SPRINT Multi-Objective Model Racing

Tiantian Zhang, Michael Georgiopoulos and Georgios C. Anagnostopoulos

zhangtt@knights.ucf.edu, michaelg@ucf.edu and georgio@fit.edu

Abstract

Multi-objective model selection, which is an important aspect of Machine Learning, refers to the problem of identifying a set of Pareto optimal models from a given ensemble of models. This paper proposes SPRINT-Race, a multi-objective racing algorithm based on the Sequential Probability Ratio Test with an Indifference Zone. In SPRINT-Race, a non-parametric ternary-decision sequential analogue of the sign test is adopted to identify pair-wise dominance and non-dominance relationship. In addition, a Bonferroni approach is employed to control the overall probability of any erroneous decisions. In the fixed confidence setting, SPRINT-Race tries to minimize the computational effort needed to achieve a predefined confidence about the quality of the returned models. The efficiency of SPRINT-Race is analyzed on artificially-constructed multi-objective model selection problems with known ground-truth. Moreover, SPRINT-Race is applied to identifying the Pareto optimal parameter settings of Ant Colony Optimization algorithms in the context of solving Traveling Salesman Problems. The experimental results confirm the advantages of SPRINT-Race for multi-objective model selection.

Keywords: Racing Algorithm, Model Selection, Multi-objective Optimization, Sequential Probability Ratio Test

1 Introduction

Given an ensemble of models, the task of Model Selection (MS) is to identify a subset of models that are optimal in terms of certain optimization criteria. There are a variety of MS problems, including feature selection, algorithm and learning strategy selection, hyper-parameter selection, etc.

The problem of Single-Objective Model Selection (SOMS), in which only one optimization criterion is considered, has received much attention in the literature of stochastic Multi-Armed Bandit (MAB) as *best arm(s) identification* problem [8, 5]. Among them, Racing Algorithms (RAs) are regarded as an elimination-type MAB, which tries to minimize the computational effort needed to achieve a predefined confidence about the quality of the returned models (fixed confidence setting). A RA is an iterative procedure that starts off with an initial ensemble of candidate models (*candidate pool*). The candidate models are evaluated by solving randomly sampled problem instances in a sequential manner and their performances are measured according to a single task-specific criterion of model optimality (e.g. prediction accuracy, convergence speed, etc). Under-performing models are eliminated as soon as sufficient statistical evidence is amassed. Consequently, a RA saves (unnecessary) computational cost spent on trying to exploit poor-performing models. In other words, a RA trades off some computational effort with the likelihood that the model(s) returned by the racing procedure is (are) indeed the optimal one(s).

Ever since the first RA was developed, RAs have been widely used in parameter tuning and configuration [2, 16], algorithm design [11], and some industrial applications [1]. Several leading RAs for SOMS have been proposed in the literature: Hoeffding Racing (HR) [17], BRACE [20], Bernstein Racing (BR) [19, 15], and F-Race [3].

However, in real-world Machine Learning (ML) applications, model selection is often multi-objective in nature [12]. In Multi-Task Learning, for instance, all task objectives should be optimized simultaneously. Moreover, in the context of evolutionary computation, computational cost is always considered as

an additional selection criterion besides the quality of the best solution obtained. Model selection in terms of multiple conflicting optimization criteria is referred to as Multi-objective Model Selection (MOMS). A straightforward way of solving a MOMS problem is to convert it to a SOMS problem, either by considering the conic linear combination or the Chebyshev scalarization of all the objectives, or by adopting other unary quality indicators [14] (e.g. hypervolume, epsilon indicator). However, Pareto optimality plays an important role in MOMS. Optimizing any single optimization objective does not provide any insight into the trade-offs among multiple conflicting objectives. The first Multi-objective Racing Algorithm (MORA), namely S-Race, which addresses the problem of MOMS in the sense of Pareto optimality was put forward by the seminal paper of Zhang et al. [25]. In S-Race, the probabilistic dominance relationship between two models is statistically inferred by the non-parametric sign test based on the observed performance vectors. Moreover, the total probability of not making any Type I errors (falsely rejecting the null hypothesis and inferring pair-wise dominance) is strictly controlled by Holm's step-down procedure, which is accomplished by adjusting the significance level of each individual hypothesis in multiple comparisons. In the same paper, S-Race was applied to selecting Support Vector Machines for binary-/ternary-classification. The obtained experimental results demonstrated that S-Race is an efficient and effective algorithm for automatic model selection.

Nevertheless, S-Race has certain limitations. First of all, in S-Race, unlike the total probability of not committing any Type I errors, the probability of making any Type II errors (falsely inferring the absence of dominance) is not strictly controlled. Moreover, the sign test adopted in S-Race can only identify dominance relationships, *i.e.*, whether model A dominates model B or the opposite. When two non-dominated models are compared and no dominance relation is discerned, the comparison continues until the available problem instances are exhausted. In other words, S-Race spends unnecessary computational cost on comparing non-dominated models. Most important of all, the sign test employed in S-Race is not an optimum test procedure, as explained in Section 2.2. If establishing dominance based on the sign test is good in terms of error probability, replacing it by the optimal test procedure should yield a MORA with improved overall sample complexity. Note that the sample complexity of a test comparing two models refers to the expectation of the final size of the sample required to reach a decision.

To overcome these limitations of S-Race, here we introduce SPRINT-Race, a MORA based on the Sequential Probability Ratio Test (SPRT) with an Indifference Zone. SPRINT-Race hinges on a ternary-decision, sequential analogue of the sign test. The comparison between two models terminates, when either dominance or non-dominance is established with certain confidence. As a result, SPRINT-Race is able to stop automatically, when no more sampling is needed for further comparisons. In addition, both Type I and Type II errors are strictly controlled via the SPRT. Therefore, SPRINT-Race is capable of strictly confining the error probability of returning dominated models and, simultaneously, abandoning non-dominated models at a predefined level. Moreover, the concept of an indifference zone is introduced in SPRINT-Race to aid in reducing the possibility of miscategorizing Pareto optimal models by chance.

SPRINT-Race infers dominance or lack thereof between a pair of models based on a ternary-decision SPRT procedure, namely dual-SPRT, which consists of two component SPRTs. Section 2 provides some necessary background on the SPRT and introduces the proposed SPRINT-Race procedure. Furthermore, the analysis of the overall error probability of SPRINT-Race, *i.e.*, the total probability of any false discoveries, is discussed in Section 2.4. The efficiency and effectiveness of SPRINT-Race for MOMS are evaluated via a series of experiments described in Section 3. First, the performance of SPRINT-Race is evaluated on artificially-constructed MOMS problems with known ground-truth solutions. Additionally, SPRINT-Race is applied to selecting Pareto optimal parameter settings of Ant Colony Optimization (ACO) algorithms in the context of Traveling Salesman Problems (TSPs). Overall, the experimental results confirm the potential advantages of SPRINT-Race for addressing MOMS problems. Finally, key findings are summarized in Section 4.

2 SPRINT-Race Description

SPRINT-Race addresses the problem of MOMS in the sense of Pareto optimality. The (non-)dominance relationship between a pair of models is determined via a ternary-decision SPRT procedure dubbed dual-SPRT. Moreover, the error probability of SPRINT-Race, meaning the total probability of falsely retaining any dominated model or removing any non-dominated model during the SPRINT-Race, is strictly controlled

at a user-specified level.

2.1 Problem Formulation

SPRINT-Race assumes the following setting: the decision maker is in possession of an ensemble of M models $\{C_1, C_2, \cdots, C_M\}$, which are to be compared according to D optimization objectives $\{f_i(C)\}_{i=1,\cdots,D}$. Without loss of generality, we will assume that all the objectives are to be maximized. Let $\mathbf{f}(C) \triangleq [f_1(C), f_2(C), \cdots, f_D(C)]$ be the performance vector of a model C. Following traditional conventions [6], $\mathbf{f}(C) \succ \mathbf{f}(C')$ is defined as follows

$$\begin{aligned} & f_i(C) \geqslant f_i(C') \quad \forall\, i \in \{1, 2, \cdots, D\} \\ & \text{and } \exists\, j \in \{1, 2, \cdots, D\} \mid f_j(C) > f_j(C') \end{aligned} \tag{1}$$

There are three possible outcomes of a pairwise comparison between model C and C', namely C dominates C', C' dominates C, and non-dominance. They are denoted as $C \succ C'$, $C \prec C'$ and $C \sim C'$, respectively. Given a model C, it is generally assumed that the objective value $f_i(C)$ is stochastic in nature. This assumption reflects the models' search process randomness or noisy performance measurements. Therefore, SPRINT-Race decides that $C \succ C'$ if $Pr\{\mathbf{f}(C) \succ \mathbf{f}(C')\} > \frac{1}{2}$; $C \prec C'$ if $Pr\{\mathbf{f}(C) \succ \mathbf{f}(C')\} < \frac{1}{2}$; and $C \sim C'$ otherwise. If there is no model that dominates model C, model C is defined as Pareto optimal. In the context of MOMS, in principle, one is interested in identifying the entire set of Pareto front models. However, if desired or if mandated by practicalities, additional subjective criteria might be employed by a decision maker to sub-sample the Pareto front and, potentially, identify a single model that reflects her/his particular preference.

2.2 Sequential Probability Ratio Test

For any testing procedure, let us denote the probability of a Type I error by $\alpha \in [0, 1]$ and the probability of a Type II error by $\beta \in [0, 1]$. Given α and β, the hypothesis test which minimizes the expected number of required samples, over the entire parameter space, is called the uniformly most efficient test of the given hypothesis test problem. It was shown that SPRT can be regarded as a locally most efficient test [23] in the sense that its expected sample complexity is minimized for some proper subset of the parameter space, for all practical purposes. Therefore, given a predefined maximum error probability, adopting SPRT for dominance inference in a MORA should result in a superior algorithm that only necessitates a near-minimum sample complexity.

Assume $\{x_1, x_2, \cdots, x_N\}$ is a sequence of independent and identically distributed observations sampled from an unknown distribution. Let $g_\theta(x_i)$ be the density of the i^{th} sample parametrized by a parameter θ. In a traditional fixed-sample test with two simple hypotheses, one of two possible decisions is made based on the observed random samples: accept the null hypothesis $H_0 : \theta = \theta_0$, or accept the alternative hypothesis $H_1 : \theta = \theta_1$. For a fixed-sample test, α is typically predetermined, but β, which is a function of α, is unknown and left uncontrolled. Therefore, given the sample size N and α, the test procedure that minimizes β is preferred; such a test is referred to as the uniformly most powerful test for the given hypothesis test problem and parameter space. The Neyman-Pearson lemma [10, p.52] implies that the likelihood ratio test yields the most powerful test for testing two simple hypotheses. To be more specific, the test statistic is computed as the ratio of the likelihood of the data under H_1 to their likelihood under H_0.

$$\lambda_N = \prod_{i=1}^{N} \frac{g(x_i|\theta_1)}{g(x_i|\theta_0)} \tag{2}$$

Distinct from fixed-sample testing, there is a third possible action in sequential testing: neither accept H_0 nor accept H_1 but continue sampling with the hope that gathering additional evidence may allow one to decide between the two hypotheses. The sample size of a sequential test is not predetermined and the samples are collected sequentially until either H_0 or H_1 is accepted. SPRT [23] is the first sequential testing procedure proposed in the literature and it proceeds as follows: assume λ_t is the test statistic at step t. Then, if $\lambda_t \leq B$, H_0 is accepted; if $\lambda_t \geq A$, H_1 is accepted; if $B \leq \lambda_t \leq A$, no decision is made and sampling resumes. As suggested in [23], by choosing $A = \frac{1-\beta}{\alpha}$ and $B = \frac{\beta}{1-\alpha}$, the probabilities of Type I and Type II errors does not exceed α and β respectively.

2.3 Dominance/Non-Dominance Inference

Due to the stochastic nature of the performance vectors, the dominance and non-dominance relationship between a pair of models can only be established via a formal test of hypothesis. A pair-wise and non-parametric test procedure is preferred due to its robustness. In a pair-wise test, each sample consists of paired observations, which effectively reduces the variability caused by external factors (e.g. differences between problem instances). Moreover, non-parametric tests have the advantage of being free of distributional assumptions. Therefore, a sequential analogue of the pair-wise non-parametric sign test [18] is adopted in SPRINT-Race to establish dominance and non-dominance between a pair of candidate models. Its motivation is straightforward: if a racing based on the sign test is good in terms of error probabilities, replacing it by the local optimal SPRT should improve it even further in terms of reducing the overall sample complexity.

Assume that a pair of candidate models C_i and C_j is compared, whose performances have been evaluated on a series of problem instances. Let N_{ij} (a random variable) denote the number of times that C_i dominates C_j and let n_{ij} be its observed value. Moreover, let $S \triangleq N_{ij} + N_{ij}$ and let its observed value be $s \triangleq n_{ij} + n_{ij}$. If $p \triangleq Pr\{N_{ij} = 1 | S = 1\}$, then, obviously, $N_{ij}|\{S = s\} \sim \text{Binomial}(s, p)$. Therefore, the relationship between C_i and C_j is established in the following manner: if $p < \frac{1}{2}$, C_i is dominated by C_j; if $p = \frac{1}{2}$, C_i and C_j are non-dominated to each other; and if $p > \frac{1}{2}$, C_i dominates C_j. In other words, the problem of inferring dominance and non-dominance between a pair of models translates to making a decision among three mutually exclusive and exhaustive hypotheses

$$H_0 : p < 1/2 \quad H_1 : p = 1/2 \quad H_2 : p > 1/2 \tag{3}$$

Subsequently, if H_0 is accepted, model C_i will be eliminated from racing; if H_2 is accepted, model C_j will eliminated; if H_1 is accepted, both C_i and C_j will be retained by the racing procedure.

Moreover, the concept of indifference zone is introduced for practical considerations: i) in real world applications, a thorough investigation is usually uneconomical and one is willing to take near-optimal decisions; and ii) it mitigates the possibility of omitting Pareto optimal models due to inaccurate performance measurements (e.g. noise). More specifically, in SPRINT-Race, the following three hypotheses are considered:

$$H_0 : p \leq 1/2 - \delta \quad H_1 : p = 1/2 \quad H_2 : p \geq 1/2 + \delta \tag{4}$$

where $\delta \in (0, 1/2)$ is chosen by the decision maker. The intervals $(1/2 - \delta, 1/2)$ and $(1/2, 1/2 + \delta)$ will be referred to as indifference zones. When $p \in (1/2 - \delta, 1/2)$, we assume that we have no strong preference between H_0 and H_1, but the rejection of H_2 is strongly preferred. Similarly when $p \in (1/2, 1/2 + \delta)$, no error is committed if either H_1 or H_2 is considered. Note that the selection of δ is not a statistical problem, but should be made based on practical concerns.

The previously described ternary-decision test, referred to as dual-SPRT, is constructed by combining two component binary-hypothesis SPRT [22] as shown in Equation (5). The decision procedure of the dual-SPRT is summarized in Table 1.

$$\begin{array}{lll} \textbf{SPRT}^{\mathbf{1}} & H_0^1 : p \leq \frac{1}{2} - \delta & H_1^1 : p \geq \frac{1}{2} \\ \textbf{SPRT}^{\mathbf{2}} & H_0^2 : p \leq \frac{1}{2} & H_1^2 : p \geq \frac{1}{2} + \delta \end{array} \tag{5}$$

Table 1: **Testing procedure of dual-SPRT**

$SPRT^1$ accepts	$SPRT^2$ accepts	dual-SPRT accepts
$H_0^1 : p \leq \frac{1}{2} - \delta$	$H_0^2 : p \leq \frac{1}{2}$	$H_0 : p \leq \frac{1}{2} - \delta$
$H_1^1 : p \geq \frac{1}{2}$	$H_0^2 : p \leq \frac{1}{2}$	$H_1 : p = \frac{1}{2}$
$H_1^1 : p \geq \frac{1}{2}$	$H_1^2 : p \geq \frac{1}{2} + \delta$	$H_2 : p \geq \frac{1}{2} + \delta$

Due to the monotonic likelihood ratio property of the binomial distribution, the component SPRTs described in Equation (5) are equivalent to the following tests

$$\begin{array}{lll} \textbf{SPRT}^{\mathbf{1}} & H_0^1 : p = \frac{1}{2} - \delta & H_1^1 : p = \frac{1}{2} \\ \textbf{SPRT}^{\mathbf{2}} & H_0^2 : p = \frac{1}{2} & H_1^2 : p = \frac{1}{2} + \delta \end{array} \tag{6}$$

Assume that, at the t^{th} step, there are n_{ij}^t times that C_i dominates C_j and n_{ji}^t times that C_j dominates C_i. In each component SPRT with $H_0 : p = p_0$ and $H_1 : p = p_1$, the test statistic is calculated as follows:

$$\lambda_t = \frac{p_1^{n_{ij}^t}(1-p_1)^{n_{ji}^t}}{p_0^{n_{ij}^t}(1-p_0)^{n_{ji}^t}} \tag{7}$$

which leads to the following simple rules

$$\begin{cases} \textbf{if } n_{ij}^t \leq \frac{\log B}{r-\bar{r}} - (n_{ij}^t + n_{ji}^t)\frac{\bar{r}}{r-\bar{r}}, \textbf{ accept } H_0 \\ \textbf{if } n_{ij}^t \geq \frac{\log A}{r-\bar{r}} - (n_{ij}^t + n_{ji}^t)\frac{\bar{r}}{r-\bar{r}}, \textbf{ accept } H_1 \\ \quad \textbf{otherwise}, \textbf{ continue sampling} \end{cases} \tag{8}$$

where $A \triangleq \frac{1-\beta}{\alpha}$, $B \triangleq \frac{\beta}{1-\alpha}$, $r \triangleq \log \frac{p_1}{p_0}$, and $\bar{r} \triangleq \log \frac{1-p_1}{1-p_0}$. Note that SPRINT-Race's two component SPRTs utilizes common α and β values in order to reduce the number of parameters that need to be specified.

2.4 SPRINT-Race and its Analysis

The whole SPRINT-Race procedure is illustrated in Algorithm 1. Initially, a dual-SPRT is established for each pair of models, which means that a total of $\binom{M}{2}$ dual-SPRTs are initialized at the beginning of the race for M candidate models. During a single step of the race, a problem instance is randomly chosen from the problem's sampling space. The performances of the two models involved in each active dual-SPRT are evaluated on the selected problem instance, and the resulting performance vectors are utilized to generate a new sample for the corresponding dual-SPRT. The test statistics are computed and, hence, decisions are made according to Equation (8). If H_0 is accepted, C_i is identified as being dominated by C_j and, thus, will be eliminated from the race. Consequently, all the dual-SPRTs involving C_i are stopped. On the other hand, if H_2 is accepted, C_i is identified as dominating C_j and, therefore, C_j is removed from the race. As a result, all the dual-SPRTs containing C_j are stopped. If H_1 is accepted, C_i and C_j are regarded as neither dominating the other. The current dual-SPRT will be terminated since no more comparisons are needed to verify the dominance relation between C_i and C_j. Otherwise, if no decision is made, the relevant dual-SPRT will be reapplied in the next step. When all the dual-SPRTs are terminated, the race concludes.

For SPRINT-Race, being a fixed confidence model selection algorithm, the total probability of making any erroneous discovery needs to be strictly controlled. In a ternary-decision dual-SPRT, there are multiple erroneous decisions that can be made. Hence, a more careful analysis is warranted, when compared to the case of a traditional binary-decision SPRT, which gives rise only to Type I and Type II errors. Similar to the analysis in [4, 22], the probability of any incorrect decision of the dual-SPRT procedure, denoted by $\gamma(p)$, is provided in Table 2.

Table 2: **Error probability analysis of dual-SPRT**

Interval	**Wrong Decisions**	$\boldsymbol{\gamma(p)}$
$p \leq \frac{1}{2} - \delta$	accept H_1 or H_2	$\gamma(p) \leq \alpha$
$\frac{1}{2} - \delta < p < \frac{1}{2}$	accept H_2	$\gamma(p) < \alpha$
$p = \frac{1}{2}$	accept H_0 or H_2	$\gamma(p) \approx \alpha + \beta$
$\frac{1}{2} < p < \frac{1}{2} + \delta$	accept H_0	$\gamma(p) < \beta$
$p \geq \frac{1}{2} + \delta$	accept H_0 or H_1	$\gamma(p) \leq \beta$

The maximum $\gamma(p)$ over the interval $[0, 1]$, denoted by γ^*, is called the true level of significance of the dual-SPRT. According to Table 2, it follows that $\gamma^* \triangleq \max_{p\in[0,1]} \gamma(p) = \max\{\alpha, \alpha + \beta, \beta\} = \alpha + \beta$.

We define the error probability Γ of SPRINT-Race to be the overall probability of falsely removing any Pareto optimal model or failing to eliminate any dominated model. Obviously, Γ is dependent on the true significance levels of a total of $\binom{M}{2}$ dual-SPRTs, where M is the size of the initial candidate pool. Using the Bonferroni inequality, we have

$$\Gamma \leq \sum_{i=1}^{\binom{M}{2}} \gamma_i^* = \sum_{i=1}^{\binom{M}{2}} (\alpha_i + \beta_i) \tag{9}$$

Algorithm 1 SPRINT-Race Pseudo-code

Initialize $Pool \leftarrow \{C_1, C_2, \cdots, C_m\}$ $(m \geq 2)$
Initialize $t = 1$
repeat
  Randomly sample a problem instance from the problem pool
  for each model $C_i \in Pool$ **do**
    for each model $C_j \in Pool \quad s.t. \ i < j$ **do**
      if the corresponding dual-SPRT continues **then**
        Evaluate C_i and C_j on the selected instance
        Update n_{ij}^t and n_{ji}^t
        if H_0 is accepted **then**
          $Pool \leftarrow Pool \setminus \{C_i\}$
          Stop all dual-SPRTs involving C_i
        else if H_2 is accepted **then**
          $Pool \leftarrow Pool \setminus \{C_j\}$
          Stop all dual-SPRTs involving C_j
        else if H_1 is accepted **then**
          Stop the dual-SPRT involving C_i and C_j
        end if
      end if
    end for
  end for
  $t = t + 1$
until All dual-SPRTs are terminated
return *Pool*

where α_i and β_i refers to the α and β values assigned for the i^{th} dual-SPRT in SPRINT-Race. To further reduce the number of parameters involved in SPRINT-Race, we set $\alpha_i = \beta_i = \epsilon$ for all $i = 1, 2, \cdots, \binom{M}{2}$. Then, Equation (9) reduces to

$$\Gamma \leq 2\binom{M}{2}\epsilon = M(M-1)\epsilon \tag{10}$$

Equation (10) suggests a way of strongly controlling the error probability of SPRINT-Race by properly assigning the ϵ value. Let Γ_{max} denote the maximum error probability allowed for SPRINT-Race. From Equation (10), we have

$$\epsilon = \frac{\Gamma_{max}}{M(M-1)}. \tag{11}$$

Obviously, the smaller Γ_{max} is, the smaller the ϵ value will be. Consequently, more samples are required to reach a decision in each component SPRT. On the contrary, larger Γ_{max} allows for less computational effort, but also more error. Therefore, the selection of the desired Γ_{max} value represents a trade-off between the probability of returning a final ensemble of models that matches as close as possible to the true Pareto front and the computational effort exerted by SPRINT-Race.

3 Experiments

In this section, the performance of SPRINT-Race [1] is first investigated by selecting the Pareto optimal models based on artificially-generated data. To further demonstrate its performance, SPRINT-Race was applied to selecting the Pareto optimal parameter settings of ACO algorithm for solving TSPs. Note that SPRINT-Race and S-Race are incomparable because (i) the concept of indifference zone of SPRINT-Race is

[1] `MATLAB`® code of SPRINT-Race is available at https://github.com/watera427/SPRINT-Race_v1. If you have any queries, please contact the primary author via email.

not employed in S-Race, (ii) the maximum number of steps is known in S-Race but unknown in SPRINT-Race, and (iii) the probability of making any Type II errors is not strictly controlled by S-Race. Therefore, it is meaningless to compare SPRINT-Race and S-Race in neither the fixed confidence setting nor the fixed sample complexity setting.

3.1 Performance Metrics

Two performance metrics were considered to measure the quality of SPRINT-Race, *retention* R and *excess* E [25], which are defined as follows:

$$R \triangleq \frac{|\mathcal{P}_R \cap \mathcal{P}_{PF}|}{|\mathcal{P}_{PF}|} \tag{12}$$

$$E \triangleq \frac{|\mathcal{P}_R \setminus \mathcal{P}_{PF}|}{|\mathcal{P}_R|} = 1 - \frac{|\mathcal{P}_R \cap \mathcal{P}_{PF}|}{|\mathcal{P}_R|} \tag{13}$$

where $\mathcal{P}_R$ is the set of models returned by SPRINT-Race, and $\mathcal{P}_{PF}$ is the ensemble of models constituting the true Pareto front.

Just as their names imply, R measures SPRINT-Race's ability of retaining Pareto optimal models. Meanwhile, E measures its ability of identifying and eliminating dominated models that do not belong to $\mathcal{P}_{PF}$. Ideally, we would like to have $R = 1$ and $E = 0$, which means SPRINT-Race is able to return exactly the ensemble of Pareto front models. However, in practice, a SPRINT-Race yielding high R and low E values is acceptable, when contrasted to the savings in computational cost. As discussed in Section 2.4, the R and E values are dependent on Γ_{max}, the predefined maximum error probability of SPRINT-Race.

R and E focus on the quality of the final ensemble identified by SPRINT-Race. Aside from these quantities, the sample complexity T, which is measured as the total number of samples used, is also considered as a performance measurement of SPRINT-Race. Note that the number of samples used by a dual-SPRT in SPRINT-Race equals the total number of times the candidate models are evaluated.
Regarding the computational complexity of SPRINT-Race, it is easy to show that SPRINT-Race takes $O(D\ \overline{V}\ \overline{M}^2)$ pair-wise comparisons for D-objectives model selection, where $\overline{M}$ is the average number of models competing throughout the entire race, and $\overline{V}$ is the average sample complexity of a dual-SPRT. In reality since pair-wise comparisons between models will be terminated once sufficient statistical evidence is collected, it is expected that $\overline{M} \ll M$ and $\overline{V} \ll V$, where V is the sample complexity of a corresponding ternary-decision fixed-sample test for dominance and non-dominance relationship inference.

3.2 Artificially Constructed MOMS Problems

We consider a few simple experiments here to illustrate the efficiency of SPRINT-Race. Assume that there are M initial models for each experiment of D objectives, represented by M randomly generated D-dimensional vectors. Subsequently, $\binom{M}{2}$ Bernoulli distributions are constructed and all relevant p values are stored in a matrix P. $P_{i,j}(j > i)$ is drawn from a uniform distribution on the open interval (0.5,1), if the i^{th} D-dimensional vector dominates the j^{th} D-dimensional vector. If neither dominates the other, $P_{i,j}$ is set to 0.5. Otherwise, $P_{i,j}$ is selected from $(0, 0.5)$, where $P_{i,j} = 1 - P_{i,j}$. Finally, $P_{i,j}$ is set to 0.5, if $i = j$. This way, $\mathcal{P}_{PF}$, the Pareto front models, are known in advance. After P is constructed, SPRINT-Race commences and, whenever a new sample is needed to be used for a dual-SPRT, the new sample is randomly generated from a Bernoulli distribution with the corresponding $P_{i,j}$ entry of P.

3.2.1 Impact of Number of Objectives

In this set of experiments, we fixed M at 100, δ at 0.05 and $\Gamma_{max} = 0.1$, but varied the number of objectives D from 2 to 14. Each experiment was repeated for 30 runs. The average R and E values for SPRINT-Race are presented in Table 3. Moreover, the average sample complexity T and the average Pareto front size $|\mathcal{P}_{PF}|$ are presented as well.

The corresponding R values are all close to 1, which means that all Pareto optimal models are kept and returned by the racing procedure. From Table 3, we also observe that the E values are all smaller than 0.1, indicating that SPRINT-Race is able to successfully identify and eliminate almost all dominated models

Table 3: **Average retention R, excess E, sample complexity T and the size of the Pareto front $|\mathcal{P}_{PF}|$ for varying D values over 30 runs**

D	R	E	T	$\lvert\mathcal{P}_{\mathbf{PF}}\rvert$
2	1.000	0.075	1.65e4	5.4
3	1.000	0.061	3.13e4	15.1
4	1.000	0.053	5.25e4	29.05
5	1.000	0.040	7.47e4	44.05
6	1.000	0.027	1.00e5	60.85
7	1.000	0.019	1.18e5	71.55
8	0.999	0.010	1.31e5	81.75
9	0.999	0.010	1.44e5	90.00
10	0.998	0.008	1.46e5	92.55
11	0.999	0.003	1.52e5	96.25
12	1.000	0.002	1.52e5	97.45
13	0.999	0.001	1.55e5	99.00
14	1.000	0.001	1.55e5	99.55

based on the predefined confidence level. Hence, SPRINT-Race returns almost exactly the problem's true Pareto front. As defined in Section 3.1, $1-R$ measures the probability of falsely removing any non-dominated models in SPRINT-Race, while E measures the probability of mistakenly retaining any dominated models in SPRINT-Race. Therefore, as we discussed before, the desired significant level Γ_{max} of SPRINT-Race, which was set to be 0.1 in this experiment, implicitly controls $1 - R + E$. It is observed that the sum of $1 - R$ and E is strictly below 0.1 as expected. However, most of them are far less than 0.1, implying that SPRINT-Race may be too conservative.

From Table 3, we also notice that when the number of objectives increases, more models are deemed Pareto optimal, as indicated by $|\mathcal{P}_{PF}|$. When $D = 14$, almost all the models are on the Pareto front; this explains why E values decrease in this case. Also, T increases monotonically with growing D. Generally, in SPRINT-Race, it takes longer to identify non-dominated models than dominated ones. When most of the models are non-dominated, it is expected that SPRINT-Race will require more samples to discover non-dominant ones. However, the growth rate of T as a function of D is not dramatic, since, once sufficient statistical evidence is collected indicating a non-dominance relation, no more sampling is needed. In Figure 1, the sample complexity of each step in several runs of SPRINT-Race with $D = 2$ is depicted. As generally observed, the sample complexity falls rapidly between the 100^{th} step and the 190^{th} step. After about 200 steps, only less than 10% of the initial models are still racing for a fine comparison.

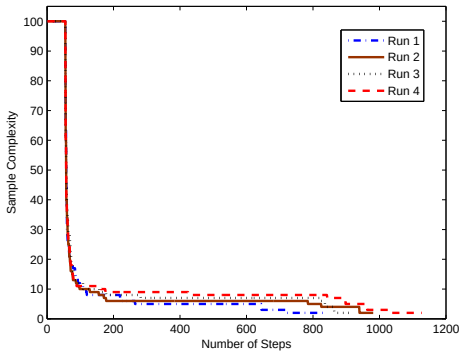

Figure 1: **The number of samples needed at each step in several runs of SPRINT-Race with binary-objective**

3.2.2 Impact of δ Values

The value of δ determines the size of the indifference zone of each ternary-decision SPRT in SPRINT-Race. In this set of experiments, we aim at understanding how δ influences the performance of SPRINT-Race. The parameter settings were $M = 10$, $\Gamma_{max} = 0.1$, $\delta \in \{0.01, 0.02, \ldots, 0.1\}$ and $D \in \{2, 3\}$ since binary-objective and ternary-objective MS are common in real world problem settings. Note that each experiment was repeated for 30 runs.

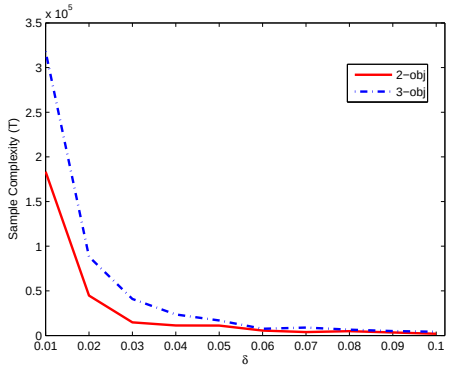

Figure 2: **Changes of sample complexity T with varying δ**

It is observed that δ has a significant impact on the sample complexity T (see Figure 2). In Figure 2, the average values of T over 30 runs are depicted for the 2-objective and 3-objective MS problem. The sample complexity of SPRINT-Race decreases with increasing δ. Such trend is expected because larger values of δ result in wider indifference zone and, when the indifference zone is broad, it is easier for the test procedure to reach a decision. As noticed in Figure 3, the error of SPRINT-Race, measured by $1 - R + E$, grows slightly with increasing δ. Because when δ grows, more dominated models will be regarded as Pareto optimal. As a result, the distinction between $\mathcal{P}_R$ of SPRINT-Race and $\mathcal{P}_{PF}$ becomes more significant.

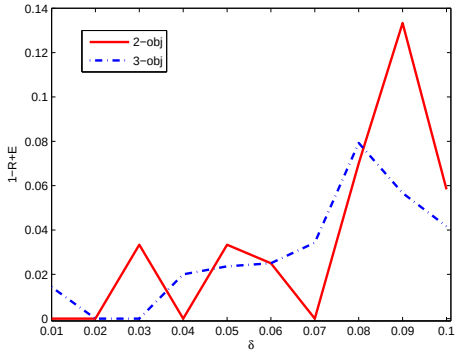

Figure 3: **Changes of $1 - R + E$ with varying δ**

3.2.3 Impact of Γ_{max} Values

In this set of experiments, the impact of the Γ_{max} value to the performance of SPRINT-Race was studied. The parameter settings were $M = 10$, $\Gamma_{max} \in \{0.01, 0.05, 0.1, 0.15, 0.2\}$, $D \in \{2, 3\}$ and $\delta = 0.01$. As the experimental results demonstrate, decreasing Γ_{max} will definitely increase the computational cost of SPRINT-Race, since more samples are required to infer any pair-wise model dominance or lack thereof. However, the R and E values only slightly varied within a reasonable range, which reflects SPRINT-Race's conservativeness.

In conclusion, the selection of δ and Γ_{max}, which are the only parameters of SPRINT-Race, results in a trade-off between the probability of returning a final ensemble of models that matches the true Pareto front and the computational effort exerted by SPRINT-Race. There is no generally-accepted "optimal" setting of the two aforementioned parameters; the determination of the exact δ and Γ_{max} values is totally depended on the user's preference pertaining to this trade-off.

3.3 ACO Selection for TSPs

In this section, we experimentally evaluated the performance of SPRINT-Race in terms of selecting the Pareto optimal parameter settings of ACO as TSP solvers. The ACO system [7] was first introduced in 1992 and is one of the most popular swarm intelligence algorithms. Ever since its inception, ACO and its variations have been widely used in a variety of combinatorial optimization problems, including scheduling problems, vehicle routing problems, assignment problems and set cover problems. Just like other swarm intelligence algorithms, the performance of ACO depends largely on its parameter settings, and, consequently, the effect of its parameters has been studied extensively in the literature of ACO [9, 21]. However, the parameter tuning process of ACO is time-consuming, when considering ever larger ensembles of parameter settings assessed on ever larger problem sets. RAs play an important role in ACO selection [2, 3] to maintain a certain level of confidence in retaining the optimal ACOs, while mitigating the computational burden. On the other hand, the TSP is one of the most famous NP-hard combinatorial optimization problems, it has been extensively studied in the literature and often serves as a standard benchmark problem. In this work, a bi-objective ACO parameter selection problem was considered. In specific, the task in question is to identify Pareto optimal ACOs, which serve as TSP solvers. An ACO parameter setting is deemed optimal, if it minimizes both the TSP tour length and the actual computation time to find this tour.

In this experiment, the `ACOTSPJava` [24] software was used, in which several ACO algorithms are implemented for solving TSPs. Prior to the race, a pool of 125 candidate models were initialized with diverse configurations in terms of different combinations of three parameters as shown in Table 4: i) α_{ACO}, the influence of pheromone trials; ii) β_{ACO}, the influence of heuristic information; and iii) ρ_{ACO}, the pheromone trail evaporation rate. The other parameters were set to the default values used in `ACOTSPJava`. Moreover, the TSP instances were all generated by the DIMACS TSP instance generator [13]. At each step, a random TSP instance was generated and the performance of each remaining models was evaluated by solving the generated TSP problem. Correspondingly, a new performance vector of each remaining model was collected, containing the length of the best tour found and the time spent of finding the best tour. SPRINT-Race was applied at each step, aiming at removing dominated models and stopping unnecessary comparison of a pair of Pareto optimal models as early as possible. 30 races were performed using $\Gamma_{max} = 0.01$ and $\delta = 0.05$.

Table 4: **ACO Parameter Description**

parameter	values
α_{ACO}	$\{0.01, 1.01, 2.01, 3.01, 4.01\}$
β_{ACO}	$\{0.1, 2.1, 4.1, 6.1, 8.1\}$
ρ_{ACO}	$\{0.1, 0.3, 0.5, 0.7, 0.9\}$

The resulting average R, E, T and $|\mathcal{P}_R|$ values over 30 runs were 1.000, 0.001, 4.46e5 and 17.933, respectively. It is observed that the R value is about 1 and the E value is close to 0, which illustrates that SPRINT-Race is able to retain exactly the ensemble of Pareto front models. Note that in this experiment the true $\mathcal{P}_{PF}$ is unknown. So the probability of dominance between pairs of models was estimated based on the collected performance vectors and the resulting ensemble of non-dominated models were regarded as $\mathcal{P}_{PF}$. To illustrate the accuracy of SPRINT-Race, we depicted the minimum probability of dominance of each model in Figure 4 of one run. For each model C_i, the minimum probability of dominance is calculated as $\min_{i \neq j} Pr\{C_i \succ C_j\}$. Considering $\delta = 0.05$, the minimum probability of dominance of any non-dominated models returned by SPRINT-Race should be no smaller than 0.45. Moreover, all the models with the minimum probability of dominance larger than 0.5 are expected to be returned by SPRINT-Race as Pareto optimal models. In Figure 4, a blue circle ($\circ$) represents a dominated model and a red bullet ($\bullet$) stands for a non-dominated model returned by SPRINT-Race. As shown in Figure 4, SPRINT-Race performs well

and returns all the Pareto optimal models as expected without erroneously including any dominated model. Furthermore, we compared the performance of the Pareto optimal models selected by SPRINT-Race and the

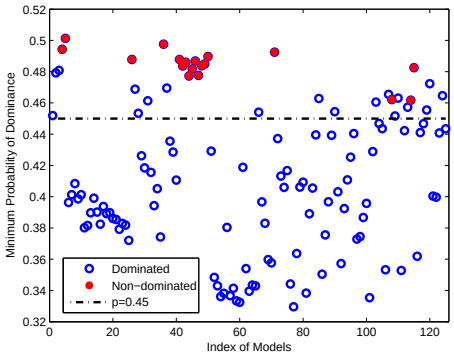

Figure 4: **Models' Minimum Probability of Dominance**

rest of models, which were identified as dominated models on a test set containing 2500 randomly generated TSP instances. The average normalized values of their performance vectors are displayed in Figure 5. It is observed that the Pareto optimal models are concentrated at the bottom left and right corners, where either the first objective or the second objective is minimized. In other words, the experimental results demonstrate that the Pareto optimal models selected based on the validation set of TSP instances are also Pareto optimal for the unseen set of TSP instances.

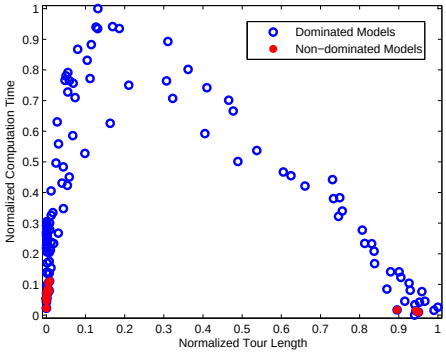

Figure 5: **Performance comparison of dominated and non-dominated models returned by SPRINT-Race on a test set**

4 Conclusions

In this paper, a new Multi-objective Racing Algorithm (MORA) for Multi-objective Model Selection (MOMS), na-med SPRINT-Race, was proffered. SPRINT-Race addresses the problem of MOMS in the proper sense of Pareto optimality by employing probabilistic dominance. Identifying dominated and non-dominated models is accomplished via a ternary-decision process, that is the sequential analogue of the non-parametric sign test. Moreover, in the fixed confidence setting of SPRINT-Race, the total probability of falsely retaining any dominated model and removing any non-dominated model is strictly controlled at a user-specified level. The racing procedure automatically stops, when sufficient statistical evidence is collected to make decisions. A key characteristic of SPRINT-Race is that it is able to balance the need for retaining all Pareto optimal models with high probability and computational cost limitations.

Experimental results were provided to illustrate the efficiency and effectiveness of SPRINT-Race. These results show that SPRINT-Race is able to return almost exactly the true Pareto front but at a reduced cost. The impact of SPRINT-Race's parameter values on its performance was analyzed on model selection problems, whose Pareto optimal solution was known beforehand. It was observed that the selection of δ and Γ_{max} results in a trade-off between the probability of returning a final ensemble of models that matches the true Pareto front and the computational effort exerted by SPRINT-Race. In addition, SPRINT-Race was applied to selecting the Pareto optimal parameter settings of Ant Colony Optimization (ACO) algorithms for solving Traveling Salesman Problems (TSPs). SPRINT-Race performs well and retains all the Pareto optimal models as expected with little error probability of including any dominated model. Overall, the experimental results confirm the potential of SPRINT-Race in MOMS.

It is worth pointing out that, while our work was singularly focused on MOMS problems, SPRINT-Race is readily applicable to optimal model initialization and configuration problems along the lines investigated in [3]; it is our intention to examine this possibility in one of our future works.

5 Acknowledgments

T. Zhang acknowledges partial support from National Science Foundation (NSF) grants No.1200566 and No.0525429. Moreover, M. Georgiopoulos acknowledges partial support from NSF grants No.0525429, No.0963146, No.1200566 and No.1161228. Additionally, G. C. Anagnostopoulos acknowledges partial support from NSF grant No.1263011. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

[1] Sven Becker, Jens Gottlieb, and Thomas Stützle. Applications of racing algorithms: An industrial perspective. In *Artificial Evolution*, volume 3871, chapter Artificial Evolution, pages 271–283. Springer, 2005.

[2] M. Birattari, Z. Yuan, P. Balaprakash, and T. Stützle. Automated algorithm tuning using F-races: Recent developments. In *Proc. of the 8th Metaheuristics Int. Conf.*, 2009.

[3] Mauro Birattari, Zhi Yuan, Prasanna Balaprakash, and Thomas Stützle. F-race and iterated F-race: An overview. Technical report, IRIDIA, 2011.

[4] J. De Boer. Sequential test with three possible decisions for testing an unknown probability. *Appl. Sci. Res., Sect. B*, 3:249 – 259, 1954.

[5] Sbastien Bubeck, Tengyao Wang, and Nitin Viswanathan. Multiple identifications in multi-armed bandits. In *Proc. of the 25th Int. Conf. on Mach. Learn.*, 2013.

[6] K. Deb. *Multi-Objective Optimization Using Evolutionary Algorithms.* John Wiley & Sons, 2009.

[7] M. Dorigo. *Optimization, Learning and Natural Algorithm.* PhD thesis, Politecnico di Milano, 1992.

[8] Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Adv. Neural. Inf. Process. Syst. 25*, pages 3221 – 3229, 2012.

[9] Dorian Gaertner and Keith Clark. On optimal parameters for ant colony optimization algorithms. In *Proc. of the Int. Conf. on Artif. Intell.*, 2005.

[10] B. K. Ghosh. *Sequential Tests of Statistical Hypotheses.* Addison-Wesley Publishing Company, Inc., 1970.

[11] Verena Heidrich-Meisner and Christian Igel. Hoeffding and bernstein races for selecting policies in evolutionary direct policy search. In *Proc. of the 26th Annual Int. Conf. on Mach. Learn.*, pages 401–408, 2009.

[12] Y. Jin and B. Sendhoff. Pareto-based multi-objective machine learning: An overview and case studies. *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, 38:397 – 415, 2008.

[13] D. S. Johnson, L. A. McGeoch, C. Rego, and F. Glover. 8$^{\text{th}}$ DIMACS Implementation Challenge: The Traveling Salesman Problem. http://dimacs.rutgers.edu/Challenges/TSP/, 2001. Accessed: Mon 24$^{\text{th}}$ Aug, 2015.

[14] J. D. Knowles, L. Thiele, and E Zitzler. A tutorial on the performance assessment of stochastic multi-objective optimizers. Technical report, CENL, ETH Zurich, 2006.

[15] Po-Ling Loh and Sebastian Nowozin. Faster hoeffding racing: Bernstein races via jackknife estimates. In *Proc. 24th Int. Conf. on Algorithm Learn.*, pages 203–217, 2013.

[16] Manuel Lopez-Ibanez and Thomas Stützle. Automatic configuration of multi-objective ACO algorithms. In *Swarm Intelligence*, pages 95–106, 2010.

[17] Oded Maron and Andrew Moore. Hoeffding races: Accelerating model selection search for classification and function approximation. *Adv. Neural Inf. Process. Syst.*, 6:59–66, 1994.

[18] W. Mendenhall, D. D. Wackerly, and R. L. Scheaffer. Nonparametric statistics. *Math. Stat. with App.*, pages 674–679, 1989.

[19] V. Mnih and C. Szepesvari. Empirical bernstein stopping. In *Proc. of the 25th Int. Conf. on Mach. Learn.*, 2008.

[20] Andrew W. Moore. Efficient algorithms for minimizing cross validation error. In *Proc. of the 11th Int. Conf. on Mach. Learn.*, pages 190–198, 1994.

[21] P. Shunmugapriya, S. Kanmani, S. Devipriya, J. Archana, and J. Pushpa. Investigation on the effects of ACO parameters for feature selection and classification. *Adv. in Comm., Netw., and Comput.*, 108:136–145, 2012.

[22] Milton Sobel and Abraham Wald. A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *The Ann. of Math. Stat.*, 20:502–522, 1949.

[23] Abraham Wald. *Sequential Analysis*. Dover, 1948.

[24] Adrian Wilke. ACOTSPJava. http://adibaba.github.io/ACOTSPJava/. Accessed: Mon 24$^{\text{th}}$ Aug, 2015.

[25] Tiantian Zhang, Michael Georgiopoulos, and Georgios C. Anagnostopoulos. S-Race: A multi-objective racing algorithm. In *Proc. of the Genet. and Evol. Comput. Conf.*, pages 1565–1572, 2013.