Variance Reduction for Better Sampling in Continuous Domains

Laurent Meunier, Carola Doerr, Jeremy Rapin, Olivier Teytaud

Variance Reduction for Better Sampling in Continuous Domains

Laurent Meunier[1,2], Carola Doerr[3], Jeremy Rapin[1], and Olivier Teytaud[1]

[1] Facebook Artificial Intelligence Research (FAIR), Paris, France
[2] PSL, Université Paris-Dauphine, Miles Team
[3] Sorbonne Université, CNRS, LIP6, Paris, France

Abstract. Design of experiments, random search, initialization of population-based methods, or sampling inside an epoch of an evolutionary algorithm use a sample drawn according to some probability distribution for approximating the location of an optimum. Recent papers have shown that the optimal *search* distribution, used for the sampling, might be more peaked around the center of the distribution than the *prior* distribution modelling our uncertainty about the location of the optimum. We confirm this statement, provide explicit values for this reshaping of the search distribution depending on the population size λ and the dimension d, and validate our results experimentally.

1 Introduction

We consider the setting in which one aims to locate an optimal solution $x^* \in \mathbb{R}^d$ for a given black-box problem $f : \mathbb{R}^d \to \mathbb{R}$ through a parallel evaluation of λ solution candidates. A simple, yet effective strategy for this *one-shot optimization* setting is to choose the λ candidates from a normal distribution $\mathcal{N}(\mu, \sigma^2)$, typically centered around an *a priori* estimate μ of the optimum and using a variance σ^2 that is calibrated according to the uncertainty with respect to the optimum. Random independent sampling is – despite its simplicity – still a very commonly used and performing good technique in one-shot optimization settings. There also exist more sophisticated sampling strategies like Latin Hypercube Sampling (LHS [20]), or quasi-random constructions such as Sobol, Halton, Hammersley sequences [7,18] – see [2,6] for examples. However, no general superiority of these strategies over random sampling can be observed when the benchmark set is sufficiently diverse [4]. It is therefore not surprising that in several one-shot settings – for example, the design of experiments [21,19,13,1] or the initialization (and sometimes also further iterations) of evolution strategies – the solution candidates are frequently sampled from random independent distributions (though sometimes improved by mirrored sampling [27]). A surprising finding was recently communicated in [6], where Cauwet et al. consider the setting in which the optimum x^* is known to be distributed according to a standard normal distribution $\mathcal{N}(0, I_d)$, and the goal is to minimize the distance of the best of the λ samples to this optimum. In the context of evolution strategies, one would formulate this problem as minimizing the sphere function

with normally distributed optimum. Intuitively, one might guess that sampling the λ candidates from the same prior distribution, $\mathcal{N}(0, I_d)$, should be optimal. This intuition, however, was disproved in [6], where it is shown that – unless the sample size λ grows exponentially fast in the dimension d – the median quality of sampling from $\mathcal{N}(0, I_d)$ is worse than that of sampling a single point, namely the center point 0. A similar observation was previously made in [22], without mathematically proven guarantees.

Our Theoretical Result. It was left open in [6] how to optimally scale the variance σ^2 when sampling the λ solution candidates from a normal distribution $\mathcal{N}(0, \sigma^2 I_d)$. While the result from [6] suggests to use $\sigma = 0$, we show in this work that a more effective strategy exists. More precisely, we show that setting $\sigma^2 = \min\{1, \Theta(\log(\lambda)/d)\}$ is asymptotically optimal, as long as λ is sub-exponential, but growing in d. Our variance scaling factor reduces the median approximation error by a $1 - \varepsilon$ factor, with $\varepsilon = \Theta(\log(\lambda)/d)$. We also prove that no constant variance nor any other variance scaling as $\omega(\log(\lambda)/d)$ can achieve such an approximation error. Note that several optimization algorithms operate with rescaled sampling. Our theoretical results therefore set the mathematical foundation for empirical rules of thumb such as, for example, used in e.g. [22,9,17,10,8,28,6].

Our Empirical Results. We complement our theoretical analyses by an empirical investigation of the rescaled sampling strategy. Experiments on the sphere function confirm the results. We also show that our scaling factor for the variance yields excellent performance on two other benchmark problems, the Cigar and the Rastrigin function. Finally, we demonstrate that these improvements are not restricted to the one-shot setting, but extend to iterative optimization strategies. More precisely, we show a positive impact on the initialization of Bayesian optimization algorithms [15] and on differential evolution [25].

Related Work. While the most relevant works for our study have been mentioned above, we briefly note that a similar surprising effect as observed here is the "Stein phenomenon" [24,14]. Although an intuitive way to estimate the mean of a standard gaussian distribution is to compute the empirical mean, Stein showed that this strategy is sub-optimal w.r.t. mean squared error and that the empirical mean needs to be rescaled by some factor to be optimal.

2 Problem Statement and Related Work

The context of our theoretical analysis is *one-shot optimization*. In one-shot optimization, we are allowed to select λ points $x_1, \ldots, x_\lambda \in \mathbb{R}^d$. The quality $f(x_i)$ of these points is evaluated, and we measure the performance of our samples in terms of simple regret [5] $\min_{i=1,\ldots,\lambda} f(x_i) - \inf_{x \in \mathbb{R}^d} f(x)$.[1] That is, we aim to

[1] This requires knowledge of $\inf_x f(x)$, which may not be available in real-world applications. In this case, the infimum can be replaced by an empirical minimum. In all applications considered in this work the value of $\inf_x f(x)$ is known.

minimize the distance – measured in *quality space* – of the best of our points to the optimum. This formulation, however, also covers the case in which we aim to minimize the distance to the optimum in the *search space*: we simply take as f the root of the sphere function $f_{x^*} : \mathbb{R}^d \to \mathbb{R}, x \mapsto ||x - x^*||^2$, where here and in the following $||.||$ denotes the Euclidean norm.

Rescaled Random Sampling for Randomly Placed Optimum. In the setting studied in Sec. 3 we assume that the optimum x^* is sampled from the standard multivariate Gaussian distribution $\mathcal{N}(0, I_d)$, and that we aim to minimize the regret $\min_{i=1,\dots,\lambda} ||x_i - x^*||^2$ through i.i.d. samples $x_i \sim \mathcal{N}(0, \sigma^2 I_d)$. That is, in contrast to the classical *design of experiments* (DoE) setting, we are only allowed to choose the scaling factor σ, whereas in DoE more sophisticated (often quasi-random and space-filling designs – which are typically not i.i.d. samples) are admissible. Intuitively, one might be tempted to guess that $\sigma = 1$ should be a good choice, as in this case the λ points are chosen from the same distribution as the optimum x^*. This intuition, however, was refuted in [6, Theorem 1], where is was shown that the middle point sampling strategy, which uses $\sigma = 0$ (i.e., all λ points collapse to $(0, \dots, 0)$) yields smaller regret than sampling from $\mathcal{N}(0, I_d)$ unless λ grows exponentially in d. More precisely, it is shown in [6] that, for this regime of λ and d, the median of $||x^*||^2$ is smaller than the median of $||x_i - x^*||^2$ for i.i.d. $x_i \in \mathcal{N}(0, I_d)$. This shows that sampling a single point can be better than sampling λ points with the wrong scaling factor, unless the budget λ is very large.
Our goal is to improve upon the middle point strategy, by deriving a scaling factor σ such that the λ i.i.d. samples yield smaller regret with a decent probability. More precisely, we aim at identifying σ such that

$$\mathbb{P}\left[\min_{1 \le i \le \lambda} ||x_i - x^*||^2 \le (1 - \varepsilon)||x^*||^2\right] \ge \delta, \tag{1}$$

for some $\delta \ge 1/2$ and $\varepsilon > 0$ as large as possible. Here, in line with [6], we have switched to regret, for convenience of notation. [6] proposed, without proof, such a scaling factor: our proposal is dramatically better in some regimes.

3 Theoretical Results

We derive sufficient and necessary conditions on the scaling factor σ such that Eq. (1) can be satisfied. More precisely, we prove that Eq. (1) holds with approximation gain $\varepsilon \approx \log(\lambda)/d$ when the variance σ^2 is chosen proportionally to $\log \lambda/d$ (and λ does not grow too rapidly in d). We then show that Eq. (1) cannot be satisfied for $\sigma^2 = \omega(\log(\lambda)/d)$. Moreover, we prove that $\varepsilon = O(\log(\lambda)/d)$, which, together with the first result, shows that our scaling factor is asymptotically optimal. The precise statements are summarized in Theorems 1, 2, and 3, respectively. Proof sketches are available in Sec. 3.1. Full proofs are left in the appendix.

Theorem 1. *(Sufficient condition on rescaling) Let* $\delta \in [\frac{1}{2}, 1)$*. Let* $\lambda = \lambda_d$*, satisfying* $\lambda_d \to \infty$ *as* $d \to \infty$ *and* $\log(\lambda_d) \in o(d)$ *(2). Then there exist two positive constants* c_1*,* c_2*, and* d_0*, such that for all* $d \geq d_0$ *it holds that* $\mathbb{P}\left[\min_{i=1,\dots,\lambda} ||x^* - x_i||^2 \leq (1-\varepsilon)\,||x^*||^2\right] \geq \delta$ *(3) when* x^* *is sampled from the standard Gaussian distribution* $\mathcal{N}(0, I_d)$*,* $x_1, \dots, x_\lambda$ *are independently sampled from* $\mathcal{N}(0, \sigma^2 I_d)$ *with* $\sigma^2 = \sigma_d^2 = c_2 \log(\lambda)/d$ *and* $\varepsilon = \varepsilon_d = c_1 \log(\lambda)/d$*.*

Theorem 1 shows that i.i.d. Gaussian sampling can outperform the middle-point strategy derived in [6] (i.e., the strategy using $\sigma^2 = 0$) if the scaling factor σ is chosen appropriately. Our next theorem summarizes our findings for the conditions that are *necessary* for the scaling factor σ^2 to outperform this middle-point strategy. This result, in particular, illustrates why neither the natural choice $\sigma = 1$, nor any other constant scaling factor can be optimal.

Theorem 2. *(Necessary condition on rescaling) Consider* $\lambda = \lambda_d$ *satisfying assumptions (2). There exists an absolute constant* $C > 0$ *such that for all* $\delta \in [\frac{1}{2}, 1)$*, there exists* $d_0 > 0$ *such that, for all* $d > d_0$ *and for all* σ *the property* $\exists \varepsilon > 0, \mathbb{P}\left[\min_{i=1,\dots,\lambda} ||x^* - x_i||^2 \leq (1-\varepsilon)\,||x^*||^2\right] \geq \delta$ *(4) for* $x^* \sim \mathcal{N}(0, I_d)$ *and* $x_1, \dots, x_\lambda$ *independently sampled from* $\mathcal{N}(0, \sigma^2 I_d)$*, implies that* $\sigma^2 \leq C \log(\lambda)/d$*.*

While Theorem 2 induces a necessary condition on the scaling factor σ to improve over the middle-point strategy, it does not bound the gain that one can achieve through a proper scaling. Our next theorem shows that the factor derived in Theorem 1 is asymptotically optimal.

Theorem 3. *(Upper bound for the approximation factor) Consider* $\lambda = \lambda_d$ *satisfying assumptions (2). There exists an absolute constant* $C' > 0$ *such that for all* $\delta \in [\frac{1}{2}, 1)$*, there exists* $d_0 > 0$ *such that, for all* $d > d_0$ *and for all* $\varepsilon, \sigma > 0$*, it holds that if* $\mathbb{P}\left[\min_{i=1,\dots,\lambda} ||x^* - x_i||^2 \leq (1-\varepsilon)\,||x^*||^2\right] \geq \delta$ *for* $x^* \sim \mathcal{N}(0, I_d)$ *and* $x_1, \dots, x_\lambda$ *independently sampled from* $\mathcal{N}(0, \sigma^2 I_d)$*, then* $\varepsilon \leq C' \log(\lambda)/d$*.*

3.1 Proof Sketches

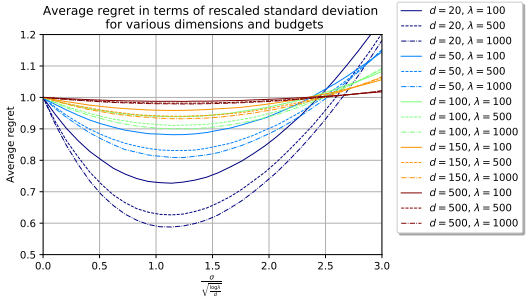
We first notice that as x^* is sampled from a standard normal distribution $\mathcal{N}(0, I_d)$, its norm satisfies $||x^*||^2 = d + o(d)$ as $d \to \infty$. We then use that, conditionally to x^*, it holds that

$$\mathbb{P}\left[\min_{i\in[\lambda]} ||x^* - x_i||^2 \leq (1-\varepsilon)\,||x^*||^2 \middle| x^*\right] = 1 - \left(1 - \mathbb{P}\left[||x - x^*||^2 \leq (1-\varepsilon)\,||x^*||^2 \middle| x^*\right]\right)^\lambda$$

We therefore investigate when the condition

$$\mathbb{P}\left[||x - x^*||^2 \leq (1-\varepsilon)\,||x^*||^2 \middle| x^*\right] > 1 - (1-\delta)^{\frac{1}{\lambda}} \tag{5}$$

is satisfied. To this end, we make use of the fact that the squared distance $||x^*||^2$ of x^* to the middle point 0 follows the central $\chi^2(d)$ distribution, whereas, for a given point $x^* \in \mathbb{R}^d$, the distribution of the squared distance $||x - x^*||^2/\sigma^2$ for $x \sim \mathcal{N}(0, \sigma^2 I_d)$ follows the non-central $\chi^2(d, \mu)$ distribution with non-centrality parameter $\mu := ||x^*||^2/\sigma^2$. Using the concentration inequalities provided in [29, Theorem 7] for non-central χ^2 distributions, we then derive sufficient and necessary conditions for condition (5) to hold. With this, and using assumptions (2), we are able to derive the results from Theorems 1, 2, and 3.



d	λ	σ^*	$\sigma = 1$
20	100	**0.73**	0.88
	500	**0.63**	0.72
	1000	**0.59**	0.66
50	100	**0.89**	1.23
	500	**0.83**	1.10
	1000	**0.81**	1.05
100	100	**0.94**	1.44
	500	**0.91**	1.33
	1000	**0.90**	1.29
150	100	**0.96**	1.53
	500	**0.94**	1.44
	1000	**0.93**	1.41
500	100	**0.99**	1.74
	500	**0.98**	1.68
	1000	**0.98**	1.66

Fig. 1. Average regret, normalized by d, on the sphere function for various dimensions and budgets in terms of rescaled standard deviation. Each mean has been estimated from $100,000$ samples. Table on the right: Average regret for $\sigma^* = \sqrt{\log(\lambda)/d}$ and $\sigma = 1$.

4 Experimental Performance Comparisons

The theoretical results presented above are in asymptotic terms, and do not specify the constants. We therefore complement our mathematical investigation with an empirical analysis of the rescaling factor. Whereas results for the setting studied in Sec. 3 are presented in Sec. 4.1, we show in Sec. 4.2 that the advantage of our rescaling factor is not limited to minimizing the distance in search space. More precisely, we show that the rescaled sampling achieves good results also in a classical DoE task, in which we aim for minimizing the regret for the Cigar and for the Rastrigin functions. Finally, we investigate in Sec. 4.3 the impact of initializing two common optimization heuristics, Bayesian Optimization (BO) and differential evolution (DE), by a population sampled from the Gaussian distribution $\mathcal{N}(0, \sigma^2 I_d)$ using our rescaling factor $\sigma = \sqrt{\log(\lambda)/d}$.

4.1 Validation of Our Theoretical Results on the Sphere Function

Fig. 1 displays the normalized average regret $\frac{1}{d}\mathbb{E}\left[\min_{i=1,\dots,\lambda} ||x^* - x_i||^2\right]$ in terms of $\sigma/\sqrt{\log(\lambda)/d}$ for different dimensions and budgets. We observe that the best parametrization of σ is around $\sqrt{\log(\lambda)/d}$ in all displayed cases. Moreover, we also see that – as expected – the gain of the rescaled sampling over the midpoint sampling ($\sigma = 0$) goes to 0 as $d \to \infty$. We also see that, for the regimes plotted in Fig. 1, the advantage of the rescaled variance grows with the budget λ. Figure 2 (on left) displays the average regret as a function of increasing values of λ for the different rescaling methods ($\sigma \in \{0, \sqrt{\log \lambda/d}, 1\}$). We remark, unsurprisingly, that the gain of rescaling is diminishing as $\lambda \to \infty$. Finally, Figure 2 (on right) shows the distribution of regrets for the different rescaling methods. The improvement of the expected regret is not at the expense of a higher dispersion of the regret.

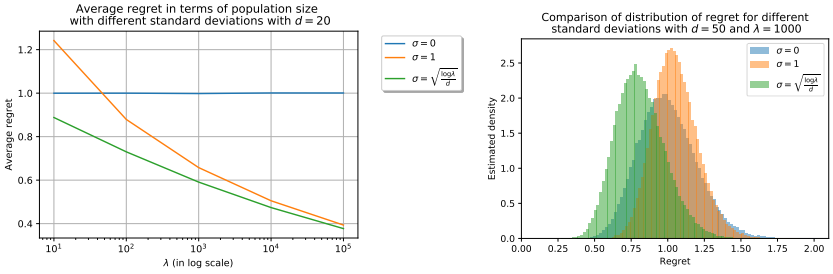


Fig. 2. Comparison of methods: without rescaling ($\sigma = 1$), midpoint sampling ($\sigma = 0$), and our rescaling method ($\sigma = \sqrt{\frac{\log \lambda}{d}}$). Each mean has been estimated from 10^5 samples. (On left) Average regret, normalized by d, on the sphere function for diverse population sizes λ at fixed dimension $d = 20$. The gain of rescaling decreases as λ increases. (On right) Distribution of the regret for the strategies on the $50d$-sphere function for $\lambda = 1000$.

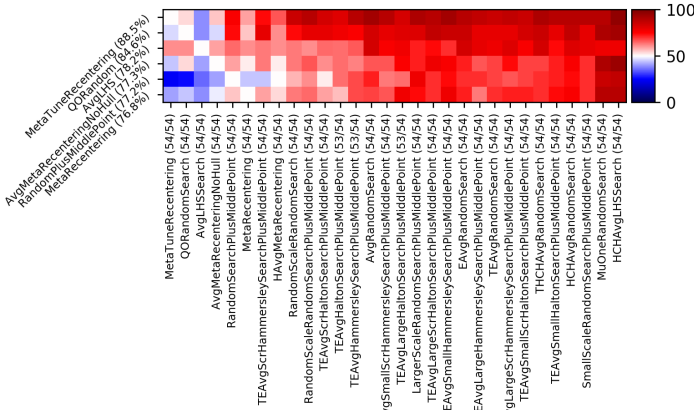
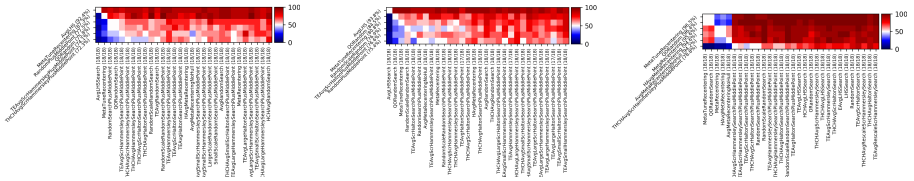


Fig. 3. Comparison of various one-shot optimization methods from the point of view of the simple regret. Reading guide in Sec. 4.2. Results are averaged over objective functions Cigar, Rastrigin, Sphere in dimension 20, 200, 2000, and budget 30, 100, 3000, 10000, 30000, 100000. `MetaTuneRecentering` performs best overall. Only the 30 best performing methods are displayed.

4.2 Comparison with the DoEs Available in Nevergrad

Motivated by the significant improvements presented above, we now investigate whether the advantage of our rescaling factor translates to other optimization tasks. To this end, we first analyze a DoE setting, in which an underlying (and typically not explicitly given) function f is to be minimized through a parallel evaluation of λ solution candidates $x_1, \ldots, x_\lambda$, and regret is measured in terms of $\min_i f(x_i) - \inf_x f(x)$. In the broader machine learning literature, and in



Sphere function Cigar function Rastrigin function

Fig. 4. Same experiment as Fig. 3, but separately over each objective function. Results are still averaged over 6 distinct budgets (30, 100, 3000, 10000, 30000, 100000) and 3 distinct dimensionalities (20, 200, 2000). `MetaTuneRecentering` performs well in each case, and is not limited to the sphere function for which it was derived. Variants of LHS are sometimes excellent and sometimes not visible at all (only the 30 best performing methods are shown).

particular in the context of hyper-parameter optimization, this setting is often referred to as *one-shot optimization* [2,6].

Experimental Setup. All our experiments are implemented and freely available in the Nevergrad platform [23]. Results are presented as shown in Fig. 3. Typically, the six best methods are displayed as rows. The 30 best performing methods are presented as columns. The order for rows and for columns is the same: algorithms are ranked by their average winning frequency, measured against all other algorithms in the portfolio. The heatmaps show the fraction of runs in which algorithm x (row) outperformed algorithm y (column), averaged over all settings and all replicas (i.e. random repetitions). The settings are typically sweepings over various budgets, dimensions, and objective functions.[2] The numbers in the captions of the columns indicate the number of settings for which the algorithms are compared against each other. That is, a bracket "(6/6)" is to be read as "the winning frequencies are averaged over all six out of a total number of six settings". For each tested (algorithm, problem) pair 20 independent runs are performed: a (6/6) case is thus based on a total number of 120 runs.

Algorithm Portfolio. Several rescaling methods are already available on Nevergrad. A large fraction of these have been implemented by the authors of [6]; in particular:

- The replacement of one sample by the center. These methods are named "`midpointX`" or "`XPlusMiddlePoint`", where `X` is the original method that has been modified that way.
- The rescaling factor `MetaRecentering` empirically derived in [6]: $\sigma = \frac{1+\log(\lambda)}{4\log(d)}$.

[2] Detailed results for individual settings are available at http://dl.fbaipublicfiles.com/nevergrad/allxps/list.html.

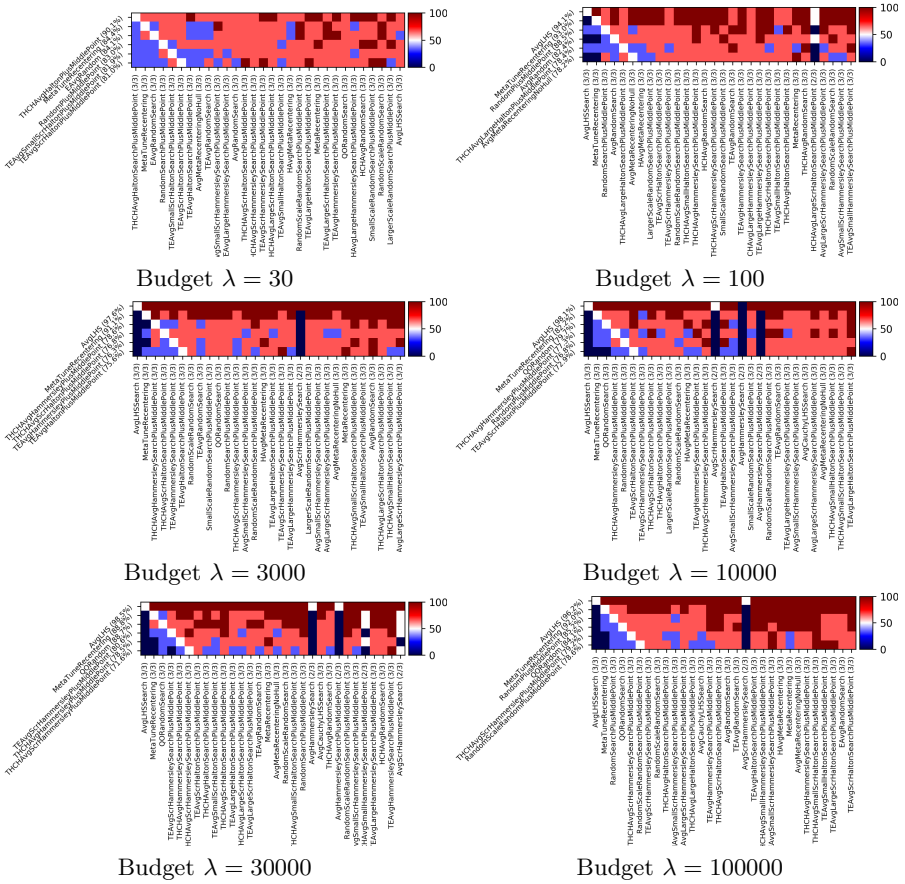
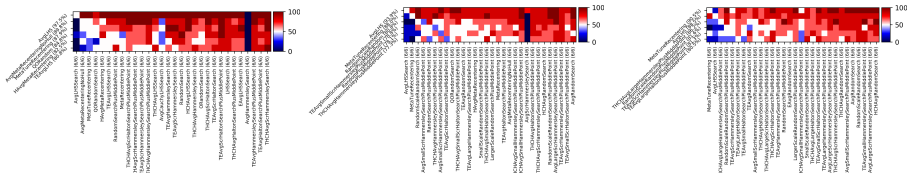


Fig. 5. Results on the sphere function, per budget. Results are averaged over dimension 20, 200, 2000. Our method `MetaTuneRecentering` performs among the best in all cases. LHS is excellent on this very simple setting, namely the sphere function.

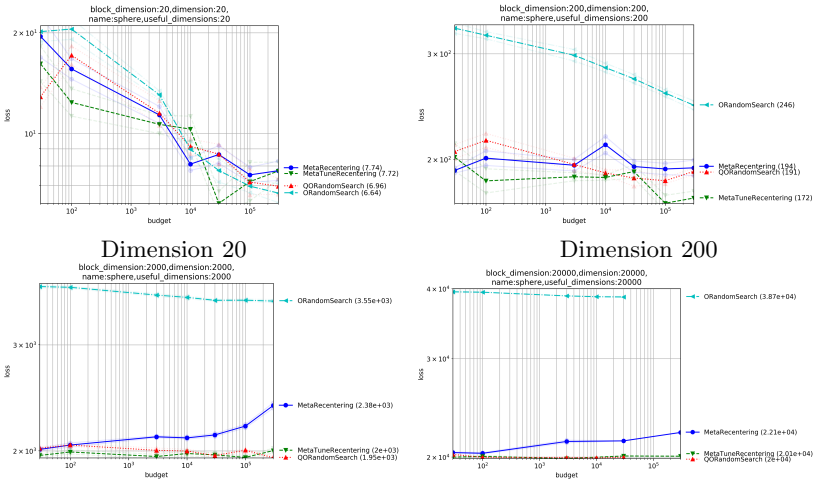
- The quasi-opposite methods suggested in [22], with prefix "`QO`": when x is sampled, then another sample $c - rx$ is added, with r uniformly drawn in $[0, 1]$ and c the center of the distribution.

We also include in our comparison a different type of one-shot optimization techniques, independent of the present work, currently available in the platform: they use the information obtained from the sampled points to recommend a point x that is not necessarily one of the λ evaluated ones. These *"one-shot+1"* strategies have the prefix "`Avg`". We keep all these and all other sampling strategies available in Nevergrad for our experiments. We add to this existing Nevergrad portfolio our own rescaling strategy, which uses the scaling factor derived in Sec. 3; i.e., $\sigma = \sqrt{\log(\lambda)/d}$. We refer to this sampling strategy as `MetaTuneRecentering`, defined below. Both scaling factors `MetaRecentering` [6]



Dimension 20 Dimension 200 Dimension 2000

Fig. 6. Results on the sphere function, per dimensionality. Results are still averaged over 6 values of the budget, namely 30, 100, 3000, 10000, 30000, 100000. Our method becomes better and better as the dimension increases.



Dimension 20 Dimension 200

Dimension 2000 Dimension 20000

Fig. 7. Same context as Fig. 6, with x-axis = budget and y-axis = average simple regret. We see the failure of `MetaRecentering` in the worsening performance as budget goes to infinity: the budget has an impact on σ which becomes worse, hence worse overall performance. We note that quasi-opposite sampling can perform decently in a wide range of values. Opposite Sampling is not much better than random search in high-dimension. Our `MetaTuneRecentering` shows decent performance: in particular, simple regret decreases as $\lambda \to \infty$.

and `MetaTuneRecentering` (our equations) are applied to quasirandom sampling (more precisely, scrambled Hammersley [13,1]) rather than random sampling. We provide detailed specifications of these methods and the most important ones below, whereas we skip the dozens of other methods: they are open sourced in Nevergrad [23].

From $[0,1]^d$ *to Gaussian quasi-random, random or LHS sampling:* Random sampling, quasi-random sampling, Latin Hypercube Sampling (or others) have a well known definition in $[0,1]^d$ (for quasi-random, see Halton [12] or Hammersley [13], possibly boosted by scrambling [1]; for LHS, see [19]). To extend to multidimen-

sional Gaussian sampling, we use that if U is a uniform random variable on $[0,1]$ and Φ the standard Gaussian CDF, then $\Phi^{-1}(U)$ simulates a $\mathcal{N}(0,1)$ distribution. We do so on each dimension: this provides a Gaussian quasi-random, random or LHS sampling.

Then, one can rescale the Gaussian quasi-random sampling with the corresponding factor σ for `MetaRecentering` ($\sigma = \frac{1+\log(\lambda)}{4\log(d)}$ [6]) and `MetaTuneRecentering` ($\sigma = \sqrt{\log(\lambda)/d}$): for $i \leq \lambda$ and $j \leq d$, $x_{i,j} = \sigma\phi^{-1}(h_{i,j})$ where $h_{i,j}$ is the j^{th} coordinate of a i^{th} Scrambled-Hammersley point.

Results for the Full DoE Testbed in Nevergrad. Fig. 3 displays aggregated results for the Sphere, the Cigar, and the Rastrigin functions, for three different dimensions and six different budgets. We observe that our `MetaTuneRecentering` strategy performs best, with a winning frequency of 80%. It positively compares against all other strategies from the portfolio, with the notable exception of `AvgLHS`, which, in fact, compares favorably against every single other strategy, but with a lower average winning frequency of 73.6%. Note here that `AvgLHS` is one of the **"oneshot+1"** strategies, i.e., it has not only one more sample, but it is also allowed to sample its recommendation adaptively, in contrast to our fully parallel `MetaTuneRecentering` strategy. It performs poorly in some cases (Rastrigin) and does not make sense as an initialization (Sect. 4.3).

Selected DoE Tasks. Figs. 4 breaks down the aggregated results from Fig. 3 by the three different functions. From this figure we see that `MetaTuneRecentering` scores second on sphere (where `AvgLHS` is winning) , third on Cigar (after `AvgLHS` and `QORandom`), and first on Rastrigin. This fine performance is quite remarkable, given that the portfolio contains quite sophisticated and highly tuned methods. In addition, the `AvgLHS` methods, sometimes performing better on the sphere, besides using more capabilities than we do as it is a "oneshot+1" method, had poor results for Rastrigin (not even in the 30 best methods). On sphere, the difference to the third and following strategies is significant (87.3% winning rate against 77.5% for the next runner-up). On Cigar, the differences between the first four strategies are greater than 4 percentage points each, whereas on Rastrigin the average winning frequencies of the first five strategies is comparable, but significantly larger than that of the sixth one (which scores 78.8% against >94.2% for the first five DoEs). Fig. 5 zooms into the results for the sphere function, and breaks them further down by available budget λ (note that the results are still averaged over the three dimensions 20, 200, 2000). `MetaTuneRecentering` scores second in all six cases. A breakdown of the results for sphere by dimension (and aggregated over the six available budgets) is provided in Fig. 6 and Fig. 7. For dimension 20, we see that `MetaTuneRecentering` ranks third, but, interestingly, the two first methods are "oneshot+1" style (`Avg` prefix). In dimension 200, `MetaTuneRecentering` ranks second, with considerable advantage over the third-ranked strategy (88.0% vs. 80.8%). Finally, for the largest tested dimension, $d = 2000$, our method ranks first, with an average winning frequency of 90.5%.

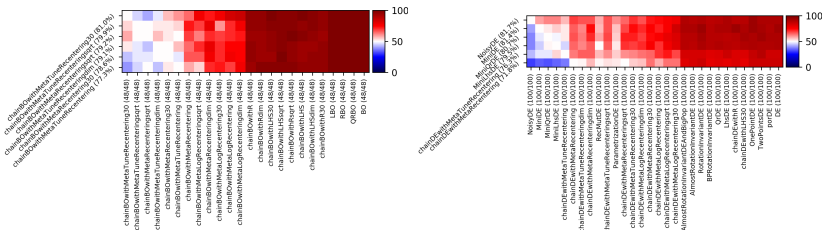


Fig. 8. Performance comparison of different strategies to initialize Bayesian Optimization (BO, left) and Differential Evolution (DE, right). A detailed description is given in Sec. 4.3. `MetaTuneRecentering` performs best as an initialization method. In the case of DE, methods different from the traditional DE remain the best on this testcase: when we compare DE with a given initialization and DE initialized with `MetaTuneRecentering`, `MetaTuneRecentering` performs best in almost all cases.

4.3 Rescaled Sampling for Better Initialization of Iterative Optimization Heuristics

We now move from the one-shot settings considered thus far to *iterative optimization*, and show that our scaling factor can also be beneficial in this context. More precisely, we analyze the impact of initializing efficient global optimization (EGO [15], a special case of Bayesian optimization) and differential evolution (DE [25]) by a population that is sampled from a distribution that uses our variance scaling scheme. It is well known that a proper initialization can be very critical for the performance of these solvers; see [11,26,22,16,3] for discussions. Fig. 8 summarizes the results of our experiments. As in the previous setups, we compare against existing methods from the Nevergrad platform, to which we have just added our rescaling factor termed `MetaTuneRecentering`. For each initialization scheme, four different initial population sizes are considered: denoting by d the dimension, by w the parallelism (i.e., the number of workers), and by b the total budget that the algorithms can spend on optimizing the given optimization task, the initial population λ is set as $\lambda = \sqrt{b}$ for `Sqrt`, as $\lambda = d$ for `Dim`, $\lambda = w$ for no suffix, and as $\lambda = 30$ when the suffix is `30`. As in Sec. 4.2 we superpose our scaling scheme on top of the quasi-random Scrambled Hammersley sequence suggested in [6], but we also consider random initialization rather than quasi-random (indicated by the suffix "`R`") and Latin Hypercube Sampling [19] (suffix "`LHS`"). The left chart in Fig. 8 is for the Bayesian optimization case. It aggregates results for 48 settings, which stem from Nevergrad's "parahdbo4d" suite. It comprises the four benchmark problems Sphere, Cigar, Ellipsoid and Hm. Results are averaged over the total budgets $b \in \{25, 31, 37, 43, 50, 60\}$, dimension $d \in \{20, 2000\}$, and parallelism $w = \max(d, \lfloor b/6 \rfloor)$. The parallelism is 20. We observe that a BO version using our `MetaTuneRecentering` performs best, and that several other variants using this scaling appear among the top-performing configurations. The chart on the right of Fig. 8 summarizes results for Differential Evolution. Since DE can handle larger budgets, we

consider here a total number of 100 settings, which correspond to the testcase named "paraalldes" in Nevergrad. In this suite, results are averaged over budgets $b \in \{10, 100, 1000, 10000, 100000\}$, dimensions $d \in \{5, 20, 100, 500, 2500\}$, parallelism $w = \max(d, \lfloor b/6 \rfloor)$, and again the objective functions Sphere, Cigar, Ellipsoid, and Hm. The parallelism is 20. Specialized versions of DE perform best for this testcase, but we see that DE initialized with our `MetaTuneRecentering` strategy ranks fifth (outperformed only by ad hoc variants of DE), with an overall winning frequency that is not much smaller than that of the top-ranked `NoisyDE` strategy (76.3% for `ChainDEwithMetaTuneRecentering` vs. 81.7% for `NoisyDE`) - and almost always outperforms the rescaling used in the original Nevergrad.

5 Conclusions and Future Work

We have investigated the scaling of the variance of random sampling in order to minimize the expected regret. While previous work [6] had already shown that the optimal scaling factor is not identical to that of the prior distribution from which the optimum is sampled (unless the sample size is exponentially large in the dimension), it did not answer the question how to scale the variance optimally. In this work, we have proven that standard deviations scaled as $\sigma = \sqrt{\log(\lambda)/d}$ gives, with probability at least $1/2$, a sample that is significantly closer to the optimum than the previous known strategies. We have also shown that the gain achieved by our rescaled sampling strategy is asymptotically optimal. Moreover, we have shown that any decent scaling factor is asymptotically at most as large as our proposed one. The empirical assessment of our rescaled sampling strategy confirms decent performance not only on the sphere function, but also on other classical benchmark problems. We have furthermore given indication that the sampling might help improve state-of-the-art numerical heuristics based on differential evolution or using Bayesian surrogate models. Our proposed one-shot method performs best in many cases, sometimes outperformed by e.g. `AvgLHS`, but is stable on a wide range of problems and meaningful also as an initialization method (as opposed to `AvgLHS`). Whereas our theoretical results can be extended to quadratic forms (by conservation of barycenters through linear transformations), an extension to wider families of functions (e.g., families of functions with order 2 Taylor expansion) is not straightforward. Apart from extending our results to broader function classes, another direction for future work comprises extensions to the multi-epoch case. Our empirical results on DE and BO gives first indication that a properly scaled variance can also be beneficial in iterative sampling. Note, however, that in the latter case, we only adjusted the initialization, not the later sampling steps. This forms another promising direction for future work.

Acknowledgements. This work was initiated at Dagstuhl seminar 19431 on Theory of Randomized Optimization Heuristics.

References

1. Atanassov, E.I.: On the discrepancy of the Halton sequences. Math. Balkanica (NS) **18**(1-2), 15–32 (2004)
2. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. J. Mach. Learn. Res. **13**, 281–305 (2012)
3. Bossek, J., Doerr, C., Kerschke, P.: Initial design strategies and their effects on sequential model-based optimization. In: Proc. of the Genetic and Evolutionary Computation Conference (GECCO'20). ACM (2020), to appear. Available at https://arxiv.org/abs/2003.13826
4. Bossek, J., Kerschke, P., Neumann, A., Neumann, F., Doerr, C.: One-shot decision-making with and without surrogates. CoRR **abs/1912.08956** (2019), http://arxiv.org/abs/1912.08956
5. Bubeck, S., Munos, R., Stoltz, G.: Pure exploration in multi-armed bandits problems. In: International conference on Algorithmic learning theory. pp. 23–37. Springer (2009)
6. Cauwet, M.L., Couprie, C., Dehos, J., Luc, P., Rapin, J., Riviere, M., Teytaud, F., Teytaud, O.: Fully parallel hyperparameter search: Reshaped space-filling. arXiv preprint arXiv:1910.08406 (2019)
7. Dick, J., Pillichshammer, F.: Digital Nets and Sequences. Cambridge University Press (2010)
8. Ergezer, M., Sikder, I.: Survey of oppositional algorithms. In: 14th International Conference on Computer and Information Technology (ICCIT 2011). pp. 623–628 (2011)
9. Esmailzadeh, A., Rahnamayan, S.: Enhanced differential evolution using center-based sampling. In: 2011 IEEE Congress of Evolutionary Computation (CEC). pp. 2641–2648 (2011)
10. Esmailzadeh, A., Rahnamayan, S.: Center-point-based simulated annealing. In: 2012 25th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE). pp. 1–4 (2012)
11. Feurer, M., Springenberg, J.T., Hutter, F.: Initializing Bayesian hyperparameter optimization via meta-learning. In: AAAI (2015)
12. Halton, J.: On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. Numerische Mathematik **2**, 84–90 (1960), http://eudml.org/doc/131448
13. Hammersley, J.M.: Monte-carlo methods for solving multivariate problems. Annals of the New York Academy of Sciences **86**(3), 844–874 (1960)
14. James, W., Stein, C.: Estimation with quadratic loss. In: Proc. of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. pp. 361–379. University of California Press (1961), https://projecteuclid.org/euclid.bsmsp/1200512173
15. Jones, D.R., Schonlau, M., Welch, W.J.: Efficient global optimization of expensive black-box functions. Journal of Global Optimization **13**(4), 455–492 (Dec 1998)
16. Maaranen, H., Miettinen, K., Mäkelä, M.: Quasi-random initial population for genetic algorithms. Computers and Mathematics with Applications **47**(12), 1885–1895 (2004)
17. Mahdavi, S., Rahnamayan, S., Deb, K.: Center-based initialization of cooperative co-evolutionary algorithm for large-scale optimization. In: 2016 IEEE Congress on Evolutionary Computation (CEC). pp. 3557–3565 (2016)
18. Matoušek, J.: Geometric Discrepancy. Springer, 2nd edn. (2010)

19. McKay, M.D., Beckman, R.J., Conover, W.J.: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics **21**(2), 239–245 (1979)
20. McKay, M.D., Beckman, R.J., Conover, W.J.: A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. Technometrics **21**, 239–245 (1979)
21. Niederreiter, H.: Random Number Generation and quasi-Monte Carlo Methods. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (1992)
22. Rahnamayan, S., Wang, G.G.: Center-based sampling for population-based algorithms. In: 2009 IEEE Congress on Evolutionary Computation. pp. 933–938 (May 2009). https://doi.org/10.1109/CEC.2009.4983045
23. Rapin, J., Teytaud, O.: Nevergrad - A gradient-free optimization platform. https://GitHub.com/FacebookResearch/Nevergrad (2018)
24. Stein, C.: Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: Proc. of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. pp. 197–206. University of California Press (1956), https://projecteuclid.org/euclid.bsmsp/1200501656
25. Storn, R., Price, K.: Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. J. of Global Optimization **11**(4), 341–359 (Dec 1997)
26. Surry, P.D., Radcliffe, N.J.: Inoculation to initialise evolutionary search. In: Fogarty, T.C. (ed.) Evolutionary Computing. pp. 269–285. Springer Berlin Heidelberg, Berlin, Heidelberg (1996)
27. Teytaud, O., Gelly, S., Mary, J.: On the ultimate convergence rates for isotropic algorithms and the best choices among various forms of isotropy. In: Proceedings of PPSN. pp. 32–41 (2006). https://doi.org/10.1007/11844297_4, https://doi.org/10.1007/11844297_4
28. Yang, X., Cao, J., Li, K., Li, P.: Improved opposition-based biogeography optimization. In: The Fourth International Workshop on Advanced Computational Intelligence. pp. 642–647 (2011)
29. Zhang, A., Zhou, Y.: On the non-asymptotic and sharp lower tail bounds of random variables (2018)

Appendix A: Relevant Concentration Bounds for χ^2 Distributions

We recall some basic definitions and properties of the central and the non-central χ^2 distributions, which are needed in the proofs of Theorems 1 and 2.

Definition 1. *(Central χ^2-distribution) Let $X_1, ..., X_d$ be d independent random variables drawn from the standard normal distribution $\mathcal{N}(0,1)$. Then the random variable $U = X_1^2 + ... + X_d^2$ follows a central $\chi^2(d)$ distribution with d degrees of freedom.*

As mentioned previously, the squared distance $||x^*||^2$ of x^* to the middle point 0 follows the central $\chi^2(d)$ distribution. This is thus also the distribution of the performance of the random sampling strategy using $\sigma^2 = 0$. In our proofs we will make use of the following properties of this distribution.

Property 1. (Properties of χ^2 distribution) Let $U \sim \chi^2(d)$. Then $\mathbb{E}(U) = d$, $\mathrm{var}(U) = 2d$, and for all $t \in [0,1]$ it holds that $\mathbb{P}\left[|\frac{U}{d} - 1| \geq t\right] \leq 2\exp(-\frac{dt^2}{8})$.

While the central χ^2 distribution suffices for the analysis of the middle point sampling strategy, *non-central χ^2 distribution* are required in the analysis of our Gaussian sampling with rescaled variance.

Definition 2. *(Non-central χ^2-distribution) Let $X_1, ..., X_d$ be independently drawn random variables satisfying $X_i \sim \mathcal{N}(\mu_i, 1)$. Let $U = X_1^2 + ... + X_d^2$. The random variable U follows a central $\chi^2(d, \mu)$ distribution with d degrees of freedom and non-centrality parameter $\mu = \sum_{i=1}^d \mu_i^2$.*

Note here that the non-central χ^2 distribution only depends on $\sum_{i=1}^d \mu_i^2$, but not on the individual values $(\mu_1, ..., \mu_d)$. Note further that, for a given point $x^* \in \mathbb{R}^d$, the distribution of the squared distance $||x - x^*||^2$ for $x \sim \mathcal{N}(0, I)$ follows the non-central $\chi^2(d, \mu)$ distribution with non-centrality parameter $\mu := ||x^*||^2$.

We recall some important properties of the non-central χ^2 distribution.

Property 2. (Properties of the non-central χ^2 distribution) Let $U \sim \chi^2(d, \mu)$. Then $\mathbb{E}(U) = d + \mu$, $\mathrm{var}(U) = 2(d + 2\mu)$, and for any $\beta > 1$ there exist positive constants C_1, C_β such that for all $x \leq (\mu + d)/\beta$ it holds that

$$P\left(U \leq -x\right) \geq C_1 \exp\left(-C_\beta \frac{x^2}{2\mu + d}\right). \tag{6}$$

Moreover, for all $x > 0$, it holds that

$$P\left(U \leq -x\right) \leq \exp\left(-\frac{1}{4}\frac{x^2}{2\mu + d}\right). \tag{7}$$

Proofs for the concentration inequalities 6 and 7 can be found in [29, Theorem 7].

Appendix B: Proof of Theorem 1 (Sufficient condition)

We now present the proof of Theorem 1, the sufficient condition for the scaling factor σ^2 to be beneficial over sampling the middle point. Let δ, λ and d satisfy the conditions of Theorem 1. Let $\varepsilon, \sigma > 0$. By the law of total probability it holds that, for all $t \leq 1$,

$$\begin{aligned}
&\mathbb{P}\left[\min_{i\in[\lambda]} ||x^* - x_i||^2 \leq (1-\varepsilon)\,||x^*||^2\right] \\
&= \mathbb{P}\left[\min_{i\in[\lambda]} ||x^* - x_i||^2 \leq (1-\varepsilon)\,||x^*||^2 \mid |\frac{||x^*||^2}{d} - 1| \leq t\right] \mathbb{P}\left[|\frac{||x^*||^2}{d} - 1| \leq t\right] \\
&+ \mathbb{P}\left[\min_{i\in[\lambda]} ||x^* - x_i||^2 \leq (1-\varepsilon)\,||x^*||^2 | |\frac{||x^*||^2}{d} - 1| > t\right] \mathbb{P}\left[|\frac{||x^*||^2}{d} - 1| > t\right].
\end{aligned}$$

Eq. 3 is therefore satisfied if

$$\mathbb{P}\left[\min_{i\in[\lambda]} ||x^* - x_i||^2 \leq (1-\varepsilon)\,||x^*||^2 \big| |\frac{||x^*||^2}{d} - 1| \leq t\right] \mathbb{P}\left[|\frac{||x^*||^2}{d} - 1| \leq t\right] \geq \delta.$$

This equation, in turn, is satisfied if for all y with $|\frac{||y||^2}{d} - 1| \leq t$ it holds that

$$\mathbb{P}\left[\min_{i\in[\lambda]} ||x^* - x_i||^2 \leq (1-\varepsilon)\,||x^*||^2 \big| x^* = y\right] \geq \frac{\delta}{\mathbb{P}\left[|\frac{||x^*||^2}{d} - 1| \leq t\right]}. \tag{8}$$

For the following computations, we fix $t := d^{-1/3}$ and we set $\delta' := \delta / \mathbb{P}\left[|\frac{||x^*||^2}{d} - 1| \leq t\right]$.

Let x^* be such that $|\frac{||x^*||^2}{d} - 1| \leq t$. Then, conditionally to x^*, we have

$$\begin{aligned}
&\mathbb{P}\left[\min_{i\in[\lambda]} ||x^* - x_i||^2 \leq (1-\varepsilon)\,||x^*||^2 \big| x^*\right] \\
&= 1 - \mathbb{P}\left[\min_{i\in[\lambda]} ||x^* - x_i||^2 \geq (1-\varepsilon)\,||x^*||^2 \big| x^*\right] \\
&= 1 - \mathbb{P}\left[||x - x^*||^2 \geq (1-\varepsilon)\,||x^*||^2 \big| x^*\right]^\lambda \\
&= 1 - \left(1 - \mathbb{P}\left[||x - x^*||^2 \leq (1-\varepsilon)\,||x^*||^2 \big| x^*\right]\right)^\lambda
\end{aligned}$$

for an x is distributed as a normal distribution $\mathcal{N}(0, \sigma^2 I)$. We recall that for such an x the distribution of the term $||x - x^*||^2/\sigma^2$ (for fixed x^*) follows the non-central $\chi^2(d, \mu)$ distribution with non-centrality parameter $\mu := ||x^*||^2/\sigma^2$. We therefore obtain (through simple algebraic manipulations) that condition (8) holds if and only if

$$\mathbb{P}\left[U \leq (1-\varepsilon)\,\frac{||x^*||^2}{\sigma^2}\right] \geq 1 - (1-\delta')^{1/\lambda},$$

with $U \sim \chi^2(d, \mu)$. Let $Y := U - \left(\frac{||x^*||^2}{\sigma^2} + d\right)$. Then the previous condition is equivalent to

$$\mathbb{P}\left[Y \leq -\left(\varepsilon\frac{||x^*||^2}{\sigma^2} + d\right)\right] \geq 1 - (1-\delta)^{1/\lambda}\,.$$

According to the concentration inequality 6, it holds that for any $\beta > 1$, there exist constants $C_1 > 0$ and $C_\beta > 0$ such that if

$$\varepsilon\frac{||x^*||^2}{\sigma^2} + d \leq \frac{1}{\beta}\left(\frac{||x^*||^2}{\sigma^2} + d\right), \tag{9}$$

then

$$\mathbb{P}\left(Y \leq -\left(\varepsilon\frac{||x^*||^2}{\sigma^2} + d\right)\right) \geq C_1 \exp\left(-C_\beta \frac{(\varepsilon\frac{||x^*||^2}{\sigma^2} + d)^2}{2\frac{||x^*||^2}{\sigma^2} + d}\right).$$

We deduce a sufficient condition for (8), by noting that it is satisfied if, for all x^* such that $|\frac{||x^*||^2}{d} - 1| \leq t$, it holds that

$$\frac{\left(\varepsilon\frac{||x^*||^2}{\sigma^2} + d\right)^2}{2\frac{||x^*||^2}{\sigma^2} + d} \leq A_\lambda, \tag{10}$$

with $A_\lambda := -\frac{1}{C_\beta}\left(\log\left(1 - (1-\delta')^{1/\lambda}\right) - \log C_1\right)$.

Let us now fix $\beta := 2$, $\varepsilon := c_1\frac{\log\lambda}{d}$ and $\sigma^2 := c_2\frac{\log\lambda}{d}$, with $c_1 := \frac{1}{3C_\beta}$ and $c_2 := c_1$. We show that, with these choices of β, ε and σ, inequalities (9) and 10 are satisfied if d is sufficiently large and x^* satisfies $|\frac{||x^*||^2}{d} - 1| \leq t$. To this end, first note that

$$\frac{\varepsilon\frac{||x^*||^2}{\sigma^2} + d}{(\frac{||x^*||^2}{\sigma^2} + d)} \leq \frac{\frac{c_1}{c_2}(1+t) + 1}{\frac{d}{c_2\log\lambda}(1-t) + 1}.$$

Under the assumptions stated in (2) the term $\frac{\frac{c_1}{c_2}(1+t)+1}{\frac{d}{c_2\log\lambda}(1-t)+1}$ converges to zero as $d \to \infty$. We therefore obtain that, for d sufficiently large and x^* satisfying $|\frac{||x^*||^2}{d} - 1| \leq t$, it holds that

$$\frac{\varepsilon\frac{||x^*||^2}{\sigma^2} + d}{\frac{||x^*||^2}{\sigma^2} + d} \leq \frac{1}{\beta},$$

which proves (9).

To show (10), we first note that

$$\frac{(\varepsilon\frac{||x^*||^2}{\sigma^2} + d)^2}{2\frac{||x^*||^2}{\sigma^2} + d} \leq \frac{\left(\frac{c_1}{c_2}(1+t) + 1\right)^2}{2\frac{d}{c_2\log\lambda}(1-t) + 1}.$$

Under the assumptions stated in (2), and since $d \to \infty$, we approximate

$$\frac{\frac{c_1}{c_2}(1+t)+1}{\frac{d}{c_2 \log \lambda}(1-t)+1} = \frac{c_2}{2}\left(\frac{c_1}{c_2}+1\right)^2 \log\lambda + o(\log\lambda) = \frac{2}{3C_\beta}\log\lambda + o(\log\lambda)$$

and $A_\lambda = \frac{1}{C_\beta}\log\lambda + o(\log\lambda)$, which shows that condition 10 holds for d sufficiently large and x^* satisfying $|\frac{||x^*||^2}{d} - 1| \leq t$.

Appendix C: Proof of Theorem 2 (Necessary condition)

We now prove the necessary condition which we have stated in Theorem 2. Let d, λ, ε, and σ satisfy the condition of Theorem 2. As in the beginning of the proof for Theorem 1, we can deduce the following necessary condition. For all $t \leq 1$ it holds that

$$\mathbb{P}\left[\min_{i\in[\lambda]} ||x^* - x_i||^2 \leq (1-\varepsilon)\,||x^*||^2 \middle| |\frac{||x^*||^2}{d} - 1| \leq t\right] \mathbb{P}\left[|\frac{||x^*||^2}{d} - 1| \leq t\right] + \mathbb{P}\left[|\frac{||x^*||^2}{d} - 1| > t\right] \geq \delta$$

Then there exists x^* such that $|\frac{||x^*||^2}{d} - 1| \leq t$ and

$$\mathbb{P}\left[\min_{i\in[\lambda]} ||x^* - x_i||^2 \leq (1-\varepsilon)\,||x^*||^2 \middle| x^*\right] \geq \frac{\delta - \mathbb{P}\left[|\frac{||x^*||^2}{d} - 1| > t\right]}{\mathbb{P}\left[|\frac{||x^*||^2}{d} - 1| \leq t\right]}. \quad (11)$$

Set $\delta' := \frac{\delta - \mathbb{P}\left[|\frac{||x^*||^2}{d} - 1| > t\right]}{\mathbb{P}\left[|\frac{||x^*||^2}{d} - 1| \leq t\right]}$. Then the necessary condition (11) can be written as

$$\mathbb{P}\left[Y \leq -\left(\varepsilon\frac{||x^*||^2}{\sigma^2} + d\right)\right] \geq 1 - (1-\delta')^{1/\lambda}$$

with $Y := U - (\frac{||x^*||^2}{\sigma^2} + d)$ and U being distributed according to a non-central χ^2 distribution with d degrees of freedom and non-centrality parameter $||x^*||^2/\sigma^2$. According to the concentration bound (7), we have

$$\mathbb{P}\left(Y \leq -\left(\varepsilon\frac{||x^*||^2}{\sigma^2} + d\right)\right) \leq \exp\left(-\frac{1}{4}\frac{(\varepsilon\frac{||x^*||^2}{\sigma^2} + d)^2}{2\frac{||x^*||^2}{\sigma^2} + d}\right).$$

Condition (11) therefore requires

$$\exp\left(-\frac{1}{4}\frac{(\varepsilon\frac{||x^*||^2}{\sigma^2} + d)^2}{2\frac{||x^*||^2}{\sigma^2} + d}\right) \geq 1 - (1-\delta')^{1/\lambda}.$$

From this we derive $\varepsilon \leq \left(\sqrt{\tilde{A}_\lambda\left(2\frac{||x^*||^2}{\sigma^2}+d\right)}-d\right)\frac{\sigma^2}{||x^*||^2}$, with $\tilde{A}_\lambda = -4\log\left(1-(1-\delta')^{1/\lambda}\right)$. As $\varepsilon > 0$, we obtain that

$$\sigma^2 < \tilde{\sigma}^2 := 2\frac{||x^*||^2/d}{\frac{d}{\tilde{A}_\lambda}-1}.$$

Fixing $t = d^{-1/3}$ and considering the requirements stated in (2) we obtain that $\tilde{\sigma} = 2\frac{\tilde{A}_\lambda}{d}+o\left(\frac{\tilde{A}_\lambda}{d}\right) = 8\frac{\log\lambda}{d}+o\left(\frac{\log\lambda}{d}\right)$, which concludes the proof of the necessary condition, as it shows $\sigma^2 \in O\left(\frac{\log\lambda_d}{d}\right)$. □

Appendix D: Proof of Theorem 3 (Upper Bound for the Gain)

The proof of Theorem 3 uses the same argument as the one of Theorem 2. We have proved that σ^2 must be between 0 and $\tilde{\sigma} = 2\frac{||x^*||^2/d}{\frac{d}{\tilde{A}_\lambda}-1}$. Then we get that:

$$\varepsilon \leq \sup_{\sigma\in[0,\tilde{\sigma}]}\left(\sqrt{\tilde{A}_\lambda\left(2\frac{||x^*||^2}{\sigma^2}+d\right)}-d\right)\frac{\sigma^2}{||x^*||^2}.$$

Noticing that:

$$\begin{aligned}&\sup_{\sigma\in[0,\tilde{\sigma}]}\left(\sqrt{\tilde{A}_\lambda\left(2\frac{||x^*||^2}{\sigma^2}+d\right)}-d\right)\frac{\sigma^2}{||x^*||^2}\\ &\quad=\sup_{\alpha\in[0,1]}\left(\sqrt{\tilde{A}_\lambda\left(2\frac{||x^*||^2}{\alpha\tilde{\sigma}^2}+d\right)}-d\right)\frac{\alpha\tilde{\sigma}^2}{||x^*||^2}\end{aligned}$$

We get after simple algebraic simplifications and for d sufficiently large under assumptions (2):

$$\begin{aligned}&\sup_{\sigma\in[0,\tilde{\sigma}]}\left(\sqrt{\tilde{A}_\lambda\left(2\frac{||x^*||^2}{\sigma^2}+d\right)}-d\right)\frac{\sigma^2}{||x^*||^2}\\ &\quad\leq\frac{d\tilde{\sigma}^2}{||x^*||^2}\sup_{\alpha\in[0,1]}\alpha\left(\sqrt{\alpha^{-1}+\frac{\tilde{A}_\lambda}{d^2}}-1\right)\\ &\quad\leq\frac{d\tilde{\sigma}^2}{||x^*||^2}\sup_{\alpha\in[0,1]}\alpha\left(\sqrt{\alpha^{-1}+1}-1\right)\\ &\quad\leq 8\frac{\log\lambda}{d}+o\left(\frac{\log\lambda}{d}\right)\end{aligned}$$

Then $\varepsilon \in O\left(\frac{\log\lambda_d}{d}\right)$, which concludes the proof of Theorem 3. □