Gaussian similarity metric for atomistic configurations

Franco Aquistapace
(Dated: March 22, 2024)

INTRODUCTION

For a given set of atomic positions $\{\vec{x}_i\}$, let us define the neighbourhood $C_i = \{\vec{x}_j\}$ of the i-th atom. This neighbourhood can be defined as the N closest atoms to atom i or as the set of atoms within a cutoff distance r from atom i. Moving forward we will use the first definition.

The aim now is to develop a metric $M(C_1, C_2)$ that can quantify the similarity between configurations C_1 and C_2. Let us establish some desirable properties for M:

- If C_1 and C_2 are equal, then $M(C_1, C_2) = 0$.
- For any C_1 and C_2, it is true that $M(C_1, C_2) \geq 0$.

Since we want our system to be scalable (i.e. comparing configurations with a big number of atoms), then we must leave behind the metrics that would require finding the best point-to-point mapping between C_1 and C_2. This rules out some of the standard ML metrics that require to have previously identified said mapping, such as Mean Squared Error (MSE), Mean Average Error (MAE) and Root Mean Squared Error (RMSE). On the other hand, point cloud metrics such as the Chamfer discrepancy,

$$d_{CD}(C_1, C_2) = \frac{1}{|C_1|} \sum_{\vec{x} \in C_1} \min_{\vec{y} \in C_2} ||\vec{x} - \vec{y}||_2^2 + \frac{1}{|C_2|} \sum_{\vec{y} \in C_2} \min_{\vec{x} \in C_1} ||\vec{x} - \vec{y}||_2^2 \quad , \tag{1}$$

require finding nearest neighbors between the two sets. This makes their complexity $\mathcal{O}(N^2)$. Overall, there seems to be an intrinsic difficulty in comparing two discrete sets of points that take on continuous values.

So... what if we change the problem itself? Let us begin by defining the *measure representation* $\mathcal{C}$ of a given atomic configuration C as:

$$\mathcal{C} = \frac{1}{\sqrt{|C|}} \sum_{\vec{x_0} \in C} f_{\vec{x}_0}(\vec{x}) \quad , \tag{2}$$

where $|C|$ is the amount of elements in the set C and $f_{\vec{x}_0} : \mathbb{R}^3 \to \mathbb{R}$. A reasonable choice for $f_{\vec{x}_0}$ is:

$$f_{\vec{x}_0}(\vec{x}) = \frac{1}{\sqrt{(2\pi\sigma)^3}} \exp\left(\frac{-1}{2\sigma} ||\vec{x} - \vec{x_0}||_2^2\right) \quad , \tag{3}$$

with $\sigma \in \mathbb{R}^+$. Notice that this choice is a probability density function, so it has the property $\langle f_{\vec{x}_0}, 1 \rangle = 1$. We are defining the inner product $\langle , \rangle : C(\mathbb{R}^3) \times C(\mathbb{R}^3) \to \mathbb{R}$ as:

$$\langle f, g \rangle = \int_{\mathbb{R}^3} f(\vec{x}) g(\vec{x}) d\vec{x} \quad . \tag{4}$$

With this inner product, we can quantify the similarity between two configurations C_1 and C_2 by means of their measure representations, $\mathcal{C}_1$ and $\mathcal{C}_2$.

We can now understand that the maximum difference between C_1 and C_2 is characterized by $\langle \mathcal{C}_1, \mathcal{C}_2 \rangle = 0$, which can be approached but never reached by the $f_{\vec{x}_0}$ functions we are using. On the other hand, we know that the maximum similarity between $f_{\vec{x}_0}$ and $f_{\vec{x}_1}$ occurs when $\vec{x}_0 = \vec{x}_1$. Let's find out what this value is in terms of $f_{\vec{x}_0}$. Notice that:

$$\begin{aligned} \langle f_{\vec{x}_0}, f_{\vec{x}_0} \rangle &= \left(\frac{1}{\sqrt{(2\pi\sigma)^3}}\right)^2 \int_{\mathbb{R}^3} \exp\left(\frac{-1}{\sigma} ||\vec{x} - \vec{x}_0||_2^2\right) d\vec{x} \\ &= \frac{1}{8\sqrt{(\pi\sigma)^3}} \quad . \end{aligned}$$

Therefore, $(8\sqrt{(\pi\sigma)^3})^{-1}$ is the maximum similarity that can be obtained for our choice of $f_{\vec{x}_0}$, if σ is the same for every configuration.

Now, we can try to estimate what is the maximum similarity that can be obtained for the configurations C_1 and C_2. By extension of our previous reasoning, we will assume that this maximum occurs when $\mathcal{C}_1 = \mathcal{C}_2$,

$$\begin{aligned}
\langle \mathcal{C}_1, \mathcal{C}_1 \rangle &= \frac{1}{|C_1|} \sum_{\vec{x}_1 \in C_1} \sum_{\vec{x}'_1 \in C_1} \int_{\mathbb{R}^3} f_{\vec{x}_1}(\vec{x}) f_{\vec{x}'_1}(\vec{x}) d\vec{x} \\
&= \frac{1}{|C_1|} \sum_{\vec{x}_1 \in C_1} \sum_{\vec{x}'_1 \in C_1} \langle f_{\vec{x}_1}, f_{\vec{x}'_1} \rangle \\
&= \frac{1}{|C_1|} \left(\sum_{\vec{x}_1 \in C_1} \langle f_{\vec{x}_1}, f_{\vec{x}_1} \rangle + \sum_{\vec{x}_1 \in C_1} \sum_{\vec{x}'_1 \in C_1 \setminus \{\vec{x}_1\}} \langle f_{\vec{x}_1}, f_{\vec{x}'_1} \rangle \right) \\
&= \frac{1}{|C_1|} \left(\frac{|C_1|}{8\sqrt{(\pi\sigma)^3}} + \sum_{\vec{x}_1 \in C_1} \sum_{\vec{x}'_1 \in C_1 \setminus \{\vec{x}_1\}} \langle f_{\vec{x}_1}, f_{\vec{x}'_1} \rangle \right) \\
&= \frac{1}{8\sqrt{(\pi\sigma)^3}} + \frac{1}{|C_1|} \sum_{\vec{x}_1 \in C_1} \sum_{\vec{x}'_1 \in C_1 \setminus \{\vec{x}_1\}} \langle f_{\vec{x}_1}, f_{\vec{x}'_1} \rangle
\end{aligned}$$

Notice that for σ sufficiently smaller than $||\vec{x}_1 - \vec{x}'_1||_2$, the second term of the expression above goes to zero. Under that assumption we get:

$$\langle \mathcal{C}_1, \mathcal{C}_1 \rangle \approx \frac{1}{8\sqrt{(\pi\sigma)^3}} \quad . \tag{5}$$

Let us choose $\sigma = (4\pi)^{-1}$, so that $\langle \mathcal{C}_1, \mathcal{C}_1 \rangle \approx 1$. Additionally, if atomic positions are given in Å, we can apply the transformation $\vec{x} \to \vec{x}/\text{Å}$ to make $\vec{x}$ unitless. Considering that transformation, we will usually get $||\vec{x}_1 - \vec{x}'_1||_2$ an order of magnitude higher than σ for $\vec{x}_1 \neq \vec{x}'_1$.

To ensure that our comparison between $\mathcal{C}_1$ and $\mathcal{C}_2$ fits exactly within the range $[0, 1]$, we will take $\min(1, \langle \mathcal{C}_1, \mathcal{C}_2 \rangle)$. We are now finally ready to define the similarity metric M, in a way that fulfills the properties that we established at the beginning:

$$M(C_1, C_2) = 1 - \min(1, \langle \mathcal{C}_1, \mathcal{C}_2 \rangle) \quad . \tag{6}$$

We can write M completely to appreciate everything that goes into it:

$$M(C_1, C_2) = 1 - \min \left\{ 1, \frac{8}{|C_1||C_2|} \sum_{\vec{x}_1 \in C_1} \sum_{\vec{x}_2 \in C_2} \int_{\mathbb{R}^3} \exp \left[-2\pi \left(||\vec{x} - \vec{x}_1||_2^2 + ||\vec{x} - \vec{x}_2||_2^2 \right) \right] \right\} \tag{7}$$