SAND2017-9526R

Bayesian Regression of Thermodynamic Models of Redox Active Materials

Katherine Johnston
kjohnsto@caltech.edu

September 1, 2017

Abstract

Finding a suitable functional redox material is a critical challenge to achieving scalable, economically viable technologies for storing concentrated solar energy in the form of a defected oxide. Demonstrating effectiveness for thermal storage or solar fuel is largely accomplished by using a thermodynamic model derived from experimental data. The purpose of this project is to test the accuracy of our regression model on representative data sets. Determining the accuracy of the model includes parameter fitting the model to the data, comparing the model using different numbers of parameters, and analyzing the entropy and enthalpy calculated from the model. Three data sets were considered in this project: two demonstrating materials for solar fuels by water splitting and the other of a material for thermal storage. Using Bayesian Inference and Markov Chain Monte Carlo (MCMC), parameter estimation was preformed on the three data sets. Good results were achieved, except some there was some deviations on the edges of the data input ranges. The evidence values were then calculated in a variety of ways and used to compare models with different number of parameters. It was believed that at least one of the parameters was unnecessary and comparing evidence values demonstrated that the parameter was need on one data set and not significantly helpful on another. The entropy was calculated by taking the derivative in one variable and integrating over another. and its uncertainty was also calculated by evaluating the entropy over multiple MCMC samples. Afterwards, all the parts were written up as a tutorial for the Uncertainty Quantification Toolkit (UQTk).

Introduction of the model

A thermodynamic model was derived from experimental data and is used to demonstrate the effectiveness for thermal storage or solar fuel. This model has three variables and seven free parameters. The variables are χ, a unitless temperature variable, u, a unitless pressure variable, and z, a unitless variable of the variable δ. This model has not yet been published, so it is not included here.

Model fitting

The first part of the project was to show that this model was demonstrative of the needed applications. The model has eight parameters: seven model parameters and σ as a hyper-parameter. Three different data sets, gathered in indpendent ways and from two different applications, were used to show this. The Zinkevich Ceria and Panlener data sets were from a material used for solar fuels by splitting water and carbon dioxide, while the Goldyreva data set was generated from a material for thermal storage. To show that the given model demonstrated this data, the ideal parameters were found using Bayesian Inference and Markov Chain Monte Carlo (MCMC) and found that these ideal parameters fit the data very well.

Bayesian Inference

Bayesian Inference is a method for determining model parameters by calibrating against a set of data points. Bayes formula is :

$$p(\theta|D, M) \propto p(D|\theta, M)p(\theta|M)$$

where $p(\theta|D, M)$ is the posterior distribution, $p(D|\theta, M)$ is the likelihood, and $p(D|\theta, M)$ is the prior. The ideal parameters for a given data set are found by maximizing the posterior.

Monte Carlo Markov Chain

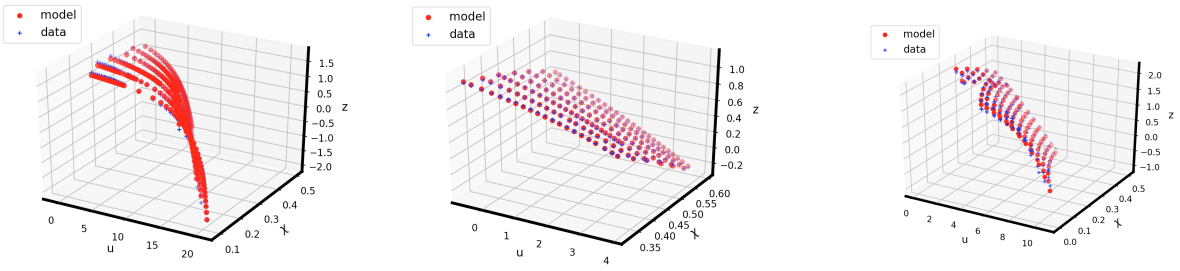
Monte Carlo Markov Chain (MCMC) is a method to sample the posterior distribution that utilizes Bayesian Inference. First, start at a given point and find the posterior probability, then select a second point based on a uniform distribution and again find the posterior probability. Then calculate α as the posterior of the new divided the old posterior:

$$\alpha = \frac{p(new|D)}{p(old|D)}.$$

If $\alpha > 1$, the new point is more likely and accept the new point. If $\alpha < 1$, then accept the new point at a probability of α. Continue this process for the desired amount of samples. This explores all areas of the given space, while focusing on areas that are more likely and have larger posteriors.

Results from model fitting

The model fitting using MCMC produced very good results. Here are some graphs from the different data sets.



a: Zinkevich Ceria b: Goldyreva c: Panlener

Figure 1: Model fitting graphs

The model fits the data very well and is fitting the Goldyreva data set best. For the Zinkevich Ceria and Panlener data sets, the model is deviating from the data around the edges of the data ranges, which maybe a concern if the model is extended far from this data ranges. To make the model fit each data set, some slight twicking was involved, including modifying initial conditions and input parameters.

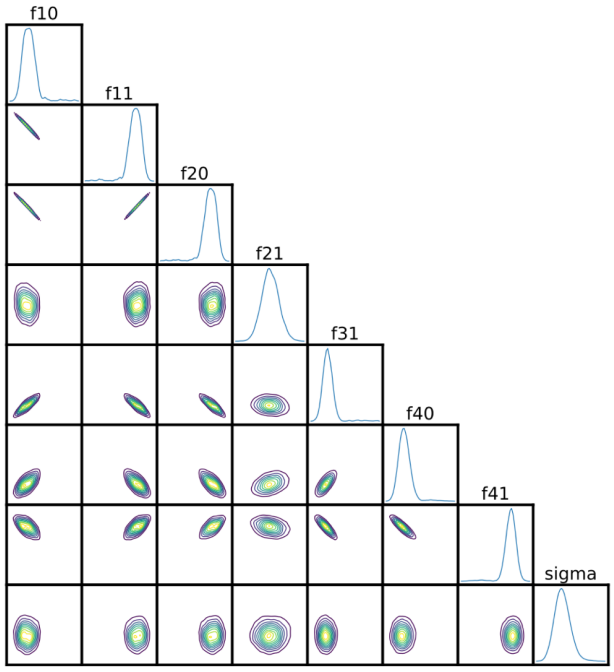


Figure 2: Zinkevich Ceria Posterior Plots for all Parameters and Correlations Between Parameters

This graph displays the posterior plots for each parameter and the correlations between parameters. The straighter lines between the parameters indicate correlation and the more circular the more independent the parameters are. From the graph, f_{10} and f_{11} form a very thin straight line and are very correlated, while sigma and all other variables form very circular shapes and appear to be mostly independent.

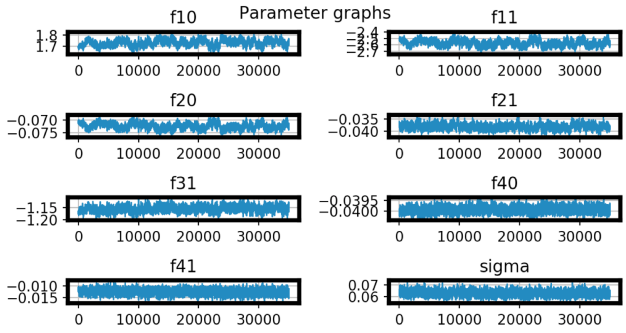


Figure 3: Zinkevich Ceria Parameter Graphs

This figure is an example of the parameter graphs for the Zinkevich Ceria data set. These graphs are taken after a burn in and with a stride because MCMC takes many steps to find the interesting part of the space and then often takes a few steps to move, thus there are often repeat points in a row. The parameters jump around a lot, which is good because the MCMC is exploring everywhere around the peak.

Model comparison

The model has 7 parameters, but not all the parameters may be necessary for the model, especially the parameter f_{21}. To demonstrate the best model fit, a bunch of models with different number of parameters were compared by calculating and comparing model evidence values.

Methods to calculate evidence values

To calculate the model evidence values, a few different methods were used. The first method was using Transitional Markov Chain Monte Carlo (tMCMC). For this method, the parameter values were calculated using tMCMC and then from those samples the evidence value was calculated. Then the evidence values were calculated from the regular MCMC in three ways: Harmonic, Gaussian, and Importance Sampling. Of the four methods, Harmonic is the most inconstant; it has large variance and is often significantly different from the other methods. The Gaussian and Importance Sampling methods agree pretty well with each other, after a burnin value. The method that utilizes tMCMC was also inconstant, but, unlike the Harmonic method, is in the same ballpark as the Gaussian and Importance Sampling methods. When using evidence values to determine the ideal number of parameters, the methods utilizing Importance Sampling and Gaussian were used.

Results

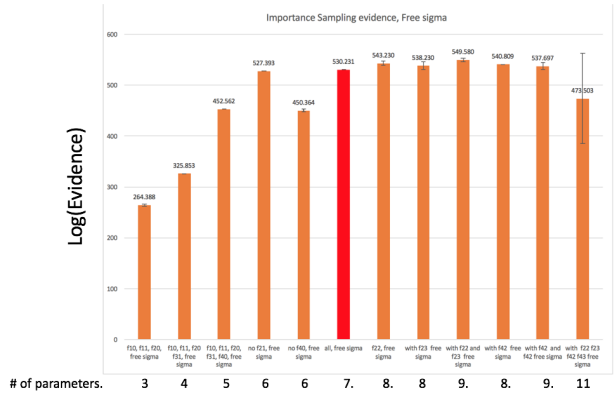


Figure 4: Goldyreva Evidence graph, by importance sampling

Here is a graph displaying the evidence values of the Goldyreva data set with models of varying number of parameters. The original model is colored in red. The graph shows that the model without f_{21} is approximately as good as the models with more parameters. The models with high number of parameters, also do not have a significantly higher evidence value. Thus, the models with higher parameter values do not provide a significantly better fit. The higher parameter values also provide more complicated models that make further calculations more complicated without providing a better fit. Also, at high numbers of parameters, there is larger variation in the evidence values.

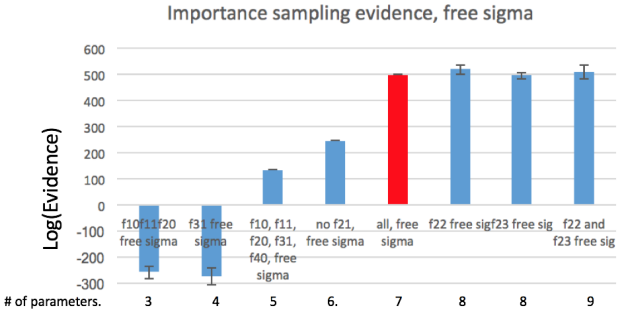


Figure 5: Zinkevich Ceria Evidence graph, by importance sampling

Here is a graph displaying the evidence values of the Zinkevich Ceria data set with models of varying number of parameters. Again, the original model is colored in red. The graph shows that the model without f_{21} is not a significantly better fit, implying that, unlike the Goldyreva data set, the parameter f_{21} is necessary. But like the Goldyreva data set, a higher number of parameters does not significantly improve the model and only increases complexity.

Finding entropy and enthalpy

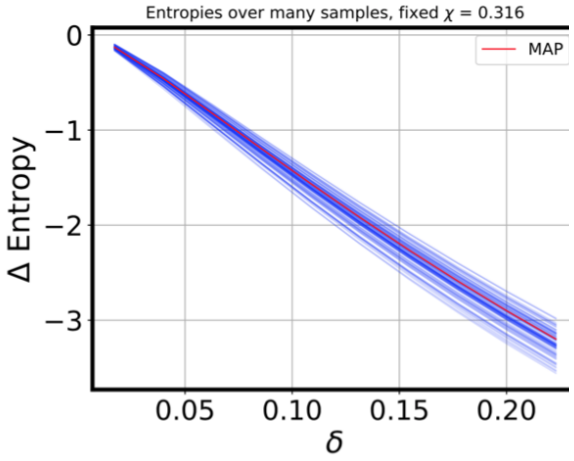
After verifying that this was the optimal model for the given data set, the next step was to calculate the entropy and enthalpy using the model. The change in entropy (ΔS)

was calculated by taking the derivative in one variable (δ) and then integrating over another variable (χ). ΔS was calculated using a numerical integration. When performing the integration, χ is treated as a constant (constant T integration). ΔS is calculated for a given number of fixed χ and δ values, and the corresponding ΔS values are graphed.

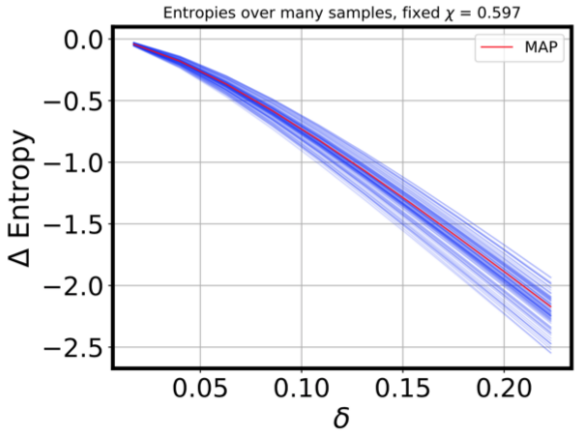
To find the uncertainty of ΔS, a desired number of samples of the parameters from the MCMC chain are taken and the entropies are calculated and graphed. Each of the blue lines in the below graphs show ΔS as calculated from different samples of the MCMC chain.

Goldyreva data

Figure 6: Goldyreva Δ Entropy with fixed χ



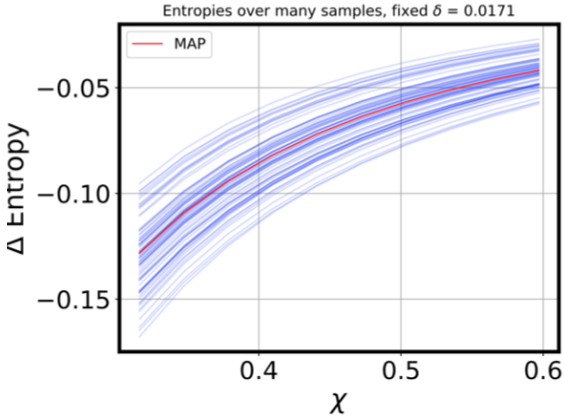
a: Goldyreva Δ Entropy with fixed $\chi = 0.316$, (lowest χ value)



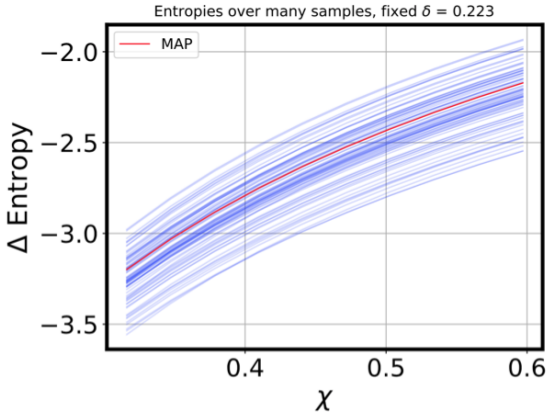
b: Goldyreva Δ Entropy with fixed $\chi = 0.597$, (highest χ value)

Here are two entropy graphs from the goldyreva data set, taking the highest and lowest value of χ. The graphs were taken with a fixed χ over all the δ values that the goldyreva data set spanned.

Figure 7: Goldyreva Δ Entropy with fixed δ



a: Goldyreva Δ Entropy with fixed $\delta = 0.0171$, (lowest δ value)

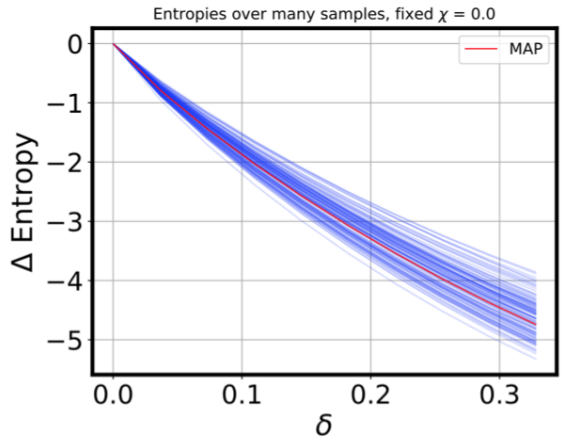


b: Goldyreva Δ Entropy with fixed $\delta = 0.223$, (highest δ value)

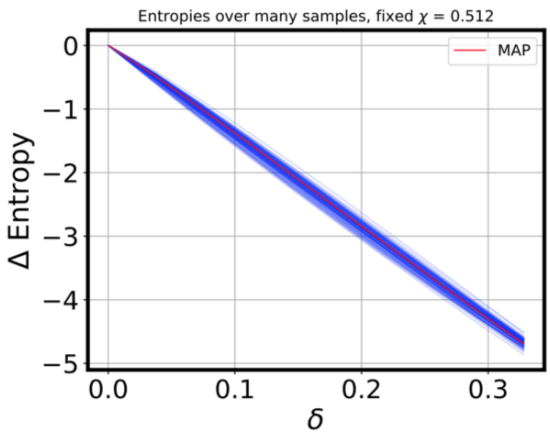
Here are some entropy graphs that are when δ is fixed and χ is varying. These graphs are of the highest and lowest δ values, and taken over the entire domain of χ that the goldyreva data encompasses. These graphs have a different trend then when χ is fixed and δ is varying. Also, the scale and ranges of the two graphs are quite different, which is different than when χ is fixed, which only changes in range a small amount.

Panlener ceria data

Figure 8: Panlener Δ Entropy with fixed χ



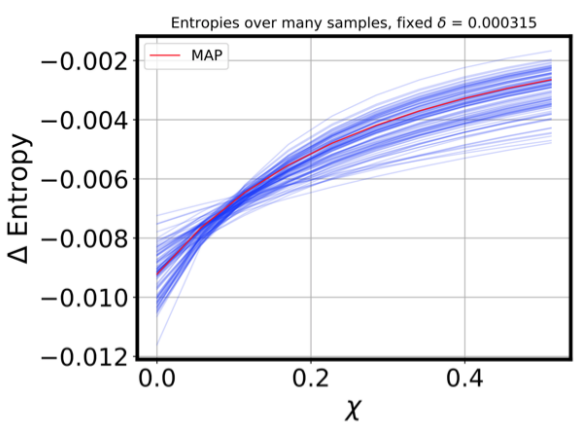
a: Panlener Δ Entropy with fixed $\chi = 0.0$, (lowest χ value)



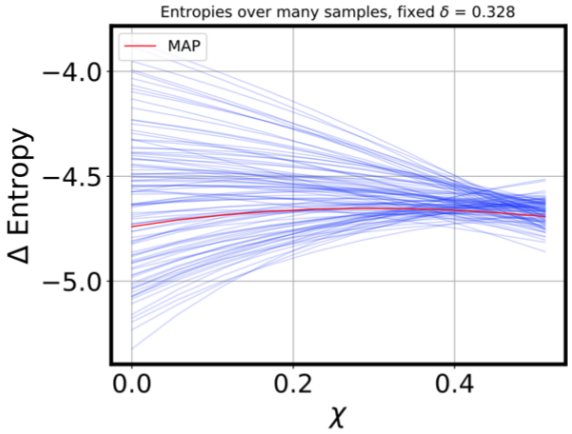
b: Panlener Δ Entropy with fixed $\chi = 0.512$, (highest χ value)

Here is the Panlener entropy graphs over a fixed χ. These are the highest and lowest chi values. The overall trend doesn't seem to change, but the uncertainty does. When χ is large, there is very little variance in the different MCMC samples; however, when χ is small, there is a considerable amount of variance, especially when δ is larger.

Figure 9: Panlener Δ Entropy with fixed δ.



a: Panlener Δ Entropy with fixed $\delta = 0.000315$, (lowest δ value)



b: Panlener Δ Entropy with fixed $\delta = 0.328$, (highest δ value)

Over the fixed δ, the Panlener entropy values are showing larger uncertainties than the others. This might present a problem because the uncertainty is significantly larger than the change in the MAP parameter values. It is believed that the when $\delta = 0.328$ it is passed the phase transition; so, it doesn't have any physical meaning and why the graph is so messy.

Zinkevich ceria data

When calculating entropy for the Zinkevich ceria data set, the results were very inconstant and different from the other data sets. A very large nagative results was sometime occurring, suggesting that a division by zero, or really small number, was occurring. Further work needs to be done to determine the cause of this problem.

Conclusion

Overall, this project was very productive and produced very good results. Good results were obtained by the model fitting by using MCMC for parameter estimate. The evidence values were able to be calculated and models were compared, and some of these results were expected. Entropy was calculated, but still has further work to be done because there is some modifications, such as the sign of ΔS may be incorrect. Reliable results of the Zinkevich ceria data set were not obtained and Enthalpy still needs be calculated.

These model fitting using MCMC will be written up for a tutorial of the Uncertainty Quantification Toolkit (UQTk). Also, will be a tutorial that calculates model evidence and uses model evidence values to compare different models. These tutorials will also explain how a user can modify the model for their own data or a different model.

1 Acknowledgments

Thank you Ellen Stechel and Anthony McDaniel for the model and work on the problem. Thank you Bert Debusschere for mentoring.

Funding was provided by the DOE Workforce Development for Teachers and Scientists (WDTS) program as well as the DOE Office of Energy Efficiency and Renewable Energy (EERE)

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

References

Goldyreva, et al. J. of Alloys and Compounds, 638, 44-49, 2015.
McDaniel, Current Opinion in Green and Sustainable Chemistry, 4, 37-43. 2017
Panlener, R. J., et al. J. of Physics and Chemistry of Solids, 36(11), 1213-1222, 1975.
Zinkevich, et al. Solid State Ionics 177, 989-1001, 2006.