Transferability theorems for materials machine learning

Marius Jonsson

December 3, 2018

1 Introduction

Andrew Ng called machine learning 'the new electricity' (Vara 2018), indicating that *artificial intelligence* is poised to transform many industries (Ng 2017). Thus, there has been a surge of activity in applying machine learning methods to quantum mechanical models. Some of the applications are to materials modelling and are fuelled by the evident success of these techniques in what is called artificial intelligence (Carleo and Troyer 2017; Butler et al. 2018; Mills et al. 2017). These successes followed from the collective experience that the scientific community has gained in fitting high volumes of data with very complex functional forms that involve a large number of free parameters, while still keeping control of regularity and thus avoid overfitting. Indeed it is generally held that fitting potential energy surfaces is the only practical way to simulate systems large enough for thousands of atoms and millions of time steps (Finnis 2004). However, machine learning is in its infancy, many methods are fragile, in the literature called the problem of transferability (Papernot et al. 2016). But as machine learning is in its infancy, it is a golden age for developing techniques that will improve the robustness of the methods; one of the most fortunate situations for scientists. The view of Andrew Ng is challenged by the most influential researchers of the field, such as 2017-NIPS Test-of-time-award winner Ali Rahimi (Hutson 2018). He asked researchers to shift their priorities in a bid to improve the understanding of the mechanics of machine learning. One common view is that the basis for many methods is experimental, often philosophical or proto-scientific and, according to Rahimi, reminiscent of Alchemic theory. The present PhD project targets such problems applied to materials modelling. Namely to the GAP framework, one of most promising materials modelling method for simulations of large systems. It is developed at the Department of Engineering, University of Cambridge, and the theorems I develop will strengthen the robustness of the method by asking «what can be done within a reasonable time frame to significantly improve the properties of the method?» and «what can we prove about the method's behaviour?»

The project description is structured by «Introduction», «Background», «Objectives and research questions», «The procedure», «The study» and «Where does this lead?».

2 Background

First principles molecular simulation, based on various approximations of electronic structure theory, is the workhorse of materials modelling. For example, *density functional theory* (DFT) (Kohn and Sham 1965; Hohenberg and Kohn 1964; Clark et al. 2005) is a prominent method, indicated by 19 000 papers according to the Web of Science database. However for larger systems, methods like GAP take over. The CASTEP creator Mike C. Payne states on his staff web page that he sees GAP as a critical ingredient in a predictive, black box multiscale modelling scheme. It works by using *Kernel fitting*, a multidimensional interpolation method that originated from statistics (Stein 1999; Rasmussen and Williams 2006), and widely applied in numerical analysis and machine learning (Vapnik 1998). The key to its success is the choice of *kernel* (Schlkopf and Smola 2002), and through it the basis functions employed (Bartók, Kermode, et al. 2018). A common theme in machine learning is training the computer to predict the value of some continuous function

$U : \mathscr{B}_1 \to \mathbb{R}$, which here is defined on a set of atomic neighbourhoods $\mathscr{B}_1$. If n is a natural number, an *atomic neighbourhood* is a function $B(\mathbf{r}_1, \cdots, \mathbf{r}_n) : D^3 \to \mathbb{R}$ from a closed disk D^3 in $\mathbb{R}^3$. Assume that $\mathbf{r}_i \in D^3$ and $\rho_i(\mathbf{r}) = \exp(-\|\mathbf{r} - \mathbf{r}_i\|^2)$ for each $1 \leq i \leq n$. Then an atomic neighbourhood is given by $B(\mathbf{r}_1, \cdots, \mathbf{r}_n)(\mathbf{r}) = \sum_{i=1}^{n} \rho_i(\mathbf{r})$ in the GAP framework. If m is another natural number, an m-tuple of such functions is called a *representative set* and the set of all such m-tuples is denoted by $\mathscr{B}_m$. Prediction of the function U can be done by providing the computer with a number of tuples from $\mathscr{B}_m \times U(\mathscr{B}_m)$ together with a loss-function. This enables the computer to minimise its predictive loss in guessing the value of U. For our case, that amounts to solving the Schrödinger equation for the representative set. If $B(\mathbf{r}_j, \cdots, \mathbf{r}_{nj})_{j=1}^{m} \in \mathscr{B}_m$, the model that is used is a $\mathscr{B}_1 \to \mathbb{R}$-function given by

$$f[B(\mathbf{s}_1, \cdots, \mathbf{s}_n)] = \sum_{j=1}^{m} x_j k[B(\mathbf{s}_1, \cdots, \mathbf{s}_n), B(\mathbf{r}_j, \cdots, \mathbf{r}_{nj})]. \tag{1}$$

The function $k : \mathscr{B}_1 \times \mathscr{B}_1 \to \mathbb{R}$ is the kernel, and is an inner product (Schlkopf and Smola 2002). In ordinary linear regression, if K is the matrix consisting of elements

$$K_{ij} = k(B(\mathbf{r}_j, \cdots, \mathbf{r}_{nj}), B(\mathbf{r}_i, \cdots, \mathbf{r}_{ni}))$$

this entails projecting the vector $\mathbf{y}$ consisting of elements $y_i = U(B(\mathbf{r}_i, \cdots, \mathbf{r}_{ni}))$ on the column space of K. This is desirable, as the normal equations coincide with the likelihood equations (Devore and Berk 2012). However, it is well known that penalising the normal equations by ridge regression can be used to regularise the fit and trade variance for bias (Agresti 2015). The GAP-framework does this, and it amounts to defining $\mathbf{x} = (x_1, \cdots, x_m)$ and $\Lambda = \lambda I_m$ for some constant λ[1] and estimating by

$$\mathbf{x} = \left(K - \Lambda\right)^{-1}\mathbf{y} \qquad \text{(Agresti 2015)}. \tag{2}$$

A recent review given by Bartók, Kermode, et al. (2018) on GAP shows its strengths, but more importantly, the authors reveal that it is critical for performance to build a theory for databases of learning material. A database of solutions is provided by DFT to train the computer, and at present, the number of configuration in the final database is a result of somewhat ad-hoc choices, driven partly by the varying computational cost of the electronic structure calculation, and partly by observed success in predicting properties, signalling sufficient amount of data. Csányi (2017) explained that 'how to build databases' is a question that will loom large in the coming years. Given that we understand kernels, basis functions and parameters, the next that is required is to get data. How do we get the data? What protocols are used to generate the data, such that we get a machine learning ansatz that is suitable for doing science? Students in the group have tackled one material at a time. Another technique used, is to apply the metric induced by the inner product (Lindstrøm 2017) k to maximise the dissimilarity of atomic neighbourhoods. As a significant effort goes into building databases, I set «development of theorems that tackles this question in a systematic way» as an objective for the doctoral project. With this background, I can introduce the research questions.

3 Objectives and research questions

Is there a best representative set in $\mathscr{B}_m$? The answer to this question would reveal if we will continue tweaking the training database indefinitely, without ever finding an optimum set. The following theorem shows that the question of how to build the optimum database of training data is, not only well defined, but also has a solution. Before I explain more, notice that the function $q : D^{2nm} \to \mathscr{B}_m$ given by

$$q(\mathbf{r}_1, \cdots, \mathbf{r}_{nm}) = \Big(B(\mathbf{r}_1, \cdots, \mathbf{r}_n), \cdots, B(\mathbf{r}_{n(m-1)}, \cdots, \mathbf{r}_{nm})\Big)$$

is surjective by construction, so it induces the quotient topology (Arkhangel'skiĭ 1991) on $\mathscr{B}_m$. This construction is decisive as atomic neighbourhoods are invariant under permutation in physics (Bartók et

[1] Increasing the value of λ decreases variance and increase bias by shrinking the estimates toward zero.

al. 2013). If only equivalence classes (Grishin 1989) of atomic neighbourhoods are considered, $\mathscr{B}_m$ is homeomorphic (Černavskiĭ 1989) to a closed subspace of $\mathbb{R}^d$. Define the *error* $\varepsilon : \mathscr{B}_m \times \mathscr{B}_1 \to \mathbb{R}$ given by

$$\varepsilon[B(\mathbf{r}_i, \cdots, \mathbf{r}_{in})_{i=1}^{m+1}] = \left| f[B(\mathbf{r}_i, \cdots, \mathbf{r}_{in})_{i=1}^{m+1}] - U[B(\mathbf{r}_{nm+1}, \cdots, \mathbf{r}_{n(m+1)})] \right|.$$

Using these definitions, this theorem follows:

Theorem. *For each representative set, the error is bounded on $\mathscr{B}_1$ and there exists a representative set that minimises its least upper bound over $\mathscr{B}_1$.*

Proof. I start by showing that ε is a continuous function. Pick some representative set $B(\mathbf{r}_i, \cdots, \mathbf{r}_{in})_{i=1}^{m}$ and atomic neighbourhood $B(\mathbf{r}_i, \cdots, \mathbf{r}_n)$. Consider $g : \mathscr{B}_m \times \mathscr{B}_1 \to \mathbb{R}$ given by

$$g[B(\mathbf{r}_i, \cdots, \mathbf{r}_{in})_{i=1}^{m+1}] = U[B(\mathbf{r}_{nm+1}, \cdots, \mathbf{r}_{n(m+1)})].$$

The function U is continuous with respect to $\mathscr{B}_1$ by construction, so if V is a basis element of the standard topology on $\mathbb{R}$, then g is continuous because

$$g^{-1}(V) = \{(B_m, B_1) \in \mathscr{B}_m \times \mathscr{B}_1 \mid g(B_m, B_1) \in V\} = \mathscr{B}_m \times U^{-1}(V).$$

In the case that K is an $m \times m$ matrix[2], the regression coefficients are continuous as a $\mathscr{B}_m \to \mathbb{R}^m$-function. To see this, decompose the coefficients in the following way:

name : domain → co domain	rule	
$r : \mathrm{GL}_m(\mathbb{R}) \to \mathrm{GL}_m(\mathbb{R})$	$r(A) = A^{-1}$	
$s : \mathscr{B}_m \to \mathrm{GL}_m(\mathbb{R})$	$s(B(\mathbf{r}_i, \cdots, \mathbf{r}_{in})_{i=1}^{m}) = K(B(\mathbf{r}_i, \cdots, \mathbf{r}_{in})_{i=1}^{m}) + \Lambda;$	K kernel matrix
$t : \mathscr{B}_m \to \mathbb{R}^m$	$t(B(\mathbf{r}_i, \cdots, \mathbf{r}_{in})_{i=1}^{m}) = \big(U(B(\mathbf{r}_1, \cdots, \mathbf{r}_n), \cdots, B(\mathbf{r}_{(m-1)n}, \cdots, \mathbf{r}_{mn}))\big)^{\mathsf{T}}$	
$u : \mathrm{GL}_m(\mathbb{R}) \times \mathbb{R}^m \to \mathbb{R}^m$	$u(A, \mathbf{a}) = A\mathbf{v}$	

$$\begin{tikzcd}
& \mathrm{GL}_m(\mathbb{R}) \arrow[r, "r"] & \mathrm{GL}_m(\mathbb{R}) & \\
\mathscr{B}_m \arrow[ur, "s"] \arrow[rr, dashed, "v=(r\circ s,t)"] \arrow[dr, "t"] \arrow[rrr, bend right=60, "u\circ v"'] & & \mathrm{GL}_m(\mathbb{R}) \times \mathbb{R}^m \arrow[u, "\pi_1"'] \arrow[r, "u"] \arrow[dl, "\pi_2"] & \mathbb{R}^m \\
& \mathbb{R}^m & &
\end{tikzcd}$$

All of these functions are continuous according to lemma 2 and corollary 1 (appendix) and by using that compositions of continuous functions are continuous (Rynne and Youngson 2007). On the other hand, the function v is continuous as each component function is (Munkres 2000). To see that the kernel is continuous, note that as k is an inner product, it is jointly continuous according to Murphy (1990). Furthermore, compositions of finite sums of continuous real functions are continuous, which proves that ε is in fact continuous.

According to the lemma 1 (appendix), $\sup_{A \in \mathscr{B}_1} \varepsilon[B(\mathbf{r}_i, \cdots, \mathbf{r}_{in})_{i=1}^{m}, A]$ is continuous, and the space $\mathscr{B}_m$ is compact as it has the quotient topology from a compact space (Munkres 2000). Hence $\sup \varepsilon$ attains its minima $B \in \mathscr{B}_m$ by the extreme value theorem (Lee 2011). By the same theorem, it is also bounded. □

[2] It is possible to fit the model using rectangular matrix K, in which case the formulas are larger, but the proof carries through in a similar way.

Table 1: Research questions that contribute to the theory of the Cambridge GAP framework and to machine learning materials modelling theory.

Research questions
1. «Can we sharpen the tools used to tackle questions about the atomic neighbourhood?».
2. «Can we sharpen the tools used to tackle potential energy surfaces?»
3. «What are the optimum material configurations to include in learning materials?»
4. «How do we address fragility of high dimensional fits?»
5. «What is the best prediction that can be made given training data?»

The theorem shows two things, first that the maximal error on the set of all configurations is bounded for all the atomic neighbourhoods (i.e. uniformly) for each representative set under the hypothesis above. Second, that the following question is meaningful: «What are the optimum material configurations to include in the training database?». The supremum in the proof above also shows that the method is potentially reasonably behaved as long as the assumptions are not violated. The behaviour can be improved further as the mathematical properties of each constituent can be refined iteratively. This is decisive, as the research group at Cambridge has explained that optimum building of training databases is one of the most important questions to tackle at the moment. In addition to that, a number of additional questions are pertinent (see table 1). By answering these questions, there is both a theoretical foundation for learning and performance, as well as novel contributions to the basic science of kernel regression methods.

4 The procedure

What is the best course of action? Taking advantage of the main theorems from mathematics, and contribute to the theory of the GAP-framework by mapping out the mathematical landscape further is part of the answer. As explained, the landscape is appealing, but can be improved further by incorporating the main symmetries from physics. If the resulting linear space is isomorphic (Ivanova and Smirnov 1990) to a space heavily researched by mathematicians, results may be plentiful. If not, this project will provide novel results of its own. Besides this, each research question has a specific procedure.

When working on the first question, regarding if it is possible to sharpen the tools used to tackle questions about the GAP framework, I will incorporate standard physics symmetries into the set $\mathscr{B}_1$ and work out which space $\mathscr{B}_1$ is isomorphic to. This is helpful as reducing the size of the space increases structure (and hence the number of available tools). Next, I will investigate the topology of $\mathscr{B}_1$ and find out which tools from topology that apply to continuous functions on $\mathscr{B}_1$ (such as the potential energy surface, for example). Then I will look at spaces homeomorphic to $\mathscr{B}_1$, and note which intuition this gives about equivalence classes of $\mathscr{B}_1$. As $(\mathscr{B}_1, k)$ is a separable (Hazewinkel 1992) Hilbert space [$q(\mathscr{Q}^3 \cap D^3)$ is a dense subset], it contains an orthonormal sequence. I will reduce the sequence to obtain a finite subspace that is useful for doing calculations. This final point is useful for linear operators on $\mathscr{B}_1$. It is not good enough for the potential energy surface directly, but for the purpose of increasing transferability, it suffices to be close. On the finite subspace, I will apply the Hahn-Banach theorem (Sobolev 1989) and write down all the results that are useful for operators on the space $\mathscr{B}_1$.

While working on the second question, regarding whether we can sharpen the tools used to tackle questions about the potential energy surface before starting calculations, I will start by using the fact that the potential energy surface can be probed at finite sets of points, and use this to make a parametric Bayesian model of every potential energy surface. I will investigate if it is possible to harvest information from each time a user runs the program to sharpen the knowledge of the types of potential energy surfaces that the user is interested in. Thereafter I will take the insight I have gained from working with potential energy surfaces to parameterise sequences of functions that converge weakly to a given potential energy surface. To get another way to manifest our knowledge of potential energy surfaces, I will evaluate the algebras of functions on $\mathscr{B}_1$ that come naturally and see if formulas can be obtained for the main features of interest of the potential energy surface. Thereafter, I will constrain the space of continuous functions that contains the potential energy surface to see what literature exists for such functions, and write down all consequences that are automatic. I will also identify stationary points of the potential energy surface.

When I get to the third question, regarding the optimum material configurations to include in the training database, I will deduce analytical constraints on the elements of $\mathscr{B}_m$. The difficulty of this problem can be tuned depending on how optimisation is done. The strongest results will arise from minimising the least upper bound of ε. I will start with this, and by working to get analytical constraints by looking at sequences in $\mathscr{B}_m$. It is easier to work with a limit rather than a supremum (Lindstrøm 2017). After I have investigated this, it is natural to introduce the Gâteaux derivative (Tikhomirov 1989) on $\mathscr{B}_m$ and work with the results I developed for questions one and two. I will work with the Gâteaux derivative until analytical constraints on $\mathscr{B}_m$ are obtained. I can gradually strengthen assumptions until suitable equations are obtained. If they are non-linear, or uneconomical to compute, further optimisation is required before they can be embedded in the GAP framework. This part of the work will implicitly answer question number four, because Devore and Berk (2012) shows that kernel fitting is a special case of multiple linear regression. Therefore given that the ridge regression is performed correctly, a poor fit shows that the projection of $\mathbf{y}$ on column space K is the problem (Agresti 2015). So if there is enough data, the only other conclusion is that the kernel poorly explains $\mathbf{y}$. This point motivates the final question.

On the final question, regarding the best prediction that can be made given this training data, I will start by investigating the statistics literature of known ways to evaluate goodness-of-fit, analogous to 'deviance' (de Jong and Heller 2008) and R^2 (DeGroot and Schervish 2014). Then I will compare the measures to the mathematical properties that I have found for the potential energy surface and the GAP-framework to decide which best determine the qualities that are decisive for materials modelling. After this, I will use statistics to investigate the relation between the values of the measure and size of the representative set (plus any free parameters that appear as a result of question number three), quantify the errors that are available with the new knowledge of GAP to indicate how much more accuracy can be obtained by refining database building, and finally use heuristic methods to study the ways that the SOAP kernel fails to convey features of the Schrödinger equation to see if this is a fruitful avenue for future research.

5 The study

I will start preparing myself in the spring 2019 while still in Norway by learning DFT at Oslo University. My goal is to start writing papers as soon as possible, using the ideas I have developed in the meantime.

The Cambridge University website states that it is common to follow lectures in the first year. The website also states that in the first year, students will attend approximately 24 hours of seminars targeted at developing their research and communications skills. After about 10-12 months I should be ready with the manuscript of the second paper, see table 2 [If Cambridge policy allows, I propose to use my existing paper on error estimation as the first paper (Jonsson 2018)]. The title might be something like «Theorems for the chemical environment». Here, I can present the first theorems for the GAP framework and expect the mathematics will be attractive. I also propose the idea of a third paper with applications to the GAP framework and the chief analytical potentials, for example intended for *Physical Review B*. «Findings for chemical potentials» would be a suitable title for such a paper.

In the second year I will begin to prepare for the most difficult research question of the thesis: «Can

Table 2: Research papers that may contribute to the theory of Cambridge GAP framework and to machine learning for material modelling. The first paper has already been published.

Title	Intended for	Due
1. «Standard error estimation by an automated blocking method»	Phys. Rev. E	Already out
2. «Theorems for the chemical environment»	Phys. Rev. Lett.	Jul 2020
3. «Findings for chemical potentials»	Phys. Rev. B	Oct 2020
4. «Observations about the potential energy surface»	J. Chem. Phys.	Jan 2021
5. «Theorems for potential energy surfaces»	Phys. Rev. B	August 2021
6. «Constructionist learning for the machine»	Nature mater.	May 2022
7. «Thesis: Constructionist learning for the machine»	Cambridge Univ.	2022-2023

we sharpen the tools used to tackle potential energy surfaces». I plan to compile the generalisations that are available by performing a Bayesian study of potential energy surfaces. This paper is intended for either *Journal of Chemical Physics* or *Physical Review Letters*. The title «Observations about the potential energy surface» should suffice and be ready in early 2021. Thereafter, I will shift my focus to study the properties of potential energy surfaces using the general intuition that I have built. Another paper containing propositions and lemmas (and theorems, if there are any) could be compiled into a manuscript intended for *Physical review B* or *Physical Review Letters*.

In the third year, I plan to work on the central parts of the project. I will work in parallel on basic machine learning theory and what has been gathered about potential energy surfaces and propose to submit the findings to *Nature materials* with the title «Constructionist learning for the machine», and submitted before graduation. After that I will work on the last question, or write the thesis.

6 Where does this lead?

This project addresses two key issues regarding the development of machine learning for materials modelling, first by mapping out the mathematical landscape of the GAP framework, and second, how the learning material for the machine should be constructed. The mechanisms governing database predictive power are unknown, but this research will reveal how the representative set depends on changes to the potential energy surface, and the mathematics that determines the parts that can be controlled a priori.

The novel contributions include investigations of the mathematical properties of atomic neighbourhoods and potential energy surfaces. By incorporating the standard physics symmetries, I will find which mathematical space $\mathscr{B}_1$ represents, and identify its topology. Thereby, what can be said about continuous functions on $\mathscr{B}_1$. As computation is the best tool for answering questions about the potential energy surface, I will map out finite subspaces of $\mathscr{B}_1$. The chief theorems of linear analysis will be applied to deduce what new can be said about the properties of linear operators on $\mathscr{B}_1$. Only a few studies of this type provide protocols for the learning processes. In addition to being novel, work of this type is sought after, see for example (Bartók et al. 2013). It is easy to understand why, as the purpose of rigourous proof is to provide sound foundations for research. The need for such research is also backed up by the most influential reseachers in basic machine learning science, such as Ali Rahimi (Hutson 2018).

The project will lead to work that is helpful for The Department of Engineering, as the theoretical foundation of the GAP framework, which it essentially owns, will grow significantly. A larger theoretical foundation will give the department more tools to explore the behaviour of the framework, allow group members to troubleshoot extensions of the framework, and provide insights into fruitful avenues for prospective research. The project is ambitious, but should be doable in 3 years in the University of Cambridge environment.

7 Appendix

Lemma 1. *Suppose that X, Y are topological spaces with Y compact and $f : X \times Y \to \mathbb{R}$ is continuous. Then the $X \to \mathbb{R}$-function given by $h(x) = \sup_{y \in Y} f(x, y)$ is continuous.*

Proof. I will prove that the inverse image of each sub basis element of $\mathbb{R}$ under h is open in X. Assume $a \in \mathbb{R}$ and pick $x \in h^{-1}(a, \infty)$. That means $a < \sup_{y \in Y} f(x, y)$. As y is compact, there is $y \in Y$ such that $\sup_{y \in Y} f(x, y) = f(x, y)$ by the extreme value theorem. Hence $(x, y) \in f^{-1}(a, \infty)$, which is to say that $x \in \pi_1(f^{-1}(a, \infty))$. Therefore $h^{-1}(a, \infty) \subseteq \pi_1(f^{-1}(a, \infty))$. For the reverse inequality, assume $x \in \pi_1(f^{-1}(a, \infty))$. Hence, there is some $y \in Y$ such that

$$a < f(x, y) \leq \sup_{y \in Y} f(x, y) = h(x).$$

That means $h^{-1}(a, \infty) \supseteq \pi_1(f^{-1}(a, \infty))$, and consequently $h^{-1}(a, \infty)$ is the image of an open set under an open map. To finish the proof, fix $b \in \mathbb{R}$ and $x \in h^{-1}(-\infty, b)$. Now define b' to be the midpoint of the interval $[h(x), b]$. As $b' > h(x) = \sup_{y \in Y} f(x, y) \geq f(x, y)$ for each $y \in Y$, I get $\{x\} \times Y \subseteq f^{-1}(-\infty, b')$. Therefore, as Y is compact, there is some open set $W \ni x$ such that $W \times Y \subseteq f^{-1}(-\infty, b')$ by the tube lemma (Munkres 2000). Moreover

$$f(W \times Y) \subseteq f(f^{-1}(-\infty, b')) \subseteq (-\infty, b') \qquad \text{only if} \qquad \text{for each } (w, y) \in W \times Y, b' > f(w, y).$$

Therefore $h(w) = \sup_{y \in Y} f(w, y) \leq b' < b$, so $w \in W \subseteq h^{-1}(-\infty, b)$. □

Lemma 2. *Suppose X is a topological space and $g_{ij} : X \to \mathbb{R}$ is a continuous function for each $1 \leq i \leq k$ and $1 \leq j \leq m$. Then the function $K : X \to \mathbb{R}^{k \times m}$ given by*

$$K(x) = \begin{bmatrix} g_{11}(x) & g_{12}(x) & \cdots & g_{1m}(x) \\ g_{21}(x) & g_{22}(x) & & \vdots \\ \vdots & & \ddots & \\ g_{k1}(x) & \cdots & & g_{km}(x) \end{bmatrix}$$

is continuous.

Proof. Pick an $x \in X$. As the operator norm (Rynne and Youngson 2007) induce the metric topology on $\mathbb{R}^{k \times m}$, there is an open set $V \in \mathbb{R}^{k \times m}$ containing $K(x)$. Pick $\varepsilon > 0$ such that $B(K(x); \varepsilon) \subseteq V$. As g_{ij} is continuous for each $1 \leq i \leq k$ and $1 \leq j \leq m$, I can pick open set $U_{ij} \ni x$ such that the open ball $B(g_{ij}(x), \varepsilon(mk)^{-1/2}) \supseteq g_{ij}(U_{ij})$ whenever $1 \leq i \leq k$ and $1 \leq j \leq m$. The set $U = \bigcap_{i=1}^{k} \bigcap_{j=1}^{m} U_{ij}$ is a finite intersection of open sets, hence it is open. Therefore, if $y \in U$, then $y \in U_{ij}$ for all $1 \leq i \leq k$ and $1 \leq j \leq m$, and so

$$g_{ij}(y) \in g_{ij}(U_{ij}) \subseteq B\Big(g_{ij}(x); \frac{\varepsilon}{(mk)^{1/2}}\Big) \qquad \text{only if} \qquad |g_{ij}(x) - g_{ij}(y)| < \frac{\varepsilon}{(mk)^{1/2}}.$$

Now, note that if $(e_1, e_2, \cdots, e_m) = \mathbf{e} \in \mathbb{R}^m$ has unit length, then according to the generalised mean (Borowski and Borwein 1989) inequality,

$$\left(\sum_{i=1}^{m} e_i\right)^2 \leq m \sum_{i=1}^{m} x_i^2 \leq m \quad \text{only if} \quad \sum_{j=1}^{k} \Big(\sum_{i=1}^{m} \underbrace{(g_{ij}(x) - g_{ij}(y))}_{< \varepsilon k^{-1/2}/m} e_i \Big)^2 = \frac{\varepsilon^2}{mk} \underbrace{\sum_{j=1}^{k} \left(\sum_{i=1}^{m} e_i\right)^2}_{\leq \sum_{j=1}^{k} m} \leq \varepsilon^2$$

and therefore $\|K(x) - K(y)\| = \sup_{\|\mathbf{e}\|<1} \|(K(x) - K(y))\mathbf{e}\| < \sup_{\|\mathbf{e}\|<1} |\varepsilon| = \varepsilon$. Which is to say $K(U) \subseteq B(K(x); \varepsilon) \subseteq V$. □

Corollary 1. *The following functions are continuous:*

$$\begin{array}{rl}
\textit{name} : \textit{domain} \to \textit{co domain} & \qquad \textit{Rule} \\
+ : \mathbb{R}^{k\times m} \times \mathbb{R}^{k\times m} \to \mathbb{R}^{k\times m} & \qquad (A, B) \mapsto A + B \\
\cdot : \mathbb{R}^{k\times m} \times \mathbb{R}^{m\times n} \to \mathbb{R}^{k\times n} & \qquad (A, B) \mapsto AB \\
a : \mathbb{R} \times \mathbb{R}^{k\times m} \to \mathbb{R}^{k\times m} & \qquad (a, B) \mapsto aB \\
^{-1} : \mathrm{GL}_m(\mathbb{R}) \to \mathrm{GL}_m(\mathbb{R}) & \qquad (A) \mapsto A^{-1}
\end{array}$$

Proof. In each case, there exists a formula $g_{ij} : \mathbb{R}^{k\times m} \to \mathbb{R}$ consisting of sums, products and quotients of real numbers, which are continuous. □

Literature cited

Agresti, Alan (2015). *Foundations of Linear and Generalized Linear Models*. 1st ed. New Jersey: Wiley & Sons, Inc.

Arkhangel'skiĭ, Alexander V. (1991). "Quotient Mapping". In: *Encyclopaedia of mathematics* 7, p. 458.

Bartók, Albert P., James Kermode, Noam Bernstein, and Gábor Csányi (2018). "Machine Learning a General Purpose Interatomic Potential for Silicon". In: *ArXiv e-prints*. arXiv: 1805.01568.

Bartók, Albert P., Risi Kondor, and Gábor Csányi (2013). "On Representing Chemical Environments". In: *Physical Review B* 87.18, p. 184115. doi: 10.1103/PhysRevB.87.184115.

Borowski, Ephraim J. and Joathan M. Borwein (1989). *Hölder Mean*. London.

Butler, Keith T, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh (2018). "Machine Learning for Molecular and Materials Science". In: *Nature* 559.7715, pp. 547–555. doi: 10.1038/s41586-018-0337-2.

Carleo, Giuseppe and Matthias Troyer (2017). "Solving the Quantum Many-Body Problem with Artificial Neural Networks". In: *Science* 355.6325. doi: 10.1126/science.aag2302.

Černavskiĭ, Aleksei V. (1989). "Homeomorphism". In: *Encyclopaedia of mathematics* 4, p. 443.

Clark, Stewart J., Matthew D. Hutson, Chris J. Pickard, Phil J. Hasnip, Matt I. J. Probert, Keith Refson, and Mike C. Payne (2005). "First Principles Methods Using CASTEP". In: *Zeitschrift für Kristallographie - Crystalline Materials* 220.5/6, pp. 567–570. doi: 10.1524/zkri.220.5.567.65075.

Csányi, Gábor (2017). *Overview of Kernel-Based Machine Learning of Atomistic Properties*. Beamer. Berlin.

de Jong, Piet and Gillian Z. Heller (2008). *Generalized Linear Models for Insurance Data*. 1st ed. Cambridge: Cambridge university press.

DeGroot, Morris H. and Mark J. Schervish (2014). *Probability and Statistics*. 4th ed. Essex: Pearson Education, Inc.

Devore, Jay L. and Kenneth L. Berk (2012). *Modern Mathematical Statistics with Applications*. 2nd ed. London: Springer.

Finnis, Mike (2004). *Interatomic Forces in Condensed Matter*. 1st ed. Oxford Series on Materials Modelling. London: Oxford University Press.

Grishin, Vyacheslav N. (1989). "Equivalence". In: *Encyclopaedia of mathematics* 3, p. 402.

Hazewinkel, Michiel (1992). "Separable Space". In: *Encyclopaedia of mathematics* 8, p. 275.

Hohenberg, Pierre and Walter Kohn (1964). "Inhomogeneous Electron Gas". In: *Physical Review* 136.3B, B864–B871. doi: 10.1103/PhysRev.136.B864.

Hutson, Matthew (2018). "AI Researchers Allege That Machine Learning Is Alchemy". In: *Science* 360.6388, p. 478. doi: 10.1126/science.aau0577.

Ivanova, Olga A. and Dmitrii M. Smirnov (1990). "Isomorphism". In: *Encyclopaedia of mathematics* 5, p. 202.

Jonsson, Marius (2018). “Standard Error Estimation by an Automated Blocking Method”. In: *Physical Review E* 98.4, p. 043304. DOI: 10.1103/PhysRevE.98.043304.

Kohn, Walter and Lu J. Sham (1965). “Self-Consistent Equations Including Exchange and Correlation Effects”. In: *Physical Review* 140.4A, A1133–A1138. DOI: 10.1103/PhysRev.140.A1133.

Lee, John M. (2011). *Introduction to Topological Manifolds*. 2nd ed. New York: Springer.

Lindstrøm, Tom (2017). *Spaces: An Introduction to Real Analysis*. 1st ed. Providence, Rhode Island: American Mathematical Society.

Mills, Kyle, Michael Spanner, and Isaac Tamblyn (2017). “Deep Learning and the Schrödinger Equation”. In: *Physical Review A* 96.4, p. 042113. DOI: 10.1103/PhysRevA.96.042113.

Munkres, James (2000). *Topology*. 2nd ed. Upper Saddle River, New Jersey: Prentice Hall, Inc.

Murphy, Gerard J. (1990). *C*-Algebras and Operator Theory*. 1st ed. London: Academic Press.

Ng, Andrew (2017). *AI Is the New Electricity*. Sebastopol.

Papernot, Nicolas, Patrick McDaniel, and Ian Goodfellow (2016). “Transferability in Machine Learning: From Phenomena to Black-Box Attacks Using Adversarial Samples”. In: *ArXiv e-prints*. arXiv: 1605.07277.

Rasmussen, Carl E. and Christopher K. I. Williams (2006). *Gaussian Processes for Machine Learning (Adaptive Computation And Machine Learning)*. 1st ed. Cambridge: The MIT Press.

Rynne, Bryan and Martin A. Youngson (2007). *Linear Functional Analysis (Springer Undergraduate Mathematics Series)*. 2nd ed. London: Springer.

Schlkopf, Bernhard and Alexander J. Smola (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. 1st ed. Cambridge, Massachusetts: The MIT Press.

Sobolev, Vladimir I. (1989). “Hahn-Banach Thorem”. In: *Encyclopaedia of mathematics* 4, p. 353.

Stein, Michael L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. 1st ed. New York: Springer.

Tikhomirov, Vladimir M. (1989). “Gâteaux Derivative”. In: *Encyclopaedia of mathematics* 4, p. 193.

Vapnik, Vladimir N. (1998). *Statistical Learning Theory*. 1st ed. New York: Wiley & Sons, Inc.

Vara, Vauhini (2018). “The Authority on A.I.” In: *Fortune* 10, pp. 34–36.