How many random restarts are enough?

Travis Dick, Eric Wong, Christoph Dann
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
tdick@cs.cmu.edu, ericwong@andrew.cmu.edu, cdann@cmu.edu

Abstract

Many machine learning problems, such as K-means, are non-convex optimization problems. Usually they are solved by performing several local searches with random initializations. How many searches should be done? Typically a fixed number is performed, but how do we know it was enough? We present a new stopping rule with non-asymptotic frequentist guarantees, which, to our knowledge, no existing rule has. By comparing all stopping rules on various benchmarks, we shed light on their effectiveness in machine-learning problems, including K-means and maximum marginal likelihood parameter selection.

1 Introduction

Many important algorithms in machine learning are based on solving convex optimization problems (e.g. support vector machines or least-squares linear regression). However, numerous important approaches also involve non-convex problems. Prominent examples are K-means clustering, maximizing the marginal likelihood of hyper-parameters (ML Type 2) in Bayesian models (Rasmussen and Williams, 2006) and maximum likelihood parameter estimation with latent variables (expectation-maximization). The prevalent approach for optimizing these problems in practice is to use descent methods (e.g. approximate Newton methods), which find local optima. As the function value of many local optima might be much worse than the true global optimum, these methods are run several times with random initializations. This immediately raises the question: *How many restarts are enough to find a good / the global optimum?*

For example, suppose that after ten restarts we see a different local optimum each time. Should we keep restarting the local optimization method? What about when we see the same local optimum ten times? Intuitively, one would keep optimizing in the first case and stop in the second. However, the most common practice in the machine learning community is to stop after a fixed number of random restarts (often 10) which might yield a local optima with bad quality or be computationally wasteful.

Several adaptive stopping rules have been proposed in the late 1980s and 1990s by the optimization community. While some rules are simple heuristics, others are derived within a sound Bayesian framework or from a frequentist perspective. Apparently, these methods have been mostly ignored by the machine learning community, despite the persisting popularity of optimizing with random restarts (*multistart* methods).[1] By providing an experimental evaluation on synthetic benchmark problems and real-world machine learning problems, this paper sheds some light on why these methods have been ignored and whether they are viable strategies in practice.

While some of the existing methods have been shown to be Bayes-optimal with respect to their respective cost function, no guarantees are given for whether the best optimum found is globally optimal. Even the frequentist rules have only asymptotic guarantees, which are of little use when

[1] According to a literature survey by Martí et al. (2013), the number of publications mentioning *multistart* has increased significantly in the last years.

stopping after a finite and usually small number of restarts. To this end, this paper presents an adaptive stopping rule based on the Good-Turing estimator (Good, 1953), that has high-probability guarantees on how much of the input space of the function has been explored during optimization, if the number of local optima can be bounded. The theoretical analysis of this rule in this paper shows that, for sufficiently smooth objective functions with box constraints, this can be extended to a guarantee on the quality of the best optimum found. Given the difficulty of the general problem, one expects that a rule with high-probability guarantees will be very conservative when compared with heuristic rules. The experimental evaluation in this paper therefore compares our novel stopping rule against the existing methods and answers the question of whether this rule is useful in practice.

1.1 Problem Formalization

The considered scenario can be formalized as follows: Given a function $f : \mathcal{S} \to \mathbb{R}$, where $\mathcal{S}$ is a finite-measure space, we seek to find a global minimum of f. The multistart approach to this problem is to start a local search method G, such as gradient descent or Newton's method, from several randomly chosen starting points y. Each run yields a solution $x = G(f, y) \in \mathcal{S}$ with function value $f(x)$, usually a local minimum of f. We assume that there are only countably many outcomes $(x^{(i)})_{\mathcal{I}}\, \mathcal{I} \subseteq \mathbb{N}$ of the local search method. This assumption holds for almost all problems in practice. For example, most functions of interests have finitely many local optima, i.e., the index set is finite, $\mathcal{I} = \{1, 2, \ldots, N\}$.

If the local search method is deterministic, the input space $\mathcal{S}$ can be partitioned into *regions of attraction* of each local optima $\mathcal{S}_i = \{y \in \mathcal{S} : G(f, y) = x^{(i)}\}$ for $i \in \mathcal{I}$. Optimizing with multiple restarts then corresponds to sampling from a discrete distribution over the local optima. More specifically, the sampling distribution is defined on the index set $\mathcal{I}$ with probabilities $\theta_i = p(i) = \mathbb{P}(y \in \mathcal{S}_i)$ where $\mathbb{P}(y \in \mathcal{S}_i)$ is the probability of initializing the local search within the region of attraction of local optimum $x^{(i)}$. This probabilistic characterization allows us to deal with distributions over a discrete set $\mathcal{I}$ rather than the more complicated space $\mathcal{S}$.

2 Overview of Existing Stopping Rules

This section discusses several existing stopping rules for the multistart method, which can be characterized as either Bayesian, heuristic or frequentist rules. The heuristic methods calculate statistics of the observed optima that are intuitively informative about whether we should stop, or not. Frequentist rules are similar to heuristic ones, but provide frequentist guarantees. In contrast, Bayesian methods directly aim at minimizing a specified loss function.

2.1 Bayesian Stopping Rules

Bayesian methods consist of three major components: a prior, likelihood and loss function. The likelihood relates the unknown true discrete distribution over local optima to the observed quantities. Together with a prior belief on the unknowns, the posterior probability of the unknown true distribution given the observations can be derived. Bayesian methods stop as soon as the expected loss under the current posterior would increase with further restarts.

Boender and Kan Rules (Boender and Kan, 1987). Boender and Kan (1987) relate the observed empirical distribution $(n_1, n_2, \ldots n_w)$ of w different minima with the unknown total number of minima k and their distribution $\theta_1, \ldots \theta_k$ by the likelihood

$$p((n_1, \ldots, n_w)|k, \theta_1, \ldots \theta_k) \frac{1}{\prod_{j=1}^{n} h_j} \sum_{(i_1, \ldots i_w) \in \mathcal{P}(k,w)} p_{\text{multin.}}(n_1, \ldots n_w; \theta_{i_1}, \theta_{i_2}, \ldots \theta_{i_w}). \quad (1)$$

This distribution generalizes the multinomial distribution $p_{\text{multin.}}$ by averaging over all possible choices of w local minima out of the k available minima as it is unknown which optima we found during optimization. The set $\mathcal{P}(k, w)$ denotes all possible tuples of w different elements from the index set $\{1, \ldots, k\}$. To account for possible reorderings of the observation $(n_1, \ldots n_w)$, the probability is normalized by $\prod_{j=1}^{n} h_j$, where h_j is the number of n_i which have the same value j.

The likelihood in Equation (1) is augmented by a prior on k and $\theta_1, \ldots, \theta_k$ which puts uniform mass on all $k \in [1, \infty)$ (improper prior) and a uniform Dirichlet prior $\mathrm{Dir}(1, \ldots, 1)$ for $\theta_1, \ldots, \theta_k$.

The rules of Boender and Kan (1987) then stop as soon as the expected loss with respect to the posterior (based on the likelihood and prior defined above) would increase when continuing to restart. The authors consider four different loss functions to trade-off computational cost of further restarts and penalties for suboptimal solutions

$$L_1 = n + cI\{k > w\} \qquad L_2 = c(k-w) + n \qquad L_3 = c\left(\frac{k-w}{k}\right) + n \qquad L_4 = cM_n + n. \quad (2)$$

The cost of restarting n is always linear in the number of restarts. The parameter c scales penalties for suboptimality in each cost function and therefore controls how early the rules stop. While the first cost function L_1 assigns a fixed penalty if we have not found all minima, the second cost L_2 gives a penalty proportional to the number of unseen local minima $k-w$. The third cost L_3 considers the fraction of unseen minima and the penalty in the last cost function is proportional to the *missing mass* M_n, i.e., the probability mass of all unseen minima (cf. Section 3.1).

Betrò and Schoen Rule. Betro and Schoen (1987) build on the work of Boender and Kan. Their method is based on a Bayesian estimate of the discrete distribution over the countable set of objective values $\{f(x^{(i)})\}_{i \in \mathcal{I}}$ rather than the local optima themselves $\{x^{(i)}\}_{i \in \mathcal{I}}$. The intuitive idea is that, by modeling the objective values, we can stop when the cost of finding a *better* local minima is too high, rather than the cost of simply finding a yet unobserved local optima.

They suppose that the true distribution over objective values belongs to the nonparametric family of so-called *neutral to the right* distributions, which have cumulative densities of the form $F(t) = 1 - \exp(-Y(t))$, where $Y(t)$ is a stochastic process that is non-decreasing almost surely, is right continuous, and satisfies $\lim_{t \to -\infty} Y(t) = 0$ and $\lim_{t \to \infty} Y(t) = \infty$. For computational tractability, they restrict themselves to Y being a homogeneous process. For this non-parametric family of distributions, they derive expressions for the posterior distribution given n observed samples from the distribution.

The loss function that the Betrò and Schoen rule uses is given by $L(f_1, \ldots, f_n; c) = nc + \min_{1 \le i \le n} f_i$, where $f_1, \ldots, f_n$ is the sequence of objective values returned by the local search procedure. Rather than finding an optimal stopping rule, Betrò and Schoen consider one and two step lookahead rules. These rules stop as soon as we expect the loss to increase after one or two more restarts.

2.2 Heuristic Stopping Rules

Assuming a box-constrained domain with a finite number of local minima, Lagaris and Tsoulos (2008) present three alternative stopping rules for finding all local minima. All rules are based on the same scheme. They stop as soon as a new minima has not been found for v restarts. The number of restarts allowed to find the next minima depends on the convergence speed of the sample variance of different statistics and generally increases with the number of minima found. The rules differ in the statistics used. The first stopping rule, the double-box heuristic, essentially considers Bernoulli variables that are independent of the actual data observed. In the second rule, v depends on an estimate of the variance of the number of times each local minima is observed after a certain number of restarts. In the third rule, the expected minimizers rule, v is determined by the variance of the estimate of the number of minima.

2.3 Frequentist Rules with Asymptotic Guarantees

Dorea's stopping rule. The stopping rule proposed by Dorea (1990) is based on estimating the probability p_ϵ that the function value of a local optima found in a single start is within ϵ of the global optimum. The probability p_ϵ is estimated by $\hat{p}_\epsilon = \rho_n(\epsilon)/n$, where $\rho_n(\epsilon)$ is the number of optima that were better than all found before and that are within ϵ of the best minimum $f_n^* = \min_{1 \le i \le n} f_i$. The justification for using this instead of the fraction of all samples within ϵ of f_n^*, is for reasons of computational efficiency. Dorea defines two stopping rules. The first one stops as soon as the probability of having already found an ϵ-optimal minimum, estimated based on $\hat{p}_\epsilon$, is high enough.

In contrast, the second rule stops conditioned on having seen no improvement in the last m steps. For infinitely many restarts, these guarantees hold with the chosen probability.

Hart's stopping rule. Using the same underlying idea, Hart (1998) proposed a modified version of Dorea's stopping rule by making two changes: first, he improved the estimate for p_ϵ by additionally counting the number of points that are ϵ-close to f_n^* ever since f_n^* was found. Next, he introduced a new parameter δ to reflect the confidence of the estimate of p_ϵ, and correspondingly modified the two stopping rules to guarantee that the probability of $|\hat{p}_\epsilon - p_\epsilon| \leq \delta$ is sufficiently large.

For both rules, only asymptotic guarantees are provided. However, those are hardly useful in practice as the whole point of stopping rules is to have a finite (and often small) set of samples. Therefore, a frequentist rule with finite-sample guarantees is highly desirable.

3 A novel high-confidence stopping rule

In general, there are no finite sample guarantees for the stopping rules presented in the previous section. We will therefore introduce now a novel rule that enjoys such guarantees for sufficiently smooth objective functions. The key idea of this rule is to stop as soon as we are certain that the *missing mass*, the probability of all unseen local minima, is small enough. In the following sections, we will first give some background on the problem of estimating the missing mass, then present the actual stopping rule, and subsequently provide guarantees in the theoretical analysis.

3.1 Estimating the missing mass

Given a finite number of samples $i_1, i_2, \dots i_n \in \mathcal{I}$ of local minima, the missing mass M_n is defined as the total probability mass of all minima that have not been observed

$$M_n = \sum_{i \in \mathcal{F}_n^0} p(i) \quad \text{where} \quad \mathcal{F}_n^r = \left\{ i \in \mathcal{I} : \sum_{j=1}^{n} \mathcal{I}\{i_k = i\} = r \right\}.$$

The set $\mathcal{F}_n^r$ denotes the indices of all minima that are observed exactly r times. Intuitively, the missing mass converges in probability to 0 as $\lim_{n\to\infty} \mathbb{P}(M_n > \epsilon) = \lim_{n\to\infty}(1-\epsilon)^n = 0$ for all $\epsilon > 0$. Historically, estimating the missing mass played a significant role in breaking the German Enigma Cipher during World War II (Orlitsky et al., 2003). The British researchers A.M. Turing and I.J. Good needed to estimate the probability with which German U-boat commanders chose a cipher from each page in a cipher book. As the sample size was small compared to the number of pages, the researchers had to explicitly account for the missing mass. To this end, they developed an estimator G_n for the missing mass[2] $G_n = |\mathcal{F}_n^1|/n$ which is simply the fraction of samples that are observed exactly once. It became known as the Good-Turing estimator (Good, 1953) and has been widely used in many applications, especially language modeling (Church and Gale, 1991).

There have been several recent advances in the theoretical analysis of the missing mass. Notably, McAllester and Schapire (2000) analyze the convergence rate of the Good Turing estimator G_n to M_n. Besides lower bounds, they prove the following upper bound on the missing mass: Let $\delta > 0$. Then with probability at least $1 - \delta$, it holds that

$$M_n \leq G_n + (2\sqrt{2} + \sqrt{3})\sqrt{\frac{\ln \frac{3}{\delta}}{n}}. \tag{3}$$

To the best of our knowledge, these results have not been leveraged for the optimization setting we are considering.

3.2 Description of the HCS rule

Inequality (3) allows us to bound the actual missing mass with high probability based on the Good-Turing estimate and the number of restarts. The bound can be computed during runtime as all

[2]The Good-Turing estimator is actually more general and can also be used for outcomes that were observed. For simplicity we restrict the exposition to the missing mass.

necessary quantities are observed. Hence, the high-confidence stopping (HCS) rule defined in Algorithm 1 is a valid stopping rule.

Algorithm 1: High-confidence stopping (HCS) rule

Parameters: $\delta \in (0,1)$: confidence level, $c \in (0,1)$: maximum acceptable missing mass
Rule: Stop after n restarts if

$$C_n := \frac{|\mathcal{F}_n^1|}{n} + (2\sqrt{2}+\sqrt{3})\sqrt{\frac{\ln\frac{3}{\delta}}{n}} < c$$

The rule stops as soon the as the missing mass is below a threshold c with probability at least $1-\delta$. The quantity C_n is exactly the bound from Equation (3). Both c and δ are parameters of the stopping rule. We will discuss the effect of different parameter choices in Section 4.

3.3 Theoretical analysis

From the construction of the high-confidence stopping rule and Theorem 9 by McAllester and Schapire (2000), one might immediately conclude the follow hypothesis.

Hypothesis 1. *When the stopping rule in Algorithm 1 stops, the total probability of all unseen local minima M_n is less than c with probability at least $1-\delta$.*

While the validity seems trivial at first, it is actually still unclear whether this hypothesis is true. To shed some light on this issue, consider the following more general statement

Proposition 1. *Let $\theta_1, \theta_2, \dots$ and $B_1, B_2, \dots$ be sequences of random variables for which for all n holds that $\mathbb{P}(\theta_n \leq B_n) \geq 1-\delta$. Consider the stopping rule $\tau = \inf\{t\,;\, B_t \leq c\}$. Then the statement $\mathbb{P}(\theta_\tau \leq B_\tau) \geq 1-\delta$ for finite τ does not necessarily hold.*

Proof. Counterexample: Take $\theta_n = 1/2$ and $B_n \sim$ Bernoulli$(1-\delta)$. Then $\mathbb{P}(\theta_n \leq B_n) = \mathbb{P}(B_n = 1) \geq 1-\delta$. If $c = 1/4$, then $\tau = \inf\{t : B_t \leq 1/4\} = \inf\{t : B_t = 0\}$ and so $\mathbb{P}(\theta_\tau \leq B_\tau) = \mathbb{P}(1/2 \leq 0) = 0 < 1-\delta$. □

In the counterexample above, the stopping time τ picks out exactly the events with mass $< \delta$ where the confidence bound does not hold. Since for the HCS rule, τ, M_n and C_n are correlated random variables, it is not clear whether τ exhibits a preference to events where the bound does not hold. In experiments, such behavior could not be observed. In order to prove Hypothesis 1, we would have to exploit further properties of the bound C_n. The proof (or refuting the hypothesis) is, however, left as future work.

Finitely many local optima: Instead, we restricted our analysis to the case when we know an upper bound N on the number of local minima. This allows us to provide upper and lower bounds on τ in the following propositions and subsequently use them to prove a high-confidence bound on the missing mass at stopping time τ in Theorem 1.

Proposition 2. *Let $\tau = \inf\{n : C_n < c\}$ be the stopping time of the HCS rule with parameters c and δ, and let*

$$g(\delta) = (2\sqrt{2}+\sqrt{3})\sqrt{\ln(3/\delta)}.$$

Then

$$\tau > \tau_{\min} = \frac{g(\delta)^2}{c^2} = \frac{\ln(3/\delta)(2\sqrt{2}+\sqrt{3})^2}{c^2}, \tag{4}$$

almost surely, that means, there are at least $\tau_{min} = O\left(\frac{\ln(3/\delta)}{c^2}\right)$ restarts with probability 1.

Proof. Since $G_n = |\mathcal{F}_n^1|/n \geq 0$, the rule can only stop after n restarts if

$$(2\sqrt{2}+\sqrt{3})\sqrt{\frac{\ln(3/\delta)}{n}} < c \Leftrightarrow (2\sqrt{2}+\sqrt{3})^2\frac{\ln(3/\delta)}{n} < c^2 \Leftrightarrow (2\sqrt{2}+\sqrt{3})^2\frac{\ln(3/\delta)}{c^2} < n. \tag{5}$$

□

For example, for $c = 0.1$ and $\delta = 0.1$, the algorithm does at least 7074 restarts.

Proposition 3. *Let $\tau = \inf\{n : C_n < c\}$ be the stopping time of the HCS rule with parameters c and δ, and let $g(\delta) = (2\sqrt{2} + \sqrt{3})\sqrt{\ln(3/\delta)}$. Furthermore, assume that there are fewer than N local minima. Then*

$$\tau < \tau_{\max} = \frac{(\sqrt{cN} + g(\delta))^2}{c^2} = \frac{\left(\sqrt{cN} + (2\sqrt{2} + \sqrt{3})\sqrt{\ln(3/\delta)}\right)^2}{c^2} \tag{6}$$

almost surely. That means there will be no more than $\tau_{\max} = O((\sqrt{cN} + \ln(3/\delta))^2/c^2)$ restarts almost surely.

Proof. Since there are no more than N local minima, the number of local minima observed exactly once is at most N with probability 1. Therefore, $G_n = |\mathcal{F}_n^1|/n \leq N/n$ and we have $C_n = G_n + g(\delta)/\sqrt{n} \leq N/n + g(\delta)/\sqrt{n}$ almost surely. Hence, whenever $N/n + g(\delta)/\sqrt{n} < c \Leftrightarrow N + g(\delta)\sqrt{n} - c\sqrt{n}^2 < 0$ the rule stops w.p. one. Solving the inequality for $\sqrt{n}$ and applying Lemma 1, we have $c > N/n + g(\delta)/\sqrt{n} \geq C_n$ almost surely for $n \geq (\sqrt{cN} + g(\delta))^2/c^2$. It follows that $\tau < \tau_{\max} = (\sqrt{cN} + g(\delta))^2/c^2$ almost surely. □

Theorem 1. *Let $\tau = \inf\{t \in \mathbb{N} : C_t < c\}$ be the stopping time of the HCS rule with parameters c and δ. Then*

$$\mathbb{P}(M_\tau < c) \geq 1 - \hat{\delta} = 1 - (\tau_{\max} - \tau_{\min})\delta = 1 - \frac{\delta(cN + 2\sqrt{cN}(2\sqrt{2} + \sqrt{3})\sqrt{\ln(3/\delta)})}{c^2}.$$

Proof. The main idea of the proof is use the upper- and lower bounds on τ and then apply the union bound on the remaining finitely many possible stopping times. The full proof is in the appendix. □

In comparison to Hypothesis 1, Theorem 1 requires an upper bound on the number of local minima and the confidence level $1 - \hat{\delta}$ where $\hat{\delta} = (\tau_{\max} - \tau_{\min})\delta$ is generally lower than $1 - \delta$. However, the confidence level can be chosen to be arbitrarily close to 1 by choosing the rule parameter δ accordingly.

For arbitrary functions, an unobserved minima can have arbitrarily low function values, as long as the missing mass is nonzero. We therefore have to put additional structural assumptions on the class of functions we are considering to relate the missing mass to the quality of unobserved minima. The following theorem provides, for example, a bound on the quality of the found minima at the stopping time, when the objective function is Lipschitz-continuous with box-constraints.

Theorem 2. *Assume that f is Lipschitz-continuous with a Lipschitz-constant L and that the region of attraction of each minima of f has at least probability $p_{\min}$ according to the uniform sampling distribution for initial points. Further assume that the space of feasible solutions $\mathcal{S} = \{\vec{x} : \vec{l} \leq \vec{x} \leq \vec{u}\} \subset \mathbb{R}^d$ is box-shaped with $u_i - l_i > 2r$ where*

$$r = \frac{2}{\sqrt{\pi}}\left(c\Gamma\left(\frac{d}{2} + 1\right)\right)^{1/d}.$$

Then once the rule in Algorithm 1 with parameters c and δ stops, with probability at least $1 - \hat{\delta} = 1 - (\tau_{\max} - \tau_{\min})\delta$, the function value of the best local optima $\tilde{x}$ found satisfies

$$f(\tilde{x}) \leq f^* + Lr$$

where $f^* = \min_{x \in \mathcal{S}} f(s)$ the global optimum.

Proof. The main idea of the proof is to use a geometric argument and show that in the worst case, the missing mass is concentrated in one of the vertices of the polytope of feasible solutions. Then, using the maximum possible distance between the corner and a point outside the missing mass and the Lipschitz-continuity, we get the desired bound on the function values. A full proof is in the appendix. □

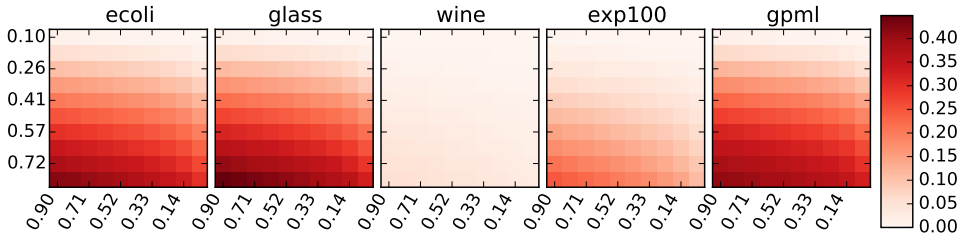

Figure 1: True missing mass M_τ at the stopping time τ according to Algorithm 1 for different parameter settings. The horizontal axis shows the confidence parameter δ and vertical axis the missing mass threshold c. The first three plots show k-means problems, the fourth plot a synthetic benchmark and the right plot a marginal log-likelihood parameter selection problem for Gaussian processes. Results are averages of 1000 independent trials.

4 Experimental Comparison

In this section, we empirically study the effects of the parameters c and δ on the performance of our proposed stopping rule. Additionally, we compare our rule against existing ones on several benchmark problems.

4.1 Benchmarks and Other Methods

We compare against the four Bayesian rules of Boender and Kan (1987) (B.K.-Fi, B.K.-N, B.K.-Fr, B.K.-M for the fixed, number, fraction, and missing mass penalties, respectively) and Betro and Schoen (1987) (B.S.), the two frequentist rules of Hart (1998) and Dorea (1990), and the two heuristic rules of Lagaris and Tsoulos (2008) (DoubleBox and Observables).

In four of the benchmark problems, we run the K-means algorithm on the Iris, E.Coli, Glass, and Wine UCI datasets (Bache and Lichman, 2013). We also include a problem where lBFGS is used to determine the maximum marginal likelihood parameters for fitting a Gaussian process to the SARCOS dataset. This dataset and the optimization procedure for this benchmark were taken from the GPML toolbox (Rasmussen and Williams, 2006). Finally, we included four synthetic problems (Exp10, Exp40, Exp80, and Exp100) which had $N \in \{10, 40, 80, 100\}$ local minima linearly spaced between 0 and 1 at $x_1, \ldots, x_N$ with function values $f(x_i) = x_i$ and where the probability of observing outcome x_i was proportional to $\exp(Nx_i/20)$.

4.2 Effect of Rule Parameters

For five representative problems, Figure 1 shows the true missing mass M_τ at the stopping time of our rule (average of 1000 independent trials). As expected the missing mass decreases with the threshold parameter c and parameter δ. We see that the expected missing mass is lower than the desired threshold, even for very large values of δ. This indicates that Hypothesis 1 might indeed be true. In addition, we observe that the missing mass for fixed parameters varies across optimization problems. In the very easy *wine* benchmark, which only has 19 local minima, the missing mass at stopping time τ is much lower than for the more challenging problems with hundreds of minima. Most likely this is caused by the fact that the original fixed-time high-confidence bound on the missing mass is very loose in such situations (see also the Good-Turing estimator analysis of Ohannessian and Dahleh (2010)).

Figure 2 shows the stopping time in $\log_{10}$ space for the same situations as Figure 1. We observe that for conservative settings of δ and c, the stopping time varies little among different problems. In contrast, for low values of c and δ, the rule adapts to the "difficulty" of different problems and stops earlier in the simple *wine* benchmark than in the challenging benchmarks *GPML* and *glass*.

4.3 Comparison Against Other Rules

Figure 3 presents a comparison of the average stopping time for each stopping rule on the *Iris* and *E.coli* K-means problems. The stopping rules perform similarly on the other benchmark problems.

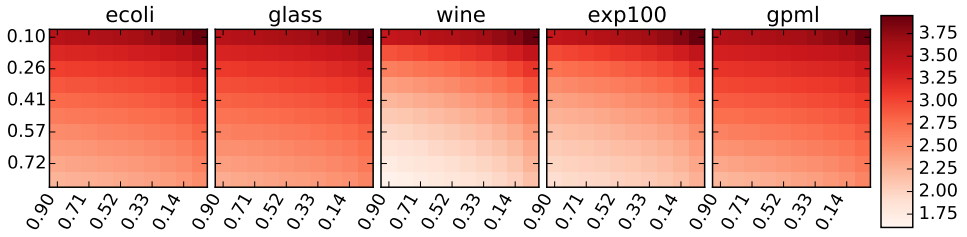

Figure 2: Stopping time τ in $\log_{10}$ space of Algorithm 1 for different parameter settings. The setting and presentation is the same as in Figure 1

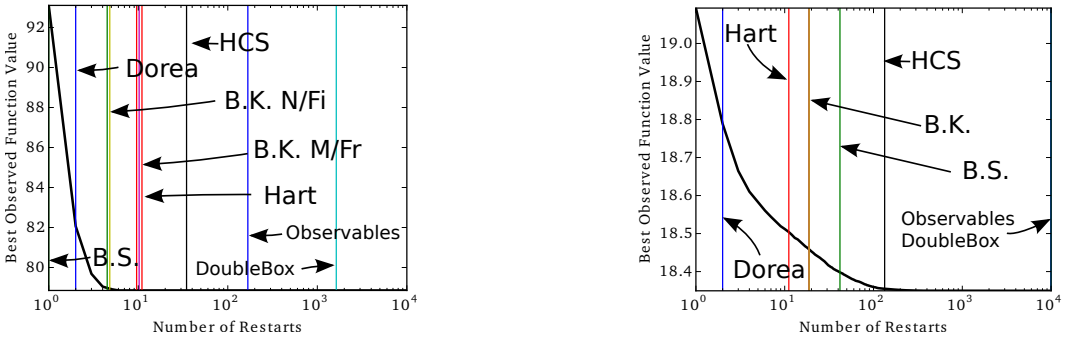

Figure 3: Comparison of all stopping rules on the *Iris* (left) and *E.coli* K-means benchmark. On the E.coli benchmark, "B.K." is used to label all four B.K.-* methods, since they do nearly the same number of restarts. We performed parameter studies for each method to determine parameter settings that stopped rather aggressively across all benchmark problems. We used these aggressive parameters in the comparison, since in practice we will not tune parameters on a per-problem basis.

For the complete comparison see Tables 1 and 2 in the appendix. The thick black curve plots the expected function value obtained as a function of the number of restarts, averaged over 800 trials. For each stopping rule, there is a vertical line that shows the mean stopping time of that rule, averaged over 800 trials. For the Iris dataset, on average 10 restarts are required to find the optimal minima, while for the E.coli dataset 100 restarts are required.

In both cases, we see that the number of restarts performed by our rule is the correct order of magnitude (34 for the Iris dataset, and 133 for the E.coli dataset). The other rules also perform more restarts for the harder E.coli dataset, but they are more aggressive and stop before the best function value is found. This is consistent with what we would expect theoretically. Our rule waits until the missing mass is low with high probability, which improves our chances of finding the optimal function value at the cost of being very conservative. In both problems, our rule is more conservative than all the other rules, except for the heuristic Double Box and Observables rules. In fact, in some of the other benchmark problems, our rule is extremely conservative.

5 Conclusion

We presented a novel stopping rule for function optimization with random restarts based on bounding the missing mass, that is the probability of finding new minima. We compared our rule against existing approaches on various benchmark problems. In contrast to existing rules, we proved meaningful finite-sample guarantees on the missing mass and the quality of the best found minima for sufficiently smooth optimization problems with box constraints. While the current results only work when a bound on the number of minima is known, it is an interesting open question as to whether the results can be extended to the general case.

References

K. Bache and M. Lichman. {UCI} Machine Learning Repository, 2013. URL http://archive.ics.uci.edu/ml.

B. Betro and F. Schoen. Sequential Stopping Rules For The Multistart Algorithm in Global Optimisation. *Mathematical Programming*, 38:271–286, 1987.

C. Boender and A. Kan. Bayesian stopping rules for multistart global optimization methods. *Mathematical Programming*, 37:59–80, 1987.

K. W. Church and W. a. Gale. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech & Language*, 5(1):19–54, 1991.

C. Dorea. Stopping rules for a random optimization method. *SIAM Journal on Control and Optimization*, 28 (4):841–850, 1990.

I. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40 (3):237–264, 1953.

W. Hart. Sequential stopping rules for random optimization methods with applications to multistart local search. *SIAM Journal on Optimization*, 9(1):270–290, 1998.

I. Lagaris and I. Tsoulos. Stopping rules for box-constrained stochastic global optimization. *Applied Mathematics and Computation*, 197(2):622–632, 2008.

R. Martí, M. G. Resende, and C. C. Ribeiro. Multi-start methods for combinatorial optimization. *European Journal of Operational Research*, 226(1):1–8, 2013.

D. McAllester and R. Schapire. On the Convergence Rate of Good-Turing Estimators. In *Conference on Learning Theory*, 2000.

M. Ohannessian and M. Dahleh. Distribution-dependent performance of the Good-Turing estimator for the missing mass. In *19th International Symposium on Mathematical Theory of Networks and Systems*, pages 679–682, 2010.

A. Orlitsky, N. Santhanam, and J. Zhang. Always good turing: Asymptotically optimal probability estimation. *Science*, 2003.

C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006. ISBN 026218253X.

A Appendix

A.1 Full Performance Comparison Data

	Ecoli	Glass	Wine	Iris	GPML
Good-Turing	133.5 / **0.0**	153.9 / **0.1**	36.2 / **28.1**	34.6 / **0.0**	114.3 / **0.0**
Dorea	2.0 / **0.5**	2.0 / **33.7**	2.0 / **132.9**	2.0 / **4.3**	2.0 / **346.8**
Hart	11.1 / **0.2**	11.3 / **3.9**	11.2 / **67.6**	11.0 / **0.0**	11.3 / **4.9**
BK-Num	18.8 / **0.1**	16.9 / **1.5**	4.0 / **107.5**	4.5 / **0.3**	6.9 / **144.2**
BK-Mass	18.8 / **0.1**	17.3 / **1.1**	8.7 / **81.3**	9.6 / **0.0**	11.8 / **20.1**
BK-Fraction	18.8 / **0.1**	17.0 / **1.4**	9.7 / **74.1**	10.2 / **0.0**	9.2 / **11.3**
BK-Fix	18.8 / **0.1**	16.9 / **1.5**	4.8 / **95.1**	4.8 / **0.1**	7.1 / **68.3**
Betro-Schoen	41.9 / **0.0**	1.0 / **66.5**	1.0 / **2342.8**	1.0 / **14.2**	1.0 / $\mathbf{2.13 \times 10^{19}}$
Double-Box	10000.0 / **0.0**	10000.0 / **0.0**	8498.0 / **7.0**	1628.4 / **0.0**	10000.0 / **-0.0**
Observables	9909.5 / **0.0**	9951.2 / **0.0**	1551.5 / **10.4**	167.5 / **0.0**	9991.3 / **-0.0**

Table 1: This table shows the results of running each of the stopping rules with aggressive parameter settings on optimization problems that arise in machine learning problems. Each table entry shows two numbers: the first is the number of restarts and the second (bold) is the suboptimality after stopping $f(x^*_\tau) - f(x^*)$. Each entry is the mean of 800 independent runs.

	Exp10	Exp40	Exp80	Exp100
Good-Turing	33.8 / **0.0**	56.3 / **0.0**	67.2 / **0.1**	68.2 / **0.1**
Dorea	2.0 / **0.4**	2.0 / **0.5**	2.0 / **0.6**	2.0 / **0.7**
Hart	11.3 / **0.1**	11.3 / **0.1**	11.3 / **0.3**	11.2 / **0.4**
BK-Num	7.0 / **0.1**	12.1 / **0.2**	13.4 / **0.3**	13.5 / **0.4**
BK-Mass	11.7 / **0.1**	12.5 / **0.1**	13.7 / **0.3**	13.7 / **0.4**
BK-Fraction	8.5 / **0.1**	12.1 / **0.2**	13.5 / **0.3**	13.5 / **0.4**
BK-Fix	7.0 / **0.1**	12.1 / **0.2**	13.4 / **0.3**	13.5 / **0.4**
Betro-Schoen	6.0 / **0.1**	6.0 / **0.2**	6.0 / **0.4**	6.0 / **0.5**
Double-Box	165.7 / **0.0**	1317.2 / **0.0**	9211.9 / **0.0**	10000.0 / **0.0**
Observables	71.7 / **0.0**	553.5 / **0.0**	3301.2 / **0.0**	7961.8 / **0.0**

Table 2: This table shows the results of running each of the stopping rules with aggressive parameter settings on synthetic distributions. The table entries are as in Table 1.

A.2 Proofs and Lemmas

Lemma 1. *Let $N, n, g(\delta) > 0$, and suppose $C_n \leq N/n + g(\delta)/\sqrt{n}$. Then, when $n > (\sqrt{cN} + g(\delta))^2/c^2$, we have $c > N/n + g(\delta)/\sqrt{n} \geq C_n$.*

Proof. Suppose $n > (\sqrt{cN} + g(\delta))^2/c^2$, and rearrange the target inequality $c > N/n + g(\delta)/\sqrt{n}$ to $nc - g(\delta)\sqrt{n} - N > 0$. By applying the quadratic formula with respect to $\sqrt{n}$, this inequality

holds when $\sqrt{n} > \frac{1}{2c}(g(\delta) + \sqrt{g(\delta)^2 - 4Nc})$. This is indeed true since

$$\begin{aligned}
n &> \frac{1}{c^2}(\sqrt{cN} + g(\delta))^2 \\
&\geq \frac{1}{c^2}(cN + \sqrt{cN}g(\delta) + g(\delta)^2) \\
&= \frac{1}{4c^2}(4cN + 2\sqrt{4cN}g(\delta) + 4g(\delta)^2) \\
&= \frac{1}{4c^2}(4cN + 2g(\delta)(\sqrt{4cN} + g(\delta)) + 2g(\delta)^2) \\
&\geq \frac{1}{4c^2}((4cN + g(\delta)^2) + 2g(\delta)\sqrt{4cN + g(\delta)^2} + g(\delta)^2) \\
&= \frac{1}{4c^2}(\sqrt{4cN + g(\delta)^2} + g(\delta))^2 \\
&= \left(\frac{1}{2c}(g(\delta + \sqrt{4cN + g(\delta)^2}))\right)^2
\end{aligned}$$

so $\sqrt{n} > \frac{1}{2c}(g(\delta + \sqrt{4cN + g(\delta)^2}))$, and thus the desired inequality holds □

Lemma 2. *Let $B_r^D = \{x \in \mathbb{R}^D : \|x\| < r\}$ be the ball of radius r centered at the origin in $\mathbb{R}^D$ and let $K = [a_1, b_1] \times \cdots \times [a_D, b_D]$ be a rectangle in $\mathbb{R}^d$ satisfying $a_i < 0$, $b_i > 0$, and $b_i - a_i > 2r$ for all $i = 1, \ldots, D$. Then $\mathrm{Vol}(B_r^D \cap K) \geq 2^{-D} \mathrm{Vol}(B_r)$.*

Proof. For each i, we must have either $a_i \leq -r$ or $b_i > r$ (otherwise we would have $b_i - a_i < 2r$ which contradicts the conditions of the lemma). Without loss of generality, suppose that $b_i > 0$ for all i. If this were not the case, we could replace $[a_i, b_i]$ with $[-b_i, -a_i]$ in the definition of K and the volume of $B_r^D \cap K$ would remain the same. It follows that $[0, r] \subset [a_i, b_i]$ for $i = 1, \ldots, D$. Now we have

$$\begin{aligned}
\mathrm{Vol}(B_r^D \cap K) &= \int_{a_1}^{b_1} \cdots \int_{a_D}^{b_D} \mathbb{I}\{\|x\| \leq r\} \, dx_D \, \ldots \, dx_1 \\
&\geq \int_0^r \cdots \int_0^r \mathbb{I}\{\|x\| \leq r\} \, dx_D \, \ldots \, dx_1 \\
&= \frac{1}{2}\int_{-r}^r \cdots \frac{1}{2}\int_{-r}^r \mathbb{I}\{\|x\| \leq r\} \, dx_D \, \ldots \, dx_1 \\
&= 2^{-D} \mathrm{Vol}(B_r^D),
\end{aligned}$$

as required. In line 2 we used the fact that the map $x \mapsto \mathbb{I}\{\|x\| \leq r\}$ is non-negative, and in line 3 we used the fact that the map is symmetric along each dimension about the point 0. □

Proof of Theorem 1

Proof. The stopping rule in Algorithm 1 guarantees that at time τ we have $C_\tau < c$, which implies that $\mathbb{P}(M_\tau \geq c) \leq \mathbb{P}(M_\tau > C_\tau)$. Propositions 2 and 3 show that $\tau_{\min} < \tau < \tau_{\max}$ with probability one, therefore $\mathbb{P}(M_\tau \geq C_\tau) = \mathbb{P}(\tau_{\min} < \tau < \tau_{\max}, M_\tau > C_\tau)$. The event $\{\tau_{\min} < \tau < \tau_{\max}, M_\tau > C_\tau\}$ is a subset of the event $\{\exists i \in \{\lfloor \tau_{\min} \rfloor + 1, \ldots, \lceil \tau_{\max} \rceil - 1\} : M_t > C_t\}$ which, together with the union bound, gives

$$\begin{aligned}
\mathbb{P}(\tau_{\min} < \tau < \tau_{\max}, M_\tau > C_\tau) &\leq \mathbb{P}(\exists i \in \{\lfloor \tau_{\min} \rfloor + 1, \ldots, \lceil \tau_{\max} \rceil - 1\} : M_t > C_t) \\
&\leq \sum_{t=\lfloor \tau_{\min} \rfloor + 1}^{\lceil \tau_{\max} \rceil - 1} \mathbb{P}(M_t > C_t) \leq (\tau_{\max} - \tau_{\min})\delta,
\end{aligned}$$

where in the last line we used the fact that $\mathbb{P}(M_t > C_t) \leq \delta$ for all fixed times t, which follows from Theorem 9 of McAllester and Schapire (2000). Putting the above inequalities together and using the fact that $\mathbb{P}(M_\tau < c) = 1 - \mathbb{P}(M_\tau \geq c)$ gives

$$\mathbb{P}(M_\tau < c) \geq 1 - (\tau_{\max} - \tau_{\min})\delta.$$

Substituting the expressions for $\tau_{\min}$ and $\tau_{\max}$ from Propositions 2 and 3 completes the proof. □

Proof of Theorem 2

Proof. Proof by contradiction: Assume that

$$\mathbb{P}(f(\tilde{x}) \leq f^* + Lr) < 1 - \hat{\delta}$$

and let $A = \{\omega \in \Omega : M_\tau(\omega) < c\}$ be the set of outcomes where the missing mass M_τ at stopping time τ is strictly less than c. According to Theorem 1, this set has probability at least $p(A) \geq 1 - \hat{\delta}$. Then there is a set of outcomes with nonzero probability for which

$$f(\tilde{x}) > f^* + Lr \quad \text{and} \quad M_\tau < c.$$

Consider now such an outcome and let x^* be a global minimum ($f(x^*) = f(x)$) and $\mathcal{K} = \bigcup_{i \in \mathcal{F}_\tau^0} S_i$ be the regions of attractions that we have observed ($p(\mathcal{K}) = M_\tau$). For any point x' in observed regions of attractions $\mathcal{S} \setminus \mathcal{K}$, we know that $f(\tilde{x}) \leq f(x')$ and so $f(x^*) + Lr < f(x')$. By Lipschitz-continuity, we get that $\|x^* - x'\| > r$. Since, this is true for all $x' \in \mathcal{S} \setminus \mathcal{K}$, we get that $r \leq \inf_{x \in \mathcal{S} \setminus \mathcal{K}} \|x^* - x\|$.

Let $\mathcal{B}_r(x^*) = \{y \in \mathcal{R}^d : \|y - x^*\| \leq r\}$ be a d-dimensional hypersphere with center x^* and radius r. Then $\mathcal{B}_r(x^*) \cap \mathcal{S} \subseteq \mathcal{K}$ since $r \leq \inf_{x \in \mathcal{S} \setminus \mathcal{K}} \|x^* - x\|$. By shifting $\mathcal{S} = [l_1, u_1] \times \cdots \times [l_d, u_d]$ by x^*, we can apply Lemma 2 and get that the probability mass of the restricted hypersphere is

$$p(\mathcal{B}_r(x^*) \cap \mathcal{S}) = \text{Vol}(\mathcal{B}_r(x^*) \cap \mathcal{S}) \tag{7}$$

$$\geq 2^{-d} \text{Vol}(\mathcal{B}_r(x^*)) = 2^{-d} r^d \pi^{d/2} \Gamma(d/2 + 1)^{-1} = c. \tag{8}$$

The missing mass is therefore $M_\tau = p(\mathcal{K}) \geq p(\mathcal{B}_r(x^*) \cap \mathcal{S}) \geq c$, but we also have that $c > M_\tau$ in the events we consider, which is a contradiction. Therefore a nonzero set of outcomes where $f(\tilde{x}) > f^* + Lr$ and $M_\tau < c$ cannot exist and so $\mathbb{P}(f(\tilde{x}) \leq f^* + Lr) \geq 1 - \hat{\delta}$. □