LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-PROC-709337

Data-driven Metric Learning for History Matching

J. Miller, J. J. Thiagarajan, P. T. Bremer, N. Hoda, D. Stern, R. Mifflin

November 15, 2016

Reservoir Simulation Conference
Montgomery, TX, United States
February 20, 2017 through February 22, 2017

Disclaimer

SPE International
Society of Petroleum Engineers

SPE-182683-MS

Data-driven Metric Learning for History Matching

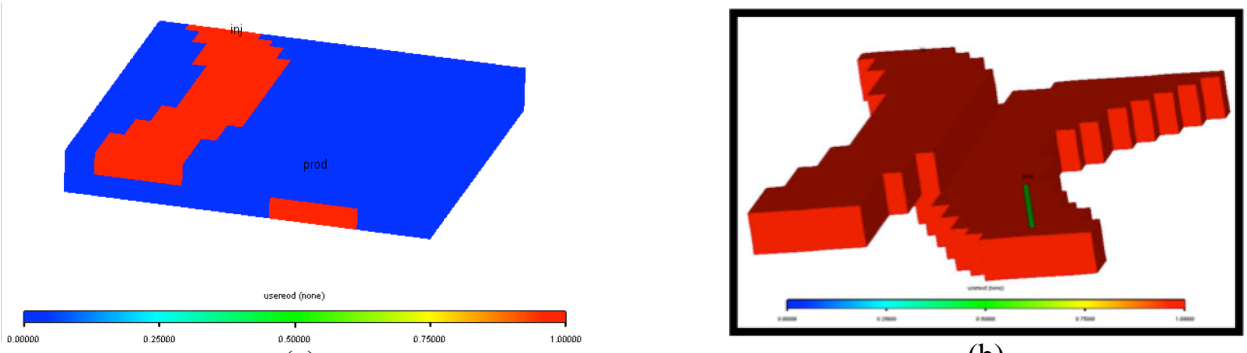
Jake Miller, University of California, Davis; Jayaraman J. Thiagarajan, Peer-TImo Bremer, Lawrence Livermore National Labs; Nazish Hoda, Dave Stern and Rick Mifflin, Exxon Mobil

Abstract

History matching, a highly non-unique inverse problem, is critical to calibrate model parameters in many scientific applications. A typical approach to history matching is to start with a uniform sampling of the high-dimensional parameter space and employ a surrogate modeling based black-box optimization to perform sequential sampling. Though this general workflow has been well studied, the problem of choosing an appropriate merit function to compare time-varying simulation outputs has been overlooked. Instead, convenient metrics such as the L2 or the L1-norm are employed. In this paper, we show that choosing an appropriate metric can significantly improve the solutions of sequential sampling. To this end, we propose a metric learning technique, develop a sequential sampling pipleline with the metric, and demonstrate its superiority to the conventional L2-norm metric.

1. Introduction

History matching is often used in reservoir simulation where model parameters are calibrated either manually or using an optimization method to match measured data that may include production data, 4D seismic, well logs, etc. The calibration exercise helps reduce the uncertainty in some of the parameters and potentially improve the predictive capability of models. One of the main challenges with history matching is that it is a highly non-unique inverse problem, i.e., different combinations of the model parameters can produce similar simulation results matching target measurements. Finding a single history matched model honoring measurements may not guarantee a reliable future performance prediction [Carter 2006]. It is crucial to perform uncertainty quantification based on all possible scenarios in order to ensure reliable predictions of future performance. Statistically, this amounts to constructing a posterior probability density function on all possible parameter settings, for a given target [Bliznyuk2008, Higdon2004]. For history matching, the input parameter space, both the number of parameters as well as their range, is very exhaustive and the input parameters have a nonlinear impact on the observations [Agbalaka2013]. To reasonably capture the structure (maxima, minima) in the observations, a typical sampling algorithm would require picking a very large number of parameter sets (may easily be of the order of thousands) to sample the parameter space and running expensive reservoir simulation for each sample. In most practical situations, running that many simulations is not possible.



(a) (b)

Fig 1. Geometry of the channel complexes (a) the color red represents the lithofacies assemblage LFA1, channel complexes, and blue indicates LFA0, the background LFA, (b) the model is filtered to show onlt the channel complexes.

The problem can be alleviated either by identifying a sampling algorithm that can effectively sample the multidimensional, large parameter space with fewer samples [White2003] or using a workflow where adaptive sampling is used to sample the parameter space starting with fewer initial samples [Keith2008, Li2014]. In this paper, we focus on the latter to perform sequential sample design, wherein a variety of merit functions are utilized to effectively explore the high-dimensional parameter space and to sample densely in regions that can potentially produce matching simulation results. Since this sequential process requires evaluation of the simulator at newly sampled points, a surrogate model is employed to predict the quality of match, which in turn guides the selection of the next set of samples.

Surrogate-based adaptive sampling approaches have previously been used by Li2015a; Li2015b. Li2015a used an adaptive importance sampling algorithm for Bayesian inversion with multimodal distributions. In this approach, the multimodal posterior is approximated using an adaptively constructed Gaussian mixture model and a surrogate model is constructued using polynomial chaos expansion. Alternately, Li2015b used an adaptively constructed Gaussian mixture model to approximate the posterior and a surrogate model is built using Gaussian process regression. The approach is tested on a non-linear history matching problem. Their approach is successful in capturing the multimodal posterior PDF of model parameters and is able to provide production prediction with uncertainty quantification.

For history matching, standard distance metrics such as the L2-norm or the L1-norm is used to quantify the mismatch between simulation results and measurements. The problem of choosing an appropriate metric for evaluating the quality of the match has been widely ignored. Surprisingly, as we demonstrate, the surrogate-based adaptive sampling is very sensitive to the chosen metric, particularly more severely when the simulation output is multivariate. In this paper, we propose a novel data-driven learning to determine the metric for evaluating the mismatch.

The rest of this paper is organized as follows: We describe the problem setup and an approach overview in Section 2. The novel data-driven metric learning is proposed in Section 3, and the application to history matching is discussed in Section 4. We present the results in Section 5 and finally summarize our findings in Section 6.

2. Problem Setup and Approach Overview

The problem considered in this paper is a two-well waterflood in a reservoir containing two stacked channel complexes. The model represents a deep-water slope channel system, in which sediment is deposited in channel complexes as a river empties into a deep basin. Each cell in the model is characterized by what lithofacies assemblage it belongs to. A lithofacies assemblage (LFA) is a collection of rocktypes associated with a particular geologic setting. In this model, there are two LFAs,

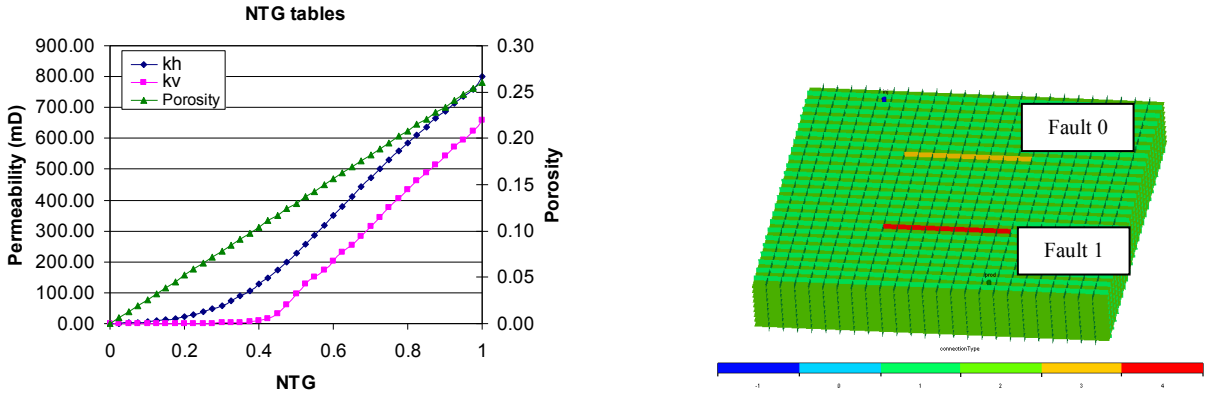


Figure 2. (a) Relationships between permeability, porosity and the net-to-gross ratio parameters, (b) Demonstration of the model (2 faults) considered in this paper.

channel complex and background – the sediment that was present before the channel complexes were deposited. Figure 1 shows the geometry of the channel complexes. In Figure 1(a), the color represents the LFAs in the model. In the figure, red represents channel complexes (LFA1), and blue is the background LFA (LFA0). Wells are also shown in the figure. Water is injected in a well labeled as *inj* at the northern end of the channel complexes, and oil and water are produced from a well labeled as *prod* located in the southern end of the reservoir. In Figure 1(b), the model has been filtered to show only the channel complexes. It is seen that they overlap only in a small part of the reservoir near the center. We use a reservoir simulator to compute a high-fidelity approximation to the exact forward model. In other words, given a set of input parameters living in some parameter space, P, the simulator outputs a triplet of time series corresponding to oil production rate (STB/day), water production rate (STB/day), and water injection pressure (psi) living in a curve space, C. The simulator can then be viewed as a function $f: P \rightarrow C$. The uncertain parameters that we adjust to improve agreement between the model result and the target are listed in Table 1. The parameters used here are representative of uncertain parameters commonly adjusted to calibrate models representing deep-water channel reservoirs in practical applications (Lun et al. SPE 159985 (2012)). Note that, our synthetic low-dimensional history matching problem may very well have a unique solution, but real history matching problems will always have an infinite number of solutions. Hence, we want to discover the topology near the target simulation, and any other reasonable local minima, since these will be relevant to the real-world case.

2.1. Model Parameters

In this model, we represent two channel complexes deposited at different times. The model is divided into two geologic units, one representing each channel complex. Unit 0, at the bottom, was deposited first. Unit 1, on top, contains a distinct channel complex that overlaps with the channel complex in Unit 0 near the center of the reservoir. Each channel complex is characterized by its' composition by specifying the net-to-gross ratio (NTG), defined as the ratio of the volume of permeable sand to total volume. Permeability and porosity are a function of NTG. The NTG is determined by the composition of the source sediment; because the two units were deposited at different times, NTG varies with both unit and LFA. NTG is determined from measurements in wells and from interpretation of seismic data. Therefore, there is some uncertainty in NTG. To represent properties in this reservoir model, there are four NTG parameters, representing the NTG associated with each unit-LFA pair. NTG00 is the NTG for LFA0 in Unit 0, NTG01 is the NTG for LAF1 in unit 0.

Once NTG is specified, the permeability and porosity are determined. To do this, we provide curves showing the relationship between permeability, porosity, and NTG. Permeability is anisotropic, so we have vertical permeability (Kv) and horizontal permeability (Kh). These relationships are shown in Figure 2(a). We assume that all the rock deposited in each LFA is characterized by a single relationship

Parameter	Minimum	Maximum	Scale	Target
flt3	0.0001	1.0	Logarithmic	0.105
flt4	0.0001	1.0	Logarithmic	0.105
NTG00	0.1	0.4	Linear	0.15
NTG01	0.7	0.9	Linear	0.75
NTG10	0.1	0.4	Linear	0.3
NTG11	0.7	0.9	Linear	0.9
Permmult0	0.5	2.0	Linear	0.5
Permmult1	0.5	2.0	Linear	2.0
Unit00	0.0001	1.0	Logarithmic	0.0001
Unit01	0.0001	1.0	Logarithmic	0.0001
Unit11	0.1	1.0	Logarithmic	0.1003

Table 1. Parameters with their corresponding upper and lower bounds and their scales within the range. In addition, we show the parameter values for the target simulation. Note that, for simplicity, we use only 7 of the 11 total parameters for our analysis.

between NTG and permeability and porosity. Because measurements of permeability and porosity are limited to a few samples taken at wells, there is uncertainty in these relationships. Therefore, we keep the relationship between permeability, porosity, and NTG as in Figure 2(a), but have a permeability multiplier that is applied to permeability in each LFA. Permmult0 applies to LFA0 and Permmult1 applies to LFA1.

In addition to setting permeability and porosity, we also need to account for thin shale zones that are not visible in seismic, but can have a significant impact on flow capacity. These are represented as transmissibility barriers. There are two types of barriers – barriers to vertical flow between geologic units, and barriers to flow across faults. Barriers to flow between units are caused by thin shales deposited between events that create channel complexes. The flow capacity of these barriers depends on LFA in each of the two units. LFA1 is coarser-grained and also more likely to erode shale as it is deposited. Therefore, areas where LFA1 in unit 1 is on top of LFA1 in unit0 are less subject to shale barriers, while the LFA0 on top of LFA0 is more likely to result in an impermeable barrier. Based on this logic, we have three transmissibility multipliers that are applied to cell faces at the interface between Unit 1 and Unit 0. The value of the multiplier depends on the LFA on either side of the cell face. We define three transmissibility multipliers: Unit00, Unit11, and Unit01. Unit00 applies to cells with LFA0 on both sides, Unit 11 applies to cells with LFA1 on both sides, and Unit01 applies to cells with LFA0 on one side and LFA1 on the other side. Barriers to flow across faults are represented as barriers to horizontal flow across selected cell faces. This model contains two faults, shown in Figure 2(b). The multipliers are flt3 and flt4.

2.2. Approach Overview

Our goal in history matching is to discover the set of uncertain model parameters, or distribution of possible parameters, that correspond to a given set of target time-series curves. Essentially, given $\mathrm{C}^* \in C$ we wish to estimate $f^{-1}(\mathrm{C}^*)$. Note that, for the rest of the paper, we will refer to C^* as the target curves and any $\boldsymbol{x}^* \in P$ such that $f(\boldsymbol{x}^*) = \mathrm{C}^*$ as the target parameters. For our analysis, we considered 7 out of the 11 total parameters in Table 1 for estimating $f^{-1}(\mathrm{C}^*)$: four NTG values (one for each Unit-LFA pair), two permeability multipliers (one for each EOD), and one fault transmissibility multiplier.

A typical solution to history matching is to adopt a surrogate-based adaptive sampling approach. Unlike several existing approaches, we construct the surrogate model to directly predict a measure of dissimilarity of a simulation output to the target curves C^*, in lieu of the actual time series output. In addition to simplifying the problem in higher dimensions (for example in [Li2015a] the authors only consider a 3D parameter space), this allows a regularized optimization with respect to a given target. However, this approach assumes knowledge of an appropriate dissimilarity function (a.k.a distance metric) $g: P \rightarrow \mathrm{R}^+$, where $g(\boldsymbol{x}) = d(f(\boldsymbol{x}), \mathrm{C}^*), for\ \boldsymbol{x} \in P$. Given that d is a true metric on the space C, we can pose the history matching problem as the following optimization problem:

$$\boldsymbol{x}^* = \ argmin\ g(\boldsymbol{x}) = argmin\ d(f(\boldsymbol{x}), \mathrm{C}^*) \tag{1}$$

In practice, popular Euclidean metrics such as the L2 distance are employed, without actually evaluating their suitability. In this paper, we show that we can obtain improved results for this non-unique inverse problem by building a suitable dissimilarity metric for effective comparison of the time series curves. Since evaluating the distance metric for a parameter configuration requires the execution of the simulator, surrogate-based adaptive sampling learns a surrogate model $\hat{g}$ using the initial set of samples. In particular, we develop a new adaptive sampling pipeline to sequentially add samples in order to improve the accuracy of the surrogate and thereby obtain solutions that better match the target.

3. Proposed Metric Learning

We consider the general problem of learning a metric for surrogate-based adaptive sampling assuming a simulator of the type $f: P \rightarrow C \subseteq R^n$ where $P \subseteq R^d$ and n denotes the length of the time series. Note that, in our case we concatenate the three time-series curves to represent the simulation output vector $\mathbf{C}$ and hence n corresponds to the length of $\mathbf{C}$. A Mahalonobis distance is a metric defined as [Bar-Hillel2006]

$$d_M(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{(\boldsymbol{a} - \boldsymbol{b})^T \boldsymbol{M} (\boldsymbol{a} - \boldsymbol{b})} \tag{2}$$

where $\boldsymbol{M} \in R^{n \times n}$ is a positive definite or positive semi-definite matrix (note in the latter case, the associated metric will only be a pseudo-metric as there may exist $\boldsymbol{a} \neq \boldsymbol{b}$ such that $d_{M(\boldsymbol{a},\boldsymbol{b})} = 0$.

We aim to learn a metric whose associated function dissimilarity function $g(\boldsymbol{x})$ is easily minimized and whose hopefully unique minimum lies at the oracle parameters. We propose to hence learn a metric such that distance between two outputs in the curve space will as accurately reflect the closeness of their corresponding parameter settings in the parameter space. This implies for each $\boldsymbol{a}, \boldsymbol{b} \in Im(f) \subseteq C$, with $\boldsymbol{a} = f(\boldsymbol{x})$ and $\boldsymbol{b} = f(\boldsymbol{y})$ and $\boldsymbol{x}, \boldsymbol{y} \in P$ we aim for

$$d_M(\boldsymbol{a}, \boldsymbol{b}) = \ d_M(f(\boldsymbol{x}), f(\boldsymbol{y})) \ \approx \ ||\boldsymbol{x} - \ \boldsymbol{y}||_2 \tag{3}$$

It is worth noting that $\boldsymbol{a}$ and $\boldsymbol{b}$ are time series and their dimension depend upon the frequency at which field measurement are performed. In order to adapt to varying importance of each parameter, rather than attempting to enforce equality of the learned metric to the standard Euclidean metric we will instead allow the metric in the parameter space to be a scaled Euclidean metric by attempting to match the equality

$$d_M^2(\boldsymbol{a}, \boldsymbol{b}) = \ d_M^2(f(\boldsymbol{x}), f(\boldsymbol{y})) = d_P^2(\boldsymbol{x}, \boldsymbol{y}) \tag{4}$$

where $\boldsymbol{P} \in R^{d \times d}$ is a positive semi-definitive diagonal matrix, which in turn amounts to $\mathbf{P}$ being diagonal with non-negative entries (when forced to be 1, it becomes the standard Euclidean metric), and

the metrics here are squared, so we are now comparing two values which are linear in the elements of the matrices.

With this aim in mind we can formulate a mathematical programming formulation to learn our metric. Given a uniformly distributed initial sample set, S_0, we attempt to enforce this equality for all pairs $\boldsymbol{x}, \boldsymbol{y} \in S_0$. However, since this is undoubtedly infeasible, we modify this construction in two ways. First, we only care to enforce this equality for pairs of parameters that are less than some tolerance value t. This implies that the we care more about the efficacy of the metric for parameter points that are a reasonable distance apart, and it reduces computational complexity by limiting the number of pairs to use (this is non-trivial since a reasonable initial sample set in higher dimensional spaces could number in the low thousands, so that the total number of pairs would be significantly large). Secondly, rather than expecting an exact solution we attempt to find a least-squares solution to these equalities. Thus we end up with the following formulation:

$$\min_{\boldsymbol{M},\boldsymbol{P}} \sum_{\substack{\boldsymbol{x},\boldsymbol{y} \\ ||\boldsymbol{x}-\boldsymbol{y}||_2<t}} [d_M^2(f(\boldsymbol{x}), f(\boldsymbol{y})) - d_P^2(\boldsymbol{x}, \boldsymbol{y})]^2 \; s.t. \; trace(\boldsymbol{P}) = 1, \boldsymbol{M}, \boldsymbol{P} \succcurlyeq 0, \boldsymbol{P} \; diagonal \tag{5}$$

In other words, the above formulation attempts to learn a Mahalonobis distance metric in the curve space and a scaled Euclidean distance (unlike the standard Euclidean distance we do not require the diagonal of $\boldsymbol{P}$ to be 1). On the other hand, there is no additional constraint on $\boldsymbol{M}$ except for its positive semi-definiteness. Note that, the constraint $trace(\boldsymbol{P})$ is required to avoid the optimization to result in the trivial zero solution for $\boldsymbol{P}$. It is often convenient to remove this constraint by reformulating this problem to

$$\min_{\boldsymbol{M},\boldsymbol{P}} \sum_{\substack{\boldsymbol{x},\boldsymbol{y} \\ ||\boldsymbol{x}-\boldsymbol{y}||_2<t}} [d_M^2(f(\boldsymbol{x}), f(\boldsymbol{y}) - d_P^2(\boldsymbol{x}, \boldsymbol{y})]^2 + (trace(\boldsymbol{P}) - 1)^2, \boldsymbol{M}, \boldsymbol{P} \succcurlyeq 0, \boldsymbol{M}, \boldsymbol{P} \; diagonal \tag{6}$$

Here, we further simplify computations by enforcing $\boldsymbol{M}$ to also be a diagonal matrix. In this case the objective function becomes a sum of squares of linear functions in the diagonal elements of $\boldsymbol{M}$ and $\boldsymbol{P}$, whose only constraints are their non-negativity. It is easy to see that any optimal solution to (5) is an optimal solution to (6) after scaling by a positive factor, and since the effectiveness of our metric in surrogate-based adaptive sampling should not be affected by positive scaling we will use the latter formulation. The reformulated problem in (6) can be efficiently solved using non-negative least-squares solvers. In order to remove noise introduced by small variations in the numerical simulations, given a pair of parameters $\boldsymbol{x}, \boldsymbol{y} \in S_0$, any coordinate i such that $|f(\boldsymbol{x})_i - f(\boldsymbol{y})_i| < 0.1$ was thresholded to 0 in the vector $f(\boldsymbol{x}) - f(\boldsymbol{y})$, without which it is common for the optimization process to overweight these coordinate values.

3.1. Effect of the Tolerance Parameter

The tolerance value is a notably hand-tuned value within this learning process, throughout this work a tolerance of $t = 1.0$ is used in the learned metrics. To justify this heuristic, given a learned metric with associated diagonal matrix $\boldsymbol{M}$ and tolerance value t, define the importance vector of $\boldsymbol{M}$, $IM(\boldsymbol{M}$,t) as

$$IM(\boldsymbol{M}, \mathrm{t}) = \left(\boldsymbol{M}_{ii} \frac{1}{k_t} \sum_{\substack{\boldsymbol{x},\boldsymbol{y} \\ ||\boldsymbol{x}-\boldsymbol{y}||_2<t}} (f(\boldsymbol{x})_i - f(\boldsymbol{y})_i)^2 \right)_{i=1}^{n} \tag{7}$$

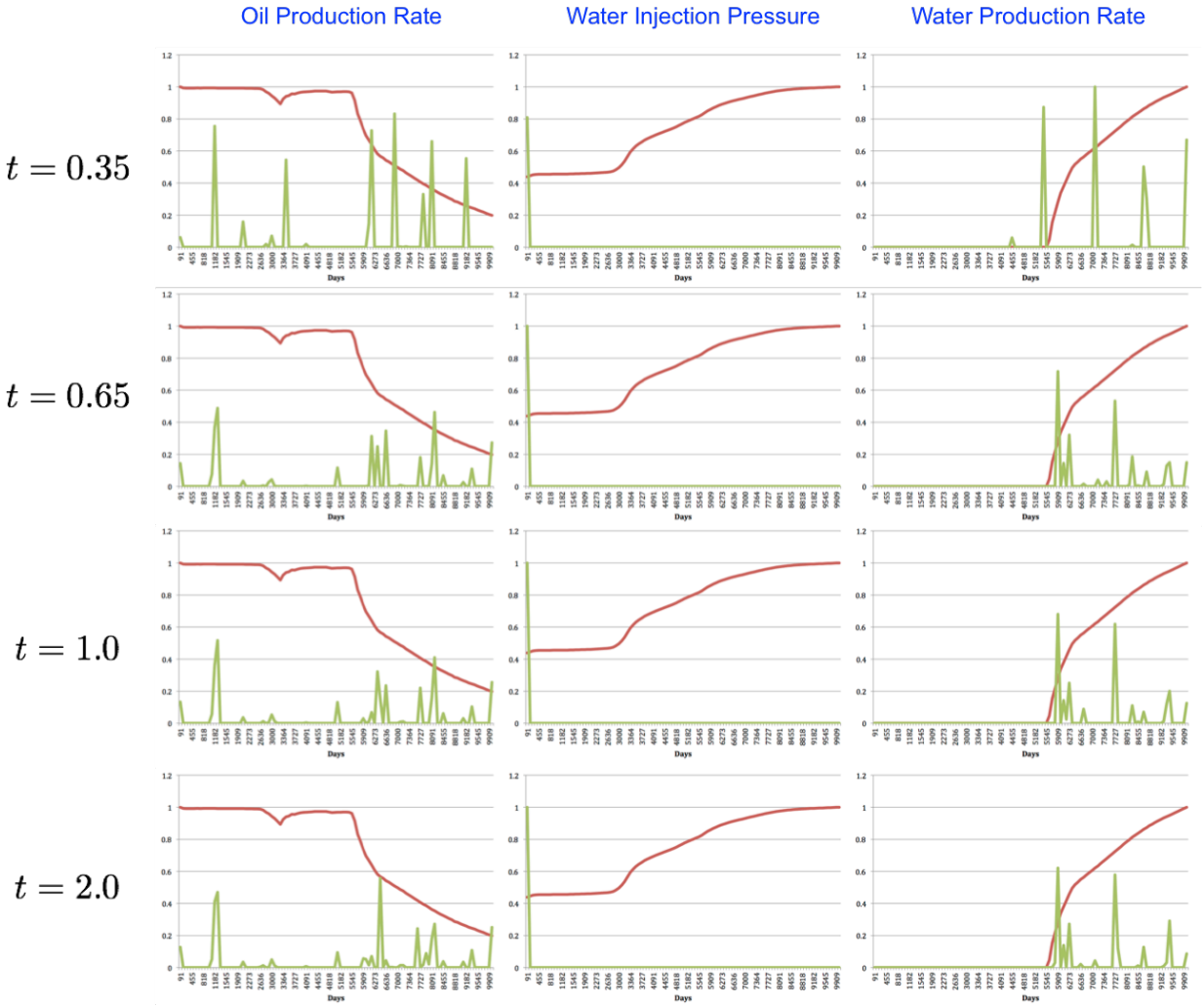


Figure 3. Shown are importance vectors at various tolerance values (t) scaled to maximum value of 1.0 (green curves) graphed over the corresponding portion of the oracle curve, each scaled to a maximum value 1.0 (red curves). The initial sample set used is the 1024 loosed box.

where $\boldsymbol{M}_{ii}$ is the i^{th} diagonal element of $\boldsymbol{M}$, and k_t is the number of pairs $\boldsymbol{x}, \boldsymbol{y} \in S_0$ such that $||\boldsymbol{x} - \boldsymbol{y}||_2 < t$. This corresponds to the importance of the i^{th} coordinate in the simulation output vector to the metric value $d_M^2(f(\boldsymbol{x}), f(\boldsymbol{y}))$. When the importance value is 0, it implies that the value of the curve at that discrete timestep is not considered during the metric computation. On the other hand, when $\boldsymbol{M}_{ii}$ is non-zero and large, the error $f(\boldsymbol{x})_i - f(\boldsymbol{y})_i$ will be penalized more. Note that, in all cases the average error with respect to all k_t pairs is considered. We note empirically that, given an initial sample set S_0, $IM(\boldsymbol{M}, \mathrm{t})$ shows convergence up to a scalar as t increases (scalar changes should correspond only to scalar changes in our metric, which in turn should not effect our adaptive sampling). In Figure 3 scaled importance vectors are graphed (superimposed on the oracle curves corresponding those coordinates) for $t = 0.35, 0.65, 1.0, 2.0$. The initial sample set used is a centroidal Voronoi tessellation (detailed in Section 4.1) of $[0,1]^7$ with 1024 samples projected to the parameter space with the bounds in Table 1. Each of the importance vectors are scaled so their maximum value is 1.0 (across all three oracle curves so that the values correspond to a scaled importance vector) and each of the oracle curves are scaled so that their maximum value is 1.0 (for each individual curves, these are used only for visualization and not for the actual analysis). As observed the scaled importance vectors show convergence, particularly in their support, as t increases and $t = 1.0$ is already similar to $t = 2.0$ in most time step. The number of pairs considered for each of the tolerance values grows rapidly; approximately 200, 40000, 260000, and

520000 pairs were considered for $t = 0.35, 0.65, 1.0, 2.0$, respectively. It is interesting to note that the number of pairs for $t = 2.0$ is just 3776 sample pairs shy of all possible pairs 1024 x (1024-1)/2 since $t = 2.0$ is close to the maximum tolerance value of $\sqrt{7} \approx 2.65$ set by the maximum distance between two points in $[0,1]^7$. Also notable in Figure 3 is the sparsity of the metrics (each of the metrics for tolerance values $0.35 - 2.0$ have a very limited support). This is due the characteristics of the data itself and not the metric learning technique, since no explicit sparsity constraint was placed. A striking feature of the learned metric in Figure 3 is that only the initial value of the water injection pressure is considered for the metric computation, while measurements at multiple timesteps describing the trends of the curves are included by the metric. Additionally, the metric learning predicts a higher weight at the water breakthrough time in agreement with the fact that matching water breakthrough time is important to attain a good history match.

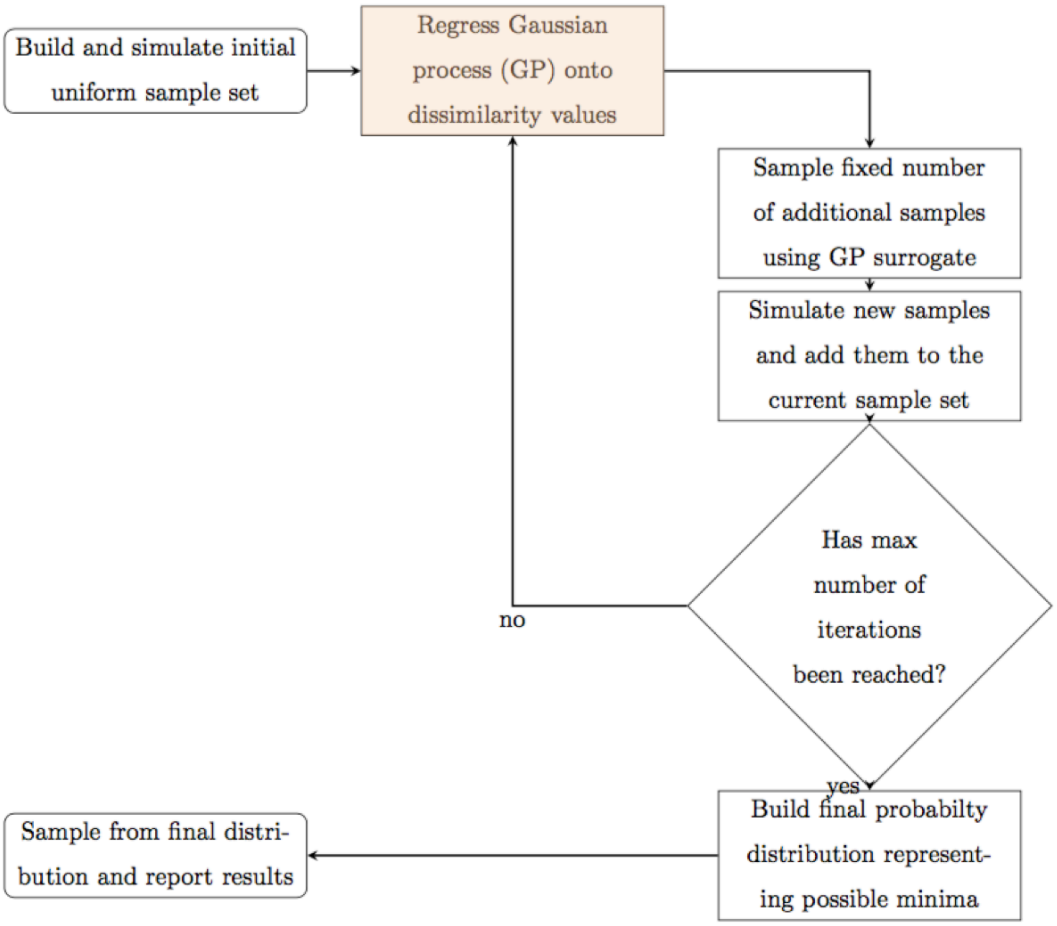


Figure 4. General workflow for the surrogate-based adaptive sampling optimization procedure

4. Application to History Matching

A natural approach to using the learned metric in history matching problem is to exploit the characteristics of the metric to perform adaptive sampling. A general workflow of this procedure is outlined in Figure 4. In contrast to existing approaches that build a surrogate model to predict the time-series curves directly using the parameter configurations, our model uses the dissimilarity measure with respect to a known target simulation as the response variable for prediction. While this allows the tuning

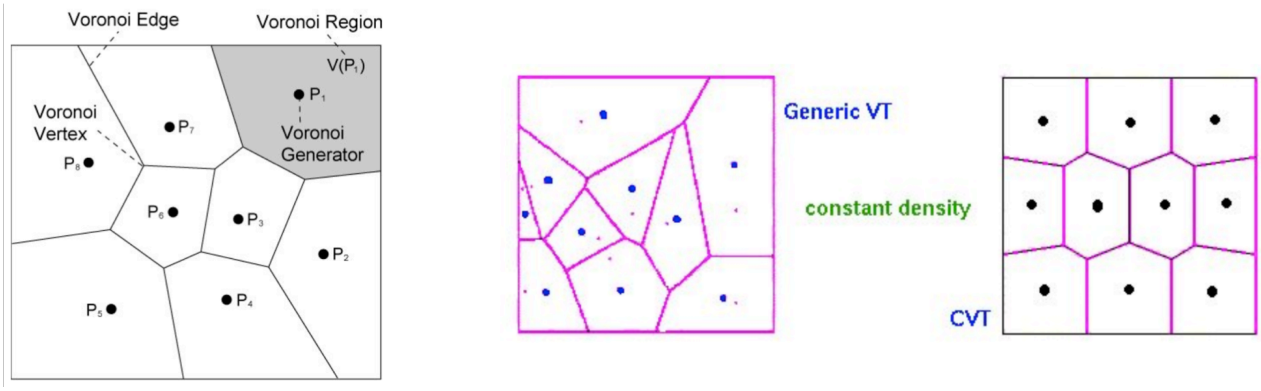


Figure 5. Definition of Voronoi tessellation and its comparison to centroidal Voronoi tessellation.

of the pipeline to a specific target (the target simulation can be switched out with any other simulation), this approach also leads to a robust optimization in high-dimensions. In general the surrogate model that adaptive sampling uses for regression can take numerous formulations, however here we fix our surrogate model to be a Gaussian Process (GP). Furthermore, we develop a new adaptive sampling strategy based on the efficacy of the surrogate model.

4.1. Initial Sampling and Surrogate Construction

In our adaptive sampling setup, the set of initial samples are obtained using centroidal Voronoi tessellation (CVT) [Du1999]. In general, a Voronoi diagram is a partition of the high-dimensional unit volume into convex polytopes. Note that, a tessellation of a flat surface refers tiling of a plane using convex polytopes, with no overlaps and no gaps. Each partition contains one generator such that every point in the partition is closer to its own generator than any other generator, where the Voronoi generators refer to the set of points that are used to form the distinct Voronoi regions. The CVT sample is a more restricted version of the Voronoi diagram definition. The constraint of the CVT sample is that each Voronoi generator must be the centroidal mass for its corresponding region. As shown with a 2-D example in Figure 5, the CVT sample has the generators exactly at the mass centroids of the convex polytopes.

Given the initial samples, we construct a surrogate model using Gaussian Processes since they are known to be highly effective in adaptive sampling. GP regression, also known as Kriging in some literature, is an interpolating regression method that extends multivariate Gaussian distributions to infinite dimensionality. Formally, a Gaussian process generates data located throughout a domain such that any finite subset of the range follows a multivariate Gaussian distribution. Given the input $\boldsymbol{x}$ and the response variable $\boldsymbol{C}$, regression attempts to infer the mapping function $\boldsymbol{C} = f(\boldsymbol{x})$. In GP regression, a Gaussian process is completely specified by its mean and covariance functions $f(\boldsymbol{x}) \approx GP(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$. A typical choice for the covariance function is the squared exponential, which corresponds to a Bayesian linear regressor. By Mercer's theorem, for every positive definite covariance matrix, there exists a linear expansion in terms of an infinite number of Gaussian basis functions. This representation of the covariance function implies a distribution over the basis functions. In other words, we can draw samples from the distribution of functions evaluated at any number of points. With the assumed form of the covariance function, GP regression solves for the mean and covariance functions using the training data.

4.2. Adaptive Sampling Pipeline

Often these surrogate models are optimized directly using either an unstructured black-box optimization method or, in the case that the regression has sufficient structure (convexity, differentiability, etc.) using a global or local optimization specific to that model. Rather than using one of these methods, which fixate on a single point whose accuracy assumes a high efficacy GP surrogate, we report a discrete

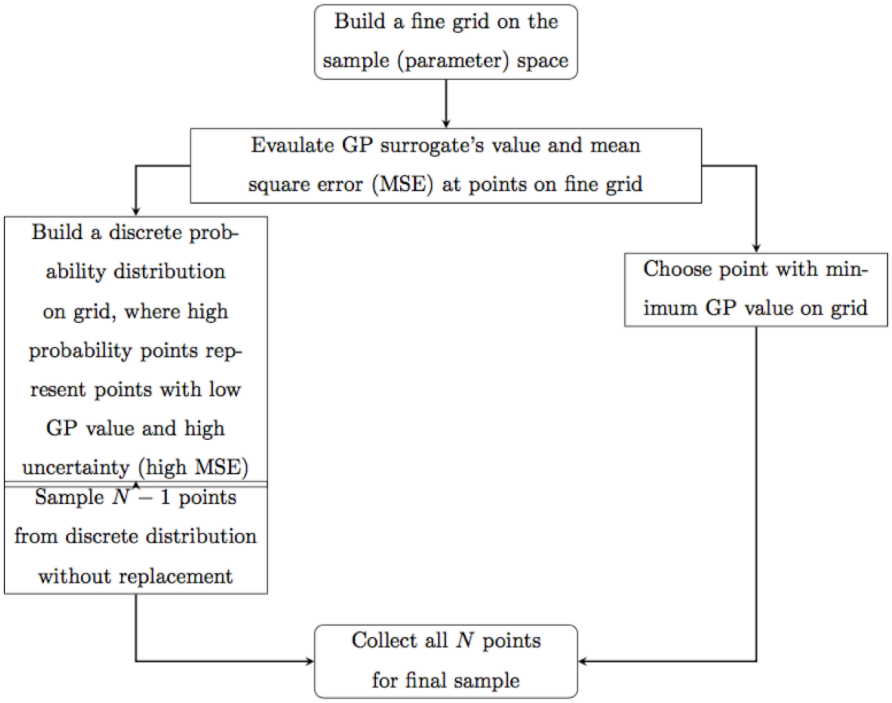


Figure 6. Proposed strategy for adaptive sampling.

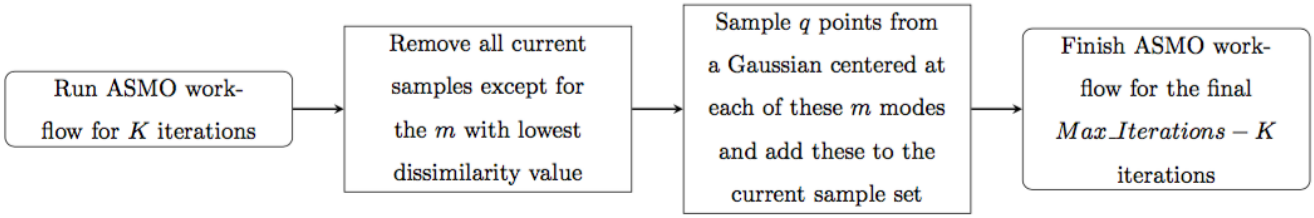


Figure 7. Modified standard adaptive sampling pipeline, including a purge step after a fixed number of iterations.

distribution of possible points near to possible local minima created using a combination of previously sampled points and the GP surrogate. Reporting our results in this manner reflects our continuing uncertainty of the surrogate model in approximating complex simulator even at the end of this workflow and also accounts for situations when there may be multiple oracle values associated with the given oracle curves, or at least curves similar to the oracle curves.

The strategy for adaptive sampling is crucial to properly training the GP surrogate. Recall that our goal is to train our surrogate $\hat{g}(\boldsymbol{x})$ so that the minima coincide with minima of our dissimilarity function $g(\boldsymbol{x})$ in both location and value. We have three goals in mind when adaptively including new samples: (i) find new minima in $g(\boldsymbol{x})$ that are currently not minima of $\hat{g}(\boldsymbol{x})$, (ii) refine the surrogate at points where minima of $\hat{g}(\boldsymbol{x})$ do not agree with those of $g(\boldsymbol{x})$, and (iii) increase efficacy of $\hat{g}(\boldsymbol{x})$ in the neighborhood of minima in $g(\boldsymbol{x})$. Figure 6 illustrates the proposed adaptive sampling strategy used at each iteration of the adaptive sampling pipeline. To initialize the sampling procedure, we build a fine grid on the parameter space that will represent a set of possible samples to take in that iteration. In order to either correct the surrogate, corresponding to our second goal, or verify its minima increase its efficacy there, corresponding to our third goal, we sample the point on the grid with the lowest GP surrogate value. This is shown on the right-hand side of Figure 6. For the remaining points to sample, we build a discrete probability distribution on this grid. This discrete probability mass function (pmf) will have larger mass at support points where the GP surrogate is lower yet its uncertainty is higher. The

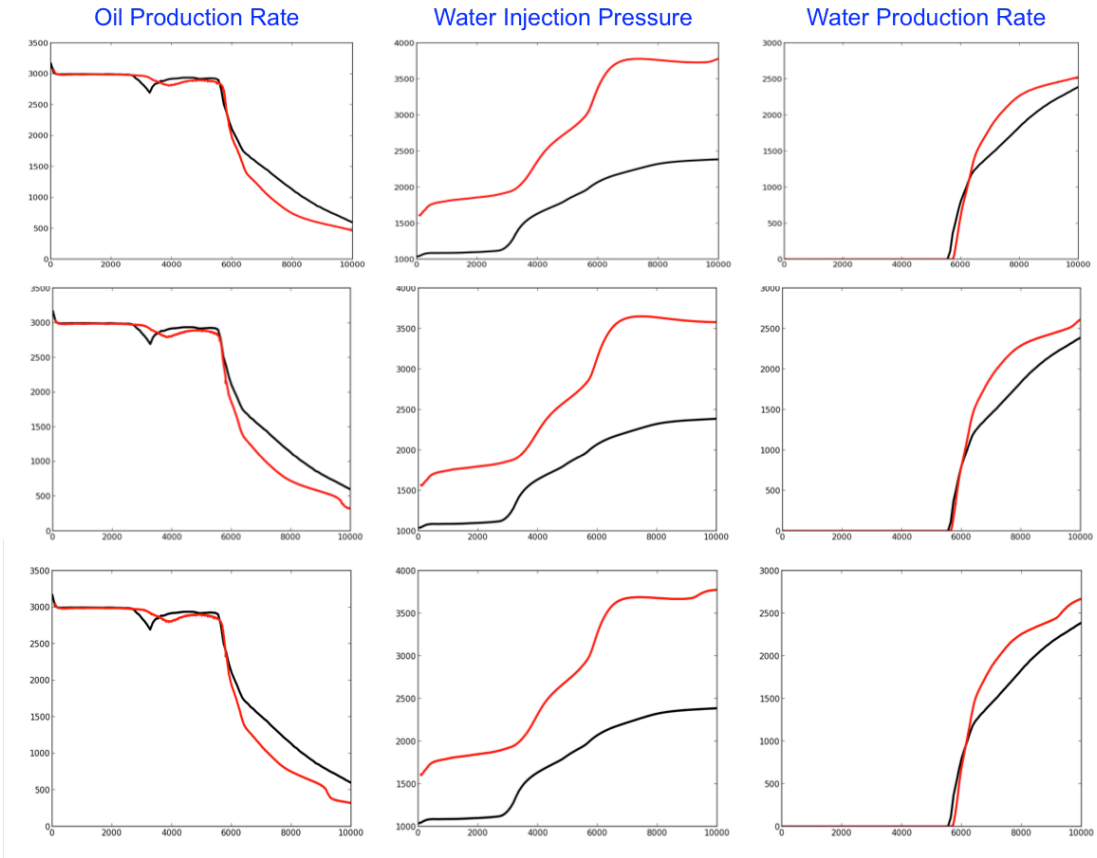


Figure 8. Performance of the L2 metric - The target curves (black) graphed against three best simulations by dissimilarity value obtained using the adaptive samplig pipeline.

mean-square error of the GP surrogate is used as a representative of uncertainty at the point, and the GP's value and MSE are weighted together such that these weight change at each iteration. At iterations where uncertainty of the surrogate is dominant, samples with high probability in the constructed pmf will satisfy our first goal, and at iterations where uncertainty is lower, high probability samples correspond to the other two goals. Finally we sample from this distribution without replacement to obtain our remaining sample points, as is shown in the left-hand side of Figure 6.

Finally, adaptive sampling suffers from the fact the sample set grows at each iteration, even when many of the samples many not be effective in training the surrogate model in the necessary areas, particularly the minima. Hence, we employ a two-stage procedure as shown in Figure 7 wherein we purge a vast majority of the samples during the middle of the workflow from the initial sample set and replace the initial sample set with what remains along with a dense collection of points around these few remaining samples. At the beginning of the second stage, we remove all the points from the current sample set except those with the m lowest dissimilarity values (set to 20 in our experiments). Then from each of these modes we sample q points from a multivariate Gaussian centered at that mode with covariance matrix $\sigma\mathbf{I}$. The points are all simulated and added to the current sample set before the workflow then continues.

5. Performance Evaluation

In this section, we evaluate the effectiveness of the proposed metric learning technique in the two-stage adaptive sampling pipeline described in Section 4. For comparison, we used the normalized L2-norm distance between the simulation outputs as the dissimilarity measure. Given a simulation output $\boldsymbol{C}$ and the target $\boldsymbol{C}^*$, the normalized L2 metric is obtained as $\frac{\|\boldsymbol{C_1}-\boldsymbol{C}^*\|_2}{\|\boldsymbol{C}^*\|_2}$. The initial sample sets used in all experiments were centroidal Voronoi tessellations of the unit cube [0,1], given a specific cardinality,

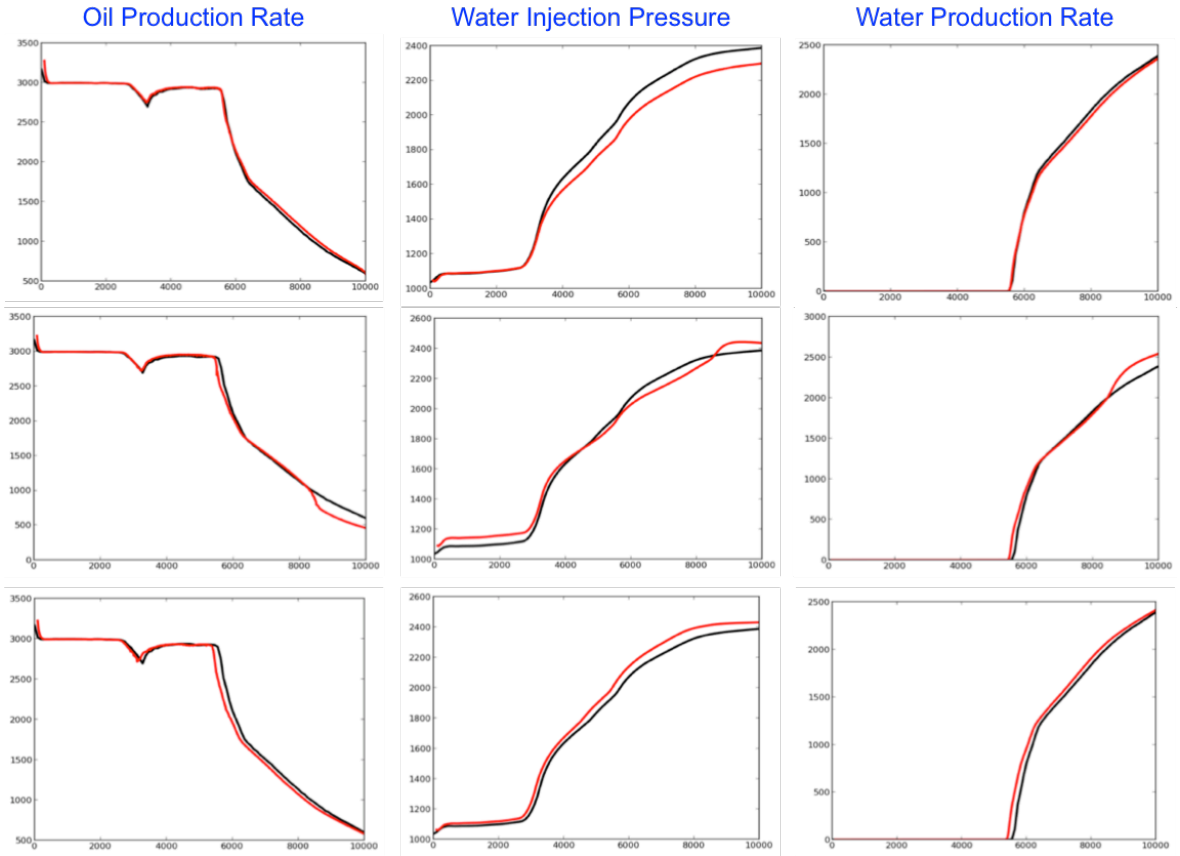


Figure 9. Performance of the proposed metric - The target curves (black) graphed against three best simulations by dissimilarity value obtained using the adaptive sampling pipeline.

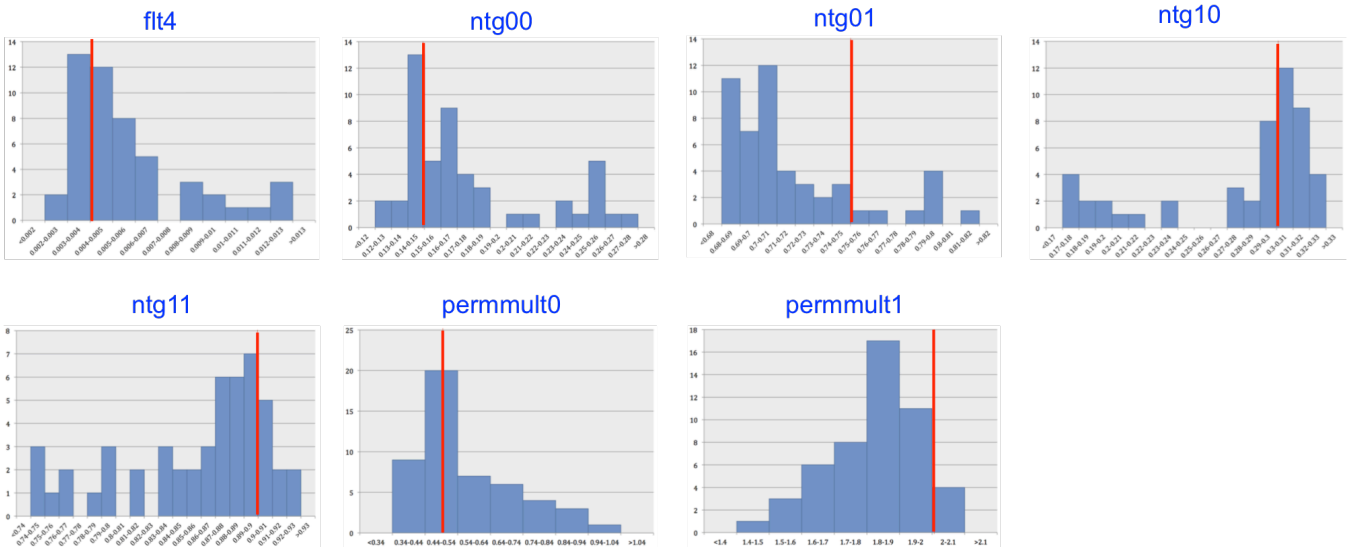


Figure 10. Histogram of parameters for the best matching simulations chosen by adaptive sampling based on the proposed metric. In each case, the target parameter is marked in red.

projected to the parameter ranges in Table 1. The learned metric used in the experiments was obtained by setting the tolerance value $t = 1.0$. The two-stage adaptive sampling pipeline was carried out for 20 iterations with 10 samples being added during each iteration.

In the case of both L2 and the proposed metrics, we used the same initial sample set and evaluated the quality of the samples obtained after adaptive sampling. While the quality of the initial samples can

impact the performance, in this paper, we focus only on the choice of the metric for fitting the GP surrogate. For evaluation, after completing the 20 iterations of adaptive sampling, we picked the three best solutions in terms of the dissimilarity to the target curves. Figure 8 shows the three solutions obtained using the L2 metric, while the results with the proposed metric are shown in Figure 9. The first observation is that the quality of proposed metric is significantly better with our metric, compared to a conventional Euclidean metric. Second, crucial features in the curves such as the dip in the oil production rate are never matched well with the L2 metric. This is indeed expected considering that in an L2 sense local discrepancies will not be reflected in the global dissimilarity measure. This can be prohibitive in history matching where compromising such local features can lead to completely different solutions in the parameter space. On the other hand, our metric in Figure 3 captures the crucial features used in characterizing the local discrepancies. This clearly demonstrates the importance of an appropriate metric while building surrogate models, which is often ignored in history matching applications. The increased efficacy of the matching directly improves the reliability of the corresponding parameter values in characterizing the behavior of the parameter space. In other words, we compute the histograms of the 7 parameters using only the best matching simulations (top 50 simulations sorted by the dissimilarity metric) from the adaptive sampling pipeline. The results in Figure 10 show that for most parameters the histograms reveal a mode near the target parameter value (marked in red) and for a few parameters the histogram shows non-trivial probabilities for values different from that of the target. From an analysis standpoint, this is an early evidence of the existence of multiple modes, atleast in few of the dimensions, in the optimization surface (obtained using the metric) in the parameter space. Further investigation of this behavior is beyond the scope of this paper and hence reserved for future work.

6. Conclusions

In this paper, we presented a novel metric learning algorithm for comparing time-varying measurements and argued its applicability in history matching. Furthermore, we developed an adaptive sampling pipeline based on the dissimilarity metric with respect to a target simulation. Though the metric is learned independent of the actual target from a limited number of simulations, we demonstrated that it could have a significant impact on the quality of samples (in terms of matching the target) inferred during sampling. This performance improvement can be attributed to the ability of the learned metric to effectively characterize even local discrepancies in the simulation outputs. Finally, comparisons with the conventional L2 metric revealed that an appropriate metric is crucial to the success of black-box optimization approaches typically adopted in history matching.

7. Acknowledgements

This work was performed under the auspices of the U.S. Dept. of Energy by Lawrence Livermore National Labora- tory under Contract DE-AC52-07NA27344.

8. References

Carter, J., Ballester, P., Tavassoli, Z., and King, P. (2006), Our calibrated model has poor predictive value: An example from the petroleum industry, Reliability Engineering and System Safety, 91(10-11), 1373-1381.

Higdon, D., Kennedy, M., Cavendish, J.C., Cafeo, J.A., Ryne, R.D. (2004), Combining field data and computer simulations for calibration and prediction, SIAM J. Sci. Comput. 26 (2) 448-466.

Bliznyuk, N., Ruppert, D., Shoemaker, C., Regis, R., Wild, S., Mugunthan, P. (2008), Bayesian calibration and uncertainty analysis for computationally expensive models using optimization and radial basis function approximation, J. Comput. Graph. Stat. 17 (2) 270-294.

Agbalaka, C. C., Stern, D., Oliver, D. S. (2013), Two-stage ensemble-based history matching with multiple modes in the objective function. Comput Geosciences; 55:28–43

White, C. D., & Royer, S. A. (2003). Experimental Design as a Framework for Reservoir Studies. Society of Petroleum Engineers. doi:10.2118/79676-MS.

Keith, J.M., Kroese, D.P., Sofronov, G.Y. (2008), Adaptive independence samplers, Stat. Comput. 18 (4) 409-420.

Li. J., Marzouk, Y.M. (2014), Adaptive construction of surrogates for the Bayesian solution of inverse problems, SIAM J. Sci. Comput. 36 A1163–A1186.

Li, W., Zhang, D., and Lin, G. (2015), A surrogate-based adaptive sampling approach for history matching and uncertainty quantification, SPE-173298-MS.

Li, W., and G. Lin (2015), An adaptive importance sampling algorithm for Bayesian inversion with multimodal distributions, J. Comput. Phys., 294, 173-190.

Bar-Hillel, A., Hertz, T., Shental, N. and Weinshall, D. (2006), Learning a Mahalanobis metric from equivalence constraints, Journal of Machine Learning Research, 6(1):937–965.

Du, Q., Faber, V., Gunzburger, M. (1999), Centroidal Voronoi Tessellations: Applications and Algorithms, SIAM Review, 41, 637-6796.