithms and comparison of software implementations," *Journal of Global Optimization*, vol. 56, no. 3, pp. 1247–1293, 2013.

[4] D. Wiestra, T. Schaul, T. Glasmachers, Y. Sun, J. Peters, and J. Schmidhubler, "Natural evolution strategies," *Journal of Machine Learning Research*, vol. 15, no. 27, pp. 949–980, 2014.

[5] 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.

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

[7] 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.

[8] A. Beskos, A. Jasra, and A. Thiery, "On the convergence of adaptive sequential Monte Carlo methods," *The Annals of Applied Probability*, vol. 26, no. 2, pp. 1111–1146, 04 2017.

[9] E. Angelino, M. J. Johnson, and R. P. Adams, "Patterns of scalable Bayesian inference," *Foundations and Trends® in Machine Learning*, vol. 9, no. 2-3, pp. 119–247, 2016.

[10] 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.

[11] T. Guilmeau, E. Chouzenoux, and V. Elvira, "Proximal-based adaptive simulated annealing for global optimization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, Singapore, 22 - 27 May 2022, pp. 5453–5457.

[12] F. Nielsen and R. Nock, "Entropies and cross-entropies of exponential families," in *Proceedings of the IEEE International Conference on Image Processing (ICIP 2010)*, Hong Kong, China, 26 - 29 Sep. 2010, pp. 3621–3624.

[13] O. Barndorff-Nielsen, *Information and Exponential Families in Statistical Theory*, John Wiley & Sons, Ltd, 2014.

[14] T. van Erven and P. Harremoes, "Rényi divergence and Kullback-Leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[15] B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro, "Kernel and rich regimes in overparametrized models," in *Proceedings of 33rd Conference on Learning Theory (COLT 2020)*, Graz, Austria, 9-12 July 2020, pp. 3635–3673.

[16] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *Journal of Machine Learning Research*, vol. 14, no. 4, pp. 1303–1347, 2013.

[17] G. Raskutti and S. Mukherjee, "The information geometry of mirror descent," *IEEE Transactions on Information Theory*, vol. 61, no. 3, pp. 1451–1457, 2015.

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

[19] M. Teboulle, "A simplified view of first order methods for optimization," *Mathematical Programming*, vol. 170, no. 1, pp. 67–96, 2018.

[20] M. Ahookhosh, H. Le Ti Khanh, N. Gillis, and P. Patrinos, "Multi-block Bregman proximal alternating linearized minimization and its application to orthogonal nonnegative matrix factorization," *Computational Optimization and Applications*, vol. 79, no. 3, pp. 681–715, 2021.

[21] 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.

[22] E. Koblents and J. Míguez, "A population Monte Carlo scheme with transformed weights and its application to stochastic kinetic models," *Statistics and Computing*, vol. 25, no. 2, pp. 407–425, 2013.