PROXIMAL BASIN HOPPING: GLOBAL OPTIMIZATION WITH GUARANTEES

Anonymous authors
Paper under double-blind review

ABSTRACT

Global optimization is a challenging problem, with plenty of algorithms displaying empirical success, but scarce theoretical backing. In this work, we propose a new theoretical framework called **Proximal Basin Hopping** (PBH), carefully tailored to combine *proximal optimization* and *local minimization*. We use it to construct a practical algorithm that converges to the global minimizer with high probability, when using a *finite amount of samples*. Proximal Basin Hopping outperforms well known algorithms on standard synthetic hard functions for moderate dimensions, and on real problems like fitting scaling laws for deep learning. Furthermore, the higher the dimension, the better the performance.

1 INTRODUCTION

Finding global minimizers of highly non-convex functions is a challenging but crucial task in various fields such as computational chemistry and biology (Prentiss et al., 2008; Alvarez & Calaminici, 2025), engineering design (Arora et al., 1995; Englander et al., 2020) and of course machine learning (Goodfellow et al., 2016). We are interested in solving the following optimization problem

$$\text{Find } \widehat{x} \in \arg\min_{x \in \mathbb{R}^d} f(x). \tag{1}$$

where f is a non-convex, continuous function. We will assume that f possesses one global minimizer and some regularity that can be exploited to compute some descent steps (or find local minimizers directly). To tackle this problem, we propose a hybrid approach that sits at the intersection between proximal algorithms (Parikh & Boyd, 2014) and basin hopping algorithms (Wales & Doye, 1997).

Standard first-order algorithms can only hope to obtain local solutions to Problem equation 1, which has prompted many to propose global optimization techniques, most of which are heuristic and do not possess strong convergence guarantees. Global optimization algorithms perform by combining exploration (through random perturbation of the current guesses) and exploitation steps (through function evaluation and/or local minimization) to reach a solution. The exploration requires a number of samples going to infinity to fit an underlying algorithm with increasing precision (Fornasier et al., 2024; Zhang et al., 2024), or a number of samples growing exponentially fast with the dimension to cover the space sufficiently enough (Hansen et al., 1992; Bouttier et al., 2020) to reach acceptable accuracy. This becomes rapidly cumbersome, even for problems of (nowadays) relatively small dimension (in the few hundreds) (Papenmeier et al., 2025). On the other hand, the inexactness of the exploitation steps through local minimization is difficult to account for in the theoretical analysis (György & Kocsis, 2011). Hence, we propose a new framework opening the path to quantify the effect of sample size and inexactness on the convergence to the global minimizer.

Contributions. The method we propose, which we call **Proximal Basin Hopping (PBH)**, explores the landscape of f with sampling and local minization. Provided that f is somewhat coercive, PBH will funnel its iterations to the global minimizer. PBH is derived from an ideal operator $\mathrm{prox}^{\gamma}_{T_f} f(x)$, where T_f is a deterministic local solver, yielding for (almost) any starting point its achievable local minimizer, and prox denotes a proximal operation. Exploration is provided with a zeroth order approximation of this prox, and local minimization with T_f (see Figure 1 for a practical illustration). The convergence is obtained by studying the ideal operator, and show that more practical ones approximate it sufficiently well. We list below these approximations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

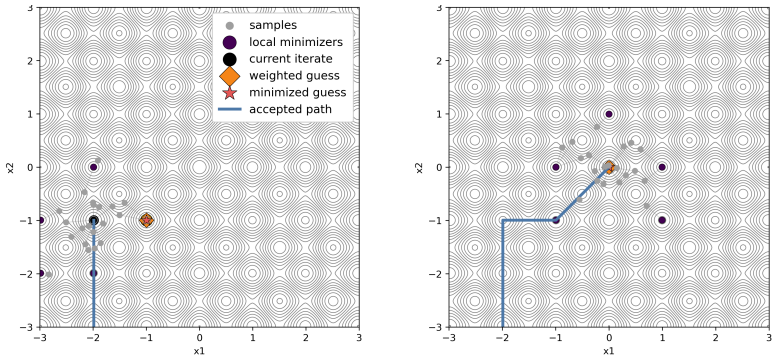


Figure 1: Proximal Basin Hopping on the 2D-Rastrigin function at iteration 3 (left) and iteration 5 (right), where convergence to the global minimizer, $x_* = (0, 0)$, has occurred. From the current iterate, we draw several samples, that are then sent to their nearest achievable local minimizer, then weighted with respect to the achieved function value. It gives a new guess that is again sent to its nearest achievable local minimizer (if it is not there already).

1. Ideal PBH: we make sense of the $\approx$ in Lemma 1,
$$S_f^\gamma(x) := \mathrm{prox}_{T_f}^\gamma f(x) \approx T_f\Big(\mathrm{prox}_{\gamma(f\circ T_f)}(x)\Big)\,.$$
2. Exact expectation PBH: the proximal operator is replaced by its zeroth-order approximation
$$S_f^\gamma(x) \approx_\delta T_f\left(\frac{\mathbb{E}_{y\sim\mathcal{N}(x,\delta\gamma\mathrm{Id})}[T_f(y)\exp(-f\left(T_f(y)\right)/\delta)]}{\mathbb{E}_{y\sim\mathcal{N}(x,\delta\gamma\mathrm{Id})}[\exp(-f\left(T_f(y)\right)/\delta)]}\right).$$
3. Approximated expectation PBH: the zeroth-order approximation of the prox is estimated using N samples. For $i = 1, \dots, N$, $y_i \sim \mathcal{N}(x, \delta\gamma\mathrm{Id})$
$$S_f^\gamma(x) \approx_\delta^N T_f\left(\frac{\sum_{i=1}^N T_f(y_i)\exp(-f\left(T_f(y_i)\right)/\delta)}{\sum_{i=1}^N \exp(-f\left(T_f(y_i)\right)/\delta)}\right).$$
4. Approximated expectation & inexact local solving PBH:
$$S_f^\gamma(x) \approx_{\delta,\varepsilon}^N T_f^\varepsilon\left(\frac{\sum_{i=1}^N T_f^\varepsilon(y_i)\exp(-f\left(T_f^\varepsilon(y_i)\right)/\delta)}{\sum_{i=1}^N \exp(-f\left(T_f^\varepsilon(y_i)\right)/\delta)}\right),$$

By allowing inexact computations of the proximal operator and of the local solver, **we obtain a practical algorithm, that can recover the global minimizer.** This relies on a structural assumption on f: *the global minimizer is sufficiently different in function value from local minimizers*.

We present extensive experiments on hard-to-optimize functions in moderate dimension (20 to 200), and apply our algorithm to the problem of fitting scaling laws on real data from (Shukor et al., 2025). Our algorithm becomes conclusively better than basin hopping from moderate to high dimension and stays competitive with state-of-the-art approach in global optimization with less theoretical guarantees.

Summary of theoretical results. Our goal is to show the convergence of our most practical algorithm. Such a result is obtained building up on the proofs of convergence of more ideal algorithms, by trying to emulate the same behavior: the second operator induces convergence because it approximates the first operator. The third operator induces convergence by emulating the second, and so on.

1. Ideal PBH: convergence holds through the properties of the potential $V_x(M) := 1/(2\gamma)\mathrm{dist}(x, \overline{\mathrm{Att}(M)})^2 + f(M)$ and of its minimizers which are the elements of $S_f^\gamma(x)$: it is sufficiently well-behaved for finite time convergence to occur.

2. Exact expectation PBH: by proving convergence of the expectation onto the minimizers of $V_x(\cdot)$ when δ goes to 0, we can reuse the convergence arguments of the first algorithm to guarantee convergence to the global minimizer (but not in finite time).
3. Approximated expectation PBH: first by standard arguments, bringing N to infinity allows us to recover the setting of the second algorithm. In a following analysis, we show that with high probability, we only need a finite N to converge.
4. Approximated expectation & inexact local solving PBH: we highlight simple conditions on the inexactness that need to hold to obtain convergence. Due to the complexity of the analysis, we assume first exact expectation.

2 Structural assumptions on f and the local solver

In this section, we present in detail the structural assumptions we make on f and on the local solver T_f in order to guarantee convergence of our algorithms. We make assumptions on f that are either standard in optimization, or reasonable given our goal of reaching the global minimizer.

Assumption 1. *The function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is proper, continuous, and coercive. We define the family of sets of local minimizers of f as*

$$\mathcal{M} := \Big\{ M \text{ connected} : \exists\, \mathcal{V} \text{ (open)} \subset \mathbb{R}^d,\ M \subsetneq \mathcal{V}, \\ M = \operatorname*{arg\,min}_{x \in \mathcal{V}} f(x), \forall x \in \mathcal{V} \setminus M,\ f(x) > f(M) \Big\}. \tag{2}$$

We assume that every $M \in \mathcal{M}$ is Borel, bounded, and only finitely many elements of $\mathcal{M}$ intersect each compact subset of $\mathbb{R}^d$.

This definition of $\mathcal{M}$ prevents redundancy, and ensures that the M are pairwise disjoints. Thus, $\mathcal{M}$ consists of possibly infinitely many disjoint, connected components of local minimizers $M \in \mathbb{R}^d$, but finitely many on any compact set. In other words, f is "tame" (see App. C.1 for some examples). In the basin hopping literature, the standard hypothesis is for f to have finitely many of these components[1]. This case is easily covered by our framework. We take the liberty of writing $f(M) < f(Q)$ for any component Q and M in $\mathcal{M}$ as f is constant on each connected component. We will denote the union of all elements of $\mathcal{M}$ as $\mathcal{M}_* = \bigcup_{M \in \mathcal{M}} M$. We also need separation between the global minimizer and other local minima:

Assumption 2. *The function f has a unique global minimizer $\{x_*\}$. $\exists\ \mu > 0$ such that for all $M \in \mathcal{M} \setminus \{x_*\}$, $f_{\min} + \mu < f(M)$.*

This assumption is relatively strong by preventing arbitrary accumulation of almost global local minimizers, but also in line with previous assumptions of the literature such as (Zhang et al., 2024, Assumption 3) which requires that no critical point $x \neq x_*$ such that $f(x) < f_{\min} + \mu$ exists.

Assumption 3. *The local solver $T_f : \mathbb{R}^d \to \mathbb{R}^d$ is measurable, single-valued, and outputs almost everywhere a local minimizer, i.e., $T_f(x) \in \mathcal{M}_*$ a.e..*

T_f maps any point to its unique achievable local minimizer. We define this achievability with the attractors of the local minimizers, which we denote by Att

$$(\forall M \in \mathcal{M}), \quad \mathrm{Att}(M) := \left\{ y \in \mathbb{R}^d,\ T_f(y) \in M \right\}. \tag{3}$$

Obviously the geometry of Att depends a lot on the solver (Asenjo et al., 2013; Levy, 2018; 2026), we can operate under the assumption that these sets are open or closed but well-behaved[2]. Furthermore, under standard assumptions the attraction sets of critical points which are not local minimizers are of measure 0 (Lee et al., 2016). As a consequence, the sets $\mathrm{Att}(M)$ form a measurable partition of $\mathbb{R}^d$ up to a null set (Lemma 6).

Assumption 4. *For every $M \in \mathcal{M}$, the restriction of T_f to $\mathrm{int}(\mathrm{Att}(M))$ is continuous.*

[1] Finite does not mean small, the number of minimizers grows exponentially with the dimension for the Lennard-Jones cluster problem, introduced in the seminal paper of basin hopping (Wales & Doye, 1997).

[2] See (Levy, 2018; 2026) for a more exhaustive discussion on basins of attraction.

This assumption is verified for a gradient flow, or a finitely many gradient steps, if the gradient is continuous. See App. C.1 for a detailed discussion when T_f is built with gradient descent.

We make one last structural assumption on the behavior of $M \mapsto \mathrm{dist}(x, \overline{\mathrm{Att}(M)})$ for $x \in \mathbb{R}^d$.

Assumption 5. *For every $x \in \mathbb{R}^d$ and $\gamma > 0$, the function*

$$V_x : M \in \mathcal{M} \mapsto f(M) + \frac{1}{2\gamma}\mathrm{dist}(x, \overline{\mathrm{Att}(M)})^2 \tag{4}$$

has a nonempty bounded set of minimizers on $\mathcal{M}$. Furthermore,

$$\mathrm{dist}(x, \mathrm{Att}(M)^\circ) = \mathrm{dist}(x, \overline{\mathrm{Att}(M)}) = \mathrm{dist}(x, \mathrm{Att}(M)). \tag{5}$$

We define:

$$S_f^\gamma(x) := \mathrm{prox}_{T_f}^\gamma f(x) := \operatorname*{arg\,min}_{M \in \mathcal{M}} V_x(M). \tag{6}$$

This assumption is relatively strong (it implies well-posedness of S_f^γ) but also reasonable: it holds for gradient flow. The coercivity of f yields the coercivity of V_x, and the assumption that f has finitely many components of local minimizers on any bounded set imply the boundedness of the $\arg\min$. The non-emptiness is always true if f is C^1 and T_f consists of finitely many gradient steps as it is continuous, yielding closedness of $\mathrm{Att}(M)$. For a general gradient flow, $\mathrm{Att}(M)$ is open: take for instance $f := \|\cdot\|^2/2$. Hence, this assumption. Finally, we define for $x \in \mathbb{R}^d$,

$$\alpha_x := \inf_{M \in \mathcal{M}} V_x(M),\ M = \bigcup_{M \in S_f^\gamma(x)} M. \tag{7}$$

3 Proposed method

We first consider our ideal operator. Its study is of great interest as it will provide the optimization problem which we hope our relaxed algorithm will emulate with the proper parameter δ. For any entry $x \in \mathbb{R}^d$, the operator S_f^γ identifies nearby local minimizers. All the proofs of this section are deferred to App. D.1. Under the assumption that x does not see $\mathrm{Att}(M)$ only through its boundary, we can prove that $S_f^\gamma(x) = T_f(\mathrm{prox}_{f \circ T_f}(x))$, which is a helpful characterization in order to approximate S_f^γ later.

Lemma 1. *Suppose that Assumptions 1, 3, and 5 hold. Set $\gamma > 0$. We have for all $x \in \mathbb{R}^d$ such that for every $M \in S_f^\gamma(x)$, $\mathrm{proj}_{\overline{\mathrm{Att}(M)}}(x) \cap \mathrm{Att}(M) \neq \varnothing$, that $S_f^\gamma(x) = T_f(\mathrm{prox}_{f \circ T_f}(x))$.*

This operator S_f^γ identifies the achievable minimizers to which x can be sent. The selection of the minimizer depends on the value of γ. Provided that γ is large enough, S_f^γ recovers the global minimizer in one iteration.

Lemma 2. *Suppose that Assumptions 1, 2, and 5 hold. For all $x \in \mathbb{R}^d$, there exists $\gamma_x > 0$ such that $\{x_*\} \in S_f^\gamma(x)$ for every $\gamma \geq \gamma_x$.*

As we will see, this relationship between γ and the elements of S_f^γ is predictive of the convergence of our more practical algorithms.

Theorem 1. ***Convergence of ideal proximal basin hopping.*** *Suppose that Assumptions 1, 2, and 5 hold. Let $x_0 \in \mathbb{R}^d$, $\gamma_0 > 0$, $\eta > 1$, iterate for $k = 0, 1, \ldots$*

$$x_{k+1} \in \operatorname*{arg\,min}_{z \in \mathcal{M}_{x_k}} f(z) \tag{8}$$

$$\gamma_{k+1} = \eta\gamma_k \mathbf{1}_{\{f(x_{k+1}) = f(x_k)\}} + \gamma_k \mathbf{1}_{\{f(x_{k+1}) < f(x_k)\}} \tag{9}$$

Then, $\{x_k\}_{k \in \mathbb{N}}$ converges to x_* in finitely many iterations.

3.1 PROXIMAL BASIN HOPPING WITH EXACT EXPECTATION.

Equipped with the proximal interpretation of S_f^γ, we introduce a relaxation by replacing the proximal operator by its zeroth-order approximation of parameter $\delta > 0$ (Osher et al., 2023; Lauga & Vaiter, 2026), playing the role of the temperature, for all $x \in \mathbb{R}^d$

$$S_f^\gamma(x) \approx_\delta T_f\left(\frac{\mathbb{E}_{y\sim\mathcal{N}(x,\delta\gamma\mathrm{Id})}[T_f(y)\exp(-f\left(T_f(y)\right)/\delta)]}{\mathbb{E}_{y\sim\mathcal{N}(x,\delta\gamma\mathrm{Id})}[\exp(-f\left(T_f(y)\right)/\delta)]}\right). \tag{10}$$

With respect to the ideal operator, we want to understand how the approximated proximal operator actually converges to a solution of $\arg\min V_x$ when δ goes to 0. In the small δ regime, this approximation behaves closely to the ideal S_f^γ, with the only caveat that if S_f^γ is set-valued, the approximation sits near the convex hull of this set. As before, proofs are deferred to App. D.2.

We have this first result, derived from large deviation theory (den Hollander, 2000) and the properties of the Gaussian distribution. It is the foundation of our analysis.

Lemma 3. *Suppose that Assumptions 1, 5 hold. Let $x \in \mathbb{R}^d$, $\gamma > 0$, $\delta > 0$, and let $Y_\delta \sim \mathcal{N}(x, \delta\gamma I_d)$. Let $M \in \mathcal{M}$. Then the family $(Y_\delta)_{\delta>0}$ satisfies, for each M,*

$$\lim_{\delta\downarrow 0} \delta \log \mathbb{P}(Y_\delta \in \mathrm{Att}(M)) = -\frac{1}{2\gamma}\mathrm{dist}(x, \overline{\mathrm{Att}(M)})^2.$$

One can see that weighting this probability by $e^{-f(M)/\delta}$ will make the potential V_x appear on the right-hand side. Hence, we can show that the ratio of expectations in Eq. equation 10 converges to the set of minimizers of V_x when δ goes to 0. This is easy to obtain if the number of minimizing components is finite (App. D.2, Proposition 5). As we have infinite number of local-minimizer components, the crucial part is to make sure that almost all the mass is in a bounded set where the finite-number-of-components argument will help us complete the proof[3]. One can see that f cannot be too flat in all directions for this property to hold, otherwise it would allow some of the mass to escape to infinity. Coercivity is not enough to contain the mass, hence we make an additional growth assumption on f.

Assumption 6. *There exist $C_1, C_2, C_3 > 0$ and $0 < \alpha < \beta$ such that*

$$(\forall z \in \mathbb{R}^d), \qquad C_1\|z\|^\alpha - C_2 \le f(z) \le C_3(1 + \|z\|)^\beta. \tag{11}$$

First, we show that a measure induced by the exponential weights goes to 1 when $\delta \downarrow 0$ on bounded sets of $\mathcal{M}_*$.

Proposition 1. *Suppose that Assumptions 1, 2, 3, 5, and 6 hold. Let $x \in \mathbb{R}^d$, $\gamma, \delta > 0$, and $Y_\delta \sim \mathcal{N}(x, \delta\gamma I_d)$. For any Borel $B \in \mathcal{M}_*$, we define*

$$\nu_{x,\delta}(B) := \mathbb{E}\left[e^{-f(T_f(Y_\delta))/\delta}\mathbf{1}_{\{T_f(Y_\delta)\in B\}}\right], \qquad \mu_{x,\delta}(B) := \frac{\nu_{x,\delta}(B)}{\nu_{x,\delta}(\mathcal{M}_*)}, \tag{12}$$

For any bounded set $K \subset \mathcal{M}_*$, $\mu_{x,\delta}(K) \underset{\delta\downarrow 0}{\longrightarrow} 1$.

In addition, we can also ensure that the mass outside bounded sets does not escape too far (App. D.2, Proposition 6). Finally, we obtain our convergence result.

Proposition 2. *Suppose that Assumptions 1, 2, 3, 5, and 6 hold. Let $x \in \mathbb{R}^d$, $\delta, \gamma > 0$ and $Y_\delta \sim \mathcal{N}(x, \delta\gamma\mathrm{Id})$. Denote by $\mathrm{conv}(\mathcal{M}_x)$ the convex hull of $\mathcal{M}_x$, the set of minimizers of V_x (Eq. 7). We have*

$$\mathrm{dist}\left(\frac{\mathbb{E}[T_f(Y_\delta)e^{-f(T_f(Y_\delta))/\delta}]}{\mathbb{E}[e^{-f(T_f(Y_\delta))/\delta}]}, \mathrm{conv}(\mathcal{M}_x)\right) \underset{\delta\downarrow 0}{\longrightarrow} 0 \tag{13}$$

In particular, if $\mathcal{M}_x = \{m_{x,*}\}$, i.e., an isolated local minimizer, then,

$$\frac{\mathbb{E}[T_f(Y_\delta)e^{-f(T_f(Y_\delta))/\delta}]}{\mathbb{E}[e^{-f(T_f(Y_\delta))/\delta}]} \underset{\delta\downarrow 0}{\longrightarrow} m_{x,*}. \tag{14}$$

[3] An easy way to make sure that all the mass is in a compact set is to truncate the attained minimizers if they have a norm above a threshold R, thus falling back on the finite minimizers case.

This result is almost equivalent to that of the finite case: the weighted barycenter onto which m_δ converges is inside the convex hull (App. D.2, Proposition 5). In general, m_δ does not need to lie inside the convex hull as some of the mass still lies outside the set of interest as long as $\delta > 0$.

The convergence now relies on playing with the γ parameter to identify competing minimizers (Remark 1) and enable concentration on one element. If, in some iterations, several components are tied, it is not really a problem as long as it does not happen indefinitely. And we can guarantee that it does not by imposing a simple rule: if the new guess is of higher function value than the current one, we increase γ. However, with respect to the first ideal algorithm, we need γ to grow comparatively slower than δ, in order to avoid breaking our concentration result, hence we decrease δ at each iteration (while γ does not change if progress is made). Note that we do not reject equivalent samples in terms of function values for the sake of exploration.

Algorithm 1 Exact expectation proximal basin hopping

Require: $x_0 \in \mathbb{R}^d$, $1 < \eta_1 < \eta_2$, $\gamma_0 > 0$, $\delta_0 > 0$ (small)
1: **for** $k = 0, 1 \ldots,$ **do**
2: $\quad m_{\delta_k,\gamma_k} \leftarrow \frac{\mathbb{E}[T_f(Y) e^{-f(T_f(Y))/\delta_k}]}{\mathbb{E}[e^{-f(T_f(Y))/\delta_k}]}, Y \sim \mathcal{N}(x_k, \delta_k \gamma_k)$
3: $\quad x_+ \leftarrow T_f\left(m_{\delta_k,\gamma_k}\right)$
4: $\quad$ **if** $f(x_+) > f(x_k)$ **then**
5: $\quad\quad x_{k+1} \leftarrow x_k, \gamma_{k+1} \leftarrow \eta_1 \gamma_k$
6: $\quad$ **else if** $f(x_+) = f(x_k)$ **then**
7: $\quad\quad x_{k+1} \leftarrow x_+, \gamma_{k+1} \leftarrow \eta_1 \gamma_k$
8: $\quad$ **else**
9: $\quad\quad x_{k+1} \leftarrow x_+, \gamma_{k+1} \leftarrow \gamma_k$
10: $\quad \delta_{k+1} \leftarrow \frac{\delta_k}{\eta_2}$
11: **return** x_{end}

Theorem 2. ***Convergence of exact expectation PBH.*** *Suppose that Assumptions 1, 2, 3, 5, and 6 hold. Then, the iterates $\{x_k\}_{k\in\mathbb{N}}$ of Algorithm 1 converge to x_* when $k \to +\infty$.*

3.2 Proximal basin hopping with approximated expectation.

In practice, we estimate our operator S_f^γ with a certain number of samples. It immediately begs the question, can this be sufficient to obtain convergence with reasonably high probability? The answer is yes, and we will show it by studying the case where N actually goes to infinity to understand it. Again, we defer the proofs to App. D.3.

$$S_f^\gamma(x) \approx_\delta^N T_f\left(\frac{\sum_{i=1}^N T_f(y_i) \exp(-f\left(T_f(y_i)\right)/\delta)}{\sum_{i=1}^N \exp(-f\left(T_f(y_i)\right)/\delta)}\right) \text{ where } y_i \sim \mathcal{N}(x, \delta\gamma \mathrm{Id}). \tag{15}$$

Convergence when $N \to +\infty$. As for the exact case, we are first concerned with the well-posedness of the operator $m_{N,\delta,\gamma}$ for all $x \in \mathbb{R}^d$ (line 2 Alg. 2). With standard arguments, we recover that $m_{N,\delta,\gamma} \to m_{\delta,\gamma}$ when $N \to +\infty$ (App. D.3, Lemma 12). The convergence of $T_f(m_{N,\delta,\gamma})$ to $T_f(m_{\delta,\gamma})$ is however not directly guaranteed as $m_{\delta,\gamma}$ may lie on the boundary of some $\mathrm{Att}(M)$ (where we would lose local continuity of T_f), or may lie in the null set, which consists of points from which T_f will not converge to a local minimizer. The first situation is covered by Lemma 13 and 14 (App. D.3). The null set is handled directly by our algorithm. With our built-in increase of γ, $m_{\delta,\gamma}$ will fall under the conditions of Lemma 13. Indeed, for sufficiently large γ, and sufficiently small δ, $m_{\delta,\gamma}$ is close to x_* (Proposition 2). Hence increasing N sufficiently does the job.

Algorithm 2 Approximated expectation proximal basin hopping

Require: $x_0 \in \mathbb{R}^d$, $1 < \eta_1 < \eta_2$, $\gamma_0 > 0$, $\delta_0 > 0$ (small), $C \geq 1$, $N_0 \in \mathbb{N} \setminus \{0\}$,

for $k = 0, 1 \ldots,$ **do**
  $m_{N_k,\delta_k,\gamma_k} \leftarrow \frac{\sum_{i=1}^{N_k} T_f(y_i) \exp(-f(T_f(y_i))/\delta)}{\sum_{i=1}^{N_k} \exp(-f(T_f(y_i))/\delta)}$, $y_1, y_2, \ldots, y_{N_k} \sim \mathcal{N}(x_k, \delta_k \gamma_k \mathrm{Id})$
  $x_+ \leftarrow T_f\left(m_{N_k,\delta_k,\gamma_k}\right)$
  if $f(x_+) > f(x_k)$ **then**
    $x_{k+1} \leftarrow x_k$, $\gamma_{k+1} \leftarrow \eta_1 \gamma_k$
  else if $f(x_+) = f(x_k)$ **then**
    $x_{k+1} \leftarrow x_+$, $\gamma_{k+1} \leftarrow \eta_1 \gamma_k$
  else
    $x_{k+1} \leftarrow x_+$, $\gamma_{k+1} \leftarrow \gamma_k$
  $\delta_{k+1} \leftarrow \frac{\delta_k}{\eta_2}$, $N_{k+1} \leftarrow C \times N_k$
return x_{end}

Theorem 3. ***Convergence of approximate expectation PBH*** *($N \to +\infty$). Suppose that Assumptions 1, 2, 3, 4, 5, and 6 hold. The iterates $\{x_k\}_{k\in\mathbb{N}}$ of Algorithm 2 (i.e., $C > 1$) converge to x_* when $k \to +\infty$.*

Convergence when $N < +\infty$. We now address the convergence of our algorithm when the number of samples remains finite. By quantifying that for N large enough $\|m_{N,\delta,\gamma} - m_{\delta,\gamma}\|$ is small with high probability, we can use the arguments of the proof of Theorem 3 to recover the convergence.

Proposition 3. *Suppose that Assumptions 1, 2, 3, 4, 5, and 6 hold. Let $x \in \mathbb{R}^d$, $\gamma, \delta > 0$. For any $\varepsilon > 0$, if $N \in \mathbb{N} \setminus \{0\}$ is large enough there exist $\alpha(\varepsilon) > 0$, such that for N i.i.d. samples y_i drawn from $Y_\delta \sim \mathcal{N}(x, \delta\gamma\mathrm{Id})$,*

$$\mathbb{P}\left(\|m_{N,\delta,\gamma} - m_{\delta,\gamma}\| \leq \varepsilon\right) \geq 1 - \frac{\alpha(\varepsilon)}{N}. \tag{16}$$

In fact, the previous result holds for all N, but is vacuous for N too small. It also holds for fixed δ and γ so we need to be careful in bringing δ to 0 too fast for theoretical and practical reasons.

Theorem 4. ***Convergence of approximate expectation PBH*** *($N < +\infty$). Suppose that Assumptions 1, 2, 3, 4, 5, and 6 hold. Then, the iterates $\{x_k\}_{k\in\mathbb{N}}$ of Algorithm 2 with finite N (i.e., $C = 1$) converge to x_* when $k \to +\infty$ with high probability if N is sufficiently large.*

The required N is path-dependent, but does not need to hold for all the iterates, only for those that have an underlying $m_{\delta,\gamma}$ able to send them to x_*. Practical considerations are discussed in App. B, notably the interplay between Proposition 3 and the values of δ and γ. Remark that the local minimizations of the y_i's can be parallelized.

3.3 Proximal basin hopping with approximated expectation and local minimization.

Now, if we also assume that we cannot do complete local minimization but only obtain approximate local solutions, then we have a third level of approximations in our algorithm.

$$S_f^\gamma(x) \approx_{\delta,\varepsilon}^N T_f^\varepsilon\left(\frac{\sum_{i=1}^N T_f^\varepsilon(y_i) \exp\left(-f\left(T_f^\varepsilon(y_i)\right)/\delta\right)}{\sum_{i=1}^N \exp\left(-f\left(T_f^\varepsilon(y_i)\right)/\delta\right)}\right), \text{ where } y_i \sim \mathcal{N}(x, \delta\gamma\mathrm{Id}) \tag{17}$$

where for all $x \in \mathbb{R}^d$, T_f^ε approximately solves f for some precision $\varepsilon \geq 0$. A first observation is that to obtain convergence we will need $\frac{\varepsilon}{\delta} \to 0$. As the analysis is much more complex, we use exact expectation. The proof is in Appendix D.4.

Theorem 5. ***Exact-expectation PBH with an inexact local solver*** *Suppose that Assumptions 1, 2, 3, 4, 5, and 6 hold. For every $\varepsilon > 0$, let $T_f^\varepsilon : \mathbb{R}^d \to \mathbb{R}^d$ be a measurable inexact local solver satisfying, for every $y \in \mathbb{R}^d$, $\left\|T_f^\varepsilon(y) - T_f(y)\right\| \leq r(\varepsilon)$, $0 \leq f(T_f^\varepsilon(y)) - f(T_f(y)) \leq \varepsilon$, with $r(\varepsilon) \underset{\varepsilon\downarrow 0}{\longrightarrow} 0$. Let*

$\varepsilon = \varepsilon(\delta)$ *satisfy* $\varepsilon(\delta) \xrightarrow[\delta \downarrow 0]{} 0$, $\frac{\varepsilon(\delta)}{\delta} \xrightarrow[\delta \downarrow 0]{} 0$. *Then, the iterates* $\{x_k\}_{k \in \mathbb{N}}$ *of Algorithm 1, with inexact solvers, converge to* x_* *when* $k \to +\infty$.

4 NUMERICAL EXPERIMENTS

In this section, we benchmark our algorithm against the zeroth-order proximal (ZOP) algorithm presented in (Zhang et al., 2024), and basin hopping (BH) algorithm (Wales & Doye, 1997). These two algorithms are the closest in terms of convergence guarantees to our proposal. We also add comparisons on synthetic functions to standard methods with less theoretical backing but an history of practical success such as Bayesian Optimization (BO) (Snoek et al., 2012), Differential Evolution (DE) (Storn & Price, 1997), Particle Swarm Optimization (PSO) (Kennedy & Eberhart, 1995), and Covariance Matrix Adaptation Evolution Strategy(CMA-ES) (Hansen & Ostermeier, 2001). Additionally, we report the result of our algorithm when always selecting the best sample at each iteration (which has no convergence guarantees), denoted PBH-b(est).

Comparison setup. Basin hopping is our Algorithm 2 with sample size of 1. The new guess, however, is accepted with probability proportional to $\exp(\frac{f(x)-f(T_f(y_i))}{\delta})$. Zeroth-order Prox is our Algorithm App. Exp. PBH with $T_f = \mathrm{Id}$. As the three algorithms are not comparable w.r.t. to iteration count, we choose to report CPU time instead. We set parameters of all algorithms to the same value at initialization, and update with the rules of our algorithm, keeping the number of samples constant for PBH and ZOP. We did not investigate multisearch, and instead focused on convergence from one initialization. After the CPU-time budget has been spent, algorithms are stopped.

Set of problems. *Hard synthetic functions.* We test our algorithm against the d-dimensional Rastrigin, its anisotropic and rotated version, the d-dimensional Styblinski-Tang function, and the d-dimensional Griewank functions (the easiest of the four). We optimized these functions with dimensions ranging from 20 to 200. The local solver uses gradient descent. An abalation study was performed to identify critical parameters, and is reported in Appendix E.

Fitting scaling laws. Fitting scaling laws is an important problem in deep learning to predict model performance as a function of model size K and the number of tokens D. In particular, the choice of dataset mixtures is of crucial importance. To derive a good scaling law, the authors of (Shukor et al., 2025) collected training runs with different domain weights h (datasets), model sizes K, and token budgets D, and recorded the resulting loss on the target domain. Three scaling law models were tested in (Shukor et al., 2025) to understand optimal data mixtures (Shukor et al., 2025, Section 2.2). The optimal parameters are estimated by minimizing a Huber objective over p observations. Finally, evaluation is performed on held-out runs with unseen triples (K, D, h) by comparing predicted and observed losses. The local solver uses LBFGS steps. The code for computing, fitting, and evaluating the scaling laws is reused from `https://github.com/apple/ml-scalefit`.

Numerical results. All the results displayed in this section are computed using 8 CPUs. We report final rankings to better highlight the increasing competitiveness of our algorithm with increasing dimension. For the synthetic functions, we can see that the performance of PBH improves when the dimension increases. ZOP failed to converge often: our hypothesis is that the number of samples needs to increase as δ decreases (Zhang et al., 2024, Section 5.2) in a much more constrained way than PBH. Other state-of-the-art algorithms tend to perform worse with increased dimension, only CMA-ES stays competitive or better than PBH. There is room for improvements for PBH with respect to CMA-ES as it performs adaptive sampling by changing its sampling distribution with respect to new information (which make it particularly effective on the rotated and anisotropic Rastrigin function). It is theoretically possible to do so with PBH, and thus a promising venue for future work. We display the results for the additive scaling law, for the three large deep learning models tested in (Shukor et al., 2025): large language model (LLM), native multi-modal model (NMM), and large vision model (LVM). We report the results for the additive scaling law, that yielded the best test error. As local minimization is performed by LBFGS, most other methods were not comparable. For almost all problems PBH with $N = 5$ is better than ZOP or BH (Table 2). Increasing the number of LBFGS steps did not yield better results (for PBH and BH), as increasing the number of samples (keeping CPU number fixed).

Rastrigin

Dim.	PBH	PBH-b	ZOP	BH	BO	DE	PSO	CMA
20	2	**1**	8	7	6	5	3	4
50	2	**1**	8	6	7	5	4	3
100	**1**	2	5	6	8	7	4	3
200	**1**	2	5	6	8	7	4	3

Griewank

PBH	PBH-b	ZOP	BH	BO	DE	PSO	CMA
4	3	8	**1**	7	2	6	5
4	3	8	5	7	6	**1**	2
3	5	3	6	8	7	**1**	**1**
1	**1**	**1**	6	8	7	5	**1**

Styblinski–Tang

Dim.	PBH	PBH-b	ZOP	BH	BO	DE	PSO	CMA
20	5	7	–	6	8	**1**	2	3
50	2	5	–	4	7	6	**1**	2
100	3	5	–	4	7	6	**1**	2
200	**1**	5	3	2	8	7	4	4

Rotated anisotropic Rastrigin

PBH	PBH-b	ZOP	BH	BO	DE	PSO	CMA
3	6	–	5	7	**1**	2	4
2	4	–	6	7	5	3	**1**
2	3	–	6	7	5	4	**1**
1	2	5	4	8	7	6	3

Table 1: Ranking of the algorithms on the four benchmark functions. Rank 1 corresponds to the best final objective value. Function values within machine precision are considered tied. $N = 10 \times d$, 10 local steps, CPU time $= 2 \times d$. We set $\gamma = 5$ and $\delta = 0.005$ at initialization.

Family	Domain	PBH (N=5)	PBH (N=10)	PBH (N=15)	ZOP (N=5)	BH
LLM ($d = 19$)	arxiv	7.244 ± 4.369	**6.848** ± 4.795	7.336 ± 5.859	7.431 ± 7.533	10.037 ± 5.472
	book	**3.671** ± 2.804	4.136 ± 2.778	3.752 ± 3.382	6.158 ± 6.523	4.915 ± 3.298
	c4	**4.105** ± 3.225	4.528 ± 3.622	4.168 ± 3.460	7.061 ± 6.521	6.977 ± 4.146
	commoncrawl	**3.964** ± 2.848	5.172 ± 3.235	4.009 ± 3.615	5.638 ± 6.520	6.843 ± 3.820
	github	**8.597** ± 7.418	9.899 ± 7.707	8.961 ± 8.241	11.131 ± 9.351	13.191 ± 8.458
	stackexchange	**6.109** ± 3.795	7.024 ± 4.270	6.374 ± 5.469	7.651 ± 6.644	8.640 ± 6.598
	wikipedia	**8.540** ± 5.088	9.484 ± 5.011	8.554 ± 5.340	10.416 ± 6.524	13.898 ± 5.064
LVM ($d = 13$)	alttext	**3.083** ± 3.745	6.037 ± 3.585	4.896 ± 3.674	6.354 ± 4.648	6.452 ± 4.459
	highquality1	5.753 ± 3.368	7.636 ± 3.903	6.474 ± 3.091	14.093 ± 6.332	**5.494** ± 3.799
	highquality2	**3.978** ± 2.948	8.189 ± 7.543	4.681 ± 7.894	14.201 ± 12.136	5.446 ± 4.691
	synthetic	11.384 ± 12.289	16.939 ± 13.689	**11.302** ± 11.621	24.611 ± 15.638	14.320 ± 7.138
NMM ($d = 11$)	captions	**3.725** ± 3.209	4.526 ± 3.400	4.074 ± 3.891	8.698 ± 7.172	5.741 ± 4.222
	interleaved	2.854 ± 1.329	3.299 ± 1.318	**2.693** ± 1.470	6.176 ± 2.263	5.618 ± 2.622
	text	2.387 ± 1.185	3.169 ± 1.320	**2.341** ± 1.313	4.500 ± 1.894	7.332 ± 3.689

Table 2: Test MRE $= \frac{|\text{prediction} - \text{observation}|}{\text{observation}}$ (%) for fitting additive scaling law. (`lbfgs steps=5`, `CPU time=300s`). We set $\gamma = 10$ and $\delta = 10^{-3}$ at initialization. Results averaged over three seeds, we report the standard deviation. On average, PBH yields better test error than other algorithms. The standard deviation is also indicative of the best possible value reachable.

5 Conclusion, Limitations and Broader Impact

Conclusion. In this paper, we present a new theoretical framework, Proximal Basin Hopping, from which we construct a practical algorithm outperforming state-of-the-art methods on various problems.

Limitations. With respect to the present convergence analysis, and the subsequent algorithm, a question can be raised about the fact that we aggregate points, instead of simply taking the best point every time. In this setting, we would lose the proximal interpretation which is vital for convergence to the global minimizer through the parameter γ, but the practical results may be better. It has been shown for the basin hopping algorithm that in some cases, better-tailored perturbation could improve optimization (Englander & Englander, 2014). Here, if we steer away from a Gaussian distribution for the stochastic exploration, we also steer away from the proximal interpretation and the convergence analysis of our algorithm. However, a variable metric (Chouzenoux et al., 2014) is possible by using $y \sim \mathcal{N}(x_k, \delta\gamma\Sigma)$ instead of $y \sim \mathcal{N}(x_k, \delta\gamma\mathrm{Id})$. It remains to choose Σ properly (Hansen & Ostermeier, 2001). Also, the impact of the inexactness of the local minimization should be investigated on a per-solver basis. A first step should be to investigate it for gradient flow. Significant work is required as the geometry of attraction basins should be described in far more detail than what we used here. Finally, if letting the number of samples be fixed simplifies greatly the computations, it could be interesting to adapt the number of samples on the fly (notably with Proposition 3), to temporarily escape some traps, or reduce the computational load if progress continues.

AI USE STATEMENT

In this work, we used generative AI tools to help implement some of the concurrent methods to our algorithm PBH. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, support qualitative and thematic data analysis, interpret results, and the rest of the required disclosure tasks related to datasets are not applicable to this work. Additionally, we used generative AI tools for identifying relevant literature, creating or editing software code, and editing the paper to improve readability and conciseness. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by the first author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All the claimed results in the paper have been proved and their proof can be found in appendix. An anonymous downloadable source code in python has been submitted as a part of supplementary materials. For the experiments on scaling laws, the data used is available openly on `https://github.com/apple/ml-scalefit`.

REFERENCES

J. A. S. Alvarez and P. Calaminici. A review of global optimization methods for molecular structures: Algorithms, applications and perspectives. *Journal of Computational Chemistry*, 46(28):e70243, 2025.

J. S. Arora, O. A. Elwakeil, A. I. Chahande, and C. C. Hsieh. Global optimization methods for engineering applications: a review. *Structural optimization*, 9(3):137–159, 1995.

D. Asenjo, J. D. Stevenson, D. J. Wales, and D. Frenkel. Visualizing basins of attraction for different minimization algorithms. *The Journal of Physical Chemistry B*, 117(42):12717–12723, 2013.

M. Audin, M. Damian, and E. Erné. *Morse theory and Floer homology*, volume 2. Springer, 2014.

W. Azizian, F. Iutzeler, J. Malick, and P. Mertikopoulos. What is the long-run distribution of stochastic gradient descent? a large deviations analysis. In *ICML 2024-41st International Conference on Machine Learning*, pp. 1–70, 2024.

W. Azizian, F. Iutzeler, J. Malick, and P. Mertikopoulos. The global convergence time of stochastic gradient descent in non-convex landscapes: Sharp estimates via large deviations. In *International Conference on Machine Learning*, pp. 1982–2044. PMLR, 2025.

M. Baioletti, V. Santucci, and M. Tomassini. A performance analysis of basin hopping compared to established metaheuristics for global optimization. *Journal of Global Optimization*, 89(3):803–832, 2024.

C. Bouttier, T. Cesari, M. Ducoffe, and S. Gerchinovitz. Regret analysis of the Piyavskii-Shubert algorithm for global Lipschitz optimization. *arXiv preprint arXiv:2002.02390*, 2020.

E. Chouzenoux, J.-C. Pesquet, and A. Repetti. Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function. *Journal of Optimization Theory and Applications*, 162(1):107–132, 2014.

P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale modeling & simulation*, 4(4):1168–1200, 2005.

M. Coste. *An introduction to o-minimal geometry*. 1999.

B. de Souza. Goat: A global optimization algorithm for molecules and atomic clusters. *Angewandte Chemie International Edition*, 64(18):e202500393, 2025.

F. den Hollander. *Large Deviations*. American Mathematical Society, 2000.

N. Di, E. C. Chi, and S. W. Fung. A monte carlo approach for nonsmooth convex optimization via proximal splitting algorithms. *arXiv preprint arXiv:2509.07914*, 2025.

A. C. Englander, J. A. Englander, and M. J. Carter. Hopping with an adaptive hop probability distribution. In *Astrodynamics Specialist Meeting*, 2020.

J. A. Englander and A. C. Englander. Tuning monotonic basin hopping: improving the efficiency of stochastic search as applied to low-thrust trajectory optimization. In *International Symposium on Space Flight Dynamics 2014*, number GSFC-E-DAA-TN14154, 2014.

J. F. Fernando. On the set of local extrema of a subanalytic function. *Collectanea mathematica*, 71 (1):1–24, 2020.

M. Fornasier, T. Klock, and K. Riedl. Consensus-based optimization methods converge globally. *SIAM Journal on Optimization*, 34(3):2973–3004, 2024.

I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

M. Goodridge, J. Moriarty, J. Vogrinc, and A. Zocca. Hopping between distant basins. *Journal of Global Optimization*, 84(2):465–489, 2022.

A. Grosso, M. Locatelli, and F. Schoen. A population-based approach for hard global optimization problems based on dissimilarity measures. *Mathematical Programming*, 110(2):373–404, 2007.

A. György and L. Kocsis. Efficient multi-start strategies for local search algorithms. *Journal of Artificial Intelligence Research*, 41:407–444, 2011.

B. Hajek. Cooling schedules for optimal annealing. *Mathematics of operations research*, 13(2): 311–329, 1988.

N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, 9(2):159–195, 2001.

P. Hansen, B. Jaumard, and S.-H. Lu. Global optimization of univariate Lipschitz functions: I. Survey and properties. *Mathematical programming*, 55(1):251–272, 1992.

C. M. Kellett. A compendium of comparison function results. *Mathematics of Control, Signals, and Systems*, 26(3):339–374, 2014.

J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks*, volume 4, pp. 1942–1948. ieee, 1995.

E. Kosygina and T. Mountford. Introductory examples and definitions. Cramér's theorem, 2018. Lecture notes.

A. Lambora, K. Gupta, and K. Chopra. Genetic algorithm-a literature review. In *2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pp. 380–384. IEEE, 2019.

G. Lauga and S. Vaiter. Characterizations of inexact proximal operators. working paper or preprint, January 2026. URL `https://hal.science/hal-05449026`.

J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pp. 1246–1257. PMLR, 2016.

A. B. Levy. *Attraction in numerical minimization: iteration mappings, attractors, and basins of attraction*. Springer, 2018.

A. B. Levy. Analyzing basins of attraction in numerical minimization. *Set-Valued and Variational Analysis*, 34(1):6, 2026.

M. Locatelli and F. Schoen. (global) optimization: historical notes and recent developments. *EURO Journal on Computational Optimization*, 9:100012, 2021.

D. Morales, P. Pérez-Aros, and E. Vilches. Convergence rates for stochastic proximal and projection estimators. *arXiv preprint arXiv:2602.06750*, 2026.

G. Ochoa, S. Verel, F. Daolio, and M. Tomassini. Local optima networks: A new model of combinatorial fitness landscapes. In *Recent advances in the theory and application of fitness landscapes*, pp. 233–262. Springer, 2014.

S. Osher, H. Heaton, and S. W. Fung. A hamilton–jacobi-based proximal operator. *Proceedings of the National Academy of Sciences*, 120(14), 2023.

L. Papenmeier, M. Poloczek, and L. Nardi. Understanding high-dimensional bayesian optimization. In *Forty-Second International Conference on Machine Learning*, 2025.

N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

M. C. Prentiss, D. J. Wales, and P. G. Wolynes. Protein structure prediction using basin-hopping. *The Journal of chemical physics*, 128(22), 2008.

Q. Rebjock and N. Boumal. Fast convergence to non-isolated minima: four equivalent conditions for c^2 functions. *Mathematical Programming*, 213(1):151–199, 2025.

R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

C. Rohilla Shalizi. General theory of large deviations. Lecture notes, 36-754 Statistical Machine Learning, Carnegie Mellon University, 2006. URL `https://www.stat.cmu.edu/~cshalizi/754/2006/notes/lecture-30.pdf`. Chapter 30.

M. Shukor, L. Béthune, D. Busbridge, D. Grangier, E. Fini, A. El-Nouby, and P. Ablin. Scaling laws for optimal data mixtures. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

J. Snoek, H. Larochelle, and R. Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

R. Storn and K. Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.

Y. Sun, Á. Baricz, and S. Zhou. On the monotonicity, log-concavity, and tight bounds of the generalized marcum and nuttall q-functions. *IEEE Transactions on Information Theory*, 56(3): 1166–1186, 2010.

M. Tomassini. A local optima network view of real function fitness landscapes. *Entropy*, 24(5):703, 2022.

D. J. Wales. Energy landscapes: some new horizons. *Current Opinion in Structural Biology*, 20(1): 3–10, 2010.

D. J. Wales and J. P. K. Doye. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A*, 101 (28):5111–5116, 1997.

Z. B. Zabinsky. Random search algorithms. *Department of Industrial and Systems Engineering, University of Washington, USA*, 34, 2009.

M. Zhang, F. Han, Y. T. Chow, S. Osher, and H. Schaeffer. Inexact proximal point algorithms for zeroth-order global optimization. *arXiv preprint arXiv:2412.11485*, 2024.

CONTENTS

1 Introduction **1**

2 Structural assumptions on f and the local solver **3**

3 Proposed method **4**

3.1 Proximal basin hopping with exact expectation. 5

3.2 Proximal basin hopping with approximated expectation. 6

3.3 Proximal basin hopping with approximated expectation and local minimization. . . 7

4 Numerical experiments **8**

5 Conclusion, limitations and broader impact **9**

A Related works **13**

B Details about our algorithm **15**

C Preliminary results and discussion of the assumptions **17**

C.1 Assumptions on f 17

C.2 Properties of the solver 18

C.3 Some preliminary results related to the local minimizers 18

D Proofs of Section 3: Proposed method **21**

D.1 Proofs for ideal basin hopping 21

D.2 Proofs for exact expectation basin hopping 23

D.3 Proofs for approximate expectation PBH 29

D.4 Proofs for inexact solver PBH 32

E Additional results **36**

E.1 Ablation study 36

E.2 Scaling laws 36

A RELATED WORKS

Global optimization and zeroth-order methods. Global optimization has a long history, with most used methods being derived from some metaheuristics (Locatelli & Schoen, 2021). A large part of this literature can be viewed as zeroth-order optimization, as the algorithms only query function values. The simplest one being pure random search, which samples candidate points and keeps the best value found (Zabinsky, 2009). More structured deterministic zeroth-order methods include Lipschitz global optimization, but are typically limited by dimension (Hansen et al., 1992; Bouttier et al., 2020). Multi-start local search is another way to combine global exploration and local descent, and its efficiency depends obviously on how starting points are allocated (György & Kocsis, 2011).

Genetic and evolutionary algorithms maintain and transform a population of candidate solutions through selection and variation mechanisms (Lambora et al., 2019). Simulated annealing instead

constructs a stochastic process whose ability to escape local minima is controlled by a temperature parameter; classical convergence guarantees require sufficiently slow cooling schedules (Hajek, 1988) if the number of states to investigate is finite. Particle swarm optimization updates a population of particles using both individual and collective best positions (Kennedy & Eberhart, 1995). Consensus-based optimization is another population-based approach in which particles concentrate around weighted averages favoring low objective values; it is a more theoretically principled way of doing particles optimization, as global convergence under suitable assumptions was derived (Fornasier et al., 2024). Bayesian optimization and related surrogate-based methods are highly effective in low to moderate dimension, but their behavior in high dimension remains delicate (Papenmeier et al., 2025).

Energy landscapes, basins, and local optima networks. A complementary line of work, that has attracted most of its attention from chemistry fields, studies the organization of nonconvex landscapes through their local minima and basins of attraction. Local optima networks encode local minimizers as nodes and transitions between them as edges, giving a graph representation of the landscape (Ochoa et al., 2014; Tomassini, 2022). Such tools have been used to visualize and quantify funneling landscapes, in which local descent tends to guide configurations toward increasingly low-energy regions (Wales, 2010). The structure of attraction basins evidently depends on the local solver: different minimization algorithms may induce different basins on the same objective (Asenjo et al., 2013). This viewpoint is closely related to the study of iteration maps, attractors, and basins of attraction in numerical minimization (Levy, 2018; 2026).

Basin-hopping algorithms. Basin hopping was introduced for the global optimization of atomic cluster energies, in particular Lennard–Jones clusters (Wales & Doye, 1997). The central idea is to compose random perturbations with local minimization, thereby transforming the original objective into an energy landscape over local minima. This transformation has proved especially effective in molecular and cluster optimization (Prentiss et al., 2008; Alvarez & Calaminici, 2025; de Souza, 2025). Despite its practical success, this scheme has limited general convergence theory and is typically analyzed or tuned empirically.

Several variants of basin hopping have since been created, due to its success. Monotonic basin hopping accepts only moves that improve the objective and has been tuned for trajectory optimization problems (Englander & Englander, 2014). Adaptive hopping strategies modify the perturbation distribution to improve transitions between promising regions (Englander et al., 2020). Population-based basin hopping maintains several candidate solutions and uses dissimilarity or diversity mechanisms to explore multiple regions of the landscape (Grosso et al., 2007). Recent comparisons show that basin hopping is competitive with established metaheuristics on global optimization benchmarks, but its performance remains problem dependent (Baioletti et al., 2024). There is also theoretical work studying the difficulty of moving between distant basins (Goodridge et al., 2022).

Proximal algorithms and stochastic proximal approximations. Proximal algorithms are a central tool in convex and variational optimization (Combettes & Wajs, 2005; Parikh & Boyd, 2014; Rockafellar & Wets, 2009). In convex settings, inexact proximal splitting methods remain convergent when the errors in the proximal evaluations are controlled, for instance by summability assumptions (Combettes & Wajs, 2005). Recent work also studies Monte Carlo approximations of proximal steps for nonsmooth convex optimization (Di et al., 2025).

A Hamilton–Jacobi-based zeroth-order approximation of the proximal operator was proposed in (Osher et al., 2023):

$$\operatorname{prox}^{\delta}_{\lambda f}(x) = \frac{\mathbb{E}_{y\sim\mathcal{N}(x,\delta\lambda I)}\left[y\exp(-f(y)/\delta)\right]}{\mathbb{E}_{y\sim\mathcal{N}(x,\delta\lambda I)}\left[\exp(-f(y)/\delta)\right]}, \qquad \operatorname{prox}_{\lambda f}(x) = \lim_{\delta\downarrow 0}\operatorname{prox}^{\delta}_{\lambda f}(x). \tag{18}$$

Rates for such stochastic proximal and projection estimators have been investigated in (Lauga & Vaiter, 2026; Morales et al., 2026). This approximation has also been used for zeroth-order global optimization in (Zhang et al., 2024), where an adaptive proximal-point scheme encourages exploration near non-global critical points. Their implementation estimates the expectation either by tensor-train methods or by Monte Carlo integration. The convergence argument relies on decreasing the smoothing parameter and increasing the number of samples sufficiently fast so that the accumulated proximal error remains controlled, for example through summability conditions (Zhang et al., 2024, Corollary 4).

B DETAILS ABOUT OUR ALGORITHM

In practice. Instead of decreasing δ at each iteration, we adjust it on the fly: if the minimum value of f identified by the sampling mechanism has not lead to concentration on this minimum value, we decrease δ accordingly. This small tweak is a practical tradeoff given that we are using small sample sizes.

Furthermore, for numerical stability, the weighted mean is computed by substracting the maximum weight value, before computing the ratio. It is also done for the ZOP algorithm (for the same reasons).

Proposition 4. *Suppose that Assumptions 1, 2, and 3 hold. Let $N \in \mathbb{N} \setminus \{0\}$, let $\{m_i\}_{1 \leq i \leq N} \in \mathcal{M}_*$, and assume that there exists $1 \leq i_x \leq N$ such that $f(m_{i_x}) < f(m_i)$ for all $i \neq i_x$, with $m_{i_x} \in \mathbb{B}(m_{i,x}, r_{i,x}) \subseteq \mathrm{Att}(m_{i_x})^\circ$. If δ is small enough, i.e., such that*

$$e^{-\Delta/\delta} \leq \frac{r_{i_x}}{(N-1)\max_{1 \leq i \leq N} \|m_i - m_{i_x}\|}, \tag{19}$$

where $\Delta = \min_{1 \leq i \leq N, i \neq i_x} f(m_i) - f(m_{i_x})$, then

$$T_f\left(\frac{\sum_{i=1}^N m_i \exp(-f(m_i)/\delta)}{\sum_{i=1}^N \exp(-f(m_i)/\delta)}\right) = m_{i_x}. \tag{20}$$

Proof. We note

$$m_\delta := \frac{\sum_{i=1}^N m_i \exp(-f(m_i)/\delta)}{\sum_{i=1}^N \exp(-f(m_i)/\delta)} \text{ and } \lambda_i = \frac{\exp(-f(m_i)/\delta)}{\sum_{i=1}^N \exp(-f(m_i)/\delta)}$$

We need to show that $m_\delta \in \mathrm{Att}(m_{i_x})$ to conclude. m_{i_x} being in the interior of $\mathrm{Att}(m_{i_x})$ there exist $r_{i_x} > 0$ such that $\mathbb{B}(m_{i_x}, r_{i_x}) \subset \mathrm{Att}(m_{i_x})$. By taking this radius as large as possible, we upper bound the value of δ required. We want:

$$\left\| \sum_{i=1}^N \lambda_i m_i - m_{i_x} \right\| \leq r_{i_x} \Leftrightarrow \left\| \sum_{i=1, i \neq i_x}^N \lambda_i (m_i - m_{i_x}) \right\| \leq r_{i_x}$$

Now,

$$\left\| \sum_{i=1, i \neq i_x}^N \lambda_i (m_i - m_{i_x}) \right\| \leq \sum_{i=1, i \neq i_x}^N \lambda_i \|m_i - m_{i_x}\| \leq (1 - \lambda_{i_x}) \max_{1 \leq i \leq N} \|m_i - m_{i_x}\|.$$

Δ is stricly greater than 0 by assumption. We have

$$\lambda_{i_x} = \frac{1}{1 + \sum_{i \neq i_x} \exp(-(f(m_i) - f(m_{i_x}))/\delta)},$$

hence,

$$1 - \lambda_{i_x} \leq \sum_{i \neq i_x} \exp(-(f(m_i) - f(m_{i_x}))/\delta) \leq (N-1)\exp(-\Delta/\delta).$$

Thus for concentration to occur, we need δ small enough so that

$$e^{-\Delta/\delta} \leq \frac{r_{i_x}}{(N-1)\max_{1 \leq i \leq N} \|m_i - m_{i_x}\|}.$$

□

This threshold on δ is a worst case, a multiplicity of m_{i_x} reduces the mass on "bad" local minimizers while increasing the mass on the correct local minimizer. An other aspect that is not directly covered by Proposition 3 is the dependence on the dimension for $\alpha(\varepsilon)$. We can at least quantify the probability of eating the basin of attraction for one sample. It is reasonable to assume that we can fit a ball of non zero radius in each basin of attraction and we can estimate the probability of one sample being in this ball explicitly, and subsequently of at least one among N. We plot in Figure 2 an estimation of these probabilities with respect to the dimension and the number of samples.

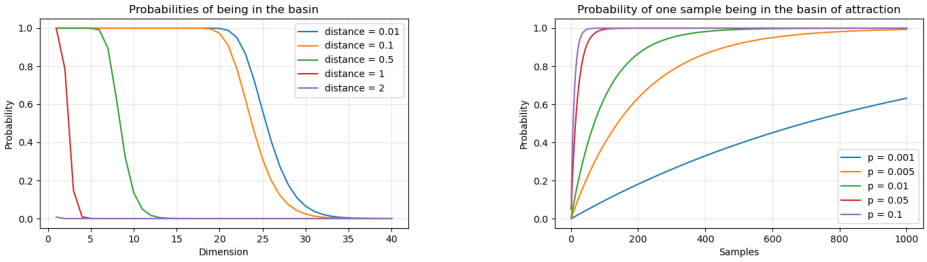


Figure 2: Probability of being in a ball in the basin of x_* and $r^2/\sigma^2 = 1$.

Lemma 4. *Let $x \in \mathbb{R}^d$, $\sigma > 0$ and $Y_\sigma \sim \mathcal{N}(x, \sigma^2 \mathrm{Id})$. Let $r > 0$. We have*

$$\mathbb{P}(Y_\sigma \in \mathbb{B}(0,r)) = \mathrm{CDF}_{\mathcal{X}'_d(\|x\|^2/\sigma^2)}(\frac{r^2}{\sigma^2}). \tag{21}$$

Moreover to have a probability $0 < \alpha < 1$ of one sample hitting this ball, we need

$$N \geq \frac{\log(1-\alpha)}{\log(1 - \mathbb{P}(Y_\sigma \in \mathbb{B}(0,r)))}. \tag{22}$$

Proof. We have

$$\mathbb{P}(Y_\sigma \in \mathbb{B}(0,r)) = \mathbb{P}(\|Y_\sigma\| \leq r) = \mathbb{P}(\|Y_\sigma\|^2 \leq r^2) = \mathbb{P}(\|Y_\sigma\|^2/\sigma^2 \leq r^2/\sigma^2)$$

Now, $\|Y_\sigma\|^2/\sigma^2$ is the sum of d independent standard normal distributions, with non zero means. Hence, it is a non central chi-square distribution:

$$\frac{\|Y_\sigma\|^2}{\sigma^2} \sim \mathcal{X}'_d(\lambda), \text{ where } \lambda = \sum_{i=1}^{d} \frac{\mu_i^2}{\sigma^2}. \tag{23}$$

Therefore,

$$\mathbb{P}(Y_\sigma \in \mathbb{B}(0,r)) = \mathrm{CDF}_{\mathcal{X}'_d(\lambda)}(\frac{r^2}{\sigma^2}).$$

Setting α as the probability of one sample to be in this ball when drawing $(y_i)_{1 \leq i \leq N}$ samples of Y_σ, it is defined as:

$$\mathbb{P}(\exists\, 1 \leq i \leq N, y_i \in \mathbb{B}(0,r)) = 1 - (1 - \mathbb{P}(Y_\sigma \in \mathbb{B}(0,r)))^N = \alpha,$$

which yields after a few rewriting:

$$N \geq \frac{\log(1-\alpha)}{\log(1 - \mathbb{P}(Y_\sigma \in \mathbb{B}(0,r)))}. \tag{24}$$

□

For instance, if $\mathbb{P}(Y_\sigma \in \mathbb{B}(0,r))$ is small and $\alpha = 0.95$ we obtain that N needs to be of the order

$$N \gtrsim \frac{3}{\mathbb{P}(Y_\sigma \in \mathbb{B}(0,r))}. \tag{25}$$

Therefore we can guarantee with high probability that we can identify basin of attractions of better local minimizers (from a local minimizer), provided that the number of samples is high enough. One should also be wary of the value of σ not to spread our samples to much. In order to quantify that more precisely we can use the fact that the cumulative distribution function of the non-central chi-squared distribution can be expressed with the generalized Marcum Q-function (Sun et al., 2010).

Lemma 5. *Let $x \in \mathbb{R}^d \setminus \{0\}$, $\sigma > 0$ and $Y_\sigma \sim \mathcal{N}(x, \sigma^2 \mathrm{Id})$. Let $r > 0$. The probability $\mathbb{P}(Y_\sigma \in \mathbb{B}(0,r))$ is, with the other parameters fixed, strictly decreasing in the dimension d and the norm $\|x\|$, and stricly increasing in the ratio r/σ. On the other hand, with other parameters fixed and $\|x\| > r$,*

$$\lim_{\sigma \to 0} \mathbb{P}(Y_\sigma \in \mathbb{B}(0,r)) = 0 \text{ and } \lim_{\sigma \to +\infty} \mathbb{P}(Y_\sigma \in \mathbb{B}(0,r)) = 0, \tag{26}$$

implying that there exists σ_* such that

$$(\forall \sigma > 0),\ \mathbb{P}(Y_{\sigma_*} \in \mathbb{B}(0,r)) \geq \mathbb{P}(Y_\sigma \in \mathbb{B}(0,r)) \tag{27}$$

Proof. We have

$$\mathrm{CDF}_{\mathcal{X}'^2_d(\|x\|^2/\sigma^2)}(\frac{r^2}{\sigma^2}) = 1 - Q_{\frac{d}{2}}\left(\frac{\|x\|}{\sigma}, \frac{r}{\sigma}\right).$$

The first result follows from (Sun et al., 2010, Theorem 1(a)) as all parameters are strictly positive. The second one is immediate from the fact that $\|x\| > r$. □

C Preliminary results and discussion of the assumptions

C.1 Assumptions on f

Finitely many minimizing components in any compact set. As stated, f is tame, meaning for any compact set K

1. K is definable in some o-minimal structure over $\mathbb{R}$, and $f|_K : K \to \mathbb{R}$ is definable: the set of local minima is a definable subset of $\mathbb{R}^d$ (Fernando, 2020, Lemma 1.7) and definable subset have finitely many connected components (Coste, 1999, Section 3.2).
2. K is semi-algebraic, and $f|_K$ is semi-algebraic: semi-algebraic function are definable (Fernando, 2020).
3. K is globally subanalytic, and $f|_K$ is globally subanalytic: same as above (Fernando, 2020)
4. $f|_K$ is a Morse function, i.e., it is C^2 with only isolated nondegenerate critical points (Audin et al., 2014).
5. $f|_K$ is a Morse-Bott function, i.e., it is C^2, and each connected component of minima is a critical submanifold (Rebjock & Boumal, 2025). This is the case of the Rastrigin function.

Graph of local minimizers. Following the basin hopping literature (Hajek, 1988; Wales, 2010; Tomassini, 2022), and some other recent works (Azizian et al., 2025), in practice we somehow assume that local minimizers of f can be represented with a graph. This assumption is however not required to prove convergence. Here, the edges depend on the local solver T_f in the following sense: for any local minimizer $P \in \mathcal{M}$, the distance from p to the attractor of a better local minimizer Q (in the sense that $f(P) > f(Q)$) is uniformly bounded for all P.

Assumption 7. *For all $P \in \mathcal{M} \setminus \{x_*\}$, there exists $Q \in \mathcal{M}$ such that*

$$\sup_{x \in P} \mathrm{dist}(x, \overline{\mathrm{Att}(Q)}) \leq d_f < +\infty \text{ and } f(P) > f(Q) \tag{28}$$

Hence, we construct our graph with vertices as local minimizers (or connected components of local minimizers) and a directed edge between two vertices P and Q if $\mathrm{dist}(P, \overline{\mathrm{Att}(Q)}) \leq d_f$. The edge is undirected if also $\mathrm{dist}(Q, \overline{\mathrm{Att}(P)}) \leq d_f$. From this assumption, the global minimizer x_* is reachable by "simply" jumping from one local minimizer to a better local minimizer. This assumption is implicitly made in basin hopping optimization so that random perturbations can reach the basin of a better local minimizer (Goodridge et al., 2022).

We can numerically compute an upper bound of d_f (Assumption 7) for the $1-D$ Rastrigin function only has isolated local minimizers, and isolated local maximizers:

$$\mathrm{Ras}(x) = x^2 + 10(1 - \cos(2\pi x)), \tag{29}$$

which are solutions of

$$\frac{x}{10\pi} = -\sin(2\pi x).$$

These solutions can be computed numerically and are approximately away from each other by a little less than 0.5, hence $d_f = \frac{1}{2}$ is a valid upper bound in this case.

C.2 Properties of the solver

The definition of the solver in the main body of the paper contains only the relevant assumption to derive our analysis, we lay here a more detailed presentation of what these assumptions entail. Assume that f is C^2-smooth (with Lipschitz continuous gradient). Let us define $\varphi : [0, +\infty), \mathbb{R}^d \mapsto \mathbb{R}^d$ the gradient flow mapping any $x \in \mathbb{R}^d$ to a local minimizer of f:

$$\begin{cases} \varphi(0, x) & = x \\ \varphi(t, x) & = -\nabla f(\varphi(t, x)) \end{cases} \tag{30}$$

Define the critical components of f as:

$$\mathrm{crit} f := \left\{ x \in \mathbb{R}^d, 0 = \nabla f(x) \right\} \tag{31}$$

and we assume it is composed of infinitely many connected components. Following this definition, we have for any component $C \in \mathrm{crit} f$ that

$$\mathrm{Att}(C) := \left\{ x \in \mathbb{R}^d, \lim_{t \to +\infty} \mathrm{dist}(\varphi(t, x), C) = 0 \right\}. \tag{32}$$

(Azizian et al., 2024, Lemma D.28) holds in our context and thus for any $x \in \mathbb{R}^d$, there exists $C \in \mathrm{crit} f$ such that

$$\lim_{t \to +\infty} d(\varphi(t, x), C) = 0. \tag{33}$$

The proof is immediate from (Azizian et al., 2024, Lemma D.28), as f is coercive and the components C are connected. The finiteness of the number of components is not important, their connectedness is. $\mathcal{M}$ is the set of asymptotically stable critical components (Azizian et al., 2024, Lemma D.29) (again the proof does not rely on finiteness of critical components), implying that for all $M \in \mathcal{M}$, $\mathrm{Att}(M)$ is open. Under standard assumptions (Lee et al., 2016) any $C \in \mathrm{crit} f \setminus \mathcal{M}$, $\mathrm{Att}(C)$ is of measure 0.

C.3 Some preliminary results related to the local minimizers

We start this presentation by listing some important properties of the local minimizers and their attractor sets.

Lemma 6. *Suppose that Assumptions 1, 3, and 5 hold. The sets $\{\mathrm{Att}(M)\}_{M \in \mathcal{M}}$ form a measurable partition of $\mathbb{R}^d$ up to a null set.*

Proof. First, the sets of attractors of critical points who are non local minimizers is of measure 0. Let us denote by N this set of points which are not attracted by a local minimizer through T_f. We have for all $y \in \mathbb{R}^d \setminus N$, $T_f(y) \in \mathcal{M}_*$ and more precisely, there exists a unique $M \in \mathcal{M}$ such that $T_f(y) \in M$ as the M are pairwise disjoints. Hence, for any $P, Q \in \mathcal{M}$, $\mathrm{Att}(P) \cap \mathrm{Att}(Q) = \emptyset$, and

$$\bigcup_j \mathrm{Att}(M_j) = T_f^{-1}(\mathcal{M}_*) = \mathbb{R}^d \setminus N. \tag{34}$$

As T_f is measurable and every M is Borel by assumption, we get the desired result. □

Lemma 7. *Let $x \in \mathbb{R}^d$. For every $\eta > 0$, there exists a finite subset $\mathcal{K}_\eta \subset \mathcal{M}$ such that*

$$\inf_{M \in \mathcal{M} \setminus \mathcal{K}_\eta} V_x(M) \geq \inf_{\widetilde{M} \in \mathcal{M}} V_x(\widetilde{M}) + \eta. \tag{35}$$

Proof. Since $\mathrm{dist}(x, \overline{\mathrm{Att}(M)})^2 \geq 0$, one has

$$V_x(M) \geq f(M), \qquad M \in \mathcal{M}.$$

Set

$$\alpha_x := \inf_{M \in \mathcal{M}} V_x(M).$$

By coercivity of f, there exists $R_\eta > 0$ such that

$$\|y\| \geq R_\eta \quad \Longrightarrow \quad f(y) \geq \alpha_x + \eta.$$

Let

$$\mathcal{K}_\eta := \{M \in \mathcal{M} : \ M \cap \overline{\mathbb{B}}(0, R_\eta) \neq \varnothing\}.$$

Since $\overline{\mathbb{B}}(0, R_\eta)$ is compact and only finitely many elements of $\mathcal{M}$ intersect a compact set, $\mathcal{K}_\eta$ is finite. Let $M \in \mathcal{M} \setminus \mathcal{K}_\eta$. Then $M \cap \overline{\mathbb{B}}(0, R_\eta) = \varnothing$, so for every $m \in M$ one has $\|m\| > R_\eta$. Hence $f(M) = f(m) \geq \alpha_x + \eta$. Therefore $V_x(M) \geq f(M) \geq \alpha_x + \eta$. Taking the infimum over $M \in \mathcal{M} \setminus \mathcal{K}_\eta$ gives

$$\inf_{M \in \mathcal{M} \setminus \mathcal{K}_\eta} V_x(M) \geq \alpha_x + \eta = \inf_{\widetilde{M} \in \mathcal{M}} V_x(\widetilde{M}) + \eta.$$

□

Any finite subset of $\mathcal{M}$ is automatically bounded by boundedness of every elements of $\mathcal{M}$ and they have a positive value gap with $\inf V_x$.

Fact 1. *Let $x \in \mathbb{R}^d$, $\alpha_x := \inf_{M \in \mathcal{M}} V_x(M)$, $\varepsilon > 0$, and*

$$\mathcal{M}_x = \bigcup_{M \in S_f^\gamma(x)} M, \text{ and } (\mathcal{M}_x)_\varepsilon = \left\{z \in \mathbb{R}^d, \ \mathrm{dist}(z, \mathcal{M}_x) < \epsilon\right\}.$$

Let $R > 0$. On $\mathbb{B}(0, R)$ there exists a positive gap of V_x between $M \in S_f^\gamma(x)$ and every other M that intersects $\mathbb{B}(0, R)$, i.e., there exists $c_{R,\varepsilon} > 0$ such that for every M in

$$M_{x,R,\varepsilon} := \left\{M \in \mathcal{M}, \ M \cap \mathbb{B}(0, R) \neq \emptyset, \ M \cap \left((\mathcal{M}_x)_\varepsilon\right)^{\complement} \neq \emptyset\right\},$$

$$V_x(M) \geq \alpha_x + c_{R,\epsilon}.$$

Lemma 8. *Suppose that Assumptions 1, 2, 3, and 5 hold. Let $x \in \mathbb{R}^d$ and $\delta, \gamma > 0$. If one of the conditions below hold*

- *There exist $C_1, C_2, C_3 > 0$ and $0 < \alpha < \beta$ such that*
$$(\forall z \in \mathbb{R}^d), \qquad C_1\|z\|^\alpha - C_2 \leq f(z) \leq C_3(1 + \|z\|)^\beta. \tag{36}$$
- *$\mathcal{M}$ has finitely many components.*

Then, $T_f : \mathbb{R}^d \to \mathcal{M}_*$ is integrable and square integrable against the Gaussian distribution, i.e.,

$$\int_{\mathbb{R}^d} \|T_f(y)\| e^{\frac{-1}{2\gamma\delta}\|y-x\|^2} dy < +\infty, \text{ and } \int_{\mathbb{R}^d} \|T_f(y)\|^2 e^{\frac{-1}{2\gamma\delta}\|y-x\|^2} dy < +\infty \tag{37}$$

Proof. Let us consider the second case first. Denote by J the number of components of $\mathcal{M}$. Then $\mathcal{M}_* = \bigcup_{j=1}^J \mathcal{M}_j$. By assumption any $M \in \mathcal{M}$ is bounded. Thus there exists some $R > 0$ such that $\mathcal{M}_* \subset \mathbb{B}(0, R)$. Hence,

$$\int_{\mathbb{R}^d} \|T_f(y)\| e^{-\frac{\|y-x\|^2}{2\delta\gamma}} dy \leq R \int_{\mathbb{R}^d} e^{-\frac{\|y-x\|^2}{2\delta\gamma}} dy < +\infty.$$

Now to the first case. We have for all $z \in \mathbb{R}^d$, $f(z) \geq f(T_f(z))$, therefore

$$C_1\|T_f(z)\|^\alpha - C_2 \leq f(T_f(z)) \leq f(z) \leq C_3(1 + \|z\|)^\beta$$

and,

$$\|T_f(z)\| \leq C_1^{-1/\alpha} \left(C_2 + C_3(1 + \|z\|^\beta)\right)^{1/\alpha}.$$

Hence,

$$\int_{\mathbb{R}^d} \|T_f(y)\| e^{-\frac{\|y-x\|^2}{2\delta\gamma}} dy \leq \int_{\mathbb{R}^d} C_1^{-1/\alpha} \left(C_2 + C_3(1 + \|y\|^\beta)\right)^{1/\alpha} e^{-\frac{\|y-x\|^2}{2\delta\gamma}} dy < +\infty.$$

The square case follows immediately from the same arguments. □

Assumption 6 and the first condition in Lemma 8 imply that f has polynomially bounded growth. This is slightly stronger than coercivity and is akin to the notion of radially unbounded function in control theory[4] (Kellett, 2014).

Lemma 9. *Suppose that Assumptions 1, 3, 5, and 6 hold. Let $x \in \mathbb{R}^d$, $Y_\delta \sim \mathcal{N}(x, \gamma\delta\mathrm{Id})$, $q > 0$ and $r > 0$. We have:*

$$\limsup_{\delta \downarrow 0} \delta \log \mathbb{E}\left[(1 + \|Y_\delta\|^q)\mathbf{1}_{\{\|Y_\delta\|>r\}}\right] \leq -\frac{(r - \|x\|^2)_+}{2\gamma}. \tag{38}$$

Proof. We have

$$\mathbb{E}\big[(1 + \|Y_\delta\|^q)\mathbf{1}_{\{\|Y_\delta\|>r\}}\big] = \frac{1}{(2\pi\delta\gamma)^{d/2}} \int_{\{\|y\|>r\}} (1 + \|y\|^q) \exp\left(-\frac{\|y - x\|^2}{2\gamma\delta}\right) dy.$$

Fix $R > r$. Splitting the domain gives $\{\|y\| > r\} = \{r < \|y\| \leq R\} \cup \{\|y\| > R\}$, hence we can split the expectation in

$$I_\delta(r) := \mathbb{E}\big[(1 + \|Y_\delta\|^q)\mathbf{1}_{\{\|Y_\delta\|>r\}}\big] = I_{\delta,1}(r, R) + I_{\delta,2}(R),$$

where

$$I_{\delta,1}(r, R) := \frac{1}{(2\pi\delta\gamma)^{d/2}} \int_{\{r<\|y\|\leq R\}} (1 + \|y\|^q) e^{-\frac{\|y-x\|^2}{2\gamma\delta}} dy$$

and

$$I_{\delta,2}(R) := \frac{1}{(2\pi\delta\gamma)^{d/2}} \int_{\{\|y\|>R\}} (1 + \|y\|^q) e^{-\frac{\|y-x\|^2}{2\gamma\delta}} dy.$$

On the bounded region $\{r < \|y\| \leq R\}$, we have $1 + \|y\|^q \leq 1 + R^q$. Moreover, if $\|y\| > r$, then $\|y - x\| \geq \|y\| - \|x\| \geq (r - \|x\|)_+$. Therefore

$$I_{\delta,1}(r, R) \leq \frac{(1 + R^q)\,\lambda_d(B(0, R))}{(2\pi\delta\gamma)^{d/2}} \exp\left(-\frac{(r - \|x\|)_+^2}{2\gamma\delta}\right).$$

Since $\delta \log(\delta^{-d/2}) \to 0$, it follows that

$$\limsup_{\delta \downarrow 0} \delta \log I_{\delta,1}(r, R) \leq -\frac{(r - \|x\|)_+^2}{2\gamma}.$$

Now onto $I_{\delta,2}(R)$. Assume $R > \|x\|$. For $\|y\| > R$, $\|y - x\| \geq R - \|x\|$, hence $e^{-\frac{\|y-x\|^2}{2\gamma\delta}} \leq e^{-\frac{(R-\|x\|)^2}{4\gamma\delta}} e^{-\frac{\|y-x\|^2}{4\gamma\delta}}$. Thus

$$I_{\delta,2}(R) \leq e^{-\frac{(R-\|x\|)^2}{4\gamma\delta}} \frac{1}{(2\pi\delta\gamma)^{d/2}} \int_{\mathbb{R}^d} (1 + \|y\|^q) e^{-\frac{\|y-x\|^2}{4\gamma\delta}} dy.$$

Using $\|y\|^q \leq \max\{1, 2^{q-1}\}(1 + \|y - x\|^q + \|x\|^q)$, we obtain $1 + \|y\|^q \leq C(1 + \|y - x\|^q)$ for some constant $C > 0$. Hence

$$I_{\delta,2}(R) \leq Ce^{-\frac{(R-\|x\|)^2}{4\gamma\delta}} \frac{1}{(2\pi\delta\gamma)^{d/2}} \int_{\mathbb{R}^d} (1 + \|y - x\|^q) e^{-\frac{\|y-x\|^2}{4\gamma\delta}} dy.$$

Making the change of variables $z = (y - x)/\sqrt{\delta}$, we obtain that the righthandside integral is in $O(\delta^{d/2}) + O(\delta^{(d+q)/2})$ hence bounded uniformly for $\delta \in (0, 1]$. Therefore there exists $C_R > 0$ such that

$$I_{\delta,2}(R) \leq C_R e^{-\frac{(R-\|x\|)^2}{4\gamma\delta}},$$

and thus

$$\limsup_{\delta \downarrow 0} \delta \log I_{\delta,2}(R) \leq -\frac{(R - \|x\|)^2}{4\gamma}.$$

[4] In this setting, ψ is also required to be positive (Kellett, 2014, Section 2).

Finally,

$$\limsup_{\delta\downarrow 0} \delta \log I_\delta(r) \leq \max\left\{ \limsup_{\delta\downarrow 0} \delta \log I_{\delta,1}(r,R),\ \limsup_{\delta\downarrow 0} \delta \log I_{\delta,2}(R) \right\}.$$

Hence

$$\limsup_{\delta\downarrow 0} \delta \log I_\delta(r) \leq \max\left\{ -\frac{(r-\|x\|)_+^2}{2\gamma},\ -\frac{(R-\|x\|)^2}{4\gamma} \right\}.$$

Choosing $R > r$ large enough so that

$$\frac{(R-\|x\|)^2}{4\gamma} > \frac{(r-\|x\|)_+^2}{2\gamma},$$

we conclude that

$$\limsup_{\delta\downarrow 0} \delta \log \mathbb{E}\big[(1+\|Y_\delta\|^q)\mathbf{1}_{\{\|Y_\delta\|>r\}}\big] \leq -\frac{(r-\|x\|)_+^2}{2\gamma}. \qquad \square$$

Remark 1. *Evolution of the set $\mathcal{M}_x$ with respect to γ. There exists approximately three regimes. When γ is small, V_x heavily penalizes distant local minimizers hence its* $\arg\min$ *is more likely to contain one unique solution. Slowly increasing γ makes it more likely that several competitors M can arise. Then as γ increases more, V_x is more likely to select a unique better local minimizer.*

The first regime is characterized in the next Lemma.

Lemma 10. *Suppose that Assumptions 1, 2, 3, 7, and 5 hold. Let $x \in \mathbb{R}^d$, $\gamma > 0$, and $P \in \mathcal{M}$ the closest (w.r.t. to* $\mathrm{Att}(P)$*) minimizing component. Denote by Q the closest (w.r.t.* $\mathrm{Att}(Q)$*) better local minimizing component of P in the sense of Assumption 7. If:*

$$\gamma < \frac{\mathrm{dist}(x, \overline{\mathrm{Att}(Q)})^2 - \mathrm{dist}(x, \overline{\mathrm{Att}(P)})^2}{2(f(P)-f(Q))} \tag{39}$$

then,

$$P = \arg\min_{M\in\mathcal{M}} V_x(M). \tag{40}$$

Proof. We have that for all $M \in \mathcal{M} \setminus \{Q\}$ such that $\mathrm{dist}(x, \overline{\mathrm{Att}(M)}) \leq d_f$ then $f(M) > f(Q)$. Furthermore, by Assumption 7, any better local minimizer in the vicinity of P has function value equal to that of Q in a radius d_f, hence we just need to make sure that Q does not belong in the $\arg\min$, giving immediately the positive upper bound on γ. $\square$

These scenarios where we have a series of equally distant and of equal function values are not possible indefinitely, first because there is a unique global minimizer, and second because the function f is coercive so it needs to grow to infinity, thus contradicting the existence of these increasingly distant local minimizers of decreasing function values.

D Proofs of Section 3: Proposed method

D.1 Proofs for ideal basin hopping

Proof of Lemma 1

Proof. We have for all $x \in \mathbb{R}^d$

$$\mathrm{prox}_{\gamma f\circ T_f}(x) = \arg\min_{y\in\mathbb{R}^d} \frac{1}{2\gamma}\|y-x\|^2 + f(T(y)),$$

and

$$S_f^\gamma(x) := \arg\min_{M\in\mathcal{M}} \frac{1}{2\gamma}\mathrm{dist}(x, \overline{\mathrm{Att}(M)})^2 + f(M).$$

Looking more closely at the first optimization problem, we have

$$\min_{y\in\mathbb{R}^d} \frac{1}{2\gamma}\|y-x\|^2 + f(T(y)) = \min_{M\in\mathcal{M}} \frac{1}{2\gamma}\mathrm{dist}(x, \overline{\mathrm{Att}(M)})^2 + f(M).$$

Hence, if for every $M \in S_f^\gamma(x)$, $\mathrm{proj}_{\overline{\mathrm{Att}(M)}}(x) \cap \mathrm{Att}(M) \neq \varnothing$,

$$\mathrm{prox}_{\gamma f\circ T_f}(x) = \bigcup_{M\in\arg\min_{M\in\mathcal{M}} \frac{1}{2\gamma}\mathrm{dist}(x,\overline{\mathrm{Att}(M)})^2+f(M)} \mathrm{proj}_{\overline{\mathrm{Att}(M)}}(x),$$

where proj denotes the orthogonal projection. Thus, as T_f maps to $\mathcal{M}$, we recover

$$T_f(\mathrm{prox}_{\gamma f\circ T_f}(x)) = \arg\min_{M\in\mathcal{M}} \frac{1}{2\gamma}\mathrm{dist}(x, \overline{\mathrm{Att}(M)})^2 + f(M). \qquad \square$$

Proof of Lemma 2

Proof. Let $x \in \mathbb{R}^d$. We have that for all $\gamma > 0$.

$$S_f^\gamma(x) = \arg\min_{M\in\mathcal{M}} \frac{1}{2\gamma}\mathrm{dist}(x, \overline{\mathrm{Att}(M)})^2 + f(M).$$

Hence,

$$\begin{aligned} x_* \in S_f^\gamma(x) &\Leftrightarrow \forall M \in \mathcal{M},\ \frac{1}{2\gamma}\mathrm{dist}(x, \overline{\mathrm{Att}(x_*)})^2 + f(x_*) \leq \frac{1}{2\gamma}\mathrm{dist}(x, \overline{\mathrm{Att}(M)})^2 + f(M) \\ &\Leftrightarrow \forall M \in \mathcal{M},\ \frac{1}{2\mu}\left[\mathrm{dist}(x, \overline{\mathrm{Att}(x_*)})^2 - \mathrm{dist}(x, \overline{\mathrm{Att}(M)})^2\right] \leq \gamma, \end{aligned}$$

where the last inequality was obtained using Assumption 2. The minimal required value of γ depends on the close and achievable minimizers. Indeed, all minimizing components such that $\mathrm{dist}(x, \overline{\mathrm{Att}(x_*)})^2 - \mathrm{dist}(x, \overline{\mathrm{Att}(M)})^2 \leq 0$ yields no constraint on γ. $\square$

Proof of Theorem 1.

Proof. First iterations are well defined as S_f^γ is non-empty and bounded (Assumption 5, Lemma 1). Second, if for any $K \in \mathbb{N}$, $x_K = x_*$ then $x_k = x_*$ for all $k > K$. Finally, by construction $f(x_{k+1}) \leq f(x_k)$, hence by coercivity of f (Assumption 1), the iterates are bounded. Denote by B_0 this bounded set. We can guarantee that there exists only finitely many failed iterations (where $f(x_+) = f(x_k)$).

Indeed, by boundedness of the iterates, there exists $d_f \geq 0$ such that for all $k \in \mathbb{N}$,

$$\mathrm{dist}(x_k, \overline{\mathrm{Att}(x_*)}) \leq d_f.$$

Let $P \in \mathcal{M}$, such that $P \cap B_0 \neq \emptyset$ and $P \neq \{x_*\}$. We have for any $x \in P$, $V_x(x_*) < V_x(P)$ if

$$\frac{\mathrm{dist}(x, \overline{\mathrm{Att}(x_*)})^2}{2\gamma} + f_{\min} < f(P)$$

which happens if $\gamma > \frac{d_f^2}{2(f(P)-f_{\min})}$. Thus, taking the infimum over all P not equal to $\{x_*\}$, we have $\gamma_s \geq \frac{d_f^2}{2\mu}$ implies that for all $x \in B_0$, $x_* \in S_f^{\gamma_s}(x)$. Let $K \in \mathbb{N}$. For any $\gamma > 0$,

$$\begin{aligned} &\gamma\eta^K \leq \gamma_s \\ \Leftrightarrow &K\log(\eta) \leq \log\left(\frac{\gamma_s}{\gamma}\right) \\ \Leftrightarrow &K \leq \frac{\log(\gamma_s/\gamma)}{\log(\eta)} \end{aligned}$$

Therefore, after at most $\left\lceil \frac{\log(\gamma_s/\gamma)}{\log(\eta)} \right\rceil_+$ failed iterations, x_* belongs to S_f^γ.

Now, by boundedness of the iterates, only finitely many $M \in \mathcal{M}$ can be visited. Hence, we conclude that convergence occurs in a finite number of iterations. $\square$

D.2 PROOFS FOR EXACT EXPECTATION BASIN HOPPING

Proof of Lemma 3.

Proof. By the small-noise Gaussian large deviation principle with speed δ^{-1}, and rate function $I_x(y) = \frac{1}{2\gamma}\|y - x\|^2$, for every Borel set $A \subset \mathbb{R}^d$ (den Hollander, 2000), [5]

$$-\inf_{y\in A^\circ} I_x(y) \le \liminf_{\delta\downarrow 0} \delta \log \mathbb{P}(Y_\delta \in A) \le \limsup_{\delta\downarrow 0} \delta \log \mathbb{P}(Y_\delta \in A) \le -\inf_{y\in \overline{A}} I_x(y).$$

Applying this with $A = \mathrm{Att}(M)$ yields

$$\begin{aligned} -\inf_{y\in \mathrm{Att}(M)^\circ} \frac{1}{2\gamma}\|y - x\|^2 &\le \liminf_{\delta\downarrow 0} \delta \log \mathbb{P}(Y_\delta \in \mathrm{Att}(M)) \\ &\le \limsup_{\delta\downarrow 0} \delta \log \mathbb{P}(Y_\delta \in \mathrm{Att}(M)) \le -\inf_{y\in \overline{\mathrm{Att}(M)}} \frac{1}{2\gamma}\|y - x\|^2. \end{aligned}$$

Which is equivalent to

$$\begin{aligned} -\frac{1}{2\gamma}\mathrm{dist}(x, \mathrm{Att}(M)^\circ)^2 &\le \liminf_{\delta\downarrow 0} \delta \log \mathbb{P}(Y_\delta \in \mathrm{Att}(M)) \\ &\le \limsup_{\delta\downarrow 0} \delta \log \mathbb{P}(Y_\delta \in \mathrm{Att}(M)) \le -\frac{1}{2\gamma}\mathrm{dist}(x, \overline{\mathrm{Att}(M)})^2. \end{aligned}$$

By the regularity assumption the lower and upper bounds coincide, and hence

$$\lim_{\delta\downarrow 0} \delta \log \mathbb{P}(Y_\delta \in \mathrm{Att}(M)) = -\frac{1}{2\gamma}\mathrm{dist}(x, \overline{\mathrm{Att}(M)})^2.$$

This proves the first claim. The equivalent exponential form follows immediately: if a scalar quantity a_δ satisfies $\delta \log a_\delta \to -c$, then $a_\delta = \exp\left(-\frac{1}{\delta}[c + o(1)]\right)$ (see (Shalizi, 2006, Definition 401)). □

First case: finite number of minimizers. If there are finitely many sets of local minimizers, then we can obtain an explicit concentration result inside the convex hull defined by the solutions in $\mathcal{M}_x$ (Fact 1), when δ goes to 0[6].

Proposition 5. *Let $x \in \mathbb{R}^d$, $\gamma > 0$, $\delta > 0$, and let $Y_\delta \sim \mathcal{N}(x, \delta\gamma I_d)$. Suppose that $\mathcal{M}$ has finitely many components, i.e.,*

$$\mathcal{M}_* = \bigcup_{1\le j\le J} M_j \tag{41}$$

for some $J \in \mathbb{N}$. Denote by I_x the set of indices of the minimizers of $\left\{f_j + \frac{1}{2\gamma}\mathrm{dist}(x, \mathrm{Att}(M_j))^2\right\}_{1\le j\le J}$, i.e., $I_x \subseteq \{1, \dots, J\}$. Then,

$$\frac{\mathbb{E}[T_f(Y_\delta)e^{-f(T_f(Y_\delta))/\delta}]}{\mathbb{E}[e^{-f(T_f(Y_\delta))/\delta}]} \xrightarrow[\delta\downarrow 0]{} \frac{\sum_{i\in I_x} e^{-f_i/\delta}\mathbb{P}(Y_\delta \in \mathrm{Att}(M_i))\mathbb{E}[T_f(Y_\delta)|Y_\delta \in \mathrm{Att}(M_i)]}{\sum_{i\in I_x} e^{-f_i/\delta}\mathbb{P}(Y_\delta \in \mathrm{Att}(M_i))}.$$

If I_x is a singleton $\{i_x\}$, and $M_{i_x} = \{m_{i_x}\}$ contains an isolated local minimizer then,

$$\frac{\mathbb{E}[T_f(Y_\delta)e^{-f(T_f(Y_\delta))/\delta}]}{\mathbb{E}[e^{-f(T_f(Y_\delta))/\delta}]} \xrightarrow[\delta\downarrow 0]{} m_{i_x}.$$

Proof. Since the local minimizers reached by T_f decompose into the finite disjoint union $\mathcal{M}$, the sets $(\mathrm{Att}_j)_{1\le j\le J}$ form a measurable partition of the subset of $\mathbb{R}^d$ on which T_f will land in $\mathcal{M}$. Denoting

[5] The small-noise Gaussian rate function can be identified directly from the explicit normal example in (Kosygina & Mountford, 2018, pp.4–5) with $X_n \sim \mathcal{N}(0, 1/n)$, which yields the rate function $I(x) = x^2/2$, and cf. (den Hollander, 2000, Exercise III.9, p.30).

[6] This result is similar in spirit to (Zhang et al., 2024, Corollary 1) even though proofs are completely different

f_j the value of f attained on M_j, we have

$$\begin{aligned}\mathbb{E}[T_f(Y_\delta)e^{-f(T_f(Y_\delta))/\delta}] &= \sum_{j=1}^{J}\mathbb{E}[T_f(Y_\delta)e^{-f(T_f(Y_\delta))/\delta}\mathbf{1}_{\{Y_\delta\in\mathrm{Att}(M_j)\}}]\\ &= \sum_{j=1}^{J}e^{-f_j/\delta}\mathbb{E}[T_f(Y_\delta)\mathbf{1}_{\{Y_\delta\in\mathrm{Att}(M_j)\}}]\\ &= \sum_{j=1}^{J}e^{-f_j/\delta}\mathbb{P}(Y_\delta\in\mathrm{Att}(M_j))\mathbb{E}[T_f(Y_\delta)|Y_\delta\in\mathrm{Att}(M_j)].\end{aligned}$$

Moreover, using Lemma 3, we obtain that

$$\delta\log\left(e^{-f_j/\delta}\mathbb{P}(Y_\delta\in\mathrm{Att}(M_j))\right) =_{\delta\to 0} -f_j - \frac{1}{2\gamma}\mathrm{dist}(x,\mathrm{Att}(M_j))^2,$$

implying

$$e^{-f_j/\delta}\mathbb{P}(Y_\delta\in\mathrm{Att}(M_j)) =_{\delta\to 0}\exp\left(-\frac{1}{\delta}\left[f_j+\frac{1}{2\gamma}\mathrm{dist}(x,\mathrm{Att}(M_j))^2+o(1)\right]\right).$$

Now, take the set I_x. For all $i\in I_x$, and $j\in J\setminus I_x$, we have

$$\delta\log\left(\frac{e^{-f_i/\delta}\mathbb{P}(Y_\delta\in\mathrm{Att}(M_i))}{e^{-f_j/\delta}\mathbb{P}(Y_\delta\in\mathrm{Att}(M_j))}\right)\xrightarrow[\delta\downarrow 0]{} f_j+\frac{1}{2\gamma}\mathrm{dist}(x,\mathrm{Att}(M_j))^2-f_i-\frac{1}{2\gamma}\mathrm{dist}(x,\mathrm{Att}(M_i))^2>0$$

implying that

$$\lim_{\delta\to 0}\frac{e^{-f_i/\delta}\mathbb{P}(Y_\delta\in\mathrm{Att}(M_i))}{e^{-f_j/\delta}\mathbb{P}(Y_\delta\in\mathrm{Att}(M_j))}=+\infty.$$

Hence, we obtain that

$$\lim_{\delta\to 0}\frac{\sum_{j\in J\setminus I_x}e^{-f_j/\delta}\mathbb{P}(Y_\delta\in\mathrm{Att}(M_j))}{\sum_{j\in J}e^{-f_j/\delta}\mathbb{P}(Y_\delta\in\mathrm{Att}(M_j))}=0.$$

The rest follows from the fact that no connected components of $\mathcal{M}$ is unbounded, guaranteeing that $\mathbb{E}[T_f(Y_\delta)|Y_\delta\in\mathrm{Att}(M_j)]$ is bounded for all j. □

This result highlights the need for the uniqueness of the global minimizer: the barycenter of several global minimizers does not make any sense unless these global minimizers belong to the same connected and convex set.

Proof of Proposition 1.

Proof. Fix $\eta>0$, and let $K_\eta\subset\mathcal{M}_*$ be a bounded set such that (Lemma 7)

$$\inf_{m\notin K_\eta}V_x(m)\geq\alpha_x+\eta.$$

Set

$$A_\eta:=\{y\in\mathbb{R}^d:\ T_f(y)\in K_\eta^{\complement}\}.$$

We look at our measure outside this bounded set

$$\begin{aligned}\nu_{x,\delta}(K_\eta^{\complement}) &= \mathbb{E}\left[e^{-f(T_f(Y_\delta))/\delta}\mathbf{1}_{\{T_f(Y_\delta)\in K_\eta^{\complement}\}}\right]\\ &= \frac{1}{(2\pi\delta\gamma)^{d/2}}\int_{A_\eta}\exp\left(-\frac{1}{\delta}\left[f(T_f(y))+\frac{\|y-x\|^2}{2\gamma}\right]\right)dy.\end{aligned}$$

We first study the term in the exponential. Set $g(y)=f(T_f(y))+\frac{\|y-x\|^2}{2\gamma}$. If $y\in A_\eta$, then $T_f(y)\in M\in K_\eta^{\complement}$, and since $y\in\mathrm{Att}(M)$,

$$\frac{\|y-x\|^2}{2\gamma}\geq\frac{\mathrm{dist}(x,\overline{\mathrm{Att}(M)})^2}{2\gamma}.$$

Hence,

$$g(y) = f(M) + \frac{\|y-x\|^2}{2\gamma} \geq f(M) + \frac{\operatorname{dist}(x, \overline{\operatorname{Att}(M)})^2}{2\gamma} = V_x(M) \geq \alpha_x + \eta.$$

Therefore, $\inf_{y \in A_\eta} g(y) \geq \alpha_x + \eta$. Now, to estimate the measure, we split our set with a ball of radius $R > 0$ in the following way:

$$A_\eta = (A_\eta \cap \mathbb{B}(0,R)) \cup \left(A_\eta \cap \mathbb{B}(0,R)^{\complement}\right).$$

The measure on the lefthandside of the union is easily bounded by the previous inequality as we have,

$$\frac{1}{(2\pi\delta\gamma)^{d/2}} \int_{A_\eta \cap \mathbb{B}(0,R)} e^{-g(y)/\delta}\, dy \leq \frac{\lambda_d(\mathbb{B}(0,R))}{(2\pi\delta\gamma)^{d/2}} e^{-(\alpha_x+\eta)/\delta}.$$

Since

$$\delta \log\left(\delta^{-d/2}\right) \xrightarrow[\delta \downarrow 0]{} 0,$$

it follows that

$$\limsup_{\delta \downarrow 0} \delta \log\left[\frac{1}{(2\pi\delta\gamma)^{d/2}} \int_{A_\eta \cap \mathbb{B}(0,R)} e^{-g(y)/\delta}\, dy\right] \leq -(\alpha_x + \eta).$$

f is bounded from below on $\mathbb{R}^d$, hence all y,

$$g(y) \geq f_{\min} + \frac{\|y-x\|^2}{2\gamma}.$$

Thus

$$\frac{1}{(2\pi\delta\gamma)^{d/2}} \int_{A_\eta \cap \mathbb{B}(0,R)^{\complement}} e^{-g(y)/\delta}\, dy \leq e^{-f_{\min}/\delta} \frac{1}{(2\pi\delta\gamma)^{d/2}} \int_{\mathbb{B}(0,R)^{\complement}} \exp\left(-\frac{\|y-x\|^2}{2\gamma\delta}\right) dy.$$

The last term is exactly

$$e^{-f_{\min}/\delta}\, \mathbb{P}(\|Y_\delta\| > R).$$

We can take $R > \|x\|$, large enough so that

$$f_{\min} + \frac{(R - \|x\|)^2}{2\gamma} > \alpha_x + \eta.$$

Since $\|y - x\| \geq R - \|x\|$ whenever $\|y\| > R$, the Gaussian tail satisfies (den Hollander, 2000)

$$\limsup_{\delta \downarrow 0} \delta \log \mathbb{P}(\|Y_\delta\| > R) \leq -\frac{(R - \|x\|)^2}{2\gamma},$$

and therefore

$$\limsup_{\delta \downarrow 0} \delta \log\left[e^{-f_{\min}/\delta} \mathbb{P}(\|Y_\delta\| > R)\right] \leq -f_{\min} - \frac{(R - \|x\|)^2}{2\gamma} < -(\alpha_x + \eta).$$

Combining the estimates on $A_\eta \cap \mathbb{B}(0,R)$ and $A_\eta \cap \mathbb{B}(0,R)^{\complement}$, we obtain $\limsup_{\delta \downarrow 0} \delta \log \nu_{x,\delta}(K_\eta^{\complement}) \leq -(\alpha_x + \eta)$. We now lower bound the denominator. Let $\varepsilon > 0$. By definition of α_x, there exists $M_\varepsilon \in \mathcal{M}$ such that

$$V_x(M_\varepsilon) \leq \alpha_x + \varepsilon.$$

Then

$$\nu_{x,\delta}(\mathcal{M}) = \mathbb{E}\left[e^{-f(T_f(Y_\delta))/\delta}\right] \geq e^{-f(m_\varepsilon)/\delta} \mathbb{P}(Y_\delta \in \operatorname{Att}(M_\varepsilon)).$$

By Lemma 3,

$$\lim_{\delta \downarrow 0} \delta \log \mathbb{P}(Y_\delta \in \operatorname{Att}(M_\varepsilon)) = -\frac{1}{2\gamma} \operatorname{dist}(x, \operatorname{Att}(M_\varepsilon))^2,$$

hence

$$\liminf_{\delta \downarrow 0} \delta \log \nu_{x,\delta}(\mathcal{M}) \geq -f(M_\varepsilon) - \frac{1}{2\gamma} \operatorname{dist}(x, \operatorname{Att}(M_\varepsilon))^2 = -V_x(M_\varepsilon) \geq -(\alpha_x + \varepsilon).$$

Letting $\varepsilon \downarrow 0$ gives

$$\liminf_{\delta \downarrow 0} \delta \log \nu_{x,\delta}(\mathcal{M}) \geq -\alpha_x.$$

Finally,

$$\mu_{x,\delta}(K_\eta^{\complement}) = \frac{\nu_{x,\delta}(K_\eta^{\complement})}{\nu_{x,\delta}(\mathcal{M})},$$

so

$$\limsup_{\delta \downarrow 0} \delta \log \mu_{x,\delta}(K_\eta^{\complement}) \leq -(\alpha_x + \eta) - (-\alpha_x) = -\eta.$$

In particular, $\mu_{x,\delta}(K_\eta^{\complement}) \to 0$, that is, $\mu_{x,\delta}(K_\eta) \to 1$. □

Proof of Proposition 6.

Proposition 6. *Let $x \in \mathbb{R}^d$, $\gamma > 0$, $q \geq 0$, and let $Y_\delta \sim \mathcal{N}(x, \delta\gamma I_d)$. Let $R > 0$. We have*

$$\limsup_{\delta \downarrow 0} \delta \log \left(\mathbb{E}[\|T_f(Y_\delta)\| e^{-f(T_f(Y_\delta))/\delta} \mathbf{1}_{\{\|T_f(Y_\delta)\| > R\}}] \right) \leq -f_{\min} - \frac{\left(\left(\frac{C_1 R^\alpha - C_2 - C_3}{C_3} \right)_+^{1/\beta} - \|x\|^2 \right)_+}{2\gamma},$$

and that

$$\liminf_{\delta \downarrow 0} \delta \log \mathbb{E}[e^{-f(T_f(Y_\delta))/\delta}] \geq -\alpha_x.$$

Then,

$$\limsup_{\delta \downarrow 0} \delta \log \left(\frac{\mathbb{E}[\|T_f(Y_\delta)\| e^{-f(T_f(Y_\delta))/\delta} \mathbf{1}_{\{\|T_f(Y_\delta)\| > R\}}]}{\mathbb{E}[e^{-f(T_f(Y_\delta))/\delta}]} \right) \leq \alpha_x - f_{\min} - \frac{\left(\left(\frac{C_1 R^\alpha - C_2 - C_3}{C_3} \right)_+^{1/\beta} - \|x\|^2 \right)_+}{2\gamma}.$$

Proof. We have that

$$\frac{\mathbb{E}[T_f(Y_\delta) e^{-f(T_f(Y_\delta))/\delta}]}{\mathbb{E}[e^{-f(T_f(Y_\delta))/\delta}]}$$

is well defined. Indeed, $0 \leq \mathbb{E}[e^{-f(T_f(Y_\delta))/\delta}] \leq e^{-f_{\min}/\delta} < +\infty$ and $\mathbb{E}[\|T_f(Y_\delta)\| e^{-f(T_f(Y_\delta))/\delta}] \leq e^{-f_{\min}/\delta} \mathbb{E}[\|T_f(Y_\delta)\|] < +\infty$ by Lemma 8. Using our measure introduced in Proposition 1 we have

$$\frac{\mathbb{E}[T_f(Y_\delta) e^{-f(T_f(Y_\delta))/\delta}]}{\mathbb{E}[e^{-f(T_f(Y_\delta))/\delta}]} = \int_{m \in \mathcal{M}_*} m d\mu_{x,\delta}(m)$$

Using similar arguments than in the proof of Lemma 8 we can have that $m \mapsto \|m\|$ is integrable against $\mu_{x,\delta}$. We have

$$\int_{m \in \mathcal{M}_*, \|m\| > R} \|m\| d\mu_{x,\delta}(m) = \frac{\mathbb{E}[\|T_f(Y_\delta)\| e^{-f(T_f(Y_\delta))/\delta} \mathbf{1}_{\{\|T_f(Y_\delta)\| > R\}}]}{\mathbb{E}[e^{-f(T_f(Y_\delta))/\delta}]}$$

As

$$\mathbb{E}[\|T_f(Y_\delta)\| e^{-f(T_f(Y_\delta))/\delta} \mathbf{1}_{\|T_f(Y_\delta)\| > R}] \leq e^{-f_{\min}/\delta} \mathbb{E}[C_1^{-1/\alpha} \left(C_2 + C_3(1 + \|Y_\delta\|^\beta) \right)^{1/\alpha} \mathbf{1}_{\|T_f(Y_\delta)\| > R}],$$

and,

$$\begin{aligned} & R < C_1^{-1/\alpha} \left(C_2 + C_3(1 + \|Y_\delta\|^\beta) \right)^{1/\alpha} \\ \Leftrightarrow & \left(\frac{C_1 R^\alpha - C_2 - C_3}{C_3} \right)^{1/\beta} < \|Y_\delta\|, \end{aligned}$$

implying that

$$\{\|T_f(Y_\delta)\| > R\} \subset \left\{ \|Y_\delta\| > \left(\frac{(C_1 R^\alpha - C_2 - C_3)_+}{C_3} \right)^{1/\beta} \right\}.$$

Therefore,

$$\mathbb{E}[\|T_f(Y_\delta)\|e^{-f(T_f(Y_\delta))/\delta}\mathbf{1}_{\|T_f(Y_\delta)\|>R}]$$
$$\leq e^{-f_{\min}/\delta}\mathbb{E}\left[C_1^{-1/\alpha}\left(C_2+C_3(1+\|Y_\delta\|^\beta)\right)^{1/\alpha}\mathbf{1}_{\left\{\|Y_\delta\|>\left(\frac{R^\alpha-C_2-C_3}{C_3}\right)^{1/\beta}\right\}}\right]$$

We can show that there exists $C_{2,\alpha}, C_{3,\alpha} > 0$ such that

$$C_1^{-1/\alpha}\left(C_2+C_3(1+\|Y_\delta\|^\beta)\right)^{1/\alpha} \leq C_{2,\alpha}+C_{3,\alpha}(1+\|Y_\delta\|^{\beta/\alpha}).$$

Now, using Lemma 9 with $q=\beta/\alpha$ and $r=\left(\frac{C_1R^\alpha-C_2-C_3}{C_3}\right)^{1/\beta}$ we have

$$\limsup_{\delta\downarrow 0}\delta\log\mathbb{E}\left[C_1^{-1/\alpha}\left(C_2+C_3(1+\|Y_\delta\|^\beta)\right)^{1/\alpha}\mathbf{1}_{\{\|Y_\delta\|>r\}}\right] \leq -\frac{(r-\|x\|^2)_+}{2\gamma}.$$

The lower bound was already shown in Proposition 1. Hence the result. □

Lemma 11. *Let $x\in\mathbb{R}^d$, $\gamma,\delta>0$, $\varepsilon>0$ and $Y_\delta\sim\mathcal{N}(x,\delta\gamma I_d)$. Let $R>0$. We have*

$$\limsup_{\delta\downarrow 0}\delta\log\mu_{x,\delta}\left(((\mathcal{M}_x)_\varepsilon)^{\complement}\cap\mathbb{B}(0,R)\right)\leq -c_{R,\varepsilon}. \tag{42}$$

Therefore for sufficiently large R

$$\limsup_{\delta\downarrow 0}\delta\log\mu_{x,\delta}\left(((\mathcal{M}_x)_\varepsilon)^{\complement}\right)<0. \tag{43}$$

hence,

$$\mu_{x,\delta}\left(((\mathcal{M}_x)_\varepsilon)^{\complement}\right)\xrightarrow[\delta\downarrow 0]{}0,\ \textit{and}\ \mu_{x,\delta}\left(((\mathcal{M}_x)_\varepsilon)\right)\xrightarrow[\delta\downarrow 0]{}1 \tag{44}$$

Proof. We have

$$\mu_{x,\delta}\left(((\mathcal{M}_x)_\varepsilon)^{\complement}\cap\mathbb{B}(0,R)\right)=\frac{\mathbb{E}\left[e^{-f(T_f(Y_\delta))/\delta}\mathbf{1}_{\left\{T_f(Y_\delta)\in((\mathcal{M}_x)_\varepsilon)^{\complement}\cap\mathbb{B}(0,R)\right\}}\right]}{\mathbb{E}\left[e^{-f(T_f(Y_\delta))/\delta}\right]}$$

and, there exists some set $\mathcal{M}_{x,R,\varepsilon}$ (Fact 1) such that

$$\left\{T_f(Y_\delta)\in((\mathcal{M}_x)_\varepsilon)^{\complement}\cap\mathbb{B}(0,R)\right\}\subset\mathcal{M}_{x,R,\varepsilon}.$$

Hence,

$$\mathbb{E}\left[e^{-f(T_f(Y_\delta))/\delta}\mathbf{1}_{\left\{T_f(Y_\delta)\in((\mathcal{M}_x)_\varepsilon)^{\complement}\cap\mathbb{B}(0,R)\right\}}\right]\leq\sum_{M\in\mathcal{M}_{x,R,\varepsilon}}e^{-f(M)/\delta}\mathbb{P}(Y_\delta\in\mathrm{Att}(M)),$$

which is sufficient to conclude for the first inequality by applying Lemma 3. The second inequality comes from the fact that

$$\mu_{x,\delta}\left(((\mathcal{M}_x)_\varepsilon)^{\complement}\right)\leq\mu_{x,\delta}\left(((\mathcal{M}_x)_\varepsilon)^{\complement}\cap\mathbb{B}(0,R)\right)+\mu_{x,\delta}(\mathbb{B}(0,R)^{\complement}),$$

where the second complement is taken in $\mathcal{M}_*$. We apply Proposition 1 to obtain the desired result. □

If R is large enough then it means that the mass outside the ball $\mathbb{B}(0,R)$ goes to 0 when $\delta\downarrow 0$.

Proof of Proposition 2.

Proof. Fix $\varepsilon > 0$ and $R > 0$. For convenience, we write

$$m_\delta := \frac{\mathbb{E}[T_f(Y_\delta)e^{-f(T_f(Y_\delta))/\delta}]}{\mathbb{E}[e^{-f(T_f(Y_\delta))/\delta}]}$$

Since the distance to a closed convex set is a convex function, Jensen's inequality gives

$$\begin{aligned}\operatorname{dist}(m_\delta, \operatorname{conv}(\mathcal{M}_x)) &= \operatorname{dist}\left(\int_{\mathcal{M}_*} z\, \mu_{\delta,x}(dz), \operatorname{conv}(\mathcal{M}_x)\right) \\ &\leq \int_{\mathcal{M}_*} \operatorname{dist}(z, \operatorname{conv}(\mathcal{M}_x))\ \mu_{\delta,x}(dz).\end{aligned}$$

We split the integral over the three regions $(\mathcal{M}_x)_\varepsilon \cap \mathbb{B}(0,R)$, $\mathbb{B}(0,R) \setminus (\mathcal{M}_x)_\varepsilon$, and $\mathbb{B}(0,R)^{\complement}$.

If $z \in (\mathcal{M}_x)_\varepsilon$, then since $\mathcal{M}_x \subset \operatorname{conv}(\mathcal{M}_x)$, $\operatorname{dist}(z, \operatorname{conv}(\mathcal{M}_x)) \leq \operatorname{dist}(z, \mathcal{M}_x) < \varepsilon$.

If $z \in \mathbb{B}(0,R) \setminus (\mathcal{M}_x)_\varepsilon$, then $\operatorname{dist}(z, \operatorname{conv}(\mathcal{M}_x)) \leq \|z\| \leq R$.

If $\|z\| > R$, then $\operatorname{dist}(z, \operatorname{conv}(\mathcal{M}_x)) \leq \|z\|$. Therefore,

$$\operatorname{dist}(m_\delta, \operatorname{conv}(\mathcal{M}_x)) \leq \varepsilon + R\,\mu_{\delta,x}\left(((\mathcal{M}_x)_\varepsilon)^{\complement}\right) + \int_{\mathcal{M}_*} \|z\| \mathbf{1}_{\{\|z\|>R\}}\, \mu_{\delta,x}(dz).$$

Now let $\delta \downarrow 0$. By Lemma 11 ,

$$\mu_\delta\left(((\mathcal{M}_x)_\varepsilon)^{\complement}\right) \to 0,$$

hence

$$\limsup_{\delta\downarrow 0} \operatorname{dist}(m_\delta, \operatorname{conv}(\mathcal{M}_x)) \leq \varepsilon + \limsup_{\delta\downarrow 0} \int_{\mathcal{M}_*} \|z\| \mathbf{1}_{\{\|z\|>R\}}\, \mu_{\delta,x}(dz).$$

Next let $R \to \infty$ and use Lemma 6 on the righthandside integral to obtain

$$\limsup_{\delta\downarrow 0} \operatorname{dist}(m_\delta, \operatorname{conv}(\mathcal{M}_x)) \leq \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, this proves

$$\operatorname{dist}(m_\delta, \operatorname{conv}(\mathcal{M}_x)) \xrightarrow[\delta\downarrow 0]{} 0.$$

The limit in the isolated local minimizer case follows immediately. □

Proof of Theorem 2.

Proof. First, by rejecting worse update and coercivity of f (Assumption 1), iterates are bounded and stay in some bounded set $B \subset \mathbb{R}^d$. Thus

$$d_f := \sup_{x\in B} \operatorname{dist}(x, \overline{\operatorname{Att}(x_*)}) < +\infty.$$

Set $\gamma > 0$. For every $x \in B$,

$$V_x(x_*) \leq \frac{1}{2\gamma} d_f^2 + f_{\min}$$

If $\gamma > \frac{d_f^2}{2\mu}$ then, for all $M \neq x_*$,

$$V_x(x_*) < f_{\min} + \mu \leq V_x(M).$$

Meaning that for γ sufficiently large, by Proposition 2 there exists $\overline{\delta} > 0$ and $r > 0$ such that for all $\delta \leq \overline{\delta}$

$$\|m_{\delta,\gamma} - x_*\| \leq r.$$

Therefore, if there exists $r > 0$, such that $\mathbb{B}(x_*, r) \subset \operatorname{Att}(x_*)$, then $x_+ = x_*$. Such statement holds in our context as x_* is an isolated local minimizer, reached by T_f. We thus obtain immediate convergence if γ is large enough and δ small enough.

Now there are several scenarios:

- if progress (i.e., strict decrease) is made at every iteration, as iterates are bounded, we visit finitely many $M \in \mathcal{M}$ until convergence (see proof of Theorem 1).
- As derived just before, we can only "not improve" the next iterate finitely many times before reaching γ_k large enough, hence we obtain convergence as soon as δ is small enough which happens eventually.

Hence, we obtain convergence of the iterates to the global minimizer. □

D.3 Proofs for approximate expectation PBH

Proof of Lemma 12.

Lemma 12. *Let $x \in \mathbb{R}^d$, $\delta, \gamma > 0$ and i.i.d. samples $\{y_i\} \sim Y_\delta \sim \mathcal{N}(x, \delta\gamma \mathrm{Id})$. Then, almost surely,*

$$m_{N,\delta,\gamma} := \frac{\sum_{i=1}^N T_f(y_i) \exp(-f\left(T_f(y_i)\right)/\delta)}{\sum_{i=1}^N \exp(-f\left(T_f(y_i)\right)/\delta)} \xrightarrow[N\to+\infty]{} m_{\delta,\gamma}. \tag{45}$$

Proof. We have first that for all $N \in \mathbb{N}$, $\sum_{i=1}^N \exp(-f\left(T_f(y_i)\right)/\delta) > 0$, each term is integrable against the Gaussian measure, and thus by the strong law of large numbers $\frac{1}{N}\sum_{i=1}^N \exp(-f\left(T_f(y_i)\right)/\delta) \xrightarrow[N\to+\infty]{} \mathbb{E}\left[e^{-f(T_f(Y_\delta))/\delta}\right]$ a.s.. On the other end, by Lemma 8, each term in $\sum_{i=1}^N T_f(y_i)\exp(-f\left(T_f(y_i)\right)/\delta)$ is integrable against the Gaussian measure, and thus by the strong law of large numbers $\frac{1}{N}\sum_{i=1}^N T_f(y_i)\exp(-f\left(T_f(y_i)\right)/\delta) \xrightarrow[N\to+\infty]{}$ $\mathbb{E}\left[T_f(Y_\delta)e^{-f(T_f(Y_\delta))/\delta}\right]$ a.s.. Hence, $m_{N,\delta,\gamma} \xrightarrow[N\to+\infty]{} m_{\delta,\gamma}$ a.s.. □

Proof of Lemma 13.

Lemma 13. *Let $x \in \mathbb{R}^d$, $\delta, \gamma > 0$ and i.i.d. samples $\{y_i\} \sim Y_\delta \sim \mathcal{N}(x, \delta\gamma \mathrm{Id})$. Suppose that there exists $M \in \mathcal{M}$ such that $\mathrm{dist}(m_{\delta,\gamma}, \partial \mathrm{Att}(M)) > 0$. Then, almost surely,*

$$T_f(m_{N,\delta,\gamma}) \xrightarrow[N\to+\infty]{} T_f(m_{\delta,\gamma}). \tag{46}$$

Proof. We have by Assumption 4 that T_f is continuous on any interior of $\mathrm{Att}(M)$, hence the result follows from Lemma 12. □

Proof of Lemma 14.

Lemma 14. *Let $(m_N)_{N\geq 1}$ be a sequence in $\mathbb{R}^d$ such that $m_N \to m, \quad N \to +\infty$. Assume that for sufficiently large N, m_N does not belong to $\mathcal{N}_s$ the measurable null set not included in the partition of $\mathbb{R}^d$ defined by $\bigcup_{M\in\mathcal{M}} \mathrm{Att}(M)$ (Lemma 6). Then every accumulation point of the sequence $(T_f(m_N))_{N\geq 1}$ belongs to*

$$\bigcup_{Q\in \mathrm{Adj}(m)} Q, \qquad \mathrm{Adj}(m) := \{Q \in \mathcal{M} :\ m \in \overline{\mathrm{Att}(Q)}\}. \tag{47}$$

Proof. Let $\overline{m}$ be a cluster point of $(T_f(m_N))_{N\geq 1}$. Then there exists a subsequence $(m_{N_k})_{k\geq 1}$ such that

$$T_f(m_{N_k}) \to \overline{m}.$$

Since $(T_f(m_{N_k}))_{k\geq 1}$ converges, it is bounded. Hence there exists a compact set $K \subset \mathbb{R}^d$ such that $T_f(m_{N_k}) \in K$ for all k. By assumption, only finitely many components of $\mathcal{M}$ intersect K. Therefore, after extracting a further subsequence, there exists a component $Q \in \mathcal{M}$ such that

$$T_f(m_{N_{k_\ell}}) \in Q \text{ for all } \ell \geq 1.$$

On the other hand, for k large enough one has $m_{N_k} \notin \mathcal{N}$, and therefore $m_{N_k} \in \mathrm{Att}(M_k)$ for a unique component $M_k \in \mathcal{M}$. Because $T_f(m_{N_k}) \in Q$ and $T_f(y) \in M$ whenever $y \in \mathrm{Att}(M)$, we must have $M_k = Q$. Hence for all sufficiently large k, $m_{N_k} \in \mathrm{Att}(Q)$. Passing to the limit and using $m_{N_k} \to m$, we obtain $m \in \overline{\mathrm{Att}(Q)}$. Thus $Q \in \mathrm{Adj}(m)$. □

Proof of Theorem 3.

Proof. First, as the number of samples grows to infinity, by Lemma 12, all $m_{N,\delta,\gamma}$ are close to $m_{\delta,\gamma}$, hence we the behavior of the algorithm is governed asymptotically by the behavior of Algorithm 1.

By rejecting worse update and coercivity of f (Assumption 1), iterates are bounded and stay in some bounded set $B \subset \mathbb{R}^d$. Thus

$$d_f := \sup_{x \in B} \operatorname{dist}(x, \overline{\operatorname{Att}(x_*)}) < +\infty.$$

Set $\gamma > 0$. For every $x \in B$,

$$V_x(x_*) \leq \frac{1}{2\gamma} d_f^2 + f_{\min}$$

If $\gamma > \frac{d_f^2}{2\mu}$ then, for all $M \neq x_*$,

$$V_x(x_*) < f_{\min} + \mu \leq V_x(M).$$

Meaning that for γ sufficiently large, by Proposition 2 there exists $\overline{\delta} > 0$ and $r_1 > 0$ such that for all $\delta \leq \overline{\delta}$

$$\|m_{\delta,\gamma} - x_*\| \leq r_1.$$

Moreover, for N sufficiently large, there exists $r_2 > 0$ such that

$$\|m_{N,\delta,\gamma} - m_{\delta,\gamma}\| \leq r_2.$$

Therefore, as there exists $r > 0$, such that $\mathbb{B}(x_*, r) \subset \operatorname{Att}(x_*)$, and $r_1 + r_2 \leq r$, then $x_+ = x_*$. We thus obtain immediate convergence if γ is large enough, N is large enough and δ small enough.

Now as in Theorem 2, there are several scenarios:

- if progress (i.e., strict decrease) is made at every iteration, as iterates are bounded, we visit finitely many $M \in \mathcal{M}$ until convergence (see proof of Theorem 1). We can however visit non minimizing point as $m_{N,\delta,\gamma}$ may sit with non zero probability in the null set of the partition defined by $\operatorname{Att}(M)$ for all $M \in \mathcal{M}$, hence finite convergence cannot be guaranteed to happen.
- As derived just before, we can only "not improve" the next iterate finitely many times before reaching γ_k large enough, hence we obtain convergence as soon as δ is small enough, and N is big enough which happens eventually.

Hence, we obtain convergence of the iterates to the global minimizer. □

Proof of Proposition 3.

Proof. Denote $m_{N,\delta,\gamma} = \frac{A_N}{B_N}$, $m_{\delta,\gamma} = \frac{A}{B}$, where

$$A_N = \frac{1}{N} \sum_{i=1}^{N} T_f(y_i) \exp(-f(T_f(y_i))/\delta), \text{ and } A = \mathbb{E}[T_f(Y_\delta) \exp(-f(T_f(Y_\delta))/\delta)],$$

while

$$B_N = \frac{1}{N} \sum_{i=1}^{N} \exp(-f(T_f(y_i))/\delta), \text{ and } B = \mathbb{E}[\exp(-f(T_f(Y_\delta))/\delta)].$$

We want to control

$$\left\| \frac{A_N}{B_N} - \frac{A}{B} \right\| = \left\| \frac{B(A_N - A) - A(B_N - B)}{BB_N} \right\|,$$

w.r.t. quantities that we control, i.e., $\|A_N - A\|$ and $|B_N - B|$. We first look at the denominator. Set $0 < \eta < B$ and the event $\{|B_N - B| \leq \eta\}$. On this event, we have $B_N \geq B - \eta > 0$, implying by the triangular inequality that

$$\|m_{N,\delta,\gamma} - m_{\delta,\gamma}\| = \left\| \frac{B(A_N - A) - A(B_N - B)}{BB_N} \right\| \leq \frac{\|A_N - A\|}{B - \eta} + \frac{\|A\| \, |B_N - B|}{B(B - \eta)}.$$

Now, to obtain $\|m_{N,\delta,\gamma} - m_{\delta,\gamma}\| \leq \varepsilon$, we bound the two terms on the righthandside of this inequality w.r.t. $B - \eta$. If $\|A_N - A\| \leq \frac{\varepsilon}{2}(B-\eta)$, and $|B_N - B| \leq \min\left(\eta, \frac{\varepsilon B(B-\eta)}{2\|A\|}\right)$, then we obtain the desired bound. It remains to quantify the probability of these two bounds holding together. Indeed, we have

$$\{\|m_{N,\delta,\gamma} - m_{\delta,\gamma}\| \leq \varepsilon\} \subset \left\{\|A_N - A\| \leq \frac{\varepsilon}{2}(B-\eta)\right\} \bigcap \left\{|B_N - B| \leq \min\left(\eta, \frac{\varepsilon B(B-\eta)}{2\|A\|}\right)\right\}.$$

The union bound on the complements then tells us that

$$\mathbb{P}\left(\|m_{N,\delta,\gamma} - m_{\delta,\gamma}\| \leq \varepsilon\right) \geq 1 - \mathbb{P}\left(\|A_N - A\| > \frac{\varepsilon}{2}(B-\eta)\right) - \mathbb{P}\left(|B_N - B| > \min\left(\eta, \frac{\varepsilon B(B-\eta)}{2\|A\|}\right)\right).$$

The two righthandside probabilities can be controlled through Markov's and Chebyshev's inequality. We can apply it through the growth control on f and Lemma 8 as $\|A_N - A\|$ and $|B_N - B|$ are both square integrable against the Gaussian measure. We have

$$\mathbb{P}(\|A_N - A\| > t) \leq \frac{\mathbb{E}[\|A_N - A\|^2]}{t^2}, \text{ and } \mathbb{P}(|B_N - B| > s) \leq \frac{\mathrm{Var}(B_N)}{s^2}.$$

As

$$\begin{aligned}
\mathbb{E}[\|A_N - A\|^2] &= \mathbb{E}\left[\frac{1}{N^2}\left\langle \sum_{i=1}^N A_N^i - A, \sum_{i=1}^N A_N^i - A\right\rangle\right] \\
&= \mathbb{E}\left[\frac{1}{N^2}\sum_{i,j=1}^N \left\langle A_N^i - A, A_N^j - A\right\rangle\right] \\
&= \frac{1}{N^2}\sum_{i,j=1}^N \mathbb{E}\left\langle A_N^i - A, A_N^j - A\right\rangle
\end{aligned}$$

For $i \neq j$ by independence, we have $\mathbb{E}\left\langle A_N^i - A, A_N^j - A\right\rangle = 0$. Remain the N, $i = j$ terms, giving

$$\mathbb{E}\left[\|A_N - A\|^2\right] = \frac{1}{N}\mathbb{E}\|T_f(y_1)\exp(-f(T_f(y_1))/\delta) - A\|^2,$$

while

$$\mathrm{Var}(B_N) = \frac{\mathrm{Var}(\exp(-f(T_f(y_1))/\delta))}{N},$$

Therefore, we obtain

$$\mathbb{P}\left(\|m_{N,\delta,\gamma} - m_{\delta,\gamma}\| \leq \varepsilon\right) \geq 1 - \frac{4\,\mathbb{E}\|T_f(y_1)\exp(-f(T_f(y_1))/\delta) - A\|^2}{N\varepsilon^2(B-\eta)^2} - \frac{\mathrm{Var}(\exp(-f(T_f(y_1))/\delta))}{N\min\left(\eta, \frac{\varepsilon B(B-\eta)}{2\|A\|}\right)^2}.$$

We deduce then the existence of α by setting N large enough. □

The dependence on δ may appear highly problematic, but also note that in practice $f(T_f(y_1))$ should be of the order of δ for our best samples.

Proof of Theorem 4.

Proof. The number of samples is finite, by Lemma 12, all $m_{N,\delta,\gamma}$ are close to $m_{\delta,\gamma}$, hence we the behavior of the algorithm is governed asymptotically by the behavior of Algorithm 1.

By rejecting worse update and coercivity of f (Assumption 1), iterates are bounded and stay in some bounded set $B \subset \mathbb{R}^d$. Thus

$$d_f := \sup_{x \in B} \mathrm{dist}(x, \overline{\mathrm{Att}(x_*)}) < +\infty.$$

Set $\gamma > 0$. For every $x \in B$,

$$V_x(x_*) \leq \frac{1}{2\gamma} d_f^2 + f_{\min}$$

If $\gamma > \frac{d_f^2}{2\mu}$ then, for all $M \neq x_*$,

$$V_x(x_*) < f_{\min} + \mu \leq V_x(M).$$

Meaning that for γ sufficiently large, by Proposition 2 there exists $\overline{\delta} > 0$ and $r_1 > 0$ such that for all $\delta \leq \overline{\delta}$

$$\|m_{\delta,\gamma} - x_*\| \leq r_1.$$

For N sufficiently large (Proposition 3), there exists $\varepsilon > 0$ and $0 < \alpha < 1$ such that

$$\mathbb{P}\left(\|m_{N,\delta,\gamma} - m_{\delta,\gamma}\| \leq \varepsilon\right) \geq \alpha.$$

Therefore, as there exists $r > 0$, such that $\mathbb{B}(x_*, r) \subset \mathrm{Att}(x_*)$, we need ε small enough so that and $r_1 + \varepsilon \leq r$, then with probability superior or equal to α, $x_+ = x_*$. Thus, if N is large enough so this ε is small enough, then the high probability follows.

The rest of the proof is the same as in Theorem 3. Hence, we obtain convergence of the iterates to the global minimizer with high probability. □

The trick here is that we do not need an uniform control on the probability on all the iterates. We simply need to wait for γ to be sufficiently large so that one of the iterates theoretically has x_* as one of its minimizers of the potential V_x, then if we have enough samples we reach the correct solution with high probability.

D.4 Proofs for inexact solver PBH

We will first prove the following statement.

Proposition 7. *Suppose that Assumptions 1, 2, 3, 4, 5, and 6 hold. For every $\varepsilon > 0$, let $T_f^\varepsilon : \mathbb{R}^d \to \mathbb{R}^d$ be a measurable inexact local solver satisfying, for every $y \in \mathbb{R}^d$,*

$$\left\|T_f^\varepsilon(y) - T_f(y)\right\| \leq r(\varepsilon), \qquad 0 \leq f(T_f^\varepsilon(y)) - f(T_f(y)) \leq \varepsilon, \tag{48}$$

where $r(\varepsilon) \xrightarrow[\varepsilon \downarrow 0]{} 0$. Denote the inexact expectation barycenters by $m_{\delta,\gamma}^\varepsilon$. Let $\varepsilon = \varepsilon(\delta)$ satisfy $\varepsilon(\delta) \xrightarrow[\delta \downarrow 0]{} 0$, $\frac{\varepsilon(\delta)}{\delta} \xrightarrow[\delta \downarrow 0]{} 0$. Then $\left\|m_{\delta,\gamma}^{\varepsilon(\delta)}(x) - m_{\delta,\gamma}(x)\right\| \xrightarrow[\delta \downarrow 0]{} 0$. If, moreover, there exists $\rho > 0$ such that $\mathbb{B}(x_\star, \rho) \subset \mathrm{Att}(x_\star)$, then $T_f^{\varepsilon(\delta)}\left(m_{\delta,\gamma}^{\varepsilon(\delta)}(x)\right) \xrightarrow[\delta \downarrow 0]{} x_\star$.

Proof. For brevity, set $Z_\delta := T_f(Y_\delta)$, $Z_\delta^\varepsilon := T_f^\varepsilon(Y_\delta)$, and $D_\delta^\varepsilon := f(Z_\delta^\varepsilon) - f(Z_\delta)$. Introduce the probability measure

$$\mathrm{d}\mu_{\delta,\gamma}(y) := \frac{\exp\left(-f(T_f(y))/\delta\right)}{\mathbb{E}\left[\exp\left(-f(T_f(Y_\delta))/\delta\right)\right]} \mathrm{d}\mathbb{P}_{Y_\delta}(y). \tag{49}$$

With this notation,

$$m_{\delta,\gamma}(x) = \int_{\mathbb{R}^d} T_f(y)\, \mathrm{d}\mu_{\delta,\gamma}(y).$$

Moreover, defining

$$A_{\delta,\varepsilon}(y) := \exp\left(-\frac{f(T_f^\varepsilon(y)) - f(T_f(y))}{\delta}\right),$$

we may rewrite the inexact barycenter as

$$m_{\delta,\gamma}^\varepsilon(x) = \frac{\displaystyle\int_{\mathbb{R}^d} T_f^\varepsilon(y) A_{\delta,\varepsilon}(y)\, \mathrm{d}\mu_{\delta,\gamma}(y)}{\displaystyle\int_{\mathbb{R}^d} A_{\delta,\varepsilon}(y)\, \mathrm{d}\mu_{\delta,\gamma}(y)}. \tag{50}$$

By equation 48, $0 \le D_\delta^\varepsilon \le \varepsilon$. Hence,

$$e^{-\varepsilon/\delta} \le A_{\delta,\varepsilon}(y) \le 1.$$

Consequently,

$$e^{-\varepsilon/\delta} \le \int_{\mathbb{R}^d} A_{\delta,\varepsilon}\,\mathrm{d}\mu_{\delta,\gamma} \le 1.$$

Therefore,

$$e^{-\varepsilon/\delta} \le \frac{A_{\delta,\varepsilon}(y)}{\displaystyle\int_{\mathbb{R}^d} A_{\delta,\varepsilon}\,\mathrm{d}\mu_{\delta,\gamma}} \le e^{\varepsilon/\delta}.$$

Define

$$q_{\delta,\varepsilon}(y) := \frac{A_{\delta,\varepsilon}(y)}{\displaystyle\int_{\mathbb{R}^d} A_{\delta,\varepsilon}\,\mathrm{d}\mu_{\delta,\gamma}}.$$

Then

$$\int_{\mathbb{R}^d} q_{\delta,\varepsilon}\,\mathrm{d}\mu_{\delta,\gamma} = 1$$

and

$$m_{\delta,\gamma}^\varepsilon(x) = \int_{\mathbb{R}^d} T_f^\varepsilon(y) q_{\delta,\varepsilon}(y)\,\mathrm{d}\mu_{\delta,\gamma}(y).$$

It follows that

$$\begin{aligned}
&\left\|m_{\delta,\gamma}^\varepsilon(x) - m_{\delta,\gamma}(x)\right\| \\
&\quad\le \left\|\int_{\mathbb{R}^d} \big(T_f^\varepsilon(y) - T_f(y)\big) q_{\delta,\varepsilon}(y)\,\mathrm{d}\mu_{\delta,\gamma}(y)\right\| \\
&\qquad + \left\|\int_{\mathbb{R}^d} T_f(y)\big(q_{\delta,\varepsilon}(y) - 1\big)\,\mathrm{d}\mu_{\delta,\gamma}(y)\right\|.
\end{aligned} \tag{51}$$

For the first term, using equation 48 and the fact that $q_{\delta,\varepsilon}$ integrates to one, we obtain

$$\left\|\int_{\mathbb{R}^d} \big(T_f^\varepsilon - T_f\big) q_{\delta,\varepsilon}\,\mathrm{d}\mu_{\delta,\gamma}\right\| \le r(\varepsilon). \tag{52}$$

For the second term, equation D.4 gives

$$|q_{\delta,\varepsilon}(y) - 1| \le e^{\varepsilon/\delta} - 1.$$

Hence

$$\begin{aligned}
&\left\|\int_{\mathbb{R}^d} T_f(y)\big(q_{\delta,\varepsilon}(y) - 1\big)\,\mathrm{d}\mu_{\delta,\gamma}(y)\right\| \\
&\quad\le \left(e^{\varepsilon/\delta} - 1\right)\int_{\mathbb{R}^d} \|T_f(y)\|\,\mathrm{d}\mu_{\delta,\gamma}(y).
\end{aligned} \tag{53}$$

The integrability estimate established in Lemma 8 implies that, for fixed x and γ,

$$\sup_{0<\delta\le\delta_0} \int_{\mathbb{R}^d} \|T_f(y)\|\,\mathrm{d}\mu_{\delta,\gamma}(y) < +\infty \tag{54}$$

for some $\delta_0 > 0$. Combining equation 51–equation 54, there exists $C_{x,\gamma} > 0$ such that

$$\left\|m_{\delta,\gamma}^\varepsilon(x) - m_{\delta,\gamma}(x)\right\| \le r(\varepsilon) + C_{x,\gamma}\left(e^{\varepsilon/\delta} - 1\right). \tag{55}$$

Taking $\varepsilon = \varepsilon(\delta)$, we have

$$r(\varepsilon(\delta)) \longrightarrow 0$$

and, because $\varepsilon(\delta)/\delta \to 0$,

$$e^{\varepsilon(\delta)/\delta} - 1 \longrightarrow 0.$$

Thus,

$$\left\| m_{\delta,\gamma}^{\varepsilon(\delta)}(x) - m_{\delta,\gamma}(x) \right\| \longrightarrow 0.$$

By the exact-expectation convergence proof (Proposition 2),

$$\operatorname{dist}\left(m_{\delta,\gamma}(x), \operatorname{conv}(\mathcal{M}_x)\right) \underset{\delta \downarrow 0}{\longrightarrow} 0.$$

Therefore,

$$\begin{aligned} &\operatorname{dist}\left(m_{\delta,\gamma}^{\varepsilon(\delta)}(x), \operatorname{conv}(\mathcal{M}_x)\right) \\ &\quad \leq \left\| m_{\delta,\gamma}^{\varepsilon(\delta)}(x) - m_{\delta,\gamma}(x) \right\| + \operatorname{dist}\left(m_{\delta,\gamma}(x), \operatorname{conv}(\mathcal{M}_x)\right) \longrightarrow 0. \end{aligned}$$

If $\mathcal{M}_x = \{x_\star\}$, then

$$m_{\delta,\gamma}^{\varepsilon(\delta)}(x) \longrightarrow x_\star.$$

Hence, for all sufficiently small δ,

$$m_{\delta,\gamma}^{\varepsilon(\delta)}(x) \in \mathbb{B}(x_\star, \rho) \subset \operatorname{Att}(x_\star).$$

It follows that

$$T_f\left(m_{\delta,\gamma}^{\varepsilon(\delta)}(x)\right) = x_\star.$$

Finally, using equation 48,

$$\begin{aligned} &\left\| T_f^{\varepsilon(\delta)}\left(m_{\delta,\gamma}^{\varepsilon(\delta)}(x)\right) - x_\star \right\| \\ &\quad = \left\| T_f^{\varepsilon(\delta)}\left(m_{\delta,\gamma}^{\varepsilon(\delta)}(x)\right) - T_f\left(m_{\delta,\gamma}^{\varepsilon(\delta)}(x)\right) \right\| \leq r(\varepsilon(\delta)) \longrightarrow 0. \end{aligned}$$

□

Proof of Theorem 5.

Proof. Let $(x_k)_{k\in\mathbb{N}}$ be generated by

$$m_{\delta_k,\gamma_k}^{\varepsilon_k}(x_k) := \frac{\mathbb{E}\left[T_f^{\varepsilon_k}(Y_k) \exp\left(-f(T_f^{\varepsilon_k}(Y_k))/\delta_k\right)\right]}{\mathbb{E}\left[\exp\left(-f(T_f^{\varepsilon_k}(Y_k))/\delta_k\right)\right]}, \qquad Y_k \sim \mathcal{N}(x_k, \delta_k\gamma_k \mathrm{Id}), \tag{56}$$

followed by

$$x_k^+ := T_f^{\varepsilon_k}\left(m_{\delta_k,\gamma_k}^{\varepsilon_k}(x_k)\right), \tag{57}$$

and the update

$$x_{k+1} \in \underset{z\in\{x_k, x_k^+\}}{\arg\min}\ f(z). \tag{58}$$

Assuming that

$$\delta_k \longrightarrow 0, \qquad \varepsilon_k \longrightarrow 0, \qquad \frac{\varepsilon_k}{\delta_k} \longrightarrow 0, \qquad \gamma_k \longrightarrow +\infty. \tag{59}$$

We will show that

$$x_k \longrightarrow x_*.$$

By the rejection of worse updates,

$$f(x_{k+1}) \leq f(x_k)$$

for every $k \in \mathbb{N}$. Hence,

$$x_k \in \left\{x \in \mathbb{R}^d \;\middle|\; f(x) \leq f(x_0)\right\}.$$

By the coercivity of f from Assumption 1, this sublevel set is compact. Consequently, the iterates remain in some bounded set $B \subset \mathbb{R}^d$.

Set

$$d_f := \sup_{x\in B} \operatorname{dist}\left(x, \overline{\operatorname{Att}(x_*)}\right) < +\infty.$$

For every $x \in B$ and every $\gamma > 0$, we have

$$V_x(x_*) = f_{\min} + \frac{1}{2\gamma}\mathrm{dist}\left(x, \overline{\mathrm{Att}(x_*)}\right)^2 \leq f_{\min} + \frac{d_f^2}{2\gamma}.$$

If

$$\gamma > \frac{d_f^2}{2\mu},$$

then

$$V_x(x_*) < f_{\min} + \mu$$

for every $x \in B$. On the other hand, by Assumption 2, for every $M \in \mathcal{M} \setminus \{x_*\}$,

$$V_x(M) \geq f(M) \geq f_{\min} + \mu.$$

Therefore,

$$\arg\min_{M \in \mathcal{M}} V_x(M) = \{x_*\} \tag{60}$$

for every $x \in B$ and every $\gamma > d_f^2/(2\mu)$.

Since $\gamma_k \to +\infty$, there exists $k_0 \in \mathbb{N}$ such that

$$\gamma_k > \frac{d_f^2}{2\mu}$$

for every $k \geq k_0$. Thus, for every sufficiently large k, x_* is the unique minimizer of V_{x_k}.

Let $\rho > 0$ be such that

$$\mathbb{B}(x_*, \rho) \subset \mathrm{Att}(x_*). \tag{61}$$

Such a radius exists because x_* is an isolated attracting local minimizer.

Denote by

$$m_{\delta_k,\gamma_k}(x_k) := \frac{\mathbb{E}\left[T_f(Y_k)\exp\left(-f(T_f(Y_k))/\delta_k\right)\right]}{\mathbb{E}\left[\exp\left(-f(T_f(Y_k))/\delta_k\right)\right]}$$

the exact-expectation barycenter associated with the exact local solver.

By Proposition 2, equation 60, and the fact that $\delta_k \to 0$, we have, for all sufficiently large k,

$$\|m_{\delta_k,\gamma_k}(x_k) - x_*\| \leq \frac{\rho}{2}. \tag{62}$$

Here, we use the locally uniform version of Proposition 2 on the bounded set B.

Moreover, by Theorem 5,

$$\left\|m^{\varepsilon_k}_{\delta_k,\gamma_k}(x_k) - m_{\delta_k,\gamma_k}(x_k)\right\| \xrightarrow[k\to+\infty]{} 0,$$

because

$$\varepsilon_k \to 0 \qquad \text{and} \qquad \frac{\varepsilon_k}{\delta_k} \to 0.$$

Consequently, for all sufficiently large k,

$$\left\|m^{\varepsilon_k}_{\delta_k,\gamma_k}(x_k) - m_{\delta_k,\gamma_k}(x_k)\right\| \leq \frac{\rho}{2}. \tag{63}$$

Combining equation 62 and equation 63, we obtain

$$\left\|m^{\varepsilon_k}_{\delta_k,\gamma_k}(x_k) - x_*\right\| < \rho$$

for every sufficiently large k. Therefore,

$$m^{\varepsilon_k}_{\delta_k,\gamma_k}(x_k) \in \mathbb{B}(x_*, \rho) \subset \mathrm{Att}(x_*).$$

It follows that

$$T_f\left(m^{\varepsilon_k}_{\delta_k,\gamma_k}(x_k)\right) = x_*. \tag{64}$$

Using equation **??** and equation 64, we obtain

$$\begin{aligned}\|x_k^+ - x_*\| &= \left\| T_f^{\varepsilon_k}\left(m_{\delta_k,\gamma_k}^{\varepsilon_k}(x_k)\right) - T_f\left(m_{\delta_k,\gamma_k}^{\varepsilon_k}(x_k)\right)\right\| \\ &\leq r(\varepsilon_k). \end{aligned} \tag{65}$$

Since $r(\varepsilon_k) \to 0$, this shows that

$$x_k^+ \longrightarrow x_*. \tag{66}$$

By continuity of f,

$$f(x_k^+) \longrightarrow f(x_*) = f_{\min}.$$

Finally, the monotonicity of the last update equation 58 gives

$$f_{\min} \leq f(x_{k+1}) \leq f(x_k^+).$$

Taking the limit yields

$$f(x_k) \longrightarrow f_{\min}.$$

The sequence $(x_k)_{k\in\mathbb{N}}$ lies in the compact set B. Therefore, it admits cluster points. If $x_{k_j} \to \overline{x}$, then the continuity of f gives

$$f(\overline{x}) = \lim_{j\to+\infty} f(x_{k_j}) = f_{\min}.$$

Since x_* is the unique global minimizer of f, we necessarily have

$$\overline{x} = x_*.$$

Thus, x_* is the unique cluster point of $(x_k)_{k\in\mathbb{N}}$, and consequently

$$x_k \longrightarrow x_*.$$

□

E Additional results

E.1 Ablation study

We performed an ablation study to identify the best set of parameters, and their consistency from one function to the other: The results are all in favor of conservative choices for the parameter: small

η_1	$1/\eta_2$	Styblinski–Tang	Aniso. Rastrigin	Rot. Aniso. Rastrigin	Griewank
1.05	**0.99**	**-684.37**	**47.76**	**44.77**	**0.01**
1.10	**0.99**	**-684.37**	**47.76**	**44.77**	**0.01**
1.20	**0.99**	**-684.37**	**47.76**	**44.77**	**0.01**
1.05	0.95	-586.03	113.25	100.85	0.92
1.10	0.95	-631.80	113.53	101.28	0.96
1.20	0.95	–	127.04	95.23	0.99
1.05	0.90	-457.47	174.96	146.30	0.89
1.10	0.90	-641.23	199.01	137.10	0.96
1.20	0.90	–	196.14	144.41	0.82

Table 3: Hyperparameter ablation study in dimension 20 for a CPU time of $2 \times d$. The best objective value for each benchmark is highlighted in bold. When a function value is missing it is because the algorithm did not converge (it was numerically instable).

increases for γ (η_1) and decreases for δ ($1/\eta_2$), from one iteration to the other.

E.2 Scaling laws

We display additional results for the full scaling law (all parameters available to be fitted Shukor et al. (2025)). The results are worse (in terms of prediction) than those of the additive law, in average. It was expected given the results of Shukor et al. (2025). Our algorithm stil outperforms the others.

Family	Domain	$N=5$	$N=10$	$N=15$	$N=20$	$N=30$	$N=40$	$N=50$	ZOP	BH
LLM ($d=47$)	arxiv	12.085	12.303	**11.681**	12.358	11.861	12.177	12.694	12.646	12.379
	book	**10.046**	10.949	10.173	10.722	10.718	10.264	10.265	10.800	10.649
	c4	10.937	10.949	10.874	10.936	11.115	11.472	**10.716**	11.309	11.314
	commoncrawl	**10.445**	11.038	11.834	11.710	11.450	11.516	11.256	11.677	11.464
	github	17.323	17.385	18.258	**17.305**	18.434	17.899	18.470	19.882	18.847
	stackexchange	**12.488**	12.891	12.943	13.162	13.110	13.024	13.073	13.620	13.374
	wikipedia	**19.658**	19.859	20.044	19.664	19.915	19.739	19.749	20.529	19.827
LVM ($d=29$)	alttext	**4.117**	6.703	11.270	10.317	12.237	14.245	14.062	13.176	10.130
	highquality1	**7.634**	7.758	9.393	12.712	18.447	16.315	19.492	21.979	11.756
	highquality2	**7.536**	10.969	9.483	11.711	15.700	18.769	19.177	21.367	12.207
	synthetic	11.435	14.736	**11.070**	13.664	14.441	16.844	21.105	29.138	15.884
NMM ($d=23$)	captions	**10.944**	12.190	11.921	12.111	12.098	12.596	12.347	11.351	11.317
	interleaved	10.839	**8.768**	10.925	11.097	10.969	10.586	10.706	10.178	10.907
	text	**6.565**	9.392	8.811	9.057	9.313	9.174	9.131	8.774	8.120

Table 4: Test $\mathrm{MRE} = \frac{|\text{prediction}-\text{observation}|}{\text{observation}}$ (%) for fitting full scaling law. (`lbfgs steps=10`, `CPU time=300s`). We set $\gamma = 10$ and $\delta = 10^{-3}$ at initialization. Results averaged over three seeds.

Family	Domain	$N=5$	$N=10$	$N=15$	$N=20$	$N=30$	$N=40$	$N=50$	ZOP	BH
LLM ($d=47$)	arxiv	12.373	12.445	**11.820**	12.478	12.318	12.561	12.387	12.646	12.269
	book	**9.910**	10.947	10.529	10.590	10.451	10.295	10.256	10.800	10.629
	c4	11.226	11.266	11.571	11.161	11.201	11.203	**10.992**	11.309	11.060
	commoncrawl	**10.650**	11.254	11.733	11.249	11.321	11.679	11.331	11.677	11.474
	github	**16.579**	18.328	18.947	16.835	18.686	18.750	18.329	19.882	18.243
	stackexchange	**11.735**	12.606	13.342	12.858	13.802	13.620	13.517	13.620	13.339
	wikipedia	**19.285**	19.407	19.909	19.398	19.976	20.077	19.930	20.529	19.744
LVM ($d=29$)	alttext	**4.438**	6.761	13.017	10.932	13.403	14.336	14.217	13.176	10.094
	highquality1	**7.228**	7.850	10.172	12.248	18.583	17.110	19.230	21.979	11.735
	highquality2	**7.305**	10.519	9.581	12.259	17.326	18.470	19.405	21.367	12.217
	synthetic	**9.318**	11.464	10.898	15.507	18.410	17.687	20.299	29.138	18.764
NMM ($d=23$)	captions	**10.727**	11.847	11.867	12.172	12.224	12.383	12.094	11.351	11.308
	interleaved	10.561	**8.522**	10.151	11.040	10.564	10.552	10.649	10.178	10.906
	text	**7.202**	8.772	8.902	9.023	9.787	9.147	9.147	8.774	8.120

Table 5: Test $\mathrm{MRE} = \frac{|\text{prediction}-\text{observation}|}{\text{observation}}$ (%) for fitting full scaling law. (`lbfgs steps=5`, `CPU time=300s`). We set $\gamma = 10$ and $\delta = 10^{-3}$ at initialization. Results averaged over three seeds.