Sampling Strategies in Ordinal Regression for Surrogate Assisted Evolutionary Optimization

Helga Ingimundardottir and Thomas Philip Runarsson
School of Engineering and Natural Sciences
University of Iceland, Reykjavik, Iceland
{hei2, tpr}@hi.is

***Abstract*—In evolutionary optimization surrogate models are commonly used when the evaluation of a fitness function is computationally expensive. Here the fitness of individuals are indirectly estimated by modeling their rank with respect to the current population by use of ordinal regression. This paper focuses on how to validate the goodness of fit for surrogate models during search and introduces a novel validation/updating policy for surrogate models, and is illustrated on classical numerical optimization functions for evolutionary computation. The study shows that for validation accuracy it is sufficient for the approximate ranking and true ranking of the training set to be sufficiently concordant or that only the potential parent individuals should be ranked consistently. Moreover, the new validation approach reduces the number of fitness evaluation needed, without a loss in performance.**

***Keywords*-surrogate models; ordinal regression; sampling; evolutionary optimization**

I. Introduction

Evolutionary optimization is a stochastic and direct search method where a population of individuals are searched in parallel. Typically only the full or partial ordering of these parallel search individuals is needed. For this reason an ordinal regression offers sufficiently detailed surrogates for evolutionary computation [1]. In this case there is no explicit fitness function defined, but rather an indirect method of evaluating whether one individual is preferable to another.

The current approach in fitness approximation for evolutionary computation involves building surrogate fitness models directly using regression. For a recent review of the state-of-the-art surrogate models see [2]–[5]. The fitness model is based on a set of evaluated solutions called the training set. The surrogate model is used to predict the fitness of candidate search individuals. Commonly a fraction of individuals are selected and evaluated within each generation (or over some number of generations [6]), added to the training set, and used for updating the surrogate. The goal is to reduce the number of costly true fitness evaluations while retaining a sufficiently accurate surrogate during evolution. When using ordinal regression a candidate search individual $\mathbf{x}_i$ is said to be preferred over $\mathbf{x}_j$ if $\mathbf{x}_i$ has a higher fitness than $\mathbf{x}_j$. The training set for the surrogate model is therefore composed of pairs of individuals $(\mathbf{x}_i, \mathbf{x}_j)_k$ and a corresponding label $t_k \in [1, -1]$, taking the value $+1$ (or -1) when $\mathbf{x}_i$ has a higher fitness than $\mathbf{x}_j$ (or vice versa). The direct fitness approximation approach does not make full use of the flexibility inherent in the ordering requirement. The technique used here for ordinal regression is kernel based and is described in section II and was first presented in [1]. The use of surrogate models and approximate ranking has made some headway, e.g. [7], however still remains relatively unexplored field of study.

The critical issue in generating surrogate models, for evolutionary strategy (ES) search [8], is the manner in which the training set is constructed. For example, in optimization it is not critical to model accurately regions of the search space with low fitness. It is, however, key to model accurately new search regions deemed potentially lucrative by the evolutionary search method. Furthermore, since the search itself is stochastic, perhaps the ranking need not to be that accurate. Indeed the best μ candidate individuals are commonly selected and the rest disregarded irrespective of their exact ranking.

In the literature new individuals are added to the training set from the new generation of unevaluated search individuals. This seems sensible since this is the population of individuals which need to be ranked. However, perhaps sampling a representative individual, for example the mean of the unevaluated search individuals, may also be useful in surrogate ranking. Typically, the unevaluated individuals are ranked using the current surrogate model and then the best of these are evaluated using the true expensive fitness function and added to the training set. Again, this seems sensible since we are not interesting in low fitness regions of the search space. Nevertheless, it remains unclear whether this is actually the case. Finally, there is the question of knowing when to stop, when is our surrogate sufficiently accurate? Is it necessary to add new search individuals to our training set at every search generation? What do we mean by sufficiently accurate? This paper describes some preliminary experiments with the aim of investigating some of these issues further.

In section III sampling methods, stopping criteria and model accuracy are discussed. Moreover, a strategy for updating the surrogate during search is presented and its effectiveness illustrated using CMA-ES on some numerical optimization functions in section IV. The paper concludes with discussion and summary in section V.

II. Ordinal Regression

Ordinal regression in evolutionary optimization has been previously presented in [1], but is given here for completeness. The ranking problem is specified by a set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell} \subset$

978-1-4577-1676-8/11/$26.00

$X \times Y$ of ℓ (solution, rank)-pairs, where $Y = \{r_1, \ldots, r_\ell\}$ is the outcome space with ordered ranks $r_1 > r_2, > \ldots > r_\ell$. Now consider the model space $\mathcal{H} = \{h(\cdot) : X \mapsto Y\}$ of mappings from solutions to ranks. Each such function h induces an ordering $\succ$ on the solutions by the following rule:

$$\mathbf{x}_i \succ \mathbf{x}_j \quad \Leftrightarrow \quad h(\mathbf{x}_i) > h(\mathbf{x}_j) \tag{1}$$

where the symbol $\succ$ denotes "is preferred to". In ordinal regression the task is to obtain function h that can for a given pair $(\mathbf{x}_i, y_i)$ and $(\mathbf{x}_j, y_j)$ distinguish between two different outcomes: $y_i > y_j$ and $y_j > y_i$. The task is, therefore, transformed into the problem of predicting the relative ordering of all possible pairs of examples [9], [10]. However, it is sufficient to consider only all possible pairs of adjacent ranks, see also [11] for yet an alternative formulation. The training set, composed of pairs, is then as follows:

$$S' = \left\{(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}), t_k = \text{sign}(y_k^{(1)} - y_k^{(2)})\right\}_{k=1}^{\ell'} \tag{2}$$

where $(y_k^{(1)} = r_i) \wedge (y_k^{(2)} = r_{i+1})$ (and vice versa $(y_k^{(1)} = r_{i+1}) \wedge (y_k^{(2)} = r_i)$) resulting in $\ell' = 2(\ell - 1)$ possible adjacently ranked training pairs. The rank difference is denoted by $t_k \in [-1, 1]$.

In order to generalize the technique to different solution data types and model spaces an implicit kernel-defined feature space with corresponding feature mapping ϕ is applied. Consider the feature vector $\phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \ldots, \phi_m(\mathbf{x})]^T \in \mathbb{R}^m$ where m is the number of features. Then the surrogate considered may be defined by a linear function in the kernel-defined feature space:

$$h(\mathbf{x}) = \sum_{i=1}^{m} w_i \phi_i(\mathbf{x}) = \left\langle \mathbf{w} \cdot \phi(\mathbf{x}) \right\rangle. \tag{3}$$

where $\mathbf{w} = [w_1, ..., w_m] \in \mathbb{R}^m$ has weight w_i corresponding to feature ϕ_i.

The aim now is to find a function h that encounters as few training errors as possible on S'. Applying the method of large margin rank boundaries of ordinal regression described in [9], the optimal $\mathbf{w}^*$ is determined by solving the following task:

$$\min_{\mathbf{w}} \quad \tfrac{1}{2}\left\langle \mathbf{w} \cdot \mathbf{w} \right\rangle + \tfrac{C}{2} \sum_{k=1}^{\ell'} \xi_k^2 \tag{4}$$

subject to $t_k \left\langle \mathbf{w} \cdot (\phi(\mathbf{x}_k^{(1)}) - \phi(\mathbf{x}_k^{(2)}) \right\rangle \geq 1 - \xi_k$ and $\xi_k \geq 0$, $k = 1, \ldots, \ell'$. The degree of constraint violation is given by the margin slack variable ξ_k and when greater than 1 the corresponding pair are incorrectly ranked. Note that

$$h(\mathbf{x}_i) - h(\mathbf{x}_j) = \left\langle \mathbf{w} \cdot (\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)) \right\rangle \tag{5}$$

and that minimizing $\left\langle \mathbf{w} \cdot \mathbf{w} \right\rangle$ maximizes the margin between rank boundaries, in our case the distance between adjacently ranked pair $h(\mathbf{x}^{(1)})$ and $h(\mathbf{x}^{(2)})$.

Furthermore, it is important to scale the features ϕ first as the evolutionary search zooms in on a particular region of the search space. A standard method of doing so is by scaling the training set such that all solutions are in some range, typically $[-1, 1]$. That is, scaled $\tilde{\phi}$ is

$$\tilde{\phi}_i = 2(\phi_i - \underline{\phi}_i)/(\overline{\phi}_i - \underline{\phi}_i) - 1 \quad i = 1, \ldots, m \tag{6}$$

where $\underline{\phi}_i$, $\overline{\phi}_i$ are the minimum and maximum i-th component of all feature vectors in the training set.

III. Sampling Methods and Improvements

In surrogate modeling, a small sample of training individuals of known fitness are needed to generate an initial surrogate. There after sampling is needed to be conducted for validating and updating the surrogate. Bearing in mind that there is generally a predefined maximum number of expensive function evaluations that can be made, the sampling of test individuals used for validating/updating the surrogate needs to be fruitful.

During evolution different regions of the space are sampled and as a consequence the surrogate ranking model may be insufficiently accurate for new regions of the search space, hence if the surrogate is not updated to reflect the original fitness function it is very probable that the ES converges to a false optimum. It is, therefore, of paramount importance to validate the surrogate during evolution. In the literature this is referred to as model management or evolution control [4].

The accuracy can be validated by generating test individuals in the new region, namely from the new candidate individuals generated at every generation of the ES by reproduction, recombination and mutation. The validation control can either be generation based, i.e. when the surrogate is converging, or individual-based, where at each generation some of the new candidate individuals are evaluated with the exact model and others are evaluated with the surrogate, see [4].

The selection of individuals to be evaluated exactly can be done randomly, however, in [12] it is reported that validating the accuracy of the ranking of potential parent individuals during evolution is most beneficial as they are critical for success. In particular, Kriging surrogate model has two main components: a drift function representing its global expected value of the true fitness function; and a covariance function representing a local influence for each data point on the model, see [13]. For Kriging models an "infill sampling criteria" is implemented by sampling the individuals which the surrogate believes to be in the vicinity of global optima, however in some cases individuals in uncertain areas are also explored, this is referred to as generalized expected improvement [14]. A performance indicator to which strategy should be focused on, i.e. following the global optima vs. getting rid of uncertainties, [15] suggests the distance between approximated optima and its real fitness value, however no obvious correlation between the two ranks could be concluded. Moreover, [13] compares 6 various sampling procedures for updating the training set using the Kriging model. Two main strategies are explored, mainly evaluating the entire candidate population or only a subset. Latter yielding a significantly fewer exact function evaluations and obtain similar goodness of fit. The former strategy mostly focuses on whether all, partial or none of

the training set should be replaced, and whether the outgoing training individuals should be the worst ranking ones (elitist) or chosen at random (universal), where the elitist perspective was considered more favorable. However, reevaluating a subset of the best ranked individuals w.r.t. the surrogate model with the exact fitness function yielded the greatest performance edge of the strategies explored.

When the training accuracy is 100% one way of evaluating the accuracy of the surrogate is through cross validation. The quality of the surrogate is measured as the rank correlation between the surrogate ranking and the true ranking on training data. Here Kendall's τ is used for this purpose [16]. Kendall's τ is computed using the relative ordering of the ranks of all $\ell(\ell-1)/2$ possible pairs. A pair is said to be concordant if the relative ranks of $h(\mathbf{x}_i)$ and $h(\mathbf{x}_j)$ are the same for $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$, otherwise they are discordant. Kendall's τ is the normalized difference in the number of concordant and discordant pairs, defined as follows,

$$\tau = \frac{C-D}{\sqrt{C+D+T(h)}\sqrt{C+D+T(f)}} \tag{7}$$

where C and D denote the number of concordant and discordant pairs, respectively, and T denotes number of ties. Two rankings are the same when $\tau = 1$, completely reversed if $\tau = -1$, and uncorrelated for $\tau \approx 0$.

The surrogate ranking validation and improvement strategy using ordinal regression is tested using a covariance matrix adaptation evolution strategy (CMA-ES) [17]. CMA-ES is a very efficient numerical optimization technique, however we still expect to reduce the number of function evaluations needed for search. In [1] the validation policy had to successfully rank all of the candidate individuals, i.e. until $\tau = 1$. If there is no limit to training size then updating the surrogate becomes too computationally expensive, hence the training size needs to be pruned to size to $\overline{\ell}$. In [1] the set was pruned to a size $\overline{\ell} = \lambda$ by omitting the oldest individuals first. These are quite stringent restrictions which can be improved upon. The pruning only considers the age of the individuals, however older individuals might still be of more interest than newer ones if their fitness ranks higher. Thus a more sophisticated way of pruning would be omitting the lowest ranking individuals first. Moreover, candidate individuals are generated randomly using a normal distribution, thus a pseudo individual representing their mean could be of interest as an indicator for the entire population, e.g. by validating this pseudo individual first could give information if the surrogate is outdated w.r.t. the current search space. Furthermore, the validation is only done on the candidate individuals for the current generation in ES where only the μ best ranked individuals will survive to become parents. In evolutionary computing one is interested in the accurate ranking of individuals generated in the neighborhood of parent individuals, hence for sufficient validation of the surrogate, only the μ best ranked individuals should be considered and evaluated, since all other individuals of lower rank will be disregarded in the next iteration of ES. Lastly, one should also investigate the frequency by which the model is validated, e.g. at each generation or every $K > 1$ generations or even have the need for validating adapt with time.

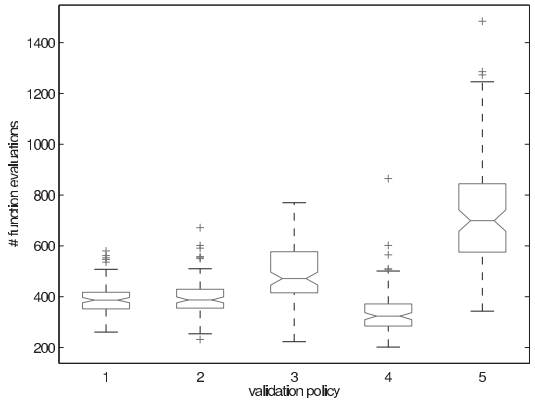

Fig. 1. Anova plot for different validation strategies: 1) prune old individuals, 2) prune bad individuals, 3) adding a pseudo mean candidate individual 4) correctly rank μ best ranked candidate individuals 5) update on every other generation for Rosenbrock's function for dimension $n = 2$

Preliminary tests were conducted on which validation method deemed fruitful, by implementing Rosenbrock's function of dimension $n = 2$, for 1) the setup presented in [1] and comparing it with the aforementioned validation improvements, which were added one at a time. Namely; 2) omitting the worst individuals during the pruning process, instead of the oldest ones; 3) initialize the validation process by using a pseudo individual that represents the mean of the new candidate individuals; 4) requiring that only the μ best candidate individuals are correctly ranked; and 5) validating on every other generation. Experimental results focusing on the number of function evaluations are shown in Fig. 1. There is no statistical difference between omitting oldest or worst ranked individuals from the training set, but this was expected, since both are believed to be representatives of a region of the search space which is no longer of interest. Adding the pseudo mean candidate individual didn't increase the performance edge. When the surrogate was updated on every other generation, it quickly became outdated and more than double function evaluations were needed to achieve the same rate of convergence. However, requiring the correct ranking for only the μ best ranked candidate individuals showed a significant performance edge.

If the training accuracy is not 100% then clearly $\tau < 1$. In this case additional training individuals would be forced for evaluation. However, enforcing a completely concordant ranking, i.e. $\tau = 1$, was deemed to be too strict due to the fact the search is stochastic. Thus the surrogate is said to be sufficiently accurate if $\tau > 0.999$.

Based on these preliminary tests, a pseudo code for the proposed model validation and improvement strategy is described in Fig. 2 where it is implemented at the end of each generation of CMA-ES. The algorithm essentially only evaluates the expensive true fitness function when the surrogate is believed

0 *Initialization*: Let $\mathcal{Y}$ denote current training set and its corresponding surrogate by h. Let $\mathcal{X}$ denote population of λ individuals of unknown fitness under inspection.
1 **for** $t := 1$ to λ **do** *(validate a test individual)*
2 Estimate ranking of $\mathcal{X}$ using h; denoted by $\bar{R}_0$.
3 $\mathbf{x}_B \leftarrow \max_{\mathbf{x}\in\mathcal{X}\setminus\mathcal{Y}} \{\bar{R}_0\}$ *(test individual)*.
4 Rank $\mathbf{x}_B$ w.r.t. individuals in $\mathcal{Y}$ using h; denoted by $\bar{R}$.
5 Evaluate $\mathbf{x}_B$ using true fitness function and evaluate its true rank among individuals in $\mathcal{Y}$; denoted by R.
6 $\mathcal{Y} \leftarrow \mathcal{Y} \cup \{\mathbf{x}_B\}$ *(add to training set)*.
7 Compare the rankings $\bar{R}$ and R by computing the rank correlation τ.
8 **if** $\tau > 0.999$ **then**
9 break (*model is sufficiently accurate*)
10 **fi**
11 Update the surrogate h using the new training set $\mathcal{Y}$.
12 **if** μ best individuals of $\bar{R}_0$ have been evaluated **then**
13 break (*model is sufficiently accurate*).
14 **fi**
15 **od**

Fig. 2. Sampling strategy to validate and improve surrogate models.

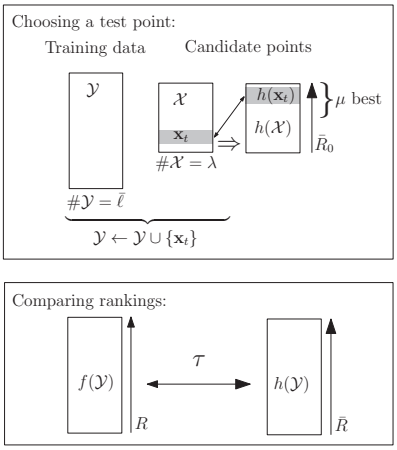

Fig. 3. Schema for the sampling strategy.

to have diverged. During each iteration of the validation process there are two sets of individuals, $\mathcal{Y}$ and $\mathcal{X}$, which are the training individuals which have been evaluated with the expensive model, and the candidate individuals (of unknown fitness) for the next iteration of CMA-ES, respectively. The test individuals of interest are those who are believed to become parent individuals in the next generation of CMA-ES, i.e. the μ best ranked candidate individuals according to the surrogate h. The method uses only a simple cross-validation on a single test individual, the one which the surrogate ranks the highest and has not yet been added to the training set. Creating more test individuals would be too costly, but plausible. Once a test individual has been evaluated it is added to the training set and the surrogate h is updated w.r.t. $\mathcal{Y}$, cf. Fig. 3. This is repeated until the surrogate is said to be sufficiently accurate, which occurs if either:

- Kendall's τ statistic between the ranking of the training set using the surrogate, $\bar{R}$, and its true ranking, R, is higher than 0.999, or
- μ best ranked candidate individuals w.r.t. the current surrogate have been added to the training set.

Note that during each update of the surrogate of the ranking of the μ best candidate individuals can change. Thus it is possible to evaluate more then μ test individuals during each validation.

Once the validation algorithm has completed, the training set is pruned to a size $\bar{\ell} = \lambda$ by omitting the lowest ranking individuals .

IV. Experimental Study

In the experimental study CMA-ES is run for several test functions, namely sphere model and Rosenbrock's function, of various dimensions $n = 2, 5, 10$ and 20. The average fitness for 100 independent runs versus the number of function evaluations is reported using the original validation procedure presented in [1] and compared with its new and improved validation procedure presented in Fig. 2, the procedures will be referred to as using "all" or only the "μ best" candidate individuals during the validation, respectively. The parameter setting for the (μ, λ) CMA-ES is as recommended in [17] with population size $\lambda = 4 + \lfloor 3\ln(n) \rfloor$ and the number of parents selected $\mu = \lambda/4$. The stopping criteria used are $1000n$ function evaluation or a fitness less than 10^{-10}. The initial mean search individual is generated from a uniform distribution between 0 and 1. It is also noted that the training set is only pruned to size $\bar{\ell} = \lambda$ subsequent to the validation and improvement procedure introduced in Fig. 2.

A. Sphere model

The first experimental results are presented for the unimodal sphere model of dimension n,

$$f(\mathbf{x}) = \sum_{i=1}^{n} x_i^2. \tag{8}$$

The average fitness versus the number of function evaluations is presented in Fig. 4. A performance edge is achieved by restricting the validation strategy to only having the surrogate correctly rank the μ highest ranking individuals, and thereby saving the algorithm of evaluating individuals that would have been disregarded in the next iteration. Fig. 5 shows the mean intermediate function evaluations that are calculated during the validation process. As one expects, requiring the method to evaluate no more than the μ best ranked candidate individuals results in a lower intermediate function evaluations, generally saving the method one function evaluation per generation, it also achieves a better mean fitness, as shown in Table I.

B. Rosenbrock's function

The first experiment is now repeated for Rosenbrock's function,

$$f(\mathbf{x}) = \sum_{i=2}^{n} 100(x_i - x_{i-1}^2)^2 + (1 - x_{i-1})^2. \tag{9}$$

The average fitness versus the number of function evaluations is presented in Fig. 6 and Fig. 7 shows the mean intermediate function evaluations that are calculated during the

		Function eval.			Generations			Fitness		
	n	mean	median	sd	mean	median	sd	mean	median	sd
all	2	130.59	132	18.33	49.02	49	6.51	2.35e-09	2.82e-10	1.15e-08
μ	2	81.53	81	9.53	48.11	48	5.02	7.01e-10	2.26e-10	1.35e-09
all	5	702.02	702	67.57	145.15	145	14.96	2.77e-10	1.82e-10	3.64e-10
μ	5	545.25	547	54.27	132.60	132	11.03	1.83e-10	1.46e-10	1.09e-10
all	10	1563.58	1553	117.09	241.83	240	18.47	1.52e-10	1.37e-10	5.03e-11
μ	10	1161.03	1158	79.98	226.60	224	13.86	1.34e-10	1.22e-10	3.80e-11
all	20	3383.83	3377	135.52	423.14	424	20.42	1.27e-10	1.21e-10	2.51e-11
μ	20	2795.28	2804	132.77	372.86	372	16.56	1.17e-10	1.12e-10	1.72e-11

TABLE I
MAIN STATISTICS OF EXPERIMENTAL RESULTS FOR UPDATING SURROGATE WITH ALL OR μ BEST INDIVIDUALS ON SPHERE MODEL.

		Function eval.			Generations			Fitness		
	n	mean	median	sd	mean	median	sd	mean	median	sd
all	2	389.85	386	63.85	132.31	130	31.25	6.24e-10	3.20e-10	1.05e-09
μ	2	344.91	336	78.58	172.16	170	49.95	7.53e-10	1.66e-10	3.64e-09
all	5	2464.22	2280	748.55	514.59	492	105.77	2.75e-01	1.74e-10	1.01e+00
μ	5	1724.89	1729	295.60	520.66	520	82.79	1.83e-10	1.53e-10	1.05e-10
all	10	6800.50	6495	1258.68	1079.82	1052	177.76	2.79e-01	1.32e-10	1.02e+00
μ	10	6138.48	6143	1398.15	1177.71	1103	310.11	1.99e-01	1.24e-10	8.73e-01
all	20	19968.80	20004	234.66	2494.00	2500	49.60	4.54e-01	2.88e-02	1.08e+00
μ	20	19645.90	20002	1086.37	2687.25	2748	230.50	3.10e-01	3.12e-07	9.97e-01

TABLE II
MAIN STATISTICS OF EXPERIMENTAL RESULTS FOR UPDATING SURROGATE WITH ALL OR μ BEST INDIVIDUALS ON ROSENBROCK'S FUNCTION.

validation process. Despite requiring more generations, the over all function evaluations are significantly lower and yield a better fitness when updating the surrogate on only the μ best individuals as shown in Table II. If all of the candidate individuals have to be ranked correctly, the method will get stuck in local minima for this problem in around 6 out of 100 experiments, however this is not a problem if only the μ best candidate individuals are ranked consistently, except at high dimensions, and even then the μ best individuals policy significantly outperforms evaluating all of the candidate individuals. Clearly the choice of validation policy will influence search performance.

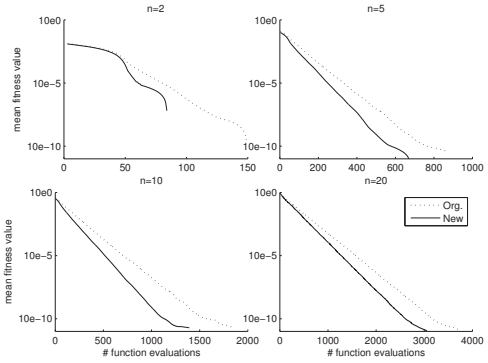

Fig. 4. Mean fitness values versus number of function evaluation by updating surrogate using all (dotted) or μ best (solid) individuals for sphere model.

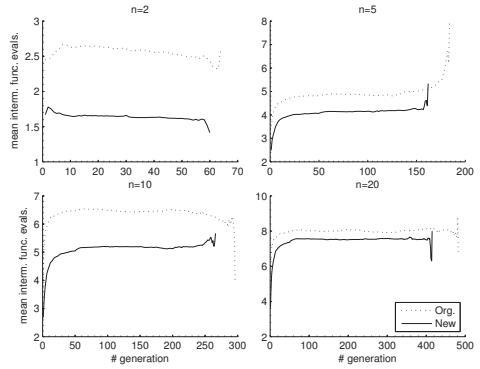

Fig. 5. Mean intermediate function evaluations versus generation by updating surrogate using all (dotted) or μ best (solid) individuals for sphere model.

V. DISCUSSION AND CONCLUSION

The technique presented in this paper to control the number of true fitness evaluations is based on a single test individual chosen from a set of candidate individuals which the surrogate ranks the highest. The approximate ranking of this test individual is compared with its true ranking in order to determine the quality of the surrogate. This is a simple form of cross-validation. An alternative approach could be to rank all candidate individuals along with the training individuals using the surrogate model. This is followed by the re-ranking of training and candidate individuals using the updated surrogate and comparing it with the previous estimate by computing Kendall's τ. Its aim is to observe a change in ranking between successive updates of the surrogate. This study has shown that during the validation process it is sufficient for τ to be close to 1 or that only the potential parent individuals should be ranked consistently. Moreover, the new validation approach reduces the number of fitness evaluation needed, without a loss in performance although it might take a few more iterations in CMA-ES. The studies presented are exploratory in nature and clearly the approach must be evaluated on a greater range of test functions. These investigations are currently underway for combinatorial optimization problem, e.g. job shop scheduling

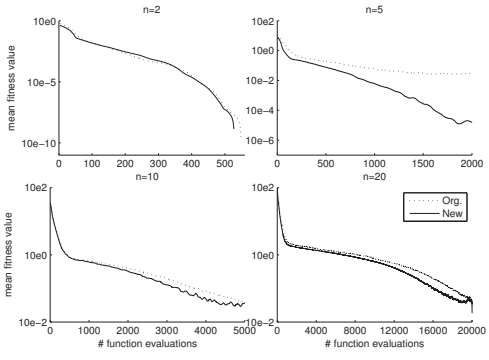

Fig. 6. Mean fitness values versus number of function evaluation by updating surrogate using all (dotted) or μ best (solid) individuals for Rosenbrock's function.

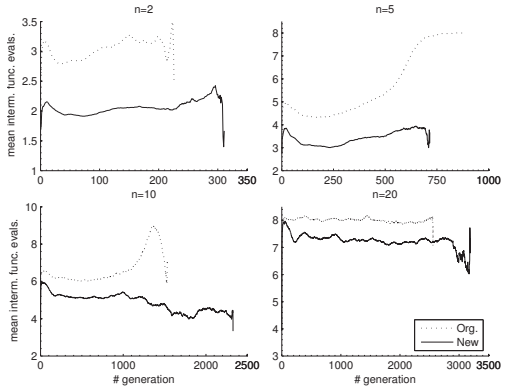

Fig. 7. Mean intermediate function evaluations versus generation by updating surrogate using all (dotted) or μ best (solid) individuals for Rosenbrock's function.

problem.

When it comes to modeling surrogates based on training data, the general rule of thumb is the bigger the training set, the more accurate a model. However, there are computational time limits thus pruning of the training set is necessary. Previous studies [13], [18] have reported that replacing random training individuals is not optimal. This study has shown that there is no statistical difference in omitting oldest or lowest-ranking individuals from the training set. Hence, for future work, further investigation on the fitness landscape is needed to determine effectively which search area is no longer of interest and thus unnecessary for the surrogate to approximate correctly. For instance it could be of interest to disregard training individuals with the largest euclidean distance away from the current candidate individuals rather than simply omitting the oldest/lowest-ranking training individuals.

When building surrogates in evolutionary computation one is interested in the quality of ranking of individuals only. For this reason the training accuracy and cross validation is a more meaningful measure of quality for the surrogate model. This is in contrast to regression, where the fitness function is modeled directly and the quality estimated in terms of measures such a least square error. This study has shown that the sampling used for validating the accuracy of the surrogate can stop once the μ best ranked candidate individuals have been evaluated, since they are the only candidate individuals who will survive to become parents in the next generation. Although in some cases the sampling could stop sooner, when the surrogate ranking and true ranking are sufficiently concordant, i.e. τ was close to 1. This slight slack in for τ is allowed due to the fact the ES search is stochastic, however the allowable range in slack for τ needs to be investigated more fully since allowing only $\tau \in [0.999, 1]$ might be too narrow an interval, resulting in an excess of expensive function evaluations needed.

However, in the context of surrogate-assisted optimization the discrepancy between the exact model and its surrogate can be translated as noise, which could be an indicator of the necessary sampling size for validation/updating the surrogate, instead of only focusing on consistently ranking the μ best candidate individuals. Therefore, one can take inspiration from a varying random walk population model suggested by [19] to approximate the population sizing to overcome unnecessary fitness evaluations.

REFERENCES

[1] T. P. Runarsson, "Ordinal regression in evolutionary computation," in *Parallel Problem Solving from Nature IX (PPSN-2006)*, ser. LNCS, vol. 4193. Springer Verlag, 2006, pp. 1048–1057.

[2] Y. Ong, P. Nair, A. Keane, and K. W. Wong, *Surrogate-Assisted Evolutionary Optimization Frameworks for High-Fidelity Engineering Design Problems*, ser. Studies in Fuzziness and Soft Computing Series. Springer, 2004, ch. 15, pp. 333–358.

[3] A. Sobester, S. Leary, and A. Keane, "On the design of optimization strategies based on global response surface approximation models," *Journal of Global Optimization*, vol. 33, no. 1, pp. 31–59, 2005.

[4] Y. Jin, "A comprehensive survey of fitness approximation in evolutionary computation," *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, vol. 9, no. 1, pp. 3–12, January 2005.

[5] D. Lim, Y.-S. Ong, Y. Jin, and B. Sendhoff, "A study on metamodeling techniques, ensembles, and multi-surrogates in evolutionary computation." New York, New York, USA: ACM Press, 2007, pp. 1288–1295.

[6] Y. Jin, M. Olhofer, and B. Sendhoff, "A framework for evolutionary optimization with approximate fitness functions," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 5, October 2002.

[7] I. Loshchilov, M. Schoenauer, and M. Sebag, "Comparison-based optimizers need comparison-based surrogates," in *Parallel Problem Solving from Nature XI (PPSN-2010)*, ser. LNCS, vol. 6239. Springer Verlag, 2010, pp. 364–373.

[8] H.-P. Schwefel, *Evolution and Optimum Seeking*. New-York: Wiley, 1995.

[9] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," *Advances in Large Margin Classifiers*, pp. 115–132, 2000.

[10] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2002.

[11] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[12] T. P. Runarsson, "Constrained evolutionary optimization by approximate ranking and surrogate models," in *Parallel Problem Solving from Nature VII (PPSN-2004)*, ser. LNCS, vol. 3242. Springer Verlag, 2004, pp. 401–410.

[13] A. Ratle, "Optimal sampling strategies for learning a fitness model," in *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*, vol. 3, 1999, pp. 2078–2085.

[14] M. Sasena, P. Papalambros, and P. Goovaerts, "Exploration of metamodeling sampling criteria for constrained global optimization," *Engineering Optimization*, vol. 34, no. 3, pp. 263–278, 2002.

[15] W. Ponweiser, T. Wagner, and M. Vincze, "Clustered multiple generalized expected improvement: A novel infill sampling criterion for surrogate models," *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*, pp. 3515–3522, June 2008.

[16] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

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

[18] Y. Jin and J. Branke, "Evolutionary Optimization in Uncertain EnvironmentsA Survey," *IEEE Transactions on Evolutionary Computation*, vol. 9, no. 3, pp. 303–317, June 2005.

[19] B. Miller, "Noise, sampling, and efficient genetic algorithms," Ph.D., University of Illinois at Urbana-Champaign, Urbana, IL, 1997.