10.64837/GSA.13.5.1

Global and Stochastic Analysis
Vol. 13 No. 5 (2026)
Received: 15th June 2026 Revised: 12th August 2026 Accepted: 05th September 2026

EXPLORATION FLOORS AND ALMOST SURE CONVERGENCE OF ADAPTIVE GAUSSIAN MIXTURE AND PARZEN-TYPE SAMPLERS IN GLOBAL OPTIMIZATION

SOONHUI LEE

Abstract. The adaptive Gaussian mixture (AGM) algorithm samples candidate solutions from a Gaussian mixture centered at the M best solutions found so far. We develop its convergence theory in a form that covers modern practice. Isolating the property that drives the argument of Lee (2018), a possibly vanishing lower bound on the sampling density that we call an *exploration floor*, we prove almost sure convergence for every elite kernel sampler whose floor sequence is non-summable. Three consequences follow: an explicit threshold for the logarithmically decaying bandwidth schedule $\sigma_j = c\sqrt{M}/(\log j)^g$: convergence for all c, M when $g < 1/2$, at $g = 1/2$ if and only if $c^2 M \geq 2n$, and provable trapping for $g > 1/2$; a shorter proof of the original theorems without the submartingale step or the positivity assumption; and, to our knowledge, the first almost sure convergence statement for unmodified prior-mixed samplers of the tree-structured Parzen estimator (TPE) family on continuous domains, which recovers the known ϵ-greedy guarantee as its constant-floor special case. Empirically, AGM dominates random search at every budget tested, matches TPE statistically on two real hyperparameter-tuning tasks at constant per-iteration cost, and a factorial sensitivity study together with a finite-budget simulation confirms the theoretical threshold exactly at $g = 1/2$.

1. Introduction

Black-box global optimization arises whenever the objective function can only be evaluated pointwise, with no access to gradients or structural information. The setting is now ubiquitous: the tuning of machine-learning hyperparameters, the calibration of simulation models, and the design of engineering systems are routinely formulated as expensive black-box problems [10]. Two families of methods dominate current practice. Evolution strategies, most prominently CMA-ES [4], adapt a full sampling covariance and are widely regarded as the strongest general-purpose continuous optimizers. Sequential model-based methods, most prominently the tree-structured Parzen estimator (TPE) [5] as implemented in Optuna [6], build density models of promising and non-promising regions from the evaluation history and are the de facto standard in hyperparameter optimization. The theory behind TPE remains thin: in contrast to classical model-based search [2, 3], the only convergence-type guarantee available applies to a *modified* sampler (TPE with ϵ-greedy candidate selection) on search spaces with at most countably many configurations, and concludes that the elite set eventually lies inside the true top-γ level

2020 *Mathematics Subject Classification.* Primary 49K99; Secondary 90-08, 62G07.
Key words and phrases. Global optimization, Gaussian mixture model, tree-structured Parzen estimator, convergence, hyperparameter optimization.

Corresponding author: SOONHUI LEE (Email address: shlee2016@hufs.ac.kr)

set [14]; no guarantee has been published for the unmodified sampler, for continuous domains, or for convergence to the global optimum itself [11].

The adaptive Gaussian mixture (AGM) algorithm [1] belongs to the model-based search family surveyed in [2]. At each iteration it draws one candidate from a mixture of Gaussian kernels centered at the M best solutions observed so far, replacing the worst member of the elite set whenever the new candidate improves upon it. AGM occupies a distinctive point in the design space: unlike model reference adaptive search [3] it requires no parameter-optimization subroutine; unlike CMA-ES it maintains no covariance matrix; unlike TPE its per-iteration cost does not grow with the length of the evaluation history. This combination makes AGM interesting beyond the specific algorithm, because it can serve as a *minimal, analyzable representative* of the entire elite-set kernel-sampling family to which TPE belongs.

We organize the paper around three research questions. **RQ1 (theory):** the convergence proofs of [1] are stated for particular sampling constructions; what is the exact property of a sampler that they require, and how far does the guarantee then extend — across AGM's own bandwidth schedules, and across the TPE family? **RQ2 (competitiveness):** how much solution quality, if any, does AGM's simplicity cost against the optimizers practitioners actually use, at realistic evaluation budgets, and at what computational overhead? **RQ3 (robustness):** the appeal of AGM rests partly on having only three parameters (M, c, g); is that claim substantiated by flat response surfaces, or does "few parameters" fail to imply "insensitive parameters"?

Our contributions follow the three questions. *First*, we isolate the driving property of the convergence machinery of [1] as an *exploration floor*, a possibly vanishing uniform lower bound δ_j on the sampling density, and prove that non-summability of $\{\delta_j\}$ suffices for almost-sure convergence of every elite kernel sampler (Theorem 3.2). The proof simultaneously shortens the original argument, dispensing with the submartingale step and with the positivity assumption on the objective. Specializing to the logarithmically decaying bandwidth schedule $\sigma_j = c\sqrt{M}/(\log j)^g$, which we call the *log-decay schedule*, yields an explicit threshold (Corollary 3.6): the guarantee holds for all $c, M > 0$ when $g < 1/2$; at the critical exponent $g = 1/2$ it holds if and only if $c^2 M \geq 2n$; and for $g > 1/2$ it genuinely fails, in that we construct an admissible instance on which the algorithm is trapped with positive probability (Proposition 3.7, proof in Appendix A). Finally, the same theorem covers prior-mixed Parzen samplers of the TPE type (Corollary 3.9): to our knowledge this is the first almost-sure convergence statement for the *unmodified* sampler on continuous domains, and the first concluding convergence to the global optimum rather than containment in a fixed quantile set. The existing ϵ-greedy guarantee of [14] is recovered within the same framework as the constant-floor special case (Remark 3.10), so the exploration floor unifies the two known exploration devices, ϵ-greedy mixing and the prior component, under one condition. The component that falls outside the argument, the candidate-ranking step of deployed TPE, is localized in a single explicit quantity (Remark 3.11). *Second*, we provide the first controlled comparison of AGM against random search [7], CMA-ES [8], and TPE [6] under a common budget, protocol, and scaled decision space, on synthetic benchmarks up to $n = 20$ and budgets up to 10^4 evaluations (Section 5) and on two real hyperparameter-tuning tasks (Section 6), including a quantification of sampler overhead. *Third*, we report a full $3 \times 3 \times 3$ factorial sensitivity study of (M, c, g),

which reveals a sharp dichotomy between multimodal and valley-shaped landscapes (Section 5.2), and a finite-budget trapping simulation that confirms the theoretical threshold exactly at $g = 1/2$ (Section 5.3).

Section 2 reviews the algorithm and assumptions. Section 3 develops the convergence theory. Section 4 describes the experimental design and Section 5 the results on synthetic benchmarks. Section 6 outlines the hyperparameter-tuning study and Section 7 concludes.

2. Preliminaries

2.1. **The AGM algorithm.** We consider $\max_{q\in Q} r(q)$ with $Q \subset \mathbb{R}^n$ compact and r treated as a black box with a unique global maximizer q^*. AGM maintains the M best decisions observed so far, $(q_{[1]}, \dots, q_{[M]})$, and at iteration $j+1$ samples a single candidate from the mixture density

$$\hat{f}\big(q \mid \bar{q}^{\,j}, \bar{r}^{\,j}\big) \;=\; \sum_{i=1}^{M} \alpha_i \, K\!\left(\frac{q - q^{j}_{[i]}}{\sigma^{j}_{[i]}}\right), \tag{1}$$

where K is the product Gaussian kernel. Two bandwidth specifications are proposed in [1]: one ties each kernel's bandwidth to the objective values of the elite set (equation (2.5) there), the other lets a common bandwidth decay with the iteration counter (equation (2.6) there). Throughout we use the latter, which we call the *log-decay schedule*, with equal weights $\alpha_i = 1/M$ and

$$\sigma_j \;=\; \frac{c\sqrt{M}}{(\log j)^g}, \qquad c, g > 0,\ j \geq 2. \tag{2}$$

If the sampled candidate improves upon the worst member of the elite set, it replaces that member; the best recorded value is therefore monotonically nondecreasing (*monotone bookkeeping*). The algorithm has three tuning parameters (M, c, g), performs one Gaussian draw and one comparison per iteration, and requires no matrix updates, no density-ratio estimation, and no inner optimization. Candidates falling outside Q are either sampled from the truncated mixture (Definition 2.7 of [1]) or projected onto the box; both variants are covered by the theory below (Remark 3.4).

2.2. **Assumptions.** Write $\mathcal{F}_j = \sigma(q^1, \dots, q^j)$, $B_\varepsilon(q^*) = \{q : \|q - q^*\| < \varepsilon\}$, and let $Y_j = \max_{i\leq j} r(q^i)$ be the best recorded value. We use the assumptions of [1]:

Assumption 1. *For every $\varepsilon > 0$, $\mu(B_\varepsilon(q^*) \cap Q) > 0$, where μ is Lebesgue measure.*

Assumption 2. *r is continuous or bounded on Q (so $\sup_Q r = r(q^*) < \infty$).*

Assumption 3. *For every $\tilde{r} < r(q^*)$ there exists $\varepsilon > 0$ such that $B_\varepsilon(q^*) \subseteq \{q \in Q : r(q) > \tilde{r}\}$.*

These are Assumptions 2.3, 2.5 and 2.6 of [1]. Positivity of r (Assumption 2.4 there) is *not* needed in this paper; it enters the original analysis only through the objective-adaptive bandwidth formula (equation (2.5) of [1]), which we do not use here.

3. Convergence theory

3.1. **Samplers with an exploration floor.** The proofs of Propositions 2.8–2.9 of [1] use exactly one property of the sampling distribution: that its density is bounded below on Q, uniformly or in a controlled iteration-dependent way. We isolate this property.

Definition 3.1 (Exploration floor)**.** *A sequential sampler on Q has* exploration floor $\{\delta_j\}_{j\geq 1}$, $\delta_j \geq 0$, *if for every j the conditional law of q^{j+1} given $\mathcal{F}_j$ admits a density f^j on Q with $f^j(q) \geq \delta_j$ for all $q \in Q$, almost surely. The floor is* persistent *if $\sum_{j\geq 1} \delta_j = \infty$.*

Theorem 3.2 (Convergence under a persistent floor)**.** *Suppose the sampler has a persistent exploration floor and uses monotone bookkeeping, and let Assumptions 1–3 hold. Then*

(a) *for every $\varepsilon > 0$, $\mathbb{P}\big(q^j \in B_\varepsilon(q^*) \ \ i.o.\big) = 1$;*
(b) *$Y_j \uparrow Y_\infty = r(q^*)$ almost surely.*

Proof. (a) Fix $\varepsilon > 0$ and set $A_{j+1} = \{q^{j+1} \in B_\varepsilon(q^*)\} \in \mathcal{F}_{j+1}$. By Definition 3.1,

$$\mathbb{P}(A_{j+1} \mid \mathcal{F}_j) \ = \ \int_{B_\varepsilon(q^*)\cap Q} f^j(q)\, dq \ \geq \ \delta_j\, \mu\big(B_\varepsilon(q^*) \cap Q\big),$$

so by Assumption 1 and persistence, $\sum_j \mathbb{P}(A_{j+1} \mid \mathcal{F}_j) = \infty$ a.s. Lévy's extension of the second Borel–Cantelli lemma (see [13], "Second Borel–Cantelli Lemma, II") states that $\{A_j \ \text{i.o.}\} = \{\sum_j \mathbb{P}(A_j \mid \mathcal{F}_{j-1}) = \infty\}$ up to a null set; hence $\mathbb{P}(A_j \ \text{i.o.}) = 1$.

(b) Y_j is nondecreasing and $Y_j \leq \sup_Q r = r(q^*) < \infty$ by Assumption 2, so Y_j converges pointwise to some $Y_\infty \leq r(q^*)$; no martingale argument is required. Fix $\eta > 0$. By Assumption 3 applied to $\tilde{r} = r(q^*) - \eta$ there is $\tilde{\varepsilon} > 0$ with $B_{\tilde{\varepsilon}}(q^*) \subseteq \{r > r(q^*) - \eta\}$. By part (a) there a.s. exists j with $q^j \in B_{\tilde{\varepsilon}}(q^*)$, whence $Y_j \geq r(q^j) > r(q^*) - \eta$ and therefore $Y_\infty \geq r(q^*) - \eta$ a.s. Taking $\eta = 1/k$ and intersecting the countably many almost-sure events gives $Y_\infty \geq r(q^*)$ a.s.; the reverse inequality was noted above. □

Remark 3.3 (Relation to [1]). With the constant floor $\delta_j \equiv (1-\lambda)/\mu(Q)$, Theorem 3.2 recovers Proposition 2.9 of [1] (uniform mixing), and conclusion (b) recovers Theorems 2.10–2.11 there with a shorter argument: monotone bookkeeping replaces the submartingale step, and part (a) replaces the contradiction argument. The positivity assumption on r is not used. We emphasize that the brevity of the proof is deliberate. The contribution of Theorem 3.2 lies not in its difficulty but in isolating the exact property that drives the argument, which is what makes the threshold and sharpness results of Section 3.2 and the transfer to the TPE family in Section 3.3 possible; the technical weight of the paper is carried by Proposition 3.7.

Remark 3.4 (Projection onto the box). Implementations often *project* an out-of-box proposal onto Q rather than sampling from the truncated density. If X has density bounded below by δ_j on Q and Π is the projection, then for any Borel $B \subseteq \operatorname{int} Q$ on which Π is the identity, $\mathbb{P}(\Pi(X) \in B) \geq \mathbb{P}(X \in B) \geq \delta_j \mu(B)$. Since Assumption 3 places $B_{\tilde{\varepsilon}}(q^*)$ in the interior of Q for small $\tilde{\varepsilon}$, the proof of Theorem 3.2 goes through verbatim for the projected sampler.

3.2. **An explicit threshold for the log-decay schedule.** Work on the scaled space $Q = [-1,1]^n$, as in the numerical study of [1], with the log-decay schedule (2) and $a := c\sqrt{M}$.

Lemma 3.5 (Mixture floor on a box)**.** *For every $q \in Q$ and every configuration of kernel centers $q_{[1]}, \dots, q_{[M]} \in Q$, the (truncated) mixture* (1) *satisfies*

$$\hat{f}^{\,j}(q) \;\geq\; (2\pi\sigma_j^2)^{-n/2} \exp\big(-2n/\sigma_j^2\big) \;=:\; \delta_j.$$

Proof. Each truncated kernel has the form $K_i(q)/\int_Q K_i(u)\,du$ with

$$K_i(q) = (2\pi\sigma_j^2)^{-n/2} \exp\big\{-\|q - q_{[i]}\|^2/(2\sigma_j^2)\big\}$$

and $\int_Q K_i \leq 1$, so truncation can only increase the density. On $[-1,1]^n$ every coordinate satisfies $|q_l - q_{[i]l}| \leq 2$, hence $\|q - q_{[i]}\|^2 \leq 4n$ and $K_i(q) \geq (2\pi\sigma_j^2)^{-n/2} e^{-2n/\sigma_j^2}$. A convex combination of terms each bounded below by δ_j is bounded below by δ_j.

Corollary 3.6 (Schedule threshold)**.** *For the log-decay schedule, with $\kappa := 2n/a^2 = 2n/(c^2 M)$,*

$$\delta_j \;=\; (2\pi)^{-n/2} a^{-n} (\log j)^{ng} \exp\big(-\kappa(\log j)^{2g}\big) \;=\; C_n\,(\log j)^{ng}\, j^{-\kappa(\log j)^{2g-1}},$$

and consequently the floor is persistent

(i) *for every $c, M > 0$ if $g < 1/2$;*
(ii) *if $g = 1/2$ and $c^2 M \geq 2n$;*

while for $g > 1/2$, and for $g = 1/2$ with $c^2 M < 2n$, one has $\sum_j \delta_j < \infty$ and Theorem 3.2 is not applicable. In cases (i)–(ii), Algorithm 1 of [1] *with the log-decay schedule satisfies conclusions (a) and (b) of Theorem 3.2.*

Proof. The displayed identity rewrites Lemma 3.5 using $e^{-\kappa(\log j)^{2g}} = j^{-\kappa(\log j)^{2g-1}}$. (i) If $g < 1/2$ then $(\log j)^{2g-1} \to 0$, so for all large j the exponent $\kappa(\log j)^{2g-1} \leq 1/2$ and $\delta_j \geq C_n j^{-1/2}$, whose sum diverges. (ii) If $g = 1/2$ then $\delta_j = C_n (\log j)^{n/2} j^{-\kappa}$, and $\sum_j j^{-\kappa}(\log j)^{n/2}$ diverges if and only if $\kappa \leq 1$, i.e. $c^2 M \geq 2n$ (at $\kappa = 1$ divergence follows from the positive logarithmic factor by the integral test). For $g > 1/2$, $(\log j)^{2g-1} \to \infty$, so eventually $\kappa(\log j)^{2g-1} \geq 2$ and $\delta_j \leq C_n (\log j)^{ng} j^{-2}$, which is summable; similarly for $g = 1/2$ with $\kappa > 1$.

The condition $c^2 M \geq 2n$ at the critical exponent has a direct interpretation: the aggregate bandwidth scale $c\sqrt{M}$ must grow at least like $\sqrt{2n}$ with dimension for exploration to persist, echoing the informal high-dimensional discussion in [1]. A quantitative caveat accompanies case (i): the floor exponent $\kappa(\log j)^{2g-1}$ drops below 1 only once $\log j > \kappa^{1/(1-2g)}$, so for $g < 1/2$ with a large constant κ the guaranteed divergence of $\sum_j \delta_j$ sets in only beyond the crossover index $j \approx \exp(\kappa^{1/(1-2g)})$ (e.g. e^{243} for $\kappa = 3$, $g = 0.4$). Subcritical decay is thus an asymptotic guarantee whose practical exploration strength still depends on the constants, a point the finite-budget experiments of Section 5.3 make concrete. The next result shows the threshold is not an artifact of the proof technique: above it, the algorithm fails with positive probability on an admissible instance. Section 5.3 confirms the threshold in finite-budget simulation.

Proposition 3.7 (Sharpness: trapping for supercritical decay)**.** *Let $g > 1/2$, or $g = 1/2$ with $\kappa > 1$. There exists a continuous $r : Q \to (0, \infty)$ with a unique global maximizer satisfying Assumptions 1–3 such that Algorithm 1 of* [1] *with the pure log-decay sampler satisfies $\mathbb{P}\big(Y_\infty < r(q^*)\big) > 0$.*

Proof. See Appendix A. □

3.3. **The TPE family.** Before stating the correspondence, we make explicit why it is worth establishing. TPE and AGM sit at opposite corners of the same design space: TPE is arguably the most widely deployed model-based sampler in practice, being the default in the leading hyperparameter optimization frameworks [6], yet its theory is thin, while AGM is theoretically clean but structurally minimal. A precise reduction between them converts this asymmetry into value in three directions. First, *theory transfers upward*: guarantees proved for the tractable member of the family become statements about the deployed one; Corollary 3.9 and Remark 3.10 below are possible only because the reduction identifies which structural properties (the exploration floor) the two algorithms share and which they do not. Second, *empirical comparisons become component ablations*: once AGM is identified as TPE stripped of specific components, a matched-protocol performance difference between the two is no longer an apples-to-oranges horse race but an estimate of what the stripped components (the non-promising density model and the candidate ranking) contribute, and at what computational price. Third, *design guidance flows both ways*: the floor principle explains which TPE modifications preserve the guarantee (the ϵ-greedy device, a non-vanishing prior weight) and which forfeit it, while TPE's engineering (quantile-based elite sets, per-dimension bandwidths) suggests principled upgrades of AGM whose effect on the guarantee can be checked against Definition 3.1. The remainder of this subsection carries out the reduction.

TPE [5, 11] partitions the evaluation history at the γ-quantile of observed objective values, fits Parzen estimators $l(x)$ to the good points and $g(x)$ to the remaining points, draws k candidates from l, and evaluates the candidate maximizing $l(x)/g(x)$. Crucially, l is not built from the observations alone: a prior component with positive weight is mixed in, which keeps l bounded away from zero on the bounded search space [11]. Table 1 aligns the two algorithms component by component.

Observation 3.8 (AGM as reduced TPE)**.** *The log-decay AGM sampler is obtained from TPE by (i) fixing the good set to the M incumbent best points instead of a quantile of the full history, (ii) deleting the density model g of non-promising points, (iii) drawing $k = 1$ candidate, which renders the ratio-ranking step vacuous, and (iv) replacing the prior component by a slowly decaying common bandwidth as the exploration device. AGM is thus a numerator-only, single-candidate member of the same elite kernel-sampling family.*

Consider then a sampler that, at iteration j, evaluates a point drawn from the good-region density

$$l^j(q) \;=\; w_0^j\, p_0(q) \;+\; \sum_{i=1}^{m_j} w_i^j\, \phi_i^j(q), \qquad w_0^j + \sum_i w_i^j = 1,$$

where p_0 is a prior density on Q and ϕ_i^j are kernel densities centered at the current good observations.

Component	**AGM (log-decay)**	**TPE**
Model of promising region	KDE on the M best points (fixed M)	KDE on the top-γ quantile (growing set)
Model of non-promising region	none	KDE $g(x)$ on remaining points
Exploration device	slowly decaying bandwidth σ_j	prior component mixed into $l(x)$
Candidates per iteration	1, drawn from the mixture	k drawn from $l(x)$; $\arg\max l/g$ retained
Per-iteration cost	$O(1)$	grows with evaluation history
Convergence guarantee	Corollary 3.6	ϵ-greedy variant [14]; Cor. 3.9

TABLE 1. Component-wise correspondence between AGM (log-decay schedule) and TPE.

Corollary 3.9 (TPE-type samplers)**.** *Suppose* $\underline{p} := \inf_{q\in Q} p_0(q) > 0$ *and* $\sum_j w_0^j = \infty$ *a.s. Then the sampler has the persistent exploration floor* $\delta_j = w_0^j \underline{p}$*, and under Assumptions 1–3 with monotone bookkeeping both conclusions of Theorem 3.2 hold. In particular this covers (i) implementations in which the good set is capped at a constant size, so that* w_0^j *is bounded below by a constant, and (ii) implementations in which the good set grows proportionally to* j *with a constant prior weight, so that* $w_0^j \asymp 1/j$ *and* $\sum_j w_0^j$ *diverges as a harmonic series.*

Proof. $l^j(q) \geq w_0^j p_0(q) \geq w_0^j \underline{p}$ for all $q \in Q$; apply Theorem 3.2. □

Remark 3.10 (Relation to the ϵ-greedy guarantee). The one existing convergence-type result for TPE, due to Watanabe et al. [14], modifies the sampler with ϵ-greedy candidate selection (with probability ϵ the evaluated point is drawn uniformly at random) and shows, for search spaces with at most countably many configurations, that the elite set is eventually contained in the true top-γ level set almost surely. The mechanism is an exploration floor: ϵ-greedy mixing gives the evaluated point's density the constant lower bound $\delta_j \equiv \epsilon/\mu(Q)$ on a bounded domain, which is precisely the uniform-mixing floor of Proposition 2.9 of [1] and the constant-floor instance of Definition 3.1. Theorem 3.2 therefore recovers the exploratory content of their result and extends it in three directions: to continuous compact domains, to *vanishing* floors (covering the unmodified sampler, whose prior weight w_0^j may decay), and to the stronger conclusion of almost-sure convergence of the best value to the global optimum rather than containment in a fixed quantile set. Conversely, their analysis handles the greedy ranking step under ϵ-greedy protection, which Corollary 3.9 does not; the two results are complementary, and the exploration floor is the common principle beneath both exploration devices.

Remark 3.11 (Scope: the candidate-ranking step). Corollary 3.9 covers the *sample-from-l* variant ($k = 1$). Deployed TPE draws $k > 1$ candidates from l^j and evaluates the maximizer of l^j/g^j. The selected point's conditional density is then $k\, l^j(q)\, F_j(q)^{k-1}$, where $F_j(q)$ is the probability that a fresh draw from l^j has a smaller ratio than q; a floor for this density requires a positive lower bound on F_j over Q, equivalently an upper bound on $\sup_Q(l^j/g^j)/\inf_Q(l^j/g^j)$. Whether

the standard constructions of g^j admit such a bound is an open question that the present correspondence makes precise; the classifier-based reformulations of the density ratio in the BORE line of work [15, 16], together with the regret analysis available for such classifier-based variants [17], supply the sharpest available theory for the ratio itself and are the natural tools for attacking it. We also note that the sample-from-l variant is not merely a theoretical device: the number of candidates per iteration is a user-configurable parameter in deployed frameworks, and setting it to one realizes exactly the sampler covered by Corollary 3.9. Note that on the event that all k candidates fall in $B_\varepsilon(q^*)$ — which has conditional probability at least $(w_0^j \underline{p}\, \mu(B_\varepsilon))^k$ — the selected point is in $B_\varepsilon(q^*)$ regardless of the ranking; hence a crude sufficient condition for conclusion (a) is $\sum_j (w_0^j)^k = \infty$, which holds in case (i) of Corollary 3.9 but not in case (ii) for $k \geq 2$.

The correspondence also generates the empirical half of RQ2: since AGM is TPE minus the denominator model and the ranking step, performance differences between the two under a matched protocol estimate what those components contribute — and their cost is precisely the overhead gap measured in Section 5.1.

4. Experimental design

4.1. **Compared algorithms.** *AGM* (log-decay schedule, with out-of-box proposals projected onto the box; $M = 20$, $c = 0.1$, $g = 1$ unless stated otherwise; these values fall in the range found effective in [1]). *Random search* [7], the standard baseline in hyperparameter optimization. *CMA-ES* via the pycma implementation [8] with initial step size 0.3 on the scaled space and default population size. *TPE* via Optuna 4.x [6] with default settings. All algorithms optimize on the scaled decision space $[-1, 1]^n$, mirroring the scaling used in the original numerical study [1], and receive an identical budget of function evaluations. No algorithm receives problem-specific tuning; this deliberately conservative protocol measures out-of-the-box behavior.

4.2. **Test problems and performance measures.** We use multimodal and ill-conditioned problems from the suite of [1]: Rastrigin, Griewank, Rosenbrock, and Zakharov in dimensions $n = 10$ and $n = 20$, and the six-dimensional Hartmann function, all in minimization form. The primary measure is the final optimality gap; the full-scale study additionally reports anytime behavior through fixed-target runtime distributions in the style of the COCO benchmarking methodology [12], pairwise Wilcoxon signed-rank tests, and total wall-clock time including sampler overhead.

4.3. **Protocol.** Each algorithm–problem pair is replicated independently with distinct random seeds. The synthetic study runs at two budgets: a *four-way* comparison at 600 evaluations (5 replications, all four samplers) and a *full-scale* comparison at 10^4 evaluations (10 replications). At the full-scale budget TPE is omitted: its per-trial cost grows with the trial history (total sampler time fitted as $T(N) \propto N^{1.6}$ over budgets $N \in \{300, 600, 900, 1800\}$ on our hardware), which projects roughly fifteen minutes of sampler overhead per replication at 10^4 evaluations, about twenty hours over the full study, against well under a second per replication for AGM; this omission is itself part of the overhead evidence discussed in Section 5.1. The sensitivity study in Section 5.2 is a complete $3 \times 3 \times 3$ factorial

Problem (n)	**AGM**	**Random**	**CMA-ES**	**TPE**
Rastrigin (10)	63.1 ± 6.1	87.2 ± 4.5	49.5 ± 2.7	52.3 ± 4.1
Griewank (10)	0.374 ± 0.030	0.750 ± 0.085	0.019 ± 0.010	0.236 ± 0.021
Rosenbrock (10)	127 ± 24	8500 ± 2074	9.36 ± 1.60	200 ± 61
Hartmann (6)	0.096 ± 0.032	0.798 ± 0.139	4.0e-6 $\pm$ 1.1e-6	0.024 ± 0.002
Zakharov (10)	49.0 ± 8.3	95.6 ± 22.4	29.2 ± 4.1	49.2 ± 18.5

TABLE 2. Four-way comparison: mean final optimality gap ($\pm$ s.e.) over 5 replications, budget 600 evaluations. The full-scale study at 10^4 evaluations is reported in Table 4.

Algorithm	**Wall-clock time (s)**	**Overhead per evaluation**
AGM	0.02–0.04	$O(1)$; a single Gaussian draw
Random	< 0.02	$O(1)$
CMA-ES	0.1–0.4	$O(n^2)$ covariance update
TPE	10–18	grows with the number of past trials

TABLE 3. Wall-clock time for one 600-evaluation run on the synthetic benchmarks (test-function cost is negligible).

over $M \in \{5, 20, 50\}$, $c \in \{0.05, 0.1, 0.5\}$, $g \in \{0.5, 1, 2\}$ with 5 replications per cell (135 runs per problem). The threshold-validation study is described in Section 5.3, and the hyperparameter-tuning study in Section 6. Software versions and reproduction commands are listed in Appendix B.

5. Numerical experiments: synthetic benchmarks

5.1. **Comparison with baseline optimizers (RQ2).** Table 2 reports the preliminary results (budget 600 evaluations, 5 replications, mean optimality gap $\pm$ standard error). Three patterns emerge and are stable across problems.

First, *AGM dominates random search on every problem*, with gap reductions ranging from a factor of about 1.4 (Rastrigin) to a factor of 67 (Rosenbrock). Since random search remains a common default in hyperparameter tuning [7], this establishes AGM as a strictly better drop-in replacement at essentially zero additional implementation or computational cost.

Second, *AGM is comparable to TPE on several multimodal landscapes.* On Rastrigin (63.1 vs. 52.3) and Zakharov (49.0 vs. 49.2) the two methods are statistically indistinguishable at this budget, and on Rosenbrock AGM is ahead (127 vs. 200). TPE holds an advantage on Griewank and Hartmann, landscapes with a strong global basin where the good/bad density-ratio model concentrates quickly. Read through Observation 3.8, this is direct evidence on RQ2: on a meaningful subset of problems, the denominator model and the candidate-ranking step, the very components AGM deletes, contribute little at small budgets, while on basin-structured problems they pay for themselves.

Third, *the overhead gap is dramatic.* Table 3 shows the wall-clock cost of a 600-evaluation run. AGM completes in roughly 0.02–0.04 seconds; TPE requires 10–18 seconds because its per-trial cost grows with the evaluation history. When individual evaluations are cheap (small models, fast simulations, per-instance tuning), sampler overhead of this magnitude dominates total runtime, and AGM's $O(1)$ iteration cost becomes a practical advantage rather than a curiosity.

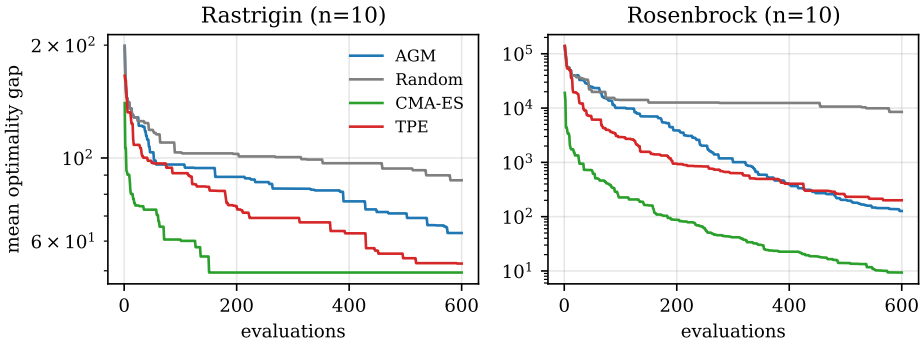


FIGURE 1. Mean best-so-far optimality gap over 5 replications at budget 600 (four samplers, log scale). AGM separates from random search within the first hundred evaluations and tracks the model-based samplers thereafter.

Problem (n)	**AGM**	**Random**	**CMA-ES**
Rastrigin (10)	27.5 ± 1.9	71.3 ± 2.7	12.4 ± 1.7
Rastrigin (20)	122 ± 4.3	198 ± 3.6	35.1 ± 3.0
Griewank (10)	0.083 ± 0.009	0.522 ± 0.030	0.010 ± 0.004
Griewank (20)	0.164 ± 0.011	1.053 ± 0.003	0.001 ± 0.001
Rosenbrock (10)	14.6 ± 0.5	2507 ± 360	2.6e-14 ± 8.5e-15
Rosenbrock (20)	92.7 ± 15.0	34530 ± 2400	4.33 ± 0.44
Zakharov (10)	0.393 ± 0.015	47.5 ± 3.8	1.8e-14 ± 4.1e-15
Hartmann (6)	0.050 ± 0.019	0.315 ± 0.038	0.012 ± 0.012

TABLE 4. Full-scale study: mean final optimality gap (± s.e.) over 10 replications, budget 10^4 evaluations. TPE omitted (projected sampler overhead $\approx$ 15 min per replication, $\approx$ 20 h for the full study; see Section 4).

Figure 1 displays the mean anytime behavior on two representative problems.

CMA-ES attains the best solution quality on all five problems at this budget, consistent with its standing as the strongest general-purpose continuous optimizer [4]. The comparison we draw is therefore not that AGM replaces CMA-ES on smooth continuous problems, but that it recovers a large fraction of the gap between random search and sophisticated samplers with almost none of their machinery — and that against TPE specifically, the trade is competitive quality at a small fraction of the overhead.

Table 4 reports the full-scale study: budget 10^4 evaluations, 10 replications, dimensions up to $n = 20$. As explained in Section 4, TPE is omitted here because its projected sampler overhead is roughly fifteen minutes per replication at this budget, some four orders of magnitude above AGM's and about twenty hours over the full study: the overhead asymmetry of Table 3, made concrete.

Figure 2 shows the corresponding anytime behavior at the full-scale budget.

Two budget-scaling effects stand out. First, AGM's advantage over random search *widens* with budget: on Zakharov the ratio grows from about 2× at 600

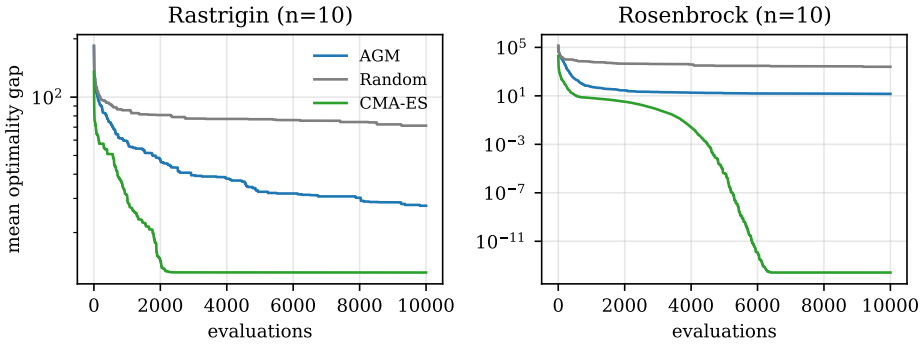


FIGURE 2. Mean best-so-far optimality gap over 10 replications at budget 10^4 (log scale). Random search stalls after the early phase while AGM continues to improve throughout; CMA-ES reaches machine precision on Rosenbrock.

	Rastrigin (10)			**Rosenbrock (10)**		
M	5: 66.1	20: 69.8	50: 75.2	5: 2182	20: 5961	50: 5940
c	0.05: 64.1	0.1: 62.3	0.5: 84.7	0.05: 4566	0.1: 1889	0.5: 7628
g	0.5: 81.7	1: 68.9	2: 60.6	0.5: 6731	1: 1649	2: 5703
Best / worst cell	54.5 / 119.0			14.7 / 31438		

TABLE 5. Main-effect means of the final optimality gap over the $3 \times 3 \times 3$ factorial (600 evaluations, 5 replications per cell), and best/worst cell means across all 27 configurations.

evaluations to 121× at 10^4; on Rosenbrock from 67× to 172× (n=10) and 372× (n=20). Random search stalls because the probability of improving on the incumbent by blind draws collapses, while AGM's elite-centered mixture keeps concentrating proposal mass. Second, on the smooth unimodal problems CMA-ES converges to machine precision, so its lead over AGM grows without bound there; on the multimodal Rastrigin landscapes the relative gap remains within a small constant factor (2.2× at n=10, 3.5× at n=20). Both effects are consistent with the structural reading of Section 3.3: AGM buys global-search robustness with no curvature model, so it should track — but not match — covariance adaptation on smooth problems and stay closest on rugged ones.

5.2. **Parameter sensitivity (RQ3).** Table 5 reports main-effect means of the final gap over the full $3 \times 3 \times 3$ factorial, together with the best and worst cell means, for one canonical multimodal problem (Rastrigin) and one canonical ill-conditioned valley (Rosenbrock).

The dichotomy is sharp. On Rastrigin the entire response surface is nearly flat: the worst of the 27 configurations (119.0) is only about twice the best (54.5), main effects are mild and monotone (faster bandwidth decay, i.e. larger g, and smaller c help; M is nearly irrelevant), and any reasonable configuration performs acceptably. On Rosenbrock the spread across configurations exceeds three orders of magnitude (14.7 to 31438), main effects are non-monotone in c and g, and a small

	$a = 1.0$ $(a^2 < 2n)$	$a = 1.5$ $(a^2 \geq 2n)$
$g = 0.4$ (subcritical)	1.00	1.00
$g = 0.5$ (critical)	0.83	1.00
$g = 1$ (supercritical)	0.12	0.23
$g = 2$ (supercritical)	0.06	0.12

TABLE 6. Escape rate to the global optimum over 200 replications (budget 5,000, initialization forced into the suboptimal basin). The theoretical threshold of Corollary 3.6 is visible exactly at $g = 1/2$: the constant condition $a^2 \geq 2n$ separates 0.83 from 1.00.

elite set ($M = 5$) is strongly preferred: mean gap 2182 for $M = 5$ against roughly 5950 for $M \in \{20, 50\}$. The latter finding quantitatively replicates the qualitative observation in [1] that larger M does not aid convergence on Rosenbrock: when the elite set spreads along a curved valley, a mixture centered on its members proposes points off the valley floor, so concentration, not exploration, is what the landscape rewards.

The answer to RQ3 is therefore nuanced: "few parameters" does not by itself imply "insensitive parameters." On multimodal landscapes the claim holds and AGM is safe to deploy untuned; on ill-conditioned landscapes the interaction structure among (M, c, g) matters as much as it would in more heavily parameterized methods. This sharpens, and gives a concrete empirical target for, the adaptive-bandwidth extension anticipated in [1]: an online schedule that contracts σ when improvements stall would need to close a measured 148× gap between the best and worst static configurations on Rosenbrock while preserving the flat-surface behavior on Rastrigin.

5.3. **Validating the schedule threshold (RQ1).** Corollary 3.6 and Proposition 3.7 predict a phase transition in the pure log-decay sampler at $g = 1/2$, with the constant condition $c^2 M \geq 2n$ deciding the critical case. We test this in a finite-budget simulation on a two-bump instance of the form used in Appendix A ($n = 1$, so the critical constant condition reads $a^2 = c^2 M \geq 2$): a global bump of height 2 and a local bump of height 1 separated by distance ≈ 1.35 on $Q = [-1, 1]$; $M = 5$; initialization forced into the local basin; budget 5,000; 200 replications. Table 6 reports the fraction of runs whose best value escaped to the global bump.

The pattern matches the theory precisely. Below the critical exponent every run escapes regardless of the constants. At $g = 1/2$ the constant condition $c^2 M \geq 2n$ is exactly discriminating: the compliant configuration escapes in all 200 runs while the non-compliant one is trapped in 17% of them. Above the critical exponent trapping dominates, as Proposition 3.7 predicts. We emphasize that these are finite-budget frequencies illustrating the mechanism of summable versus non-summable exploration mass, rather than a direct test of the asymptotic statement.

6. Numerical experiments: hyperparameter tuning

We evaluate the samplers on two realistic tuning tasks with 3-fold cross-validated log-loss as the objective, a budget of 200 evaluations, and 10 independent replications per sampler. *Task A* tunes XGBoost (30 trees) on the breast-cancer dataset over 6 hyperparameters: the learning rate on a log scale over $[10^{-3}, 1]$, maximum

	AGM	Random	CMA-ES	TPE
Task A: XGBoost	0.0855 ± 0.0013	0.0975 ± 0.0027	0.0849 ± 0.0013	0.0842 ± 0.0007
Task B: MLP	0.1969 ± 0.0045	0.2174 ± 0.0035	0.1972 ± 0.0020	0.1907 ± 0.0019

TABLE 7. Tuned 3-fold CV log-loss (mean ± s.e., 10 replications, budget 200 evaluations).

depth ($\{2, \dots, 10\}$), subsample ratio ($[0.5, 1]$), column subsample ($[0.3, 1]$), minimum child weight ($\{1, \dots, 20\}$), and L_2 regularization (log-scale, $[10^{-3}, 10^2]$). *Task B* tunes a multilayer perceptron (SGD, 40 epochs) on a 500-sample subset of the digits dataset over 5 hyperparameters: layer width ($\{16, \dots, 96\}$), depth ($\{1, 2\}$), initial learning rate and L_2 penalty (both log-scale), and momentum ($[0.5, 0.99]$). All samplers see the identical $[-1, 1]^d$ encoding with the same decoders. We examine three questions: (i) whether AGM improves on random search uniformly; (ii) whether AGM is within the confidence band of TPE on at least half of the tasks; and (iii) how the samplers compare in total tuning time.

Table 7 reports the tuned losses. On question (i), AGM improves significantly on random search on both tasks (two-sided Mann–Whitney, $p = 0.001$ on Task A, $p = 0.007$ on Task B). On question (ii), the answer is affirmative on both tasks rather than half: AGM is statistically indistinguishable from TPE on Task A ($p = 0.121$) and Task B ($p = 0.473$), and likewise from CMA-ES ($p = 0.678$ and $p = 0.473$). On these realistic, low-effective-dimensional response surfaces [10], the components AGM deletes from TPE — the non-promising density model and the candidate ranking — produced no detectable quality difference at this budget, which is the field version of the RQ2 finding of Section 5.1. On question (iii), at a budget of 200 evaluations the objective cost dominates for every sampler, and total tuning times per replication were comparable across samplers (roughly 10–13 s on Task A and 32–36 s on Task B). The overhead distinction is a large-budget phenomenon: TPE's per-trial cost grows with the history (Table 3), which is why it is already impractical at 10^4 evaluations (Table 4), while AGM's cost per iteration is constant. The overhead comparison is therefore budget-dependent: AGM's advantage is negligible at small budgets and decisive at large ones.

Two caveats bound these conclusions. Even with 10 replications the nonparametric tests have limited power, so "indistinguishable from TPE" should be read as an absence of evidence for a difference rather than evidence of equivalence; and both tasks are small classification problems chosen for reproducibility. Larger models and budgets, and regression tasks, remain for future work.

7. Discussion and conclusion

This paper positioned the AGM algorithm of [1] in the contemporary black-box optimization landscape, theoretically and empirically. The theoretical side isolated the exploration floor as the single property driving the convergence argument, extended the guarantee to every elite kernel sampler with a non-summable floor, made the log-decay schedule's guarantee exact, including the interpretable critical condition $c^2 M \geq 2n$ and a matching impossibility result above the threshold, and transferred the guarantee to prior-mixed Parzen samplers of the TPE family, for which no convergence statement had previously been available. The precise gap

that remains (the candidate-ranking step of deployed TPE) is now an explicitly stated open problem, localized in a single density ratio.

The empirical side showed that AGM uniformly improves on random search, significantly on both tuning tasks and by factors growing to two orders of magnitude at the 10^4 budget on synthetic problems, while remaining statistically indistinguishable from TPE on both hyperparameter-tuning tasks at constant per-iteration cost, and trailing covariance-adaptive methods on smooth continuous problems, as expected. The sensitivity study disciplined the simplicity narrative: AGM is robust to its parameters exactly where its mixture model matches the landscape and sensitive where it does not. And the trapping simulation confirmed the theoretical threshold exactly at the critical exponent.

The theory and the experiments jointly delineate a design trade-off that we believe deserves emphasis. Aggressive bandwidth decay ($g > 1/2$) buys short-run exploitation, and indeed was the best static choice at budget 600 on multimodal problems (Section 5.2), at the price of a positive probability of permanent trapping (Proposition 3.7, Table 6); conservative decay preserves the global guarantee at some short-run cost. A floored variant, mixing a fixed or slowly vanishing uniform component into the log-decay mixture, enjoys both: the finite-budget behavior of the chosen schedule and the guarantee of Theorem 3.2. Our comparison experiments use the pure log-decay sampler with box projection exactly as specified in Section 4; the asymptotic guarantee attaches to the floored variant, which differs from the implemented sampler only by an occasional uniform draw.

Two extensions follow naturally. First, the bandwidth constants (c, g) could be adapted online from the observed improvement rate; Theorem 3.2 accommodates any such schedule that respects a persistent floor, Section 5.2 supplies the quantitative target such a schedule must meet, and the recent trend toward online-adaptive TPE variants [11] indicates demand for exactly this kind of guarantee-preserving adaptation. A natural companion question is the convergence *rate*: for the constant-floor (ϵ-greedy-type) case, rates are available under surrogate-model assumptions [19], and extending such rate analysis to vanishing persistent floors, where Corollary 3.6 pins down the boundary of almost-sure convergence, is open. Second, the noisy-objective case, in which the elite ranking itself is uncertain, would connect the algorithm to the sequential decision-making motivation of [1]. More broadly, the correspondence of Section 3.3 suggests treating AGM as the analyzable core of the TPE family and asking which TPE components (quantile-based elite sets, the denominator model, candidate ranking, per-dimension bandwidths, and the metric kernels of recent combinatorial extensions [18]) preserve the exploration-floor property and hence the guarantee, and which trade it for empirical speed.

Appendix A. Proof of Proposition 3.7

Fix $n = 1$, $Q = [-1, 1]$ (the construction extends to any n). Let $0 < \rho < 1/8$ and define the continuous positive function

$$r(q) \;=\; \max\Big\{\, 2\,\psi\big((q - q^*)/\rho\big),\; \psi\big((q - s)/\rho\big),\; \tfrac{1}{20}\Big\}, \qquad q^* = \tfrac{3}{4},\; s = -\tfrac{3}{4},$$

where $\psi(u) = (1 - |u|)_+$ is the triangular bump. Then r has the unique global maximizer q^* with $r(q^*) = 2$, a strict local maximizer s with $r(s) = 1$, and $r \le 1/20$ outside the two bumps; Assumptions 1–3 hold by continuity. Write $G := \{q : r(q) >$

$1/2\} \cap (q^* - \rho,\, q^* + \rho)$, note $\{q : r(q) > 1\} \subseteq G$, and set $d = \operatorname{dist}\big(G, [s-\rho, s+\rho]\big) \geq \frac{3}{2} - 2\rho \geq 1$.

Initialization event. Let $E_0 = \{q_i^1 \in (s - \rho/2,\, s + \rho/2) \text{ for } i = 1, \ldots, M\}$; if the initial M points are drawn uniformly on Q (or by any rule assigning positive mass to that interval), $\mathbb{P}(E_0) =: p_0 > 0$.

Elites cannot leave the local bump except through G. On E_0, every initial elite value lies in $[\psi(1/2), 1] = [1/2, 1]$. A later sample q enters the elite set only if $r(q)$ exceeds the current minimum elite value, which is always $\geq 1/2$; since $r \leq 1/20$ outside the two bumps and $r \leq 1$ on the local bump, any sample entering the elite set lies either in the local bump $[s - \rho, s + \rho]$ or in G. Define the stopping time $\tau = \inf\{j : q^j \in G\}$; for all $j < \tau$ every kernel center lies in $[s-\rho, s+\rho]$, at distance at least d from G.

Escape probability is summable. For $j < \tau$ the sampling density at any $q \in G$ is a convex combination of truncated kernels with centers at distance $\geq d$ from q. Since each center lies in $Q = [-1, 1]$ and $\sigma_j \leq \sigma_2 =: \bar{\sigma}$, each one-dimensional normalizer satisfies $\int_{-1}^{1} K_i \geq \nu > 0$ with the explicit constant $\nu = \Phi(0) - \Phi(-2/\bar{\sigma})$, where Φ is the standard normal distribution function: the normalizer $\Phi((1-c)/\sigma) - \Phi((-1-c)/\sigma)$ is minimized over centers $c \in [-1, 1]$ at the endpoints and is decreasing in σ on $(0, \bar{\sigma}]$ there. Hence

$$\hat{f}^{\,j}(q) \;\leq\; \nu^{-1}(2\pi\sigma_j^2)^{-1/2} \exp\big(-d^2/(2\sigma_j^2)\big) \qquad (q \in G),$$

so with $\beta_j := \nu^{-1}\mu(G)\,(2\pi\sigma_j^2)^{-1/2} e^{-d^2/(2\sigma_j^2)}$,

$$\mathbb{P}(\tau = j+1 \mid \mathcal{F}_j,\ \tau > j,\ E_0) \;\leq\; \beta_j.$$

Substituting $\sigma_j = a/(\log j)^g$ gives $\beta_j = C(\log j)^g \exp\{-(d^2/2a^2)(\log j)^{2g}\}$, which is summable exactly as in the proof of Corollary 3.6: for $g > 1/2$ this holds for any $d > 0$, while in the critical case $g = 1/2$ it holds whenever $d^2/(2a^2) > 1$. To enforce the latter, place the bump centers at $q^* = 1 - 2\rho$ and $s = -(1 - 2\rho)$ instead; the bump supports remain in the interior of Q and are separated by $d \geq 2 - 6\rho$, so $d^2/(2a^2) = \kappa\, d^2/(4n) = \kappa(1 - 3\rho)^2 \to \kappa > 1$ as $\rho \downarrow 0$, and the exponent strictly exceeds 1 for ρ small. Moreover $\sum_{j \geq j_0} \beta_j \to 0$ as $j_0 \to \infty$; and by shrinking ρ (which shrinks $\mu(G)$ and enlarges d) we may ensure $S := \sum_{j \geq 2} \beta_j < 1$.

Conclusion. By the layered union bound over the stopping time,

$$\mathbb{P}(\tau < \infty \mid E_0) \;\leq\; \sum_{j \geq 2} \mathbb{E}\big[\mathbb{P}(\tau = j+1 \mid \mathcal{F}_j)\, \mathbf{1}\{\tau > j\} \,\big|\, E_0\big] \;\leq\; S \;<\; 1.$$

On $E_0 \cap \{\tau = \infty\}$ no evaluated point ever lies in G; since $Y_j > 1$ requires an evaluated point in $\{r > 1\} \subseteq G$, we get $Y_\infty \leq 1 < 2 = r(q^*)$ there, and $\mathbb{P}\big(Y_\infty < r(q^*)\big) \geq p_0(1 - S) > 0$. □

Appendix B. Reproducibility

All experiments were run with Python 3, NumPy 2.4.4, pycma 4.4.4, Optuna 4.9.0, scikit-learn 1.8.0, SciPy 1.17.1, and XGBoost 3.4.0, with seeds $0, \ldots, \text{reps} - 1$ per algorithm–problem pair. The synthetic comparison and sensitivity studies are reproduced by the accompanying script `agm_experiments.py`: `--budget 600 --reps 5` reproduces Table 2 and `--budget 10000 --reps 10 --algos AGM Random CMA-ES` reproduces Table 4. The TPE overhead projection at 10^4 evaluations uses single-run wall-clock measurements at $N \in \{300, 600, 900, 1800\}$ trials

on the same machine; a log–log fit gives total sampler cost $T(N) \propto N^{\alpha}$ with $\alpha \approx 1.6$, hence $T(10^4) \approx 9 \times 10^2$ s per replication. The hyperparameter-tuning study is reproduced by `hpo_experiments.py` (e.g., `--task A --algo AGM --reps 10`); search-space decoders are defined in that script exactly as described in Section 6. The trapping study of Section 5.3 uses the two-bump instance defined in Appendix A with smoothed bumps of widths 0.05 and 0.15, escape threshold $Y_j > 1.5$, and forced initialization uniform on $[-1, -0.2]$.

Acknowledgment

This research was supported by the Hankuk University of Foreign Studies Research Fund.

References

[1] S. Lee, Adaptive Gaussian mixture model in global optimization, *Global and Stochastic Analysis* 5(2) (2018) 153–170.

[2] M. Zlochin, M. Birattari, N. Meuleau, M. Dorigo, Model-based search for combinatorial optimization: A critical survey, *Annals of Operations Research* 131(1) (2004) 373–395.

[3] J. Hu, M. C. Fu, S. I. Marcus, A model reference adaptive search method for global optimization, *Operations Research* 55(3) (2007) 549–568.

[4] N. Hansen, A. Ostermeier, Completely derandomized self-adaptation in evolution strategies, *Evolutionary Computation* 9(2) (2001) 159–195.

[5] J. Bergstra, R. Bardenet, Y. Bengio, B. Kégl, Algorithms for hyper-parameter optimization, *Advances in Neural Information Processing Systems* 24 (2011).

[6] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, *Proceedings of KDD* (2019) 2623–2631.

[7] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, *Journal of Machine Learning Research* 13 (2012) 281–305.

[8] N. Hansen, Y. Akimoto, P. Baudis, CMA-ES/pycma on GitHub, Zenodo (2019).

[9] R. Turner et al., Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020, *Proceedings of the NeurIPS 2020 Competition and Demonstration Track* (2021).

[10] M. Feurer, F. Hutter, Hyperparameter optimization, in: *Automated Machine Learning*, Springer, 2019, 3–33.

[11] S. Watanabe, Tree-structured Parzen estimator: Understanding its algorithm components and their roles for better empirical performance, arXiv:2304.11127 (2023).

[12] N. Hansen, A. Auger, R. Ros, O. Mersmann, T. Tušar, D. Brockhoff, COCO: A platform for comparing continuous optimizers in a black-box setting, *Optimization Methods and Software* 36(1) (2021) 114–144.

[13] R. Durrett, *Probability: Theory and Examples*, Duxbury Press, 1995.

[14] S. Watanabe, N. Awad, M. Onishi, F. Hutter, Speeding up multi-objective hyperparameter optimization by task similarity-based meta-learning for the tree-structured Parzen estimator, *Proceedings of IJCAI-23* (2023) 4380–4388. doi:10.24963/ijcai.2023/487.

[15] L. Tiao, A. Klein, M. Seeger, E. Bonilla, C. Archambeau, F. Ramos, BORE: Bayesian optimization by density-ratio estimation, *Proceedings of ICML* (2021).

[16] J. Song, L. Yu, W. Neiswanger, S. Ermon, A general recipe for likelihood-free Bayesian optimization, *Proceedings of ICML* (2022).

[17] R. Oliveira, L. Tiao, F. Ramos, Batch Bayesian optimisation via density-ratio estimation, with guarantees, *Advances in Neural Information Processing Systems* 35 (2022).

[18] K. Abe, Y. Wang, S. Watanabe, Tree-structured Parzen estimator can solve black-box combinatorial optimization more efficiently, arXiv:2507.08053 (2025).

[19] J. Li, I. O. Ryzhov, Convergence rates of epsilon-greedy global optimization under radial basis function interpolation, *Stochastic Systems* 13(1) (2023) 59–92.

HUFS Business School, Hankuk University of Foreign Studies, Seoul, 02450, Republic of Korea
Email address: shlee2016@hufs.ac.kr