Temperature-agnostic analytic continuation of conductivity with neural nets

S. Verret

Institute for Learning Algorithms (MILA), Montréal, Canada and

Institut Quantique (IQ), Sherbrooke, Canada

(Dated: June 30, 2022)

CONTENTS

I. TEMPERATURE-AGNOSTIC ANALYTIC CONTINUATION

A. Analytic continuation of conductivity

The starting point of this report is the following relation (partly derived in appendix) between the Matsubara response function $\Pi(i\omega_n)$ and the optical conductivity $\sigma(\omega)$,

$$\Pi(i\omega_n) = \int_{-\infty}^{\infty} \frac{d\omega}{\pi} \frac{\omega^2}{\omega^2 + \omega_n^2} \mathrm{Re}\{\sigma(\omega)\}. \tag{1}$$

Here, ω is a frequency or energy (we set $\hbar = 1$), $\omega_n = 2\pi k_B T n$ are Matsubara frequency, T is the temperature, and k_B is the Boltzmann constant. Four our purpose, the term "analytic continuation" refers to the problem of inverting equation (1). That is, recover a numerical estimate of $\mathrm{Re}\{\sigma(\omega)\}$ given a numerical estimate for $\Pi(i\omega_n)$.

Two special cases of the above equation will be of particular importance. First, the value $\Pi(i\omega_n = 0)$ which corresponds to the normalization of $\mathrm{Re}\{\sigma(\omega)\}$,

$$\Pi(0) = \int_{-\infty}^{\infty} \frac{d\omega}{\pi} \mathrm{Re}\{\sigma(\omega)\}. \tag{2}$$

Second, values of $\Pi(i\omega_n)$ at high ω_n and values of $\mathrm{Re}\{\sigma(\omega)\}$ at high ω (the tails) are related by

$$\lim_{\omega_n \to \infty} \omega_n^2 \Pi(i\omega_n) = \int_{-\infty}^{\infty} \frac{d\omega}{\pi} \omega^2 \mathrm{Re}\{\sigma(\omega)\}. \tag{3}$$

B. Dimensionless formulation

We can obtain a dimensionless version of (1) by dividing it by (2). To clarify the dimensional analysis,we substitute $\omega_n = 2\pi k_B T n$, and temporarily restore $\hbar$,

$$\frac{\Pi(i\omega_n)}{\Pi(0)} = \int_{-\infty}^{\infty} d\omega \frac{\omega^2}{\omega^2 + \left(\frac{2\pi k_B}{\hbar} T\right)^2 n^2} \left(\frac{\mathrm{Re}\{\sigma(\omega)\}}{\int_{-\infty}^{\infty} d\omega \mathrm{Re}\{\sigma(\omega)\}} \right). \tag{4}$$

We will denote the left-hand side as a vector $\mathbf{\Pi}$,

$$\Pi_n = \frac{\Pi(i\omega_n)}{\Pi(0)} \tag{5}$$

and the term in parenthesis as a density function $p(\omega)$,

$$p(\omega) = \frac{\mathrm{Re}\{\sigma(\omega)\}}{\int_{-\infty}^{\infty} d\omega \mathrm{Re}\{\sigma(\omega)\}}. \tag{6}$$

These are natural dimensionless input $\mathbf{\Pi}$ and output $p(\omega)$ for the analytic continuation problem. We then rewrite (4) with a change of variable $\omega \to s\omega$

$$\Pi_n = \int_{-\infty}^{\infty} d\omega \frac{\omega^2}{\omega^2 + \left(\frac{2\pi k_B}{\hbar} T\right)^2 n^2} p(\omega), \tag{7}$$

$$= \int_{-\infty}^{\infty} d\omega \frac{\omega^2}{\omega^2 + \left(\frac{2\pi k_B}{\hbar} \frac{T}{s}\right)^2 n^2} sp(s\omega). \tag{8}$$

This allow to isolate an alternative integrand argument $sp(s\omega)$, where s fits the definition of a *scale parameter*. If we consider that s carries the units of frequency, ω becomes a dimensionles continuous variable (analogous to the integer n in matsubara frequencies) and the term $\frac{2\pi k_B}{\hbar} \frac{T}{s}$ then corresponds to a dimensionless temperature.

C. Temperature-agnostic formulation

The above dimensionless formulation shows that there is a degenerancy between analytic continuation problems at different temperatures. We can make this degenerancy explicit by writing integral (7) as a functional $\Pi_n = \Pi_n[p(\omega)](T)$, such that lines (7) and (8) become

$$\Pi_n[p(\omega)](T) = \Pi_n[sp(s\omega)](T/s). \tag{9}$$

This equality tells us that the same response function components Π_n can be found at different temperatures T and T/s for respective distributions $p(\omega)$ and $sp(s\omega)$. In other words, the input $\mathbf{\Pi}$ has a continuum of compatible output $sp(s\omega)$, since s can take any value. Thus, without the temperature T or the frequency scale s specified, the analytic continuation task is ambiguous; For a well defined problem, we must either fix T or s.

Fixing T Working from a single temperature T_{ref} and considering $s = T_{\mathrm{ref}}/T$ in (9)

$$\Pi_n[p(\omega)](T_{\mathrm{ref}}) = \Pi_n[\tfrac{T_{\mathrm{ref}}}{T} p(\tfrac{T_{\mathrm{ref}}}{T}\omega)](T), \tag{10}$$

we see that results at any T can be obtained by rescaling the distribution $p(\omega)$ obtained at T_{ref}. However, with T_{ref} fixed in this way, the space of functions $p(\omega)$ we have to search to find the one compatible with $\mathbf{\Pi}$ includes all $sp(s\omega)$. In other words, any function $sp(s\omega)$ has to be "outputable" by the algorithm, and (10) tells us how to switch from one $sp(s\omega)$ to the other when changing temperature.

Fixing s By fixing s instead of T, we can actually narrow down that search space to distribution of a single scale (namely fixed second central moment, as shown below). We hypothesize that narrowing the search space is desirable, and thus concentrate on fixing s for the rest of this section.

We define *temperature-agnostic analytic continuation* as the task of mapping the response function $\Pi_n = \Pi_n[sp(s\omega)](T/s)$, obtained from temperature T/s and density function $sp(s\omega)$ for any s, to a single *reference* density $p(\omega)$, same for all s:

$$\Pi_n[sp(s\omega)](T/s) \longrightarrow p(\omega) \qquad\qquad \text{for any } s. \tag{11}$$

With analytic continuation defined this way, we can ignore temperature alltogether and work in terms of $\mathbf{\Pi}$ and $p(\omega)$ only. As we shall explain next (section I E), once the reference $p(\omega)$ is obtained, the missing scale s can be recovered from $\mathbf{\Pi}$. Also, there is freedom in the specific choice of reference $p(\omega)$ to represent all other $sp(s\omega)$, and this choice requires care.

D. Rescaling to fixed second moment

The above procedures depend on the existence of the neural network $p_{\boldsymbol{\theta}}(\omega, \mathbf{\Pi})$. Training such neural networks requires training data, more specifically several thousand pairs of inputs $\mathbf{\Pi}$ and output targets $p(\omega)$. Here we only explain how to remove any $sp(s\omega)$ from the training dataset to get only reference distributions $p(\omega)$.

To create each $(\mathbf{\Pi}, p(\omega))$ pair, we start by generating an arbitrary distribution $p_0(\omega)$ as described in section III. However, we presume the latter is not a reference distribution; it is inherently scaled as $p_0(\omega) = sp(s\omega)$, for some unknown s. To find s, we make a choice. We define all reference $p(\omega)$ so that they have the same second central moment μ_2. This μ_2 becomes a constant of the dataset,

$$\mu_2 = \int d\omega\omega^2 p(\omega). \tag{12}$$

Performing the change of variable $\omega \to s\omega$ in (12), we get

$$\mu_2 = s^2 \int d\omega\omega^2 p_0(\omega), \tag{13}$$

from which we isolate $1/s$ as

$$\frac{1}{s} = \sqrt{\frac{1}{\mu_2} \int d\omega\omega^2 p_0(\omega),} \tag{14}$$

letting us prepare the reference $p(\omega)$ from $p_0(\omega)$ as

$$p(\omega) = \frac{1}{s} p_0\left(\frac{1}{s}\omega\right). \tag{15}$$

Preparing every conductivity like this produces a dataset with only reference $p(\omega)$ as targets.

Fixing μ_2 have three noteworthy advantages. First, all reference distributions will have similar widths in the ω space; there will be no exessively narrow or wide $p(\omega)$, this is useful when working on a frequency grid. Second, if the generation process for $p_0(\omega)$ allows one to control the second central moment of the distribution analytically, the correct scale can be enforced without numerical overhead (we did not take advantage of this in the current work). And third, the constant μ_2 simplifies the recovery of conductivity explained in the next section.

E. Recovering conductivity

Suppose we have access to a neural network $p_{\boldsymbol{\theta}}(\omega, \mathbf{\Pi})$ (with $\boldsymbol{\theta}$ denoting all adjustable parameters) which takes $\mathbf{\Pi}$ as an input and approximates a single reference $p(\omega)$ as an output,

$$p_{\boldsymbol{\theta}}(\omega, \mathbf{\Pi}) \approx p(\omega). \tag{16}$$

To perform analytic continuation, a user starts from a response function $\Pi(i\omega_n)$ and prepare the according dimensionless input vector $\mathbf{\Pi}$, with

$$\Pi_n = \frac{\Pi(i\omega_n)}{\Pi(0)}. \tag{17}$$

Using the latter as an input for the algorithm, the user obtains the reference $p_{\boldsymbol{\theta}}(\omega, \mathbf{\Pi}) \approx p(\omega)$. However, the user desires not the reference distribution $p(\omega)$, but one of the many possible $sp(s\omega)$ corresponding to their particular temperature. From the change of variable $s\omega \to \omega$ in (3),

$$\lim_{\omega_n \to \infty} \omega_n^2 \frac{\Pi(i\omega_n)}{\Pi(0)} = \int_{-\infty}^{\infty} d\omega \omega^2 sp(s\omega) \tag{18}$$

$$= \frac{1}{s^2} \int\limits_{-\infty}^{\infty} d\omega \omega^2 p(\omega) \tag{19}$$

$$= \frac{\mu_2}{s^2}, \tag{20}$$

s can be isolated as

$$s = \sqrt{\frac{1}{\mu_2} \lim_{\omega_n \to \infty} \omega_n^2 \frac{\Pi(i\omega_n)}{\Pi(0)}}. \tag{21}$$

This expression depends on the user's temperature through $\omega_n = 2\pi n k_B T$. The user can then recover the desired conductivity by combining this s with (2) and (6),

$$\text{Re}\{\sigma(\omega)\} = \pi\Pi(0) s p_{\boldsymbol{\theta}}(s\omega, \mathbf{\Pi}). \tag{22}$$

II. USING NEURAL NETWORKS

A. Discrete formulation

Analytic continuation with neural networks is typically done with vector-to-vector neural networks [1–4]. The output of the neural network is then a vector $\hat{\boldsymbol{p}} = p_{\boldsymbol{\theta}}(\boldsymbol{\Pi})$ which targets the desired function $p(\omega)$ on a grid of real frequencies,

$$\omega_m = \Delta\omega m. \qquad\qquad \text{with } m \in \{-M, ...1, 0, 1, ..., M\}. \tag{23}$$

the spacing between frequencies $\Delta\omega$ can also be expressed as $\Delta\omega = \omega_{\max}/M$ with $\omega_{\max}$ marking the end of the sampling. Discretizing (7) on this grid,

$$[\boldsymbol{\Pi}]_n \approx \sum_m \Delta\omega \frac{\omega_m^2}{\omega_m^2 + \omega_n^2} p(\omega_m) \tag{24}$$

$$\approx \sum_m \frac{m^2}{m^2 + (\frac{2\pi k_B}{\hbar}\frac{T}{\Delta\omega}n)^2} \Delta\omega p(\Delta\omega m) \tag{25}$$

$$\approx \sum_m \frac{m^2}{m^2 + (\frac{2\pi k_B}{\hbar}\frac{T}{\Delta\omega}n)^2} [\boldsymbol{p}]_m \tag{26}$$

we define the target vector $\boldsymbol{p}$ as

$$[\boldsymbol{p}]_m = \Delta\omega p(\Delta\omega m). \tag{27}$$

Training the neural network then amounts to minimize the expected distance between $\hat{\boldsymbol{p}}$ and $\boldsymbol{p}$ over a large training set of $(\boldsymbol{\Pi}, \boldsymbol{p})$ pairs (see section II D).

One must be careful and remember that discretizing integrals as above is only an approximation. As such, we recommend to avoid sums like (26) and perform more accurate integration of $p(\omega)$ whenever possible. In fact, seeing the kernel in (26) as a matrix and inverting that matrix is not the same as solving the continuous inversion problem. This can be understood by imagining the limiting case in which one uses only a handful of points to approximate the integral. The additional difficulty due to the discretization error which is part of the matrix problem, but it is not intrinsic to the continuous problem. Equating the two can be misleading.

Note that $\Delta\omega$ occupies the same position as the scale parameter s which was the focus of the previous section, and indeed, it will play precisely this role in what follows.

B. Normalization

For various reasons (see section II D), it is beneficial to make the neural network's output components p_m sum to one. Considering the discretized normalization integral for $p(\omega)$,

$$\int d\omega p(\omega) = 1 \approx \sum_m \Delta\omega p(\Delta\omega m), \tag{28}$$

we see that p_m must correspond to $[\boldsymbol{p}]_m = \Delta\omega p(\Delta\omega m)$, as defined in the previous section. There is no problem in combining the rescaling $p(\omega) \to sp_0(s\omega)$ discussed in previous section with the above normalization. We will note it $\boldsymbol{p} \to \boldsymbol{p}_s$.

$$[\boldsymbol{p}_s]_m = s\Delta\omega p_0(s\Delta\omega m). \tag{29}$$

Both $\boldsymbol{p}$ and $\boldsymbol{p}_s$ sum to one, because both $p(s)$ and $sp(s\omega)$ integrate to one. Note, however, that the ω_m values at which those functions are sampled are not the same; respectively $\Delta\omega m$ and $s\Delta\omega m$.

C. Preparing data

Training the neural network requires to prepare a dataset. This section covers four approaches (numerated) to do so, and discuss them. In all cases, the idea is to generate a large number of random normalized functions $p_0(\omega)$ and use each of them to obtain one input-target pair, $(\boldsymbol{\Pi}, \boldsymbol{p})$.

Fixing T The simplest way is to work at fixed temperature.

1. $\boldsymbol{p}$: Sample $p_0(\omega)$ on the frequency grid $\omega_m = \Delta\omega m$ and normalize, $[\boldsymbol{p}]_m = \Delta\omega p_0(\Delta\omega m)$.

   $\boldsymbol{\Pi}$: Using $p_0(\omega)$, integrate (7) numerically, at chosen temperature T_{ref}.

Note that generating $(\boldsymbol{\Pi}, p_0(\omega))$-pairs at fixed temperature naturally produce multiple scales s. By this, we mean that multiple $p_0(\omega) = sp(s\omega)$ from the dataset could share the same reference $p(\omega)$ but with different values of s. Thus, using a single temperature T_{ref} to generate a dataset is sufficient to cover multiple dimensionless temperature $\mathcal{T} \propto \frac{T}{s}$. This is why the fixed T_{ref} scheme can be used to get results at other temperatures as explained in section I C.

Fixing s The better way is to work at fixed scale s. Here we consider the rescaling explained in section I D for which there are two equivalent ways to proceed.

2. $\boldsymbol{p}$: Get the reference distribution $p(\omega) = \frac{1}{s}p_0(\frac{1}{s}\omega)$ as explained in section I D. Sample it with the frequency grid $\omega_m = \Delta\omega m$, and normalize, $[\boldsymbol{p}]_m = \frac{1}{s}\Delta\omega p_0(\frac{1}{s}\Delta\omega m)$.

   $\boldsymbol{\Pi}$: Using the rescaled $p(\omega)$, integrate (7) numerically at any temperature T.

3. $\boldsymbol{p}$: Sample the original distribution $p_0(\omega)$ on a new frequency grid $\omega_m = \frac{\Delta\omega}{s} m$ with $\frac{1}{s}$ obtained as described in section I D and normalize, $[\boldsymbol{p}]_m = \frac{1}{s}\Delta\omega p_0(\frac{1}{s}\Delta\omega m)$.

   $\boldsymbol{\Pi}$: Using the original distribution $p_0(\omega)$, integrate (7) at adjusted temperature T/s.

Approaches 2 and 3 yield the same $\boldsymbol{p}$ and $\boldsymbol{\Pi}$ because of the degenerancy covered in section I C. Indeed, approach 2 uses the distribution $p(\omega)$ at temperature T whereas approach 3 uses the distribution $p_0(\omega) = sp(s\omega)$ at temperature T/s. Another important difference is the sampling grid, which must match the scale s to obtain the same $\boldsymbol{p}$ out of $p(\omega)$ and $p_0(\omega) = sp(s\omega)$.

Spurious correlations Let us now consider a fourth way of generating data, an hybrid bewtween 1 and 3, where we use the rescaled target $\boldsymbol{p}$ but fixed T_{ref} input $\boldsymbol{\Pi}$.

4. $\boldsymbol{p}$: Sample the original distribution $p_0(\omega)$ on a new frequency grid $\omega_m = \frac{\Delta\omega}{s} m$ with $\frac{1}{s}$ obtained as described in section I D and normalize $[\boldsymbol{p}]_m = \frac{1}{s}\Delta\omega p_0(\frac{1}{s}\Delta\omega m)$.

   $\boldsymbol{\Pi}$: Using the original distribution $p_0(\omega)$, Integrate (7) at temperature T_{ref}.

We will now argue that, because of the fixed T_{ref}, approaches 1 and 4 both lead to spurious correlations. Indeed, when all $\boldsymbol{\Pi}$ are computed at the same temperature, a narrow distribution $p_0(\omega)$ leads to a fast-decreasing $\boldsymbol{\Pi}$, and a wide $p_0(\omega)$ leads to slow-decreasing $\boldsymbol{\Pi}$, as expected from the tail relation (3). In approach 1, this correlation should manifest directly: fast-decreasing $\boldsymbol{\Pi}$ are paired with narrow $\boldsymbol{p}$ (meaning $\boldsymbol{p}$ has only a few large components near $\omega = 0$) while slow-decreasing $\boldsymbol{\Pi}$ are paired with wide $\boldsymbol{p}$ (many non-negligible components). This is what we mean by "spurious correlation" in that case. In approach 4, the spurious correlations get more subtle. Since targets $\boldsymbol{p}$ are rescaled to the same width (the rescaling leads to fixed second moment), the correlations between the width of $\boldsymbol{p}$ and decreasing rate of $\boldsymbol{\Pi}$ is removed. However, other spurious correlations might remain depending on the process used to generate $p_0(\omega)$. For example, if this process lead to simpler structure (fewer peaks) for narrow $p_0(\omega)$, and more complex structure (many peaks) for wide $p_0(\omega)$ (as in section III), then the these correlations between the structure of $p_0(\omega)$ and its width gets transferred through the one between the width of $p_0(\omega)$ and the decreasing rate of $\boldsymbol{\Pi}$. As a result, fast-decreasing $\boldsymbol{\Pi}$ are paired with simpler-structured $\boldsymbol{p}$ and slow-decreasing $\boldsymbol{\Pi}$ are paired with more complex-structured $\boldsymbol{p}$. This is what we mean by "spurious correlation" in that case. Such spurious correlations may or may not be desirable, but one must be aware they exists.

Approaches 2 and 3 completely avoid these uncontrolled correlations between the decrease of $\boldsymbol{\Pi}$ and the width or the structure of the output. Approach 2 do so by rescaling $p_0(\omega)$ and working only with the new function $p(\omega)$ which has fixed width (fixed second moment), and approach 3, which is equivalent, instead adjusts the temperature T/s and the frequency grid to the second moment of $p_0(\omega)$. With these approaches, if a single value of T is used, all $\mathbf{\Pi}$ end up with the same decrease rate (because all $\boldsymbol{p}(\omega)$ have the same width). In order to get a dataset which correctly spans $\mathbf{\Pi}$ space, randomly sampled values of T should be used. One can then control spurious correlations, if desired, by varying T as a function of $p_0(\omega)$.

The above considerations clarify the kind of bias which is induced when preparing the dataset. Constraining the output space of analytic continuation by choosing the training domain is the main benefit of using machine-learning for analytic continuation [5]. This bias in the space of allowed $p(\omega)$ is desirable since it acts as a form of regularization. Similarly, a bias in the way this space changes with temperature (the spurious correlations discussed above) might also be desirable. Indeed, high temperature $\mathbf{\Pi}$ don't provide as much information as low temperature $\mathbf{\Pi}$ and therefore simpler $p(\omega)$ might be easier to learn at high temperatures. As described in this section, by being careful with the dataset generation, it is possible to control these biases.

D. Training details

Neural network We considered the case where the input and output sizes are respectively $N = 128$ and $M = 512$. The optimal neural network we found is fully connected, with four hidden layers roughly of dimension 1000, 1350, 1700, and 1700. We use rectified linear unit (ReLU) activation functions and a Softmax output unit (producing a normalized $\boldsymbol{p}$). The Softmax requires normalized target outputs as explained in section II B. Weights are initialized using Xavier initialization.

Standardized inputs Any data that enters the neural network is standardized as $(\mathbf{\Pi} - \boldsymbol{\mu})/\boldsymbol{\sigma}$, where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{\Pi}]$ and $\boldsymbol{\sigma} = \sqrt{\mathbb{E}[\mathbf{\Pi}^2] - \mathbb{E}[\mathbf{\Pi}]^2}$ are respectively the average and the standard deviation of the training inputs dataset(element-wise), with expectation values taken over the training set only. The random noise added to $\mathbf{\Pi}$ to simulate Quantum Monte-Carlo response functions is renewed at every epoch and it is added before standardization.

Loss functions As a training objective, we compared the performance of cross-entropy,

$$\mathrm{CE}(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \frac{1}{M} \sum_m p_m \log \hat{p}_m \tag{30}$$

mean-square error,

$$\mathrm{MSE}(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \frac{1}{M} \sum_m \sqrt{(p_m - \hat{p}_m)^2} \tag{31}$$

and mean-absolute error

$$\mathrm{MAE}(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \frac{1}{M} \sum_m |p_m - \hat{p}_m|. \tag{32}$$

We found that using the mean-absolute error as the training objective leads to better final performance for all three losses.

Optimizer We use the Adam optimizer with learning rate of $\alpha = 8 \times 10^{-5}$ and a scheduler decreasing the rate as $\alpha \to 0.216\alpha$ whenever the validation loss doesn't improve for five straight epochs. That decrase stops when the learning rate reach $\alpha \sim 10^{-10}$. We also use learning rate warm-up, linearly increasing α from zero to its value of $\alpha = 8 \times 10^{-5}$ during the first epoch.

Regularization Our random search for optimal hyper-parameters revealed that neither L2 regularization, dropout, nor batchnorm improve results. However, using slightly larger input noise at training time than at test time leads to better performance, and training on more complex data (Beta peaks) can lead to better performance on simple data (Gaussian peaks) than training on simple data.

E. Usage details (inference)

Once the user obtain the estimate $\hat{\boldsymbol{p}}$ with the neural network, the only thing left to do is to find the frequency grid $\Delta\omega$ on which to interpret that estimate. Considering the discretization (24) for the special case of the tail relation (3),

$$2\pi k_B T \lim_{n\to\infty} n^2 \Pi_n = \Delta\omega^2 \sum_m m^2 p_m \tag{33}$$

We see that $\boldsymbol{\Pi}$ and $\boldsymbol{p}$ provide a closed relation between the temperature and the frequency spacing. If we assume that the discretization is a good approximation for the integral and that the last component Π_N correctly captures the limit $n \to \infty$, then we can cast this relation as

$$\Delta\omega^2 = \frac{2\pi k_B N^2 \Pi_N}{\sum_m m^2 p_m} T. \tag{34}$$

This is a discrete version of (21) which can be used to recover $\Delta\omega$ from the user's temperature T. In principle, however, the continuous version should be preferred. In particular, a simple way to improve on (34) is to use a better estimate for $\lim_{\omega_n\to\infty} \omega_n^2 \Pi_n$, than $2\pi k_B T N^2 \Pi_N$. Note that if the fixed T_{ref}

III. GENERATING DENSITY FUNCTIONS

A. Gaussian peaks

The first way to generate random distributions is a weighted sum of Gaussian peaks

$$p(\omega) = \sum_{j=1}^{N} \frac{\pi A_j}{\sqrt{2\pi}\sigma_j} \exp\Big(-\frac{1}{2}\Big(\frac{\omega - \omega_j}{\sigma_j}\Big)^2 \Big). \tag{35}$$

Each peak parametrized by weight A_j, center ω_j, and width σ_j.

Most machine learning papers on analytic continuation were trained on sets of Gaussian peaks as described above [1–3, 5]. The most influential paper in that respect is arguably that of Arsenault et al [5], which correctly identifies the importance of the data as a mean for regulation. Unfortunately, some of the constraints used for their data generation are described qualitatively, which prevents faithful reproduction. The work of Yoon et al. [3] uses a somewhat complicated and arbitrary distribution for the number of peaks. The dataset defined by Fournier et al.[1] and re-used in Xie et al.[2], are more straigthforward to generate. Note however, that their definition implies that $A_j = \sigma_j$ which favours very rounded distributions.

B. Beta peaks

The second way to generate random distributions is a weigthed sum of β-peaks

$$p(\omega) = \sum_{j=1}^{N} \frac{\pi A_j}{\tilde{\sigma}_j} \mathrm{Beta}\Big(\frac{\omega - \tilde{\omega}_j}{\tilde{\sigma}_j}; \alpha_j, \beta_j\Big). \tag{36}$$

Each peak is parametrized by the amplitude A_j, the α_j and β_j of the Beta distribution,

$$\mathrm{Beta}(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{1-\alpha}(1-x)^{1-\beta}, \tag{37}$$

and, since the latter is only non-zero for $0 \leq x \leq 1$, the center ω_j and the width σ_j of the peaks we generate must be adjusted to the mean and standard deviation of the distribution, as

$$\tilde{\omega}_j = \omega_j + \sigma_j \frac{\alpha_j}{\alpha_j + \beta_j} \tag{38}$$

$$\tilde{\sigma}_j = \sigma_j \Big(\frac{(\alpha_j + \beta_j + 1)(\alpha_j + \beta_j)^2}{\alpha_j \beta_j} \Big)^{\frac{1}{2}}. \tag{39}$$

Note that for large $\alpha = \beta$, the Beta distribution becomes very similar to the Gaussian distribution. Values of $\alpha = \beta \gtrsim 10$ are sufficient to make them indistinguishable on a plot. However, the tails of the Gaussian extend to infinity, whereas the tails of the Beta distribution are always bounded within $0 \leq \frac{\omega-\tilde{\omega}_j}{\tilde{\sigma}_j} \geq 1$. Nevertheless, we hypothesize that learning with Beta-distribution will easily generalize to Gaussian distributions.

Note also that for disctrete values of α and β, the beta distribution is identical to Bernstein polynomials

$$b_{m,n}(x) = \binom{m}{n} x^n (1-x)^{m-n} \tag{40}$$

where $\binom{m}{n}$ is the binomial coefficient, and m and n are integer parameters. They can be related to the real-valued α and β of the Beta distribution by

$$m = \alpha + \beta - 2 \tag{41}$$

$$n = \alpha - 1. \tag{42}$$

The normalization, mean, and standard deviation of these polynomials are respectively given by

$$a_{m,n} = \frac{1}{m+1}, \tag{43}$$

$$\mu_{m,n} = \frac{1+n}{2+m}, \tag{44}$$

$$\sigma_{m,n} = \left(\frac{(1+n)(2+n)}{(2+m)(3+m)} - \frac{(1+n)^2}{(2+m)^2} \right)^{\frac{1}{2}}. \tag{45}$$

so it is possible to adjust them to control their center and width as we did for the Beta distribution. It is worth noting that Bernstein polynomials are used to write Bezier curves in polynomial form, and they provide a constructive proof of the Weierstrass approximation theorem.

ACKNOWLEDGMENTS

This project was started in close collaboration with Reza Nourafkan and André-Marie Tremblay. I must also acknowledge additional useful discussions with Yoshua Bengio, Maxime Charlebois, Samuel Desrosiers, Sébastien Lachapelle, Vincent Mai and Francesco Rotella.

[1] R. Fournier, L. Wang, O. V. Yazyev, and Q. S. Wu, Phys. Rev. Lett. **124**, 1 (2020), arXiv:1810.00913.

[2] X. Xie, F. Bao, T. Maier, and C. Webster, , 1 (2019), arXiv:1905.10430.

[3] H. Yoon, J. H. Sim, and M. J. Han, Phys. Rev. B **98**, 1 (2018), arXiv:1806.03841.

[4] L. Kades, J. M. Pawlowski, A. Rothkopf, M. Scherzer, J. M. Urban, S. J. Wetzel, N. Wink, and F. Ziegler, , 1 (2019), arXiv:1905.04305.

[5] L. F. Arsenault, R. Neuberg, L. A. Hannah, and A. J. Millis, Inverse Probl. **33**, 1 (2017), arXiv:1612.04895.

Appendix A: Refreshers on fundamentals

1. Complex analysis

Residues The *residue theorem* expresses a contour integral in the complex plane as the sum of residues of the poles (sigularities) in the region enclosed by the counter-clockwise contour

$$\oint dz f(z) = 2\pi i \sum_k \mathrm{Res}[f(z), z_k] \tag{A1}$$

$$= 2\pi i \sum_k \frac{1}{(m_k+1)!} \lim_{z\to z_k} \frac{d^{m_k-1}}{dz^{m_k-1}} \Big[(z-z_k)f(z)\Big] \qquad \text{for poles of order } m_k \tag{A2}$$

$$= 2\pi i \sum_k \lim_{z\to z_k} (z-z_k)f(z) \qquad \text{for poles of order 1} \tag{A3}$$

As an example, it allows to evaluate following integral

$$\oint dz \frac{f(z)}{z-z_0} = 2\pi i f(z_0). \tag{A4}$$

Causality When considering a spectrum $f(z)$ (Fourier transform of $f(t)$) which is analytical in the upper-half complex frequency plane, except at certain poles, and for which $\lim_{|z|\to\infty} f(z) = 0$, the *lemme of Jordan* allows us to complete the Fourier integral for all $t < 0$ with a non-contributing half-circle contour. This allow to turn the integral into the sum of residues

$$f(t<0) = \int_{-\infty}^{\infty} dz e^{-izt} f(z) + \underbrace{\int_{\frown} dz e^{-izt} f(z)}_{=0} \tag{A5}$$

$$= 2\pi i \sum_k \mathrm{Res}[f(z), z_k]. \tag{A6}$$

As a consequence, a pole-free spectrum $f(z)$ in the upper-half complex plane ($z = z' + iz''$ with $z'' > 0$ such that $e^{-z''t}$ ensures convergence) correspond to a causal function in time ($f(t<0) = 0$).

Sokhotski–Plemelj theorem Finally, we remind the useful theorem

$$\lim_{\eta\to 0} \int_{\infty}^{\infty} \frac{f(z)}{z \pm i\eta} dz = \mathcal{P} \int_{\infty}^{\infty} \frac{f(z)}{z} dz \mp \pi i f(z). \tag{A7}$$

2. Linear response theory

In classical systems, a "linear response function" $\chi(\omega)$ relates a driving force $F(t)$ to the solution $x(t)$ of a given differential equation. It is defined from a linear relation in the frequency domain as

$$x(\omega) = \chi(\omega) F(\omega). \tag{A8}$$

In the time domain, because of the the convolution theorem, this definition becomes

$$x(t) = \int_{-\infty}^{\infty} dt' \chi(t-t')F(t'). \tag{A9}$$

In particular, any differential equations of the form

$$\sum_{n=0}^{\infty} a_n \frac{\partial^n x(t)}{\partial t^n} = \sum_{n=0}^{\infty} b_m \frac{\partial^m F(t)}{\partial t^m}, \tag{A10}$$

can be Fourier transformed such that the linear response is the exact response

$$x(\omega) = \underbrace{\frac{\sum_m b_m(-i\omega)^m}{\sum_n a_n(-i\omega)^n}}_{\chi(\omega)} F(\omega), \tag{A11}$$

In quantum systems, response functions show up as $\chi^R_{A,B}(\omega)$ in time-dependent perturbation theory. Considering the perturbation $\delta H(t) = a(t)A(t)$, proportional to observable $A(t)$ (hence the subscript A in $\chi^R_{A,B}(\omega)$), and its effect on the expectation value of another observable $\langle B(t)\rangle$ (hence the subscript B in $\chi^R_{A,B}(\omega)$), the linear approximation for time evolution in the interaction picture (not shown here) leads to an expression equivalent to (A9),

$$\langle B(t)\rangle - \langle B(t)\rangle_0 = \frac{i}{\hbar}\int_{-\infty}^{t} dt' \langle [B(t), \delta H(t')]\rangle \tag{A12}$$

$$= \int_{-\infty}^{\infty} dt' \underbrace{\frac{i}{\hbar}\langle [B(t), A(t')]\rangle \theta(t-t')}_{\chi^R_{A,B}(t-t')} a(t'). \tag{A13}$$

Note that the response is said to be *causal*, or *retarded* (hence the R superscript) because the Heaviside step function $\theta(t-t')$ ensures the response is zero if $t < t'$. This lead to a pole-free spectrum in the upper-half complex plane (as discussed at equations (A5)-(A6)).

3. Spectral representation

Taking the Fourier transform of the above $\chi^R_{A,B}(t-t')$ leads to the spectral representation. Omitting the A, B subscript for simplicity, we can rewrite it as

$$\chi^R(t-t') = 2i\chi''(t-t')\theta(t-t'), \tag{A14}$$

with

$$\chi''(t-t') = \frac{1}{2\hbar}\langle [A_I(t), B_I(t')]\rangle. \tag{A15}$$

The Fourier transform of (A14) is then

$$\chi(z) = \int_{-\infty}^{\infty} \frac{d\omega}{\pi} \frac{\chi''(\omega)}{\omega - z} \tag{A16}$$

where $\chi''(\omega)$ is the Fourier transform of $\chi''(t-t')$, and is real-valued if the latter is even. Since $\chi^R(t-t')$ is causal, $\chi(z)$ is analytic in the upper half plane of the complex frequency z (as discussed at equations (A5)-(A6)). The above $\chi(z)$ unifies various definitions of the response function, notably the advanced response function $\chi^A(\omega) = \chi(\omega - i\eta)$, the retarded response functions $\chi^R(\omega) = \chi(\omega+i\eta)$, and most important, the the Matsubara response functions $\chi(i\omega_n)$. The latter is sampled at discrete Matsubara frequencies $\omega_n = \frac{2\pi}{\beta}n$ with n integer and the inverse temperature $\beta = \frac{1}{k_B T}$.

Finally, our definitions ensure that $\chi''(\omega)$ is the imaginary part of $\chi^R(\omega)$, as seen with (A7)

$$\begin{aligned}
\chi^R(\omega) = \chi(\omega + i\eta) &= \int_{-\infty}^{\infty} \frac{d\omega'}{\pi} \frac{\chi''(\omega')}{\omega' - \omega - i\eta} && \text{(A17)}\\
&= \mathcal{P}\int_{-\infty}^{\infty} \frac{d\omega'}{\pi} \frac{\chi''(\omega')}{\omega' - \omega} + i\chi''(\omega) && \text{(A18)}\\
&= \chi'(\omega) + i\chi''(\omega). && \text{(A19)}
\end{aligned}$$

The first term of (A18) is one of the two *Kramer-Kronig relations* (*Hilbert transforms*):

$$\begin{aligned}
\chi'(\omega) &= \mathcal{P}\int_{-\infty}^{\infty} \frac{d\omega'}{\pi} \frac{\chi''(\omega')}{\omega' - \omega} && \text{(A20)}\\
\chi''(\omega) &= -\mathcal{P}\int_{-\infty}^{\infty} \frac{d\omega'}{\pi} \frac{\chi'(\omega')}{\omega' - \omega}. && \text{(A21)}
\end{aligned}$$

4. Optical conductivity

One particularly important response function is the current-current correlation function $\Pi^R(\omega) = \chi^R_{j_x,j_x}(\omega)$, which is related to the optical conductivity $\sigma(\omega)$ by the *Kubo formula*

$$\sigma(\omega) = \frac{1}{i(\omega + i\eta)}\left[\Pi^R(\omega) - \frac{ne^2}{m}\right]. \tag{A22}$$

From there, we can develop by substituting the spectral representation of $\Pi^R(\omega)$,

$$\begin{aligned}
\sigma(\omega) &= \frac{1}{i(\omega + i\eta)}\left[\int_{-\infty}^{\infty} d\omega' \frac{\Pi''(\omega')}{\omega' - (\omega + i\eta)} - \int_{-\infty}^{\infty} d\omega' \frac{\Pi''(\omega')}{\omega'}\right] && \text{(A23)}\\
&= \frac{1}{i(\omega + i\eta)}\int_{-\infty}^{\infty} d\omega' \Pi''(\omega') \frac{\omega' - (\omega' - (\omega + i\eta))}{\omega'(\omega' - (\omega + i\eta))} && \text{(A24)}\\
&= \int_{-\infty}^{\infty} d\omega' \frac{\Pi''(\omega')/i\omega'}{\omega' - (\omega + i\eta)} && \text{(A25)}\\
&= \frac{\Pi''(\omega')}{\omega'} - i\mathcal{P}\int_{-\infty}^{\infty} d\omega' \frac{\Pi''(\omega')/\omega'}{\omega' - \omega}. && \text{(A26)}
\end{aligned}$$

The last terms in lines (A22) and (A23) are equal as a result of the *f-sum rule*. The last line was again obtained with (A7) and allows us to extract the important relation

$$\mathrm{Re}\{\sigma(\omega)\} = \frac{\Pi''(\omega)}{\omega}. \tag{A27}$$

For Matsubara frequencies $z = i\omega_n$, the spectral representation $\Pi(z)$ then simplifies to

$$\begin{aligned}\Pi(i\omega_n) &= \int_{-\infty}^{\infty} \frac{d\omega}{\pi} \frac{\omega}{\omega - i\omega_n} \mathrm{Re}\{\sigma(\omega)\} && \text{(A28)}\\ &= \int_{-\infty}^{\infty} \frac{d\omega}{\pi} \frac{\omega^2}{\omega^2 + \omega_n^2} \mathrm{Re}\{\sigma(\omega)\} + \underbrace{i \int_{-\infty}^{\infty} \frac{d\omega}{\pi} \frac{\omega\omega_n}{\omega^2 + \omega_n^2} \mathrm{Re}\{\sigma(\omega)\}}_{=0}. && \text{(A29)}\end{aligned}$$

The last term vanishes because $\mathrm{Re}\{\sigma(\omega)\}$ is even (not shown), which guarantees that $\Pi(i\omega_n)$ is real. Equation (A29) is the same as (1).

Appendix B: Miscellaneous

1. Lorentz combs with analytic integration

An alternative way to generate a random distributions is to express $\Pi''(\omega)$ as a sum of Dirac delta functions approximated by Lorentzians (Cauchy distributions)

$$\begin{aligned}\mathrm{Re}\{\sigma(\omega)\} = \frac{\Pi''(\omega)}{\omega} &= \frac{\pi}{\omega} \sum_{j=i}^{J} \Big[a_j \delta(\omega - \omega_j) - a_j \delta(\omega + \omega_j)\Big] \qquad \text{with } \omega_j \geq 0 && \text{(B1)}\\ &\approx \sum_{j=i}^{J} \frac{\pi a_j}{\omega} \Big[\frac{1}{\pi} \frac{\eta}{(\omega - \omega_j)^2 + \eta^2} - \frac{1}{\pi} \frac{\eta}{(\omega + \omega_j)^2 + \eta^2}\Big] && \text{(B2)}\\ &\approx \sum_{j=i}^{J} \frac{a_j \eta}{\omega} \frac{\big((\omega + \omega_j)^2 + \eta^2\big) - \big((\omega - \omega_j)^2 + \eta^2\big)}{\big((\omega - \omega_j)^2 + \eta^2\big)\big((\omega + \omega_j)^2 + \eta^2\big)} && \text{(B3)}\\ &\approx \sum_{j=i}^{J} \frac{a_j \eta}{\omega} \frac{4\omega\omega_j}{(\omega^2 + \omega_j^2 + \eta^2)^2 - (2\omega\omega_j)^2}. && \text{(B4)}\end{aligned}$$

This form has the advantage of removing the integral when computing $\Pi(i\omega_n)$, as hinted by the Dirac delta. We can even track the finite width η by substituting (B2) into (A28)

$$\Pi(i\omega_n) \approx \sum_{j=i}^{J} a_j \int_{-\infty}^{\infty} \frac{d\omega}{\pi} \frac{1}{\omega - i\omega_n} \Big[\frac{\eta}{(\omega - \omega_j)^2 + \eta^2} - \frac{\eta}{(\omega + \omega_j)^2 + \eta^2}\Big], \tag{B5}$$

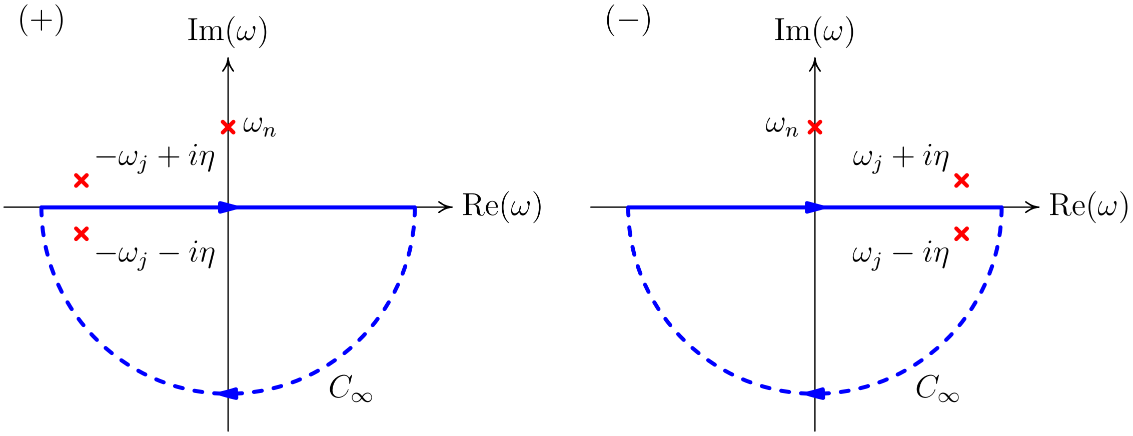

FIG. 1. Contours for integral $I_\pm$ (B7). The C_∞ arcs do not contribute thanks to Jordan's lemma.

where the residue theorem (A1) let us solve the two integrals using the contours shown in Fig. 1

$$I_\pm = \int_{-\infty}^{\infty} \frac{d\omega}{\pi} \frac{1}{\omega - i\omega_n} \frac{\eta}{(\omega \pm \omega_j)^2 + \eta^2} \tag{B6}$$

$$= \frac{\eta}{\pi} \int_{-\infty}^{\infty} d\omega \frac{1}{\omega - i\omega_n} \frac{1}{(\omega \pm \omega_j - i\eta)(\omega \pm \omega_j + i\eta)} \tag{B7}$$

$$= \frac{\eta}{\pi} \left[-2\pi i \frac{1}{(\mp\omega_j - i\eta - i\omega_n)(-2i\eta)} \right] \tag{B8}$$

$$= \frac{\mp\omega_j + i\eta + i\omega_n}{\omega_j^2 + (\omega_n + \eta)^2}, \tag{B9}$$

simplifying (B5) to

$$\Pi(i\omega_n) \approx \sum_{j=i}^{J} \frac{2a_j\omega_j}{\omega_j^2 + (\omega_n + \eta)^2}. \tag{B10}$$

Therefore, (B4) and (B10) can be used to generate $p(\omega)$ and $\Pi(i\omega_n)$ data without having to perform costly numerical integrals. We can further get arbitrary $p(\omega)$ either by modulating the peak amplitudes a_j, or we can distribute their centers ω_j non-uniformly. The resulting function will be smooth as long as the width of the peaks are at least twice as large as the spacing between them, and the moments of the spectrum will be well-defined if a large enough number J of Lorentzians are summed. Unfortunately, in practice, this sum can become as costly as a numerical integral.