SENSITIVITY ANALYSIS FROM EVOLUTIONARY ALGORITHM SEARCH PATHS

G. Paul[a], **C. L. Müller**[a] and **I. F. Sbalzarini**[a]

[a]Institute of Theoretical Computer Science and Swiss Institute of Bioinformatics,
ETH Zurich, Universitätstrasse 6, CH–8092 Zürich, Switzerland.
grpaul@inf.ethz.ch, christian.mueller@inf.ethz.ch, ivos@ethz.ch

Many problems of practical importance can be translated into the study of a real-valued objective function of several continuous variables. The graph of such an objective function defines a surface that we term "landscape". Different aspects of the same landscape might be of interest, but attainable only through independent means. For example, when the main objective is optimization, one can use Evolutionary Algorithms (EAs) to generate a search path designed to be biased toward the set of optima of the landscape. If the main purpose is to compute the global sensitivity [1] of the objective function with respect to its parameters, sampling strategies tailored to this problem are available in order to avoid a bias in the estimate due to bad sampling of the landscape. Nevertheless, when evaluations of the objective function are costly, the ability to achieve such conflicting tasks at once becomes crucial. We present here a generic estimation procedure for off-line global and local sensitivity indices from samples generated by an EA optimizer.

Sensitivity analysis (SA) quantifies the impact of variations in the input parameters on the objective function using either local or global indices. Local indices (scaled partial derivatives, Morris' method, *etc.*) quantify the sensitivity around a nominal value, resulting in easy-to-compute quantities at the cost of locality. Global indices are computed over a wide range of possible input values, which is computationally harder. In the following, we focus on variance-based global sensitivity indices (GSIs) [1]. Let us denote $f\colon \mathcal{D} \subset \mathbb{R}^d \to \mathbb{R}$ the objective function and $\mathbf{X} = (X_i)_{i=1,\cdots,d}$ a d-dimensional random variable. The first order GSIs $(S_k)_{k=1,\cdots,d}$ read:

$$S_k = \frac{\mathbb{V}[\,\mathbb{E}[f(\mathbf{X})|X_k]]}{\mathbb{V}[f(\mathbf{X})]}\ ,$$

where $\mathbb{V}[\cdot]$ and $\mathbb{E}[\cdot]$ denote the variance and the expectation, respectively. S_k measures the ratio between the output variance of the best (in a mean-square error sense) univariate predictor $\mathbb{E}[f(\mathbf{X}) \mid X_k]$ of f and the total output variance. It is clear that if f were a function of X_k only, S_k would be 1. Higher-order indices, built by higher-order conditioning, lead to an ANOVA-like decomposition of the contributions of the input parameters to the objective function. We restrict our presentation to first-order indices and to a box domain $\mathcal{D} = [a_1, b_1] \times \cdots \times [a_i, b_i] \times \cdots \times [a_d, b_d]$.

One way of assessing the sensitivities of f only is to sample the domain $\mathcal{D}$ uniformly. If the samples are produced by an EA, however, direct use of the above formula leads to wrong sensitivities, in the sense that the indices would be computed with respect to the sampling generated by the EA search. Different searches and different algorithms would hence return different estimates. In order to compute the sensitivities of the objective function only, we need to "normalize" the biased sampling of the landscape in some sense. To this end, we express the targeted moments under uniform sampling with respect to the sampling probability distribution "used" by the EA. This is inspired by importance sampling. Let us denote by $\mathbb{E}_\mu[\cdot]$ the expectation under probability law μ. The absence of a subscript indicates sampling with a uniform probability distribution. We make use of the following equality:

$$\mathbb{E}[f(\mathbf{X}) \mid X_k] = \mathbb{E}_{\mu_{-k|k}}[f(\mathbf{X})w_k(\mathbf{X})|X_k]$$

with the importance weights

$$w_k(\mathbf{X}) \coloneqq \frac{\prod_{i=\{1,\cdots,d\}\setminus\{k\}} u_{a_i,b_i}(X_i)}{\mu_{-k|k}\left(\mathbf{X}_{-k}|X_k\right)}\ ,$$

where u_{a_i,b_i} is the probability distribution function (p.d.f.) of a uniform continuous random variable in $[a_i, b_i]$, $\mu_{-k|k}$ is the conditional p.d.f. of μ with respect to X_k, and $\mathbf{X}_{-k} \coloneqq (X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_d)$.

Suppose that a given EA has generated the data $(f(\mathbf{X}_j), \mathbf{X}_j)_{j=\{1,\cdots,N\}}$ and that we know how each datum has been generated according to the p.d.f. μ_j. Estimation of the first order indices $S_k \coloneqq \frac{N_k}{D_k}$ involves four steps:

1. Weight each data point with its appropriate importance weight, such that we have for each dimension k the new data set $(Y_{kj} = f(\mathbf{X}_j)w_k(\mathbf{X}_j), \mathbf{X}_j)_{j=\{1,\cdots,N\},k=\{1,\cdots,d\}}$.
2. Estimate at the N_q query points $(x_{k1}, \cdots, x_{kq}, \cdots, x_{kN_q})_{k=\{1,\cdots,d\}}$ the conditional expectations $m_k(x_{kq}) = \mathbb{E}[Y_k | X_k = x_{kq}]$ using local polynomial regression [3].
3. Compute the variance in the numerator using a numerical quadrature rule on the quadrature points $(x_{kq})_{q=\{1,\cdots,N_q\},k=\{1,\cdots,d\}}$ to get $\widehat{N_k} \approx \int \left(\widehat{m_k}(x) - \int \widehat{m_k}(x) u_{a_k,b_k}(x)\,\mathrm{d}x\right)^2 u_{a_k,b_k}(x)\,\mathrm{d}x$.
4. Compute $\widehat{D_k}$ with the classical unbiased empirical variance estimate on the data $\left(f(\mathbf{X}_j)w_{\mathcal{D}}(\mathbf{X}_j)\right)_{j=\{1,\cdots,N\}}$ with the whole domain weight $w_{\mathcal{D}} \coloneqq \frac{\prod_{i=\{1,\cdots,d\}} u_{a_i,b_i}(\mathbf{X}_j)}{\mu_j(\mathbf{X}_j)}$.

The first step uses the known distributions μ_j that generated the data in order to build in the second step an unbiased estimation of the conditional expectation under uniform sampling, in a way similar to importance sampling. The computational efficiency of this step relies on the ability to compute the conditional p.d.f. $\mu_{j,-k|k}(\cdot)$. This is easily done if the EA used a multivariate Gaussian to generate the samples at each generation, as is the case for example in the Covariance Matrix Adaptation Evolutionary Strategy (CMA-ES) [2].

The second step uses a non-parametric procedure to estimate the conditional expectations. We could have used any linear smoother to do so, but we chose local polynomials for their nice theoretical and computational properties [3]. This technique approximates locally the regression function by a polynomial of degree p:

$$P_p(u; \boldsymbol{\beta}) = \sum_{l=0}^{p} \frac{\beta_l}{l!} u^l \ .$$

The estimation of the conditional expectation at the query point x_q involves solving the weighted least squares problem

$$\widehat{\boldsymbol{\beta}}(x_q; h_N) \coloneqq \arg\min_{\boldsymbol{\beta}} \sum_{j=1}^{N} \frac{1}{h_N} K\left(\frac{X_{kj} - x_q}{h_N}\right) \left(Y_{kj} - P_p(X_{kj} - x_q; \boldsymbol{\beta})\right)^2$$

with $K(\cdot)$ a kernel function and h_N the so-called bandwidth, controlling the size of the local neighborhood. From this estimate, derivative estimates up to order p are recovered via:

$$\widehat{m_k^{(\nu)}}(x_q; h_N) \coloneqq \nu!\, \widehat{\beta}_\nu(x_q) \tag{1}$$

for $\nu \in \{0, \cdots, p\}$. It has been shown that such estimates have nice asymptotic properties well suited for our task [3]. The method adapts to both fixed and random sampling designs, and it lacks boundary effects in the sense that the bias is of the same order at both interior and boundary points, hence no boundary handling is necessary. A third appealing property is the fact that the asymptotic bias of $\widehat{m_\cdot^{(\nu)}}(\cdot)$ is free of the sampling density when $p - \nu$ is odd. Finally, even if this estimation procedure looks computationally expensive, fast and efficient numerical methods exist to select the bandwidth and estimate $\boldsymbol{\beta}$ [3].

Each estimated conditional expectation $\widehat{m_k}(\cdot)$ is a one-dimensional function. Hence, computing the variance in the numerator can be efficiently done to arbitrary precision via numerical quadrature, instead of basing the estimation on a Monte-Carlo strategy as proposed by Da-Veiga *et al.* [4]. The estimation of the denominator D_k involves a d-dimensional integral, and hence any quadrature is forbidden if the dimension d is large. Instead, we rely on a simple Monte-Carlo estimate.

We provide now a first asymptotic result for the estimate of the numerator in order to reveal the terms involved. We do so by working conditionally on the search path generated by the EA and looking at the samples as an i.i.d. realization of the joint p.d.f. that generated them. Combining results from Fan and Gijbels [3] with the continuous mapping theorem, we can express the asymptotic behavior of the estimator of N_k using local polynomials of degree p odd, provided $h_N \to 0$, $Nh_N \to \infty$, and $m_k^{p+1}(\cdot)$ is integrable:

$$\widehat{N_k}(p, h_N) = N_k + 2e_1^T S_p^{-1} c_p \frac{1}{(p+1)!} C(m_k, p) h_N^{p+1}(1 + o_P(1)) \tag{2}$$

with

$$C(m_k, p) \coloneqq \int m_k(x) m_k^{(p+1)}(x) u_{a_k,b_k}(x)\,\mathrm{d}x - \int m_k(x) u_{a_k,b_k}(x)\,\mathrm{d}x \int m_k^{(p+1)}(x) u_{a_k,b_k}(x)\,\mathrm{d}x\ ,$$

where e_1 is the $(p+1)$-dimensional unit vector with the first element equal to 1, $S_p \coloneqq (K_{r+c})_{0\le r,c\le p}$ is the Hankel matrix of the kernel moments $K_n \coloneqq \int u^n K(u)\,\mathrm{d}u$, $c_p = (K_{p+1}, \cdots, K_{2p+1})^T$, and $o_P(1)$ denotes a random variable converging to zero in probability. We see that the nice properties of the local polynomial estimator have been lifted to the estimator of the numerator, in particular that the asymptotic bias is free of μ. Without any additional work, we can also recover local sensitivity indices from the fitted coefficients $\boldsymbol{\beta}$ using formula (1). An indirect way of doing a local sensitivity analysis would be to integrate the conditional expectations in step 3 only in a neighborhood around a nominal value of interest.

Numerical experiments with the present strategy allow assessing the finite-sample properties of the estimator. We used $f = x_1 + x_2^2$ as a simple test case for which analytical global first-order sensitivity indices are known. We find that the present estimator is able to recover the global sensitivity indices with relative errors around 10% using 5 samples from each of 200 independently placed Gaussians with correlations of the covariances uniformly in $[-1.1]$. This mimics the situation where the samples are produced by an EA with a Gaussian proposal distribution performing a random walk over the landscape.

An alternative strategy to the one proposed here would be to first reconstruct the objective function f based on the samples generated by the EA, and then compute the sensitivity indices using kriging or response-surface methods. If the dimensionality of the problem is high, however, the curse of dimensionality disfavors such a strategy. In our method, only low-dimensional quantities need to be estimated, namely d one-dimensional non-parametric regressions, and d one-dimensional integrals. This is a significant advantage compared to first estimating the whole landscape, which would require estimating a d-dimensional surface, and then recovering sensitivity indices by integration. As a result, the computational complexity of our algorithm scales linearly with the number of dimensions, for a fixed number of samples. Nevertheless, if the conditional distributions needed to compute the importance weights are unknown or hard to access, we face the same issue as we would then need to estimate a $(d-1)$-dimensional density from the data. For the sake of conciseness, we only presented first order indices here, but there is no theoretical nor practical limitation to computing also higher-order indices. These are of importance when assessing the effect of interactions between the parameters on the objective function. In our setting, second order indices would require the estimation of 2-dimensional regression functions, for which the local polynomial regression methodology is still very efficient.

We could equally have used smoothing techniques other than local polynomials, such as splines or orthogonal functions. One can show [3] that all linear smoothers can be unified and differ only by their smoothing

kernels. The choice of a given smoother relies on mainly two criteria: optimality properties with respect to a given underlying function class, and spatial adaptivity. It has been shown that local polynomials are nearly optimal in a linear minimax sense over the space of smooth functions, whereas wavelets are better suited to spaces where this is not true. Adaptivity can artificially be achieved with local polynomials based on selection schemes for the bandwidth parameter. Wavelets implement spatial adaptivity in a much more natural way, via a multi-resolution decomposition of the state space.

From a theoretical point of view, the asymptotic analysis presented here could be improved in many directions. The first one could be to compute also the asymptotic properties of the full estimator $\widehat{N_k}/\widehat{D_k}$. Second, one could study the fluctuations of the proposed estimator, such as its convergence in distribution, in order to derive asymptotic confidence bounds. Another research direction is to investigate the properties of the estimator unconditionally on the path. This is more involved since it requires taking into account the correlated way in which the samples are proposed by the EA. In the statistics literature, there exist extensions of local polynomial regression to the case of correlated samples, such as in time-series analysis. It is shown there that for stationary processes the form of the asymptotic bias does not change [3]. Such a generalization is crucial if one wants to use our framework in an on-line setting, where the EA uses previous samples to learn landscape features and accordingly adapt the search proposal.

References

[1] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola, "Global Sensitivity Analysis. The Primer", *John Wiley & Sons*, Ltd, 2008.

[2] N. Hansen. "The CMA evolution strategy: a comparing review.", *In: Towards a new evolutionary computation. Advances on estimation of distribution algorithms*, J. Lozano, P. Larranaga, I. Inza, and E. Bengoetxea, editors, pages 75–102. Springer, 2006.

[3] J. Fan and I. Gijbels, "Local polynomial modelling and its applications, Vol. 66 of Monographs on Statistics and Applied Probability", *London: Chapman Hall,*, 1996.

[4] S. Da Veiga, F. Wahl, and F. Gamboa, "Local polynomial estimation for sensitivity analysis on models with correlated inputs.", *Technometrics*, 51(4):452–463, 2009.