Predicting the future of complex signals

The information bottleneck applied to the generalized Langevin equation

Internship report

Geert Herman Albert Schulpen

Nanomaterials Science and Experimental Physics

Faculty of Science

Utrecht University

Supervisors:

Age Tjalma, MSc

Dr. Jenny Poulton

Biochemical networks, AMOLF

November 18, 2022

Abstract

Biological systems live in dynamic environments and thus need to adapt to changes in that environment. However, adapting costs time, so the system needs to predict its future environment based on past environments. Therefore, the system needs to store some information about this history from which it is possible to make this prediction. What information is optimal to extract from the signal history depends on the covariance function of the signal. The information bottleneck provides a framework to determine what information is relevant for this prediction.

We look at a system that predicts the value of a signal generated by the generalized Langevin equation at one specific future time. The generalized Langevin equation provides a non-Markovian signal with a non-trivial memory function. We analyze two specific types of memory functions; an exponential decay, and a combination between an exponential decay and a Dirac delta function. These memory functions can generate a wide variety of signals. However, we find that the optimal response for each memory function does not vary strongly with parameter changes. The information bottleneck method requires discretization of the input signal. This discretization induces information loss, which masks what information is necessary to perform the prediction.

A vector autoregressive model is constructed, which generates a discrete signal. This eliminates the information loss by the information bottleneck. As such, we find the information required to predict the future signal value for both signal types. For the exponential decay memory function, the future is predicted optimally from the most recent three signal values. The prediction seems based on determining the signal's instantaneous position, velocity and acceleration.

For the combined exponential decay and Dirac delta function memory function, the future cannot be predicted optimally from a limited set of past signal values. If the exponential decay is the more important memory function, the response mirrors the previous case but is enhanced by a strongly decaying exponential tail. This tail seems to elucidate the effect of the instantaneous Dirac delta function. On the other hand, if the Dirac delta function is the more important memory function, the response mirrors the result of the information bottleneck for the stochastic damped harmonic oscillator (Lotte Slim, 2020). However, it is enhanced by a slowly decaying exponential tail. The tail seems to elucidate the effect of the history-dependent exponential decay memory function.

Contents

1 Introduction

Biological systems exist in a dynamic environment. The system might rely on different machinery depending on the state of the environment. Several strategies are available to select what machinery is active. The simplest strategy is to keep everything active. In that case, the system needs to spend more energy on the building and upkeep of all the machinery, but it will be ready for any environment. On the other hand, the system might statistically switch from one set of machinery to another one. In this case, the system spends less energy on maintaining all the machinery, but it might not have the correct machinery ready at the right time. Finally, the system might use some sensing mechanism to gather information about the environment and choose what machinery to activate to adapt to changes in the environment. For this strategy to be viable, the system needs to be able to predict its future environment based on the information it gathered in the past. This is needed because activating the correct machinery takes some finite length of time. In this case, the system spends less energy maintaining the machinery and instead uses some energy to sense the environment and predict the future environment. Which of these strategies is the best strategy in any specific case depends on the energy costs and accuracy required.

The last strategy is relevant in gene expression [1, 2], vision [3], and the chemotaxis network [4]. In these systems, prediction plays a part in the optimal survival strategy. However, a trade-off exists between the prediction's accuracy and the prediction system's resource cost. In general, the information a system has about the past limits the amount it has about the future. A system can make the best possible prediction about the future environment if everything about the past environments is known. However, storing all this information would be highly resource-intensive. We assume that a cell has limited resources for prediction. Therefore, the system must store a compressed representation of past environments to limit resource costs.

How should the signal be compressed? An ideal system would only sense relevant information for prediction, discarding irrelevant information. Given that many biological systems have existed for a long time, evolution suggests that we might expect these systems to be nearly ideal [5]. We need a definition of relevant information to determine if these systems are indeed ideal.

For this, we will use the information bottleneck method. This framework has been used to study biological systems in several contexts before. For example, in determining optimal sensing networks for transcription factors in fly-embryo development [2], prediction of output in neurons in the retina and visual cortex [3, 6, 7], allele frequencies in different generations [7] and evaluating the performances of the push-pull network and chemotaxis network as cellular sensing devices [8, 9].

The information bottleneck framework builds on information theory. This theory, originally introduced by C. Shannon, provides a mathematical framework for considering information transmission [10]. However, it does not distinguish between information and relevant information, which is what the information bottleneck adds [11].

The information bottleneck establishes a method to find the optimal compression of information to do a certain task. For example, a task could be to determine the value of a variable y. If a second variable, x, is correlated with y, then knowing x can aid in determining the value of y; there is mutual information between x and y $I(x;y) > 0$. If x must be compressed to $\tilde{x}$, then the mutual information between that compression and y is reduced relative to the uncompressed representation x; $I(x;y) >= I(\tilde{x};y)$. An optimal compression would maximise $I(\tilde{x};y)$ while minimising $I(\tilde{x};x)$. We return to the information bottleneck in more detail in section 3. This framework is powerful for defining relevant information and considering the trade-off between prediction accuracy and resource costs.

In this work, we consider the task of predicting a signal value at some specific future time. As input we consider the signal values at discretized times in the past. The compression variable is a scalar that is a linear function of the past signal values. In other words, this is a weighted average of the input. The information bottleneck in this context requires us to find the correlation function of the signal. This allows us to determine the optimal weights for each of these historical signal values to determine the compression variable such that the compression has a maximal correlation with the future signal value. We call these optimal weights the information bottleneck kernel. This information bottleneck kernel thus shows what past signal values are most important in predicting the future of the signal. In one previous work, this analysis was performed for a Markovian signal in the form of the Ornstein-Uhlenbeck model [8]. In another previous work, the analysis was done for a non-Markovian signal in the form of the stochastic damped harmonic oscillator [9]. For the Markovian signal, the optimal prediction network should only

remember the current signal value, which is also the best prediction for the signal value [8]. Meanwhile, the two most recent measurements are important for the stochastic damped harmonic oscillator. The weights of these measurements are such that the best prediction is the current signal value added to the best guess of the instantaneous derivative, which is determined from the difference between the two measurements [9]. In this work, we will extend the framework to fully non-Markovian signals with non-trivial memory functions, which we generate using the generalized Langevin equation. Studying these signals will allow us to investigate whether more complex compressions of the past trajectory are required to predict signals with more complex correlation structures.

In the first part of the report, we will introduce the generalized Langevin equation and, in particular, the form of several non-trivial memory functions. We will examine signal trajectories generated by the generalized Langevin equation to aid in understanding the behaviour of these signals. Next, we will introduce the information bottleneck method in detail and show how the autocovariance of a signal is used to determine the optimal compression. Subsequently, we will find the autocovariance of the generalized Langevin signal for our non-trivial memory functions and use this to find the optimal compression. We will find that the discretization of the continuous signal introduces artefacts in the information bottleneck kernel, so it is more useful to consider a truly discrete signal. To this end, we will introduce a vector autoregressive model that models the generalized Langevin equations and determine its autocovariance. This discrete model allows us to eliminate the artefacts in the information bottleneck kernel and show the optimal compression of a generalized Langevin system.

2 Memory kernels

The Langevin equation was introduced by Paul Langevin as a theoretical description of Brownian motion [12, 13]. The equation describes the movement of a particle in a fluid, which exerts a viscous resistance force on the particle. The equation for the coordinate x at time t of a particle of mass m that experiences, on average, a friction force of magnitude $-\lambda\frac{dx}{dt}$ is written as [12]

$$m\frac{d^2x}{dt^2} = -\lambda\frac{dx}{dt} + \eta, \tag{1}$$

where η is a random force due to the irregularity of collisions with solvent molecules. Both the friction force and the random force stem from collisions with solvent molecules. It can be shown that the magnitudes of both forces are intimately related as a manifestation of the broader fluctuation-dissipation theorem [14]. The Langevin equation has proven to be a powerful tool to describe the Brownian motion process. Moreover, it has established new physics and mathematics, being one of the first stochastic differential equations [13]. An important assumption needed for the Langevin equation is the assumption that the particle performing the Brownian motion is much heavier than the solvent molecules. As such, the molecular relaxation time is very fast compared to the timescales of the Brownian particles and collisions of the solvent molecules with the Brownian particle are not correlated to each other [14, 15]. While this assumption is mathematically convenient, it breaks down for short times or light-weight particles [14]. In those cases, the relaxation times in the solvent are important, so collisions with solvent molecules have to be considered correlated in time. The generalized Langevin equation is a method to include this correlation in the model of Brownian motion. Instead of an instantaneous friction force and uncorrelated noise, the equation assumes that the noise is correlated with itself at different times, and the resistance force is based on the history of the velocity, as enforced by the fluctuation-dissipation theorem. The generalized Langevin equation in one coordinate x for a particle with mass m, in a harmonic potential with force constant κ for $t \geq 0$ is written as [16]

$$m\ddot{x}(t) = -\int_{t_m}^{t} dt^0\ \Gamma(t-t^0)\dot{x}(t^0) - \kappa x(t) + \eta(t), \tag{2}$$

where we write $\dot{x}$ for the time derivative, and $\Gamma(t)$ is the memory kernel that defines a corrective force based on the past trajectory between $t_m < 0$ and 0, $x(t)$ where $t_m \leq t < 0$. The memory kernel prioritizes some aspects of this past trajectory over others. The fluctuation-dissipation theorem in this case stipulates $\langle\eta(t)\eta(t^0)\rangle = k_BT\Gamma(|t-t^0|)$.

Only after choosing the memory function $\Gamma(t)$ is the system fully specified. In this report, we consider several memory kernels in depth. We first introduce a selection of memory kernels to consider as input statistics for a biological system and comment on their behavior. For this section, we will interpret the generalized Langevin equation as an equation of motion.

2.1 Dirac delta function

In the case that the kernel is a Dirac delta function, only the current value of the velocity influences the corrective force on the system. In this case, this corrective force is exactly a friction force. Here, the original Langevin equation of a damped harmonic oscillator with stochastic driving is recovered. The Langevin equation in one dimension x for a particle with mass m is given as

$$m\ddot{x} = -2\zeta\sqrt{\kappa m}\dot{x}(t) - \kappa x(t) + \eta(t), \tag{3}$$

where ζ sets the magnitude of the friction force. This equation of motion includes three forces, the friction force $-\zeta\sqrt{\kappa m}\dot{x}(t)$, the harmonic potential $-\kappa x(t)$, and the random white noise force $\eta(t)$. The damping parameter ζ tunes the system into one of three distinct regimes [17]. For $\zeta < 1$, the friction force is small compared to the harmonic potential. Here, $x(t)$ traces a decaying oscillation around the equilibrium position at $x = 0$. In this case, the system is in the underdamped regime. For $\zeta > 1$, the friction is high compared to the potential and $x(t)$ slowly decays to equilibrium without oscillating. In this case, the system is in the overdamped regime. Finally, the system is in the critical damping regime for $\zeta = 1$. In this regime, the particle reaches and stays in the equilibrium position the fastest (assuming m and κ are fixed). The stochastic damped harmonic oscillator (SDHO) model is well known, so we use this as a reference case for comparison to the other memory kernels.

2.2 Exponential

One set of commonly used memory kernels is a sum over several exponential decays [18–20]. In these kernels, the most recent time point is prioritized, but the past trajectory's behaviour influences the force on the system in the present. Memory kernels of this type are used to model turbulent hydrodynamics [20] and the behavior of tracer particles in viscoelastic liquids [19,21]. We will be studying a single exponential decay to get a feel for how this memory kernel behaves. We define the memory kernel

$$\Gamma(t) = \frac{B}{\tau} e^{-t/\tau}, \tag{4}$$

where B is the strength of the memory effect, and τ is the timescale over which the memory decays. From Equation (2), we find the generalized Langevin equation for this case

$$m\ddot{x} = -\frac{B}{\tau}\int_{t_m}^{t} dt' \, e^{-(t-t')/\tau}\dot{x}(t') - \kappa x(t) + \eta(t). \tag{5}$$

While for the SDHO we have some intuition for what to expect, this memory term adds significant complexities. In Figures 1, 2 and 3, we plot deterministic trajectories for three sets of parameters and compare these to trajectories for the SDHO. The trajectories were generated by discretizing the equation using an Euler method and assuming starting conditions $x(t < 0) = 7$. To recover a value for ζ that yields a fair comparison we equate $2\zeta\sqrt{\kappa m} = \frac{B}{\tau}\int_{1}^{t} dt' \, e^{-(t-t')} = B$. With this choice, the corrective force experienced by a particle with constant velocity is the same using either equation. In Figure 1, both trajectories look like an underdamped harmonic oscillator but with different parameters. Indeed, in this case, the corrective force is small compared to the potential.

Furthermore, in Figure 2, both trajectories look similar to an overdamped harmonic oscillator. Here the corrective force is large compared to the potential, and τ is small, so the memory function highly favors the most recent velocities. Indeed, in the limit $\tau \to 0$, the exponential memory kernel recovers the delta function, such that for sufficiently small values for τ, this kernel will yield similar results to the Langevin equation.

Finally, in Figure 3, we see a distinctly different behavior between the two trajectories. Here the corrective force is again large. In the SDHO system, this means the system is overdamped and does not oscillate. However, in the exponential kernel case, where now τ is large, the system oscillates, but not around zero. Here, we see the consequences of the memory kernel prioritizing velocities in the past. Consider the point labelled with the green dot at $t = 1.36$. Here the current velocity is zero. Thus if the corrective force were a simple frictional force, the corrective force would also be zero. However, the corrective force is influenced by the past for a system with an exponential memory kernel. At all recent points in the past

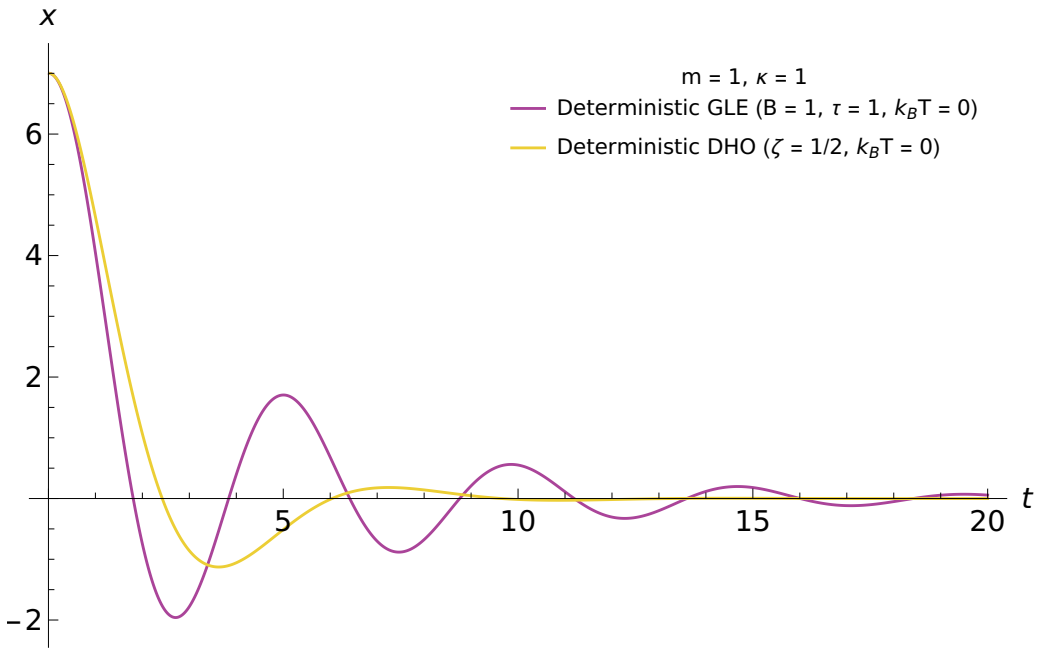

Figure 1: A trajectory of the general Langevin equation with an exponential memory kernel at low corrective force and a trajectory of an underdamped harmonic oscillator. The friction force in the harmonic oscillator was chosen so that the corrective force in both cases is equal for a constant velocity. Trajectories were calculated by discretizing the (generalized) Langevin equation with an Euler method and time steps $\delta t = 0.02$. The starting conditions were taken as $x(t < 0) = 7$.

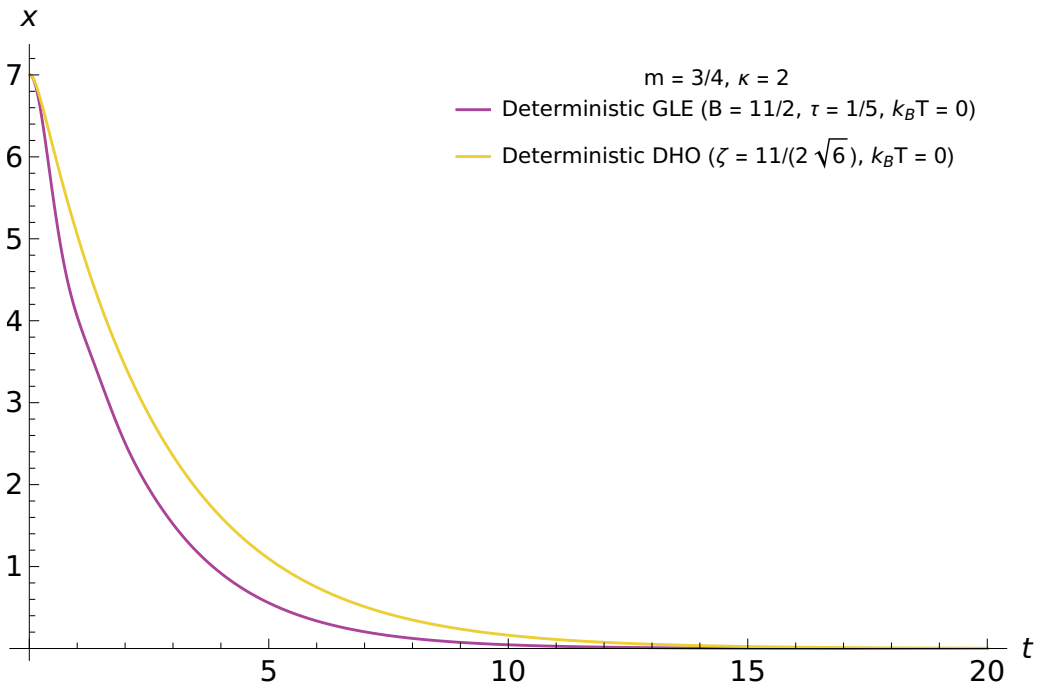

Figure 2: A trajectory of the general Langevin equation with an exponential memory kernel at high corrective force and short memory decay length, and a trajectory of an overdamped harmonic oscillator. The friction force in the harmonic oscillator was chosen so that the corrective force in both cases is equal for a constant velocity. Trajectories were calculated by discretizing the (generalized) Langevin equation with an Euler method and time steps $\delta t = 0.02$. The starting conditions were taken as $x(t < 0) = 7$.

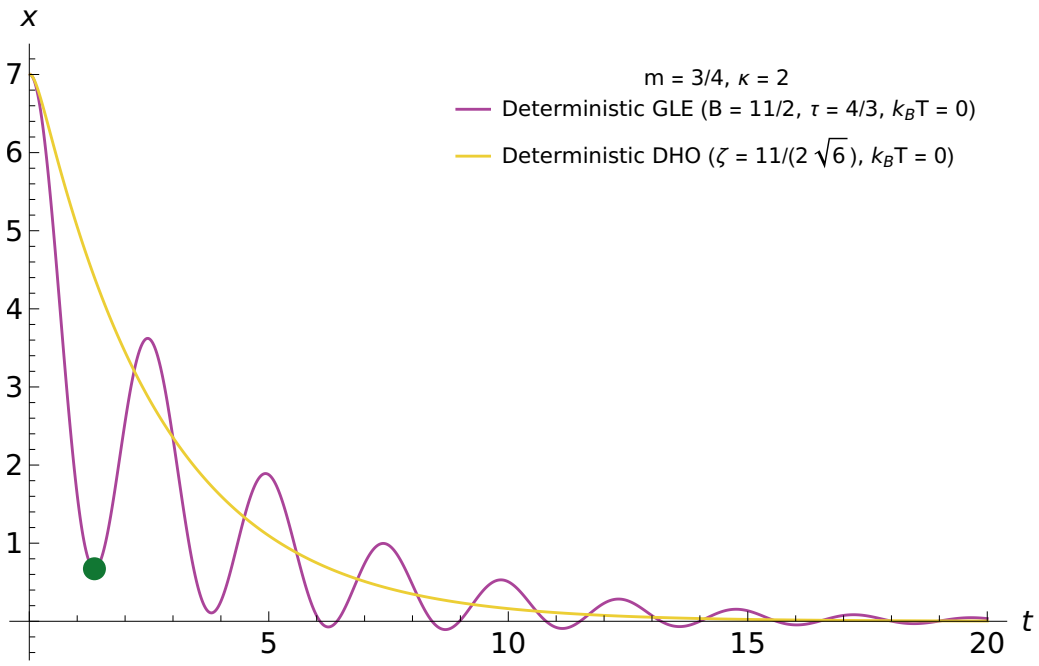

Figure 3: A trajectory of the general Langevin equation with an exponential memory kernel at high corrective force and long memory decay length, and a trajectory of an overdamped harmonic oscillator. The friction force in the harmonic oscillator was chosen so that the corrective force in both cases is equal for a constant velocity. Trajectories were calculated by discretizing the (generalized) Langevin equation with an Euler method and time steps $\delta t = 0.02$. The starting conditions were taken as $x(t < 0) = 7$.

(those prioritized by the kernel), the velocity has been negative. Thus the corrective force, which opposes velocity, will be positive even though the current velocity is zero. This effect is sometimes known as a delay oscillation.

The same effect occurs for the low B case (Figure 1). Here, the period of oscillations is lower, but the amplitude is larger than for the system with delta memory. We explain this because while the velocity is slowing (before the minima in the trajectory), the corrective force is larger. After all, it includes larger velocity contributions from the recent past. Thus it decelerates faster. However, at $v = 0$ and shortly thereafter, the acceleration is still positive, meaning that the amplitude of the next oscillation remains high.

To illustrate this, we calculate the force due to the potential and the corrective force separately so that we can compare the behavior of the velocity and corrective acceleration. For the deterministic trajectories from Figure 1 and Figure 3 this is visualized in, respectively, Figure 4 and Figure 5.

For the system with exponential memory, we see that the corrective acceleration is not in phase with the velocity but lags behind it, which is in contrast with the corrective acceleration for the SDHO system, which is always opposite to the velocity.

It is instructive to compare the location of the extrema in velocity. In Figure 4, the extremum is earlier in the period of the velocity than in Figure 5. This difference is because of the relative time it takes for velocity to go to zero in both cases. In figure 4, B is small, so the corrective force is small. Hence the system oscillated around zero. Here the period of the oscillation is approximately 5. The decay time of the memory kernel is 1. Since the memory kernel is smaller than half the period of the oscillation, the extremum in the corrective force will come shortly after the extremum in velocity. In figure 5, B is large, so the corrective force is larger. In this case, the system does not oscillate around zero but around a decaying exponential. The period of these oscillations is approximately 2.5. The decay time of the memory kernel is 4/3, larger than half the period of the oscillation. The contributions of the opposing sign part of the velocity have a larger effect when the decay time of the memory kernel is long compared to the period of the oscillations. Thus, the extremum in velocity will move further from the preceding velocity sign change as memory kernel decay time increases relative to the oscillation period.

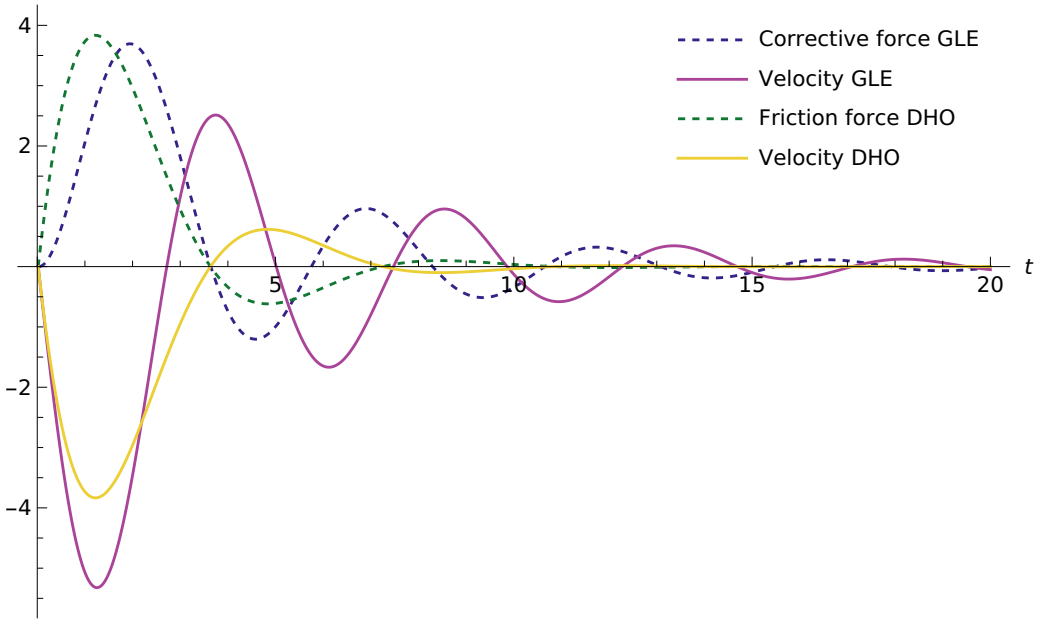

Figure 4: The velocity and corrective acceleration of the deterministic trajectories in Figure 1. We note that the corrective force and the velocity of the generalized Langevin equation are not in phase.

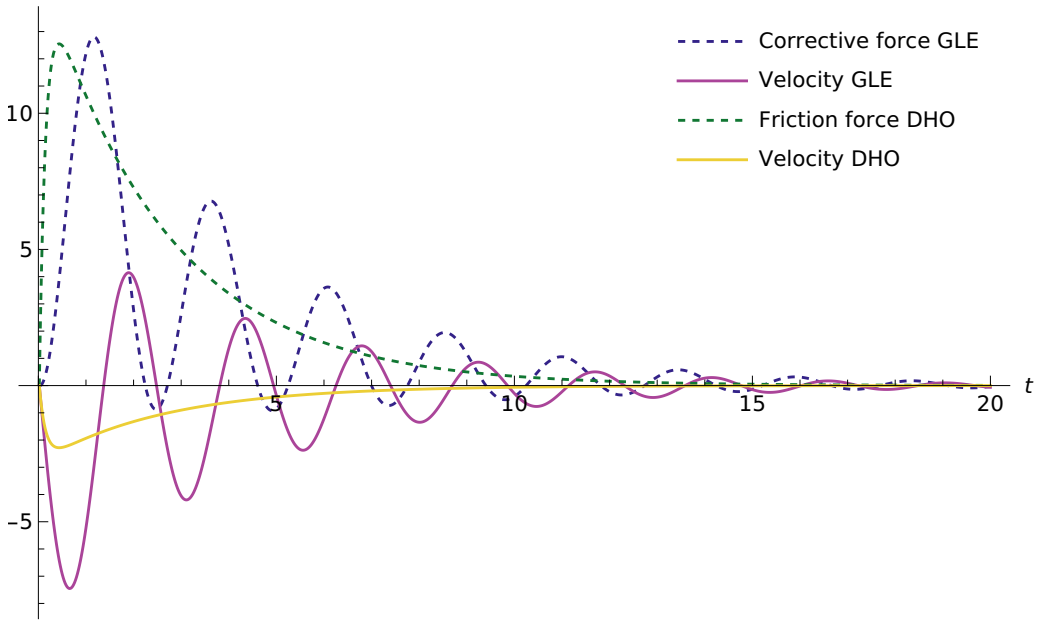

Figure 5: The velocity and corrective acceleration of the deterministic trajectories in Figure 3. We note that the corrective force and the velocity of the generalized Langevin equation are not in phase.

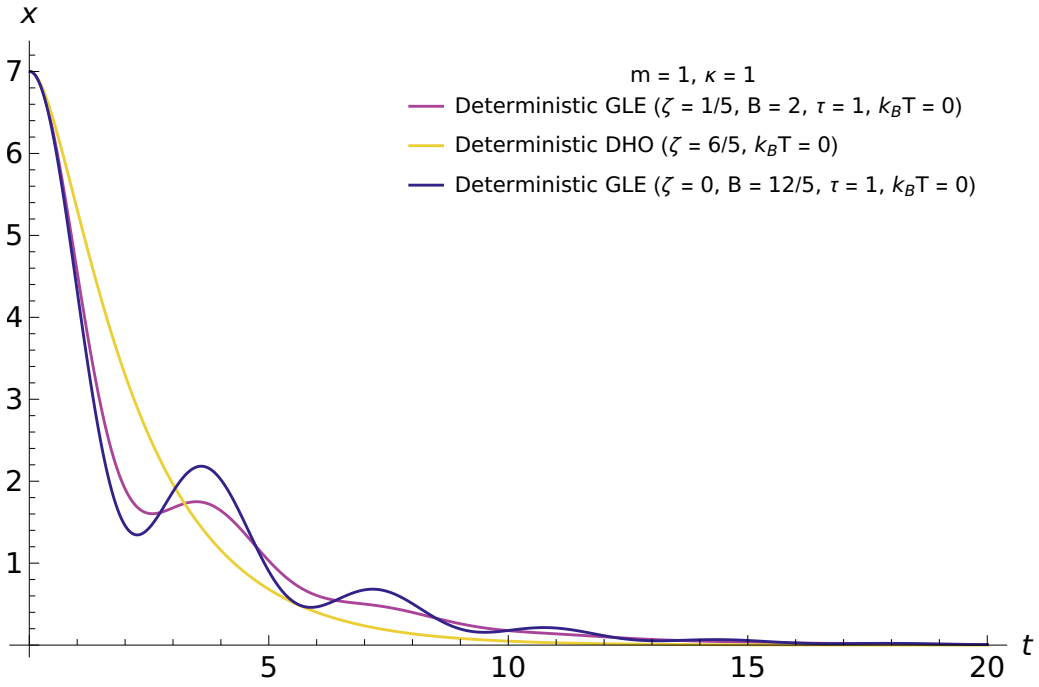

Figure 6: Comparison of trajectories of the general Langevin equation with a combined delta and exponential memory kernel to the general Langevin equation with exponential memory kernel and the SDHO. The corrective force magnitudes were chosen so that the corrective force in each case is equal at a constant velocity. Trajectories were calculated by discretizing the (generalized) Langevin equation with an Euler method and time steps $\delta t = 0.02$. The starting conditions were taken as $x(t < 0) = 7$.

2.3 Combined delta function and exponential

An alternative extension of the Langevin equation is realized by combining the delta function in the memory with an exponential decay, such that the relative importance of the current time point and the past in determining the corrective force can be straightforwardly adjusted. The history kernel is taken to be

$$\Gamma(t) = 4\zeta\sqrt{\kappa m}\,\delta(t) + \frac{B}{\tau}e^{-t/\tau}. \tag{6}$$

We use the Stratonovich convention to integrate a delta function at the boundary of the integral domain. The generalized Langevin equation is

$$m\ddot{x} = -2\zeta\sqrt{\kappa m}\dot{x}(t) - \frac{B}{\tau}\int_{t_m}^{t} dt' \, e^{-(t-t')/\tau}\dot{x}(t') - \kappa x(t) + \eta(t). \tag{7}$$

Tuning the ratio of $2\zeta\sqrt{\kappa m}$ with B tunes the importance between the instantaneous friction and the history-dependent corrective force. Some example trajectories are plotted in Figures 6 and 7, together with an equivalent SDHO trajectory and an equivalent generalized Langevin equation with exponential memory kernel trajectory. Again, the total corrective force for a particle at constant velocity is kept constant to compare the trajectories.

The features of these trajectories look similar to either the exponential memory kernel discussed before or the SDHO. For the exponential memory kernel, we argued that the period of oscillations was shorter and the amplitude higher than in the SDHO case. Here the period and amplitude are in between these two extremes. This tuning between the importance of both corrective forces allows for a distinctive feature that is present specifically with the combined memory kernel, shown in Figure 7. Here the first through of the oscillation is at $x < 0$, while subsequent ones are at $x > 0$.

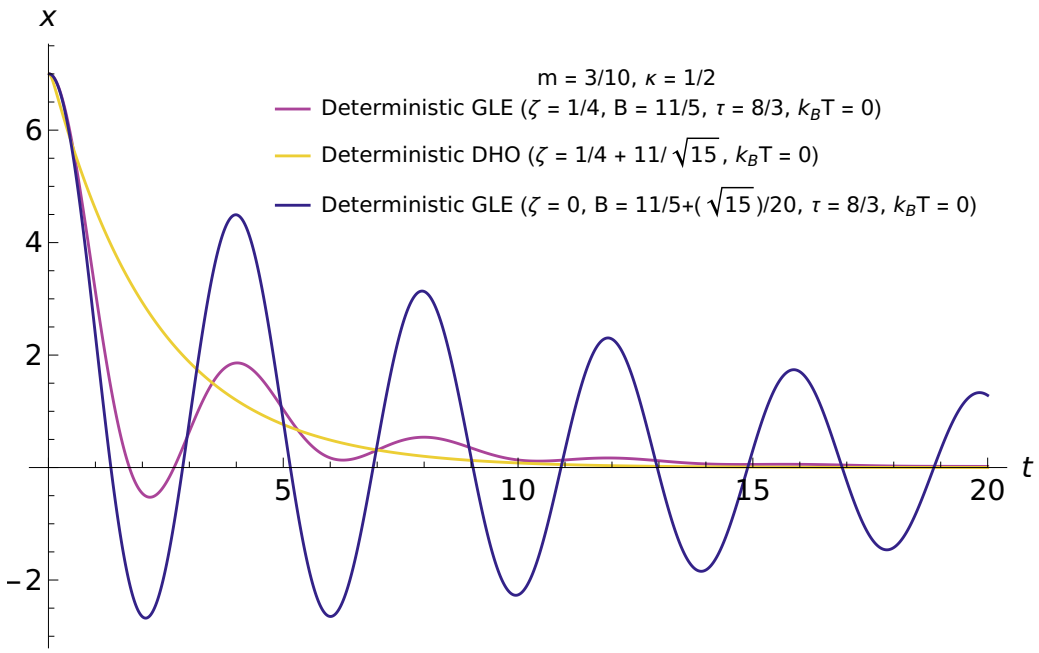

Figure 7: Comparison of trajectories of the general Langevin equation with a combined delta and exponential memory kernel to the general Langevin equation with exponential memory kernel and the SDHO. The corrective force magnitudes were chosen so that the corrective force in each case is equal at a constant velocity. Trajectories were calculated by discretizing the (generalized) Langevin equation with an Euler method and time steps $\delta t = 0.02$. The starting conditions were taken as $x(t < 0) = 7$.

3 Gaussian information bottleneck

As discussed in the introduction, this report aims to determine the optimal compression or representation of an input signal to determine the future value of that signal. We will use the Gaussian information bottleneck method to determine the kernel that achieves this optimal compression. The Gaussian information bottleneck is a framework in information theory developed by Tishby *et al.* [11, 22]. In this section, we will first present some needed background on information theory, then present the information bottleneck method, in particular the Gaussian information bottleneck method, and finally, develop how the framework is used in this work.

3.1 Information theory

Information theory provides useful tools to talk about information quantitatively. In particular, the extension of the concept of entropy to non-physical systems is a powerful tool in this context. The Shannon entropy of the probability distribution $p(x)$ of event X, which takes a discrete value $x \in \mathsf{X}$ is given as [10, 23]

$$H(X) = \sum_{x 2 \mathsf{X}} -p(x) \log(p(x)). \tag{8}$$

This entropy is a measure of the uncertainty in the outcome of X, such that abroad probability distribution $p(x)$ has high entropy. In contrast, a probability distribution that is sharply peaked has low entropy. In the limit where the outcome of X is completely determined, the entropy will be 0, while it will be positive in every other case.

For a set of two events X and Y the joint probability distribution $p(x, y)$ sets the probabilities of finding the outcomes $x \in \mathsf{X}$ and $y \in \mathsf{Y}$. The joint entropy is then defined as

$$H(X, Y) = \sum_{x 2 \mathsf{X}, y 2 \mathsf{Y}} -p(x, y) \log(p(x, y)). \tag{9}$$

If the two events are independent, such as when rolling two dice, the joint probability factorizes as $p(x, y) = p(x) \cdot p(y)$, such that the joint entropy can be rewritten as $H(X, Y) = H(X) + H(Y)$. On

the other hand, if the events are not independent, such as if two cards are drawn from a deck without replacement, the joint probability distribution does not factorize. In fact, since the card that is drawn first cannot be drawn again, the probability distribution is more peaked than if the events were independent, such that $H(X,Y) < H(X) + H(Y)$. In this report, we are interested in events that are not independent.

In that case, determining X yields information about Y, such that we can define a conditional probability $p(y|x)$, such that $p(x)p(y|x) = p(x,y)$. This is the probability that $Y = y$ given that $X = x$. The uncertainty that now is left is found using Equation (8), $H(Y|x) = \sum_{y2\mathsf{Y}} -p(y)\log(p(y|x))$. In general, we know that X contains information about Y, but we do not know the value of X. In this case we average $H(Y|x)$ over all possible x giving the conditional entropy:

$$H(Y|X) = \sum_{x2\mathsf{X}} -p(x)H(Y|x) = \sum_{x2\mathsf{X}} -p(x)\sum_{y2\mathsf{Y}} p(y|x)\log(p(y|x)) = \sum_{y2\mathsf{Y},x2\mathsf{X}} -p(x,y)\log(p(y|x)). \tag{10}$$

Finally, we define the mutual information,

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X), \tag{11}$$

which is a measure of how much information one variable contains about the other variable. As implicitly shown in the above definition, the mutual information is symmetric, i.e. X contains the same amount of information about Y as Y does about X.

In this work, we will consider Gaussian variables and thus want to determine the entropy, the joint and conditional entropies and the mutual information for Gaussian variables. However, Gaussian variables are continuous, so we need to consider an extension of entropy to continuous probability density functions. For a continuous variable X with probability density function $p(x)$ the differential entropy is defined [23]

$$h(X) = -\int_{1}^{1} dx\, p(x)\log(p(x)). \tag{12}$$

Where we take $0\log 0 = 0$, such that regions where $p(x) = 0$ will not influence the result. The definitions for conditional entropies, joint entropies, and mutual information extend similarly to continuous variables. We note that differential entropy is not a mathematically rigorous extension of the entropy of a discrete variable to a continuous variable [23]. However, it is similar enough in the context we are interested in [22]. The probability density function for a Gaussian a d dimensional Gaussian variable X with mean μ and covariance matrix Σ_x is given as [23]

$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^d\sqrt{|\Sigma_x|}} e^{\ \frac{1}{2}(\mathbf{x}\ \ \mu)^T{\Sigma_x}^{\ 1}(\mathbf{x}\ \ \mu)}, \tag{13}$$

where $|A|$ is the determinant and $A^{\ 1}$ the inverse of matrix A, and $\mathbf{v}^T$ is the transpose of vector $\mathbf{v}$. We find the differential entropy [23]

$$\begin{aligned}
h(X) &= -\int d\mathbf{x}\, p(\mathbf{x})\log\left(\frac{1}{\sqrt{2\pi}^d\sqrt{|\Sigma_x|}} e^{\ \frac{1}{2}(\mathbf{x}\ \ \mu)^T{\Sigma_x}^{\ 1}(\mathbf{x}\ \ \mu)}\right) \\
&= \frac{1}{2}\log\left(2^d\pi^d|\Sigma_x|\right)\int d\mathbf{x}\, p(\mathbf{x}) + \frac{1}{2}\int d\mathbf{x}\, p(\mathbf{x})(\mathbf{x}-\mu)^T{\Sigma_x}^{\ 1}(\mathbf{x}-\mu) \\
&= \frac{1}{2}\log\left(2^d\pi^d|\Sigma_x|\right) + \frac{1}{2}\sum_{i,j}({\Sigma_x}^{\ 1})_{i,j}\int d\mathbf{x}\, p(\mathbf{x})(x_i-\mu_i)(x_j-\mu_j) \\
&= \frac{1}{2}\log\left(2^d\pi^d|\Sigma_x|\right) + \frac{1}{2}\sum_{i,j}({\Sigma_x}^{\ 1})_{i,j}(\Sigma_x)_{j,i} \\
&= \frac{1}{2}\log\left(2^d\pi^d|\Sigma_x|\right) + \frac{d}{2} = \frac{1}{2}\log\left((2\pi e)^d|\Sigma_x|\right).
\end{aligned} \tag{14}$$

Here we use the definition of the covariance matrix, $(\Sigma_x)_{j,i} = (\Sigma_x)_{i,j} = \int d\mathbf{x}\, p(\mathbf{x})(x_i-\mu_i)(x_j-\mu_j)$, and we note that $\sum_j({\Sigma_x}^{\ 1})_{i,j}(\Sigma_x)_{j,i} = I_{i,i}$. Specifically, this shows that the differential entropy for a Gaussian variable is found directly from its covariance and number of dimensions.

3.2 Information bottleneck

The information bottleneck is used to find the optimal compression $\widetilde{X}$ of some input X while preserving as much information about the variable Y as possible. This is achieved by comparing the mutual information between $\widetilde{X}$ and X, $I(\widetilde{X};X)$, with the mutual information between $\widetilde{X}$ and Y, $I(\widetilde{X};Y)$ [11]. The optimal compression thus has two goals, maximizing $I(\widetilde{X};Y)$ while minimizing $I(\widetilde{X};X)$. In doing so, the compressed variable is as informative as possible about the task, while the input is maximally compressed.

In this work, the only task we consider is determining the signal value some time in the future $t_{fut} > 0$. As input we consider a Gaussian signal $x(t)$ discretized in N timesteps Δ, such that the input is a vector of signal values at times $t = -n \cdot \Delta$ where $n \in \{0, 1, 2, ..., N-1\}$. Explicitly, the input variable $\boldsymbol{s}_p$ is a vector that collects these signal values,

$$\boldsymbol{s}_p = \big(x(0), x(-\Delta), x(-2\Delta), ..., x(-(N-1)\Delta)\big)^T. \tag{15}$$

The task variable s_f is the signal value at $t = t_{fut}$, $s_f = x(t_{fut})$. Since the task is to predict exactly one value, the best compression strategy is calculating some scalar related to the best prediction by some bijective function. So that each value that the compression variable can take maps to a unique prediction. As such, we will be considering a scalar compression variable $\tilde{x}$, which is found as a linear transformation [22]

$$\tilde{x} = Ax + \xi, \tag{16}$$

where A is the transformation matrix and ξ is the noise introduced by the compression mechanism, which is a Gaussian variable with mean 0 and covariance σ_ξ. We note that the noise ξ is independent from both $\boldsymbol{s}_p$ and s_f, such that the correlations are zero, $\langle \xi \boldsymbol{s}_p \rangle = \langle \xi s_f \rangle = 0$. A compression is then found by specifying A and σ_ξ. In particular, the ideal compression is found by minimizing the functional [22]

$$\min_{A,\sigma_\xi} \mathcal{L}[A, \sigma_\xi] = I(\widetilde{x}; \boldsymbol{s}_p) - \beta I(\widetilde{x}; s_f), \tag{17}$$

where β is a parameter indicating the relative importance of compression compared to information retention. We find for $\beta \leq 0$ a trivial solution to Equation (17), which is $I(\widetilde{x}; \boldsymbol{s}_p) = I(\widetilde{x}; s_f) = 0$. In this case, the need for compression is greater than the need to retain information, so it is better not to store any information. On the other hand, when β is large, the retention of some information will be preferred.

Solving this minimization problem comes down to finding the information bottleneck kernel A and noise level σ_ξ that minimize $\mathcal{L}$. To do so, we by rewriting $\mathcal{L}$ in terms of (differential) entropies

$$\mathcal{L}[A, \sigma_\xi] = h(\widetilde{x}) - h(\widetilde{x}|\boldsymbol{s}_p) - \beta \left(h(\widetilde{x}) - h(\widetilde{x}|s_f)\right) = (1-\beta)h(\widetilde{x}) - h(\widetilde{x}|\boldsymbol{s}_p) + \beta h(\widetilde{x}|s_f). \tag{18}$$

As in Equation (14), the entropy of a Gaussian variable is found directly from its covariance and dimension. We note that the dimension for all three variables is the same, such that the problem of minimizing $\mathcal{L}$ requires us to find $\Sigma_{\tilde{x}}$, $\Sigma_{\tilde{x}|\boldsymbol{s}_p}$, and $\Sigma_{\tilde{x}|s_f}$.

We find for $\Sigma_{\tilde{x}}$, remembering the definition in Equation (16) and noting $\langle \xi \rangle = 0$,

$$\begin{aligned}
\Sigma_{\tilde{x}} &= \langle [\tilde{x} - \langle \tilde{x} \rangle][\tilde{x} - \langle \tilde{x} \rangle] \rangle = \big\langle [A\boldsymbol{s}_p + \xi - \langle A\boldsymbol{s}_p + \xi \rangle][A\boldsymbol{s}_p + \xi - \langle A\boldsymbol{s}_p + \xi \rangle]^T \big\rangle \\
&= \big\langle [A(\boldsymbol{s}_p - \langle \boldsymbol{s}_p \rangle) + \xi][(\boldsymbol{s}_p - \langle \boldsymbol{s}_p \rangle)^T A^T + \xi] \big\rangle \\
&= A\langle (\boldsymbol{s}_p - \langle \boldsymbol{s}_p \rangle)(\boldsymbol{s}_p - \langle \boldsymbol{s}_p \rangle)^T \rangle A^T + \langle \xi (\boldsymbol{s}_p - \langle \boldsymbol{s}_p \rangle)^T \rangle A^T + A\langle (\boldsymbol{s}_p - \langle \boldsymbol{s}_p \rangle)\xi \rangle + \langle \xi\xi \rangle \\
&= A\Sigma_{\boldsymbol{s}_p} A^T + 0 + 0 + \sigma_\xi = A\Sigma_{\boldsymbol{s}_p} A^T + \sigma_\xi.
\end{aligned} \tag{19}$$

For $\Sigma_{\tilde{x}|\boldsymbol{s}_p}$ we note that once $\boldsymbol{s}_p$ is determined the variance of $\tilde{x}$ is fully determined by the variance of ξ, such that $\Sigma_{\tilde{x}|\boldsymbol{s}_p} = \sigma_\xi$.

For $\Sigma_{\tilde{x}|s_f}$ we use the Schur complement formula [22] to write

$$\Sigma_{\tilde{x}|s_f} = \Sigma_{\tilde{x}} - \Sigma_{\tilde{x}s_f} \Sigma_{s_f}^{\ 1} \Sigma_{s_f\tilde{x}}.$$

The covariance $\Sigma_{s_f\tilde{x}} = \Sigma_{\tilde{x}s_f}^T$ is found as

$$\Sigma_{s_f\tilde{x}} = \langle s_f \tilde{x}^T \rangle = \langle s_f (A\boldsymbol{s}_p + \xi)^T \rangle = \langle s_f \boldsymbol{s}_p^T A^T \rangle + \langle s_f \xi^T \rangle = \langle s_f \boldsymbol{s}_p^T \rangle A^T = \Sigma_{s_f \boldsymbol{s}_p} A^T = \big(A\Sigma_{\boldsymbol{s}_p s_f}\big)^T.$$

We thus find, using the result from Equation (19) for $\Sigma_{\tilde{x}}$,

$$\Sigma_{\tilde{x}js_f} = A\Sigma_{\boldsymbol{s}_p}A^T + \sigma_\xi - A\Sigma_{\boldsymbol{s}_p s_f}\Sigma_{s_f}{}^1\Sigma_{s_f\boldsymbol{s}_p}A^T = A(\Sigma_{\boldsymbol{s}_p} - \Sigma_{\boldsymbol{s}_p s_f}\Sigma_{s_f}{}^1\Sigma_{s_f\boldsymbol{s}_p})A^T + \sigma_\xi = A\Sigma_{\boldsymbol{s}_p js_f}A^T + \sigma_\xi. \tag{20}$$

Here we use the Schur complement formula again in the last step, but backwards this time.

The minimization problem is now written as

$$\mathcal{L}[A, \sigma_\xi] = \frac{(1-\beta)}{2}\log(|A\Sigma_{\boldsymbol{s}_p}A^T + \sigma_\xi|) - \frac{1}{2}\log(|\sigma_\xi|) + \frac{\beta}{2}\log(|A\Sigma_{\boldsymbol{s}_p js_f}A^T + \sigma_\xi|).$$

The minima of $\mathcal{L}[A, \sigma_\xi]$ can only be found at those points where both of its first derivatives are 0. We first consider $\frac{\partial \mathcal{L}}{\partial A}$, and come back to σ_ξ later. Using the identity $\frac{\partial}{\partial A}\log(|ACA^T|) = 2(ACA^T)^{-1}AC$ [22] we differentiate $\mathcal{L}$ with respect to A

$$\frac{\partial \mathcal{L}}{\partial A} = \frac{1-\beta}{2}\frac{2A\Sigma_{\boldsymbol{s}_p}}{A\Sigma_{\boldsymbol{s}_p}A^T + \sigma_\xi} + \frac{\beta}{2}\frac{2A\Sigma_{\boldsymbol{s}_p js_f}}{A\Sigma_{\boldsymbol{s}_p js_f}A^T + \sigma_\xi}.$$

We find the extrema by equating the above to 0. After rearranging we find

$$\frac{\beta-1}{\beta}\frac{A\Sigma_{\boldsymbol{s}_p js_f}A^T + \sigma_\xi}{A\Sigma_{\boldsymbol{s}_p}A^T + \sigma_\xi}A = A\Sigma_{\boldsymbol{s}_p js_f}\Sigma_{\boldsymbol{s}_p}{}^1 = A\left(I_N - \Sigma_{\boldsymbol{s}_p s_f}\Sigma_{s_f}{}^1\Sigma_{s_f\boldsymbol{s}_p}\Sigma_{\boldsymbol{s}_p}{}^1\right). \tag{21}$$

This we recognize as an eigenvalue problem for the left eigenvectors of $\Sigma_{\boldsymbol{s}_p js_f}\Sigma_{\boldsymbol{s}_p}{}^1$, which can, as before, be rewritten using the Schur complement formula. We note that the eigenvalues of this matrix are all real and satisfy $0 \leq \lambda \leq 1$ [22]. By finding these eigenvectors, we can partially solve the minimization problem by determining the information bottleneck kernel A as function of the noise level σ_ξ.

We note that in eigenvalue problems the eigenvectors have a remaining degree of freedom in their norm. However, in this case the eigenvalue is written as $\lambda = \frac{\beta}{\beta}{}^1\frac{A\Sigma_{\boldsymbol{s}_p js_f}A^T + \sigma_\xi}{A\Sigma_{\boldsymbol{s}_p}A^T + \sigma_\xi}$, such that we might express the norm of A using the eigenvalue. We solve for the norm by splitting A into its normalised vector ν and its norm $\|A\|$, such that $A = \|A\|\nu$. We have

$$\begin{aligned}
\lambda &= \frac{\beta-1}{\beta}\frac{\|A\|\nu\Sigma_{\boldsymbol{s}_p js_f}\|A\|\nu^T + \sigma_\xi}{\|A\|\nu\Sigma_{\boldsymbol{s}_p}\|A\|\nu^T + \sigma_\xi} \\
\lambda\beta\left(\|A\|^2\nu\Sigma_{\boldsymbol{s}_p}\nu^T + \sigma_\xi\right) &= (\beta-1)\left(\|A\|^2\nu\Sigma_{\boldsymbol{s}_p js_f}\nu^T + \sigma_\xi\right) \\
\|A\|^2\left(\lambda\beta\nu\Sigma_{\boldsymbol{s}_p}\nu^T - (\beta-1)\nu\Sigma_{\boldsymbol{s}_p js_f}\nu^T\right) &= \sigma_\xi\left((\beta-1) - \lambda\beta\right) \\
\|A\| &= \sqrt{\frac{\sigma_\xi\left(\beta - \lambda\beta - 1\right)}{\lambda\beta\nu\Sigma_{\boldsymbol{s}_p}\nu^T - (\beta-1)\nu\Sigma_{\boldsymbol{s}_p js_f}\nu^T}}.
\end{aligned}$$

To simplify we note that $\nu\Sigma_{\boldsymbol{s}_p js_f}\nu^T = \nu\Sigma_{\boldsymbol{s}_p js_f}\Sigma_{\boldsymbol{s}_p}{}^1\Sigma_{\boldsymbol{s}_p}\nu^T = \nu\Sigma_{\boldsymbol{s}_p js_f}\Sigma_{\boldsymbol{s}_p}{}^1\nu^T\nu\Sigma_{\boldsymbol{s}_p}\nu^T = \lambda\nu\Sigma_{\boldsymbol{s}_p}\nu^T$, where we use that $\Sigma_{\boldsymbol{s}_p}{}^1\Sigma_{\boldsymbol{s}_p} = I_d$ and $\nu^T\nu = \|\nu\|^2 = 1$. Plugging this in, we find

$$\|A\| = \sqrt{\frac{\sigma_\xi\left(\beta - \lambda\beta - 1\right)}{\beta\lambda\nu\Sigma_{\boldsymbol{s}_p}\nu^T - \beta\lambda\nu\Sigma_{\boldsymbol{s}_p}\nu^T + \lambda\nu\Sigma_{\boldsymbol{s}_p}\nu^T}} = \sqrt{\frac{\sigma_\xi\left(\beta - \lambda\beta - 1\right)}{\lambda\nu\Sigma_{\boldsymbol{s}_p}\nu^T}}. \tag{22}$$

We note that this equation provides a constraint on β and λ, as the argument of the square root cannot be negative. It can be shown that $\nu\Sigma_{\boldsymbol{s}_p}\nu^T$ is positive [22] and the covariance of ξ is by definition positive. As such, the sign of the square root argument is fully decided by the rest of the numerator. Thus we find from $\beta(1-\lambda) - 1 \geq 0$ either $\lambda \leq \frac{\beta}{\beta}{}^1$ or $\beta \geq \frac{1}{1}{}^1\frac{1}{\lambda}$ [22]. This criterion can be used to determine which eigenvalue and eigenvector is the one that minimizes $\mathcal{L}$. However, this criterion might be satisfied by more than one eigenvalue for certain values of β. To decide which eigenvalue we should choose, we plug our result into the optimization problem of Equation (17) and determine a strategy to choose an appropriate eigenvalue from the result [22]. To do this, we first determine the mutual information between the compression $\widetilde{x}$ and the signal $\boldsymbol{s}_p$,

$$\begin{aligned}
I(\widetilde{x}; \boldsymbol{s}_p) &= h(\widetilde{x}) - h(\widetilde{x}|\boldsymbol{s}_p) = \frac{1}{2}\log\left((2\pi e)|A\Sigma_{\boldsymbol{s}_p}A^T + \sigma_\xi|\right) - \frac{1}{2}\log(2\pi e\sigma_\xi) \\
&= \frac{1}{2}\log\left|\|A\|^2\nu\Sigma_{\boldsymbol{s}_p}\nu^T + \sigma_\xi\right| - \frac{1}{2}\log(\sigma_\xi) = \frac{1}{2}\log\left|\frac{\sigma_\xi\left(\beta - \lambda\beta - 1\right)}{\lambda\nu\Sigma_{\boldsymbol{s}_p}\nu^T}\nu\Sigma_{\boldsymbol{s}_p}\nu^T + \sigma_\xi\right| - \frac{1}{2}\log(\sigma_\xi) \\
&= \frac{1}{2}\log\left|\frac{(\beta - \lambda\beta - 1) + \lambda}{\lambda}\right| = \frac{1}{2}\log\left|\frac{(\beta-1)(1-\lambda)}{\lambda}\right|,
\end{aligned} \tag{23}$$

and the mutual information between $\tilde{x}$ and s_f,

$$\begin{aligned}
I(\widetilde{x}; s_f) = h(\widetilde{x}) - h(\widetilde{x}|s_f) &= \frac{1}{2}\log\left(2\pi e|A\Sigma_{\boldsymbol{s}_p}A^T + \sigma_\xi|\right) - \frac{1}{2}\log(2\pi e|A\Sigma_{\boldsymbol{s}_p js_f}A^T + \sigma_\xi|) \\
&= \frac{1}{2}\log\left|\frac{\sigma_\xi\left(\beta-\lambda\beta-1\right)}{\lambda\nu\Sigma_{\boldsymbol{s}_p}\nu^T}\nu\Sigma_{\boldsymbol{s}_p}\nu^T + \sigma_\xi\right| - \frac{1}{2}\log\left|\frac{\sigma_\xi\left(\beta-\lambda\beta-1\right)}{\lambda\nu\Sigma_{\boldsymbol{s}_p}\nu^T}\nu\Sigma_{\boldsymbol{s}_p js_f}\nu^T + \sigma_\xi\right| \\
&= \frac{1}{2}\log\left|\frac{(\beta-\lambda\beta-1)}{\lambda}+1\right| - \frac{1}{2}\log\left|\frac{(\beta-\lambda\beta-1)}{\lambda\nu\Sigma_{\boldsymbol{s}_p}\nu^T}\lambda\nu\Sigma_{\boldsymbol{s}_p}\nu^T+1\right| \\
&= \frac{1}{2}\log\left|\frac{(\beta-1)(1-\lambda)}{\lambda}\right| - \frac{1}{2}\log|\beta-\lambda\beta| = \frac{1}{2}\log\left|\frac{(\beta-1)(1-\lambda)}{\lambda(\beta-\lambda\beta)}\right| \\
&= \frac{1}{2}\log\left|\frac{(\beta-1)(1-\lambda)}{\lambda\beta(1-\lambda)}\right| = \frac{1}{2}\log\left|\frac{\beta-1}{\lambda\beta}\right|. \qquad (24)
\end{aligned}$$

By plugging these results into Equation (17), we have

$$\begin{aligned}
\mathcal{L} = I(\widetilde{X};X) - \beta I(\widetilde{X};Y) &= \frac{1}{2}\log\left|\frac{(\beta-1)(1-\lambda)}{\lambda}\right| - \beta\left(\frac{1}{2}\log\left|\frac{(\beta-1)(1-\lambda)}{\lambda}\right| - \frac{1}{2}\log|\beta-\lambda\beta|\right) \\
&= \frac{1-\beta}{2}\log\left|\frac{(\beta-1)(1-\lambda)}{\lambda}\right| + \frac{\beta}{2}\log|\beta-\lambda\beta|.
\end{aligned}$$

We consider the derivative $\frac{d}{d\lambda}\mathcal{L}$ on the domain $\beta \geq \frac{1}{1-\lambda}$. We find

$$\begin{aligned}
\frac{d}{d\lambda}\mathcal{L} &= \frac{d}{d\lambda}\left(\frac{1-\beta}{2}\log\left|\frac{(\beta-1)(1-\lambda)}{\lambda}\right| + \frac{\beta}{2}\log|\beta-\lambda\beta|\right) \\
&= \frac{1-\beta}{2}\frac{\lambda}{1-\lambda}\frac{d}{d\lambda}\left(\frac{1}{\lambda}-1\right) + \frac{\beta}{2}\frac{1}{\beta-\lambda\beta}(-\beta) \\
&= -\frac{1-\beta}{2}\frac{\lambda}{1-\lambda}\frac{1}{\lambda^2} - \frac{\beta}{2}\frac{1}{1-\lambda} = \frac{1}{2}\left(\frac{1-\beta}{\lambda(\lambda-1)} + \frac{\beta\lambda}{\lambda(\lambda-1)}\right) \\
&= \frac{1+\beta(\lambda-1)}{2\lambda(\lambda-1)} \qquad (25) \\
&\geq \frac{1+\frac{1}{1-\lambda}(\lambda-1)}{2\lambda(\lambda-1)} = \frac{1-1}{2\lambda(\lambda-1)} = 0. \qquad (26)
\end{aligned}$$

As this derivative is non-negative in this domain, the function $\mathcal{L}$ becomes bigger for bigger λ. The minimal value for $\mathcal{L}$ is thus found for the smallest eigenvalue.

We also note that the variance of the noise, σ_ξ, disappears from $\mathcal{L}$ once the transformation matrix A is determined. Since the variance of the noise does also not influence the relative weights in the normalized transformation matrix ν, as this is one of the left eigenvectors of the matrix $I_N - \Sigma_{\boldsymbol{s}_p s_f}\Sigma_{s_f}{}^1\Sigma_{s_f \boldsymbol{s}_p}\Sigma_{\boldsymbol{s}_p}{}^1$. This shows that the variance of the noise is not specified by the information bottleneck method, such that this is a free variable that needs to be specified as part of the system under consideration. The choice of variance then sets the scale for the problem. For this work we will choose $\sigma_\xi = 1$.

3.3 This work

In conclusion, we find the optimal mapping of our incoming signal $\boldsymbol{s}_p$, which is a vector with the N most recent discretized signal values, onto a scalar $\tilde{x} = A\ \boldsymbol{s}_p + \xi$, given that we want to predict the future signal value s_f. To find this information bottleneck kernel A we need to determine the left eigenvectors of $\Sigma_{\boldsymbol{s}_p js_f}\Sigma_{\boldsymbol{s}_p}{}^1 = I_N - \Sigma_{\boldsymbol{s}_p s_f}\Sigma_{s_f}{}^1\Sigma_{s_f \boldsymbol{s}_p}\Sigma_{\boldsymbol{s}_p}{}^1$. If, for a given value of the trade-off parameter β, the smallest eigenvalue λ satisfies $\lambda \leq \frac{\beta-1}{\beta}$, the kernel A should be $A = \sqrt{\frac{\beta-\lambda\beta-1}{\lambda\nu\Sigma_{\boldsymbol{s}_p}\nu^T}}\nu$, where ν is the normalised eigenvector. In this case, the information retention has a large enough weight to be incorporated. On the other hand, if the smallest eigenvalue is bigger than $\frac{\beta-1}{\beta}$, compression is much more important than information retention, and the trivial solution $A = 0$ should be used. In that case, one can easily see that all the information is mapped into the same point (or more accurately same region, as the white noise process ξ will make the compression spread out a little bit around $\tilde{x} = 0$).

In the following, we present a general method of solving the generalized Langevin equation, which allows us to determine the autocovariance of the signal once the memory function is specified. Subsequently, we

look at the different memory functions introduced in the previous section and the resulting information bottleneck kernels these memory functions induce.

4 Continuous Signals

To determine the information bottleneck kernels for the memory kernels presented in Section 2, we thus need to find the autocorrelation function of the generalized Langevin equation. The formal solution to the generalized Langevin equation has been known for some time [16, 24]. The solution is found in Laplace space and subsequently transformed back into real space. First, we present the general form of the autocorrelation function, and subsequently, we find the particular solutions for the presented kernels.

4.1 General solution

The full solution for the generalized Langevin equation without specifying a memory function, following a paper from Di Terlizzi *et al.* in 2020 [16], is presented in appendix B. Here we present the main results needed to find the correlation functions, which we need to solve the information bottleneck.

The Laplace transform of a function $f(t)$ is found as

$$L_s\{f(t)\} = \int_0^\infty dt\ e^{-st} f(t) = \widehat{F}(s), \tag{27}$$

while the inverse Laplace transform of a function $\widehat{F}(s)$ is found either by inspection or by the complex integral

$$L_t^{-1}\{F(s)\} = \frac{1}{2\pi i}\int_{\alpha-i\infty}^{\alpha+i\infty} ds\ e^{st}\widehat{F}(s), \tag{28}$$

where α is chosen such that all the singularities of $\widehat{F}(s)$ are to the left of the contour.

To find the formal solution for a generalized Langevin equation,

$$m\ddot{x} = -\int_{t_m}^{t} dt^0\ \Gamma(t-t^0)\dot{x}(t^0) - \kappa x(t) + \eta(t), \tag{29}$$

where the noise η obeys the fluctuation-dissipation theorem $\langle \eta(t)\eta(t^0)\rangle = k_B T\Gamma(|t-t^0|)$, we determine the Laplace transform of both sides. This allows us to define the Laplace transform of a quantity called the position susceptibility $\chi_x(t)$,

$$\widehat{\chi}_x(s) = \frac{1}{ms^2 + \widehat{\Gamma}(s)s + \kappa}, \tag{30}$$

where $\widehat{\Gamma}(s)$ is the Laplace transform of the memory kernel in the generalized Langevin equation. Determining the inverse Laplace transform of this position susceptibility will turn out to be the most important step in finding correlation functions for the signals under consideration. In fact, the formal solution for $x(t)$ of Equation (29) is found in terms of the position susceptibility and an anti-derivative of it,

$$\chi(t) = \int_0^t \chi_x(\tau)d\tau \qquad \text{or, equivalently in Laplace space} \qquad \widehat{\chi}(s) = \frac{\widehat{\chi}_x(s)}{s}. \tag{31}$$

The formal solution is

$$x(t) = x(t_m)\left(1 - \kappa\chi(t-t_m)\right) + m\dot{x}(t_m)\chi_x(t-t_m) + \int_{t_m}^{t} d\tau\ \chi_x(t-\tau)\eta(\tau). \tag{32}$$

To find the autocorrelation function, we determine $\langle x(t)x(t^0)\rangle$. This expression contains nine different correlations. Specifically, we have

$$\langle x(t)x(t^0)\rangle = \langle x(t_m)x(t_m)\rangle\,(1-\kappa\chi(t-t_m))\,(1-\kappa\chi(t^0-t_m)) \tag{33}$$
$$+ m^2\langle \dot{x}(t_m)\dot{x}(t_m)\rangle\chi_x(t-t_m)\chi_x(t^0-t_m) \tag{34}$$
$$+ m\langle x(t_m)\dot{x}(t_m)\rangle\big(\chi_x(t^0-t_m)\,[1-\kappa\chi(t-t_m)] + \chi_x(t-t_m)\,[1-\kappa\chi(t^0-t_m)]\,\big) \tag{35}$$
$$+ \int_{t_m}^{t} d\tau\ \chi_x(t-\tau)\big(\,[1-\kappa\chi(t^0-t_m)]\,\langle x(t_m)\eta(\tau)\rangle + m\chi_x(t^0-t_m)\langle \dot{x}(t_m)\eta(\tau)\rangle\big) \tag{36}$$
$$+ \int_{t_m}^{t^0} d\tau^0\ \chi_x(t^0-\tau^0)\big(\,[1-\kappa\chi(t-t_m)]\,\langle x(t_m)\eta(\tau^0)\rangle + m\chi_x(t-t_m)\langle \dot{x}(t_m)\eta(\tau^0)\rangle\big) \tag{37}$$
$$+ \int_{t_m}^{t} d\tau \int_{t_m}^{t^0} d\tau^0\ \chi_x(t-\tau)\chi_x(t^0-\tau^0)\langle \eta(\tau)\eta(\tau^0)\rangle. \tag{38}$$

The four terms on lines (36) and (37) contain the correlation between the noise at times $t > t_m$ and the initial conditions. These are not correlated, so the integrals evaluate to 0. The four terms on lines (33), (34), and (35) contain the initial condition correlations and do not simplify further without assuming anything about these initial conditions. The final term, the double integral over the correlations in the noise, is calculated using the fluctuation-dissipation theorem, $\langle \eta(t)\eta(t^0)\rangle = k_BT\Gamma(|t-t^0|)$. This calculation is shown in Section B.3. The resulting autocorrelation function is

$$\begin{aligned}\langle x(t)x(t^0)\rangle =& \langle x(t_m)x(t_m)\rangle\,(1-\kappa\chi(t-t_m))\,(1-\kappa\chi(t^0-t_m)) + m^2\langle \dot{x}(t_m)\dot{x}(t_m)\rangle\chi_x(t-t_m)\chi_x(t^0-t_m)\\ &+ m\langle x(t_m)\dot{x}(t_m)\rangle\big(\chi_x(t^0-t_m)\,[1-\kappa\chi(t-t_m)] + \chi_x(t-t_m)\,[1-\kappa\chi(t^0-t_m)]\,\big)\\ &+ k_BT\big(\chi(t-t_m)+\chi(t^0-t_m)-\chi(|t-t^0|)-\kappa\chi(t-t_m)\chi(t^0-t_m)-m\chi_x(t-t_m)\chi_x(t^0-t_m)\big).\end{aligned} \tag{39}$$

Since we are interested in the equilibrium properties of the signals, we take the limit $t_m \to -\infty$, such that we find for the position

$$x(t) = \int_{-\infty}^{t} d\tau\ \chi_x(t-\tau)\eta(\tau), \tag{40}$$

and for the autocorrelation

$$\langle x(t)x(t^0)\rangle = k_BT\left(\frac{1}{\kappa} - \chi\left(|t-t^0|\right)\right). \tag{41}$$

Given this, we will determine the position susceptibility and, subsequently, the autocorrelation function for the generalized Langevin equations introduced in Section 2. Each time we will compare the results to the stochastic harmonic oscillator results from Reference [9].

4.2 Exponential kernel

4.2.1 Autocorrelation function

Let us first take a look at the exponential memory kernel. The Laplace transform of the memory kernel is

$$\Gamma(s) = L_s\left\{\Gamma(t)\right\} = \frac{B}{\tau}\int_0^{\infty} dt\ e^{-st}e^{-t/\tau} = \frac{B}{\tau}\left[\frac{e^{-t(s+1/\tau)}}{-(s+1/\tau)}\right]_0^{\infty} = \frac{B}{\tau s+1}. \tag{42}$$

We thus find the Laplace transform of the position susceptibility,

$$\widehat{\chi}_x(s) = \frac{1}{ms^2 + \frac{B}{\tau s+1}s + \kappa} = \frac{\tau s+1}{m\tau s^3 + ms^2 + (B+\kappa\tau)s + \kappa}. \tag{43}$$

To determine the position susceptibility in real space, we must evaluate the complex integral of Equation (28). This is done using the Cauchy residue theorem, where we note that extending the contour in

infinity around the left half plane adds 0 to the integral. Thus, evaluating the complex integral is equivalent to determining the poles of the Laplace transform of the position susceptibility and summing their residues. The poles of the Laplace transform of the position susceptibility are located at the points where the denominator becomes 0, such that finding them is equivalent to solving the third order equation $m\tau s^3 + ms^2 + (B + \kappa\tau)s + \kappa = 0$. If the equation has three non-equal solutions c_1, c_2, c_3 the position susceptibility would be

$$\chi_x(t) = \sum_{n=1}^{3} \frac{(\tau c_n + 1)e^{tc_n}}{3m\tau c_n^2 + 2mc_n + B + \kappa\tau}. \tag{44}$$

If two or three of the solutions are equal, the residues cannot be expressed in the same manner. In that case, the position susceptibility can be expressed as either

$$\chi_x(t) = A_1 e^{tc_1} + e^{tc_2}(A_2 + A_3 t), \qquad \text{or} \qquad \chi_x(t) = e^{tc_1}(A_1 + A_2 t + A_3 t^2), \tag{45}$$

where A_1, A_2, and A_3 are constants that are found by calculating the residues at the solution(s).

Thus, we note that it is not possible to get an exact symbolic solution for $\chi_x(t)$ in terms of m, κ, B and τ. Instead, we choose a, c_1, c_2 and c_3, which correspond to exact values of m, κ, B and τ. These are found by setting $m\tau s^3 + ms^2 + (B + \kappa\tau)s + \kappa = a(s + c_1)(s + c_2)(s + c_3)$ giving

$$\begin{aligned} m &= a(c_1 + c_2 + c_3) \\ \kappa &= a c_1 c_2 c_3 \\ B &= a\frac{(c_1 + c_2)(c_1 + c_3)(c_2 + c_3)}{c_1 + c_2 + c_3} \\ \tau &= \frac{1}{c_1 + c_2 + c_3}. \end{aligned}$$

We always choose $a = \frac{1}{c_1 c_2 c_3}$, such that $\kappa = 1$ and the autocorrelation in Equation 41 at $t = t^0$ equals $k_B T$. This makes sure that all the autocorrelations have the same scaling. Using these identifications, we can write the position susceptibilities using either Equation (44) or one of the options in Equation (45). We then determine the autocorrelation function of the signal using Equations (31) and (41).

So, while the position susceptibility and autocorrelation function can be analytically calculated for every choice of parameters, we will not be able to find a common symbolic expression for it. This is, for our purposes, not a great loss, as in the process of calculating the information bottleneck a matrix inversion needs to be performed that cannot easily be performed symbolically either. As such, let us continue with the knowledge that we have a recipe for calculating the correlation function but cannot write it down symbolically.

4.2.2 Information bottleneck

Now that we have the statistics for our signal, we can use this to determine the important information in the signal using the information bottleneck method. For that, we need to discretize the signal. We define the past signal as the column vector

$$\boldsymbol{s}_p = (x(0), x(-\Delta), x(-2\Delta), ..., x(-(N-1)\Delta))^T, \tag{46}$$

where Δ is the discretization timestep and $(N-1)\Delta$ the length of the signal. The future signal value at time t_{fut}, the object which we seek to predict, is

$$s_f = x(t_{fut}). \tag{47}$$

We analytically calculate the autocorrelation function at the requisite time lags and subsequently construct the matrix $I_d - \Sigma_{\boldsymbol{s}_p s_f}\Sigma_{s_f}{}^{1}\Sigma_{s_f \boldsymbol{s}_p}\Sigma_{\boldsymbol{s}_p}{}^{1}$. The left eigenvectors and corresponding eigenvalues are found as the eigensystem of the transpose of said matrix. The normalized eigenvector corresponding to the smallest eigenvalue is the information bottleneck kernel indicating the relative importance of the corresponding past signal values in predicting the future signal value.

We plot the normalized information bottleneck kernels for several combinations of parameters in Figure 8. We see that the kernels have many non-zero components and that the kernels for different signals are

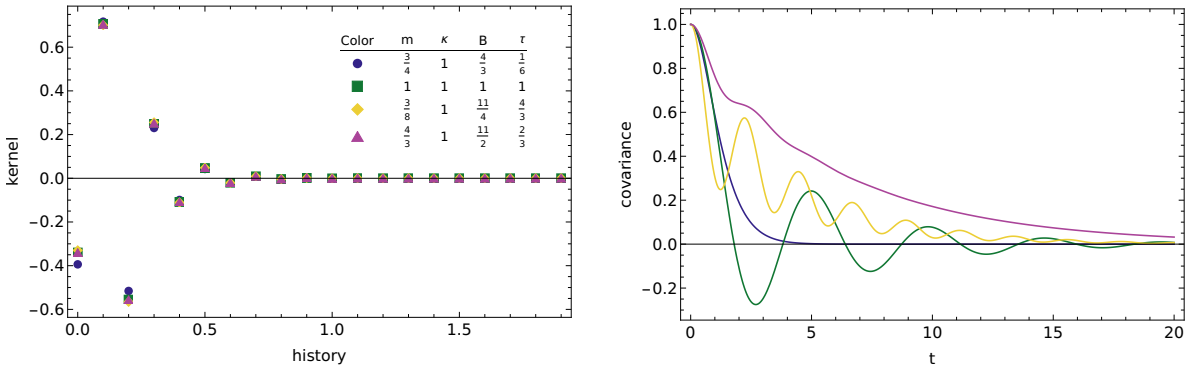

Figure 8: Normalized information bottleneck kernels and correlation functions for several different choices of parameters for the generalized Langevin equation with exponential memory. The information bottleneck kernels were calculated for discretization timesteps of $\Delta = 0.1$, with a signal length $(N-1)\Delta = 1.9$, and to predict the signal value a time $t_{fut} = 1/2$ into the future.

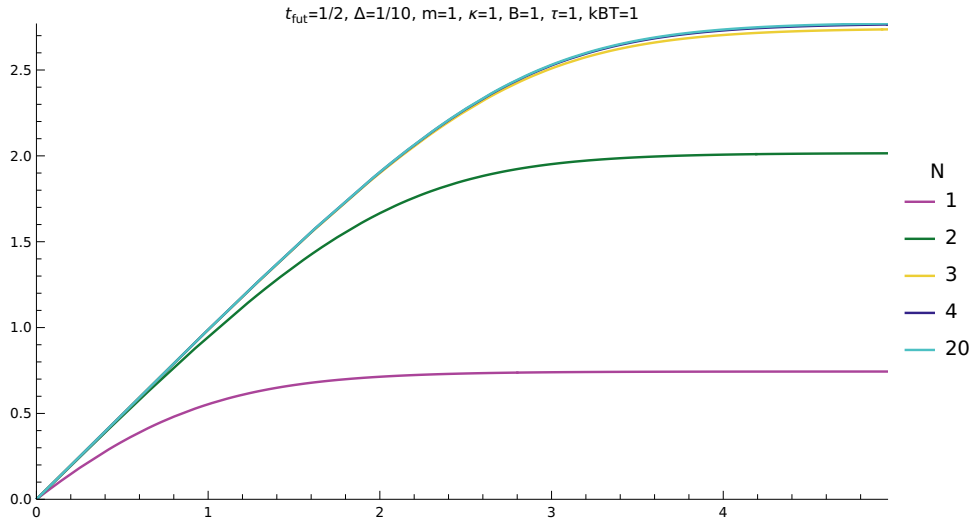

Figure 9: Future information as a function of past information for the signal with the same parameters as the green curve in Figure 8 for several values of the vector length.

similar, even for distinctly different covariance functions. These observations are similar to the kernels found for the continuous signal for the Langevin equation in Reference [9].

In Reference [9], the authors note that, in the SDHO case, the information bottleneck kernel contains artifacts when discretizing a continuous signal. This might also be the case in our system. To check this, we plot the future information as a function of the past information for different lengths of the vector $\vec{A}$. This is shown in figure 9. We see that the curves for a vector length of 3 and a vector length of 20 almost overlap and that the curve for a vector length of 4 is indistinguishable from the curve for a vector length of 20. This suggests that the most recent 4 components contain by far the most relevant information. In the SDHO case, a similar plot shows that a vector of just two entries contains almost all the information [9]. In that case, it seems that the kernel needs to approximate the instantaneous velocity at the current time, and the best approximation of this can be found by including more than just the two most recent values of the discretized signal. We believe that the clearest way to separate these two effects is to have a truly discrete version of the signal, which we will discuss in Section 5.

4.3 Combined delta and exponential kernel

4.3.1 Autocorrelation function

For the combined memory kernel, the Laplace transform is, again using the Stratonovich convention for the integral of the delta function at the domain boundary,

$$\begin{aligned}\Gamma(s) = L_s\left\{\Gamma(t)\right\} &= \int_0^1 dt\ e^{-st}\left(4\zeta\sqrt{\kappa m}\,\delta(t) + \frac{B}{\tau}e^{-t/\tau}\right) = 4\zeta\sqrt{\kappa m}\int_0^1 dt\ e^{-st}\delta(t) + \frac{B}{\tau}\int_0^1 dt\ e^{-st}e^{-t/\tau} \\ &= 2\zeta\sqrt{\kappa m} + \frac{B}{\tau s+1}. \end{aligned} \tag{48}$$

We thus find the Laplace transform of the position susceptibility,

$$\widehat{\chi}_x(s) = \frac{1}{ms^2 + \left(2\zeta\sqrt{\kappa m} + \frac{B}{\tau s+1}\right)s + \kappa} = \frac{\tau s+1}{m\tau s^3 + (m + 2\tau\zeta\sqrt{\kappa m})\,s^2 + (B + \kappa\tau + 2\zeta\sqrt{\kappa m})s + \kappa}. \tag{49}$$

This leads us to a similar place as for the exponential memory kernel in the previous section. To find the autocorrelation function for the signal, we have to determine the complex integral from Equation (28), for which we can use the Cauchy integration theorem to change it into a sum over the poles of the position susceptibility. To find these, we have to solve a cubic equation. Depending on the number of distinct poles, this yields different forms for the residues at those poles, so finding a closed form of the autocorrelation function is impossible. In the case of the exponential memory kernel, we argued that choosing a scaling factor and three solutions to the cubic equation (i.e. fixing 4 degrees of freedom) was able to fix all the parameters. As such, the scaling factor and solutions fully determine the correlation function's behavior. In this case, however, the introduction of the additional parameter ζ means that we cannot fix all the parameters based on the scaling factor and solutions of the cubic equation alone. Indeed, we fix the overall magnitude of the correlation function, which fixes κ. However, we can now still vary the importance of the two corrective forces for a fixed set of solutions to the cubic equation.

4.3.2 Information bottleneck

We determine the information bottleneck kernel using the same discretization and calculation methods from the exponential memory kernel case. In figure 10, normalized information bottleneck kernels and corresponding autocovariance functions are shown. Similarly, as in the previous case, with a purely exponential memory function, we see that the information bottleneck kernels for different autocovariance functions have similar shapes. Moreover, also similar to the exponential memory kernel, many entries are close to 0 but not actually 0.

To get some intuition about how many entries contain information in this case, let us once again plot the future information as a function of the past information for different lengths of the vector $\vec{A}$ in Figure 11. In this case, we see a distinct difference between the curves for $N = 4$ and $N = 20$, while the curve for $N = 3$ is almost indistinguishable from the former. This suggests that information about more than 4 distinct measurements is needed. However, this might again be influenced by the discretization we perform to go from the continuous signal to the discrete information bottleneck. So, let us now turn to

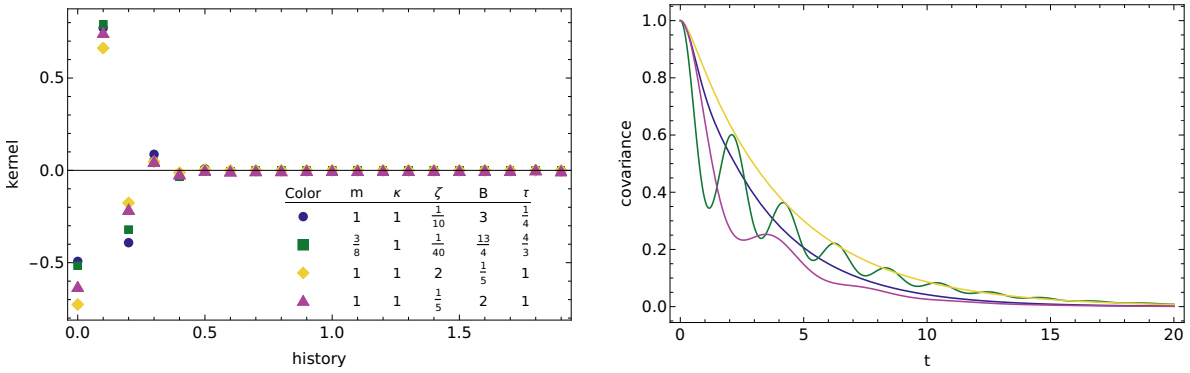

Figure 10: Normalized information bottleneck kernels and correlation functions for several different choices of parameters for the generalized Langevin equation with combined delta function and exponential memory. The information bottleneck kernels were calculated for discretization timesteps of $\Delta = 0.1$, with a signal length $(N-1)\Delta = 1.9$, and to predict the signal value a time $t_{fut} = 1/2$ into the future.

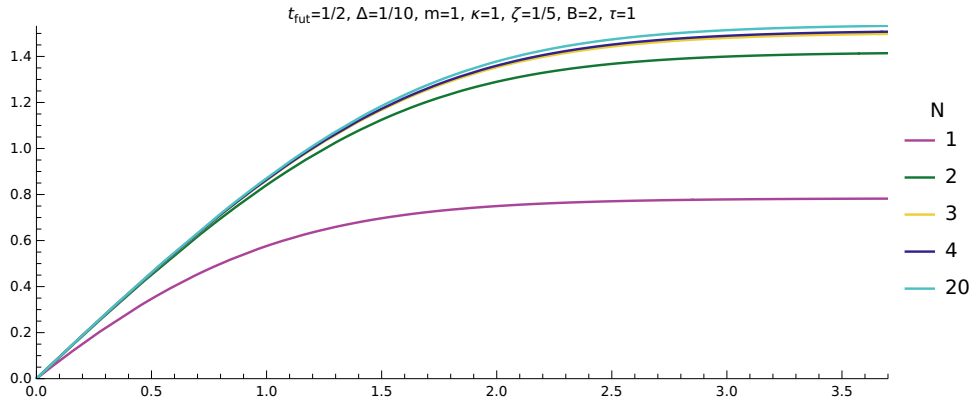

Figure 11: Future information as a function of past information for the signal with the same parameters as the pink curve in Figure 10 for several values of the vector length.

a discrete model such that this discretization-induced error cannot be the reason for the extension of the information bottleneck kernel.

5 Autoregressive models

Since, in the continuous models that we discussed, the discretization of the signal might introduce errors in the kernel, let us consider a method to create a discrete signal from the start. Autoregressive models are discrete-time models that determine the new output value as a linear combination of past outputs and a stochastic term. These models are useful when working with a discrete version of the information bottleneck since the output of these models is discrete. As such, we expect that the model will be able to show whether certain patterns in the (information bottleneck) kernel are discretization errors or actual effects based on the type of signal. In previous work on the information bottleneck, for signals generated by the SDHO, discretization effects fully disappeared when transitioning to an autoregressive signal [9]. We will first set out some basics on time series in general and autoregressive models in particular.

5.1 Time series and autoregressive models

A time series $y_t, t \in \mathsf{Z}$ is a series of numbers describing something at discrete times t [25, 26]. Examples of time series include the daily price of a plate of food in the canteen or the daily temperature at noon in Amsterdam. For certain applications, it is useful to consider multiple time series simultaneously, such that we could think of a time series that includes the daily temperature, humidity, wind direction and wind speed at noon in Amsterdam. The number of combined variables d is the dimensionality of the time series. Autoregressive models are a specific type of time series in which the future value can be expressed based on the past values, some constant and a noise term. In particular, a (vector) autoregressive model of order p, (V)AR(p), depends on the p past output values of the model, such that we can write the current state of the model, y_n, as a linear combination of the past states of the model $y_{n\ \ 1}, y_{n\ \ 2}, ...y_{n\ \ p}$. We write [26]

$$y_n = \phi_0 + \sum_{j=1}^{p} A_j y_{n\ \ j} + \epsilon_{n\ \ 1}. \tag{50}$$

Here y_n is a d-dimensional vector which includes one entry for each of the included variables, the linear transformation matrices $A_1, A_2, ...A_p$ are $d \times d$ matrices, and $\epsilon_{n\ \ 1}$ is a d-dimensional white noise process, which might be called an innovation process in literature [26], such that $\langle \epsilon_n \rangle = 0$ and $\langle \epsilon_{n1} \epsilon_{n2}^T \rangle = \Sigma_\epsilon \delta_{n1,n2}$. We note that Σ_ϵ here is the d by d same-time covariance matrix of the noises, not some correlation across different time points, for which we used the symbol Σ in section 3. In our case, we are only interested in a process with mean value 0, which is achieved by equating $\phi_0 = 0$. By repeatedly evaluating Equation (50) we find that y_n is uniquely determined by $y_0, y_1, ..., y_p, \epsilon_0, \epsilon_1, ..., \epsilon_{n\ \ 1}$, such that an order p autoregressive model needs p initial conditions to determine the dynamics.

To determine the autocovariance of vector autoregressive models we need an equivalent vector autoregressive model of order 1, as these models are more convenient to work with. As such, we define a $d \cdot p$-dimensional VAR(1) model that is equivalent to the d-dimensional VAR(p) model from Equation (50) [26]

$$Y_n = \mathbf{A} Y_{n\ \ 1} + E_{n\ \ 1}, \tag{51}$$

where we define

$$Y_n = \begin{pmatrix} y_n \\ y_{n\ \ 1} \\ \vdots \\ y_{n\ \ p+1} \end{pmatrix}, \qquad E_{n\ \ 1} = \begin{pmatrix} \epsilon_{n\ \ 1} \\ 0_d \\ \vdots \\ 0_d \end{pmatrix}, \text{ and } \quad \mathbf{A} = \begin{pmatrix} A_1 & A_2 & \cdots & A_{p\ \ 1} & A_p \\ I_d & 0_{d,d} & \cdots & 0_{d,d} & 0_{d,d} \\ 0_{d,d} & I_d & & 0_{d,d} & 0_{d,d} \\ \vdots & & \ddots & & \vdots \\ 0_{d,d} & 0_{d,d} & \cdots & I_d & 0_{d,d} \end{pmatrix}. \tag{52}$$

Here I_d is the d dimensional identity matrix, 0_d a vector of d zeros, and $0_{d,d}$ a $d \times d$ matrix of zeros. We note that as each of the ys in Y has d entries, the vectors are both dp entries long. Similarly, each of the matrices in $\mathbf{A}$ is a $d \times d$ matrix, so that $\mathbf{A}$ has $dp \times dp$ entries. We note that in writing out the expression for y_n from Equation (51) we recover Equation (50), while the expression for $y_{n\ \ 1}$ from Equation (51) is simply $y_{n\ \ 1} = I_d \cdot y_{n\ \ 1} + 0_d$. We effectively keep copies of the most recent p entries of the time series to be able to calculate the next entry in just one step.

5.1.1 Autocovariance for autoregressive models

This section extracts the most relevant parts of section 2.1.4 of Reference [26], writing them in the form explicitly useful for our work. The autocovariance of an autoregressive model with mean value 0 is found as $\mathbf{K}_{yy}(h) = \langle y_n y_{n\ h}^T \rangle$, where $h \geq 0$ is the time-lag of the autocovariance. We note that for this autocovariance we have $\mathbf{K}_{yy}(h)^T = \mathbf{K}_{yy}(-h)$. Returning to the d dimensional order p process from Equation (50), we calculate the autocovariance by postmultiplication with $y_{n\ h}^T$ and finding the expectation value;

$$\begin{aligned}\mathbf{K}_{yy}(h) = \langle y_n y_{n\ h}^T \rangle &= \sum_{j=1}^{p} A_j \langle y_{n\ j} y_{n\ h}^T \rangle + \langle \epsilon_{n\ 1} y_{n\ h}^T \rangle \\ &= \sum_{j=1}^{p} A_j \mathbf{K}_{yy}(h-j) + \left\langle \epsilon_{n\ 1} \left(\sum_{l=1}^{p} A_l y_{n\ h\ l} + \epsilon_{n\ h} \right)^T \right\rangle \\ &= \sum_{j=1}^{p} A_j \mathbf{K}_{yy}(h-j) + \langle \epsilon_{n\ 1} \epsilon_{n\ h}^T \rangle . \end{aligned} \tag{53}$$

Here we use that the noise is not correlated with the previous signal values, $\langle \epsilon_n y_{n^0\ n}^T \rangle = 0$, such that all terms $\left\langle \epsilon_{n\ 1} \left(\sum_{l=1}^{p} A_l y_{n\ h\ l} \right) \right\rangle$ are equal to 0, no matter what time-lag we choose for h. We note that ϵ_n is a white noise process, such that $\langle \epsilon_{n1} \epsilon_{n2}^T \rangle = \Sigma_\epsilon \delta_{n1,n2}$. Thus, after separating the cases $h = 0$ and $h > 0$, we find the so-called Yule-Walker equations,

$$\mathbf{K}_{yy}(0) = \sum_{j=1}^{p} A_j \mathbf{K}_{yy}(0-j) + \Sigma_\epsilon = \sum_{j=1}^{p} A_j \mathbf{K}_{yy}(j)^T + \Sigma_\epsilon, \tag{54}$$

$$\mathbf{K}_{yy}(h) = \sum_{j=1}^{p} A_j \mathbf{K}_{yy}(h-j). \tag{55}$$

Where we use that $\mathbf{K}_{yy}(-h) = \mathbf{K}_{yy}(h)^T$. The second of these equations (Equation (55)) looks very similar to the original autoregressive model. Indeed, if the autocovariances $\mathbf{K}_{yy}(0), \mathbf{K}_{yy}(1), ..., \mathbf{K}_{yy}(p-1)$ are known, we can use that equation to calculate the higher time-lag autocovariances iteratively. To be able to determine those low time-lag autocovariances we need to consider the dp dimensional order 1 autoregressive model corresponding to this model using Equation (52). We can express the autocovariance of that model, $\mathbf{K}_{YY}(h)$, in terms of the autocovariances of the original model as

$$\begin{aligned}\mathbf{K}_{YY}(h) = \langle Y_n Y_{n\ h}^T \rangle &= \left\langle \begin{pmatrix} y_n \\ y_{n\ 1} \\ \vdots \\ y_{n\ p+1} \end{pmatrix} \begin{pmatrix} y_{n\ h}^T & y_{n\ h\ 1}^T & \cdots & y_{n\ h\ p+1}^T \end{pmatrix} \right\rangle \\ &= \begin{pmatrix} \langle y_n y_{n\ h}^T \rangle & \langle y_{n\ 1} y_{n\ h\ 1}^T \rangle & \cdots & \left\langle y_n y_{n\ h\ p+1}^T \right\rangle \\ \langle y_{n\ 1} y_{n\ h}^T \rangle & \langle y_{n\ 1} y_{n\ 1\ h}^T \rangle & \cdots & \left\langle y_{n\ 1} y_{n\ h\ p+1)}^T \right\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle y_{n\ p+1} y_{n\ h}^T \rangle & \langle y_{n\ p+1} y_{n\ h\ 1}^T \rangle & \cdots & \left\langle y_{n\ p+1} y_{n\ p+1\ h}^T \right\rangle \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{K}_{yy}(h) & \mathbf{K}_{yy}(h+1) & \cdots & \mathbf{K}_{yy}(h+p-1) \\ \mathbf{K}_{yy}(h-1) & \mathbf{K}_{yy}(h) & \cdots & \mathbf{K}_{yy}(h+p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{K}_{yy}(h-p+1) & \mathbf{K}_{yy}(h-p+2) & \cdots & \mathbf{K}_{yy}(h) \end{pmatrix}, \end{aligned} \tag{56}$$

such that for $h = 0$, we find the autocovariances $\mathbf{K}_{yy}(0), \mathbf{K}_{yy}(1), ..., \mathbf{K}_{yy}(p-1)$ as the first d rows of $\mathbf{K}_{YY}(0)$, which is the first row in the visualization in Equation (56). For this VAR(1) model the Yule-Walker equations from Equation (54) and Equation (55) simplify to

$$\mathbf{K}_{YY}(0) = \mathbf{A}\mathbf{K}_{YY}(1)^T + \Sigma_\epsilon \text{ and } \mathbf{K}_{YY}(h) = \mathbf{A}\mathbf{K}_{YY}(h-1).$$

We can plug $\mathbf{K}_{YY}(1) = \mathbf{A}\mathbf{K}_{YY}(0)$ into the first Yule-Walker equation and find an equation only depending on $\mathbf{K}_{YY}(0)$,

$$\mathbf{K}_{YY}(0) = \mathbf{A}\mathbf{K}_{YY}(0)^T\mathbf{A}^T + \Sigma_\epsilon = \mathbf{A}\mathbf{K}_{YY}(0)\mathbf{A}^T + \Sigma_\epsilon. \tag{57}$$

To solve this equation for $\mathbf{K}_{YY}(0)$, we recast it by unraveling the covariance matrices into vectors. We define the operator *vec*, which stacks the columns of a matrix on top of each other to form a vector. Equation (57) can be rewritten to

$$vec\left(\mathbf{K}_{YY}(0)\right) = \left(\mathbf{A}\otimes\mathbf{A}\right) vec\left(\mathbf{K}_{YY}(0)\right) + vec\left(\Sigma_\epsilon\right), \tag{58}$$

where $\otimes$ denotes the Kronecker product, which multiplies each entry in the first matrix with the second matrix. This equation can be solved to find the autocovariance "vector."

$$vec\left(\mathbf{K}_{YY}(0)\right) = \left(I_{(Kp)^2} - \mathbf{A}\otimes\mathbf{A}\right)^{-1} vec\left(\Sigma_\epsilon\right). \tag{59}$$

Finally, we can reshape the autocovariance vector into the autocovariance matrix and extract the p autocovariance matrices for the original d dimensional VAR(p) model, $\mathbf{K}_{yy}(0), \mathbf{K}_{yy}(1), ..., \mathbf{K}_{yy}(p-1)$. These can then be used in the second Yule-Walker equation (Equation (55)) to iteratively find the autocovariance matrix of the original model for arbitrary time lags. Using this information, we can look at a vector autoregressive model to model the generalized Langevin equation.

5.2 Application to the generalized Langevin equation

To model the generalized Langevin equation using such a vector autoregressive model, we follow the ideas presented in Reference [20]. There, an autoregressive model is constructed to study turbulence in fluids. For this purpose, a non-Markovian process is considered that is equivalent to the generalized Langevin equation with a combined delta function and exponential memory kernel, but without an external potential. Let us consider this as the starting point to build our autoregressive model. The process under consideration is

$$m\dot{v} = -2\zeta\sqrt{\kappa m}\, v(t) - \int_1^t dt' \frac{B}{\tau} e^{-(t-t')/\tau} v(t') + F_1(t) + F_2(t), \tag{60}$$

where the noise is split into two parts; $F_1 \sim \mathcal{N}(0, 4k_BT\zeta\sqrt{\kappa m})$ is a Gaussian white noise and $F_2 \sim \mathcal{N}(0, k_BT\frac{B}{\tau}(1-e^{-2t/\tau}))$ is a colored noise generated by an Ornstein-Uhlenbeck process.

We now discretize the time in steps of length δt by defining $t_n = n\,\delta t$ and $v_n = v(t_n)$. We use the Euler method, such that, for the left-hand side of the equation, we have $\dot{v} \approx (v_n - v_{n-1})/\delta t$. Meanwhile, we evaluate the right-hand side of the equation at $t = t_{n-1}$, and we replace the integral by a sum with the integrand evaluated at times $t_{n-1}, t_{n-2}, t_{n-3}, ...$ We then find

$$m(v_n - v_{n-1})/\delta t = -2\zeta\sqrt{\kappa m}\, v_{n-1} - \sum_{j=1}^{1} \frac{B\delta t}{\tau} e^{-(j-1)\delta t/\tau} v_{n-j} + \eta_{n-1} + w_{n-1} \tag{61}$$

$$v_n = \left(1 - 2\delta t\zeta\sqrt{\kappa/m} - \frac{B\delta t^2}{\tau m}\right) v_{n-1} - \frac{B\delta t^2}{\tau m}\sum_{j=2}^{1} e^{-(j-1)\delta t/\tau} v_{n-j} + \frac{\delta t}{m}\eta_{n-1} + \frac{\delta t}{m} w_{n-1}, \tag{62}$$

where $\eta \sim \mathcal{N}\left(0, 4k_BT\zeta\sqrt{\kappa m}\right)$ is the white noise and w_{n-1} is noise generated by an Ornstein-Uhlenbeck process;

$$w_n = e^{-\delta t/\tau} w_{n-1} + \eta'_{n-1}, \qquad \eta' \sim \mathcal{N}\left(0, \frac{Bk_BT}{\tau}(1 - e^{-2\delta t/\tau})\right). \tag{63}$$

A problem with this process is that we need an infinite number of previous outputs to calculate it. As this is not feasible, we choose an order for our model by calculating a finite number of terms and using that result. We note that this makes it so that our model describes a different integro-differential equation, one where the lower limit of the integral is equal to $t - p\delta t$ instead of the fixed value t_m. However, if $p\delta t$ is sufficiently big compared to the decay rate of the memory kernel τ, this should not make a large difference, as the exponential memory function will be small at that point. Finally, to add an overall

harmonic potential to the process, we introduce a third variable x, such that $\dot{x}(t) = v(t)$. Discretizing this, adding a $-\delta t \frac{\kappa}{m} x_{n\ 1}$ term into equation (62), and combining everything into one vector autoregressive model we have

$$\begin{pmatrix} v_n \\ x_n \\ w_n \end{pmatrix} = \begin{pmatrix} 1 - 2\zeta\delta t\sqrt{\frac{\kappa}{m}} - \frac{B\delta t^2}{\tau m} & -\frac{\kappa\delta t}{m} & \frac{\delta t}{m} \\ \delta t & 1 & 0 \\ 0 & 0 & e^{\ \frac{\delta t}{\tau}} \end{pmatrix} \begin{pmatrix} v_{n\ 1} \\ x_{n\ 1} \\ w_{n\ 1} \end{pmatrix} - \sum_{j=2}^{p} \begin{pmatrix} \frac{B\delta t^2}{\tau m} e^{\ (j\ \ 1)\frac{\delta t}{\tau}} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} v_{n\ j} \\ x_{n\ j} \\ w_{n\ j} \end{pmatrix} + \begin{pmatrix} \eta_{n\ 1} \\ 0 \\ \eta^{0}_{n\ 1} \end{pmatrix}. \tag{64}$$

This discrete model does converge to the continuous case for the combined limits $\delta t \to 0$ and $p \to \infty$. However, for finite δt, the model introduces errors of order δt^2 with the continuous case. As such, quantitative comparisons between the autoregressive model and the generalized Langevin equation will not be possible. However, we will show that the autoregressive model exhibits qualitatively similar behaviour to the continuous model.

Using the previous result for the autocovariance, we can determine the covariance matrix of this model, and thus, in particular, find the covariance $\langle x(t^n)x(t^{n\ \ k})\rangle$. To do so, we construct the corresponding vector autoregressive model of order 1;

$$\begin{pmatrix} v_n \\ x_n \\ w_n \\ v_{n\ 1} \\ x_{n\ 1} \\ w_{n\ 1} \\ v_{n\ 2} \\ \vdots \\ w_{n\ \ p+1} \end{pmatrix} = \mathbf{A}_{3p,3p} \begin{pmatrix} v_{n\ 1} \\ x_{n\ 1} \\ w_{n\ 1} \\ v_{n\ 2} \\ \\ \vdots \\ \\ \\ w_{n\ \ p} \end{pmatrix} + \begin{pmatrix} \eta_{n\ 1} \\ 0 \\ \eta^{0}_{n\ 1} \\ 0 \\ 0 \\ 0 \\ \vdots \\ \\ 0 \end{pmatrix}, \tag{65}$$

where $\mathbf{A}_{3p,3p}$ is a square matrix with $3p$ columns and $3p$ rows and is given as

$$\begin{pmatrix} 1 - 2\zeta\delta t\sqrt{\frac{\kappa}{m}} - \frac{B\delta t^2}{\tau m} & -\frac{\kappa\delta t}{m} & \frac{\delta t}{m} & \frac{B\delta t^2}{\tau m} e^{\ 1\frac{\delta t}{\tau}} & 0 & 0 & \cdots & \frac{B\delta t^2}{\tau m} e^{\ (p\ \ 1)\frac{\delta t}{\tau}} & 0 & 0 \\ \delta t & 1 & 0 & 0 & 0 & 0 & & 0 & 0 & 0 \\ 0 & 0 & e^{\ \frac{\delta t}{\tau}} & 0 & 0 & 0 & & 0 & 0 & 0 \\ & I_3 & & & 0_{3,3} & & \cdots & & 0_{3,3} & \\ & & & & & & \ddots & & & \\ & 0_{3,3} & & & I_3 & & & & & \\ & & & & & & & & \vdots & \\ & \vdots & & & & & \ddots & \ddots & & \\ & 0_{3,3} & & \cdots & & & I_3 & & 0_{3,3} & \end{pmatrix}.$$

As a reminder, here, I_3 denotes the identity matrix of size 3 and $0_{3,3}$ stands for the square zero matrix of size 3.

Now, following Equation (59), we find the covariance matrix at time-lag 0 for this model, $\mathbf{K}_{GLE}(0)$, as

$$vec\left(\mathbf{K}_{GLE}(0)\right) = \left(I_{(3p)^2} - \mathbf{A}_{3p,3p} \otimes \mathbf{A}_{3p,3p}\right)^{\ 1} vec\left(\Sigma_\eta\right), \tag{66}$$

where Σ_η is the covariance matrix of the noise processes in Equation (65). Since the two noise processes are independent this is a $3p \times 3p$ matrix that is empty except for the $(1,1)$ and $(3,3)$ components, which are $\langle \eta_n \eta_n \rangle$ and $\langle \eta^0_n \eta^0_n \rangle$, respectively.

However, as finding the inverse of the sparse matrix $\left(I_{(3p)^2} - \mathbf{A}_{3p,3p} \otimes \mathbf{A}_{3p,3p}\right)$ is very inefficient, even numerically, a more practical solution allows for numerically solving

$$\left(I_{(3p)^2} - \mathbf{A}_{3p,3p} \otimes \mathbf{A}_{3p,3p}\right) vec\left(\mathbf{K}_{GLE}(0)\right) = vec\left(\Sigma_\eta\right), \tag{67}$$

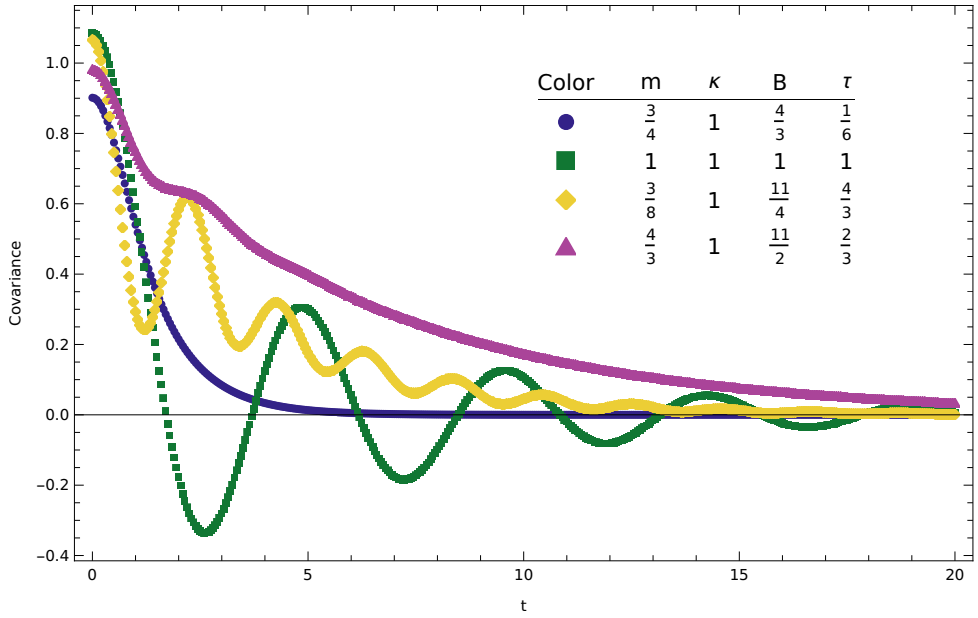

Figure 12: Correlation functions for the discretized signal with parameters that are the same as in Figure 8. The discretization timestep $\delta t = 1/20$ and the autoregressive model order $p = 60$. We note that the correlation functions show the same qualitative behavior in both the discrete and continuous signals.

for the unknown $\mathbf{K}_{GLE}(0)$. From this covariance matrix we find the p 3 by 3 covariance matrices for the original model $\mathbf{K}_{gle}(h)$ with $0 \leq h < p$. The second Yule-Walker equation for the original model now allows us to find the longer time-lag covariance matrices

$$\mathbf{K}_{gle}(h) = \begin{pmatrix} 1 - 2\zeta\delta t\sqrt{\frac{\kappa}{m}} - \frac{B\delta t^2}{\tau m} & -\frac{\kappa\delta t}{m} & \frac{\delta t}{m} \\ \delta t & 1 & 0 \\ 0 & 0 & e^{\frac{\delta t}{\tau}} \end{pmatrix} \mathbf{K}_{gle}(h-1) - \sum_{j=2}^{p} \begin{pmatrix} \frac{B\delta t^2}{\tau m} e^{(j-1)\frac{\delta t}{\tau}} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \mathbf{K}_{gle}(h-j). \tag{68}$$

In particular, the autocovariance of the x variable, $\langle x(t^n)x(t^{n-h})\rangle$, is important as the input for the Information Bottleneck. This is found as the middle entry in $\mathbf{K}_{gle}(h)$.

5.3 Results

Using the above method, we determine the covariance for both the purely exponential and mixed delta-exponential memory cases. We can compare these to the continuous case and use them in the information bottleneck framework.

5.3.1 Exponential memory kernel

To get the purely exponential memory kernel, we use $\zeta = 0$ in the derived autoregressive model. Using the same parameter values as for the continuous model and choosing discretization timestep $\delta t = 1/20$ and autoregressive model order $p = 60$, we find qualitatively similar covariances as for the continuous case, as shown in Figure 12. In comparison with the correlation functions from Figure 8 we see that the qualitative behavior of the correlation functions is the same. However, quantitatively some differences show up. These arise due to discretization and are an expected effect.

Since we find qualitatively similar behavior we will interpret this autoregressive model as a good discretization of the generalized Langevin equation with an exponential memory term. We use the generated covariance functions in the information bottleneck framework to find the respective kernels, shown in Figure 13. Once again, the information bottleneck kernels are similar for different covariance functions. The kernel's first and last three components are non-zero, while the rest are zero.

We suspect that the last three components are non-zero is due to the history not extending to $-\infty$ in the discrete model. To estimate the length of the history, the kernel incorporates information from as far back as possible. Indeed, we see that the future information as a function of the past information slightly

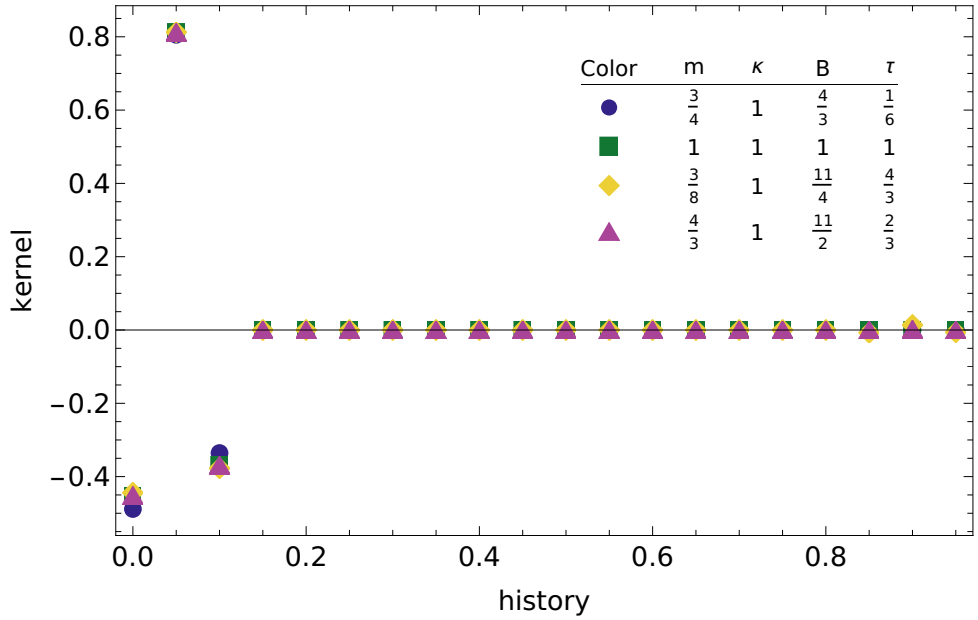

Figure 13: Information bottleneck kernels for the discretized signal with parameters that are the same as in Figure 8. We use a discretization timestep $\delta t = 1/20$, an autoregressive model order $p = 60$ and an information bottleneck signal length of $N = 20$.

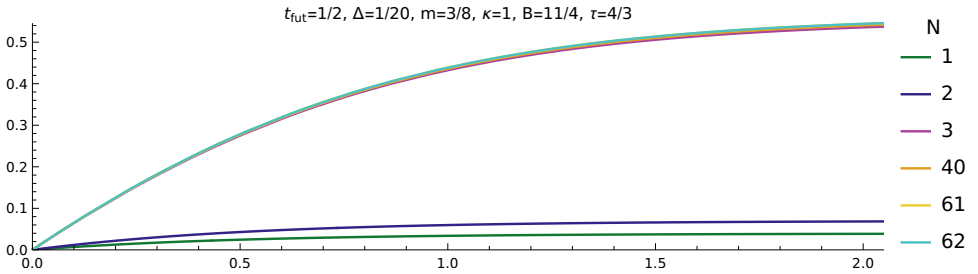

Figure 14: Future information as a function of past information for the signal with the same parameters as the yellow signal in Figures 12 and 13 for several values of the vector length.

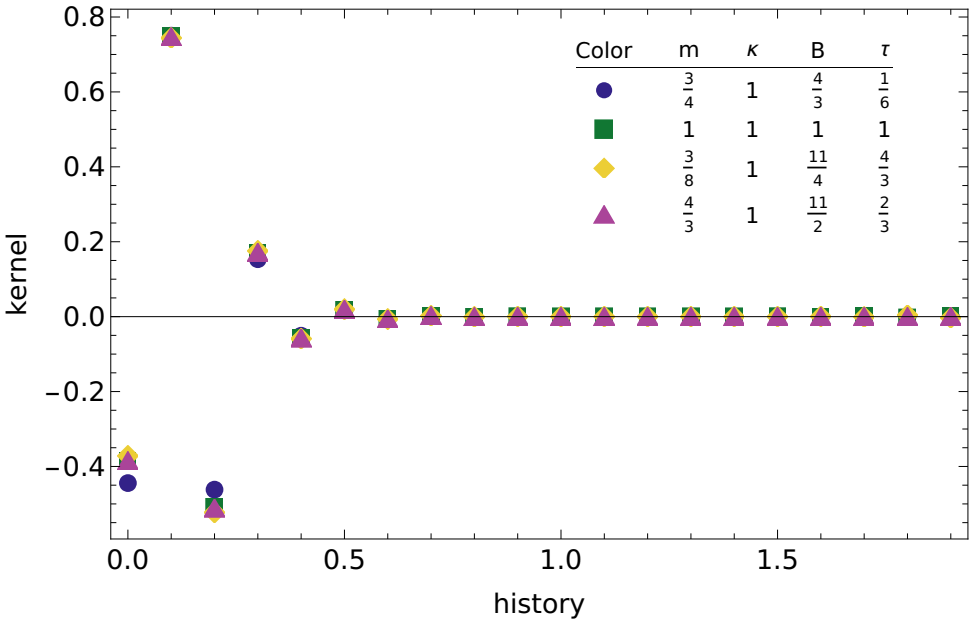

Figure 15: Information bottleneck kernels for the discretized signal with parameters that are the same as in Figure 13, but using a different discretization timestep for the autoregressive model and the information bottleneck. We use an autoregressive model discretization timestep $\delta t = 1/20$, an autoregressive model order $p = 60$, and an information bottleneck discretization timestep $\Delta = 1/10$.

increases with increasing kernel/vector length, but only up to a kernel length of $N = 62$, after which this information bound stays the same. However, this effect is very small compared to the information bound changes we see between a kernel length of 1 and 3. The information bound for a selected number of values for N is shown in Figure 14.

Here, the discretization timesteps of the autoregressive model δt and the information bottleneck Δ are equal. In the discussion for the continuous signal, we proposed that more than 3 entries in the vector are non-zero to find better estimates of the instantaneous velocity and acceleration of the signal. If this is the case, we should expect that if we take a different discretization for the information bottleneck for a discrete signal, the same behavior reemerges. We indeed see that the kernel becomes extended when we use every other signal point as the input for the information bottleneck, as shown in Figure 15.

5.3.2 Combined delta function and exponential memory kernel

To get a combined memory kernel, we use $\zeta > 0$ in the autoregressive model. Using the same parameter values as for the continuous model and choosing discretization timestep $\delta t = 1/20$ and autoregressive model order $p = 60$, we find qualitatively similar covariances as for the continuous case, as shown in Figure 16.

Information bottleneck kernels corresponding to the covariance functions are shown in Figure 17. Strikingly, for the green and blue models, multiple values at the start of the information bottleneck kernel are non-zero. In contrast, for the pink and yellow models, some values at the end of the information bottleneck kernel are bigger than those in the middle. The non-zero values at the end suggest that the information bottleneck kernel in this example is not long enough for these last kernels.

The absolute values of the entries in an information bottleneck kernel of length $N = 65$, such that it is longer than the autoregressive model order $p = 60$, are plotted in Figure 18 with a logarithmic vertical axis. Apart from the first two entries, we see two different behaviors. The more recent values for the blue and green signals have a high magnitude, but the subsequent decay is fast. On the other hand, the decay is slow for the pink and yellow signals, while the more recent entries have a low magnitude. This suggests that the relative importance of the two types of corrective force present here plays an important role in shaping the kernel.

To check if this is the case, let us look at the future information as a function of the past information plot for several different information bottleneck kernel lengths. In Figures 19 and 20, this plot is shown

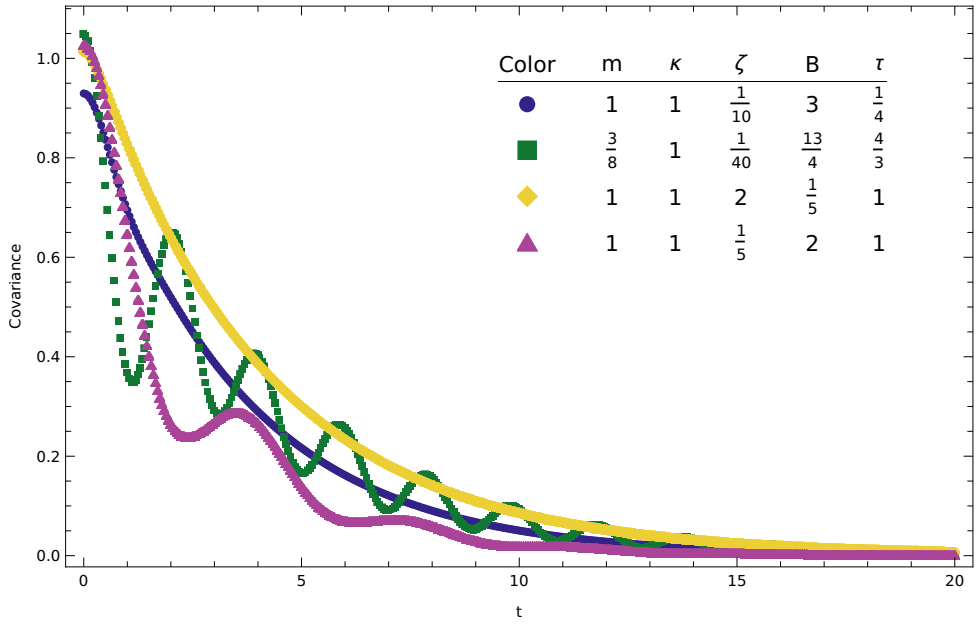

Figure 16: Correlation functions for the discretized signal with parameters that are the same as in Figure 10. The discretization timestep $\delta t = 1/20$ and the autoregressive model order $p = 60$. We note that the correlation functions show the same qualitative behavior in both the discrete and continuous signals.

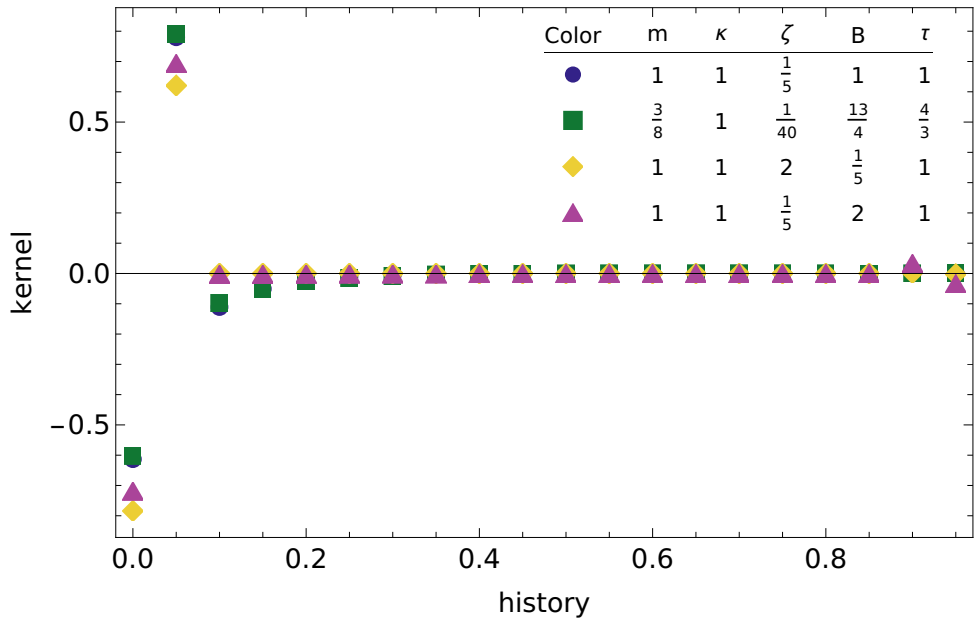

Figure 17: Information bottleneck kernels for the discretized signal with parameters that are the same as in Figure 10. We use a discretization timestep $\delta t = 1/20$, an autoregressive model order $p = 60$, and a information bottleneck signal length of $N = 20$.

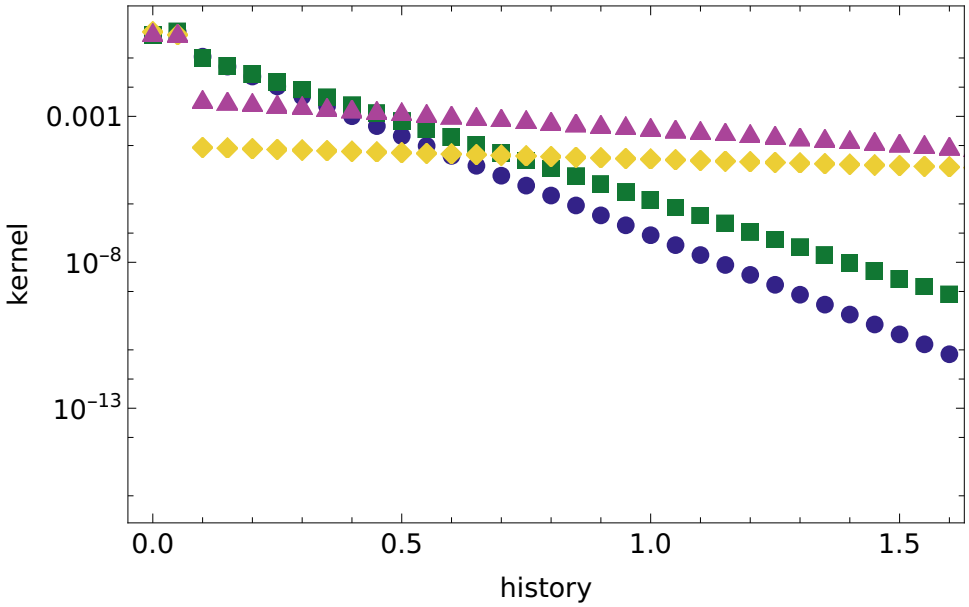

Figure 18: Absolute values of information bottleneck kernels for the discretized signal with parameters that are the same as in Figure 10. We use a discretization timestep $\delta t = 1/20$, an autoregressive model order $p = 60$, and a information bottleneck signal length of $N = 65 > p$. Note that the vertical axis is on a logarithmic scale. We note that not the full information bottleneck kernel is shown here, as the interplay between the finite autoregressive model order and the finite information bottleneck kernel length introduces artefacts, which are explained in the main text.

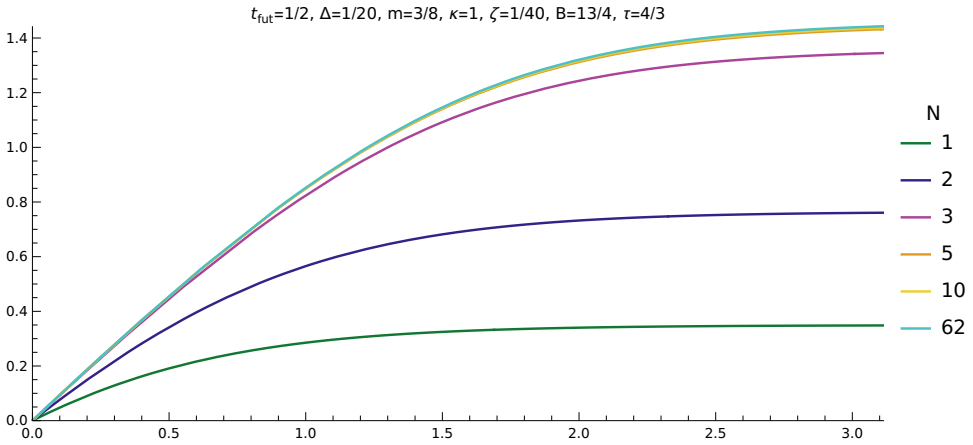

Figure 19: Future information as a function of past information for the signal with the same parameters as the green signal in Figure 16 for several values of the vector length.

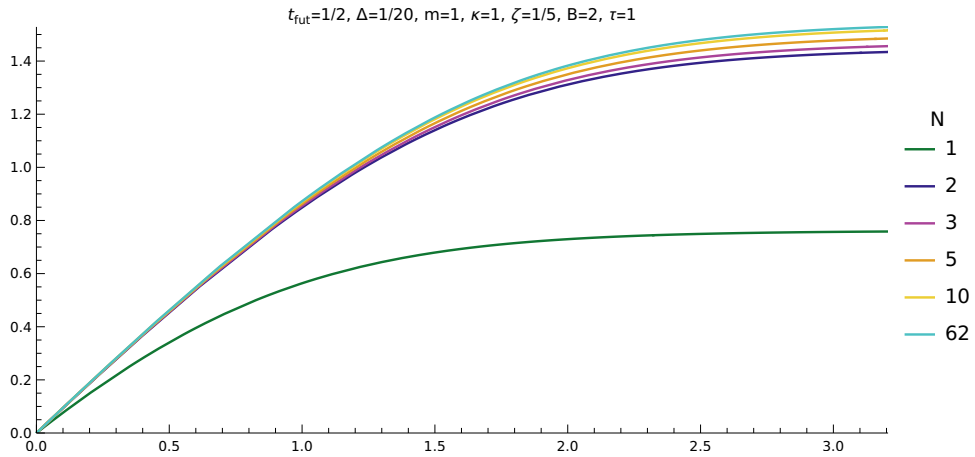

Figure 20: Future information as a function of past information for the signal with the same parameters as the pink signal in Figure 16 for several values of the vector length.

for the green and pink signals, respectively. We choose the green signal as a case where the exponential memory function corrective force is much more important than the instantaneous friction. On the other hand, the pink signal has both corrective forces of similar importance.

In the case where the corrective force with exponential memory is more important (Figure 19), the most information is found in the most recent five measurements, particularly in the first three. The high importance of the first three components mirrors the results of the purely exponential memory function. This suggests that the first three entries in the information bottleneck kernel perform a similar function as the three components in the exponential memory function case. The rest of the kernel now elucidates the correction needed for the influence of the instantaneous friction on this "baseline" prediction. Since friction is an instantaneous effect, it seems natural that the most recent history of the signal is much more informative of this correction than the farther past.

In the case where the instantaneous friction is more important (Figure 20), the most information is found in the most recent two measurements, with significant additional information available with higher information bottleneck kernel lengths. Indeed, we can clearly distinguish the curves for $N = 10$ and $N = 62$ in this case, whereas in Figure 19 these curves are indistinguishable. The high importance of just two components mirrors the information bottleneck kernels found in Reference [9]. So, it seems that in contrast to the previous case, a baseline prediction is made for instantaneous friction, with the rest of the kernel elucidating the influence of the exponential memory function. Here, the slower decay of the magnitude suggests that the history of the signal is also informative of the correction needed, which we rationalize by the fact that the exponential memory function does depend on the history of the signal.

This difference suggests that for regions in parameter space where both corrective forces are important or the instantaneous friction is more important the information bottleneck kernels will be longer than in regions where the exponential memory is more important.

6 Discussion and conclusion

In this project, we have studied the discrete-time information bottleneck for a generalized Langevin equation. This equation yields signals that are Gaussian, stationary and non-Markovian. The generalized Langevin equation has a memory kernel, such that the instantaneous friction force, which is present in the stochastic damped harmonic oscillator, is replaced by a history-dependent corrective force. In particular, we considered the cases of a memory kernel that includes a single exponential decay or a combination of exponential decay with a Dirac delta function. We have chosen these memory kernels as they provide an analytically tractable equation and are, in some ways, the simplest version of a generalized Langevin equation. Other possible kernels are sums of exponential decays or (sums of) fractions. For the latter, a first part of the analysis is given in Appendix C, which led to intractable infinite sums and thus was not pursued further.

We considered the generation of signals in continuous time for the two memory kernels that were pursued

in section 4. We found that the information bottleneck kernels look extended, while the majority of the relevant information is available in the most recent past. The extension of the information bottleneck kernels is due to the information losses introduced by the discretization in the information bottleneck method. As such, part of the prediction that is done in this case is to recover this lost information.

We constructed a discrete version of the signal to eliminate the information loss introduced by the discretization. In section 5 we constructed a vector autoregressive model that generates a discrete signal which qualitatively corresponds to the continuous-time signal. The information bottleneck kernels using this method do not suffer the same information loss. This is a better indicator of what aspects of the signal are necessary for prediction in the absence of discretization.

We find that the information bottleneck kernel has three non-zero entries for the exponential memory function. These are specifically the three most recent values in the signal. This is analogous to the previous result on the stochastic harmonic oscillator, in which the two most recent signal values contain relevant information. For this signal, it seems very likely that the optimal kernel determines the instantaneous position, velocity and acceleration to determine the compression. It is, in this case, possible to write a system of equations $\dot{x} = v$, $\dot{v} = a$, and $\dot{a} = f(a, v, x) + \eta$ with some linear function f and a white noise process η. This follows a similar method as rewriting the deterministic generalized Langevin equation without the memory integral, which is shown in Appendix A. In that case, the fact that these instantaneous measures provide all the information to predict the future of the signal is rationalized easily.

We find that the information bottleneck kernel for the combined exponential decay and Dirac delta function memory kernel is also extended when using the autoregressive model. In this case, we distinguish two types of information bottleneck kernels. In the first type, most information is available in two measurements, but more information is available with an extended exponential tail. This is the case if the instantaneous delta function memory is important relative to the exponential decay memory. It looks like the information bottleneck kernel mimics the strategy for the stochastic damped harmonic oscillator but needs an extended tail to incorporate the second memory effect. Alternatively, in the other type of information bottleneck kernels found for the combined memory kernel, three measurements are needed to find most of the information, with additional information mostly in the next two to three measurements. This is the case if the exponential decay memory is (much) more important than the instantaneous delta function memory. In this case, it seems that the information bottleneck kernel mimics the strategy for the exponential memory kernel but needs slightly more information to determine the effect of the additional instantaneous friction force.

We thus find the optimal information bottleneck kernels for the generalized Langevin equation with either an exponential or a combined exponential and delta function memory kernel. Further research into this subject has two distinct paths forward. On the one hand, it is interesting to consider if any known biological networks use kernels with the same form we find in this work, such as earlier research found that the push-pull network can approximate the optimal kernel for an Ornstein-Uhlenbeck signal and the chemotaxis network can approximate the optimal kernel for the underdamped stochastic harmonic oscillator [8,9]. Alternatively, if no biological candidate is obvious, we ask what kind of chemical reaction network could implement this kernel. Additionally, it might be interesting to consider other, more complicated, memory kernels in the generalized Langevin equation. For example, the analysis for the fractional memory kernel could be completed, and the behavior of sums of exponential decays might be interesting and easily accessible from the current work. Alternatively, one could consider designing a memory kernel from a covariance function with desired properties. Specifically, in determining the covariance functions of the signals in the current work, a characteristic cubic equation was found. It might be interesting to consider what happens were one to construct a quartic equation instead, which might serve as a next generalization.

A Deterministic solution

We note that for both of the memory kernels considered in this report we can find the solution to the deterministic problem analytically. The noise-less equation with both corrective forces is written as

$$m\ddot{x} = -2\zeta\sqrt{\kappa m}\dot{x}(t) - \frac{B}{\tau}\int_{t_m}^{t} dt^0\ e^{\ (t\ \ t^0)/\tau}\dot{x}(t^0) - \kappa x(t). \tag{69}$$

We differentiate both sides to find

$$m\,\dddot{x} = -\kappa\dot{x} - 2\zeta\sqrt{\kappa m}\,\ddot{x} - \frac{B}{\tau}\dot{x}(t) - \frac{B}{\tau}\int_{0}^{t}\frac{e^{(t\ \ s)/\tau}\dot{x}(s)}{-\tau}ds. \tag{70}$$

Here we notice that the integral term is the same as before differentiation, such that we can use Equation (69) again to eliminate the integral. We find

$$m\,\dddot{x} = -\kappa\dot{x} - 2\zeta\sqrt{\kappa m}\,\ddot{x} - \frac{B}{\tau}\dot{x}(t) - \frac{1}{\tau}\left(2\zeta\sqrt{\kappa m}\,\dot{x}(t) + \kappa x(t) + m\ddot{x}\right), \tag{71}$$

which rearranges to

$$\dddot{x} = -\frac{1}{m}\left(\frac{\kappa}{\tau}x + \left(\kappa + \frac{B}{\tau} + \frac{2\zeta\sqrt{\kappa m}}{\tau}\right)\dot{x} + \left(2\zeta\sqrt{\kappa m} + \frac{m}{\tau}\right)\ddot{x}\right). \tag{72}$$

As such, the problem is simplified from an integro-differential equation to an ordinary differential equation, which is much more tractable.

B Solving the GLE

The generalized Langevin equation for times $t \geq 0$ with a harmonic potential well is given as

$$m\ddot{x} = -\int_{t_m}^{t} dt^0\ \Gamma(t - t^0)\dot{x}(t^0) - \kappa x(t) + \eta(t). \tag{73}$$

With m the particle mass (signal inertia), x, $\dot{x}$, and $\ddot{x}$ the particle position and its derivatives (signal and its derivatives), $\Gamma(t)$ the memory kernel (we note that $\Gamma(t) = 0$ for $t < 0$), κ the force constant for the harmonic potential, $t_m \leq 0$ the time to which memory effects extend, and $\eta(t)$ the random force which is Gaussian and obeys $\langle\eta(t)\rangle = 0$, and

$$\langle\eta(t)\eta(t^0)\rangle = k_B T\Gamma(|t - t^0|), \tag{74}$$

according to the dissipation-fluctuation theorem [14, 16].

To solve this equation we will closely follow a paper from Di Terlizzi *et al.* from 2020 [16], which in turn extends the treatment of Fox in 1977 [24] with an overall force.

B.1 Modified Laplace transformation

We use their definition of a modified Laplace transform

$$L_s^{t_m}\{g(t)\} = \int_{t_m}^{1} dt\ e^{\ st}g(t) = \widetilde{G}^{t_m}(s), \tag{75}$$

which is much alike a Laplace transform but has a different lower integration limit. For this transformation we consider some identities

Differentiation in time domain follows a similar rule to the normal Laplace transform, which we derive using partial integration

$$L_s^{t_m}\{\dot{g}(t)\} = \int_{t_m}^{1} dt\ e^{\ -st}\dot{g}(t) = [e^{\ -st}g(t)]_{t_m}^{1} + \int_{t_m}^{1} dt\ se^{\ -st}g(t) = s\widetilde{G}^{t_m}(s) - e^{\ -st_m}g(t_m). \tag{76}$$

For the second derivative we perform a second partial integration to find

$$L_s^{t_m}\{\ddot{g}(t)\} = s^2\widetilde{G}^{t_m}(s) - se^{\ -st_m}g(t_m) - e^{\ -st_m}\dot{g}(t_m). \tag{77}$$

Integration in time domain follows exactly like in ordinary Laplace transformation, again using partial integration we have

$$L_s^{t_m}\left\{\int_{t_m}^{t} g(\tau)d\tau\right\} = \int_{t_m}^{1} dt\ e^{\ -st}\int_{t_m}^{t} d\tau\ g(\tau) = \left[\frac{e^{\ -st}}{-s}\int_{t_m}^{t} d\tau\ g(\tau)\right]_{t_m}^{1} - \int_{t_m}^{1} dt\ \frac{e^{\ -st}}{-s}g(t) = \frac{\widetilde{G}^{t_m}(s)}{s}. \tag{78}$$

Convolution of some function $g(t)$ with some other function $f(t)$ where $f(t<0)=0$ (we note that the memory kernel obeys this structure). We note identity for changing order of integration $\int_a^b dx\int_a^x dy = \int_a^b dy\int_y^b dx$ and write

$$L_s^{t_m}\left\{\int_{t_m}^{t} d\tau\ f(t-\tau)g(\tau)\right\} = \int_{t_m}^{1} dt\int_{t_m}^{t} d\tau\ e^{\ -st}f(t-\tau)g(\tau) = \int_{t_m}^{1} d\tau\int_{\tau}^{1} dt\ e^{\ -st}f(t-\tau)g(\tau).$$

Substituting $u = t-\tau$ we get

$$L_s^{t_m}\left\{\int_{t_m}^{t} d\tau\ f(t-\tau)g(\tau)\right\} = \int_{t_m}^{1} d\tau\int_{0}^{1} du\ e^{\ -s(u+\tau)}f(u)g(\tau) = \int_{t_m}^{1} d\tau\ e^{\ -s\tau}g(\tau)\int_{0}^{1} du\ e^{\ -su}f(u).$$

Where we recognize the modified Laplace transform of $g(t)$ and the original Laplace transform of $f(u)$, such that

$$L_s^{t_m}\left\{\int_{t_m}^{t} d\tau\ f(t-\tau)g(\tau)\right\} = \widetilde{G}^{t_m}(s)\widehat{F}(s). \tag{79}$$

Functions Some basic functions will turn out to be useful to be able to recognize to find inverse transformations. Assuming $b > t_m$ we find

$$L_s^{t_m}\{a\} = a\int_{t_m}^{1} dt\ e^{\ -st} = \frac{ae^{\ -st_m}}{s}, \tag{80}$$

$$L_s^{t_m}\{\delta(t-b)\} = \int_{t_m}^{1} dt\ \delta(t-b)e^{\ -st} = e^{\ -sb}, \tag{81}$$

$$L_s^{t_m}\{g(t-t_m)\} = \int_{t_m}^{1} dt\ e^{\ -st}g(t-t_m) = \int_{0}^{1} du\ e^{\ -su}g(u)e^{\ -st_m} = \widehat{G}(s)e^{\ -st_m}. \tag{82}$$

Meanwhile, for the delta function at t_m, $\delta(t-t_m)$, we use the so-called Stratonovich convention for the integral of the delta function, which states $\int_a^b dx f(x)\delta(x-a) = f(a)/2$. We will use this convention throughout this report. As such, we get

$$L_s^{t_m}\{\delta(t-t_m)\} = \int_{t_m}^{1} dt\ \delta(t-t_m)e^{\ -st} = \frac{e^{\ -st_m}}{2}. \tag{83}$$

B.2 Solving the GLE

We apply this modified Laplace transform to the equation to find

$$mL_s^{t_m}\{\ddot{x}(t)\} = -L_s^{t_m}\left\{\int_{t_m}^{t} dt^0\ \Gamma(t-t^0)\dot{x}(t^0)\right\} - \kappa L_s^{t_m}\{x(t)\} + L_s^{t_m}\{\eta(t)\}$$

$$m\left(s^2\widetilde{X}^{t_m}(s) - sx(t_m)e^{\ st_m} - \dot{x}(t_m)e^{\ st_m}\right) = -\widehat{\Gamma}(s)\left(s\widetilde{X}^{t_m}(s) - x(t_m)e^{\ st_m}\right) - \kappa\widetilde{X}^{t_m}(s) + \widetilde{H}^{t_m}(s). \tag{84}$$

We now collect all the $\widetilde{X}^{t_m}(s)$ terms to find the solution for the Laplace transform of $x(t)$

$$\widetilde{X}^{t_m}(s) = \frac{msx(t_m)e^{\ st_m} + m\dot{x}(t_m)e^{\ st_m} + \widehat{\Gamma}(s)x(t_m)e^{\ st_m} + \widetilde{H}(s)}{ms^2 + \widehat{\Gamma}(s)s + \kappa}. \tag{85}$$

To solve without stating an explicit memory kernel we define a quantity that Di Terlizzi *et al.* call a position susceptibility $\chi_x(t)$, which Laplace transform is

$$\widehat{\chi}_x(s) = \frac{1}{ms^2 + \widehat{\Gamma}(s)s + \kappa}. \tag{86}$$

We find

$$\begin{aligned}
\widetilde{X}^{t_m}(s) &= \left(x(t_m)e^{\ st_m}\left(ms + \widehat{\Gamma}(s)\right) + m\dot{x}(t_m)e^{\ st_m} + \widetilde{H}(s)\right)\widehat{\chi}_x(s) \\
&= \left(\frac{x(t_m)e^{\ st_m}}{s}\left(ms^2 + \widehat{\Gamma}(s)s + \kappa - \kappa\right) + m\dot{x}(t_m)e^{\ st_m} + \widetilde{H}(s)\right)\widehat{\chi}_x(s) \\
&= \left(\frac{x(t_m)e^{\ st_m}}{s}\left(\widehat{\chi}_x(s)^{\ -1} - \kappa\right) + m\dot{x}(t_m)e^{\ st_m} + \widetilde{H}(s)\right)\widehat{\chi}_x(s) \\
&= \frac{x(t_m)e^{\ st_m}}{s}(1 - \kappa\widehat{\chi}_x(s)) + m\dot{x}(t_m)e^{\ st_m}\widehat{\chi}_x(s) + \widetilde{H}(s)\widehat{\chi}_x(s).
\end{aligned} \tag{87}$$

To find the solution for $x(t)$ we now need to perform the inverse modified Laplace transform. The inverse modified Laplace transform is defined as [16]

$$L_t^{\ -1,t_m}\{\widetilde{F}(s)\} = \frac{1}{2\pi i}\int_{\alpha\ i\,1}^{\alpha\ i\,1} ds e^{st}\widetilde{F}(s), \tag{88}$$

where α is chosen such that all the singularities of $\widetilde{F}(s)$ are on the 'left' of the vertical contour, i.e. the real part of the singularities is smaller than α. However, instead of performing this inverse transformation directly we will invert by inspection. We have

$$\begin{aligned}
x(t) = x(t_m)L_t^{\ -1,t_m}\left\{\frac{e^{\ st_m}}{s}\right\} - \kappa x(t_m)L_t^{\ -1,t_m}\left\{\frac{e^{\ st_m}}{s}\widehat{\chi}_x(s)\right\} \\
+ 2m\dot{x}(t_m)L_t^{\ -1,t_m}\left\{\frac{e^{\ st_m}}{2}\widehat{\chi}_x(s)\right\} + L_t^{\ -1,t_m}\left\{\widetilde{H}^{t_m}(s)\widehat{\chi}_x(s)\right\},
\end{aligned}$$

here we recognize (in sequence) the modified Laplace transformations of a constant and the convolutions of $\chi_x(t)$ with a constant, a Dirac-delta function, and $\eta(t)$. Filling in the identifications we find

$$x(t) = x(t_m) - \kappa x(t_m)\int_{t_m}^{t} d\tau\ \chi_x(t-\tau) + 2m\dot{x}(t_m)\int_{t_m}^{t} d\tau\chi_x(t-\tau)\delta(\tau - t_m) + \int_{t_m}^{t} d\tau\ \chi_x(t-\tau)\eta(\tau) \tag{89}$$

$$= x(t_m) - \kappa x(t_m)\int_{0}^{t\ t_m} du\ \chi_x(u) + 2m\dot{x}(t_m)\frac{\chi_x(t-t_m)}{2} + \int_{t_m}^{t} d\tau\ \chi_x(t-\tau)\eta(\tau) \tag{90}$$

$$= x(t_m)\left(1 - \kappa\chi(t-t_m)\right) + m\dot{x}(t_m)\chi_x(t-t_m) + \int_{t_m}^{t} d\tau\ \chi_x(t-\tau)\eta(\tau). \tag{91}$$

Here we change the integration variable in the first integral ($u = t - \tau$) and flip the integration direction, such that the minus signs cancel out. We also define

$$\chi(t) = \int_0^t \chi_x(\tau) d\tau, \tag{92}$$

the anti-derivative of $\chi_x(t)$ in the last line. We also note that we can use Equation (78) to find a relationship between the Laplace transforms of χ and χ_x,

$$\widehat{\chi}(s) = \frac{\widehat{\chi}_x(s)}{s}. \tag{93}$$

From here we can find the velocity and the expectation value/average of $x(t)$ and $v(t)$ quite straightforwardly. We note that for $t \gg t_m$ the expectation values of both become 0, as should be expected for stationary signals.

B.3 Correlation

For the position autocorrelation function we now get

$$\begin{aligned}
\big\langle x(t)x(t^0)\big\rangle = {} & \big\langle x(t_m)x(t_m)\big\rangle \left(1 - \kappa\chi(t - t_m)\right)\left(1 - \kappa\chi(t^0 - t_m)\right) && (94)\\
& + m^2\big\langle \dot{x}(t_m)\dot{x}(t_m)\big\rangle \chi_x(t - t_m)\chi_x(t^0 - t_m) && (95)\\
& + m\big\langle x(t_m)\dot{x}(t_m)\big\rangle \big(\chi_x(t^0 - t_m)\left[1 - \kappa\chi(t - t_m)\right] + \chi_x(t - t_m)\left[1 - \kappa\chi(t^0 - t_m)\right]\big) && (96)\\
& + \int_{t_m}^{t} d\tau\ \chi_x(t - \tau)\big(\left[1 - \kappa\chi(t^0 - t_m)\right]\langle x(t_m)\eta(\tau)\rangle + m\chi_x(t^0 - t_m)\langle \dot{x}(t_m)\eta(\tau)\rangle\big) && (97)\\
& + \int_{t_m}^{t^0} d\tau^0\ \chi_x(t^0 - \tau^0)\big(\left[1 - \kappa\chi(t - t_m)\right]\langle x(t_m)\eta(\tau^0)\rangle + m\chi_x(t - t_m)\langle \dot{x}(t_m)\eta(\tau^0)\rangle\big) && (98)\\
& + \int_{t_m}^{t} d\tau \int_{t_m}^{t^0} d\tau^0\ \chi_x(t - \tau)\chi_x(t^0 - \tau^0)\big\langle \eta(\tau)\eta(\tau^0)\big\rangle. && (99)
\end{aligned}$$

The four terms on lines (97) and (98) contain the correlation between the noise at times $t > t_m$ and the initial conditions. These are not correlated, so the integrals evaluate to 0. The four terms on lines (94), (95) and (96) contain the initial condition correlations and do not simplify further without assuming anything about these initial conditions. The remaining double integral on line (99) is the only part we can simplify. As a shorthand, we call this term $\phi(t, t^0)$. The autocorrelation of the random force is given by the memory kernel due to the fluctuation dissipation theorem as in Equation (74),

$$\langle \eta(t)\eta(t^0)\rangle = k_B T \Gamma(|t - t^0|).$$

To calculate the double integral we apply the modified Laplace transform twice, once to transform t into s and once to transform t^0 into s^0, such that

$$\begin{aligned}
L_{s^0}^{t_m}\left\{L_s^{t_m}\left\{\phi(t, t^0)\right\}\right\} &= \int_{t_m}^{\infty} dt^0 \int_{t_m}^{\infty} dt\ e^{-s^0 t^0} e^{-st} \int_{t_m}^{t} d\tau \int_{t_m}^{t^0} d\tau^0\ \chi_x(t - \tau)\chi_x(t^0 - \tau^0)\Gamma(|\tau - \tau^0|)\\
&= \int_{t_m}^{\infty} d\tau \int_{\tau}^{\infty} dt \int_{t_m}^{\infty} d\tau^0 \int_{\tau^0}^{\infty} dt^0\ e^{-s^0 t^0} e^{-st}\chi_x(t - \tau)\chi_x(t^0 - \tau^0)\Gamma(|\tau - \tau^0|).
\end{aligned}$$

Here we switch the order of integration twice, thus resulting in differing integration boundaries. We substitute $u = t - \tau$ and $u^0 = t^0 - \tau^0$ for the t and t^0 integrals respectively to separate out the position

susceptibilities,

$$L_{s^0}^{t_m}\left\{L_s^{t_m}\left\{\phi(t,t^0)\right\}\right\} = \int_{t_m}^{1} d\tau \int_{t_m}^{1} d\tau^0\ e^{\ s\tau} e^{\ s^0\tau^0}\Gamma(|\tau-\tau^0|)\int_0^1 du\ e^{\ su}\chi_x(u)\int_0^1 du^0\ e^{\ s^0u^0}\chi_x(u^0)$$

$$= \widehat{\chi}_x(s)\widehat{\chi}_x(s^0)\int_{t_m}^{1} d\tau \int_{t_m}^{1} d\tau^0\ e^{\ s(\tau\ \ \tau^0)\ \ \tau^0(s+s^0)}\Gamma(|\tau-\tau^0|) \tag{100}$$

$$= \widehat{\chi}_x(s)\widehat{\chi}_x(s^0)\Phi^0(s,s^0). \tag{101}$$

We define another shorthand for the remaining integrals

For the remaining integrals we substitute $y = \tau - \tau^0$ for τ to find

$$\int_{t_m}^{1} d\tau \int_{t_m}^{1} d\tau^0\ e^{\ s(\tau\ \ \tau^0)\ \ \tau^0(s+s^0)}\Gamma(|\tau-\tau^0|) = \int_{t_m}^{1} d\tau^0\ e^{\ \tau^0(s+s^0)}\int_{t_m\ \ \tau^0}^{1} dy\ e^{\ sy}\Gamma(|y|)$$

$$= \int_{t_m}^{1} d\tau^0\ e^{\ \tau^0(s+s^0)}\left(\int_0^1 \underline{dy\ e^{\ sy}\Gamma(y)} + \int_{t_m\ \ \tau^0}^{0} dy\ e^{\ sy}\Gamma(-y)\right).$$

Here we split the y integral in two pieces, such that we may remove the absolute value signs around y. We recognize the Laplace transformation for $\Gamma(y)$ as the underlined integral and perform integration by parts on the τ^0 integral, integrating the exponent, differentiating the part between the parentheses. We get

$$= \left[\frac{e^{\ \tau^0(s+s^0)}}{-s-s^0}\left(\widehat{\Gamma}(s) + \int_{t_m\ \ \tau^0}^{0} dy\ e^{\ sy}\Gamma(-y)\right)\right]_{t_m}^{1}$$

$$- \int_{t_m}^{1} d\tau^0 \frac{e^{\ \tau^0(s+s^0)}}{-s-s^0}\frac{d}{d\tau^0}\left(\widehat{\Gamma}(s) + \int_{t_m\ \ \tau^0}^{0} dy\ e^{\ sy}\Gamma(-y)\right)$$

$$= \left(\frac{e^{\ 1\,(s+s^0)}}{-s-s^0}\left(\widehat{\Gamma}(s) + \int_{t_m\ \ 1}^{0} dy\ e^{\ sy}\Gamma(-y)\right)\right) - \left(\frac{e^{\ t_m(s+s^0)}}{-s-s^0}\left(\widehat{\Gamma}(s) + \int_{t_m\ \ t_m}^{0} dy\ e^{\ sy}\Gamma(-y)\right)\right)$$

$$- \int_{t_m}^{1} d\tau^0 \frac{e^{\ \tau^0(s+s^0)}}{-s-s^0}\left(\frac{d}{d\tau^0}\widehat{\Gamma}(s) + \frac{d}{d\tau^0}\int_{t_m\ \ \tau^0}^{0} dy\ e^{\ sy}\Gamma(-y)\right)$$

$$= 0 - \frac{e^{\ t_m(s+s^0)}}{-s-s^0}\left(\widehat{\Gamma}(s)+0\right) - \int_{t_m}^{1} d\tau^0 \frac{e^{\ \tau^0(s+s^0)}}{-s-s^0}\left(0 + e^{\ s(t_m\ \ \tau^0)}\Gamma(-(t_m-\tau^0))\right)$$

$$= \frac{e^{\ t_m(s+s^0)}}{s+s^0}\widehat{\Gamma}(s) + \frac{e^{\ st_m}}{s+s^0}\int_{t_m}^{1} d\tau^0\ e^{\ \tau^0 s^0}\Gamma(\tau^0-t_m).$$

Where we use that $\int_0^0 dx f(x) = 0$ and $\frac{d}{dx}\int_a^b{}_{\ x}\, dy f(y) = f(a-x)$. Now we substitute $y^0 = \tau^0 - t_m$ and get

$$= \frac{e^{\ t_m(s+s^0)}}{s+s^0}\widehat{\Gamma}(s) + \frac{e^{\ st_m}}{s+s^0}\int_0^1 dy^0\ e^{\ t_m s^0} e^{\ y^0 s^0}\Gamma(y^0)$$

$$= \frac{e^{\ t_m(s+s^0)}}{s+s^0}\widehat{\Gamma}(s) + \frac{e^{\ t_m(s+s^0)}}{s+s^0}\widehat{\Gamma}(s^0) = \frac{e^{\ t_m(s+s^0)}}{s+s^0}\left(\widehat{\Gamma}(s)+\widehat{\Gamma}(s^0)\right)$$

$$\int_{t_m}^{1} d\tau \int_{t_m}^{1} d\tau^0\ e^{\ s(\tau\ \ \tau^0)\ \ \tau^0(s+s^0)}\Gamma(|\tau-\tau^0|) = \frac{e^{\ t_m(s+s^0)}}{s+s^0}\left(\widehat{\Gamma}(s)+\widehat{\Gamma}(s^0) + \frac{(s+s^0)(\kappa+mss^0)}{ss^0} - \frac{(s+s^0)(\kappa+mss^0)}{ss^0}\right)$$

Plugging our result back into Equation (100), adding 0, and remembering the definition of $\widehat{\chi}_x(s)$ we get

$$\begin{aligned}
L_{s^0}^{t_m}\left\{L_s^{t_m}\left\{\phi(t,t^0)\right\}\right\} &= \widehat{\chi}_x(s)\widehat{\chi}_x(s^0)\frac{e^{\ t_m(s+s^0)}}{s+s^0}\left(\widehat{\Gamma}(s)+\widehat{\Gamma}(s^0)+\frac{(s+s^0)(\kappa+mss^0)}{ss^0}-\frac{(s+s^0)(\kappa+mss^0)}{ss^0}\right)\\
&= \frac{e^{\ t_m(s+s^0)}}{s+s^0}\widehat{\chi}_x(s)\widehat{\chi}_x(s^0)\frac{s^0(ms^2+\widehat{\Gamma}(s)s+\kappa)+s(ms^{02}+\widehat{\Gamma}(s^0)s^0+\kappa)-(s+s^0)(\kappa+mss^0)}{ss^0}\\
&= \frac{e^{\ t_m(s+s^0)}}{(s+s^0)ss^0}\left(s^0\widehat{\chi}_x(s^0)+s\widehat{\chi}_x(s)-\kappa(s+s^0)\widehat{\chi}_x(s)\widehat{\chi}_x(s^0)-m(s+s^0)ss^0\widehat{\chi}_x(s)\widehat{\chi}_x(s^0)\right)\\
&= e^{\ t_m(s+s^0)}\left(\frac{s^0\widehat{\chi}_x(s^0)}{ss^0(s+s^0)}+\frac{s\widehat{\chi}_x(s)}{ss^0(s+s^0)}-\kappa\frac{\widehat{\chi}_x(s)}{s}\frac{\widehat{\chi}_x(s^0)}{s^0}-m\widehat{\chi}_x(s)\widehat{\chi}_x(s^0)\right).
\end{aligned}$$

Such that we now almost have the equation in a form in which we can recognize all the parts to do the inverse modified Laplace transform. We first note that

$$\frac{b}{ab(a+b)}=\frac{a+b}{ab(a+b)}-\frac{a}{ab(a+b)}=\frac{1}{ab}-\frac{1}{b(a+b)}, \tag{102}$$

which we use rewrite the first two terms between the brackets. After distributing the exponent we have

$$\begin{aligned}
L_{s^0}^{t_m}\left\{L_s^{t_m}\left\{\phi(t,t^0)\right\}\right\} = &-\kappa e^{\ t_m s}\frac{\widehat{\chi}_x(s)}{s}e^{\ t_m s^0}\frac{\widehat{\chi}_x(s^0)}{s^0}\\
&- me^{\ t_m s}\widehat{\chi}_x(s)e^{\ t_m s^0}\widehat{\chi}_x(s^0)\\
&+\left(\frac{e^{\ t_m s}}{s}\frac{e^{\ t_m s^0}\widehat{\chi}_x(s^0)}{s^0}-\frac{e^{\ t_m(s+s^0)}\widehat{\chi}_x(s^0)}{s^0(s+s^0)}\right)\\
&+\left(\frac{e^{\ t_m s^0}}{s^0}\frac{e^{\ t_m s}\widehat{\chi}_x(s)}{s}-\frac{e^{\ t_m(s+s^0)}\widehat{\chi}_x(s)}{s(s+s^0)}\right).
\end{aligned}$$

We note that for the κ and m terms the two transformations work separately. For the m term we recognize Equation (82), while for the κ term we recognize a combination of Equations (78) and (82). For the first term in both sets of brackets we recognize a combination of Equations (80), (78) and (82). For the last term in both sets of brackets we consider the double modified Laplace transformation of $\theta(t-t^0)\chi(t-t^0)$, with $\theta(x)$ the Heaviside step function. We use the substitution $u=t-t^0$ and find

$$\begin{aligned}
L_{s^0}^{t_m}\left\{L_s^{t_m}\left\{\theta(t-t^0)\chi(t-t^0)\right\}\right\} &= \int_{t_m}^{1}dt^0\int_{t_m}^{1}dt\ e^{\ s^0t^0}e^{\ st}\theta(t-t^0)\chi(t-t^0)\\
&= \int_{t_m}^{1}dt^0\int_{t_m}^{1}\int_{t^0}^{1}du\ e^{\ s^0t^0}e^{\ s(u+t^0)}\theta(u)\chi(u)\\
&= \int_{t_m}^{1}dt^0\ e^{\ (s+s^0)t^0}\int_{0}^{1}du\ e^{\ su}\int_0^u d\tau\ \chi_x(\tau)\\
&= \frac{e^{\ t_m(s+s^0)}}{s+s^0}\frac{\widehat{\chi}_x(s)}{s}.
\end{aligned}$$

Plugging in all the identifications we have

$$\begin{aligned}
\phi(t,t^0) = &-\kappa\chi(t-t_m)\chi(t^0-t_m)-m\chi_x(t-t_m)\chi_x(t^0-t_m)\\
&+\chi(t^0-t_m)-\theta(t^0-t)\chi(t^0-t)+\chi(t-t_m)-\theta(t-t^0)\chi(t-t^0).
\end{aligned} \tag{103}$$

Thus, substituting Equation (103) into the position autocorrelation (at line (99)) we find

$$\begin{aligned}
\langle x(t)x(t^0)\rangle = &\big\langle x(t_m)x(t_m)\big\rangle\left(1-\kappa\chi(t-t_m)\right)\left(1-\kappa\chi(t^0-t_m)\right)+m^2\big\langle\dot{x}(t_m)\dot{x}(t_m)\big\rangle\chi_x(t-t_m)\chi_x(t^0-t_m)\\
&+m\big\langle x(t_m)\dot{x}(t_m)\big\rangle\left(\chi_x(t^0-t_m)\left[1-\kappa\chi(t-t_m)\right]+\chi_x(t-t_m)\left[1-\kappa\chi(t^0-t_m)\right]\right)\\
&+k_BT\big(\chi(t-t_m)+\chi(t^0-t_m)-\chi(|t-t^0|)-\kappa\chi(t-t_m)\chi(t^0-t_m)-m\chi_x(t-t_m)\chi_x(t^0-t_m)\big).
\end{aligned} \tag{104}$$

As a reference, $\chi_x(t)$ is defined by its Laplace transform

$$\widehat{\chi}_x(s) = \frac{1}{ms^2 + \widehat{\Gamma}(s)s + \kappa}, \tag{105}$$

and $\chi(t)$ is defined as the integral of $\chi_x(t)$

$$\widehat{\chi}(t) = \int_0^t \chi_x(\tau)d\tau = L_t^{\ 1}\left\{\frac{s^{\ 1}}{ms^2 + \widehat{\Gamma}(s)s + \kappa}\right\}, \tag{106}$$

B.4 Limit for equilibrium statistics

For our applications we are mostly interested in the equilibrium statistics of the signal x. As such we remove the initial conditions by taking the limit $t_m \to -\infty$. To do so we determine the limits $\lim_{t\,!\,\infty} \chi_x(t)$ and $\lim_{t\,!\,\infty} \chi(t)$.

We note the final value theorem for Laplace transforms. This theorem states that if a function $f(t)$ has a Laplace transform $\widehat{F}(s)$ with at most one of its poles at the origin and the others in the open-left-half plane then $\lim_{t\,!\,\infty} f(t) = \lim_{s\,!\,0} s\widehat{F}(s)$ [27]. As such we will calculate the limits $\lim_{s\,!\,0} s\widehat{\chi}_x(s)$ and $\lim_{s\,!\,0} s\widehat{\chi}(s)$,

$$\lim_{t\,!\,\infty} \chi_x(t) = \lim_{s\,!\,0} s\widehat{\chi}_x(s) = \lim_{s\,!\,0} \frac{s}{ms^2 + \widehat{\Gamma}(s)s + \kappa} = 0, \tag{107}$$

$$\lim_{t\,!\,\infty} \chi(t) = \lim_{s\,!\,0} s\widehat{\chi}(s) = \lim_{s\,!\,0} \frac{s}{s}\frac{1}{ms^2 + \widehat{\Gamma}(s)s + \kappa} = \frac{1}{\kappa}. \tag{108}$$

Where, for the $\chi(t)$ limit we assume that $\lim_{s\,!\,0} \widehat{\Gamma}(s)s = \lim_{t\,!\,\infty} \Gamma(t)$ exists and equals 0. This is a reasonable assumption, as this essentially states that our memory cannot be infinitely long.

Using our results we can determine $\lim_{t_m\,!\,\infty} x(t)$,

$$\begin{aligned}
\lim_{t_m\,!\,\infty} x(t) &= \lim_{t_m\,!\,\infty} (x(t_m))\left(1 - \kappa \lim_{t_m\,!\,\infty}\left(\chi(t-t_m)\right)\right) + m \lim_{t_m\,!\,\infty}\left(\dot{x}(t_m)\chi_x(t-t_m)\right) \\
&\qquad + \lim_{t_m\,!\,\infty}\left(\int_{t_m}^{t} d\tau\ \chi_x(t-\tau)\eta(\tau)\right) \\
&= \lim_{t_m\,!\,\infty} (x(t_m))\left(1 - \kappa\frac{1}{\kappa}\right) + m \lim_{t_m\,!\,\infty} (\dot{x}(t_m))\cdot 0 + \lim_{t_m\,!\,\infty} \int_{t_m}^{t} d\tau\ \chi_x(t-\tau)\eta(\tau) \\
&= \lim_{t_m\,!\,\infty} \int_{t_m}^{t} d\tau\ \chi_x(t-\tau)\eta(\tau).
\end{aligned} \tag{109}$$

For $\lim_{t_m\,!\,\infty} \langle x(t)x(t^0)\rangle$ in Equation (104) we note that both the $(1-\kappa\chi(t-t_m))$ terms and the $\chi_x(t-t_m)$ terms will go to 0 in the limit. From the rest of the terms we get

$$\lim_{t_m\,!\,\infty} \langle x(t)x(t^0)\rangle = k_BT\left(\frac{1}{\kappa} + \frac{1}{\kappa} - \chi(|t-t^0|) - \kappa\frac{1}{\kappa}\frac{1}{\kappa} - m\cdot 0\cdot 0\right) = k_BT\left(\frac{1}{\kappa} - \chi(|t-t^0|)\right) \tag{110}$$

B.5 Comparison to Langevin equation

The delta function memory kernel is important, as this is the kernel that recovers the normal Langevin equation. As such, we should make sure that here the solution found in the previous subsections corresponds to the solution of the stochastic damped harmonic oscillator. We define the memory kernel

$$\Gamma(t) = 4\zeta\sqrt{\kappa m}\,\delta(t), \tag{111}$$

such that, plugging this into Equation (73) yields

$$\begin{aligned} m\ddot{x} &= -\int_{t_m}^{t} dt^0\, 4\zeta\sqrt{\kappa m}\ \delta(t-t^0)\dot{x}(t^0) - \kappa x(t) + \eta(t), \\ &= -2\zeta\sqrt{\kappa m}\,\dot{x}(t) - \kappa x(t) + \sigma\xi(t). \end{aligned} \tag{112}$$

Here $\xi(t)$ is a normalized white noise process, such that this is indeed the Langevin equation. The fluctuation-dissipation theorem is obeyed if $\sigma^2 = 4\zeta\sqrt{\kappa m}\,k_BT$. In this form the solution for $x(t)$ in the case $\zeta \neq 1$ is [9,17]

$$x(t) = \sqrt{\frac{\zeta k_BT}{\sqrt{\kappa m}(\zeta^2-1)}}\int_1^t d\tau\ \xi(\tau)\left(e^{\ \sqrt{\frac{\kappa}{m}}(\zeta\ \ \sqrt{\zeta^2\ \ 1})(t\ \ \tau)} - e^{\ \sqrt{\frac{\kappa}{m}}(\zeta+\sqrt{\zeta^2\ \ 1})(t\ \ \tau)}\right), \tag{113}$$

while the auto-correlation is found as

$$\langle x(t)x(t^0)\rangle = \frac{k_BT}{2\kappa\sqrt{\zeta^2-1}}\left(\frac{e^{\ \sqrt{\frac{\kappa}{m}}(\zeta\ \ \sqrt{\zeta^2\ \ 1})jt\ \ t^0j}}{\zeta-\sqrt{\zeta^2-1}} - \frac{e^{\ \sqrt{\frac{\kappa}{m}}(\zeta+\sqrt{\zeta^2\ \ 1})jt\ \ t^0j}}{\zeta+\sqrt{\zeta^2-1}}\right). \tag{114}$$

To use the framework established in the previous section we determine the Laplace transform of the memory kernel

$$\widehat{\Gamma}(s) = L_s\left\{4\zeta\sqrt{\kappa m}\,\delta(t)\right\} = 4\zeta\sqrt{\kappa m}\int_0^1 dt\ \delta(t)e^{\ \ st} = 2\zeta\sqrt{\kappa m}. \tag{115}$$

We thus find the Laplace transform of the position susceptibility, as introduced in Equation (86), as

$$\widehat{\chi}_x(s) = \frac{1}{ms^2+2\zeta\sqrt{\kappa m}\,s+\kappa} = \frac{1}{m(s+\sqrt{\frac{\kappa}{m}}\left(\zeta-\sqrt{\zeta^2-1}\right))(s+\sqrt{\frac{\kappa}{m}}\left(\zeta+\sqrt{\zeta^2-1}\right))}. \tag{116}$$

To find the position auto-correlation function we are interested in the function $\chi(t)$, whose Laplace transform is $\widehat{\chi}(s) = \widehat{\chi}_x(s)/s$. We have

$$\widehat{\chi}(s) = \frac{1}{ms}\frac{1}{s+\sqrt{\frac{\kappa}{m}}\left(\zeta-\sqrt{\zeta^2-1}\right)}\frac{1}{s+\sqrt{\frac{\kappa}{m}}\left(\zeta+\sqrt{\zeta^2-1}\right)}, \tag{117}$$

which is in a form that we recognize as a multiplication of the Laplace transforms of a constant (Equation (80) for $t_m = 0$) and two exponential decays. We find

$$\begin{aligned} \chi(t) &= \int_0^t dt^0\ \frac{1}{m}\int_0^{t^0} d\tau e^{\ \sqrt{\frac{\kappa}{m}}\left(\zeta\ \ \sqrt{\zeta^2\ \ 1}\right)(t^0\ \ \tau)}e^{\ \sqrt{\frac{\kappa}{m}}\left(\zeta+\sqrt{\zeta^2\ \ 1}\right)\tau} \\ &= \frac{1}{m}\int_0^t dt^0\ e^{\ \sqrt{\frac{\kappa}{m}}\left(\zeta\ \ \sqrt{\zeta^2\ \ 1}\right)t^0}\int_0^{t^0} d\tau e^{\ 2\sqrt{\frac{\kappa}{m}}\sqrt{\zeta^2\ \ 1}\,\tau} \\ &= \frac{1}{m}\int_0^t dt^0\ e^{\sqrt{\frac{\kappa}{m}}\left(\ \zeta+\sqrt{\zeta^2\ \ 1}\right)t^0}\frac{e^{\ 2\sqrt{\frac{\kappa}{m}}\sqrt{\zeta^2\ \ 1}\,t^0}-1}{-2\sqrt{\frac{\kappa}{m}}\sqrt{\zeta^2-1}} \\ &= \frac{1}{2\sqrt{\kappa m}\sqrt{\zeta^2-1}}\int_0^t dt^0\ e^{\ \sqrt{\frac{\kappa}{m}}\left(\zeta\ \ \sqrt{\zeta^2\ \ 1}\right)t^0} - e^{\ \sqrt{\frac{\kappa}{m}}\left(\zeta+\sqrt{\zeta^2\ \ 1}\right)t^0} \\ &= \frac{1}{2\sqrt{\kappa m}\sqrt{\zeta^2-1}}\left(\frac{e^{\ \sqrt{\frac{\kappa}{m}}\left(\zeta\ \ \sqrt{\zeta^2\ \ 1}\right)t}-1}{-\sqrt{\frac{\kappa}{m}}\left(\zeta-\sqrt{\zeta^2-1}\right)} - \frac{e^{\ \sqrt{\frac{\kappa}{m}}\left(\zeta+\sqrt{\zeta^2\ \ 1}\right)t}-1}{-\sqrt{\frac{\kappa}{m}}\left(\zeta+\sqrt{\zeta^2-1}\right)}\right) \\ &= \frac{1}{2\kappa\sqrt{\zeta^2-1}}\left(\frac{1}{\zeta-\sqrt{\zeta^2-1}} - \frac{1}{\zeta+\sqrt{\zeta^2-1}} - \frac{e^{\ \sqrt{\frac{\kappa}{m}}\left(\zeta\ \ \sqrt{\zeta^2\ \ 1}\right)t}}{\zeta-\sqrt{\zeta^2-1}} + \frac{e^{\ \sqrt{\frac{\kappa}{m}}\left(\zeta+\sqrt{\zeta^2\ \ 1}\right)t}}{\zeta+\sqrt{\zeta^2-1}}\right) \\ &= \frac{1}{2\kappa\sqrt{\zeta^2-1}}\left(2\sqrt{\zeta^2-1} + \frac{e^{\ \sqrt{\frac{\kappa}{m}}\left(\zeta+\sqrt{\zeta^2\ \ 1}\right)t}}{\zeta+\sqrt{\zeta^2-1}} - \frac{e^{\ \sqrt{\frac{\kappa}{m}}\left(\zeta\ \ \sqrt{\zeta^2\ \ 1}\right)t}}{\zeta-\sqrt{\zeta^2-1}}\right). \end{aligned} \tag{118}$$

We plug this into Equation (110) and find

$$\begin{aligned}\langle x(t)x(t^0)\rangle &= k_BT\left(\frac{1}{\kappa}-\frac{1}{2\kappa\sqrt{\zeta^2-1}}\left(2\sqrt{\zeta^2-1}+\frac{e^{\sqrt{\frac{\kappa}{m}}\left(\zeta+\sqrt{\zeta^2-1}\right)jt\ t^0j}}{\zeta+\sqrt{\zeta^2-1}}-\frac{e^{\sqrt{\frac{\kappa}{m}}\left(\zeta-\sqrt{\zeta^2-1}\right)jt\ t^0j}}{\zeta-\sqrt{\zeta^2-1}}\right)\right)\\ &= k_BT\left(\frac{1}{\kappa}-\frac{1}{\kappa}-\frac{1}{2\kappa\sqrt{\zeta^2-1}}\left(\frac{e^{\sqrt{\frac{\kappa}{m}}\left(\zeta+\sqrt{\zeta^2-1}\right)jt\ t^0j}}{\zeta+\sqrt{\zeta^2-1}}-\frac{e^{\sqrt{\frac{\kappa}{m}}\left(\zeta-\sqrt{\zeta^2-1}\right)jt\ t^0j}}{\zeta-\sqrt{\zeta^2-1}}\right)\right)\\ &= \frac{k_BT}{2\kappa\sqrt{\zeta^2-1}}\left(\frac{e^{\sqrt{\frac{\kappa}{m}}\left(\zeta-\sqrt{\zeta^2-1}\right)jt\ t^0j}}{\zeta-\sqrt{\zeta^2-1}}-\frac{e^{\sqrt{\frac{\kappa}{m}}\left(\zeta+\sqrt{\zeta^2-1}\right)jt\ t^0j}}{\zeta+\sqrt{\zeta^2-1}}\right),\end{aligned} \tag{119}$$

which is indeed exactly the same as for the normal Langevin equation in Equation (114). As such, we conclude that the framework established in Section B allows for determining the auto-correlation function needed to determine the Gaussian information bottleneck for generalized Langevin equations. The information bottleneck in this context is discussed in Ref. [9].

C Fractional kernel

Memory kernels of type $t^{\ \lambda}$ turn out to be rather difficult to solve, as the inverse Laplace transform of $\chi(s)$ yields so-called Mittag-Leffler functions, which are infinite sums. Thus evaluating the solutions will take significant amounts of time. We follow a paper by Viñales and Despósito to present the correlation function [28]. We define a memory kernel as

$$\Gamma(t) = \frac{B}{\Gamma(1-\lambda)}t^{\ \lambda}, \tag{120}$$

for $0<\lambda<1$. Here $\Gamma(x)$ is the gamma function. The fact λ cannot be an integer is due to the gamma function not existing for non-positive integers[1]. The Laplace transform of the memory kernel will now be

$$\widehat{\Gamma}(s) = \frac{B}{\Gamma(1-\lambda)}\int_0^\infty dt\, e^{\ st}t^{\ \lambda} = \frac{B}{\Gamma(1-\lambda)}\int_0^\infty \frac{du}{s}e^{\ u}\left(\frac{u}{s}\right)^{\ \lambda} = \frac{B}{s^{1\ \lambda}}\frac{\int_0^\infty du\, e^{\ u}u^{(1\ \lambda)\ 1}}{\Gamma(1-\lambda)} = Bs^{\lambda\ \ 1}, \tag{121}$$

where we use the definition of the gamma function $\Gamma(z)=\int_0^\infty dt\, t^{z\ \ 1}e^{\ t}$. We then find the Laplace transform of the position susceptibility as

$$\widehat{\chi}_x(s) = \frac{1}{ms^2+Bs^\lambda+\kappa} \tag{122}$$

The Laplace transform of this is found using the theory of fractional integration and differentiation. In particular the solution to this three term equation can be found in Section 5.4 in a book by Igor Podlubny [29]. We find

$$\chi_x(t) = \frac{1}{m}\sum_{k=0}^{\infty}\frac{(-1)^k}{k!}\left(\frac{\kappa}{m}\right)^k t^{2k+1}E_{2\ \lambda,2+\lambda k}^{(k)}\left(-\frac{B}{m}t^{2\ \ \lambda}\right), \tag{123}$$

where $E_{\alpha,\beta}^{(k)}$ is the k^{th} derivative of the Mittag-Leffler function in two parameters. The Mittag-Leffler function is defined as [29]

$$E_{\alpha,\beta}(z) = \sum_{k=0}^{\infty}\frac{z^k}{\Gamma(\alpha k+\beta)}, \quad \alpha>0,\ \beta>0. \tag{124}$$

Such that the k^{th} derivative can be found by differentiating each term separately [29]

$$E_{\alpha,\beta}^{(k)}(z) = \sum_{j=0}^{\infty}\frac{(j+k)!\,z^j}{j!\,\Gamma(\alpha(j+k)+\beta)}. \tag{125}$$

[1]Note that we cannot get around this restriction by defining $\Gamma(t)=Bt^{\ \lambda}$, as in this case the Laplace transform would introduce a factor $\Gamma(1-\lambda)$, as shown in equation (121)

Before inserting the definitions we first concern ourselves with integrating Equation (123). We note an useful derivative given in Section 1.2.3 in the book by Podlubny [29], adapted for a non-fractional derivative,

$$\frac{d}{dt}t^{\alpha k+\beta}\ {}^{1}E^{(k)}_{\alpha,\beta}(at^{\alpha}) = t^{\alpha k+\beta}\ {}^{2}E^{(k)}_{\alpha,\beta}\ {}_{1}(at^{\alpha}). \tag{126}$$

We identify, by comparing Equation (123) to the right-hand-side of the above, $\alpha = 2-\lambda$, $\beta = 3+\lambda k$, and $a = -\frac{B}{m}$, such that indeed $\alpha k+\beta-2 = (2-\lambda)k+3+\lambda k-2 = 2k+1$. Thus, we can use the left-hand-side as an antiderivative in Equation (92) to find

$$\begin{aligned}
\chi(t) &= \int_0^t \chi_x(\tau)d\tau = \frac{1}{m}\sum_{k=0}^{1}\frac{(-1)^k}{k!}\left(\frac{\kappa}{m}\right)^k\left[\tau^{2k+2}E^{(k)}_{2\ \lambda,3+\lambda k}\left(-\frac{B}{m}\tau^{2\ \ \lambda}\right)\right]_0^t \\
&= \frac{1}{m}\sum_{k=0}^{1}\frac{(-1)^k}{k!}\left(\frac{\kappa}{m}\right)^k\left(t^{2k+2}E^{(k)}_{2\ \lambda,3+\lambda k}\left(-\frac{B}{m}\tau^{2\ \ \lambda}\right) - 0\cdot E^{(k)}_{2\ \lambda,3+\lambda k}(0)\right) \\
&= \frac{1}{m}\sum_{k=0}^{1}\frac{(-1)^k}{k!}\left(\frac{\kappa}{m}\right)^k t^{2k+2}E^{(k)}_{2\ \lambda,3+\lambda k}\left(-\frac{B}{m}\tau^{2\ \ \lambda}\right).
\end{aligned}$$

We thus find $\chi(t)$ as [28]

$$\chi(t) = \frac{1}{m}\sum_{k=0}^{1}\frac{(-1)^k}{k!}\left(\frac{\kappa}{m}\right)^k t^{2k+2}\sum_{j=0}^{1}\frac{(j+k)!\ \left(-\frac{B}{m}t^{2\ \ \lambda}\right)^j}{j!\,\Gamma\big((2-\lambda)j+2k+3\big)}, \tag{127}$$

which leads to the correlation function

$$\langle x(t)x(t^{\emptyset})\rangle = k_BT\left(\frac{1}{\kappa} - \frac{1}{m}\sum_{k=0}^{1}\frac{(-1)^k}{k!}\left(\frac{\kappa}{m}\right)^k|t-t^{\emptyset}|^{2k+2}\sum_{j=0}^{1}\frac{(j+k)!\ \left(-\frac{B}{m}|t-t^{\emptyset}|^{2\ \ \lambda}\right)^j}{j!\,\Gamma\big((2-\lambda)j+2k+3\big)}\right). \tag{128}$$

This correlation function has an unfortunate structure. While the inner infinite sum (over j) can be written as a closed function, the same is not true for the outer sum (over k). Moreover, for this outer sum the absolute value of the terms in the sum at first grows exponentially before shrinking exponentially. While neighboring terms have different signs in the sum, this "bulge" does not sum to 0, such that we need to consider a large amount of terms in that sum to a high precision to be able to find the correct result.

D References

[1] F. Jacob and J. Monod (1961), "Genetic regulatory mechanisms in the synthesis of proteins," *Journal of Molecular Biology*, **3**(3):318–356, ISSN 00222836, doi:10.1016/S0022-2836(61)80072-7.
URL https://linkinghub.elsevier.com/retrieve/pii/S0022283661800727

[2] M. Bauer, M. D. Petkova, T. Gregor, E. F. Wieschaus, and W. Bialek (2021), "Trading bits in the readout from a genetic network," *Proceedings of the National Academy of Sciences*, **118**(46), ISSN 0027-8424, doi:10.1073/pnas.2109011118, arXiv: 2012.15817.
URL https://pnas.org/doi/full/10.1073/pnas.2109011118

[3] A. J. Sederberg, J. N. MacLean, and S. E. Palmer (2018), "Learning to make external sensory stimulus predictions using internal correlations in populations of neurons," *Proceedings of the National Academy of Sciences*, **115**(5):1105–1110, ISSN 0027-8424, doi:10.1073/pnas.1710779115.
URL https://pnas.org/doi/full/10.1073/pnas.1710779115

[4] K. Nakamura and T. J. Kobayashi (2021), "Connection between the Bacterial Chemotactic Network and Optimal Filtering," *Physical Review Letters*, **126**(12):128102, ISSN 0031-9007, 1079-7114, doi:10.1103/PhysRevLett.126.128102.
URL https://link.aps.org/doi/10.1103/PhysRevLett.126.128102

[5] C. C. Govern and P. R. ten Wolde (2014), "Optimal resource allocation in cellular sensing systems," *Proceedings of the National Academy of Sciences*, **111**(49):17486–17491, ISSN 0027-8424, doi:10.1073/pnas.1411524111.
URL https://pnas.org/doi/full/10.1073/pnas.1411524111

[6] S. E. Palmer, O. Marre, M. J. Berry, and W. Bialek (2015), "Predictive information in a sensory population," *Proceedings of the National Academy of Sciences*, **112**(22):6908–6913, ISSN 0027-8424, doi:10.1073/pnas.1506855112, arXiv: 1307.0225.
URL https://pnas.org/doi/full/10.1073/pnas.1506855112

[7] V. Sachdeva, T. Mora, A. M. Walczak, and S. E. Palmer (2021), "Optimal prediction with resource constraints using the information bottleneck," *PLOS Computational Biology*, **17**(3):e1008743, ISSN 1553-7358, doi:10.1371/journal.pcbi.1008743, publisher: Public Library of Science.
URL https://dx.plos.org/10.1371/journal.pcbi.1008743

[8] J. M. Goedhart (2019), *Measuring the Past, Predicting the Future: The Push-Pull Network*, Master's thesis, University of Amsterdam.

[9] L. Slim (2020), *Optimal Cellular Prediction and the Push-Pull Network*, Master's thesis, University of Amsterdam.

[10] C. E. Shannon (1948), "A Mathematical Theory of Communication," *Bell System Technical Journal*, **27**(4):623–656, ISSN 00058580, doi:10.1002/j.1538-7305.1948.tb00917.x.
URL https://ieeexplore.ieee.org/document/6773067

[11] N. Tishby, F. C. Pereira, and W. Bialek (1999), "The Information Bottleneck Method," in "The 37th annual Allerton Conference on Communication, Control, and Computing," (pages 368–377), doi:10.1108/eb040537, iSSN: 03055728.

[12] M. P. Langevin (1908), "Sur la théorie du mouvement brownien," *Comptes rendus hebdomadaires des seances de l'Academie des sciences*, **146**:530–533.
URL https://gallica.bnf.fr/ark:/12148/bpt6k3100t/f530.item

[13] D. S. Lemons and A. Gythiel (1997), "Paul Langevin's 1908 paper "On the Theory of Brownian Motion" ["Sur la théorie du mouvement brownien," C. R. Acad. Sci. (Paris) 146 , 530–533 (1908)]," *American Journal of Physics*, **65**(11):1079–1081, ISSN 0002-9505, doi:10.1119/1.18725.
URL http://aapt.scitation.org/doi/10.1119/1.18725

[14] R. Kubo (1966), "The fluctuation-dissipation theorem," *Reports on Progress in Physics*, **29**(1):306, ISSN 00344885, doi:10.1088/0034-4885/29/1/306.
URL https://iopscience.iop.org/article/10.1088/0034-4885/29/1/306

[15] H. Mori (1965), "Transport, Collective Motion, and Brownian Motion," *Progress of Theoretical Physics*, **33**(3):423–455, ISSN 0033-068X, doi:10.1143/PTP.33.423.
URL https://academic.oup.com/ptp/article-lookup/doi/10.1143/PTP.33.423

[16] I. Di Terlizzi, F. Ritort, and M. Baiesi (2020), "Explicit Solution of the Generalised Langevin Equation," *Journal of Statistical Physics*, **181**(5):1609–1635, ISSN 0022-4715, doi:10.1007/s10955-020-02639-4, arXiv: 2005.04012 Publisher: Springer.
URL https://link.springer.com/10.1007/s10955-020-02639-4

[17] S. F. Nørrelykke and H. Flyvbjerg (2011), "Harmonic oscillator in heat bath: Exact simulation of time-lapse-recorded data and exact analytical benchmark statistics," *Physical Review E*, **83**(4):041103, ISSN 1539-3755, doi:10.1103/PhysRevE.83.041103, arXiv: 1102.0524.
URL https://link.aps.org/doi/10.1103/PhysRevE.83.041103

[18] J. Fricks, L. Yao, T. C. Elston, and M. G. Forest (2009), "Time-Domain Methods for Diffusive Transport in Soft Matter," *SIAM Journal on Applied Mathematics*, **69**(5):1277–1308, ISSN 0036-1399, doi:10.1137/070695186.
URL http://epubs.siam.org/doi/10.1137/070695186

[19] M. Lysy, N. S. Pillai, D. B. Hill, M. G. Forest, J. W. R. Mellnik, P. A. Vasquez, and S. A. McKinley (2016), "Model Comparison and Assessment for Single Particle Tracking in Biological Fluids," *Journal of the American Statistical Association*, **111**(516):1413–1426, ISSN 0162-1459, doi:10.1080/01621459.2016.1158716, arXiv: 1407.5962 Publisher: Taylor & Francis.
URL https://doi.org/10.1080/01621459.2016.1158716

[20] W. Lee (2018), "Generalized Langevin equation and the linear regression model with memory," *Physical Review E*, **98**(2):022137, ISSN 2470-0045, doi:10.1103/PhysRevE.98.022137, publisher: American

Physical Society.
URL https://link.aps.org/doi/10.1103/PhysRevE.98.022137

[21] S. A. McKinley, L. Yao, and M. G. Forest (2009), "Transient anomalous diffusion of tracer particles in soft matter," *Journal of Rheology*, **53**(6):1487–1506, ISSN 0148-6055, 1520-8516, doi:10.1122/1.3238546.
URL http://sor.scitation.org/doi/10.1122/1.3238546

[22] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss (2004), "Information bottleneck for Gaussian variables," *Journal of Machine Learning Research*, **6**:165–188, ISSN 1533-7928.
URL https://www.jmlr.org/papers/v6/chechik05a.html

[23] T. M. Cover and J. A. Thomas (2006), *Elements of information theory*, Wiley India, New Delhi, ISBN 978-81-265-4194-2.

[24] R. F. Fox (1977), "The generalized Langevin equation with Gaussian fluctuations," *Journal of Mathematical Physics*, **18**(12):2331–2335, ISSN 0022-2488, doi:10.1063/1.523242.
URL http://aip.scitation.org/doi/10.1063/1.523242

[25] N. Wiener (1949), *Extrapolation, Interpolation and Smoothing of Stationary Time Series With Engineering Applications*, The Technology Press of the Massachusetts Institute of Technology and John Wiley & Sons, Inc., New York, ISBN 978-0-262-25719-0.

[26] H. Lütkepohl (2005), *New Introduction to Multiple Time Series Analysis*, Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN 978-3-540-27752-1, doi:10.1007/978-3-540-27752-1.
URL https://books.google.nl/books?id=COUFCAAAQBAJ&lpg=PR4&ots=wG6BisZIEs&dq=multipletimeseriesanalysis&lr&hl=nl&pg=PA26#v=onepage&q&f=false

[27] J. Chen, K. H. Lundberg, D. E. Davison, and D. S. Bernstein (2007), "The Final Value Theorem Revisited - Infinite Limits and Irrational Functions," *IEEE Control Systems*, **27**(3):97–99, ISSN 1066-033X, doi:10.1109/MCS.2007.365008.
URL https://ieeexplore.ieee.org/document/4213171/

[28] A. D. Viñales and M. A. Despósito (2006), "Anomalous diffusion: Exact solution of the generalized Langevin equation for harmonically bounded particle," *Physical Review E*, **73**(1):016111, ISSN 1539-3755, doi:10.1103/PhysRevE.73.016111.
URL https://link.aps.org/doi/10.1103/PhysRevE.73.016111

[29] I. Podlubny (1999), *Fractional Di erential Equations : An Introduction to Fractional Derivatives, Fractional Di erential Equations, to Methods of Their Solution and Some of Their Applications*, Academic Press, San Diego, ISBN 0-12 558840-2, series Title: Mathematics in Science and Engineering.
URL https://web-p-ebscohost-com.proxy.library.uu.nl/ehost/detail?sid=82d1bd9b-1860-406a-91ba-87fd3bf073b0@redis&vid=0&format=EB&lpid=lp_iv&rid=0#AN=227563&db=nlebk