Surrogate Based Optimization (SBO) in pSeven

Surrogate based optimization capabilities in pSeven cover all problem types listed here including robust formulations. Moreover, it provides the base for mixed-integer variables support in pSeven. SBO methods are typically applied for any computationally expensive input problem.

I. Overview

Surrogate based optimization in pSeven is based on Gaussian Processes (GP) modeling technique and originates from *Probability of Improvement* (PI) approach. Still, pSeven implements numerous changes with respect to conventional formulations. In particular, GP construction and philosophy have been drastically altered, so a qualitative review of corresponding aspects of pSeven optimizer is needed. Conventional surrogate based optimization approaches have two prime drawbacks, that can be illustrated on the well-known GP modeling:

1. Model training takes too much time and has large memory footprint. Model training implies maximization of some derived quantity, usually likelihood or some cross validation criterion. In either case, it reduces to multiple evaluation of model predictions with varying model parameters, which requires to invert the correlation matrix on every parameters change. In the conventional implementation, the cost of every evaluation scales like $O\left(N_{sample}^3\right)$ (N_{sample} being the current sample size), because the correlation matrix is of size $N_{sample} \times N_{sample}$ and becomes computationally expensive even for moderate samples $N_{sample} \sim O(10^3)$. Model training is more time-consuming in problems with large dimensional design scape situation with the large number of involved model parameters. The minimal required number of sampled designs scales like $N_{sample} \sim N$ in N dimensions. The total number of model parameters to be determined is of the same order, $N_{param} \sim N$. Additionally, the number of iterations required to locate likelihood maximum is of order N, $N_{iter} \sim N$, each of which costs N evaluations at different parameters values. Overall conclusion is that the cost of model training scales with problem dimensionality as $N_{iter} \cdot N \cdot N_{sample}^3 \sim N^5$. The above implies that conventional GP-based SBO strategies are not applicable to large scale problems. Recently, there were a few notable algorithmic achievements, aimed to overcome these limitations. For instance, one could exclusively utilize iterative matrix methods for large samples and thus reduce the cost of every elementary step to $O(N_{sample}^2)$. Although this helps to push maximal manageable sample sizes to a few thousand, this solution is not entirely satisfactory because the above N^5 estimate is simply changed to something like N^4, which is still prohibitive at large N.

2. None of conventional surrogate models properly takes into account possible multi-scale dependencies of underlying model. Primarily, the fact that virtually all the surrogate models used nowadays have only a small number of tunable parameters. Hence, they cannot be adequate for models, which exhibit multi-scale behavior. The need to invent multi-resolution models had become evident long ago (see, for instance, recent activities in this field in mathematical statistics literature), however, we are not aware of any satisfactory solution up to now. Note that the wish to have multi-resolution surrogates contradicts previously discussed complexity of model construction process: the more parameters surrogate model possess

(multi-resolution) the more costly it becomes to construct (training time). To summarize, the need to apply SBO inspired techniques to large scale problems requires to resolve two contradictory issues: reduce to admissible level the cost of model training (both the time and memory requirements) and simultaneously incorporate multi-resolution capabilities into the model, because generically high dimensional underlying models do exhibit different length scales.

Although the above challenge might seem unsolvable, there is a crucial simplifying observation. Namely, we are not going to construct full-fledged surrogate models for underlying responses as it is usually the goal in mathematical statistics. In optimization context, the goal is to find optimal solution and not to predict responses away from optimality. This means, in particular, that surrogates must be accurate only in promising regions of the design space. Technically, this might be best illustrated by the generic sequence of steps performed during surrogate-based optimization (SBO). Here, one starts with initial design of experiment (DoE) plan aimed to produce well separated set of points having good space filling properties. By its very definition, DoE generated sample is almost uniform in the design space and therefore there is a single number characterizing DoE sample: characteristic (mean) distance L_{mean} between nearest designs. Suppose now that we want to train GP model on DoE generated sample. Silent feature of virtually any GP model is that its tunable parameters reflect the correlation lengths along different coordinate axes in the design space. Then it follows just from dimensional analysis that properly trained GP model ought to have tuned parameters equal to L_{mean} up to dimensionless factors of order one. An immediate and surprising conclusion is that model training is in fact redundant once training set is uniformly distributed and has single characteristic length scale: optimal parameter values at least in case of Kriging-like GP models might be guessed a priori without any actual likelihood optimization. To be on the safe side, one should, of course, check the adequacy of guessed optimal parameters and, perhaps, perform a few likelihood optimization steps. However, this does not invalidate the prime message of performed analysis: full fledged likelihood optimization should not be conducted, relevant surrogate model parameters might be guessed in advance.

The above conclusion is a great step towards reducing the cost of surrogate model training. However, it is operational only at the initial DoE stage and seems to be not applicable after that. To proceed, we need to consider (in general terms) the sequence of SBO steps performed to get optimal solution.

These are very simple ideologically: given current surrogate model optimizer predicts a few promising locations and evaluates underlying model at these designs. Then training set is augmented with newly discovered responses, surrogate model is retrained and optimization process proceeds to the next iteration.

Prime observations here:

1. New designs are added at distinguished locations only, hopefully, near the (locally) optimal solution.
2. Only a few designs are added at a time (number of added designs is much smaller than the current sample size).
3. Characteristic nearest neighbor distances in the vicinity of added designs is smaller than that of original sample.

It follows then that initial DoE sample is in fact augmented with well-localized clusters of new

solutions. Qualitatively, upon accounting for new designs surrogate models should only change in the vicinity of added points, far away from added clusters surrogate are expected to remain intact. Moreover, due to relatively small distance scale within each cluster compared to that of underlying sample correlations within each cluster are expected to be stronger than that between new solutions and points from current sample. Thus we should explicitly account for correlations within each cluster.

Moreover, each of these in-cluster correlations are to be described by new cluster-specific GP models, ultimate reason being that distances within each cluster are small and hence corresponding responses might experience multi-resolution properties of underlying model.

In more details, the above reasoning might be reduced to the following formulation:

1. Evaluated designs eventually cluster in promising regions of the design space
2. Hierarchy of length scales could be observed:

Let $\langle L\rangle_x$ denotes characteristic distance between nearest sampled designs around *x*. Then

- DoE stage: $\langle L\rangle_x = L_0 \forall x$
- After a few iterations (Ω is some promising region):

$$\langle L\rangle_x = L_0 x \notin \Omega \langle L\rangle_x = L_1 x \in \Omega L_1 \lesssim L_0$$

- At later stages (Ω_i are the nested promising regions):

$$\begin{array}{cc} \langle L\rangle_x = L_0 & x \notin \Omega \\ \langle L\rangle_x = L_1 & x \in \Omega_1 \\ & \cdots \\ \langle L\rangle_x = L_k & x \in \Omega_k \\ L_k \lesssim \cdots & \lesssim L_1 \lesssim L_0 \end{array}$$

Therefore, at the expense of additional evaluations we enforce length scales hierarchy at every iteration: instead of single candidate evaluation we perform DoE sampling in candidate's vicinity, determined by upper region length scale. Consequences:

- Underlying model is not only probed at candidate location x_c, but is explored in candidate's vicinity $\Omega(x_c)$

$F(x_c) \to \{F(x_i)\}, i \in \Omega(x_c)$

- Every iteration induces well-defined smaller length scale L_k

$L_k \lesssim \cdots \lesssim L_1 \lesssim L_0$,

each L_k being associated with particular nested regions.

As a crucial side effect of the above reasoning we note that for each new submodel the training process becomes essentially simple as it was for the first model constructed at DoE stage. The only thing which is yet to be discussed is how to unify various surrogates into one global model describing underlying responses in whole design space. For only one added cluster the answer seems to be simple:

in addition to correlations $K^{(0)}(x,y)$, present before new solutions were added, additional contribution looks like composition of three terms:

1. From point *x* to some clustered design z_i
2. Within new cluster correlations of points z_i and z_j
3. From point z_j to considered location y

Formally, updated correlation function looks like:

$K(x,y) = K^{(0)}(x,y) +$

$\alpha \sum_{i,j} K^{(1)}(x,z_i)[K^{(1)}]^{-1}(z_i,z_j)K^{(1)}(z_j,y)$ where the only undetermined parameter is the relative magnitude of new correlations with respect to old ones. This is important: the above generic reasoning valid in SBO context lead us to the conclusion that surrogate model updating reduces to simple one-dimensional optimization subproblem to determine relative weight factor *α*. All other parameters are determined a priory thanks to the specific design space resolution properties of SBO methodology.

The above construction trivially generalizes to the case of several nested regions in which separate GP models are defined. Namely, anzats for multi-resolution GP correlation function, which reflects the above hierarchy of length scales, reads

$K(x,y) = K^{(0)}(x,y) +$

$\sum_{\mu} \quad \alpha_{\mu} \sum_{i,j} \quad K^{(\mu)}(x, x_{\mu}^{i})[K^{(\mu)}]_{ij}^{-1} K^{(\mu)}(x_{\mu}^{j}, y),$

where $K^{(\mu)}$ are Ω_{μ}-specific correlation vector/matrix. Parameters to be determined here include only relative amplitudes $\alpha_{\mu} \geq 0$.

To summarize, our proposition is to radically reduce the computational cost of surrogate-based optimization simultaneously introducing multi-resolution capabilities into the surrogate models. Underlying idea is based on the specifics of virtually any SBO setup, namely, the fact that uniformity of sampled designs could easily be achieved at every SBO step. Price to pay is the necessity to explicitly maintain the hierarchy of surrogate models, each describing subsample correlations at every SBO step. However, this should be considered as advantage, not the drawback: hierarchical structure of surrogates naturally admits multi-resolution capabilities of total surrogate model.

Prime distinctive features of our approach:

- Prime gross features of every correlation function $K^{(\mu)}$ are known in advance once length scale hierarchy is respected
- Seems that one could avoid $K^{(\mu)}$-parameters tuning (``training'') altogether
- Only amplitudes α_{μ} are to be determined for every new region (every iteration)
- α_{μ} determination is cheap (no inversions of large matrices is involved)
- Knowledge of length scale hierarchy allows to predict the domains where model is changing upon the sample augmentation.

II. Single-Objective Constrained SBO

We have to distinguish two vastly different cases:

1. Expensive objective function (perhaps, supplemented with cheap constraints).
2. Single expensive constraint entering the problem with cheap objective function.

Generic combination of cheap/expensive observables is a natural generalization of these two extremes.

Let's consider the first case first.

1. The case of single expensive objective function

Our approach is modeled around the well-known "Probability of Improvement" treatment, in which auxiliary internal subproblem to be solved reads

$x^{*} = \arg\max_{x} PI(x) PI(x) = \Phi[u] u = \frac{f^{*} - \hat{f}_{x}}{\hat{\sigma}_{x}}$ where $(\hat{f}, \hat{\sigma})$ is the surrogate model prediction and uncertainty. Note that numerically it is complete disaster to consider $\Phi[u]$. Instead pSeven solves equivalent problem

$x^{*} = \arg\max_{x} u_{x}$

Meaning of PI criterion is simple: solution $x*$ is the point at which probability to improve current best value f^{*} is maximal (including prediction uncertainties). There are a few weak points of PI strategy:

- Performance crucially depends upon the choice of f^{*} value. Indeed, f^{*} is to large extent arbitrary, there are two limiting cases:

  $f^{*} = -\epsilon + min\hat{f}$: algorithm often ``hangs'' in small vicinity of already known solutions.

  $f^{*} = -\infty$: algorithm essentially find $x^{*} = \arg\max\hat{\sigma}$, which is a badly posed problem (multiple equivalent solutions).

- Algorithm is not sufficiently robust with respect to (hopefully, small) inadequacy of surrogates. When surrogate model predictions deviate

significantly from true responses PI criterion suggests wrong evaluation candidates.

To ameliorate both the above deficiencies pSeven considers "continuous" family of PI-like criteria $PI(x,t)$:

$f^* \to f^*(t) = -t|\Delta f| + \hat{minf}t \in [0:1]$ where Δf is estimated range of objective function variation and is chosen such that $t \in [0:1]$ provides a homotopy between local and global search modes. The solution x^* also becomes *t*-dependent $x^* \to x^*(t)$ and it is crucial to investigate the continuity of the path $x^*(t)$ in the design space. Note that the above mentioned switch from local to global search with rising *t* manifests itself in discontinuity of $x^*(t)$ at some t-values $x^*(t_i - 0) \neq x^*(t_i + 0)$. Technically, we consider sufficiently large number of *t*-values:

- Discontinuity of $x^*(t)$ is revealed by the appearance of several well-separated clusters of evaluation candidates.
- Within each cluster we could pick up essentially any point as the next candidate to be evaluated. The above ensures the presence of both local and global search modes in pSeven SBO strategy. Overall algorithm performance crucially depends upon the solution of internal auxiliary problem. We utilize multi-start strategy which gradually reaches optimal solution and allows to keep candidates to be reused on next SBO iterate:
- Take sufficiently large number of initial guesses and sufficiently crude termination tolerances.
- Push current candidates towards local optimal solutions.
- Cluster resulting set and select only one candidate from each cluster
- Diminish termination tolerances and close the cycle

Note: intermediate locally optimal designs are the natural candidates for multi-starts at next SBO iteration thanks to multi-resolution capabilities of utilized surrogate models.

It remains to discuss how the surrogate models are updated during the course of algorithm. Normally, underlying model is evaluated for each obtained evaluation candidate x_i^*, surrogates are updated with new responses. However, implemented hierarchical GP appeal to slightly different strategy, which requires the following actions for each candidate point x_i^*:

- Establish local characteristic length scale λ_i between nearby sampled designs
- Generate DoE plan in the vicinity of x_i^* with characteristic length scale $\overline{\lambda}_i < \lambda_i$
- Evaluate underlying model at generated locations. These responses are to be used to train next hierarchical GP level.

2. The case of single expensive constraint and cheap objective function.

Consider the case of cheap objective supplemented with one expensive constraint

$$\min_x f \; c_L \le c \le c_U$$

for which current surrogate model *c^* is available. pSeven establishes next evaluation candidate in two stages:

- Solve

$$\min_x f \; c_L \le \hat{c} \le c_U$$

infeasibility of which says nothing about infeasibility of original problem. If it is infeasible, design with minimal constraints violations is taken as x^*.

- Otherwise, pSeven takes into account model uncertainties by considering

$$\min_x f \; \Phi\left[\frac{c_L - \hat{c}}{\hat{\sigma}}\right] \le \alpha \Phi\left[\frac{\hat{c} - c_U}{\hat{\sigma}}\right] \le \alpha$$

with predefined small *α*-parameter (quantile).

III. Multi-Objective Constrained SBO

Multi-Objective SBO optimization of pSeven is build on top of corresponding single-objective counterpart, hence directly utilizes all the advantages of respective algorithms (see above). We only need to overview how the original multi-objective problem is reduced to the sequence of single-objective treatments.

Underlying idea is simple: first, pSeven establishes complete set of anchor points in close analogy to what is done in non-surrogate based approaches. Anchor points determination allows to estimate global geometry of Pareto frontier, to detect degenerate cases earlier and to proceed to the second stage of Pareto front discovery.

Second stage utilizes the notion of Chebyshev convolution to construct particular instance of single-objective problem, the solution of which provides new Pareto optimal design. In more details, respective single objective function looks like:

$$\max_i [\alpha_i \cdot f_i] \sum_i \quad \alpha_i = 1 i = 1, \dots, K$$

where the coefficients α_i parameterizing *K*-dimensional simplex, are varied from iteration to iteration in a way, which ensures even coverage of Pareto frontier.

IV. Summary

Surrogate based optimization in pSeven provides an universal set of algorithms aimed to deal with virtually any expensive input problems, ranging from simplest unconstrained single-objective case to the most difficult to handle multi-objective robust formulations. It incorporates cutting-edge efficient methods for both the surrogate models construction and their forthcoming exploration/exploitation. Prime advantage of surrogate modeling in pSeven optimizer is its multi-resolution ability, which being supplemented with specific hierarchical optimization strategies, allows to address large-scale expensive optimization problems, which are difficult to consider with conventional methods.