On Integration of Multi-Point Improvements

R. Girdziušas[1], **J. Janusevskis**[1,2], **and R. Le Riche**[1,3]

[1] École Nationale Supérieure des Mines de Saint-Étienne
[2] Riga Technical University, Latvia
[3] CNRS UMR 6158

Abstract

Kriging with maximization of a multi-point expected improvement presents an interesting direction in the parallel budgeted optimization. However, this statistical criterion is a difficult-to-evaluate integral, and all the presently available methods are either extremely slow or not accurate enough. We introduce an extremely simple and fast integration method which is also accurate when the covariance matrices of kriging responses are diagonally-dominant, a frequently occuring case in practice. In addition, we state preliminary results with a novel importance sampling method which is developed as a more accurate alternative to the Monte Carlo based methods. Our tests are carefully designed to represent early and late stages of a budgeted optimization. We emphasize that integrating 4-variate functions can be extremely challenging.

1 Introduction

We are interested in the parallel budgeted optimization algorithms which rely on the multi-point expected improvement (EI) [1, 2, 3], which is an integral defined as

$$\mathrm{EI}(\mathbf{x}) = \int_{\mathbb{R}^d} (f_{\text{best}} - \min \mathbf{y})_+ \, p_{\mathbf{m}(\mathbf{x}),\mathbf{C}(\mathbf{x})}(\mathbf{y}) d\mathbf{y}, \tag{1}$$

where $(\cdot)_+ \equiv \max(0, \cdot)$. The integration variables $\mathbf{y} \in \mathbb{R}^{d=\mu+\lambda}$ are multiple kriging responses observed at d locations $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_d\}$, $\mathbf{x} \in \mathbb{R}^{d_0}$. The responses are distributed according to the conditional normal density with mean $\mathbf{m}(\mathbf{x}) \in \mathbf{R}^d$ and covariance $\mathbf{C}(\mathbf{x}) \in \mathbf{R}^{d\times d}$, the conditioning being performed on all the observed values of an expensive-to-evaluate cost function $f : \mathbf{R}^{d_0} \mapsto \mathbf{R}$. The parameter $f_{\text{best}} \in \mathbb{R}$ is the minimal known value of the cost function which needs to be further minimized.

Considering the asynchronous cloud access, one splits the location set $\mathbf{x}$ into the "active subset" of μ known locations whose cost is being evaluated by the nodes, and the uknown "candidate subset" of λ locations that need to be further evaluated by readily available nodes. The "active subset" may also include previously observed locations of erroneous cost evaluations, or locations that do not satisfy additional requirements. In both of the cases the multi-point EI criterion keeps the new candidate locations away from the active subset without introducing any additional parameters to the optimization algorithm.

It is the high variability of the possible kriging mean, covariance, and f_{best} values what makes the integration hard. The kriging means can be on the scale of $O(10^5)$ while the value $f_{\text{best}} = O(10^{-1})$ might still be far from the optimal one. The mean sample Monte Carlo method might not hit a single nonzero value, or, with a few hits, the estimated EI values may exhibit extremely large variances which makes the subsequent ranking of candidate locations $\mathbf{x}$ unreliable.

The multi-point EI criterion also needs to be evaluated quickly as the integration is often repeated a million times during a single run of the budgeted optimization. The use of the cloud resources to parallelize any Monte Carlo method becomes nontrivial as a single second wasted in the parallel communication of subaverages may add days of delay to the overall optimization.

2 Integration Methods

2.1 A Simple Integration Method

S. Let us introduce the method which dramatically improves a variety of fast inaccurate general purpose techniques, such as unscented transforms, and the mean sample Monte Carlo method whose sample size is, say, $\mathcal{O}(10^3)$. We will show that the multi-point EI can be approximately evaluated with the method whose "sample size" is barely $2d$.

In what follows, we will work with the Cholesky decomposition $\mathbf{C} = \mathbf{L}\mathbf{L}^T$:

$$\mathrm{EI}(\mathbf{x}) = \int_{\mathbb{R}^d} \left(f_{\text{best}} - \min(\mathbf{m} + \mathbf{L}\mathbf{u})\right)_+ p(\mathbf{u}) d\mathbf{u}. \tag{2}$$

Hereafter $p(\mathbf{u})$ is the standard d-variate normal probability density function, and the integrand is not equal to zero only in the feasible region defined as

$$S = \{\mathbf{u} \in \mathbb{R}^d : \cup(-f_{\text{best}}\mathbf{1} + \mathbf{m} + \mathbf{L}\mathbf{u} \leq \mathbf{0})\}, \tag{3}$$

where the set union $\cup$ acts on the halfplanes presented as the inequalities (row-wise).

We can now introduce the method which relies on the following rule:

$$\int_{\mathbb{R}^d} g(\mathbf{u}) p(\mathbf{u})\, d\mathbf{x} = \sum_{i=1}^{2d} w_i g(\mathbf{u}_i). \tag{4}$$

Here one could have preserved the full product $g(\mathbf{u})p(\mathbf{u})$ on the right hand side, but it is better to drop out the density for the reasons that will become clear soon.

Our first proposition is to choose $2d$ nodes in the regions where one practically observes most substantial amounts of "mass" of the integrand:

$$\mathbf{u}_i = \mathrm{col}_i\big[\mathrm{diag}(\mathbf{u}_0 - c_1\mathbf{1}), \mathrm{diag}(\mathbf{u}_0 - c_2\mathbf{1})\big], \;\; \mathbf{u}_0 = \mathbf{L}^{-1}(f_{\text{best}}\mathbf{1} - \mathbf{m}). \tag{5}$$

The operator $\mathrm{diag}(\cdot)$ builds a diagonal square matrix from a given vector, and the integration nodes are $2d$ columns of two such matrices. The vector $\mathbf{u}_0$ is the intersection point of all the hyperplanes indicated in Eq. (3). The values $c_1 = 1$ and $c_2 = 2$ shift the integration nodes from a finitely-valued edge of a feasible region towards its interior by one and two standard deviations.

The choice of nodes according to Eq. (5) is hardly the most optimal in all the integration cases, but it works well when the covariance matrix is diagonally-dominant, and thus it is a definite improvement over the node placements dictated by the exact monomial integration or "scaling arguments" related to the normal measure concentration phenomenon. The latter does not apply here as there can be quite a mismatch between the standard normal density and the rest of the integrand in Eq. (2). Practically observed coordinates of the vector $\mathbf{u}_0$ may reach values such as $\mathcal{O}(100)$, which indicates that the feasible region is too far away from the maximal density value of $p(\mathbf{u})$. A direct evaluation of the latter is thus not possible, and for this reason the values $p(\mathbf{u}_i)$ are excluded from Eq. (4).

Our second advancement in building the integration rule is to choose the weights w_i so that the following $2d$ test integrals are evaluated exactly:

$$\mu_i = \int_{\mathbb{R}^d} (m'_i - \tilde{\mathbf{l}}_i\mathbf{u})_+ p(\mathbf{u}) d\mathbf{u} = m'_i \Phi\left(\frac{m'_i}{\sigma'_i}\right) + \sigma'_i\, \phi\left(\frac{m'_i}{\sigma'_i}\right), \;\; i = 1, \ldots d, \tag{6}$$

$$\int_{\mathbb{R}^d} \left((m'_i - \tilde{\mathbf{l}}_i\mathbf{u})_+ - \mu_i\right)^2 p(\mathbf{u}) d\mathbf{u} = \sigma'^2_i \Phi^2\left(\frac{m'_i}{\sigma'_i}\right) + \sigma'^2_i \phi^2\left(\frac{m'_i}{\sigma'_i}\right) - m'_i\sigma'_i \Phi\left(\frac{m'_i}{\sigma'_i}\right)\phi\left(\frac{m'_i}{\sigma'_i}\right), \;\; i = 1, \ldots d. \tag{7}$$

Here $\tilde{\mathbf{l}}_i$ is the ith row of $\mathbf{L}$, $m'_i = f_{\text{best}} - m_i$, and $\sigma'_i = \|\tilde{\mathbf{l}}_i\|_2$. The symbols Φ and ϕ denote the standard normal univariate distribution and density, resp.

The substitution of $2d$ test integrands given by Eqs. (6) and (7) into Eq. (4) yields a linear system of $2d$ equations for the determination of $2d$ weights w_i, which must be solved every time a new integrand is presented. In the presence of singularities, the pseudo-inverse is applied.

The choice of these particular test integrals is dictated by the ability to use domain-specific functions which can also be integrated exactly. There are not that many choices, and this is a clear improvement over the use monomials which would lead to very poor integration results. Notably, the inclusion of one more node with the demand to integrate, in addition to these $2d$ integrals, the unity constant function, would already lead to performance deteriorations.

2.2 Monte Carlo Based Methods

M1–M9. The numbers "1–9" will correspond to the mean sample Monte Carlo method applied with $10^1 \dots 10^9$ samples, resp. Presently, the absolute borderline between fast and slow methods is the sample size of $10^3 \dots, 10^4$ points, which may demand 1–10 ms. of time when using an average desktop computer.

I2–I4. We introduce the Monte Carlo variance reduction method by sampling strictly on the regions of a non-zero improvement by employing a truncated multivariate mormal distribution. The problem is decomposed so that the samples from the truncated density can be obtained by making d synchronous draws from a one-sided truncated univariate normal distribution. Fast sampling strategies from such densities are available [4, 5]. The weights of the importance sampling are obtainted by aggregating probabilities provided by the univariate variables. The numbers "2–4" correspond to the method applied with $10^2 \dots 10^4$ samples, resp.

T. The Tallis formulae can be used to map the original problem to the evaluation of $O(d^2)$ multivariate normal distribution integrals [6]. The latter, despite the notable advances (see the references in [6]), can still be very demanding, as indicated by the recommended practical sample sizes of the existing quasi Monte Carlo methods, such as $\mathcal{O}(1000d)$. Numerical guarantees of a strict positive definiteness of the resulting covariance matrices have not been provided yet, and we have encountered sets of integrals where this is indeed a problem. However, when it works, the method is especially useful in theoretical precision studies. We thank the authors who provided us their early R implementation, which we use as the reference method in many test sets discussed in Section 3. This decision is motivated by the following observations: (i) the best results obtained with the T method strongly correlate with the complexity and accuracy of the M9 method, and (ii) the T method does not produce zero integrated improvement values which are still observable in the M9 method (they cause problems when evaluating relative errors).

3 Integration Accuracy and Sample Complexity

We have extracted the integrals from the EI-based minimization of the Rosenbrock and Rastrigin functions defined on the domains $[0, 5]^{d_0}$ and $[0, 2.5]^{d_0}$, resp. Both cost functions are minimized in $d_0 = 2$ (2D) and $d_0 = 9$ (9D) cases, their optimal values are equal to zero, except in the 2D case of Rastrigin where the value is unity. The kriging models are created by using the Matérn (5/2) kernel [7] whose hyper-parameters are estimated by maximizing the likelihood criterion. Sets I and II signify the beginning (after 5 and 90 function evaluations in 2D and 9D, resp.), while the sets III and IV correspond to the end of the minimization (55 and 490 function evaluations in 2D and 9D, resp.). Sets I and III signify the beginning of the CMA-ES [8] maximization of the EI criterion (estimated with the M5 method), while the sets II and IV correspond to the end-phase of the maximization. Each particular set represents the CMA-ES population, and it contains one hundred 4-variate integrals, each described by the 4-point kriging mean, 4x4 covariance matrix, and a scalar value f_{best}.

A statistical description of all the sets is given in Table 1. We provide the average minimal and maximal elements of the mean of kriging responses, root spectral norms of the kriging covariance matrices, and the diagonal-dominance D which is the average ratio between the absolute value of the diagonal element and the sum of the absolute values of the off-diagonal elements. The other remaining quantities are the minimal cost value reached, the average multi-point EI value obtained with the M9 method, and the root-mean-squared error of the leave-one-out cross-validation (CV) of the kriging model, relative to the standard deviation of the actual cost function.

The accuracy of the integration methods is reported in Table 2. The T method [6] is used a reference method whenever possible. The parameters of the latter are chosen so that the sample complexity approximately matches that of the M9 method. We report the median of the relative errors and Spearman's rank correlations. The latter is "bootstrapped" by generating one hundred subsets of ten randomly (uniformly) chosen integrals out of one hundred integrals given for each set, followed by the ranking of the estimated EI values and calculation of the average Spearman's rank correlation. We have chosen such a measure to emphasize that the integration is actually needed in the maximization of the multi-point EI values, but the impact of various integration accuracy measures on the global optimization remains unknown. The accuracy of the rank correlations is ± 0.1, and the low correlation value interval $[-0.2, 0.2]$ indicates that the method ranks the values of the multi-point EI criterion randomly. Notably, the integration setting with the Rosenbrock 9D set IV does not exist in order to emphasize that even the M9 method (M5 method was used to generate the other cases) cannot resolve the incredibly small improvements, and the CMA-ES maximization becomes random.

Table 2 indicates that the S method may attain unusually small median relative errors, but the ranking may not always correlate with the results of the T method which is believed to be more accurate. Notably, the S-method can still

Table 1: Statistics of the problem sets I–IV. All the quantities, except the cross-validation error, are averages over 100 integrals constructed for each optimization problem and its particular set.

Set	$\min \mathbf{m}$	$\max \mathbf{m}$	$\sigma^{0.5}_{\max}(\mathbf{C})$	D	f_{best}	$\text{EI}(\mathbf{x})$	CV,%	$\min \mathbf{m}$	$\max \mathbf{m}$	$\sigma^{0.5}_{\max}(\mathbf{C})$	D	f_{best}	$\text{EI}(\mathbf{x})$	CV,%
			Rosenbrock, 2D							Rosenbrock, 9D				
I	375.5	15749.1	313.3	1.9	1.4	46.0	6.3	16212.6	80847.5	8804.2	5.2	13648.1	5037.5	28.5
II	-115.0	8764.4	295.6	0.7	1.4	196.1	6.3	-19151.6	63916.4	12872.0	3.0	13648.1	35546.8	28.5
III	311.7	13077.6	93.2	5.0	0.0	0.0	4.8	26023.5	75492.0	2825.6	6.8	0.6	0	8.9
IV	2.0	7864.0	34.4	3.4	0.0	0.6	4.8	–	–	–	–	–	–	–
			Rastrigin, 2D							Rastrigin, 9D				
I	23.7	50.7	9.5	16.5	18.9	1.3	58.2	188.8	3659.2	852.3	5.7	166.6	386.0	29.8
II	15.9	19.9	6.9	4.4	18.9	6.5	58.2	-1019.4	226.6	1637.0	7.2	166.6	1976.2	29.8
III	27.3	54.5	12.6	1525.5	1.0	0.0	45.3	753.6	4143.8	239.4	9.1	16.9	0.0	21.4
IV	32.2	54.4	19.3	8.9	1.0	0.0	45.3	74.3	6101.3	364.9	13.3	16.9	1.6	21.4

Table 2: Relative median errors and Spearman's rank correlations by using the T method [6] as the reference. In the case with Rosenbrock 2D, sets I–IV, the T method is not numerically stable, and the M9 method is used as the reference.

	Rosenbrock, 2D								Rosenbrock, 9D							
	I		II		III		IV		I		II		III		IV	
S	1.4e-03	0.8	4.4e-01	0.0	1.8e-03	0.9	1.2e-04	1.0	1.8e-04	0.9	3.8e-03	0.0	4.9e-08	0.6	–	–
M4	2.5e-02	0.9	5.6e-03	0.6	4.3e-02	0.8	1.8e-02	1.0	8.4e-03	1.0	1.4e-03	0.7	1.0e+00	-0.0	–	–
IS4	1.1e-02	1.0	6.3e-03	0.0	8.9e-03	0.9	1.4e-02	0.9	1.2e-02	1.0	1.6e-03	0.1	5.8e-02	0.9	–	–
M9	0.0e+00	1.0	0.0e+00	1.0	0.0e+00	1.0	0.0e+00	1.0	4.4e-05	1.0	5.2e-06	1.0	1.0e+00	0.0	–	–
	Rastrigin, 2D								Rastrigin, 9D							
	I		II		III		IV		I		II		III		IV	
S	2.2e-02	0.9	8.3e-02	0.0	5.2e-06	0.4	9.9e-01	-0.0	4.2e-02	0.8	4.6e-02	-0.1	1.5e-07	0.6	6.5e-03	0.9
M4	1.6e-02	0.9	5.5e-03	0.6	1.0e+00	0.2	8.1e-02	0.2	8.3e-03	1.0	4.0e-03	0.8	1.0e+00	0.1	4.2e-02	1.0
IS4	8.3e-03	1.0	8.9e-03	0.1	8.2e-03	1.0	9.7e-02	0.0	8.1e-03	0.9	5.2e-03	0.1	1.3e-02	1.0	1.5e-02	0.5
M9	7.7e-05	1.0	2.2e-05	1.0	4.6e-03	1.0	2.3e-03	1.0	4.3e-05	1.0	7.5e-06	1.0	6.3e-01	0.5	1.5e-04	1.0

rank the points even in some hard cases, and so can the IS4 method (clf. Rosenbrock 9D, set III). The IS4 method outperforms its M4 counterpart on the sets I and III. However, as the EI value increases on the sets II and IV, the performance of IS4 deteriorates. The method still provides good results at the end of the optimization, when there is a mismatch between the kriging mean and f_{best} (clf. Rosenbrock 9D set III). Essentially, the S and IS4 methods do not rank well the EI values according to the T method on the set II, when the kriging model of the actual cost function is inadequate. Both methods work well in the remaining sets, except the case with Rastrigin 2D, set IV, when the optimal cost function value is already reached and the global optimization progress is no longer possible.

4 Conclusions

The multi-point improvement integral is challenging already in the 4-variate case as the use of the mean sample Monte Carlo method with one billion samples may not hit a single nonzero value of the improvement variable. The integrands vary a lot depending on the cost function, global optimization stage, whether the points are evaluated at the high improvement regions, and if the kriging covariances are diagonally-dominant. We have introduced two relatively accurate integration methods of low sample complexity that improve the mean sample Monte Carlo method.

One could emphasize a clear distinction between: (i) fast techniques which may integrate a four variate integral in microseconds (the S method), the border-line cases such as the M4 method and its IS4 counterpart, which may already operate on the scale of hundreds of milliseconds, and (iii) the "heavy weights" (M5–9, T-method) that are mostly useful in theoretical accuracy studies, and which may still not be accurate enough.

Further research is needed to investigate the impact of the integration methods on the global optimization performance.

References

[1] M. Schonlau. *Computer Experiments and Global Optimization*. PhD thesis, Univ. of Waterloo, 1997.

[2] J. Janusevskis, R. Le Riche, and D. Ginsbourger. Parallel expected improvements for global optimization: Summary, bounds and speed-up. Technical report, École Nationale Supérieure des Mines de St-Étienne, 2010.

[3] S. C. Clark and P. I. Frazier. Parallel Global Optimization Using An Improved Multi-points Expected Improvement Criterion. Presentation slides, INFORMS Optimization Society Conference, Miami, 2012.

[4] N. Chopin. Fast simulation of truncated Gaussian distributions. *Statistics and Computing*, 21, 2011.

[5] V. Mazet. Simulation d'une distribution Gaussienne tronquée sur un intervalle fini. Technical report, Université de Strasbourg, 2012.

[6] C. Chevalier and D. Ginsbourger. Fast computation of the multi-points expected improvement with applications in batch selection. Technical report, IMSV, Bern, 2012.

[7] M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, 1999.

[8] N. Hansen, S. D. Müller, and P. Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evolutionary Comp.*, 11:1–18, 2003.