

A Mori-Zwanzig Formalism

In this section, we provide a detailed derivation of the Mori-Zwanzig formalism.

The evolution of system observations. We provide two complementary views of a dynamical system: the microscopic Cauchy problem and the macroscopic MZ formalism.

Let $\Phi \in \mathbb{R}^n$ be the full *phase-state* of the system, evolving under the autonomous ODE

$$\frac{d\Phi(t)}{dt} = S(\Phi(t)), \quad \Phi(0) = x_0, \quad (3)$$

where $S : \mathcal{M} \rightarrow \mathbb{R}^n$ is C^1 (hence locally Lipschitz). By the Picard-Lindelöf [9] theorem, for each initial condition $x_0 \in \mathcal{M}$ there is a unique solution $\Phi(t)$ on the interval $T \subseteq \mathbb{R}$. This defines the flow

$$\Phi_t : \mathcal{M} \rightarrow \mathcal{M}, \quad \Phi_t(x_0) = \Phi(t).$$

Define the measure space $(\mathcal{M}, \mathcal{F}, \mu)$ with the phase-state manifold \mathcal{M} , a σ -algebra \mathcal{F} (typically the Borel σ -algebra $\mathcal{B}(\mathcal{M})$), and a finite, flow-invariant probability measure μ . The Hilbert space of observable functions (i.e. *observables*) is defined as $\mathcal{H} = L^2(\mathcal{M}, \mathcal{F}, \mu)$ consisting of real-valued, square-integrable functions $g : \mathcal{M} \rightarrow \mathbb{R}$, with inner product

$$\langle g, h \rangle = \int_{\mathcal{M}} g(x)h(x)d\mu(x).$$

Note that since $\mathcal{M} \subset \mathbb{R}^n$ is a separable metric space with finite measure, then \mathcal{H} is a separable Hilbert space. In addition, these scalar-valued observables may be arbitrary non-linear functions of the phase-space variable Φ .

Each observable $g \in \mathcal{H}$ evolves under the Koopman semigroup $\{U^t\}_{t \geq 0}$ via $U^t g(x) = g(\Phi_t(x))$ and the Liouville operator $\mathcal{L} : \mathcal{H} \rightarrow \mathcal{H}$ is the infinitesimal generator, defined by

$$(\mathcal{L}g)(x) = \lim_{t \rightarrow 0} \frac{g(\Phi_t(x)) - g(x)}{t} = \frac{d}{dt} g(\Phi_t(x)) \Big|_{t=0},$$

so that formally $U^t = e^{t\mathcal{L}}$.

The decomposition into resolved and unresolved observables. Central to the MZ formalism is the orthogonal decomposition of \mathcal{H} :

$$\mathcal{H} = \mathcal{V} \oplus \mathcal{V}^\perp = \text{ran}(P) \oplus \text{ran}(I - P)$$

where P projects onto the *resolved* subspace (i.e., the observables we retain) and $I - P$ onto the unresolved subspace. The choice of P is the sole degree of freedom in the Mori-Zwanzig formalism, determining which components of the full dynamics are treated as resolved.

Mori and Zwanzig offer differing canonical projections: Mori's projection selects \mathcal{V} as the finite-dimensional span of observables $\{g_i\}$, projecting via inner products [55]; Zwanzig's projection takes \mathcal{V} to be the (typically infinite-dimensional) subspace of functions measurable with respect to a σ -algebra \mathcal{G} , projecting via the conditional expectation $Pg = \mathbb{E}[g|\mathcal{G}]$ [87].

The derivation of the GLE. The instantaneous evolution of g is given by

$$\frac{d}{dt} e^{t\mathcal{L}} g(0) = \mathcal{L} e^{t\mathcal{L}} g(0),$$

which can be decomposed into its two projected dynamics yielding two coupled equations

$$\begin{aligned} \frac{d}{dt} P e^{t\mathcal{L}} g(0) &= P \mathcal{L} P e^{t\mathcal{L}} g(0) + P \mathcal{L} Q e^{t\mathcal{L}} g(0), \\ \frac{d}{dt} Q e^{t\mathcal{L}} g(0) &= Q \mathcal{L} Q e^{t\mathcal{L}} g(0) + Q \mathcal{L} P e^{t\mathcal{L}} g(0). \end{aligned}$$

We rewrite the second equation for $v(t) = Q e^{t\mathcal{L}} g(0)$ where $A(t) = Q e^{t\mathcal{L}} g(0)$ and $F(t) = Q \mathcal{L} P e^{t\mathcal{L}} g(0)$,

$$\frac{d}{dt} v(t) = A(t)v(t) + F(t).$$

The solution is given by Dyson’s identity

$$v(t) = e^{tA}v(0) + \int_0^t e^{(t-s)A}F(s)ds.$$

Notice that $v(0) = Qg(0)$. Substituting for v, A, F , we have

$$Qe^{t\mathcal{L}}g(0) = e^{tQ\mathcal{L}}Qg(0) + \int_0^t e^{(t-s)Q\mathcal{L}}Pe^{s\mathcal{L}}g(0)ds = e^{tQ\mathcal{L}}g(0) + \int_0^t e^{(t-s)Q\mathcal{L}}Pg(s)ds.$$

The GLE results from substituting the prior result into the dynamics for $\frac{d}{dt}Pg(t)$

$$\frac{\partial}{\partial t}Pg(t) = P\mathcal{L}Pg(t) + \int_0^t P\mathcal{L}e^{(t-s)Q\mathcal{L}}Q\mathcal{L}Pg(s)ds + P\mathcal{L}e^{tQ\mathcal{L}}Qg(0).$$

The connection to Koopman operator theory. The Koopman operator $\mathcal{K}^t : \mathcal{H} \rightarrow \mathcal{H}$ is a bounded linear operator that evolves any observable $g \in \mathcal{H}$ along the flow $T \subset \mathbb{R}$ on the phase manifold

$$\mathcal{K}^t g(x_0) = g(T(x_0, t)).$$

Because \mathcal{H} is infinite dimensional, in practice one often restricts attention to a finite resolved subspace $\mathcal{V} = \text{Span}\{g^{(1)}, \dots, g^{(r)}\} \subset \mathcal{H}$ with orthogonal complement \mathcal{V}^\top .

The evolution of $\hat{g} \in \mathcal{V}$ in this reduced subspace, with restricted evolution operator $\hat{\mathcal{K}}$, accumulates an error term

$$\begin{aligned} \hat{g} \circ T &= \hat{\mathcal{K}}^t \hat{g} + r, \\ r &\in \mathcal{V}^\top. \end{aligned}$$

where $\hat{g} \in \mathcal{V}$. The residual r is the closure problem, which is addressed via the Mori–Zwanzig formalism by projecting onto \mathcal{V} while accounting for the influence of \mathcal{V}^\top .

B Theoretical Details

In this section, we provide proofs of the corresponding propositions from Section 3.

B.1 Assumptions

For completeness, we restate our assumptions below. In addition, we will provide some more context to the significance of these assumptions.

Assumption 3.1. (Differentiability of P_{μ_t}) Suppose the measure curve $t \rightarrow \mu_t$ is a C^1 curve in total-variation norm and write $\rho_t = \frac{d\mu_t}{d\mu_*} \in L^1(\mu_*)$, with $\dot{\rho}_t$ its time derivative in $L^1(\mu_*)$. Then the map

$$t \rightarrow P_{\mu_t} = \left[f \mapsto \mathbb{E}_{\mu_t}[f | \mathcal{G}] \right] \in \mathcal{B}(L^2(\mu_*), L^2(\mu_t))$$

is Fréchet-differentiable, and its derivative at time t is the bounded operator \hat{P}_{μ_t} .

This assumption is critical to ensuring that the GLE is well-posed. In practice it forces us to choose a feature-map basis whose dependence on t makes P_{μ_t} a smooth function of time—only then can the model reliably learn the evolving dynamics.

Assumption 3.2. (Support Coverage Assumption) Let $\tilde{\mu} = \sum_{t=1}^T \mu_t$. We require $\mu_* \ll \tilde{\mu}$ or equivalently $\text{supp}(\mu_*) \subseteq \bigcup_{t=1}^T \text{supp}(\mu_t)$.

This condition ensures that the projected dynamics P_{μ_t} can act on the entire latent state: there are no hidden modes in μ_* that fall completely outside the supports of the training measures. Equivalently, it removes any degrees of freedom from the latent state, so that our GLE truly governs all of the relevant latent dynamics.

B.2 Proofs

Proposition 3.1. (Generalized GLE on a Measure Bundle) Let $g(t)$ evolve under the Liouville operator \mathcal{L} on a fixed Hilbert space $\mathcal{H} = L^2(\mathcal{M}, \mathcal{F}, \mu_*)$. Let $P_{\mu_*} : \mathcal{H} \rightarrow \mathcal{V} \subset \mathcal{H}$ be an orthogonal projection onto $\mathcal{V} = L^2(\mathcal{M}, \mathcal{G}, \mu_*)$ with $\mathcal{G} \subset \mathcal{F}$. For a family of C^1 measures $\{\mu_t\}_{t \in [0, T]}$ let $P_{\mu_t} : \mathcal{V} \rightarrow \mathcal{V}_t$ be the corresponding family of projections defining a Hilbert bundle $\{\mathcal{V}_t\}_{t \in [0, t]}$ with $\mathcal{V}_t = L^2(\mathcal{M}, \mathcal{G}, \mu_t)$. The evolution of the resolved variable $P_{\mu_t} g(t)$ satisfies the following GLE

$$\frac{d}{dt} P_{\mu_t} g(t) = P_{\mu_t} \dot{P}_{\mu_t} Q_{\mu_t} g(t) + P_{\mu_t} \mathcal{L} P_{\mu_*} g(t) + \int_0^t P_{\mu_t} \mathcal{L} e^{(t-s)Q_{\mu_*}} \mathcal{L} P_{\mu_*} g(s) ds + P_{\mu_t} e^{tQ_{\mu_*}} \mathcal{L} g(0).$$

Proof. By Assumption 3.1 P_{μ_t} is differentiable, so that the GLE is given by chain rule as

$$\frac{d}{dt} (P_{\mu_t} g(t)) = \dot{P}_{\mu_t} g(t) + P_{\mu_t} \frac{d}{dt} g(t) = \dot{P}_{\mu_t} g(t) + P_{\mu_t} \mathcal{L} g(t).$$

Let \mathcal{H} and \mathcal{V} be decomposed as $\mathcal{H} = \mathcal{V} \oplus \mathcal{V}^\perp = \text{ran}(P_{\mu_*}) \oplus \text{ran}(Q_{\mu_*})$, and $\mathcal{V} = \text{ran}(P_{\mu_t}) \oplus \text{ran}(Q_{\mu_t})$ for all t . First, using the decomposition of \mathcal{V} , we rewrite

$$\dot{P}_{\mu_t} g(t) = \dot{P}_{\mu_t} P_{\mu_t} g(t) + \dot{P}_{\mu_t} Q_{\mu_t} g(t) = \dot{P}_{\mu_t} Q_{\mu_t} g(t) = P_{\mu_t} \dot{P}_{\mu_t} Q_{\mu_t} g(t)$$

using the identities in Section 2.

Inserting the fixed-time decomposition for $\mathcal{L} g(t)$, we see

$$P_{\mu_t} \mathcal{L} g(t) = P_{\mu_t} \mathcal{L} (P_{\mu_*} + Q_{\mu_*}) g(t)$$

hence

$$\frac{d}{dt} (P_{\mu_t} g(t)) = \dot{P}_{\mu_t} g(t) + P_{\mu_t} \mathcal{L} P_{\mu_*} g(t) + P_{\mu_t} \mathcal{L} Q_{\mu_*} g(t)$$

Finally, using Dyson's identity to solve for $v(t) = Q_{\mu_*} g(t)$ as in the standard MZ formalism, we find

$$\frac{d}{dt} P_{\mu_t} g(t) = P_{\mu_t} \dot{P}_{\mu_t} Q_{\mu_t} g(t) + P_{\mu_t} \mathcal{L} P_{\mu_*} g(t) + \int_0^t P_{\mu_t} \mathcal{L} e^{(t-s)Q_{\mu_*}} \mathcal{L} P_{\mu_*} g(s) ds + P_{\mu_t} e^{tQ_{\mu_*}} \mathcal{L} g(0).$$

□

Proposition 3.2. (Optimal Task Projector) Let Assumption 3.2 hold for $\tilde{\mu} = \frac{1}{T} \sum_{t=1}^T \mu_t$ and define

$$P_{\text{train}} = \underset{G \in \mathcal{G} - \text{measurable}}{\text{argmin}} \mathbb{E}_{\tilde{\mu}} \|y(t) - G(g(t))\|_2^2$$

where $y(t)$ is the target at time t . Then $P_{\text{train}} = P_{\mu_*}$.

Proof. The space of all \mathcal{G} -measurable functions is a closed subspace of $L^2(\mathcal{M}, \tilde{\mu})$. By the uniqueness of the projection operator, then the unique minimizer of

$$\underset{G \in \mathcal{G} - \text{measurable}}{\text{argmin}} \mathbb{E}_{\tilde{\mu}} \|y(t) - G(g(t))\|_2^2$$

is $G^*(g) = \mathbb{E}_{\tilde{\mu}} [y | G]$. Equivalently $P_{\text{train}} = P_{\tilde{\mu}}$ is the unique orthogonal projector in $L^2(\mathcal{M}, \tilde{\mu})$ onto the \mathcal{G} -measurable subspace.

By Assumption 3.2, then the conditional expectation operators $\tilde{\mu}$ and μ_* coincide almost everywhere. Concretely,

$$\mathbb{E}_{\tilde{\mu}} [y | G](x) = \mathbb{E}_{\mu_*} [y | G](x) \quad \text{for } \tilde{\mu}\text{-a.e. } x.$$

□

Proposition 3.3. (Vanishing Drift Under an Invariant Trivialization) Suppose the Radon-Nikodym densities satisfy

$$\rho_t(x) = \frac{d\mu_t}{d\mu_0} = \alpha(t),$$

i.e., each μ_t differs from μ_0 only by a scalar factor $\alpha > 0$ independent of x , then $\dot{P}_{\mu_t} = 0$.

Proof. For any $g \in L^2(\mathcal{M}, \mu_t)$, P_{μ_t} is defined by the requirement

$$\int_G f d\mu_t = \int_G (P_{\mu_t} f) d\mu_t \quad \text{for all measurable } G.$$

Since $\mu_t = \alpha(t)\mu_0$

$$\int_G f d\mu_t = \alpha(t) \int_G f d\mu_0, \quad \int_G (P_{\mu_t} f) d\mu_t = \alpha(t) \int_G (P_{\mu_t} f) d\mu_0$$

Therefore

$$\int_G f d\mu_0 = \int_G (P_{\mu_t} f) d\mu_0 \quad \text{for all measurable } G,$$

which by the uniqueness of the conditional-expectation operator in L^2 characterizes P_{μ_0} . We thus conclude that $P_{\mu_t} = P_{\mu_0}$ for all t , and as a result the time-derivative vanishes, i.e., $\dot{P}_{\mu_t} = 0$. \square

Proposition 3.4. (Time-dependent Bundle Delay-Coordinate Closure) *Let $K(t, s)$ be the memory kernel on \mathcal{V} described in Section 3.2, fix a delay $\tau > 0$ and integer $r \geq 1$, and assume*

$$\sup_{t \in [0, T]} \int_0^{t-r\tau} \|K(t, s)\|_{\text{op}} ds \leq \epsilon.$$

For $k = 1, \dots, r$ define

$$M_k(t) = \int_{t-k\tau}^{t-(k-1)\tau} K(t, s) ds.$$

Then for any bounded latent trajectory $h(s)$ one has

$$\left\| \int_0^t K(t, s) a(s) ds - \sum_{k=1}^r M_k(t) h(t - k\tau) \right\| \leq \epsilon \sup_{s \in [0, T]} \|h(s)\|.$$

In other words, the full memory integral is approximated up to $O(\epsilon)$ by an r -dimensional delay-coordinate embedding.

Proof. The stated assumption

$$\sup_{t \in [0, T]} \int_0^{t-r\tau} \|K(t, s)\|_{\text{op}} ds \leq \epsilon.$$

is an assumption on the tail of the memory operator. Functionally, it acts as a bound on the far past. In addition to the stated assumption, from Assumption 3.1 the resolved trajectory $h(t) = P_{\mu_t} g(t)$ is Lipschitz in t . In particular, there exists $L > 0$ so that

$$\|h(s) - h(t - k\tau)\| \leq L|s - (t - k\tau)| \leq L\tau, \quad \forall s \in [t - k\tau, t - (k-1)\tau].$$

Using this, we split the integral as

$$\int_0^t K(t, s) h(s) ds = \int_0^{t-r\tau} K(t, s) h(s) ds + \sum_{k=1}^r \int_{t-k\tau}^{t-(k-1)\tau} K(t, s) h(s) ds.$$

By the state assumption the first term is bounded by $\epsilon \sup_{s \in [0, T]} \|h(s)\|$.

Now for each k

$$\int_{t-k\tau}^{t-(k-1)\tau} K(t, s) h(s) ds = M_k(t) h(t - k\tau) + \int_{t-k\tau}^{t-(k-1)\tau} K(t, s) [h(s) - h(t - k\tau)] ds.$$

Using the stated assumption

$$\begin{aligned} & \left\| \int_{t-k\tau}^{t-(k-1)\tau} K(t, s) [h(s) - h(t - k\tau)] ds \right\| \\ & \leq \sup_{s \in [t-k\tau, t-(k-1)\tau]} \|h(s) - h(t - k\tau)\| \cdot \left\| \int_{t-k\tau}^{t-(k-1)\tau} K(t, s) ds \right\|_{\text{op}} \\ & \leq L\tau \int_{t-k\tau}^{t-(k-1)\tau} \|K(t, s)\|_{\text{op}} ds \end{aligned}$$

Under the finite integrability assumption let $\int_{t-r\tau}^t \|K(t, s)\|_{\text{op}} ds = M$. Taking the sum over $k = 1, \dots, r$ then

$$\sum_{k=1}^r L\tau \int_{t-k\tau}^{t-(k-1)\tau} \|K(t, s)\|_{\text{op}} ds = L\tau \int_{t-r\tau}^t \|K(t, s)\|_{\text{op}} ds = L\tau M$$

Now by choosing τ sufficiently small, i.e., by choosing the block size of the memory quadrature to be sufficiently small so that $L\tau M \leq \epsilon$ then the block error and the tail error each contribute at most $\epsilon \sup \|h\|$, yielding

$$\left\| \int_0^t K(t, s) a(s) ds - \sum_{k=1}^r M_k(t) h(t - k\tau) \right\| \leq 2\epsilon \sup_{s \in [0, T]} \|h(s)\|.$$

□

Corollary 3.1. (*Toroidal Latent Manifold*) Suppose we constrain each latent coordinate $h_i(t)$ to live on a circle of period L_i and we enforce that both the learned drift and memory-kernel parameters depend on h only through these periodic coordinates. Then the entire latent trajectory $h(t)$ evolves on the m -dimensional torus \mathbb{T}^m . As a result, the network can only represent—and learn—functions defined on this compact, boundary-free manifold.

Proof. By the assumption of periodicity then each of the MZ terms descent to well-defined maps on the quotient $\mathbb{R}^m / (L_1\mathbb{Z} \times \dots \times L_m\mathbb{Z})$, and the initial condition $h(0) \in S_{L_1}^1 \times \dots \times S_{L_m}^1$ uniquely determines a solution $h(t)$ that never leaves the torus.

Therefore any decoder $F : \mathbb{R}^m \rightarrow Y$ must descent to a well-defined map $\hat{F} : \mathbb{T}^m \rightarrow Y$, i.e., those maps that are periodic in each coordinate. □

C Methodological Details

C.1 Architectural Details

Neural Wave Field The Neural Wave Field maintains two coupled latent state $h_t \in \mathbb{R}^n$ and $\mu_t \in \mathbb{R}^n$, which evolve under a Mori-Zwanzig inspired network and an accompanying measure-update expert. At each time step t the raw input x_t is first embedded into the feature space as a ghost boundary point. That is, it is available to be uptaken by the memory kernel provided the gating mechanism allows it.

For this reason, the MZ-NET σ_{mem} and σ_{force} are critical for determining the amount of long history information to retain, and the amount of new information to incorporate into the memory state. Whether the information is ultimately taken into the latent state is governed by σ_{closure} . These signals jointly determine a convolutional kernel C_{h_t} and padded hidden state \tilde{h}_t for updating $h_{t+1} = C_{h_t} \star \tilde{h}_t$.

A measure-dynamics expert network D_μ determines the update for the measure between two time periods. This module enforces that μ_t remains a valid probability density via softmax with a large temperature of 100.

Given our assumptions on the conditional-expectation projections of P_{μ_t} , we train using the MSE loss across all tasks.

WaveRNN The WaveRNN architecture is most similar to the Neural Wave Field in its construction of a latent state. There are two particular differences in the approaches. First, the WaveRNN utilizes periodic boundary conditions which are a limiting factor as described by Corollary 3.1. Moreover, the architecture relies on a static decoder and encoder which forces the projection dynamics to be invariant. As a result, the architecture will be unable to achieve a minimal latent state representation. Furthermore, it will be prohibited from accurately learning the selective copy task.

Mamba The Mamba architecture is a state-of-the-art structured state-space model. It has achieved particular success in modeling long-range tasks. It has done so by balancing long-range and short range updates to the latent state.

Transformers The positional encoding-based (or replacement) transformers aim to use various methods to replace fixed positional encoding mechanisms with relative positional encoding mechanisms. These have shown strong results in memory tasks such as the copy task.

C.2 Task Details

Lorenz Attractor We simulate the Lorenz system

$$\dot{x} = \sigma(y - x), \quad \dot{y} = x(\rho - z) - y, \quad \dot{z} = xy - \beta z$$

with standard parameters $(\sigma, \rho, \beta) = (10, 28, 8/3)$ using a fourth-order Runge-Kutta integrator at step size $\Delta t = 0.01$. At each time step only the x -coordinate is provided as input; the models must reconstruct the full state (x_t, y_t, z_t) .

For all experiments, we use a training batch size of 128 and test using a batch size of 32. All batches are generated randomly to obtain the trajectory of 300 time-steps. The loss is only computed on the last 280 time-steps. For all models we use the Adam optimizer with a learning rate of 0.001 for 1000 batches.

For our comparisons, we use the following configurations. For WaveRNN [35], we use one channel, an identity activation, and a hidden dimension of 20 to have a more direct comparison to our model. The loss is mean squared error (MSE).

Copy For all experiments, we use a training batch size of 128 and test using a batch size of 50. All batches are generated randomly to obtain the sequence of 10 tokens to be memorized. We use $T = 20$, so the total sequence length is 30. The loss is only computed on the last 10 tokens; the intermediate outputs are not considered. That is, we only care about the model’s ability to reproduce the sequence of 10 tokens at the final 10 timesteps. For all models we use the Adam optimizer with a learning rate of 0.001 for 1000 batches.

For our comparisons, we use the following configurations. For WaveRNN [35], we use one channel and an identity activation to have a more direct comparison to our model. The loss is mean squared error (MSE). For Mamba and the transformer models, we use cross entropy loss, as they naturally output logits over the vocabulary size. We found that these models needed at least 2 layers to perform on the task, which we use in our experiments. For the transformers, we use a single attention head.

Selective Copy By randomizing token positions and focusing evaluation solely on the terminal outputs, this task highlights each model’s ability to selectively attend to and retain the correct information. Our architecture’s time-dependent projection and delay-coordinate closure enable it to isolate the N informative tokens with minimal overhead, even as memory capacity is constrained.

C.3 Assumptions Note

As a note on the practical implications of the assumptions made. When the size of the latent state is larger than the minimal representation but not large enough to trivialize the dynamics of the measure, then the additional degrees of freedom provide many non-unique and non-trivial solutions. In this case, we experience large standard deviations in the training loss between runs with differing initial conditions. In the case where memory is sufficiently large to trivialize the measure dynamics, the learning became significantly more consistent.

In addition, the continuity assumptions on the measure make it impossible to use the current framework to effectively learn a version of the copy task where the predicted output is required to be placed in order. However, on this task, we observe that the Mamba and transformer architectures perform exceptionally well.