Center-based Sampling in Optimization and Machine Learning: A Theory and Review

Rasa Khosrowshahli*, Shahryar Rahnamayan, SMIEEE†, and Beatrice Ombuki-Berman*
*Department of Computer Science, †Yousef Haj-Ahmad Department of Engineering
Brock University
St. Catharines, ON, Canada
{rkhosrowshahli, bombuki, srahnamayan}@brocku.ca

***Abstract*—High-dimensional optimization and machine learning (ML) pipelines frequently rely on randomized proposals within box-constrained domains, yet it is often unclear when concentrating proposals near the box center increases the likelihood of generating useful candidates. We study center-based sampling (CBS) in the D-dimensional unit hypercube, where candidates are drawn from a central subcube rather than uniformly from the full box. Under a simple single-draw model, we prove that CBS reduces expected squared distance to an independent target by a fixed per coordinate amount, so the total reduction grows linearly with the dimension. When the target is uniform on the hypercube, this means advantage upgrades to a high probability regime in which the probability that a center draw is closer than a uniform draw converges to one exponentially fast in D, with explicit Hoeffding and Bernstein guarantees and an accurate central limit theorem (CLT) approximation. The expectation comparison also extends beyond the uniform target model to any independent target with a finite second moment, for which the mean advantage is unchanged because the uniform and CBS samplers share the same center while CBS has a smaller variance. We complement the theory with the same budget random search experiments on shifted Sphere, Rastrigin, and Ackley functions over box-constrained domains, where CBS consistently improves the best objective under a fixed sample budget, with the strongest gains when the optimum is centrally located. These results position CBS as a useful geometric baseline for box-constrained randomized search, while clarifying that the concentration theorem applies only to the uniform target cube model.**

***Index Terms*—high-dimensional search space, center-based sampling, optimization, machine learning**

I. Introduction

Random sampling is a workhorse in modern optimization and ML pipelines. It appears directly in black-box optimization methods for parameter initialization in trust region [1] and partition-based global search [2] and in population-based heuristics such as differential evolution (DE) [3], particle swarm optimization (PSO) [4], and evolution strategies (ES) [5]. It also underlies practical routines such as bounded hyperparameter search via random search [6] and Bayesian optimization [7], randomized initialization (Glorot/Xavier [8] and He [9]), and stochastic evaluation with mini-batches [10]. In these settings, one repeatedly proposes candidates in a bounded domain and hopes that some of them land close to an unknown optimum.

In black-box problems, the optimum can lie anywhere in the search landscape. Without strong prior information, treating its location as random or poorly predictable under noise and model mismatch is a standard idealization. A natural question is whether some proposal distributions produce candidates that are typically closer to useful solutions than uniform box sampling. We answer this question in a box-constrained reference model. Many practical routines, including bounded hyperparameter search, trust region methods, and population initialization, operate over explicit lower and upper bounds and can be affinely rescaled to $[0, 1]^D$. The hypercube is therefore not meant to encode all feasible sets and it is the cleanest box model in which the effect of centering can be isolated from algorithm adaptation, surrogate modeling, and constraint handling. We compare (i) uniform sampling over $[0, 1]^D$ and (ii) CBS from a central subcube, and show that the center-based choice becomes strongly favored as dimension grows. This helps explain why center-based heuristics have been used for population initialization and variation in evolutionary computation, including center-based population sampling [11], enhanced DE with center-based sampling [12], centralized DE initialization [13], and opposite-center learning in DE [14]. More broadly, this line of work is connected to opposition-based learning (OBL) [15], which made explicit the value of exploiting opposite or central counterparts in bounded search spaces. Our analysis makes the scaling with dimension explicit under a probabilistic model.

This hypercube model with a uniform random target is deliberately simple. It serves as a baseline when target locations vary across instances at the domain scale, or when the relevant configuration is poorly localized due to limited prior information, noise, or model mismatch. The model isolates one geometric mechanism, namely that at fixed mean, shrinking sampling variance toward the center reduces mean squared distance, and in high dimensions, this per-coordinate improvement can accumulate. The same simplicity is also a limitation. Real feasible sets can have nonlinear constraints or variable couplings, objectives need not behave like squared distance to an independent target, and practical algorithms usually compare many samples or adapt their proposal distribution over time rather than relying on a single independent draw. We therefore treat the hypercube result as a clean reference point for box-constrained randomized search, not as a universal prescription. The single-draw comparison isolates the centering mechanism by comparing the best of N center samples

to the best of N uniform samples, a practically important next step. This perspective also helps interpret the recent center-bias critique of evolutionary benchmarking by Kudela [16]. If benchmark optima are disproportionately central, algorithms with implicit center bias can appear stronger than they would on less center-favoring problems. Our results identify a geometric mechanism underlying center-biased gains, while making clear that the advantage is conditional, not universal.

Our main message is that, in high dimensions, even a modest per coordinate improvement compounds into a decisive overall effect. Concretely, we show that sampling uniformly from a central subcube reduces the expected squared distance to the random target by a fixed amount per coordinate, hence by an amount that grows linearly in D. Next, we show that this mean advantage becomes a high-probability guarantee. Specifically, the probability that a center-sampled point is closer than a uniformly sampled point approaches one exponentially fast as the dimension grows, with an explicit exponent depending on the subcube width.

The analysis is also interpretable through the lens of centering and variance control. In one dimension, the expected squared distance to a uniform target is governed by second moments. Shrinking the sampling region around the target mean reduces variance and, therefore, reduces mean-squared error. This viewpoint connects the bounded-domain result to widely used centered sampling schemes in ML, such as variance-scaled Glorot/Xavier [8] and He [9] initializations, and to evolution strategies that sample from Gaussians centered at an iteratively updated mean, as in the covariance matrix adaptation evolution strategy (CMA-ES) [5] and natural evolution strategies (NES) [17].

Contributions.

- A closed form comparison of expected squared distances in the hypercube model, together with the observation that this mean advantage extends to any independent target with finite second moment because CBS and uniform sampling share the same center, while CBS has smaller variance.
- Explicit Hoeffding and Bernstein lower bounds on the probability that a central sample is closer than a uniform-hypercube sample, showing exponential convergence to near certainty under the uniform target cube model.
- Monte Carlo validation of the probability bounds and a same budget benchmark study on shifted Sphere, Rastrigin, and Ackley functions showing that CBS also improves fixed budget random search on standard box-constrained landscapes.
- A discussion of scope, including which parts of the analysis depend on the uniform target cube assumption and which do not, and of the best of N, adaptive, and non-box settings that remain open.

Organization. Section II situates our theory within related center-based ideas across optimization and ML. Section III-A establishes the hypercube model and proves the high-dimensional advantage for central subcube sampling, and Section V concludes.

II. Background and Related Work

Centering and variance control recur across high-dimensional optimization and ML, but they play different roles in different communities. Our aim in this section is not to provide an exhaustive survey of every center-based method. Instead, we organize the literature around a single theme that is directly relevant to our analysis. When uncertainty is high, sampling or aggregating around a center can reduce variance and improve reliability. We therefore move from the geometric intuition behind the theory to optimization algorithms that explicitly maintain centers, and finally to ML settings where centered distributions or averages improve stability. This organization is intended to clarify both why CBS is natural and where the limits of our theorems lie.

A. High-dimensional geometry and sampling baselines

The theoretical core of our paper is geometric. In high dimensions, a small per coordinate reduction in expected squared distance accumulates into a large aggregate effect. Our main theorem formalizes this intuition. If the center-based sampler has a fixed one-dimensional advantage, concentration across coordinates upgrades that mean advantage into a high-probability guarantee that strengthens exponentially with dimension. This phenomenon comes from the product structure of the sampling model rather than from any optimization specific mechanism.

Uniform i.i.d. sampling is the cleanest baseline for isolating this effect, but it is not the only reasonable baseline in practice. Derivative-free optimization often relies on space filling designs such as Latin hypercube sampling (LHS) and quasi-Monte Carlo (QMC) sequences to reduce discrepancy and improve global coverage. Those designs change the dependence structure and coverage properties, so their interaction with center bias depends on the objective and target model. Because our goal is to isolate the centering effect itself, the present theory compares single independent draws. An important next step is to study the likelihood that a center sample is closer under dependent designs such as LHS and QMC, rather than treating that question as already settled by our argument.

B. Center-based mechanisms in derivative-free and population-based optimization

In optimization, center-based behavior usually appears adaptively rather than through a fixed central subcube. The simplest examples arise in local modeling such as trust region methods sample and build models around a current iterate [1], and partition-based global methods such as DIviding RECTangles (DIRECT) which use rectangle centers as representative points to balance exploration and exploitation [2]. Both implicitly concentrate proposals near a current best estimate, which is an adaptive form of CBS built on the belief that promising solutions are nearby.

Population-based methods make the center role explicit. Evolution strategies (ES) draw candidates from a distribution centered at the current population mean, then update that mean

by a fitness weighted recombination of the best samples, as in CMA-ES [5] and natural evolution strategies (NES) [17]. In a canonical ES/CMA-ES iteration,

$$\theta_k^{(t)} \sim \mathcal{N}(\mu_t, \Sigma_t), \quad k = 1, \ldots, \lambda, \tag{1}$$

and after evaluation, the mean is updated by weighted recombination,

$$\mu_{t+1} = \sum_{i=1}^{\lambda} w_i \, \theta_{i:\lambda}^{(t)}, \qquad w_i \geq 0, \ \sum_{i=1}^{\lambda} w_i = 1, \tag{2}$$

where $\theta_{i:\lambda}^{(t)}$ denotes the i-th best sample under the fitness. This loop is a natural algorithmic counterpart to our bounded-domain theory in which candidates are sampled around a center, and the center is then moved toward stronger performers.

The same pattern appears in several families of population-based algorithms. In differential evolution, CBS has been used explicitly in initialization and variation operators, from early center-based population initialization [11] to enhanced DE variants [12], centralized initialization [13], opposite-center learning [14], opposition-based DE [18], and recent large-scale studies [19], [20]. Estimation of distribution algorithms (EDAs) [21], [22] and cross-entropy methods [23] go one step further by repeatedly fitting and resampling a parametric distribution whose location parameter serves as a learned center. Swarm methods also exploit central tendencies, sometimes implicitly as in PSO [4], and sometimes explicitly as in center PSO [24], center-based particle swarm optimization (CenPSO) [25], and consensus-based optimization, where particles are pulled toward a weighted consensus point [26]–[28]. In multi-objective optimization, centroid ideas appear both in standard frameworks such as the non-dominated sorting genetic algorithm II (NSGA-II) [29], strength Pareto evolutionary algorithm 2 (SPEA2) [30], multiobjective evolutionary algorithm based on decomposition (MOEA/D) [31], and non-dominated sorting genetic algorithm III (NSGA-III) [32], and in explicitly center-guided variants such as CGDE3 [33] and parent-centric crossover (PCX) [34]. Taken together, these examples suggest that the center is not an isolated trick but a recurring organizing device in search algorithms.

OBL is a particularly relevant ancestor for this literature. Tizhoosh [15] formalized the idea of evaluating an estimate together with an opposite counterpart, and later DE variants translated that intuition into opposition- and center-based sampling schemes. This history is important when interpreting recent benchmarking critiques. Kudela [16] argues that methods with implicit center bias can look artificially strong on benchmark suites whose optima are disproportionately central. Our theory complements rather than contradicts that critique, which states that under a centered target model, center bias should help, and when targets move toward the boundary, the advantage can shrink or reverse.

C. Centering and variance control in machine learning

In ML, the same logic appears less as search in a box and more as variance control in parameter or representation space. Neural network initialization is the closest analogy to our hypercube result. Schemes such as Glorot/Xavier [8] and He [9] are centered and primarily calibrated by variance, ensuring stable signal propagation. Section III-A makes the corresponding point in the bounded-domain setting. When means are matched, the expected squared distance is determined by variances, so shrinking the variance toward a meaningful center improves the average distance.

Center-based representations also appear across learning systems, from k-means [35], Lloyd's algorithm [36], and Kohonen self-organizing maps [37] to radial basis function (RBF) networks [38], prototypical networks [39], and variational autoencoders in which the latent mean acts as a representative embedding [40]. In Bayesian optimization, Gaussian process surrogates [41] provide posterior means that define natural exploitation centers, while acquisition functions such as expected improvement [42] balance that exploitation with uncertainty driven exploration [7].

Centered aggregation also appears in noisy learning settings, where averaging can improve robustness and generalization.

III. Central sampling is closer in high dimensions

A. Setup

Let $D \in \mathbb{N}$ be the dimension and consider the unit hypercube $[0,1]^D$. Let the (random) target be $T = (T_1, \ldots, T_D)$ with i.i.d. coordinates $T_i \sim \mathrm{Unif}[0,1]$. We compare two independent candidate points.

- **Uniform cube sampling.** $X = (X_1, \ldots, X_D)$ with i.i.d. $X_i \sim \mathrm{Unif}[0,1]$.
- **Center-based sampling.** Fix a half-width $a \in (0, \frac{1}{2})$ and let $Y = (Y_1, \ldots, Y_D)$ with i.i.d. $Y_i \sim \mathrm{Unif}[\frac{1}{2} - a, \frac{1}{2} + a]$.

Assume X, Y, T are mutually independent. We measure closeness using Euclidean distance $\|\cdot\|_2$.

B. Center minimizes expected squared distance

The next lemma quantifies the per coordinate expected squared distance to a uniformly random target.

Lemma 1 (Expected squared distance in one dimension)**.** *Let $U \sim \mathrm{Unif}[0,1]$ and $V \sim \mathrm{Unif}[0,1]$ be independent. Here V represents a uniformly random target location, while U represents a sample drawn uniformly from the entire interval. Then*

$$\mathbb{E}\big[(U-V)^2\big] = \frac{1}{6}. \tag{3}$$

Let $W \sim \mathrm{Unif}[\frac{1}{2} - a, \frac{1}{2} + a]$ be independent of V. Here W represents a sample drawn from a centered subinterval of half-width a around the midpoint $\frac{1}{2}$. Then

$$\mathbb{E}\big[(W-V)^2\big] = \frac{a^2}{3} + \frac{1}{12}. \tag{4}$$

Consequently, the center-based sample W is strictly closer to the random target V in expected squared distance than the complete interval sample U.

$$\mathbb{E}\big[(U-V)^2\big] - \mathbb{E}\big[(W-V)^2\big] = \frac{1-4a^2}{12} > 0 \quad \text{for all } a \in (0, \tfrac{1}{2}). \tag{5}$$

Proof. We use the identity

$$\mathbb{E}\big[(A-B)^2\big] = \mathrm{Var}(A) + \mathrm{Var}(B) + \big(\mathbb{E}[A] - \mathbb{E}[B]\big)^2$$

for independent real random variables A, B.

Uniform vs. uniform. For U, V i.i.d. $\mathrm{Unif}[0,1]$, we have $\mathbb{E}[U] = \mathbb{E}[V] = \frac{1}{2}$ and $\mathrm{Var}(U) = \mathrm{Var}(V) = \frac{1}{12}$, hence $\mathbb{E}[(U-V)^2] = \frac{1}{12} + \frac{1}{12} = \frac{1}{6}$

Center-subinterval vs. uniform. For $W \sim \mathrm{Unif}[\frac{1}{2}-a, \frac{1}{2}+a]$ independent of V, we have $\mathbb{E}[W] = \frac{1}{2} = \mathbb{E}[V]$ and $\mathrm{Var}(W) = \frac{(2a)^2}{12} = \frac{a^2}{3}$, so

$$\mathbb{E}[(W-V)^2] = \mathrm{Var}(W) + \mathrm{Var}(V) = \frac{a^2}{3} + \frac{1}{12}. \quad (6)$$

Subtracting yields $(1-4a^2)/12 > 0$ for $a \in (0, \frac{1}{2})$. □

By independence across the D coordinates, the expected total squared distance advantage is

$$\mathbb{E}\big[\|X-T\|_2^2 - \|Y-T\|_2^2\big] = D \cdot \frac{1-4a^2}{12} = D\mu(a), \quad (7)$$

where $\mu(a) = (1-4a^2)/12 > 0$.

The expectation comparison is driven by second moments (variance control) and is therefore elementary once the target mean is known. The main quantitative message of the paper is not merely that the mean squared distance improves, but that this improvement becomes highly reliable in high dimensions. In particular, the probability that a center draw is closer than a full cube draw converges to 1 at an explicit exponential rate (Theorem 1).

C. Gaussian connection in centering and variance control

The preceding computation is an instance of a more general identity for independent real random variables A, B,

$$\mathbb{E}\big[(A-B)^2\big] = \mathrm{Var}(A) + \mathrm{Var}(B) + \big(\mathbb{E}[A] - \mathbb{E}[B]\big)^2. \quad (8)$$

Thus, when $\mathbb{E}[A] = \mathbb{E}[B]$ (centered with respect to each other), the expected squared distance depends only on the sum of variances. In our hypercube model, $\mathbb{E}[T_i] = \frac{1}{2}$, and the center-based sampler is constructed so that $\mathbb{E}[Y_i] = \frac{1}{2}$, while $\mathrm{Var}(Y_i) = a^2/3 < \mathrm{Var}(X_i) = 1/12$. Eq. (8) shows that the advantage comes directly from variance reduction at fixed mean.

This mirrors the role of initialization in neural networks. Both Gaussian and uniform initializations are typically zero mean, and their qualitative behavior is strongly governed by the chosen variance.

D. Scope beyond the uniform-target cube

Expected distance advantage is not restricted to uniform targets. The variance identity from Eq. (8) yields a broader statement than Lemma 1. Let $T = (T_1, \ldots, T_D)$ be any random target vector that is not necessarily uniform where its coordinates are independent of X and Y and has finite second moment. Since $\mathbb{E}[X_i] = \mathbb{E}[Y_i] = \frac{1}{2}$, $\mathrm{Var}(X_i) = \frac{1}{12}$, and $\mathrm{Var}(Y_i) = \frac{a^2}{3}$, applying Eq. (8) coordinatewise gives

$$\mathbb{E}\big[\|X-T\|_2^2 - \|Y-T\|_2^2\big] = \sum_{i=1}^{D} \left(\frac{1}{12} - \frac{a^2}{3}\right) = D\,\mu(a). \quad (9)$$

Hence, the expected squared-distance advantage of CBS does not rely on the target being uniform or centrally concentrated. The uniform target assumption is used only for the explicit probability bounds developed below.

What the uniform target cube buys. Theorems 1 and 2 require more than an expectation comparison. They exploit the fact that under the uniform target cube model the coordinate gap variables Z_i are i.i.d. and bounded, so classical concentration inequalities apply directly. For general target laws, or for non-product sampling schemes, the mean advantage can survive while the concentration exponent changes or becomes harder to characterize.

Where the model stops being a realistic performance model. The uniform-target cube is a principled reference model for box-constrained search under weak prior information, not a full description of practical optimization. Real problems can have nonlinear feasibility regions, active constraints, curved low-dimensional manifolds, or objectives that are not well approximated by squared distance to an independent target. Likewise, many algorithms operate in a best of N or adaptive regime rather than the single draw setting analyzed here. In those richer settings, CBS can still help or hurt depending on the geometry and the algorithmic context. Kudela's benchmark critique [16] remains relevant at that algorithmic level. If benchmark suites place optima disproportionately near the coordinate center, methods with implicit center bias can look stronger than they would on less-favored center tasks. Our theorem isolates a single geometric mechanism underlying center-biased gains, not by claiming benchmark neutrality or universal superiority.

E. Concentration upgrades mean advantage to near certainty

Define, for $i = 1, \ldots, D$,

$$Z_i := (X_i - T_i)^2 - (Y_i - T_i)^2,$$
$$\Delta := \|X-T\|_2^2 - \|Y-T\|_2^2 = \sum_{i=1}^{D} Z_i. \quad (10)$$

By independence across coordinates, the Z_i are i.i.d. and each satisfies $Z_i \in [-1, 1]$ (since each squared term is in $[0,1]$). By Lemma 1, $\mathbb{E}[Z_i] = (1-4a^2)/12 =: \mu(a) > 0$.

Theorem 1 (Central sample is closer with probability $1 - e^{-\Theta(D)}$)**.** *Fix $a \in (0, \frac{1}{2})$ and define $\mu(a) = (1-4a^2)/12$. Then for all $D \in \mathbb{N}$,*

$$\mathbb{P}\big(\|Y-T\|_2 \le \|X-T\|_2\big) \;\ge\; 1 - \exp\Big(-\tfrac{1}{2}\,D\,\mu(a)^2\Big). \quad (11)$$

In particular, $\mathbb{P}(\|Y-T\|_2 \le \|X-T\|_2) \to 1$ exponentially fast as $D \to \infty$.

Proof. Since Euclidean norms are nonnegative, comparing distances is equivalent to comparing squared distances.

$$\{\|Y-T\|_2 \le \|X-T\|_2\} = \{\|Y-T\|_2^2 \le \|X-T\|_2^2\} = \{\Delta \ge 0\}. \tag{12}$$

Thus, it suffices to lower bound $\mathbb{P}(\Delta \ge 0)$.

We have $\mathbb{E}[\Delta] = \sum_{i=1}^{D} \mathbb{E}[Z_i] = D\mu(a)$ with $\mu(a) > 0$, and $Z_i \in [-1, 1]$. Hoeffding's inequality for sums of independent bounded variables yields [43].

$$\begin{aligned}\mathbb{P}(\Delta - \mathbb{E}[\Delta] \le -\mathbb{E}[\Delta]) &\le \exp\left(-\frac{2\,\mathbb{E}[\Delta]^2}{4D}\right) \\ &= \exp\left(-\frac{2(D\mu(a))^2}{4D}\right) \\ &= \exp\Big(-\tfrac{1}{2}\,D\,\mu(a)^2\Big),\end{aligned} \tag{13}$$

where we used $\sum_{i=1}^{D}(1-(-1))^2 = 4D$. Therefore $\mathbb{P}(\Delta \ge 0) \ge 1 - \exp(-\frac{1}{2}D\mu(a)^2)$, which implies the claimed bound. □

Corollary 1 (Explicit exponent)**.** *With $\mu(a) = (1-4a^2)/12$, Theorem 1 gives*

$$\mathbb{P}\big(\|Y-T\|_2 \le \|X-T\|_2\big) \ge 1 - \exp\left(-\frac{D(1-4a^2)^2}{288}\right). \tag{14}$$

Corollary 2 (Repeated instances)**.** *Fix $a \in (0, \frac{1}{2})$ and $D \in \mathbb{N}$. Let $(X^{(k)}, Y^{(k)}, T^{(k)})$, $k = 1, \dots, n$, be i.i.d. copies of (X, Y, T) from Section III-A. Then the probability that CBS wins at least once across n independent instances satisfies*

$$\begin{aligned}\mathbb{P}\Big(\exists k \in \{1,\dots,n\} :\ &\|Y^{(k)} - T^{(k)}\|_2 \le \|X^{(k)} - T^{(k)}\|_2\Big) \\ &\ge 1 - \big(1 - p_D(a)\big)^n \\ &\ge 1 - \Big(\exp\big(-\tfrac{1}{2}D\mu(a)^2\big)\Big)^n,\end{aligned} \tag{15}$$

where $\mu(a) = (1-4a^2)/12$ and

$$p_D(a) := \mathbb{P}(\|Y-T\|_2 \le \|X-T\|_2). \tag{16}$$

Corollary 2 treats n independent repetitions with independent targets. A sharper question is how the best of n center draws compares to the best of n cube draws for a single fixed target. We defer this order statistics comparison to future work. Many practical optimizers operate in precisely this best of N regime, so the present theorem should be read as isolating the centering effect before selection enters.

This lower bound tends to 1 rapidly with D, since the failure probability is at most $\exp\big(-\frac{1}{2}D\,\mu(a)^2\big)$, i.e., it decays exponentially in the dimension at rate $\Theta(\mu(a)^2)$.

Hoeffding's bound can be conservative because it uses only the bounded range $Z_i \in [-1, 1]$ and ignores the actual variance of Z_1, which is typically much smaller than the worst-case variance implied by the range. Bernstein's inequality incorporates $\mathrm{Var}(Z_1)$ and yields a significantly tighter bound, as formalized below.

Theorem 2 (Bernstein bound: a tighter variance-aware guarantee)**.** *Let $Z_1 := (X_1 - T_1)^2 - (Y_1 - T_1)^2$, where $X_1, T_1 \sim \mathrm{Unif}[0,1]$ and $Y_1 \sim \mathrm{Unif}[\frac{1}{2} - a, \frac{1}{2} + a]$ are mutually independent, and define $\sigma^2(a) := \mathrm{Var}(Z_1)$. Then for all $D \in \mathbb{N}$,*

$$\mathbb{P}\big(\|Y-T\|_2 \le \|X-T\|_2\big) \ \ge\ 1 - \exp\left(-\frac{D\,\mu(a)^2}{2\sigma^2(a) + \frac{2\mu(a)}{3}}\right). \tag{17}$$

Proof. We apply Bernstein's inequality for sums of bounded random variables [44]. For i.i.d. random variables Z_i with $|Z_i| \le 1$, $\mathbb{E}[Z_i] = \mu$, and $\mathrm{Var}(Z_i) = \sigma^2$, Bernstein's inequality gives

$$\mathbb{P}\left(\sum_{i=1}^{D} Z_i - D\mu \le -t\right) \le \exp\left(-\frac{t^2}{2D\sigma^2 + \frac{2t}{3}}\right).$$

Setting $t = D\mu$ (so that $\sum_i Z_i - D\mu \le -D\mu$ is equivalent to $\Delta < 0$) gives

$$\mathbb{P}(\Delta < 0) \le \exp\left(-\frac{(D\mu)^2}{2D\sigma^2 + \frac{2D\mu}{3}}\right) = \exp\left(-\frac{D\mu^2}{2\sigma^2 + \frac{2\mu}{3}}\right).$$

□

A direct integration of the one-dimensional gap law gives

$$\mathbb{E}[Z_1^2] = \frac{144a^4 + 40a^2 + 29}{720}, \qquad \sigma^2(a) = \frac{(2a^2+1)(4a^2+3)}{90}. \tag{18}$$

Empirically, $\sigma^2(a)$ ranges from approximately 0.034 at $a = 0.1$ to 0.054 at $a = 0.4$, which is far below the worst-case variance of $1/3$ implied by the range $[-1, 1]$. This explains why the Bernstein bound is much tighter than Hoeffding at moderate D.

Heuristically, since $\Delta = \sum_{i=1}^{D} Z_i$ is a sum of i.i.d. terms, a CLT approximation suggests [45] $\mathbb{P}(\Delta < 0) \approx \Phi\big(-\sqrt{D}\,\mu(a)/\sigma(a)\big)$, so the effective z-score grows like $\sqrt{D}\,\mu(a)/\sigma(a)$. The CLT approximation closely matches empirical results in Fig. 2. Berry–Esseen bounds can formalize this with an explicit finite D error term, which we leave for future work.

Monte Carlo validation. We empirically estimate $p_D(a) = \mathbb{P}(\|Y-T\|_2 \le \|X-T\|_2)$ under the model of Section III-A using Monte Carlo simulation with 2×10^4 trials per (D, a). The empirical mean gap $\widehat{\mathbb{E}}[\Delta]$ agrees closely with the prediction $D \cdot \mu(a)$ from Lemma 1. For $(D, a) = (100, 0.1)$ the theory gives 8.0 and we observe ≈ 8.00, and for $(1000, 0.2)$ the theory gives 70.0 and we observe ≈ 69.96. We sweep $D \in \{2, 5, 10, 20, 50, 100, 200, 500, 1000, 10^4, 10^5\}$ and half-widths $a \in \{0.1, 0.2, 0.3, 0.4\}$, equivalently subcube side lengths $s = 2a \in \{0.2, 0.4, 0.6, 0.8\}$. Fig. 1 shows that the empirical probability $\hat{p}_D$ increases rapidly with the dimension D and approaches 1. Moreover, narrower central subcubes, with smaller side length $s = 2a$, yield faster convergence, consistent with the per-coordinate mean gap $\mu(a) = (1-4a^2)/12$. Table I summarizes the smallest tested dimension at which $\hat{p}_D$ exceeds 0.95, 0.99, and 0.999. At large dimensions (e.g. $D = 10^4$ and 10^5), we observe zero failures in 2×10^4 trials

TABLE I

EMPIRICAL DIMENSIONS REQUIRED TO REACH NEAR CERTAIN ADVANTAGE FOR CENTRAL SAMPLING. ENTRIES REPORT THE SMALLEST TESTED D SUCH THAT THE MONTE CARLO ESTIMATE $\hat{p}_D$ (EQ. (16)) EXCEEDS THE INDICATED THRESHOLD, USING 2×10^4 TRIALS PER (D, s). HERE s IS THE CENTRAL SUBCUBE SIDE LENGTH AND $a = s/2$ AND THE TESTED GRID IS $D \in \{2, 5, 10, 20, 50, 100, 200, 500, 1000, 10^4, 10^5\}$.

s	a	$\mu(a)$	D		
			$\hat{p}_D \geq 0.95$	$\hat{p}_D \geq 0.99$	$\hat{p}_D \geq 0.999$
0.2	0.1	0.08	20	50	50
0.4	0.2	0.07	20	50	100
0.6	0.3	0.0533	50	100	200
0.8	0.4	0.03	200	500	1000

across all tested s, so the empirical failure probability falls below the Monte Carlo resolution.

CLT approximation vs. Hoeffding and Bernstein bounds. Fig. 1 compares the CLT approximation and concentration bounds against the empirical probability. The CLT approximation $\mathbb{P}(\Delta \geq 0) \approx \Phi(\sqrt{D}\mu/\sigma)$ closely matches the empirical results across all tested dimensions and subcube widths. For instance, at $D = 50$ and $a = 0.1$ the CLT predicts 99.88% versus an empirical 99.95%, and at $D = 20$ and $a = 0.1$ it predicts 97.3% versus 97.8%. This demonstrates that the CLT approximation accurately captures the true probability. The Hoeffding lower bound (Theorem 1) is conservative at $D = 50$, $a = 0.1$, it guarantees only $p_D \geq 14.8\%$. The Bernstein bound (Theorem 2) is tighter, guaranteeing $p_D \geq 92.6\%$ for the same setting. The empirical results never violate either bound, validating Theorems 1 and 2. The remaining gap between the CLT approximation and the Bernstein bound arises because concentration inequalities provide worst-case guarantees rather than distributional estimates. Fig. 2 shows the same comparison on a logarithmic scale for $\log_{10}(1-\hat{p}_D)$. On this scale, the CLT closely tracks the empirical failure probability across all dimensions, while the Hoeffding and Bernstein lower bounds on p_D remain rigorous guarantees.

F. Benchmark validation on standard box-constrained test functions

To complement the single draw theory with a fixed budget search experiment, we compare the same budget random search on three standard box-constrained test functions, such as Sphere, Rastrigin, and Ackley. In each trial, we draw $N = 32$ candidates either from the full cube $[0,1]^D$ or from the central subcube of side length $s = 0.8$, evaluate all samples, and compare the best objective value produced by each sampler. Sphere uses $f(x) = \|x - c\|_2^2$, while Rastrigin and Ackley are evaluated in their standard shifted forms on linearly rescaled coordinates so that $x = c$ is the global minimizer inside $[0,1]^D$. We report the empirical win probability

$$p_{D,f}^{\text{best}} := \mathbb{P}\left(\min_{1 \leq j \leq N} f(Y^{(j)}) \leq \min_{1 \leq j \leq N} f(X^{(j)}) \right), \qquad (19)$$

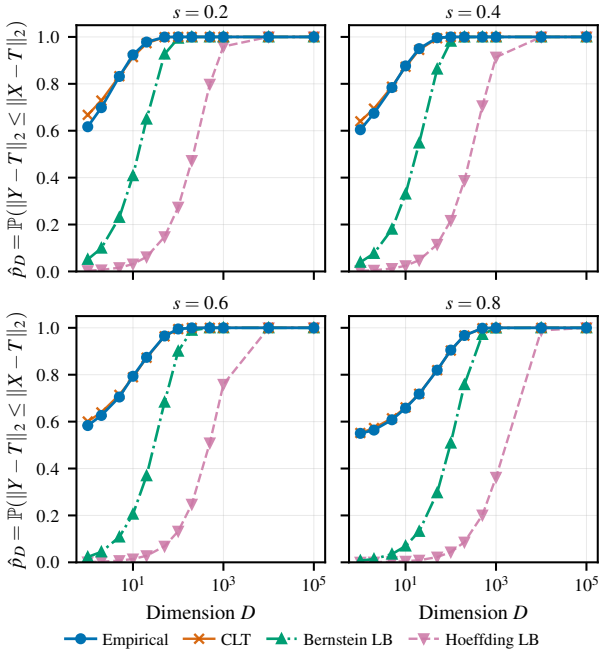

Fig. 1. **Advantage probability.** Monte Carlo validation of empirical $\hat{p}_D$ versus D for several subcube side lengths s (with $a = s/2$). Here $\hat{p}_D = \mathbb{P}(\|Y - T\|_2 \leq \|X - T\|_2)$ is the probability that the center draw is at least as close to the target as the uniform draw. Curves show $\hat{p}_D$, a CLT approximation, and Hoeffding and Bernstein lower bounds on p_D.

estimated from 5×10^3 trials. For each function, we test both a centered optimum ($c = 0.5$ in every coordinate) and a corner optimum ($c = 0$ in every coordinate).

Table II shows two consistent patterns. First, CBS is decisively better when the optimum is centered, and the advantage strengthens with dimension. Second, even when the optimum is shifted to a corner, CBS remains competitive and usually superior under this same budget random search protocol. This is consistent with the variance reduction mechanism identified above, which states that under squared distance box geometries, shrinking proposal variance can still outweigh location mismatch. These benchmark results are empirical rather than theorem level. They do not imply universal best of N guarantees for arbitrary objectives, but they do provide optimization function validation beyond the purely geometric Monte Carlo study.

IV. CENTER-BASED FINAL SELECTION UNDER MINI-BATCH EVALUATION NOISE

In population-based ML or noisy black-box optimization, the final deployment step can introduce its own bias. If several late stage candidates are close in true objective value and one selects the individual with the best score on a fresh mini-batch, the winner is often partly the candidate that received the most favorable noise realization rather than the one with the best full dataset objective. This optimistic selection effect is well known in noisy evolutionary settings [46].

When the final population is already concentrated near a good solution, returning a center-based aggregate, such as the

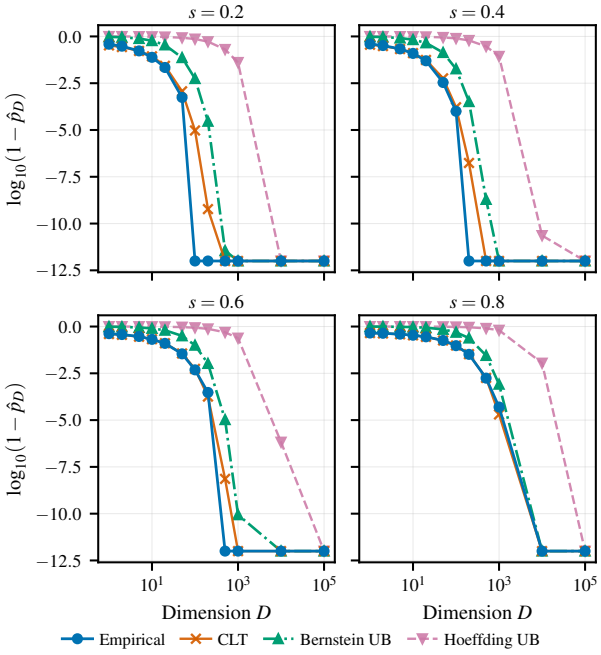

Fig. 2. **Failure probability.** Monte Carlo validation of $\log_{10}(1-\hat{p}_D)$ versus D (failure probability on log scale). Curves show empirical values, a CLT approximation, and Hoeffding and Bernstein upper bounds on $1-p_D$. When $\hat{p}_D = 1$, $1-\hat{p}_D$ is floored at 10^{-12} before $\log_{10}$.

TABLE II

SAME BUDGET RANDOM SEARCH BENCHMARK VALIDATION. EACH ENTRY IS THE EMPIRICAL PROBABILITY THAT THE BEST OF $N = 32$ CENTER SAMPLES FROM THE CENTRAL SUBCUBE WITH SIDE LENGTH $s = 0.8$ ATTAINS AN OBJECTIVE VALUE NO WORSE THAN THE BEST OF $N = 32$ UNIFORM CUBE SAMPLES, OVER 5×10^3 TRIALS. FUNCTIONS ARE SHIFTED SO THAT THE OPTIMUM IS EITHER CENTERED ($c = 0.5$ IN EVERY COORDINATE) OR AT THE CORNER ($c = 0$ IN EVERY COORDINATE).

Function	Optimum	$D = 20$	$D = 50$	$D = 100$
Sphere	center	0.981	1.000	1.000
Sphere	corner	0.544	0.708	0.853
Rastrigin	center	0.956	0.996	1.000
Rastrigin	corner	0.544	0.696	0.840
Ackley	center	0.977	1.000	1.000
Ackley	corner	0.533	0.705	0.836

population mean or a robust center, can be more stable than returning the single mini-batch winner. Under a local quadratic view of the objective, averaging shrinks dispersion around the local optimum and can therefore improve the expected full dataset objective. This is consistent with the broader role of averaging in learning, including Polyak–Ruppert averaging [47] and stochastic weight averaging [48], [49].

Practical recommendation. When the final population is tight and mini-batch noise is non-negligible, do not rely solely on the best single mini-batch score to choose the returned solution. A better default is to re-evaluate a center-based aggregate of the final population, or at least compare that aggregate against the best individual before deployment.

V. CONCLUSION

We studied a simple but informative model of high-dimensional randomized search in the unit hypercube and proved that sampling from a central region is typically closer, in expected squared distance, than sampling uniformly over the full domain. The key geometric mechanism is variance reduction at a shared center. Under the uniform-target cube model, this means advantage upgrades to an explicit exponential high-probability guarantee via Hoeffding and Bernstein inequalities.

An important clarification is that the expectation comparison is more general than the concentration theorem. Because the uniform and CBS samplers have the same mean, the expected squared distance advantage persists for any independent target with a finite second moment. The uniform target assumption is only needed for the clean i.i.d. bounded gap structure behind the explicit probability bounds.

We also connected this bounded domain phenomenon to centered distributions used throughout ML and optimization. Evolution strategies such as CMA-ES can be viewed as Gaussian, fitness weighted variants of CBS. Complementing the theory, our same budget benchmark study on shifted Sphere, Rastrigin, and Ackley functions shows that CBS also improves fixed budget random search on standard box-constrained landscapes, with the largest gains for centrally located optima.

Our future work includes formal best of N guarantees, extensions to non-product and non-box feasible sets, characterizing optimal central widths under richer prior information, and sharper finite D approximations via Berry–Esseen refinements or exact moment computations. Broader empirical studies on neural network training pipelines and benchmark suites with varying optimum locations would further clarify when center-based sampling yields the largest gains.

Code availability. The code for experiments is available at: https://github.com/rkhosrowshahi/cbs-theory.

REFERENCES

[1] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust Region Methods*, ser. MOS-SIAM Series on Optimization. Philadelphia, PA: SIAM, 2000.

[2] D. R. Jones, C. D. Perttunen, and B. E. Stuckman, "Lipschitzian optimization without the lipschitz constant," *Journal of Optimization Theory and Applications*, vol. 79, no. 1, pp. 157–181, 1993.

[3] R. Storn and K. Price, "Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[4] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks (ICNN)*. IEEE, 1995, pp. 1942–1948.

[5] N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evolutionary Computation*, vol. 9, no. 2, pp. 159–195, 2001.

[6] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.

[7] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, 2012.

[8] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, ser. Proceedings of Machine Learning Research, vol. 9. PMLR, 2010, pp. 249–256.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[11] S. Rahnamayan and G. G. Wang, "Center-based sampling for population-based algorithms," in *2009 IEEE Congress on Evolutionary Computation*, 2009, pp. 933–938.

[12] S. Esmailzadeh and S. Rahnamayan, "Enhanced differential evolution using center-based sampling," in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2011.

[13] H. Salehinejad and S. Rahnamayan, "Effects of centralized population initialization in differential evolution," in *Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2016.

[14] H. Xu, C. D. Erdbrink, and V. V. Krzhizhanovskaya, "How to speed up optimization? opposite-center learning and its application to differential evolution," *Procedia Computer Science*, vol. 51, pp. 805–814, 2015.

[15] H. R. Tizhoosh, "Opposition-based learning: A new scheme for machine intelligence," in *Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC)*, vol. 1. IEEE, 2005, pp. 695–701.

[16] J. Kudela, "A critical problem in benchmarking and analysis of evolutionary computation methods," *Nature Machine Intelligence*, vol. 4, no. 12, pp. 1238–1245, 2022.

[17] D. Wierstra, T. Schaul, T. Glasmachers, Y. Sun, J. Peters, and J. Schmidhuber, "Natural evolution strategies," *Journal of Machine Learning Research*, vol. 15, pp. 949–980, 2014.

[18] S. Rahnamayan, H. R. Tizhoosh, and M. M. Salama, "Opposition-based differential evolution," *IEEE Transactions on Evolutionary Computation*, vol. 12, no. 1, pp. 64–79, 2008.

[19] H. Hiba, S. Rahnamayan, A. A. Bidgoli, A. Ibrahim *et al.*, "A comprehensive investigation on novel center-based sampling for large-scale global optimization," *Swarm and Evolutionary Computation*, vol. 73, p. 101105, 2022.

[20] R. Khosrowshahli, S. Rahnamayan, A. Ibrahim, A. A. Bidgoli, and M. Makrehchi, "Population-level center-based sampling for meta-heuristic algorithms," *Swarm and Evolutionary Computation*, vol. 92, p. 101827, 2025.

[21] H. Mühlenbein and G. Paass, "From recombination of genes to the estimation of distributions I. binary parameters," in *Parallel Problem Solving from Nature – PPSN IV*, ser. Lecture Notes in Computer Science. Springer, 1996, vol. 1141, pp. 178–187.

[22] P. Larrañaga and J. A. Lozano, Eds., *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation*. Kluwer Academic Publishers, 2001.

[23] R. Y. Rubinstein and D. P. Kroese, *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning*. Springer, 2004.

[24] Y. Liu, Z. Qin, Z. Shi, and J. Lu, "Center particle swarm optimization," *Neurocomputing*, vol. 70, no. 4–6, pp. 672–679, 2007.

[25] S. J. Mousavirad and S. Rahnamayan, "CenPSO: A novel center-based particle swarm optimization algorithm for large-scale optimization," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2020, pp. 2066–2071.

[26] R. Pinnau, C. Totzeck, O. Tse, and S. Martin, "A consensus-based model for global optimization and its mean-field limit," *Mathematical Models and Methods in Applied Sciences*, vol. 27, no. 1, pp. 183–204, 2017.

[27] J. A. Carrillo, Y.-P. Choi, C. Totzeck, and O. Tse, "An analytical framework for a consensus-based global optimization method," *arXiv preprint arXiv:1602.00220*, 2018.

[28] M. Fornasier, H. Huang, L. Pareschi, and P. Sünnen, "Consensus-based optimization on the sphere: Convergence to global minimizers and machine learning," *Journal of Machine Learning Research*, vol. 22, no. 237, pp. 1–55, 2021.

[29] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[30] E. Zitzler, M. Laumanns, and L. Thiele, "SPEA2: Improving the strength pareto evolutionary algorithm," ETH Zurich, TIK – Computer Engineering and Networks Laboratory, Tech. Rep. 103, 2001.

[31] Q. Zhang and H. Li, "MOEA/D: A multiobjective evolutionary algorithm based on decomposition," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 6, pp. 712–731, 2007.

[32] K. Deb and H. Jain, "An evolutionary many-objective optimization algorithm using reference-point based nondominated sorting approach, part I: Solving problems with box constraints," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 4, pp. 577–601, 2014.

[33] H. Hiba, A. A. Bidgoli, A. Ibrahim, and S. Rahnamayan, "CGDE3: An efficient center-based algorithm for solving large-scale multi-objective optimization problems," in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2019.

[34] K. Deb, A. Anand, and D. Joshi, "A computationally efficient evolutionary algorithm for real-parameter optimization," *Evolutionary Computation*, vol. 10, no. 4, pp. 371–395, 2002.

[35] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. University of California Press, 1967, pp. 281–297.

[36] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[37] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[38] D. S. Broomhead and D. Lowe, "Multivariable functional interpolation and adaptive networks," *Complex Systems*, vol. 2, pp. 321–355, 1988.

[39] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, 2017.

[40] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[41] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[42] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient global optimization of expensive black-box functions," *Journal of Global Optimization*, vol. 13, no. 4, pp. 455–492, 1998.

[43] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.

[44] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[45] P. Billingsley, *Probability and Measure*, 3rd ed. Wiley, 1995.

[46] C. Bian, C. Qian, Y. Yu, and K. Tang, "On the robustness of median sampling in noisy evolutionary optimization," *Science China Information Sciences*, vol. 64, no. 5, p. 150103, 2021.

[47] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM Journal on Control and Optimization*, vol. 30, no. 4, pp. 838–855, 1992.

[48] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," in *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

[49] T. Garipov, P. Izmailov, D. Podoprikhin, D. Vetrov, and A. G. Wilson, "Loss surfaces, mode connectivity, and fast ensembling of DNNs," in *Advances in Neural Information Processing Systems*, 2018.