Inference of a Universal Ornstein-Zernike Closure Relationship with Machine Learning

Rhys E. A. Goodall
Cavendish Laboratory
University of Cambridge
Cambridge, CB3 0HE, UK
reag2@cam.ac.uk

Alpha A. Lee
Cavendish Laboratory
University of Cambridge
Cambridge, CB3 0HE, UK
aal44@cam.ac.uk

Abstract

The Ornstein-Zernike framework provides an elegant route for solving the inverse problem of determining a pairwise interaction potential for a liquid given its structure. However, in order to realise the potential of the formalism superior closure relationships are required. Current approximate closure relationships have been shown to have restricted universality and give rise to thermodynamic inconsistencies. In this work rather than attempting to analytically derive a new closure relationship we return to the point of the approximation and investigate whether machine learning can be used to infer a universal closure for the framework directly from simulation data. We show preliminary results that indicate this is a fruitful approach and identify areas for further work.

1 Introduction

While nature has fine tuned its ability to produce complicated and highly specific systems through self-assembly scientific efforts are rudimentary in comparison. Within liquid state systems recent work on the inverse design problem, how to design pairwise interaction potentials that give rise to particular liquid structures, has made use of Iterative Boltzmann Inversion (IBI) [1, 2]. Although powerful this approach leaves much to be desired as it relies on iterative use of computationally costly molecular dynamics simulations.

In many ways use of IBI can be seen as a step backwards given that the Ornstein-Zernike formalism [3] already provides the theoretical framework for direct inversion. However, for the formalism to really be useful new closure relationships, with more general applicability, are required. In this work we explore the application of machine learning to learn such a closure directly from data – data can be easily generated by solving the associated forward problem. The need for improved closures is inhibiting progress in many areas including, but not limited to, inverse design, coarse-graining multi-component atomistic systems [4] and calculation of chemically accurate solvation energies [5].

2 Background Theory

Ornstein-Zernike framework gives that the total correlation function, $h(r)$, for an isotropic fluid of density ρ can be decomposed into direct correlations and indirect correlations between particles:

$$h(r_{12}) = c(r_{12}) + \rho \int c(r_{13})h(r_{32})dr_3 \tag{1}$$

This equation defines the direct correlation function, $c(r_{12})$. $h(r_{12})$ is related to the more commonly used radial distribution function by $h(r_{12}) = g(r_{12}) - 1$. The second term in the Ornstein-Zernike

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

equation is the convolution of $h(r)$ and $c(r)$, therefore taking the Fourier transform of the equation yields an algebraic equation in Fourier space.

$$H(q) = C(q) + \rho H(q)C(q) \tag{2}$$

This equation can be broken apart and re-written it in terms of the static structure factor, $S(q)$. Doing so establishes a straightforward link between the formalism in Fourier space and readily available experimental information in the form of the static structure factor.

$$H(q) = \frac{1}{\rho}\Big(S(q) - 1\Big), \qquad C(q) = \frac{1}{\rho}\left(1 - \frac{1}{S(q)}\right) \tag{3}$$

To solve the inverse design of determining $\phi(r)$ from this structural information a second equation, a closure relationship, coupling $h(r)$ and $c(r)$ with $\phi(r)$ is needed. The generally accepted form of the closure function is [6]:

$$h(r) + 1 = \exp(-\beta\phi(r) + \gamma(r) + B(r)) \tag{4}$$

where $B(r)$ is the bridge function and $\gamma(r) = h(r) - c(r)$ is the indirect correlation function. Traditionally closure relationships approximate $B(r)$ using local functions of $\gamma(r)$. For long-range potentials the Hyper-netted Chain approximation (HNC), $B(r) \simeq 0$, is typically used to close the system [7]. Whilst for short-range purely-repulsive systems the Percus-Yevick approximation (PY), $B(r) \simeq \ln(1 + \gamma(r)) - \gamma(r)$, is a common choice [8].

3 Learning Generally Applicable Closure Relationships

Machine learning offers an ever improving suite of powerful tools that can be used for function approximation. Given this we are no longer tied to analytically tractable closures such as HNC or PY. However, before we can apply machine learning to this problem the most important question to answer is not whether we've picked the right model/architecture but whether the feature set we give the model contains sufficient information to determine the system. From a theoretical standpoint $B(r)$ can be expanded as an infinite series in $\gamma(r)$:

$$B(r) = \frac{\bar{F}_3}{2!}\gamma^2(r) + \frac{\bar{F}_4}{3!}\gamma^3(r) + ... \tag{5}$$

where the average modification functions, $\bar{F}_n$, are dependant on the density, ρ, and the temperature, T [9]. By unit analysis the bridge function can only be expressed in terms of dimensionless reduced quantities ρ^* and T^*. However, for complicated pairwise potentials, where multiple length and energy scales are required to define the system, comparable reduced quantities are ill-defined preventing formulation of a general closure in terms of $\gamma(r)$ only. If instead we choose to express our bridge function as $B(r) = B(h(r), c(r), ...)$, we see that, given Equation 1, we recover the degree of freedom corresponding to the density in the average modulation function description.

For a complete closure it would also be necessary to recover the degree of freedom corresponding to the reduced temperature. For this we suggest that a 'softened' distance parameter, r/ξ, where ξ contains information about the overlap behaviour (often divergent) of the potential, would be a reasonable choice. This argument is consistent with the boost in performance observed from introducing switching functions in both the Rogers-Young and Zerah-Hansen closures [10, 11]. However, for this approach to be useful such a parameter must be constructed without prior knowledge of the interaction potential. This is a non-trivial problem and has been left as an area for further work.

Despite this, to explore whether increased efficacy learnt closures are possible we adopt a closure of the form $\phi(r) = \phi(h(r), c(r))$ – implying here that $B(r) = B(h(r), c(r))$ also. We note that such a closure doesn't guarantee path independence for the chemical potential and Helmholtz free energy [12]. Ensuring such independence is important if the new closure is to out perform current closures when applied to real physical problems.

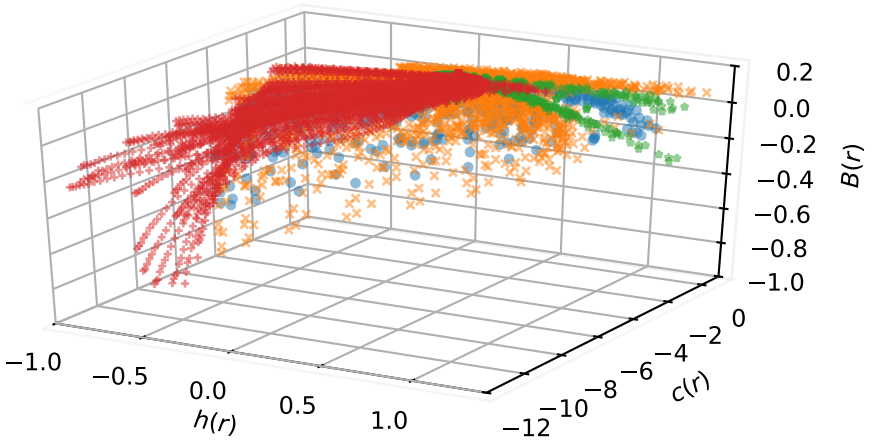

Figure 1: The lack of a one-to-one mapping for $B(r)$ as a function of $h(r)$ and $c(r)$ suggests that the local basis used is insufficient to fully describe $B(r)$. In the plot different classes of interaction potentials are differentiated as follows: blue circles - Hard-Sphere, orange crosses - Core-Softened, green stars - Soft-Sphere and red plus signs - Soft-Core.

Despite remarkable successes most machine learning approaches are essentially powerful interpolation frameworks. Therefore, in order to infer a generally applicable closure we need to explore a wide variety of the possible interaction space. We have investigated thirteen different interaction potentials. The potentials used can be grouped loosely into four classes: Hard-Sphere - potentials containing strong divergences that prevent particles from overlapping, Core-Softened - hard-sphere models where a repulsive plateau is added before the divergence to introduce complex multi-lengthscale structure, Soft-Sphere - weakly divergent systems analogous to hard-sphere systems, and Soft-Core - potentials that do not diverge and allow particles to overlap. We will refer to Hard-Sphere and Core-Softened potentials as hard potentials and Soft-Sphere and Soft-Core potentials as soft potentials. The molecular dynamics package 'ESPResSo' [13, 14] was used to determine $h(r)$ and $S(q)$ for systems of 4096 particles interacting under these chosen potentials at various temperatures and densities.

The symbolic regression software 'Eureqa' [15, 16] and deep learning library 'Keras' [17] were used to train approximate universal closures using this generated data. For our closure networks we trained simple multi-layer perceptions consisting of 5 layers each with 16 hidden units and ReLU activations[1]. We randomly reserved 20% of the available data as a test set to evaluate model performance with.

4 Results and Discussion

4.1 Symbolic Regression Recovers HNC

To start our discussion we eschew the current theoretical framework and attempt to use symbolic regression to learn an analytical closure for $\phi(r)$ as function of $h(r)$ and $c(r)$ directly from our simulation data. After training we are left with an expression functionally equivalent to Equation 4.

$$\beta\phi(r) = A \times h(r) - B \times c(r) - C \times \ln(1 + h(r)) + D \tag{6}$$

Where $A = 0.85$, $B = 0.98$, $C = 0.88$ and $D = 7.0 \times 10^{-5}$. Given the closeness of these coefficients to integer values this data driven analysis suggests that, for the available data, HNC is the best simple analytical closure. Given this we take HNC as our benchmark for comparing our learnt closures and re-cast the learning problem from that of learning the full closure for $\phi(r)$ to just learning $B(r)$ i.e. the correction term to HNC.

[1]The data set and code for training the closure networks is available at https://github.com/CompRhys/ornstein-zernike.

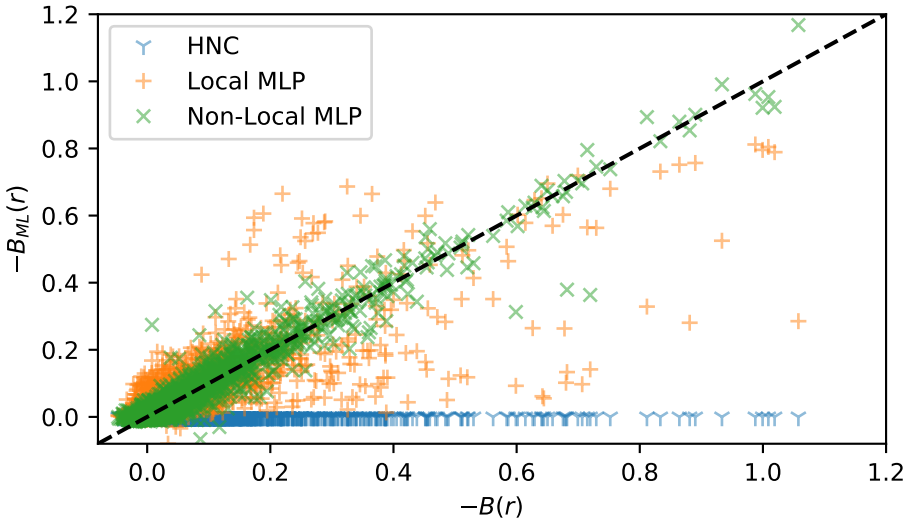

Figure 2: The predicted bridge function against the simulation ground truth for different closure approaches on the test set. The desired one-to-one equivalence line is shown in black. The figure clearly shows that both the local and non-local closure networks out perform HNC.

Table 1: Performance of HNC and the inferred closure networks on the test set.

Closure	RMSE	R^2
HNC	0.046	0.00
Local	0.025	0.63
Non-Local	0.010	0.95

4.2 Validity of the Locality Constraint for Multi-Layer Perceptron Closures

Plotting $B(r)$ as a function of $h(r)$ and $c(r)$ in Figure 1 shows clearly that a many-to-one mapping exists from $h(r)$ and $c(r)$ to $B(r)$. If we limit ourselves to considering closures in terms of $h(r)$ and $c(r)$ this observation motivates the hypothesis that non-local closure relationships are necessary. This can be done by extending the training features to include the first order derivatives with respect to the radial displacement, $h'(r)$ and $c'(r)$. By comparison consideration of non-locality within a theoretical framework is often analytically intractable.

Table 1 shows that both the local and non-local closure networks are significantly better than HNC for predicting the bridge function on the test set. In Figure 2 the local closure clearly exhibits two branches. These branches correspond to the network saturating under the locality constraint at different points for hard and soft potentials. By contrast the non-local network is able to unify the soft and hard potentials.

5 Conclusions and Future Work

The preliminary results of this work are promising and show that machine learning is an effective tool for tackling inverse problems of this type. We have not yet tested our closure downstream but believe that the path independence of our learnt closures, for instance when calculating the system pressure, as well as their performance when coarse-graining multi-component systems will be good testing grounds to explore in future work.

In addition we hope to improve upon the following areas; Firstly, further theoretical and explorative work is needed to determine appropriate radial functions or collective variables that could be used to recover the remaining missing degrees of freedom from the formalism. Secondly, we would like to experiment with physically motivated loss functions e.g. loss functions based on the update rule

of Iterative Boltzmann Inversion. Finally, we would like to explore models that map between entire distribution functions in the input and the desired potential all at once, rather than in a point-wise manner. Such models are desired as they allow non-local effects to be learnt implicitly. However, before such models could be applied rigorously it would be necessary to first define and resample the data in units of some natural lengthscale – potentially the cube root of the Kirkwood–Buff integral [18] or the radius of the first co-ordination sphere. This procedure should ensure the input features are consistently defined and therefore allow the model to deployed generally regardless of the data generation methodology.

Acknowledgements

The authors would like to acknowledge the support of the Winton Programme for the Physics of Sustainability.

References

[1] R. B. Jadrich, J. A. Bollinger, B. A. Lindquist, and T. M. Truskett. Equilibrium cluster fluids: pair interactions via inverse design. *Soft Matter*, 11:9342–9354, 2015.

[2] Beth A. Lindquist, Ryan B. Jadrich, and Thomas M. Truskett. Assembly of nothing: equilibrium fluids with designed structured porosity. *Soft Matter*, 12:2663–2667, 2016.

[3] L.S. Ornstein and F. Zernike. Accidental deviations of density and opalescence at the critical point of a single substance. *Proc. Akad. Sci.*, 17:793, 1914.

[4] Qifei Wang, David J Keffer, Donald M Nicholson, and J Brock Thomas. Use of the ornstein-zernike percus-yevick equation to extract interaction potentials from pair correlation functions. *Physical Review E*, 81(6):061204, 2010.

[5] J Richardi, PH Fries, and H Krienke. The solvation of ions in acetonitrile and acetone: A molecular ornstein–zernike study. *The Journal of chemical physics*, 108(10):4079–4089, 1998.

[6] J.P. Hansen and I.R. McDonald. *Theory of Simple Liquids*. Elsevier Science, 2006.

[7] Tohru Morita. Theory of classical fluids: Hyper-netted chain approximation, formulation for a one-component system. *Progress of Theoretical Physics*, 20(6):920–938, 1958.

[8] Jerome K. Percus and George J. Yevick. Analysis of classical statistical mechanics by means of collective coordinates. *Phys. Rev.*, 110:1–13, 1958.

[9] L.L. Lee. *Molecular Thermodynamics of Electrolyte Solutions*. World Scientific Publishing Company, 2008.

[10] Forrest J. Rogers and David A. Young. New, thermodynamically consistent, integral equation for simple fluids. *Phys. Rev. A*, 30:999–1007, 1984.

[11] Gilles Zerah and Jean-Pierre Hansen. Self consistent integral equations for fluid pair distribution functions: Another attempt. *The Journal of Chemical Physics*, 84(4):2336–2343, 1986.

[12] Stefan M. Kast. Free energies from integral equation theories: Enforcing path independence. *Phys. Rev. E*, 67:041203, Apr 2003.

[13] H. J. Limbach, A. Arnold, B. A. Mann, and C. Holm. ESPResSo – an extensible simulation package for research on soft matter systems. *Comp. Phys. Comm.*, 174(9):704–727, 2006.

[14] A. Arnold, O. Lenz, S. Kesselheim, R. Weeber, F. Fahrenberger, D. Roehm, P. Košovan, and C. Holm. ESPResSo 3.1 — Molecular Dynamics Software for Coarse-Grained Models. In M. Griebel and M. A. Schweitzer, editors, *Meshfree Methods for Partial Differential Equations VI*, volume 89 of *Lecture Notes in Computational Science and Engineering*, pages 1–23. Springer, 2013.

[15] Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.

[16] Michael Schmidt and Hod Lipson. Eureqa. www.nutonian.com, 2014.

[17] François Chollet et al. Keras. https://keras.io, 2015.

[18] Kenneth E Newman. Kirkwood–buff solution theory: derivation and applications. *Chemical Society Reviews*, 23(1):31–40, 1994.