Contour location via entropy reduction leveraging multiple information sources

Alexandre N. Marques  
Department of Aeronautics and Astronautics  
Massachusetts Institute of Technology  
Cambridge, MA 02139  
noll@mit.edu

Remi R. Lam  
Center for Computational Engineering  
Massachusetts Institute of Technology  
Cambridge, MA 02139  
rlam@mit.edu

Karen E. Willcox  
Institute for Computational Engineering and Sciences  
University of Texas at Austin  
Austin, TX 78712  
kwillcox@ices.utexas.edu

Abstract

We introduce an algorithm to locate contours of functions that are expensive to evaluate. The problem of locating contours arises in many applications, including classification, constrained optimization, and performance analysis of mechanical and dynamical systems (reliability, probability of failure, stability, etc.). Our algorithm locates contours using information from multiple sources, which are available in the form of relatively inexpensive, biased, and possibly noisy approximations to the original function. Considering multiple information sources can lead to significant cost savings. We also introduce the concept of contour entropy, a formal measure of uncertainty about the location of the zero contour of a function approximated by a statistical surrogate model. Our algorithm locates contours efficiently by maximizing the reduction of contour entropy per unit cost.

1 Introduction

In this paper we address the problem of locating contours of functions that are expensive to evaluate. This problem arises in several areas of science and engineering. For instance, in classification problems the contour represents the boundary that divides objects of different classes. Another example is constrained optimization, where the contour separates feasible and infeasible designs. This problem also arises when analyzing the performance of mechanical and dynamical systems, where contours divide different behaviors such as stable/unstable, safe/fail, etc. In many of these applications, function evaluations involve costly computational simulations, or testing expensive physical samples. We consider the case when multiple information sources are available, in the form of relatively inexpensive, biased, and possibly noisy approximations to the original function. Our goal is to use information from all available sources to produce the best estimate of a contour under a fixed budget.

We address this problem by introducing the CLoVER (Contour Location Via Entropy Reduction) algorithm. CLoVER is based on a combination of principles from Bayesian multi-information source optimization [1][2] and information theory [3]. Our new contributions are:

- The concept of contour entropy, a measure of uncertainty about the location of the zero contour of a function approximated by a statistical surrogate model.

• An acquisition function that maximizes the reduction of contour entropy per unit cost.
• An algorithm that locates contours of functions using multiple information sources via reduction of contour entropy.

This work is related to the topic of Bayesian multi-information source optimization (MISO) [9, 10]. Specifically, we use a statistical surrogate model to fit the available data and estimate the correlation between different information sources, and we choose the location for new evaluations as the maximizer of an acquisition function. However, we solve a different problem than Bayesian optimization algorithms. In the case of Bayesian optimization, the objective is to locate the global maximum of an expensive-to-evaluate function. In contrast, we are interested in the entire set of points that define a contour of the function. This difference is reflected in our definition of an acquisition function, which is fundamentally distinct from Bayesian optimization algorithms.

Other algorithms address the problem of locating the contour of expensive-to-evaluate functions, and are based on two main techniques: Support Vector Machine (SVM) and Gaussian process (GP) surrogate. CLOVER lies in the second category.

SVM [7] is a commonly adopted classification technique, and can be used to locate contours by defining the regions separated by them as different classes. Adaptive SVM [8, 10] and active learning with SVM [11, 12] improve the original SVM framework by adaptively selecting new samples in ways that produce better classifiers with a smaller number of observations. Consequently, these variations are well suited for situations involving expensive-to-evaluate functions. Furthermore, Dribusch et al. [14] propose an adaptive SVM construction that leverages multiple information sources, as long as there is a predefined fidelity hierarchy between the information sources.

Algorithms based on GP surrogates [15–20] use the uncertainty encoded in the surrogate to make informed decisions about new evaluations, reducing the overall number of function evaluations needed to locate contours. These algorithms differ mainly in the acquisition functions that are optimized to select new evaluations. Bichon et al. [15], Ranjan et al. [16], and Picheny et al. [17] define acquisition functions based on greedy reduction of heuristic measures of uncertainty about the location of the contour, whereas Bect et al. [18] and Chevalier et al. [19] define acquisition functions based on one-step look ahead reduction of quadratic loss functions of the probability of an excursion set. In addition, Stroh et al. [21] use a GP surrogate based on multiple information sources, under the assumption that there is a predefined fidelity hierarchy between the information sources. Opposite to the algorithms discussed above, Stroh et al. [21] do not use the surrogate to select samples. Instead, a pre-determined nested LHS design allocates the computational budget throughout the different information sources.

CLOVER has two fundamental distinctions with respect to the algorithms described above. First, the acquisition function used in CLOVER is based on one-step look ahead reduction of contour entropy, a formal measure of uncertainty about the location of the contour. Second, the multi-information source GP surrogate used in CLOVER does not require any hierarchy between the information sources. We show that CLOVER outperforms the algorithms of Refs. [15–20] when applied to two problems involving a single information source. One of these problems is discussed in Sect. 4, while the other is discussed in the supplementary material.

The remainder of this paper is organized as follows. In Sect. 2, we present a formal problem statement and introduce notation. Then, in Sect. 3 we introduce the details of the CLOVER algorithm, including the definition of the concept of contour entropy. Finally, in Sect. 4 we present examples that illustrate the performance of CLOVER.

2 Problem statement and notation

Let \( g : D \rightarrow \mathbb{R} \) denote a continuous function on the compact set \( D \subseteq \mathbb{R}^d \), and \( g_\ell : D \rightarrow \mathbb{R}, \ell \in [M] \), denote a collection of the \( M \) information sources (IS) that provide possibly biased estimates of \( g \). (For \( M \in \mathbb{Z}_+ \), we use the notation \([M] = \{1, \ldots, M\}\) and \([M]_0 = \{0, 1, \ldots, M\}\)). In general, we assume that observations of \( g_\ell \) may be noisy, such that they correspond to samples from the normal distribution \( \mathcal{N}(g_\ell(x), \lambda_\ell(x)) \). We further assume that, for each IS \( \ell \), the query cost function,

\[^1\text{The statistical model used in the present algorithm is the same introduced in [8], and we attempt to use a notation as similar as possible to this reference for the sake of consistency.}\]
We introduce the concept of \( c_\ell : D \to \mathbb{R}^+ \), and the variance function \( \lambda_\ell \) are known and continuously differentiable over \( D \). Finally, we assume that \( g \) can also be observed directly without bias (but possibly with noise), and refer to it as information source 0 (ISO), with query cost \( c_0 \) and variance \( \lambda_0 \).

Our goal is to find the best approximation, within a fixed budget, to a specific contour of \( g \) by using a combination of observations of \( g_\ell \). In the remainder of this paper we assume, without loss of generality, that we are interested in locating the zero contour of \( g \), defined as the set \( Z = \{ z \in D \mid g(z) = 0 \} \).

3 The CLoVER algorithm

In this section we present the details of the CLoVER (Contour Location Via Entropy Reduction) algorithm. CLoVER has three main components: (i) a statistical surrogate model that combines information from all \( M + 1 \) information sources, presented in Sect. [3.1] (ii) a measure of the entropy associated with the zero contour of \( g \) that can be computed from the surrogate, presented in Sect. [3.2], and (iii) an acquisition function that allows selecting evaluations that reduce this entropy measure, presented in Sect. [3.3]. In Sect. [3.4] we discuss the estimation of the hyperparameters of the surrogate model, and in Sect. [3.5] we show how these components are combined to form an algorithm to locate the zero contour of \( g \). We discuss the computational cost of CLoVER in the supplementary material.

3.1 Statistical surrogate model

CLoVER uses the statistical surrogate model introduced by Poloczek et al. [3] in the context of multi-information source optimization. This model constructs a single Gaussian process (GP) surrogate that approximates all information sources, presented in Sect. [3.1], (ii) a measure of the entropy defined by

\[
H(W) = -\sum_{i=1}^{k} P(w_i) \ln P(w_i),
\]

where \( P(w_i) \) denotes the probability mass of value \( w_i \). It follows from this definition that lower values of entropy are associated to processes with little uncertainty (\( P(w_i) \approx 1 \) for one of the possible outcomes).

We introduce the concept of contour entropy as the entropy of a discrete random variable associated with the uncertainty about the location of the zero contour of \( g \), as follows. For any given \( x \in D \), the posterior distribution of \( f(0, x) \) (surrogate model of \( g(x) \)), conditioned on all the available evaluations, is a normal random variable with known mean \( \mu(0, x) \) and variance \( \sigma^2(0, x) \). Given \( g(x) \in \mathbb{R}^+ \), an observation \( y \) of this random variable can be classified as one of the following three events: \( y < -\epsilon(x) \) (denoted as event \( L \)), \( |y| < \epsilon(x) \) (denoted as event \( C \)), or \( y > \epsilon(x) \) (denoted as event \( U \)). These three events define a discrete random variable, \( W_x \), with probability mass...
where

with each of them. In particular, $P$ measures uncertainty at parameter value $\ell$, and $V$ measures the uncertainty in whether $g(\ell)$ lies below, within, or above the tolerance $\epsilon(\ell)$, and is given by

$$H(W_\ell; f) = -P(L) \log P(L) - P(C) \log P(C) - P(U) \log P(U).$$

This entropy measures uncertainty at parameter value $\ell$ only. To characterize the uncertainty of the location of the zero contour, we define the contour entropy as

$$\mathcal{H}(f) = \frac{1}{V(D)} \int_D H(W_\ell; f) \, d\ell,$$

where $V(D)$ denotes the volume of $D$.

### 3.3 Acquisition function

CLOVER locates the zero contour by selecting samples that are likely to reduce the contour entropy at each new iteration. In general, samples from IS0 are the most informative about the zero contour of $g$, and thus are more likely to reduce the contour entropy, but they are also the most expensive to evaluate. Hence, to take advantage of the other $M$ IS available, the algorithm performs observations that maximize the expected reduction in contour entropy, normalized by the query cost.

Consider the algorithm after $n$ samples evaluated at $X_n = \{(\ell^i, x^i)\}_{i=1}^n$, which result in observations $Y_n = \{y^i\}_{i=1}^n$. We denote the posterior GP of $f$, conditioned on $\{X_n, Y_n\}$, as $f^n$, with mean $\mu^n$ and covariance matrix $\Sigma^n$. Then, the algorithm selects a new parameter value $\ell \in D$, and IS $\ell \in [M]_0$ that satisfy the following optimization problem.

$$\max_{\ell \in [M]_0, x \in D} u(\ell, x; f^n),$$

where

$$u(\ell, x; f^n) = \frac{\mathbb{E}_y[\mathcal{H}(f^n) - \mathcal{H}(f^{n+1}) | \ell^{n+1} = \ell, x^{n+1} = x]}{c_\ell(x)},$$

and the expectation is taken over the distribution of possible observations, $y^{n+1} \sim \mathcal{N}(\mu^n(\ell, x), \Sigma^n((\ell, x), (\ell, x)))$. To make the optimization problem tractable, the search domain is replaced by a discrete set of points $A \subset D$, e.g., a Latin Hypercube design. We discuss how to evaluate the acquisition function $u$ next.
Given that $f^n$ is known, $\mathcal{H}(f^n)$ is a deterministic quantity that can be evaluated from (4–5). Namely, $H(W_x; f^n)$ follows directly from (4), and the integration over $D$ is computed via a Monte Carlo-based approach (or regular quadrature if the dimension of $D$ is relatively small).

Evaluating $E_\nu[\mathcal{H}(f^{n+1})]$ requires a few additional steps. First, the expectation operator commutes with the integration over $D$. Second, for any $x' \in D$, the entropy $H(W_x; f^{n+1})$ depends on $y^{n+1}$ through its effect on the mean $\mu^{n+1}(0, x')$ (the covariance matrix $\Sigma^{n+1}$ depends only on the location of the samples). The mean is affine with respect to the observation $y^{n+1}$ and thus is distributed normally: $\mu^{n+1}(0, x') = \mathcal{N}(\mu^{n}(0, x'), \sigma^2(x'; \ell, x))$, where $\sigma^2(x'; \ell, x) = (\Sigma^n + \Sigma^m((0, x'), (\ell, x)))$, Hence, after commuting with the integration over $D$, the expectation with respect to the distribution of $y^{n+1}$ can be equivalently replaced by the expectation with respect to the distribution of $\mu^{n+1}(0, x')$, denoted by $E_{\nu}[\mu]$.

Third, in order to compute the expectation operator analytically, we introduce the following approximations.

\[ \Phi(x) \ln \Phi(x) \approx \sqrt{2\pi} \psi(x - \bar{x}), \]
\[ (\Phi(x + d) - \Phi(x - d)) \ln(\Phi(x + d) - \Phi(x - d)) \approx \sqrt{2\pi} \psi(x - d + \bar{x}) + \psi(x + d - \bar{x}), \]

where $\psi$ is the normal probability density function, $\bar{x} = \Phi^{-1}(e^{-1})$, and $c = \Phi(\bar{x}) \ln \Phi(\bar{x})$. Figure 2 shows the quality of these approximations. Then, we can finally write

\[ E_\nu[\mathcal{H}(f^{n+1}) | \ell^{n+1} = \ell, x^{n+1} = x] = \frac{1}{V(D)} \int_D E_\mu[H(W_x; f^{n+1}) | \ell^{n+1} = \ell, x^{n+1} = x] dx' \]
\[ \approx -\frac{c}{V(D)} \int_D \left[ \frac{1}{2} \left( \frac{\mu^{n}(0, x') + (-1)^{\ell} \epsilon}{\sigma(x'; \ell, x)} + (-1)^{\ell} \bar{x} \right)^2 \right] dx', \]

where

\[ \sigma^2(x'; \ell, x) = \Sigma^{n+1}((0, x'), (0, x')) + \sigma^2(x'; \ell, x), \]
\[ r^2(x'; \ell, x) = \frac{\Sigma^{n+1}((0, x'), (0, x'))}{\sigma^2(x'; \ell, x)}. \]

### 3.4 Estimating hyperparameters

Our experience indicates that the most suitable approach to estimate the hyperparameters depends on the problem. Maximum a posteriori (MAP) estimates normally perform well if reasonable guesses are available for the priors of hyperparameters. On the other hand, maximum likelihood estimates (MLE) may be sensitive to the randomness of the initial data, and normally require a larger number of evaluations to yield appropriate results.

Given the challenge of estimating hyperparameters with small amounts of data, we recommend updating these estimates throughout the evolution of the algorithm. We adopt the strategy of estimating the hyperparameters whenever the algorithm makes a new evaluation of ISO. The data obtained by evaluating ISO is used directly to estimate the hyperparameters of $\mu_0$ and $\Sigma_0$. To estimate the hyperparameters of $\mu_\ell$ and $\Sigma_\ell$, $\ell \in [M]$, we evaluate all other $M$ information sources at the same location and compute the biases $\delta_\ell = y_\ell - y_0$, where $y_\ell$ denotes data obtained by evaluating IS $\ell$. The biases are then used to estimate the hyperparameters of $\mu_\ell$ and $\Sigma_\ell$. 

---

Figure 2: Comparison between functions involving products of $\Phi$ and $\ln \Phi$ and approximations (8–9).
3.5 Summary of algorithm

1. Compute an initial set of samples by evaluating all $M + 1$ IS at the same values of $x \in D$. Use samples to compute hyperparameters and the posterior of $f$.

2. Prescribe a set of points $\mathcal{A} \subset D$ which will be used as possible candidates for sampling.

3. Until budget is exhausted, do:
   (a) Determine the next sample by solving the optimization problem (6).
   (b) Evaluate the next sample at location $x^{n+1}$ using IS $\ell^{n+1}$.
   (c) Update hyperparameters and posterior of $f$.

4. Return the zero contour of $\mathbb{E}[f(0, x)]$.

4 Numerical results

In this section we present three examples that demonstrate the performance of CLoVER. The first two examples involve multiple information sources, and illustrate the reduction in computational cost that can be achieved by combining information from multiple sources in a principled way. The last example compares the performance of CLoVER to that of competing GP-based algorithms, showing that CLoVER can outperform existing alternatives even in the case of a single information source.

4.1 Multimodal function

In this example we locate the zero contour of the following function within the domain $D = [-4, 7] \times [-3, 8]$.

$$g(x) = \frac{(x_1^2 + 4)(x_2 - 1)}{20} - \sin\left(\frac{5x_1}{2}\right) - 2.$$  \hspace{1cm} (11)

This example was introduced in Ref. [15] in the context of reliability analysis, where the zero contour represents a failure boundary. We explore this example further in the supplementary material, where we compare CLoVER to competing algorithms in the case of a single information source.

To demonstrate the performance of CLoVER in the presence of multiple information sources, we introduce the following biased estimates of $g$:

$$g_1(x) = g(x) + \sin\left(\frac{5}{22} (x_1 + \frac{x_2}{2}) + \frac{5}{4}\right), \quad g_2(x) = g(x) + 3 \sin\left(\frac{5}{11} (x_1 + x_2 + 7)\right).$$

We assume that the query cost of each information source is constant: $c_0 = 1$, $c_1 = 0.01$, $c_2 = 0.001$. We further assume that all information sources can be observed without noise.

Figure 3 shows predictions made by CLoVER at several iterations of the algorithm. CLoVER starts with evaluations of all three IS at the same 10 random locations. These evaluations are used to compute the hyperparameters using MLE, and to construct the surrogate model. The surrogate model is based on zero mean functions and squared exponential covariance kernels [22]. The contour entropy of the initial setup is $H = 0.315$, which indicates that there is considerable uncertainty in the estimate of the zero contour. CLoVER proceeds by exploring the parameter space using mostly IS2, which is the model with the lowest query cost. The algorithm stops after 123 iterations, achieving a contour entropy of $H = 4 \times 10^{-9}$. Considering the samples used in the initial setup, CLoVER makes a total of 17 evaluations of IS0, 68 evaluations of IS1, and 68 evaluations of IS2. The total query cost is 17.8. We repeat the calculations 100 times using different values for the initial 10 random evaluations, and the median query cost is 18.1. In contrast, the median query cost using a single information source (IS0) is 38.0, as shown in the supplementary material. Furthermore, at query cost 18.0, the median contour entropy using a single information source is $H = 0.19$.

We assess the accuracy of the zero contour estimate produced by CLoVER by measuring the area of the set $S = \{x \in D \mid g(x) > 0\}$ (shaded region shown on the top left frame of Figure 3). We estimate the area using Monte Carlo integration with $10^6$ samples in the region $[-4, 7] \times [-3, 8]$. We compute a reference value by averaging 20 Monte Carlo estimates based on evaluations of $g$: $\text{area}(S) = 36.5541$. Figure 4 shows the relative error in the area estimate obtained with 100 evaluations of CLoVER. This figure also shows the evolution of the contour entropy.
Figure 3: Locating the zero contour of the multimodal function (11). Upper left: Zero contour of IS0, IS1, and IS2. Other frames: Samples and predictions made by CLoVER at several iterations. Dashed black line: Zero contour predicted by the surrogate model. Colors: Mean of the surrogate model \( f(0, x) \). CLoVER obtains a good approximation to the zero contour with only 17 evaluations of expensive IS0.

Figure 4: Left: Relative error in the estimate of the area of the set \( S \). Right: Contour entropy. Median, 25, and 75 percentiles.

### 4.2 Stability of tubular reactor

We use CLoVER to locate the stability boundary of a nonadiabatic tubular reactor with a mixture of two chemical species. This problem is representative of the operation of industrial chemical reactors, and has been the subject of several investigations, e.g. [23]. The reaction between the species releases heat, increasing the temperature of the mixture. In turn, higher temperature leads to a nonlinear increase in the reaction rate. These effects, combined with heat diffusion and convection, result in complex dynamical behavior that can lead to self-excited instabilities. We use the dynamical model described in Refs. [24, 25]. This model undergoes a Höpf bifurcation, when the response of the system transitions from decaying oscillations to limit cycle oscillations. This transition is controlled by the Damköhler number \( D \), and here we consider variations in the range \( D \in [0.16, 0.17] \) (the...
bifurcation occurs at the critical Damköhler number $D_{cr} = 0.165$. To characterize the bifurcation, we measure the temperature at the end of the tubular reactor ($\theta$), and introduce the following indicator of stability.

$$g(D) = \begin{cases} 
\alpha(D), & \text{for decaying oscillations}, \\
(\gamma r(D))^2, & \text{for limit cycle oscillations}.
\end{cases}$$

$\alpha$ is the growth rate, estimated by fitting the temperature in the last two cycles of oscillation to the approximation $\theta \approx \theta_0 + \bar{\theta} e^{\alpha t}$, where $t$ denotes time. Furthermore, $r$ is the amplitude of limit cycle oscillations, and $\gamma = 25$ is a parameter that controls the intensity of the chemical reaction.

Our goal is to locate the critical Damköhler number using two numerical models of the tubular reactor dynamics. The first model results from a centered finite-difference discretization of the governing equations and boundary conditions, and corresponds to IS0. The second model is a reduced-order model based on the combination of proper orthogonal decomposition and the discrete empirical interpolation method, and corresponds to IS1. Both models are described in details by Zhou [24].

Figure 5 shows the samples selected by CLoVER, and the uncertainty predicted by the GP surrogate at several iterations. The algorithm starts with two random evaluations of both models. This information is used to compute a MAP estimate of the hyperparameters of the GP surrogate, using the procedure recommended by Poloczek et al. [3] and to provide an initial estimate of the surrogate. In this example we use covariance kernels of the Matérn class [22] with $\nu = 5/2$, and zero mean functions. After these two initial evaluations, CLoVER explores the parameter space using 11 evaluations of IS1. This behavior is expected, since the query cost of IS0 is 500-3000 times the query cost of IS1. Figure 5 shows the evolution of the contour entropy and query cost along the iterations. After an exploration phase, CLoVER starts exploiting near $D = 0.165$. Two evaluations of IS0, at iterations 12 and 14, allow CLoVER to gain confidence in predicting the critical Damköhler number at $D_{cr} = 0.165$. After eight additional evaluations of IS1, CLoVER determines that other bifurcations are not likely in

---

2For the length scales of the covariance kernels, Poloczek et al. [3] recommend using normal distribution priors with mean values given by the range of $D$ in each coordinate direction. We found this heuristics to be only appropriate for functions that are very smooth over $D$. In the present example we adopt $d_0 = 0.002$ and $d_1 = 0.0005$ as the mean values for the length scales of $\Sigma_0$ and $\Sigma_1$, respectively.

---

![Figure 5: Locating the Hopf bifurcation of a tubular reactor (zero contour of stability indicator). Shaded area: ±3σ around the mean of the GP surrogate. CLoVER locates the bifurcation after 22 iterations, using only 4 evaluations of IS0.](image-url)
the parameter range under consideration. CLoVER concludes after a total of 22 iterations, achieving \( H = 6 \times 10^{-9} \).

### 4.3 Comparison between CLoVER and existing algorithms for single information source

Here we compare the performance of CLoVER with a single information source to those of algorithms EGRA [15], Ranjan [16], TMSE [17], TIMSE [18], and SUR [18]. This comparison is based on locating the contour \( g = 80 \) of the two-dimensional Branin-Hoo function [26] within the domain \( \mathcal{D} = [-5, 10] \times [0, 15] \). We discuss a similar comparison, based on a different problem, in the supplementary material.

The algorithms considered here are implemented in the R package KrigInv [19]. Our goal is to elucidate the effects of the distinct acquisition functions, and hence we execute KrigInv using the same GP prior and schemes for optimization and integration as the ones used in CLoVER. Namely, the GP prior is based on a constant mean function and a squared exponential covariance kernel, and the hyperparameters are computed using MLE. The integration over \( \mathcal{D} \) is performed with the trapezoidal rule on a 50 × 50 uniform grid, and the optimization set \( \mathcal{A} \) is composed of a 30 × 30 uniform grid. All algorithms start with the same set of 12 random evaluations of \( g \), and we repeat the computations 100 times using different random sets of evaluations for initialization.

We compare performance by computing the area of the set \( S = \{ x \in \mathcal{D} \mid g(x) > 80 \} \). We compute the area using Monte Carlo integration with 10⁶ samples, and compare the results to a reference value computed by averaging 20 Monte Carlo estimates based on evaluations of \( g \): area(\( S \)) = 57.8137. Figure 7 compares the relative error in the area estimate computed with the different algorithms. All algorithms perform similarly, with CLoVER achieving a smaller error on average.
Acknowledgments

This work was supported in part by the U.S. Air Force Center of Excellence on Multi-Fidelity Modeling of Rocket Combustor Dynamics, Award FA9550-17-1-0195, and by the AFOSR MURI on managing multiple information sources of multi-physics systems, Awards FA9550-15-1-0038 and FA9550-18-1-0023.

References


