Notes on Operator Learning

August 10, 2026

Contents

Chapter 1

Stochastic Calculus

For background on stochastic calculus and Langevin diffusions, see [Särkkä and Solin, 2019] and [Chewi, 2024].

We start with a diffusion process in $\mathbb{R}^d$. Let

$$d\boldsymbol{x}(t) = f(\boldsymbol{x}(t), t)\, dt + L(\boldsymbol{x}(t), t)\, d\boldsymbol{B}(t),$$

where $f : (\boldsymbol{x}, t) \mapsto \mathbb{R}^d$, $L : (\boldsymbol{x}, t) \mapsto \mathbb{R}^{d\times d}$, and $\boldsymbol{B}(t)$ is a standard d-dimensional Brownian motion. Throughout, the functions are assumed sufficiently regular so that the expressions below are well-defined.

1.1 Itô Formula and the Generator

The first object we need is the differential of an observable along the stochastic dynamics. This leads directly to the infinitesimal generator, which records the deterministic drift of the observable.

Proposition 1 (Itô formula for the observable)**.** *Let $\phi : (\boldsymbol{x}, t) \mapsto \mathbb{R}$ be sufficiently regular. Then*

$$\begin{aligned} d\phi(\boldsymbol{x}(t), t) = \Big[& \nabla_t \phi(\boldsymbol{x}(t), t) + \nabla_x \phi(\boldsymbol{x}(t), t)^\top f(\boldsymbol{x}(t), t) \\ & + \frac{1}{2} \operatorname{tr}\left\{ \left(\nabla_x \nabla_x^\top \phi(\boldsymbol{x}(t), t) \right) L(\boldsymbol{x}(t), t) L(\boldsymbol{x}(t), t)^\top \right\} \Big] dt \\ & + \nabla_x \phi(\boldsymbol{x}(t), t)^\top L(\boldsymbol{x}(t), t)\, d\boldsymbol{B}(t). \end{aligned}$$

Proof. By Itô's chain rule,

$$d\phi(\boldsymbol{x}(t), t) = \nabla_t \phi(\boldsymbol{x}(t), t)\, dt + \nabla_x \phi(\boldsymbol{x}(t), t)^\top d\boldsymbol{x}(t) + \frac{1}{2} d\boldsymbol{x}(t)^\top \nabla_x^2 \phi(\boldsymbol{x}(t), t)\, d\boldsymbol{x}(t).$$

Substituting

$$d\boldsymbol{x}(t) = f(\boldsymbol{x}(t), t)\, dt + L(\boldsymbol{x}(t), t)\, d\boldsymbol{B}(t)$$

and using the formal Itô algebra

$$d\boldsymbol{B}(t)\, dt = 0, \qquad dt\, d\boldsymbol{B}(t) = 0, \qquad d\boldsymbol{B}(t)\, d\boldsymbol{B}(t)^\top = dt,$$

gives the claimed expression. □

The generator is the operator that appears after the martingale part of Itô's formula vanishes under conditional expectation.

Definition 1 (Generator)**.** *The time-dependent infinitesimal generator $\mathcal{A}_s$ is defined by*

$$\mathcal{A}_s\phi \coloneqq \lim_{\delta\to 0} \frac{\mathbb{E}\left[\phi(\boldsymbol{x}(s+\delta), s+\delta) \mid \boldsymbol{x}(s)\right] - \phi(\boldsymbol{x}(s), s)}{\delta},$$

whenever this limit exists.

Proposition 2 (Expression for the generator)**.** *For the diffusion above,*

$$\mathcal{A}_s\phi = \nabla_s\phi(\boldsymbol{x}(s), s) + \nabla_x\phi(\boldsymbol{x}(s), s)^\top f(\boldsymbol{x}(s), s) + \frac{1}{2}\operatorname{tr}\left\{\left(\nabla_x\nabla_x^\top\phi(\boldsymbol{x}(s), s)\right) L(\boldsymbol{x}(s), s)L(\boldsymbol{x}(s), s)^\top\right\}.$$

Proof. Integrating Itô's formula from s to t, we get

$$\begin{aligned}\phi(\boldsymbol{x}(t), t) - \phi(\boldsymbol{x}(s), s) = \int_s^t \Bigg[&\nabla_r\phi(\boldsymbol{x}(r), r) + \nabla_x\phi(\boldsymbol{x}(r), r)^\top f(\boldsymbol{x}(r), r) \\ &+ \frac{1}{2}\operatorname{tr}\left\{\left(\nabla_x\nabla_x^\top\phi(\boldsymbol{x}(r), r)\right) L(\boldsymbol{x}(r), r)L(\boldsymbol{x}(r), r)^\top\right\}\Bigg] dr \\ &+ \int_s^t \nabla_x\phi(\boldsymbol{x}(r), r)^\top L(\boldsymbol{x}(r), r)\, d\boldsymbol{B}(r).\end{aligned}$$

The stochastic integral

$$M_t \coloneqq \int_0^t \nabla_x\phi(\boldsymbol{x}(r), r)^\top L(\boldsymbol{x}(r), r)\, d\boldsymbol{B}(r)$$

is a martingale under the usual square-integrability assumptions. Therefore

$$\mathbb{E}\left[M_t - M_s \mid \mathcal{F}_s\right] = 0.$$

Since $\boldsymbol{x}(s)$ is $\mathcal{F}_s$-measurable, the tower property implies

$$\mathbb{E}\left[M_t - M_s \mid \boldsymbol{x}(s)\right] = \mathbb{E}\left[\mathbb{E}\left[M_t - M_s \mid \mathcal{F}_s\right] \mid \boldsymbol{x}(s)\right] = 0.$$

Thus, for $t = s + \delta$,

$$\begin{aligned}&\mathbb{E}\left[\phi(\boldsymbol{x}(s+\delta), s+\delta) \mid \boldsymbol{x}(s)\right] - \phi(\boldsymbol{x}(s), s) \\ &= \Bigg[\nabla_s\phi(\boldsymbol{x}(s), s) + \nabla_x\phi(\boldsymbol{x}(s), s)^\top f(\boldsymbol{x}(s), s) \\ &\quad + \frac{1}{2}\operatorname{tr}\left\{\left(\nabla_x\nabla_x^\top\phi(\boldsymbol{x}(s), s)\right) L(\boldsymbol{x}(s), s)L(\boldsymbol{x}(s), s)^\top\right\}\Bigg]\delta + O(\delta^2).\end{aligned}$$

Dividing by δ and taking $\delta \to 0$ gives the result. □

1.2 Semigroups and Evolution Equations

The generator is naturally connected with the Markov semigroup. In this notation, we write

$$P_{s,t}(\phi) \coloneqq \mathbb{E}\left[\phi(\boldsymbol{x}(t), t) \mid \boldsymbol{x}(s)\right].$$

Then the generator can also be written as

$$\mathcal{A}_s\phi = \lim_{\delta\to 0} \frac{P_{s,s+\delta}(\phi) - P_{s,s}(\phi)}{\delta}.$$

Proposition 3 (Backward equation)**.** *For $s < t$,*

$$\frac{d}{dt}P_{s,t}(\phi) = P_{s,t}\mathcal{A}_t\phi.$$

Proof. Using the semigroup property,

$$\begin{aligned} P_{s,t}\mathcal{A}_t\phi &= P_{s,t}\lim_{\delta\to 0}\frac{P_{t,t+\delta}(\phi)-P_{t,t}(\phi)}{\delta} \\ &= \lim_{\delta\to 0}\frac{P_{s,t}P_{t,t+\delta}(\phi)-P_{s,t}P_{t,t}(\phi)}{\delta} \\ &= \lim_{\delta\to 0}\frac{P_{s,t+\delta}(\phi)-P_{s,t}(\phi)}{\delta} \\ &= \frac{d}{dt}P_{s,t}(\phi). \end{aligned}$$

□

In the time-homogeneous case,

$$P_{s,t}\phi = P_{0,t-s}\phi \coloneqq P_{t-s}\phi.$$

The generator no longer depends on the initial time:

$$\mathcal{A}\phi \coloneqq \lim_{\delta\to 0}\frac{P_\delta(\phi)-P_0(\phi)}{\delta}.$$

Proposition 4 (Commutation with the semigroup)**.** *In the time-homogeneous case,*

$$P_t\mathcal{A}\phi = \mathcal{A}P_t\phi = \frac{d}{dt}P_t\phi.$$

This is the Kolmogorov backward equation.

Proof. We compute directly:

$$\begin{aligned} P_t\mathcal{A}\phi &= \lim_{\delta\to 0}P_t\frac{P_\delta(\phi)-P_0(\phi)}{\delta} \\ &= \lim_{\delta\to 0}\frac{P_tP_\delta(\phi)-P_tP_0(\phi)}{\delta} \\ &= \lim_{\delta\to 0}\frac{P_{t+\delta}(\phi)-P_t(\phi)}{\delta} \\ &= \frac{d}{dt}P_t\phi. \end{aligned}$$

Similarly,

$$\mathcal{A}P_t\phi = \lim_{\delta\to 0}\frac{P_\delta(P_t(\phi))-P_0(P_t(\phi))}{\delta} = \frac{d}{dt}P_t\phi.$$

□

Formally, this resembles a homogeneous linear ODE. Therefore,

$$P_t\phi = e^{t\mathcal{A}}P_0\phi = e^{t\mathcal{A}}\phi.$$

In this sense, knowing the generator determines the dynamics.

1.3 Forward Equation and Adjoint Generator

There is also a dual point of view, where the semigroup acts on the law of the process instead of on observables. Let $Q_t(d\boldsymbol{x})$ be the law of $\boldsymbol{x}(t)$. Then

$$\int P_{0,t}(\phi)(\boldsymbol{x})\, Q_0(d\boldsymbol{x}) = \int \phi(\boldsymbol{x})\, Q_t(d\boldsymbol{x}).$$

Proposition 5 (Forward equation)**.** *The law Q_t satisfies*

$$\frac{d}{dt} Q_t = \mathcal{A}_t^* Q_t,$$

where $\mathcal{A}_t^*$ is the adjoint of $\mathcal{A}_t$.

Proof. Differentiating the duality relation gives

$$\begin{aligned}
\frac{d}{dt} \int P_{0,t}(\phi)(\boldsymbol{x})\, Q_0(d\boldsymbol{x}) &= \int \frac{d}{dt} P_{0,t}(\phi)(\boldsymbol{x})\, Q_0(d\boldsymbol{x}) \\
&= \int P_{0,t} \mathcal{A}_t(\phi)(\boldsymbol{x})\, Q_0(d\boldsymbol{x}) \\
&= \int \mathcal{A}_t(\phi)(\boldsymbol{x})\, Q_t(d\boldsymbol{x}).
\end{aligned}$$

Hence,

$$\begin{aligned}
\frac{d}{dt} \int \phi(\boldsymbol{x})\, Q_t(d\boldsymbol{x}) &= \left\langle \phi, \frac{d}{dt} Q_t \right\rangle \\
&= \int \mathcal{A}_t(\phi)(\boldsymbol{x})\, Q_t(d\boldsymbol{x}) \\
&= \langle \mathcal{A}_t \phi, Q_t \rangle \\
&= \langle \phi, \mathcal{A}_t^* Q_t \rangle .
\end{aligned}$$

Since this holds for every sufficiently regular ϕ, we obtain

$$\frac{d}{dt} Q_t = \mathcal{A}_t^* Q_t.$$

□

Proposition 6 (Adjoint generator)**.** *Suppose that $Q_t(d\boldsymbol{x}) = p_t(\boldsymbol{x}) d\boldsymbol{x}$, and define*

$$D(\boldsymbol{x}, t) \coloneqq L(\boldsymbol{x}, t) L(\boldsymbol{x}, t)^\top.$$

Assuming that the boundary terms vanish,

$$\mathcal{A}_t^* p_t(\boldsymbol{x}) = -\nabla_x^\top \left(f(\boldsymbol{x}, t) p_t(\boldsymbol{x}) \right) + \frac{1}{2} \mathrm{tr} \left\{ \nabla \nabla^\top \left(D(\boldsymbol{x}, t) p_t(\boldsymbol{x}) \right) \right\}.$$

Proof. By integration by parts,

$$\begin{aligned}
\int \mathcal{A}_t(\phi)(\boldsymbol{x})\, Q_t(d\boldsymbol{x}) &= \int \nabla_x \phi(\boldsymbol{x})^\top f(\boldsymbol{x}, t) p_t(\boldsymbol{x})\, d\boldsymbol{x} \\
&\quad + \frac{1}{2} \sum_{i,j=1}^{d} \int D_{ij}(\boldsymbol{x}, t) \partial_{x_i x_j} \phi(\boldsymbol{x}) p_t(\boldsymbol{x})\, d\boldsymbol{x} \\
&= - \int \phi(\boldsymbol{x}) \nabla_x^\top \left(f(\boldsymbol{x}, t) p_t(\boldsymbol{x}) \right) d\boldsymbol{x} \\
&\quad + \frac{1}{2} \sum_{i,j=1}^{d} \int \phi(\boldsymbol{x}) \partial_{x_i x_j} \left(D_{ij}(\boldsymbol{x}, t) p_t(\boldsymbol{x}) \right) d\boldsymbol{x}.
\end{aligned}$$

The expression multiplying ϕ is precisely $\mathcal{A}_t^* p_t$. □

Chapter 2

Functional Analytic Background

2.1 Reproducing Kernel Hilbert Spaces

We say that a symmetric function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive semidefinite kernel if, for any finite set of points $\{x_1, \dots, x_m\} \subset \mathcal{X}$, the matrix

$$\mathbf{K} = (K(x_i, x_j))_{i,j=1}^m$$

is positive semidefinite. Equivalently, for every $a = (a_1, \dots, a_m) \in \mathbb{R}^m$,

$$\sum_{i,j=1}^m a_i a_j K(x_i, x_j) \geq 0.$$

Definition 2 (Positive semidefinite kernel)**.** *A symmetric function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called positive semidefinite if, for every finite set $\{x_1, \dots, x_m\} \subset \mathcal{X}$, the Gram matrix*

$$(K(x_i, x_j))_{i,j=1}^m$$

is positive semidefinite.

The kernel K induces an inner product on a space of functions. For each $x \in \mathcal{X}$, define

$$\Phi_x : \mathcal{X} \to \mathbb{R}, \qquad \Phi_x(z) = K(x, z).$$

Let

$$\mathcal{H}_0 = \operatorname{span}\{\Phi_x : x \in \mathcal{X}\}.$$

For

$$f = \sum_{i=1}^A a_i \Phi_{x_i}, \qquad g = \sum_{j=1}^B b_j \Phi_{x'_j},$$

we define

$$\langle f, g \rangle_{\mathcal{H}_0} \coloneqq \sum_{i=1}^A \sum_{j=1}^B a_i b_j K(x_i, x'_j).$$

Using the definition of Φ_x, this can be rewritten as

$$\begin{aligned} \langle f, g \rangle_{\mathcal{H}_0} &= \sum_{j=1}^B b_j \sum_{i=1}^A a_i K(x_i, x'_j) \\ &= \sum_{j=1}^B b_j f(x'_j). \end{aligned}$$

By symmetry of K, it also follows that

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{A} a_i g(x_i).$$

If the kernel is only positive semidefinite, this inner product may have nonzero elements with zero norm. In that case, one first quotients by the null space and then completes the resulting pre-Hilbert space. The completion is a Hilbert space $\mathcal{H}$, called the reproducing kernel Hilbert space associated with K.

Theorem 1. *Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive semidefinite kernel. Then there exists a unique Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$ such that, for every $x \in \mathcal{X}$, $K(x, \cdot) \in \mathcal{H}$, and for every $f \in \mathcal{H}$,*

$$f(x) = \langle f, K(x, \cdot) \rangle_{\mathcal{H}}.$$

In particular,

$$\langle K(x, \cdot), K(z, \cdot) \rangle_{\mathcal{H}} = K(x, z), \qquad x, z \in \mathcal{X}.$$

Two fundamental properties follow from the construction. First,

$$\langle \Phi_x, \Phi_z \rangle_{\mathcal{H}} = K(x, z).$$

Second, for $f = \sum_{i=1}^{A} a_i \Phi_{x_i} \in \mathcal{H}_0$,

$$\begin{aligned} \langle f, \Phi_x \rangle_{\mathcal{H}} &= \sum_{i=1}^{A} a_i K(x_i, x) \\ &= \sum_{i=1}^{A} a_i \Phi_{x_i}(x) \\ &= f(x). \end{aligned}$$

By continuity, this identity extends to every $h \in \mathcal{H}$:

$$h(x) = \langle h, K(x, \cdot) \rangle_{\mathcal{H}}.$$

This is called the reproducing property.

The importance of this construction is that it allows us to replace inner products in a possibly high-dimensional feature space by evaluations of the kernel. To see this, consider first ridge regression in Euclidean space. Given data $(x_i, y_i)_{i=1}^{n}$, collected in a design matrix X and a vector Y, ridge regression solves

$$\theta_* = \arg\min_{\theta} \|Y - X\theta\|_2^2 + \lambda \|\theta\|_2^2.$$

The solution is

$$\theta_* = (X^\top X + \lambda I)^{-1} X^\top Y.$$

Using the identity

$$X^\top (X X^\top + \lambda I) = (X^\top X + \lambda I) X^\top,$$

we obtain

$$(X^\top X + \lambda I)^{-1} X^\top = X^\top (X X^\top + \lambda I)^{-1}.$$

Therefore, the fitted values can be written as

$$X\theta_* = X X^\top (X X^\top + \lambda I)^{-1} Y.$$

This expression depends on the data only through the Gram matrix

$$XX^\top = (\langle x_i, x_j \rangle)_{i,j=1}^n.$$

This suggests replacing Euclidean inner products by kernel evaluations. In the RKHS setting, we consider the regularized empirical risk

$$f_* = \arg\min_{f \in \mathcal{H}} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2.$$

Equivalently,

$$f_* = \arg\min_{f \in \mathcal{H}} \|Y - f(X)\|_2^2 + \lambda \|f\|_{\mathcal{H}}^2,$$

where

$$f(X) = (f(x_1), \dots, f(x_n))^\top.$$

Theorem 2 (Representer theorem)**.** *Let $\mathcal{H}$ be the RKHS associated with $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, let $x_1, \dots, x_n \in \mathcal{X}$, let*

$$\mathcal{L} : \mathbb{R}^n \to \mathbb{R}$$

be an arbitrary loss function depending only on the values of f at the sample points, and let

$$\Omega : [0, \infty) \to \mathbb{R}$$

be a strictly increasing function. If the optimization problem

$$f_* \in \arg\min_{f \in \mathcal{H}} \left\{ \mathcal{L}(f(x_1), \dots, f(x_n)) + \Omega(\|f\|_{\mathcal{H}}) \right\},$$

has a minimizer, then every minimizer belongs to the finite-dimensional subspace generated by the kernel sections at the data points:

$$f_* \in \operatorname{span}\{K(x_i, \cdot) : i = 1, \dots, n\}.$$

Equivalently, there exist coefficients $a_1, \dots, a_n$ such that

$$f_*(x) = \sum_{i=1}^n a_i K(x_i, x).$$

Let

$$\mathcal{H}_\parallel = \operatorname{span}\{K(x_i, \cdot) : i = 1, \dots, n\}.$$

Every $f \in \mathcal{H}$ can be decomposed as

$$f = f_\parallel + f_\perp,$$

where $f_\parallel \in \mathcal{H}_\parallel$ and $f_\perp \perp \mathcal{H}_\parallel$. Then, for every sample point x_i,

$$f_\perp(x_i) = \langle f_\perp, K(x_i, \cdot) \rangle_{\mathcal{H}} = 0.$$

Therefore,

$$f(x_i) = f_\parallel(x_i).$$

Moreover, by Pythagoras,

$$\|f\|_{\mathcal{H}}^2 = \|f_\parallel\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2 \geq \|f_\parallel\|_{\mathcal{H}}^2.$$

Hence the minimizer belongs to $\mathcal{H}_\parallel$. Thus,

$$f_*(x) = \sum_{i=1}^n a_i K(x_i, x)$$

for some coefficients $a_1, \ldots, a_n$.

Define

$$\mathbf{K} = (K(x_i, x_j))_{i,j=1}^n, \qquad k_x = (K(x_1, x), \ldots, K(x_n, x))^\top.$$

If

$$f(x) = \sum_{i=1}^n a_i K(x_i, x),$$

then

$$f(x) = a^\top k_x.$$

In particular, on the training data,

$$f(X) = \mathbf{K}a.$$

Also,

$$\|f\|_{\mathcal{H}}^2 = \sum_{i,j=1}^n a_i a_j K(x_i, x_j) = a^\top \mathbf{K} a.$$

Therefore, the RKHS optimization problem becomes

$$\arg\min_{a \in \mathbb{R}^n} \|Y - \mathbf{K}a\|_2^2 + \lambda a^\top \mathbf{K} a.$$

Differentiating gives

$$(\mathbf{K} + \lambda I)a = Y,$$

and hence

$$a_* = (\mathbf{K} + \lambda I)^{-1} Y.$$

Thus the learned function is

$$f_*(x) = \mathbf{K}a = k_x^\top (\mathbf{K} + \lambda I)^{-1} Y.$$

This is useful, because we are solving an optimization problem over a possibly infinite dimensional Hilbert Space only using information of $\mathbf{K}$.

2.2 Spectral and Resolvent Theory

The generator of a diffusion is typically an unbounded operator. For this reason, we need a small amount of spectral theory for unbounded operators, their domains, adjoints, and resolvents. The reference point for this part is [Davies, 1995].

The main point to keep in mind is that an unbounded differential expression, such as a Laplacian or a Langevin generator, does not define an operator until we specify the space on which it acts and its domain. In the applications below the ambient space will typically be a weighted space

$$\mathcal{L}^2_\pi(\mathcal{X}) = \left\{ f : \mathcal{X} \to \mathbb{R} : \int_{\mathcal{X}} |f(x)|^2 \, \pi(dx) < \infty \right\},$$

where π is an invariant probability measure. The first-order Sobolev space associated with π is

$$H^1_\pi(\mathcal{X}) = \left\{ f \in \mathcal{L}^2_\pi(\mathcal{X}) : \nabla f \in \mathcal{L}^2_\pi(\mathcal{X}; \mathbb{R}^d) \right\}.$$

Its natural norm is

$$\|f\|^2_{H^1_\pi} = \|f\|^2_{\mathcal{L}^2_\pi} + \|\nabla f\|^2_{\mathcal{L}^2_\pi} = \int_{\mathcal{X}} |f(x)|^2 \, \pi(dx) + \int_{\mathcal{X}} \|\nabla f(x)\|^2 \, \pi(dx),$$

Here ∇f is understood in the weak sense.

For a generator $\mathcal{A}$ that is non-positive and symmetric in $\mathcal{L}^2_\pi$, we will also use the energy space

$$\mathcal{W}^\mu_\pi = \left\{ f : \|f\|^2_{\mathcal{W}^\mu_\pi} \coloneqq \mu\|f\|^2_{\mathcal{L}^2_\pi} - \langle \mathcal{A}f, f\rangle_{\mathcal{L}^2_\pi} < \infty \right\}.$$

For overdamped Langevin dynamics, this becomes the weighted Sobolev norm

$$\|f\|^2_{\mathcal{W}^\mu_\pi} = \mu\|f\|^2_{\mathcal{L}^2_\pi} + \frac{1}{\beta}\int_{\mathcal{X}} \|\nabla f(x)\|^2 \, \pi(dx).$$

Thus, the energy norm combines the equilibrium size of an observable with the Dirichlet energy of its gradient. This is precisely the geometry used later to learn the resolvent without evaluating it directly.

Definition 3. *A linear operator on a Hilbert space $\mathcal{H}$ is a pair consisting of a dense linear subspace $L \subseteq \mathcal{H}$ and a linear map $A : L \to \mathcal{H}$. We denote by* $\mathrm{Dom}(A) = L$ *the domain of the operator A.*

Example 1. *Let $\mathcal{H} = \mathcal{L}^2[0,1]$ and consider the differential expression $Af = -f''$. This expression becomes an operator only after a domain is chosen. In the informal notation used above, two natural choices are*

$$L_D = \left\{ f \in C^2[0,1] : f(0) = f(1) = 0 \right\}$$

and

$$L_N = \left\{ f \in C^2[0,1] : f'(0) = f'(1) = 0 \right\}.$$

Thus $A_D f = -f''$ on L_D, while $A_N f = -f''$ on L_N. For the closed Hilbert-space operators, these domains are usually completed in Sobolev spaces. Recall that $H^k(0,1)$ is the space of functions whose weak derivatives up to order k belong to $\mathcal{L}^2(0,1)$, and $H^1_0(0,1)$ is the closure of $C^\infty_c(0,1)$ in $H^1(0,1)$. With this notation, the closed Dirichlet domain is

$$\mathrm{Dom}(A_D) = H^2(0,1) \cap H^1_0(0,1),$$

and the closed Neumann domain is

$$\mathrm{Dom}(A_N) = \{ f \in H^2(0,1) : f'(0) = f'(1) = 0 \}.$$

Both operators act as $-f''$, but they encode different boundary conditions and therefore have different spectra. This example is the basic model for why domains cannot be treated as a technical afterthought.

Definition 4. *A linear operator A on $\mathcal{H}$ is bounded if*

$$\|A\| \coloneqq \sup_{f \in \mathrm{Dom}(A),\, f \neq 0} \frac{\|Af\|}{\|f\|} < \infty.$$

A complex number λ is an eigenvalue of an operator A if there exists a non-zero $f \in \mathrm{Dom}(A)$ such that $Af = \lambda f$.

Definition 5. *Let A be a linear operator on $\mathcal{H}$ with domain* $\mathrm{Dom}(A) = L$*. A complex number z does not lie in the spectrum* $\mathrm{Spec}(A)$ *if the operator $(z - A)$ maps L one-to-one onto $\mathcal{H}$, and the inverse operator*

$$R(z, A) = (z - A)^{-1}$$

is bounded.

Definition 6. *We say that an operator A with dense domain L in a Hilbert space $\mathcal{H}$ is symmetric if, for all $f, g \in L$,*

$$\langle Af, g\rangle_{\mathcal{H}} = \langle f, Ag\rangle_{\mathcal{H}}.$$

Definition 7. *If A is a linear operator on a Hilbert space $\mathcal{H}$, then the adjoint operator A^* is determined by*

$$\langle Af, g\rangle_{\mathcal{H}} = \langle f, A^*g\rangle_{\mathcal{H}}$$

for all $f \in \mathrm{Dom}(A)$ and $g \in \mathrm{Dom}(A^*)$. More precisely, $\mathrm{Dom}(A^*)$ is the set of all $g \in \mathcal{H}$ for which there exists $k \in \mathcal{H}$ such that

$$\langle Af, g\rangle_{\mathcal{H}} = \langle f, k\rangle_{\mathcal{H}}$$

for all $f \in \mathrm{Dom}(A)$, in which case $A^*g = k$.

We say that A is self-adjoint if A is symmetric and $\mathrm{Dom}(A) = \mathrm{Dom}(A^*)$.

Example 2. *By integration by parts,*

$$\begin{aligned}\int_0^1 -f''g &= -[f'g]_0^1 + \int_0^1 f'g' \\ &= -[f'g]_0^1 + [fg']_0^1 - \int_0^1 fg''.\end{aligned}$$

This computation shows why boundary conditions are part of the definition of the operator. The domains L_D and L_N, or their Sobolev closures, should be checked carefully when identifying $\mathrm{Dom}(A^*)$ and deciding whether the resulting operator is self-adjoint.

For the Dirichlet operator A_D, the condition $f(0) = f(1) = 0$ kills the $[fg']_0^1$ term. The remaining boundary term $-[f'g]_0^1$ vanishes for all admissible f exactly when $g(0) = g(1) = 0$. Hence

$$\mathrm{Dom}(A_D^*) = H^2(0,1) \cap H_0^1(0,1) = \mathrm{Dom}(A_D).$$

For the Neumann operator A_N, the condition $f'(0) = f'(1) = 0$ kills the $[f'g]_0^1$ term. The remaining boundary term $[fg']_0^1$ vanishes for all admissible f exactly when $g'(0) = g'(1) = 0$. Hence

$$\mathrm{Dom}(A_N^*) = \{g \in H^2(0,1) : g'(0) = g'(1) = 0\} = \mathrm{Dom}(A_N).$$

Thus, with these Sobolev domains, both A_D and A_N are self-adjoint. If one starts from the smaller C^2-domains L_D and L_N, the integration-by-parts computation still identifies the boundary conditions, but the adjoint naturally lives in the larger Sobolev space $H^2(0,1)$. In that sense, the self-adjoint operators are the closed realizations of the original differential expression.

Theorem 3. *The spectrum of any self-adjoint operator A is real and non-empty. If $z \notin \mathbb{R}$, then*

$$\|R(z,A)\| = \|(z-A)^{-1}\| \leq |\mathrm{Im}(z)|^{-1},$$

which implies that the resolvent operator is bounded.

A key point for the learning problem is that the largest eigenvalues of the resolvent correspond to the eigenvalues of the generator closest to zero. These are precisely the slow modes of the dynamics, which makes low-rank approximation of the resolvent physically relevant.

Theorem 4. *Let $\mathcal{H}$ be a Hilbert space, and let A be a compact self-adjoint operator. Then there exists a complete orthonormal basis of $\mathcal{H}$ composed of eigenfunctions of A.*

Suppose that

$$A\psi_i = \lambda_i \psi_i.$$

Let

$$R(\mu, A) = (\mu I - A)^{-1}.$$

Then ψ_i is also an eigenfunction of $R(\mu, A)$. Indeed,

$$\begin{aligned} R(\mu, A)\psi_i &= \nu_i \psi_i \\ \Rightarrow \psi_i &= \nu_i(\mu I - A)\psi_i \\ \Rightarrow \psi_i &= \nu_i(\mu - \lambda_i)\psi_i. \end{aligned}$$

Thus,

$$\nu_i = \frac{1}{\mu - \lambda_i}.$$

The next result explains why the resolvent can be approximated through its leading spectral components.

Theorem 5 (Eckart–Young–Mirsky Theorem)**.** *Let $A : \mathcal{H}_1 \to \mathcal{H}_2$ be a compact operator between Hilbert spaces with singular value decomposition*

$$A = \sum_{i \geq 1} \sigma_i \, u_i \otimes v_i,$$

where $\{v_i\} \subset \mathcal{H}_1$ and $\{u_i\} \subset \mathcal{H}_2$ are orthonormal sets, and $(\sigma_i)_i$ are the singular values of A, which satisfy

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq 0.$$

For all $d \geq 1$,

$$A_d \coloneqq \sum_{i=1}^{d} \sigma_i \, u_i \otimes v_i$$

satisfies

$$\|A - A_d\| = \min_{\operatorname{rank}(B) \leq d} \|A - B\|.$$

Moreover, if $\sigma_d > \sigma_{d+1}$, then A_d is the unique minimizer.

For a compact self-adjoint operator, the singular values are the absolute values of the eigenvalues. In the case of the Langevin generator below, the resolvent $R(\mu, A)$ is compact, self-adjoint, and positive, so its singular values are exactly its eigenvalues $\nu_i = (\mu - \lambda_i)^{-1}$. Therefore, applying the Eckart–Young–Mirsky theorem to the resolvent means truncating the spectral expansion of $R(\mu, A)$. Through the relation $\lambda_i = \mu - \nu_i^{-1}$, this keeps precisely the generator modes whose eigenvalues are closest to zero, i.e. the slow modes.

Let $A : \operatorname{Dom}(A) \subseteq \mathcal{H} \to \mathcal{H}$ be self-adjoint and non-positive, and assume that A has compact resolvent. For $\mu > 0$, define

$$R(\mu, A) \coloneqq (\mu I - A)^{-1}.$$

This is the situation for the Langevin generator we discuss later. Under this convention, modes with λ_i close to zero decay slowly, while very negative eigenvalues correspond to fast modes. Working with $R(\mu, A)$ instead of A has the following advantages:

- A may be unbounded and defined only on $\operatorname{Dom}(A) \subsetneq \mathcal{H}$, whereas $R(\mu, A)$ is bounded and defined on all of $\mathcal{H}$. Moreover,
$$\|R(\mu, A)\| \leq \frac{1}{\mu}.$$
- $R(\mu, A)$ is compact, self-adjoint, and positive. Hence, it admits finite-rank approximations and the spectral theory of compact operators applies.

- A and $R(\mu, A)$ have the same eigenfunctions. Indeed, if

$$A\psi_i = \lambda_i \psi_i,$$

then

$$R(\mu, A)\psi_i = \nu_i \psi_i, \qquad \nu_i = \frac{1}{\mu - \lambda_i},$$

and

$$\lambda_i = \mu - \frac{1}{\nu_i}.$$

- Under this non-positivity assumption, $0 < \nu_i \leq 1/\mu$. Thus, eigenvalues of A close to zero are mapped to the largest eigenvalues of $R(\mu, A)$, while increasingly negative eigenvalues are mapped close to zero.

- Since $R(\mu, A)$ is positive,

$$\sigma_i(R(\mu, A)) = \nu_i.$$

Therefore, the leading singular components of $R(\mu, A)$ correspond to the eigenvalues of A closest to zero. By the Eckart–Young–Mirsky theorem, a low-rank approximation of $R(\mu, A)$ recovers precisely this leading spectral subspace.

Hence, the resolvent preserves the spectral information of A, while transforming the problem into the approximation of a bounded compact operator.

Chapter 3

Learning Conditional Expectation Operator

3.1 Conditional Mean Embedding

TODO: dar uma ref

In Section 2.1, we introduced reproducing kernel Hilbert spaces. The idea is that, given a set $\mathcal{X}$ and a positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we can associate to K a Hilbert space $\mathcal{H}$ and a feature map

$$\Phi : \mathcal{X} \to \mathcal{H}, \qquad x \mapsto \Phi_x := K(x, \cdot),$$

such that

$$K(x, z) = \langle \Phi_x, \Phi_z \rangle_{\mathcal{H}}.$$

Moreover, for every $h \in \mathcal{H}$, we have the reproducing property

$$h(x) = \langle h, \Phi_x \rangle_{\mathcal{H}}.$$

Now suppose that $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ is distributed according to $P_{X,Y}$, and suppose that we have kernels K_X and K_Y on $\mathcal{X}$ and $\mathcal{Y}$, with corresponding RKHSs $\mathcal{H}_X$ and $\mathcal{H}_Y$. Denote their feature maps by

$$\Phi_x := K_X(x, \cdot) \in \mathcal{H}_X, \qquad \Psi_y := K_Y(y, \cdot) \in \mathcal{H}_Y.$$

The question is how to represent statistics of the distribution $P_{X,Y}$ in this Hilbert space representation. For example, if $f \in \mathcal{H}_X$, how can we express $\mathbb{E}_X [f(X)]$? More generally, if $g \in \mathcal{H}_Y$, how can we express the conditional expectation

$$\mathbb{E}_{Y \mid X=x} [g(Y)]?$$

The kernel mean embedding of the marginal distribution P_X is defined as the expectation of the feature map:

$$\mu_X := \mathbb{E}_X [\Phi_X] \in \mathcal{H}_X.$$

Then, for every $f \in \mathcal{H}_X$, by the reproducing property,

$$\begin{aligned} \langle f, \mu_X \rangle_{\mathcal{H}_X} &= \langle f, \mathbb{E}_X [\Phi_X] \rangle_{\mathcal{H}_X} \\ &= \mathbb{E}_X [\langle f, \Phi_X \rangle_{\mathcal{H}_X}] \\ &= \mathbb{E}_X [f(X)] . \end{aligned}$$

Hence, the expectation of any function $f \in \mathcal{H}_X$ can be written as an inner product in $\mathcal{H}_X$ with the mean embedding μ_X.

We can also represent functions on the product space $\mathcal{X} \times \mathcal{Y}$ using tensor product RKHSs. If

$$h \in \mathcal{H}_X \otimes \mathcal{H}_Y,$$

then the feature map on $\mathcal{X} \times \mathcal{Y}$ is given by

$$(x, y) \mapsto \Phi_x \otimes \Psi_y,$$

and the reproducing property becomes

$$h(x, y) = \langle h, \Phi_x \otimes \Psi_y \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y}.$$

Thus, tensor products allow us to represent functions of both variables (x, y) in a Hilbert space associated with the joint space $\mathcal{X} \times \mathcal{Y}$.

We now apply the same idea to conditional expectations. We want to find an element $\mu_{Y \mid x} \in \mathcal{H}_Y$ such that, for every $g \in \mathcal{H}_Y$,

$$\langle g, \mu_{Y \mid x} \rangle_{\mathcal{H}_Y} = \mathbb{E}_{Y \mid X = x} [g(Y)].$$

Using the reproducing property in $\mathcal{H}_Y$, we have

$$\begin{aligned} \mathbb{E}_{Y \mid X = x} [g(Y)] &= \mathbb{E}_{Y \mid X = x} [\langle g, \Psi_Y \rangle_{\mathcal{H}_Y}] \\ &= \left\langle g, \mathbb{E}_{Y \mid X = x} [\Psi_Y] \right\rangle_{\mathcal{H}_Y}. \end{aligned}$$

Therefore, it is natural to define the conditional mean embedding by

$$\mu_{Y \mid x} := \mathbb{E}_{Y \mid X = x} [\Psi_Y] \in \mathcal{H}_Y.$$

With this definition,

$$\mathbb{E}_{Y \mid X = x} [g(Y)] = \langle g, \mu_{Y \mid x} \rangle_{\mathcal{H}_Y}.$$

It remains to represent the dependence of $\mu_{Y \mid x}$ on x. For this, we introduce a conditional embedding operator

$$U_{Y \mid X} : \mathcal{H}_X \to \mathcal{H}_Y$$

such that

$$\mu_{Y \mid x} = U_{Y \mid X} \Phi_x.$$

To define this operator, we first introduce covariance operators in RKHS. Define

$$C_{XX} := \mathbb{E}_X [\Phi_X \otimes \Phi_X], \qquad C_{YX} := \mathbb{E}_{X,Y} [\Psi_Y \otimes \Phi_X].$$

Here, $\Phi_X \otimes \Phi_X$ is an operator from $\mathcal{H}_X$ to $\mathcal{H}_X$, while $\Psi_Y \otimes \Phi_X$ is an operator from $\mathcal{H}_X$ to $\mathcal{H}_Y$.

Recall that if $a \in \mathcal{H}_Y$ and $b \in \mathcal{H}_X$, then the tensor product

$$a \otimes b$$

defines a rank-one linear operator from $\mathcal{H}_X$ to $\mathcal{H}_Y$ by

$$(a \otimes b) f = a \langle b, f \rangle_{\mathcal{H}_X}, \qquad f \in \mathcal{H}_X.$$

Thus, the second element b is used to test the input function f through the inner product, while the first element a gives the direction of the resulting vector.

In particular, for $f' \in \mathcal{H}_X$,

$$(\Phi_X \otimes \Phi_X) f' = \Phi_X \langle \Phi_X, f' \rangle_{\mathcal{H}_X}.$$

By the reproducing property,

$$\langle \Phi_X, f' \rangle_{\mathcal{H}_X} = f'(X).$$

Therefore,

$$(\Phi_X \otimes \Phi_X) f' = \Phi_X f'(X).$$

Taking expectation gives

$$C_{XX} f' = \mathbb{E}_X \left[(\Phi_X \otimes \Phi_X) f' \right] = \mathbb{E}_X \left[\Phi_X f'(X) \right].$$

Now, taking the inner product with $f \in \mathcal{H}_X$, we obtain

$$\begin{aligned} \langle f, C_{XX} f' \rangle_{\mathcal{H}_X} &= \left\langle f, \mathbb{E}_X \left[\Phi_X f'(X) \right] \right\rangle_{\mathcal{H}_X} \\ &= \mathbb{E}_X \left[\langle f, \Phi_X \rangle_{\mathcal{H}_X} f'(X) \right] \\ &= \mathbb{E}_X \left[f(X) f'(X) \right]. \end{aligned}$$

Thus, C_{XX} satisfies

$$\langle f, C_{XX} f' \rangle_{\mathcal{H}_X} = \mathbb{E}_X \left[f(X) f'(X) \right].$$

Similarly, $\Psi_Y \otimes \Phi_X$ defines an operator from $\mathcal{H}_X$ to $\mathcal{H}_Y$. For $f \in \mathcal{H}_X$,

$$(\Psi_Y \otimes \Phi_X) f = \Psi_Y \langle \Phi_X, f \rangle_{\mathcal{H}_X}.$$

Again, by the reproducing property,

$$\langle \Phi_X, f \rangle_{\mathcal{H}_X} = f(X).$$

Hence,

$$(\Psi_Y \otimes \Phi_X) f = \Psi_Y f(X).$$

Taking expectation gives

$$C_{YX} f = \mathbb{E}_{X,Y} \left[(\Psi_Y \otimes \Phi_X) f \right] = \mathbb{E}_{X,Y} \left[\Psi_Y f(X) \right].$$

Now, taking the inner product with $g \in \mathcal{H}_Y$, we get

$$\begin{aligned} \langle g, C_{YX} f \rangle_{\mathcal{H}_Y} &= \langle g, \mathbb{E}_{X,Y} \left[\Psi_Y f(X) \right] \rangle_{\mathcal{H}_Y} \\ &= \mathbb{E}_{X,Y} \left[\langle g, \Psi_Y \rangle_{\mathcal{H}_Y} f(X) \right] \\ &= \mathbb{E}_{X,Y} \left[g(Y) f(X) \right]. \end{aligned}$$

Therefore, C_{YX} satisfies

$$\langle g, C_{YX} f \rangle_{\mathcal{H}_Y} = \mathbb{E}_{X,Y} \left[g(Y) f(X) \right].$$

These identities show that C_{XX} represents second-order information about X in $\mathcal{H}_X$, while C_{YX} represents the interaction between X and Y through the two RKHSs.

We now derive the expression for the conditional embedding operator. The defining property of conditional expectation gives, for $f \in \mathcal{H}_X$ and $g \in \mathcal{H}_Y$,

$$\mathbb{E}_{X,Y} \left[f(X) g(Y) \right] = \mathbb{E}_X \left[f(X) \mathbb{E}_{Y \mid X} \left[g(Y) \right] \right].$$

Using the conditional mean embedding,

$$\mathbb{E}_{Y \mid X} \left[g(Y) \right] = \langle g, \mu_{Y \mid X} \rangle_{\mathcal{H}_Y}.$$

If $\mu_{Y \mid X} = U_{Y \mid X} \Phi_X$, then

$$\mathbb{E}_{Y \mid X} \left[g(Y) \right] = \langle g, U_{Y \mid X} \Phi_X \rangle_{\mathcal{H}_Y}.$$

Therefore,

$$\begin{aligned}\mathbb{E}_{X,Y}\left[f(X)g(Y)\right] &= \mathbb{E}_X\left[f(X)\langle g, U_{Y\,|\,X}\Phi_X\rangle_{\mathcal{H}_Y}\right] \\ &= \mathbb{E}_X\left[\langle f, \Phi_X\rangle_{\mathcal{H}_X}\langle g, U_{Y\,|\,X}\Phi_X\rangle_{\mathcal{H}_Y}\right].\end{aligned}$$

Using the adjoint operator $U^*_{Y\,|\,X} : \mathcal{H}_Y \to \mathcal{H}_X$, we have

$$\langle g, U_{Y\,|\,X}\Phi_X\rangle_{\mathcal{H}_Y} = \langle U^*_{Y\,|\,X} g, \Phi_X\rangle_{\mathcal{H}_X}.$$

Hence,

$$\begin{aligned}\mathbb{E}_{X,Y}\left[f(X)g(Y)\right] &= \mathbb{E}_X\left[\langle f, \Phi_X\rangle_{\mathcal{H}_X}\langle U^*_{Y\,|\,X} g, \Phi_X\rangle_{\mathcal{H}_X}\right] \\ &= \mathbb{E}_X\left[f(X)\langle U^*_{Y\,|\,X} g, \Phi_X\rangle_{\mathcal{H}_X}\right] \\ &= \mathbb{E}_X\left[\langle U^*_{Y\,|\,X} g, f(X)\Phi_X\rangle_{\mathcal{H}_X}\right] \\ &= \langle U^*_{Y\,|\,X} g, \mathbb{E}_X\left[f(X)\Phi_X\right]\rangle_{\mathcal{H}_X} \\ &= \langle U^*_{Y\,|\,X} g, C_{XX} f\rangle_{\mathcal{H}_X} \\ &= \langle g, U_{Y\,|\,X} C_{XX} f\rangle_{\mathcal{H}_Y}.\end{aligned}$$

On the other hand, by the definition of C_{YX},

$$\mathbb{E}_{X,Y}\left[f(X)g(Y)\right] = \langle g, C_{YX} f\rangle_{\mathcal{H}_Y}.$$

Since this holds for every $g \in \mathcal{H}_Y$, we obtain

$$C_{YX} f = U_{Y\,|\,X} C_{XX} f.$$

Since this holds for every $f \in \mathcal{H}_X$, we have

$$C_{YX} = U_{Y\,|\,X} C_{XX}.$$

Formally, if C_{XX} is invertible, this gives

$$U_{Y\,|\,X} = C_{YX} C^{-1}_{XX}.$$

Therefore,

$$\mu_{Y\,|\,x} = U_{Y\,|\,X}\Phi_x = C_{YX} C^{-1}_{XX}\Phi_x.$$

In practice, C_{XX} may not be invertible, so one usually uses a regularized version:

$$\mu_{Y\,|\,x} = C_{YX}(C_{XX} + \lambda I)^{-1}\Phi_x,$$

where $\lambda > 0$ is a regularization parameter.

In practice, the operators C_{XX} and C_{YX} are unknown, since they depend on the distribution $P_{X,Y}$. Given a sample

$$\{(x_i, y_i)\}_{i=1}^n \sim P_{X,Y},$$

we estimate them by

$$\widehat{C}_{XX} := \frac{1}{n}\sum_{i=1}^n \Phi_{x_i} \otimes \Phi_{x_i}, \qquad \widehat{C}_{YX} := \frac{1}{n}\sum_{i=1}^n \Psi_{y_i} \otimes \Phi_{x_i}.$$

The empirical regularized estimator of the conditional mean embedding is then

$$\widehat{\mu}_{Y\,|\,x} = \widehat{C}_{YX}(\widehat{C}_{XX} + \lambda I_{\mathcal{H}_X})^{-1}\Phi_x.$$

Although this expression is written in terms of operators on RKHSs, it can be computed using only Gram matrices. Define

$$\mathbf{\Phi} : \mathbb{R}^n \to \mathcal{H}_X, \qquad \mathbf{\Phi} a := \sum_{i=1}^{n} a_i \Phi_{x_i},$$

and

$$\mathbf{\Psi} : \mathbb{R}^n \to \mathcal{H}_Y, \qquad \mathbf{\Psi} a := \sum_{i=1}^{n} a_i \Psi_{y_i}.$$

Their adjoints satisfy

$$\mathbf{\Phi}^* f = \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}, \qquad \mathbf{\Psi}^* g = \begin{pmatrix} g(y_1) \\ \vdots \\ g(y_n) \end{pmatrix}.$$

With this notation,

$$\widehat{C}_{XX} = \frac{1}{n} \mathbf{\Phi}\mathbf{\Phi}^*, \qquad \widehat{C}_{YX} = \frac{1}{n} \mathbf{\Psi}\mathbf{\Phi}^*.$$

Therefore,

$$\begin{aligned} \widehat{\mu}_{Y\,|\,x} &= \frac{1}{n} \mathbf{\Psi}\mathbf{\Phi}^* \left(\frac{1}{n} \mathbf{\Phi}\mathbf{\Phi}^* + \lambda I_{\mathcal{H}_X} \right)^{-1} \Phi_x \\ &= \mathbf{\Psi}\mathbf{\Phi}^* \left(\mathbf{\Phi}\mathbf{\Phi}^* + n\lambda I_{\mathcal{H}_X} \right)^{-1} \Phi_x. \end{aligned}$$

We now use the identity

$$\mathbf{\Phi}^* \left(\mathbf{\Phi}\mathbf{\Phi}^* + n\lambda I_{\mathcal{H}_X} \right)^{-1} = \left(\mathbf{\Phi}^*\mathbf{\Phi} + n\lambda I_n \right)^{-1} \mathbf{\Phi}^*.$$

Indeed, this follows from

$$\left(\mathbf{\Phi}\mathbf{\Phi}^* + n\lambda I_{\mathcal{H}_X} \right) \mathbf{\Phi} = \mathbf{\Phi} \left(\mathbf{\Phi}^*\mathbf{\Phi} + n\lambda I_n \right),$$

together with the fact that $n\lambda > 0$. Hence,

$$\widehat{\mu}_{Y\,|\,x} = \mathbf{\Psi} \left(\mathbf{\Phi}^*\mathbf{\Phi} + n\lambda I_n \right)^{-1} \mathbf{\Phi}^* \Phi_x.$$

Let

$$K := \mathbf{\Phi}^*\mathbf{\Phi} \in \mathbb{R}^{n \times n}, \qquad K_{ij} = K_X(x_i, x_j),$$

and define

$$k_x := \mathbf{\Phi}^* \Phi_x = \begin{pmatrix} K_X(x_1, x) \\ \vdots \\ K_X(x_n, x) \end{pmatrix}.$$

Then

$$\widehat{\mu}_{Y\,|\,x} = \mathbf{\Psi}(K + n\lambda I_n)^{-1} k_x.$$

Equivalently, if

$$\alpha(x) := (K + n\lambda I_n)^{-1} k_x,$$

then

$$\widehat{\mu}_{Y\,|\,x} = \sum_{i=1}^{n} \alpha_i(x) \Psi_{y_i} = \sum_{i=1}^{n} \alpha_i(x) K_Y(y_i, \cdot).$$

Thus, the conditional distribution of Y given $X = x$ is represented empirically by a weighted combination of the output features associated with the observed values $y_1, \ldots, y_n$.

For any $g \in \mathcal{H}_Y$, we estimate the conditional expectation by

$$\begin{aligned}\widehat{\mathbb{E}_{Y \mid X=x}[g(Y)]} &:= \langle g, \widehat{\mu}_{Y \mid x} \rangle_{\mathcal{H}_Y} \\ &= \left\langle g, \sum_{i=1}^{n} \alpha_i(x) \Psi_{y_i} \right\rangle_{\mathcal{H}_Y} \\ &= \sum_{i=1}^{n} \alpha_i(x) g(y_i).\end{aligned}$$

Therefore, once the weights $\alpha(x)$ are computed, the conditional expectation of $g(Y)$ given $X = x$ is estimated by a weighted average of the observed values $g(y_i)$.

In summary, the theoretical expression

$$\mu_{Y \mid x} = C_{YX}(C_{XX} + \lambda I_{\mathcal{H}_X})^{-1} \Phi_x$$

leads to the empirical estimator

$$\widehat{\mu}_{Y \mid x} = \sum_{i=1}^{n} \alpha_i(x) \Psi_{y_i}, \qquad \alpha(x) = (K + n\lambda I_n)^{-1} k_x.$$

Thus, conditional mean embeddings can be estimated using only kernel evaluations and the solution of a regularized linear system.

Chapter 4

Learning Langevin Generators

4.1 Langevin Generators and Energy Geometry

The main object of our study is the overdamped time-homogeneous Langevin dynamics

$$d\boldsymbol{x}(t) = -\nabla_x U(\boldsymbol{x})\, dt + \sqrt{2\beta^{-1}}\, d\boldsymbol{B}(t),$$

where $U : \mathbb{R}^d \to \mathbb{R}$ is sufficiently regular and $\beta > 0$.

The general formula for the generator and the adjoint now gives the following specialization.

Proposition 7 (Generator and stationary density)**.** *For every sufficiently regular spatial observable $\phi : \mathbb{R}^d \to \mathbb{R}$, the infinitesimal generator is*

$$\mathcal{A}\phi = \langle -\nabla U, \nabla\phi\rangle + \beta^{-1}\Delta\phi.$$

Moreover, if p_t is the density of $\boldsymbol{x}(t)$, then

$$\mathcal{A}^* p_t(\boldsymbol{x}) = -\nabla^\top \left(-\nabla U(\boldsymbol{x}) p_t(\boldsymbol{x}) + \beta^{-1}\nabla p_t(\boldsymbol{x})\right).$$

The stationary density is

$$\pi(\boldsymbol{x}) = \frac{e^{-\beta U(\boldsymbol{x})}}{Z_\beta},$$

where Z_β is the normalizing constant.

Proof. The expression for $\mathcal{A}$ follows from the general diffusion generator by taking $f = -\nabla U$ and $L = \sqrt{2\beta^{-1}}I$. The formula for $\mathcal{A}^*$ is the corresponding Fokker–Planck operator (Proposition 6). If π is stationary, then $\mathcal{A}^*\pi = 0$, hence

$$-\nabla^\top \left(-\nabla U(\boldsymbol{x})\pi(\boldsymbol{x}) + \beta^{-1}\nabla\pi(\boldsymbol{x})\right) = 0.$$

The zero-flux solution satisfies

$$-\nabla U(\boldsymbol{x})\pi(\boldsymbol{x}) + \beta^{-1}\nabla\pi(\boldsymbol{x}) = 0,$$

or equivalently

$$-\beta\nabla U(\boldsymbol{x}) = \frac{\nabla\pi(\boldsymbol{x})}{\pi(\boldsymbol{x})}.$$

Integrating this identity gives $\pi(\boldsymbol{x}) = e^{-\beta U(\boldsymbol{x})}/Z_\beta$. □

The next lemma rewrites the generator in the geometry of its stationary measure. This is the computation that explains both the symmetry of $\mathcal{A}$ and the sign of its spectrum.

Lemma 1 (Dirichlet form of the Langevin generator)**.** *Let $\pi(\boldsymbol{x}) = Z_\beta^{-1} e^{-\beta U(\boldsymbol{x})}$. For smooth test functions ϕ, φ for which boundary terms vanish,*

$$\mathcal{A}\phi = \frac{1}{\beta\pi}\nabla^\top(\pi\nabla\phi),$$

and

$$\langle \mathcal{A}\phi, \varphi\rangle_{\mathcal{L}^2_\pi} = -\frac{1}{\beta}\int \nabla\phi^\top\nabla\varphi\,\pi.$$

In particular, $\mathcal{A}$ is symmetric and non-positive on this class of functions.

Proof. Since

$$-\nabla U = \frac{1}{\beta}\frac{\nabla\pi}{\pi},$$

we have

$$\frac{1}{\beta\pi}\nabla^\top(\pi\nabla\phi) = \frac{1}{\beta}\frac{\nabla\pi}{\pi}^\top\nabla\phi + \frac{1}{\beta}\Delta\phi = -\nabla U^\top\nabla\phi + \frac{1}{\beta}\Delta\phi = \mathcal{A}\phi.$$

Therefore, integrating by parts,

$$\begin{aligned}\int \mathcal{A}\phi\,\varphi\,\pi &= \int \frac{1}{\beta}\nabla^\top(\pi\nabla\phi)\,\varphi \\ &= -\frac{1}{\beta}\int \nabla\phi^\top\nabla\varphi\,\pi.\end{aligned}$$

The same expression with ϕ and φ exchanged gives symmetry. Taking $\varphi = \phi$ gives

$$\langle \mathcal{A}\phi, \phi\rangle_{\mathcal{L}^2_\pi} = -\frac{1}{\beta}\int \|\nabla\phi\|^2\pi \le 0.$$

□

Remark 1. *We have shwon that*

$$\frac{1}{\beta\pi}\nabla^\top(\pi\nabla\phi) = \mathcal{A}\phi.$$

therefore,

$$\sum_{i=1}^{\infty}\lambda_i\psi_i(x)\langle\psi_i,\phi\rangle_{\mathcal{L}^2_\pi} = \mathcal{A}\phi(x) = \frac{1}{\beta}e^{\beta U(x)}\frac{d}{dx}\left(e^{-\beta U(x)}\phi'(x)\right)$$

Taking $\phi = \psi_i$,

$$\beta\lambda_i e^{-\beta U(x)}\psi_i(x) = \frac{d}{dx}\left(e^{-\beta U(x)}\psi_i'(x)\right)$$

which can be written as

$$\left(-\pi\psi_i'\right)' = -\beta\lambda_i\pi\psi_i,$$

which is a Sturm-Liouville Problem [Bakry et al., 2014, Sec. 2.6].

The lemma is initially a computation on test functions. To obtain the operator used in spectral theory, define

$$H^1_\pi = \left\{\phi \in \mathcal{L}^2_\pi : \nabla\phi \in \mathcal{L}^2_\pi(\mathbb{R}^d)\right\}$$

This Sobolev space is the natural form domain of the Langevin generator. The Dirichlet form is the bilinear energy obtained by closing the integration by parts identity on this domain: it measures the gradient part of the H^1_π norm and represents $-\langle\mathcal{A}\phi,\varphi\rangle_{\mathcal{L}^2_\pi}$.

$$\mathcal{E}(\phi,\varphi) = \frac{1}{\beta}\int \nabla\phi^\top\nabla\varphi\,\pi.$$

When U is sufficiently regular and confining, this form is closed on H^1_π. The self-adjoint realization of $\mathcal{A}$ is the operator associated with $-\mathcal{E}$: ϕ belongs to $\mathrm{Dom}(\mathcal{A})$ if there exists $h \in \mathcal{L}^2_\pi$ such that

$$-\mathcal{E}(\phi, \varphi) = \langle h, \varphi \rangle_{\mathcal{L}^2_\pi} \qquad \text{for all } \varphi \in H^1_\pi,$$

and then $\mathcal{A}\phi = h$. On smooth functions in this domain, this abstract operator coincides with the differential expression above.

Under these assumptions, and when the resolvent is compact, the spectral theorem allows us to write

$$\mathcal{A} = \sum_i \lambda_i \psi_i \otimes \psi_i,$$

for some orthonormal basis $\{\psi_i\}_i \subset \mathcal{L}^2_\pi$. Since $\mathcal{A}$ is non-positive, all eigenvalues satisfy $\lambda_i \leq 0$, and

$$\mathcal{A}\phi = \sum_i \lambda_i \psi_i \langle \psi_i, \phi \rangle_{\mathcal{L}^2_\pi}.$$

Given any observable ϕ, let

$$P_t\phi = \sum a_i(t)\psi_i$$

Using the Kolmogorov backward equation,

$$\frac{d}{dt} P_t\phi = \mathcal{A} P_t \phi,$$

hence

$$\begin{aligned} \sum a_i'(t)\psi_i &= \sum \lambda_i a_i(t) \langle \psi_i, \psi_i \rangle_{\mathcal{L}^2_\pi} \psi_i \\ &= \sum \lambda_i a_i(t) \psi_i, \end{aligned}$$

so, $a_i(t) = e^{\lambda_i t} a_i(0)$ and since $P_0\phi = \phi = \sum a_i(0)\psi_i$, we have that

$$a_i(t) = e^{\lambda_i t} \langle \phi, \psi_i \rangle_{\mathcal{L}^2_\pi}.$$

Therefore, if we know the spectral decomposition of $\mathcal{A}$, then

$$P_t\phi = \sum_i e^{\lambda_i t} \langle \phi, \psi_i \rangle_{\mathcal{L}^2_\pi} \psi_i.$$

This connection between Markov diffusion generators, Dirichlet forms, and the associated Sobolev geometry is standard; see [Bakry et al., 2014] for the functional-analytic background.

4.2 General Framework for Learning the Generator

We assume we have a Langevin process and a sample $x_1, \ldots, x_n \sim \pi$, where π is the stationary distribution of the process.

The target object is the generator $\mathcal{A}$, but $\mathcal{A}$ is unbounded. Therefore, we work with the shifted resolvent

$$R(\mu, \mathcal{A}) = (\mu I - \mathcal{A})^{-1}, \qquad \mu > 0,$$

and keep this convention throughout. If $\mathcal{A}\psi_i = \lambda_i \psi_i$, then

$$R(\mu, \mathcal{A})\psi_i = \frac{1}{\mu - \lambda_i}\psi_i, \qquad \lambda_i = \mu - \frac{1}{\nu_i},$$

where ν_i denotes the corresponding resolvent eigenvalue.

The natural space for this resolvent regression problem is the energy space

$$\mathcal{W}_\pi^\mu = \left\{\phi \in \mathcal{L}_\pi^2 : \|\phi\|_{\mathcal{W}_\pi^\mu} < \infty\right\},$$

with inner product

$$\langle \phi, \varphi \rangle_{\mathcal{W}_\pi^\mu} = \mu \langle \phi, \varphi \rangle_{\mathcal{L}_\pi^2} - \langle \mathcal{A}\phi, \varphi \rangle_{\mathcal{L}_\pi^2}.$$

TODO: talk about lax milgran For overdamped Langevin dynamics this becomes

$$\langle \phi, \varphi \rangle_{\mathcal{W}_\pi^\mu} = \mu \langle \phi, \varphi \rangle_{\mathcal{L}_\pi^2} + \frac{1}{\beta} \int_{\mathcal{X}} \nabla\phi(x)^\top \nabla\varphi(x)\, \pi(dx).$$

This inner product is useful because it can be estimated from samples using only function values and gradients of the chosen features.

We want to learn the eigenpairs (λ_i, ψ_i) associated with the spectral decomposition of the resolvent operator $R(\mu, \mathcal{A})$.

Since learning the full operator is infinite-dimensional, we restrict the problem to learning how $R(\mu, \mathcal{A})$ acts on a finite-dimensional family of functions

$$\phi \in \mathcal{F} \coloneqq \operatorname{span}\{z_0, z_1, \ldots, z_m\},$$

where $z_i : \mathcal{X} \to \mathbb{R}$ and $z_i \in \mathcal{W}_\pi^\mu$.

Define the operator

$$Z : \mathbb{R}^{m+1} \to \mathcal{W}_\pi^\mu, \qquad Zu \coloneqq \sum_{i=0}^{m} u_i z_i.$$

Then any function $\phi \in \mathcal{F}$ can be written as

$$\phi = Zu$$

for some vector $u \in \mathbb{R}^{m+1}$. Thus, we want to learn $R(\mu, \mathcal{A})Zu$. Since we still want to represent the image of the operator as a function in $\mathcal{F}$, we seek a matrix $G \in \mathbb{R}^{(m+1)\times(m+1)}$ such that

$$R(\mu, \mathcal{A})Zu \approx ZGu.$$

A natural loss would be something like:

$$\|R(\mu, \mathcal{A})Z - ZG\|_2.$$

However, we do not know how to evaluate $R(\mu, \mathcal{A})Z$ pointwise. The key observation behind the energy formulation of [Kostic et al., 2024] is the following: the unknown resolvent term can be replaced by quantities that are computable from inner products.

An important property of this inner product is that, for sufficiently regular functions, as shown in the previous section,

$$\begin{aligned}
\langle \phi, \varphi \rangle_{\mathcal{W}_\mu^\pi} &= \mu \langle \phi, \varphi \rangle_{\mathcal{L}_\pi^2} - \langle \mathcal{A}\phi, \varphi \rangle_{\mathcal{L}_\pi^2} \\
&= \langle (\mu I - \mathcal{A})\phi, \varphi \rangle_{\mathcal{L}_\pi^2} \\
&= \langle R(\mu, \mathcal{A})^{-1}\phi, \varphi \rangle_{\mathcal{L}_\pi^2}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\langle R(\mu, \mathcal{A})\phi, \varphi \rangle_{\mathcal{W}_\mu^\pi} &= \left\langle R(\mu, \mathcal{A})^{-1} R(\mu, \mathcal{A})\phi, \varphi \right\rangle_{\mathcal{L}_\pi^2} \\
&= \langle \phi, \varphi \rangle_{\mathcal{L}_\pi^2}.
\end{aligned}$$

That is, if we change the geometry of the space using this inner product, the action of the resolvent over functions simplifies. Indeed, consider the Hilbert–Schmidt loss

$$\mathcal{R}(G) = \|R(\mu, \mathcal{A})Z - ZG\|^2_{\mathrm{HS}(\mathbb{R}^{m+1}, \mathcal{W}^\pi_\mu)}.$$

Expanding the squared norm gives

$$\begin{aligned}\mathcal{R}(G) = \|R(\mu, \mathcal{A})Z\|^2_{\mathrm{HS}} + \|ZG\|^2_{\mathrm{HS}} \\ - 2\langle R(\mu, \mathcal{A})Z, ZG\rangle_{\mathrm{HS}}.\end{aligned}$$

If $S, T : \mathbb{R}^{m+1} \to \mathcal{W}^\mu_\pi$ are Hilbert–Schmidt operators and $\{e_j\}_{j=0}^m$ is the standard basis of $\mathbb{R}^{m+1}$, then

$$\langle S, T\rangle_{\mathrm{HS}(\mathbb{R}^{m+1}, \mathcal{W}^\mu_\pi)} = \sum_{j=0}^{m} \langle Se_j, Te_j\rangle_{\mathcal{W}^\mu_\pi} = \mathrm{tr}\left\{(S^*T)\right\}.$$

The adjoint operator

$$Z^* : \mathcal{W}^\mu_\pi \to \mathbb{R}^{m+1}$$

is characterized by

$$\langle Zu, \phi\rangle_{\mathcal{W}^\mu_\pi} = \langle u, Z^*\phi\rangle_{\mathbb{R}^{m+1}},$$

and is explicitly given by

$$Z^*\phi = \begin{pmatrix} \langle z_0, \phi\rangle_{\mathcal{W}^\mu_\pi} \\ \vdots \\ \langle z_m, \phi\rangle_{\mathcal{W}^\mu_\pi} \end{pmatrix}.$$

Define the matrices

$$W \coloneqq Z^*Z \qquad \text{and} \qquad C \coloneqq Z^*R(\mu, \mathcal{A})Z.$$

Their entries follow by testing the operators on the canonical basis. Since $Ze_j = z_j$,

$$W_{ij} = e_i^\top W e_j = e_i^\top Z^* Z e_j = \langle Ze_i, Ze_j\rangle_{\mathcal{W}^\pi_\mu} = \langle z_i, z_j\rangle_{\mathcal{W}^\pi_\mu}.$$

Similarly,

$$\begin{aligned}C_{ij} &= e_i^\top C e_j \\ &= e_i^\top Z^* R(\mu, \mathcal{A}) Z e_j \\ &= \langle R(\mu, \mathcal{A}) z_j, z_i\rangle_{\mathcal{W}^\pi_\mu} \\ &= \langle z_j, z_i\rangle_{\mathcal{L}^2_\pi} = \langle z_i, z_j\rangle_{\mathcal{L}^2_\pi},\end{aligned}$$

where the last equality uses the energy identity.

For Langevin dynamics,

$$W_{ij} = \mathbb{E}_{X\sim\pi}\left[\mu z_i(X) z_j(X) + \frac{1}{\beta}\nabla z_i(X)^\top \nabla z_j(X)\right],$$

whereas

$$C_{ij} = \mathbb{E}_{X\sim\pi}\left[z_i(X) z_j(X)\right].$$

The second term of the loss satisfies

$$\begin{aligned}\|ZG\|^2_{\mathrm{HS}} &= \mathrm{tr}\left\{((ZG)^* ZG)\right\} \\ &= \mathrm{tr}\left\{\left(G^\top Z^* Z G\right)\right\} \\ &= \mathrm{tr}\left\{\left(G^\top W G\right)\right\}.\end{aligned}$$

For the cross term, using the Hilbert–Schmidt inner product and then the energy identity,

$$\begin{aligned}
\langle R(\mu,\mathcal{A})Z, ZG\rangle_{\mathrm{HS}} &= \sum_{j=0}^{m} \langle (R(\mu,\mathcal{A})Z)\, e_j, (ZG)\, e_j\rangle_{\mathcal{W}_\mu^\pi} \\
&= \sum_{j=0}^{m} \langle R(\mu,\mathcal{A})Ze_j, ZGe_j\rangle_{\mathcal{W}_\mu^\pi} \\
&= \sum_{j=0}^{m} \left\langle R(\mu,\mathcal{A})z_j, Z\left(\sum_{i=0}^{m} G_{ij}e_i\right)\right\rangle_{\mathcal{W}_\mu^\pi} \\
&= \sum_{i,j=0}^{m} G_{ij}\, \langle R(\mu,\mathcal{A})z_j, z_i\rangle_{\mathcal{W}_\mu^\pi} \\
&= \sum_{i,j=0}^{m} G_{ij}\, \langle z_j, z_i\rangle_{\mathcal{L}_\pi^2} \\
&= \sum_{i,j=0}^{m} G_{ij}C_{ij} \\
&= \mathrm{tr}\left\{\left(G^\top C\right)\right\}.
\end{aligned}$$

Hence,

$$\mathcal{R}(G) = \|R(\mu,\mathcal{A})Z\|_{\mathrm{HS}}^2 + \mathrm{tr}\left\{\left(G^\top W G\right)\right\} - 2\mathrm{tr}\left\{\left(G^\top C\right)\right\}.$$

Since the first term does not depend on G, minimizing $\mathcal{R}(G)$ is equivalent to solving

$$\min_G \left\{\mathrm{tr}\left\{\left(G^\top W G\right)\right\} - 2\mathrm{tr}\left\{\left(G^\top C\right)\right\}\right\}.$$

If W is invertible, the first-order optimality condition is

$$WG = C,$$

and therefore

$$G = W^{-1}C.$$

Now, suppose that $Gu = \nu u$ for some $u \in \mathbb{R}^{m+1}$. Then

$$R(\mu,\mathcal{A})Zu \approx ZGu = \nu Zu.$$

Therefore, $\psi = Zu$ is an approximate eigenfunction of the resolvent. Recovering the eigenpairs of $R(\mu,\mathcal{A})$ then gives the corresponding eigenfunctions of $\mathcal{A}$, with eigenvalues transformed through the resolvent relation described above.

4.2.1 Using RKHS representation

In the RKHS formulation of [Kostic et al., 2024], the dictionary functions z_i are chosen from a reproducing kernel Hilbert space [Mohri et al., 2012]. This provides a controlled finite-dimensional approximation of the resolvent and allows one to state statistical guarantees for the estimated spectral quantities. Let $\mathcal{H}$ be an RKHS with kernel k and feature map $\Phi : \mathcal{X} \to \mathcal{H}$, so that

$$k(x,y) = \langle \Phi(x), \Phi(y)\rangle_{\mathcal{H}}.$$

An important assumption is not merely $\mathcal{H} \subset \mathcal{L}_\pi^2$, but rather

$$\mathcal{H} \subset \mathcal{W}_\pi^\mu,$$

so that the features have finite energy.

Concretely, every $\phi \in \mathcal{H}$, viewed as a function on $\mathcal{X}$, must satisfy

$$\|\phi\|^2_{\mathcal{W}^\mu_\pi} = \mu\|\phi\|^2_{\mathcal{L}^2_\pi} - \langle \mathcal{A}\phi, \phi\rangle_{\mathcal{L}^2_\pi} < \infty.$$

For overdamped Langevin dynamics, this is the requirement that both the $\mathcal{L}^2_\pi$ norm of ϕ and the weighted gradient energy $\beta^{-1}\|\nabla\phi\|^2_{\mathcal{L}^2_\pi}$ are finite.

TODO: fazer com mais detalhes

Denote by

$$Z_\mu : \mathcal{H} \to \mathcal{W}^\mu_\pi$$

the canonical injection.

This means that Z_μ is the inclusion map: for $h \in \mathcal{H}$, $Z_\mu h$ is the same function h, but regarded as an element of the energy space $\mathcal{W}^\mu_\pi$.

In this setting, the population covariance operators are

$$W_\mu = Z^*_\mu Z_\mu = S^*_\pi(\mu I - \mathcal{A})S_\pi, \qquad C = S^*_\pi S_\pi,$$

where $S_\pi : \mathcal{H} \to \mathcal{L}^2_\pi$ is the canonical inclusion. The identity

$$Z^*_\mu R(\mu, \mathcal{A}) Z_\mu = C$$

is the RKHS analogue of the finite-dimensional identity used above. Given empirical estimates $\widehat{W}_\mu$ and $\widehat{C}$, and a regularization parameter $\gamma > 0$, define

$$\widehat{W}_{\mu,\gamma} = \widehat{W}_\mu + \mu\gamma I.$$

The full-rank regularized estimator is the kernel ridge estimator

$$\widehat{G}_{\mu,\gamma} = \widehat{W}^{-1}_{\mu,\gamma}\widehat{C}.$$

The reduced-rank estimator imposes $\operatorname{rank}(G) \le r$. It is obtained by truncating the singular value decomposition:

$$\widehat{G}^r_{\mu,\gamma} = \widehat{W}^{-1/2}_{\mu,\gamma}\left[\left[\widehat{W}^{-1/2}_{\mu,\gamma}\widehat{C}\right]\right]_r,$$

where $[[\cdot]]_r$ denotes the rank-r truncated SVD. This is the infinite-dimensional version of replacing $G = W^{-1}C$ by a low-rank operator that keeps only the leading resolvent modes.

4.2.2 Using Neural Network representation

In [Kostic et al.,], neural networks are used to learn the functions z_i. In this parameterization, the target approximation has the form

$$R(\mu, \mathcal{A}) \approx Z_\theta D_\theta Z^*_\theta.$$

This direction connects generator learning with representation learning and with settings where only biased samples or black-box feedback are available.

Here $Z_\theta : \mathbb{R}^{m+1} \to \mathcal{W}^\mu_\pi$ is defined by neural features

$$Z_\theta u = \sum_{i=0}^{m} u_i z^\theta_i,$$

and D_θ is a diagonal matrix representing the leading resolvent eigenvalues. The neural network is trained so that the span of $\{z^\theta_i\}^m_{i=0}$ approximates the leading invariant subspace of the resolvent.

The population objective mirrors the Hilbert–Schmidt loss used above:

$$\mathcal{L}_\alpha(\theta) = \|R(\mu,\mathcal{A}) - Z_\theta D_\theta Z_\theta^*\|^2_{\mathrm{HS}(\mathcal{W}^\mu_\pi)} - \|R(\mu,\mathcal{A})\|^2_{\mathrm{HS}(\mathcal{W}^\mu_\pi)} + \alpha\,\|Z_\theta^* Z_\theta - I\|_F^2\,.$$

The first term asks the neural features to span a good low-rank approximation of the resolvent. The subtraction removes a constant independent of θ. The last term, controlled by $\alpha \geq 0$, encourages the learned features to be orthonormal in the energy geometry.

As in the finite-dimensional derivation, the loss can be rewritten using only matrices of inner products. Define

$$W_\theta = Z_\theta^* Z_\theta, \qquad C_\theta = Z_\theta^* R(\mu,\mathcal{A}) Z_\theta.$$

The energy identity turns C_θ into an $\mathcal{L}^2_\pi$-covariance matrix, whereas W_θ is the Sobolev energy covariance matrix involving both values and gradients of the neural features. Therefore the computable loss has the schematic form

$$\mathcal{L}_\alpha(\theta) = \mathrm{tr}\left\{\left((D_\theta W_\theta)^2\right)\right\} - 2\mathrm{tr}\left\{(D_\theta C_\theta)\right\} + \alpha\|W_\theta - I\|_F^2.$$

If the samples come from a biased distribution, the same matrices are estimated with the appropriate importance weights. This is the bridge between the matrix problem $G = W^{-1}C$ and the neural representation-learning objective: the finite dictionary is replaced by a trainable dictionary, but the geometry and the resolvent identity are the same.

4.3 Worked Example: Double-Well Potential

As a concrete one-dimensional example, consider the double-well potential:

$$U(x) = H(x^2-1)^2, \qquad U'(x) = 4Hx(x^2-1), \qquad x \in \mathbb{R}.$$

The corresponding overdamped Langevin dynamics is

$$dX_t = -4HX_t(X_t^2-1)\,dt + \sqrt{\frac{2}{\beta}}\,dB_t.$$

The potential has two global minima at $x = -1$ and $x = 1$, where $U(-1) = U(1) = 0$, and a local maximum at $x = 0$, where $U(0) = H$. Thus H is the energy barrier between the two wells, while the dimensionless quantity βH determines how difficult it is for the diffusion to move from one well to the other. This geometry is illustrated in fig. 4.1.

The computational illustrations in this section are generated from direct stationary samples and Euler–Maruyama simulations of this one-dimensional system.

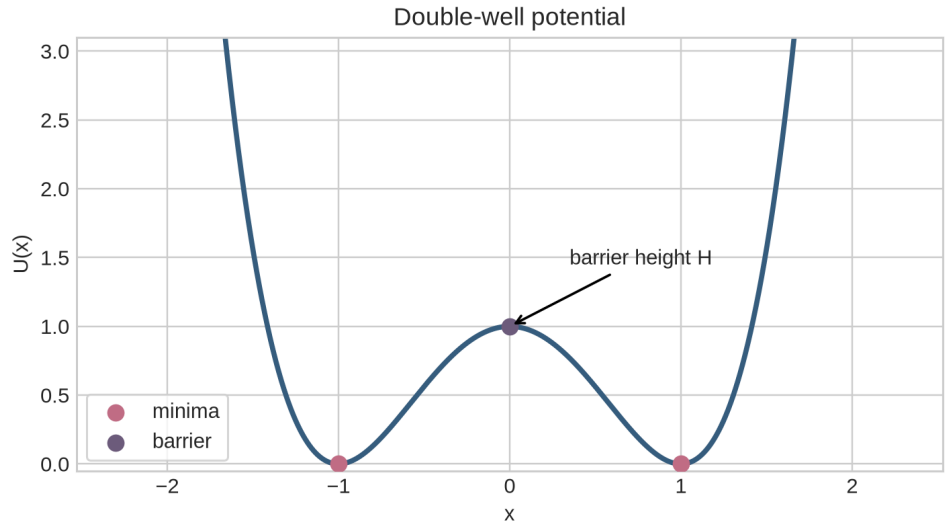

Figure 4.1: Double-well potential $U(x) = H(x^2-1)^2$. The two minima correspond to the metastable wells, and the barrier at $x = 0$ has height H.

The invariant distribution is

$$\pi(dx) = \frac{1}{Z_\beta} e^{-\beta H(x^2-1)^2}\, dx,$$

where Z_β is the normalizing constant. This density is bimodal, with one mode near each minimum of the potential. When βH is large, most of the mass is concentrated near -1 and 1, and transitions between the two wells become rare. Figure 4.2 compares this stationary density for two values of βH.

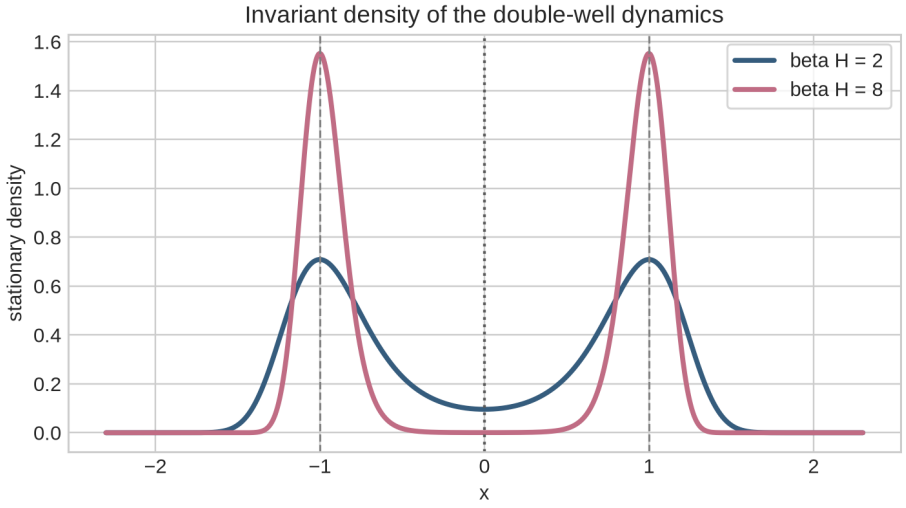

Figure 4.2: Invariant densities for two values of βH. Increasing βH concentrates the mass near the wells and makes the density near the barrier smaller.

Computationally, this example should exhibit two distinct time scales. On short time scales, a trajectory quickly relaxes toward the closest well. On much longer time scales, the trajectory occasionally crosses the barrier and switches wells. The slow spectral modes of the generator should

therefore capture the metastable transition between the two regions. The simulated trajectories in fig. 4.3 show this contrast.

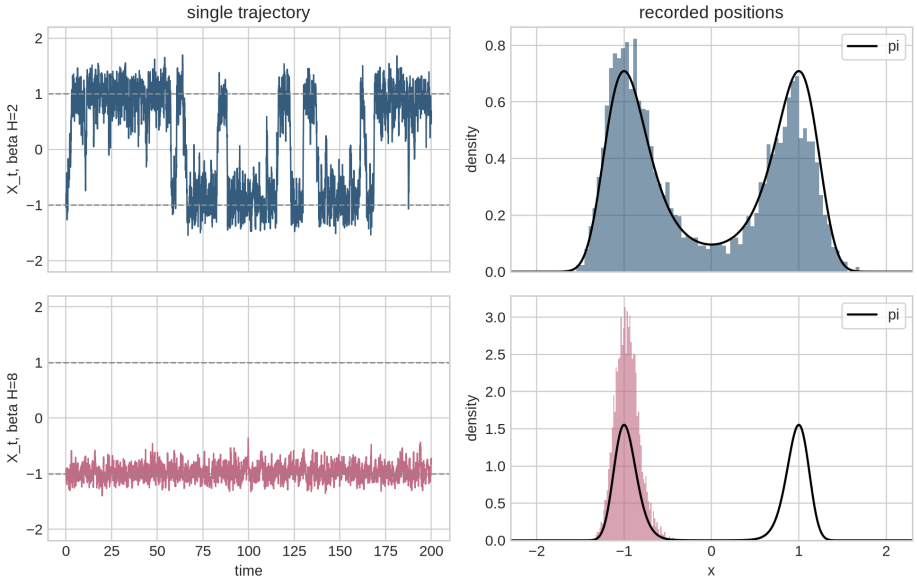

Figure 4.3: Euler–Maruyama simulations for two barrier regimes. For smaller βH, the trajectory moves between wells on the simulated time window. For larger βH, a single trajectory may remain trapped in one well even though the invariant density is bimodal.

For smooth observables $\phi : \mathbb{R} \to \mathbb{R}$, the generator is

$$\mathcal{A}\phi(x) = -4Hx(x^2 - 1)\phi'(x) + \beta^{-1}\phi''(x).$$

The constant function is the eigenfunction associated with the eigenvalue $\lambda_0 = 0$. The first non-trivial eigenfunction is expected to separate the two wells: it changes sign between the left and right modes and has an eigenvalue close to zero when the barrier is high. Higher eigenfunctions describe faster fluctuations inside the wells.

To apply the finite-dimensional method, choose a dictionary consisting of the constant feature $z_0(x) = 1$ and Gaussian radial features

$$z_i(x) = \exp\left(-\frac{(x - c_i)^2}{2\ell^2}\right), \qquad i = 1, \dots, m,$$

with centers c_i equally spaced over $[-2, 2]$. Given samples $x_1, \dots, x_n \sim \pi$, estimate the matrices

$$\widehat{C}_{ij} = \frac{1}{n}\sum_{k=1}^{n} z_i(x_k) z_j(x_k)$$

and

$$\widehat{W}_{ij} = \frac{1}{n}\sum_{k=1}^{n}\left[\mu z_i(x_k) z_j(x_k) + \frac{1}{\beta} z_i'(x_k) z_j'(x_k)\right].$$

The empirical resolvent estimator is then

$$\widehat{G} = \widehat{W}^{-1}\widehat{C},$$

or its regularized analogue when $\widehat{W}$ is ill-conditioned. If $\widehat{G}u_i = \widehat{\nu}_i u_i$, then

$$\widehat{\psi}_i = \sum_{j=0}^{m} (u_i)_j z_j \qquad \text{and} \qquad \widehat{\lambda}_i = \mu - \frac{1}{\widehat{\nu}_i}$$

give approximate eigenfunctions and eigenvalues of the Langevin generator. In the numerical experiment below, we take $H = 1$, $\beta = 2$, and therefore $\beta H = 2$. This lower effective barrier is deliberate: it makes the stationary sample cover both wells, instead of being dominated by the mixing difficulty of a single long trajectory. Concretely, we used $n = 8000$ representative points obtained from inverse-CDF quantiles of π, rather than samples from one simulated path. The leading non-constant eigenfunction then reveals the transition coordinate between the two wells. The corresponding estimated eigenvalues are reported in table 4.1, and the leading eigenfunctions are shown in fig. 4.4.

mode i	resolvent eigenvalue $\widehat{\nu}_i$	generator eigenvalue $\widehat{\lambda}_i$
0	1.0000	0.0000
1	0.8148	-0.2273
2	0.1969	-4.0783
3	0.1162	-7.6041
4	0.0772	-11.9517
5	0.0548	-17.2475
6	0.0412	-23.3012
7	0.0322	-30.0395

Table 4.1: Estimated spectral values for the finite-dimensional resolvent approximation with $H = 1$, $\beta = 2$, $\beta H = 2$, $\mu = 1$, and a representative stationary sample of size $n = 8000$. The generator eigenvalues are recovered through $\widehat{\lambda}_i = \mu - \widehat{\nu}_i^{-1}$.

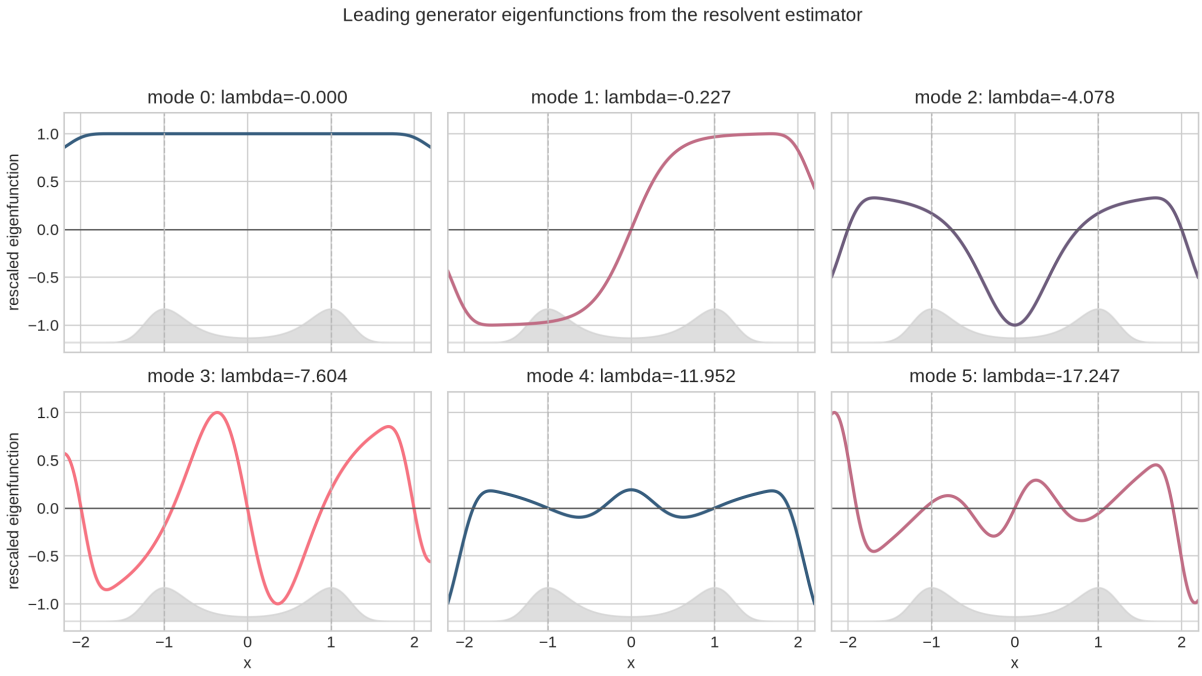

Figure 4.4: First six eigenfunctions obtained from the finite-dimensional resolvent approximation for $\beta H = 2$. The gray density in each panel shows the stationary distribution used to construct the representative sample. The constant eigenfunction corresponds to $\lambda_0 = 0$, while mode 1 separates the two wells and represents the slow transition coordinate.

Finally, we can repeat the same stationary-sample procedure for several values of the effective barrier βH. In this comparison, $H = 1$ is kept fixed and $\beta = \beta H$ is varied. For each value, the sample used by the algorithm is constructed directly from the stationary density by inverse-CDF

quantiles. This avoids confusing the spectral estimate with the poor mixing of a single trajectory at large βH.

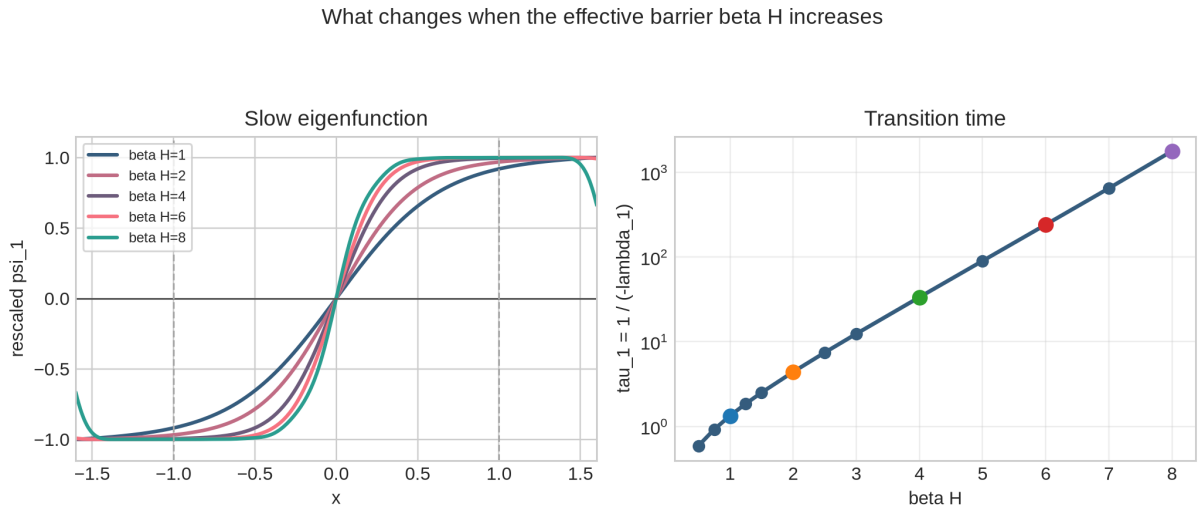

Figure 4.5: Effect of increasing the effective barrier βH. Left: the first non-constant eigenfunction becomes the slow coordinate separating the two wells. Right: the associated transition time $\tau_1 = 1/(-\widehat{\lambda}_1)$ increases rapidly.

This sweep makes the metastability mechanism visible. As βH increases, the first non-constant eigenfunction becomes closer to a signed indicator of the left and right wells, and the transition time

$$\tau_1 = \frac{1}{-\widehat{\lambda}_1}$$

grows rapidly. This qualitative picture is summarized in fig. 4.5; the same spectral trend is shown more directly in fig. 4.6.

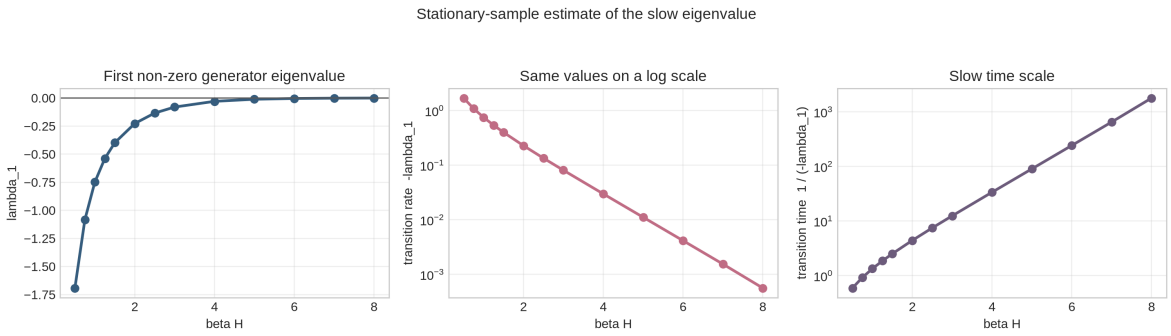

Figure 4.6: Estimated first non-zero generator eigenvalue as a function of the effective barrier βH. As the barrier increases, $\widehat{\lambda}_1$ approaches zero, meaning that the transition between wells becomes slower. The middle panel shows the corresponding transition rate $-\widehat{\lambda}_1$, while the right panel shows the transition time $1/(-\widehat{\lambda}_1)$.

4.4 Biased Dynamics

In this section we follow [Devergne et al., 2024, Kostic et al.,]. In the previous example, since the potential U is explicitly known and the state space is one-dimensional, it is relatively easy to

generate samples from its Boltzmann distribution

$$\pi(dx) = \frac{1}{Z} e^{-\beta U(x)}\, dx.$$

This allowed us to approximate the covariance and energy matrices directly from samples distributed according to π.

In realistic applications, however, knowing the potential U does not necessarily imply that samples from π can be generated efficiently. For instance, when U contains several metastable wells separated by high energy barriers, a trajectory following the target dynamics may remain trapped in one region of the state space for a long time. Consequently, even a large dataset may contain only a small number of transitions between the relevant metastable states.

A possible strategy is therefore to generate samples from a different distribution π', chosen so that the relevant regions of the state space are explored more efficiently. We assume that

$$\pi \ll \pi',$$

so that expectations under the target distribution can be recovered from samples distributed according to π' by importance sampling. More precisely, for every suitable function f,

$$\mathbb{E}_{X\sim\pi}[f(X)] = \mathbb{E}_{X'\sim\pi'}\left[\frac{d\pi}{d\pi'}(X') f(X')\right].$$

Suppose that π' is also a Boltzmann distribution at inverse temperature β, associated with a modified potential U':

$$\pi'(dx) = \frac{1}{Z'} e^{-\beta U'(x)}\, dx.$$

The potential U' can be written as

$$U'(x) = U(x) + V(x),$$

where V is called the bias potential. The corresponding sampling distribution is therefore

$$\pi'(dx) = \frac{1}{Z'} e^{-\beta(U(x)+V(x))}\, dx.$$

The Radon–Nikodym derivative between the target and sampling distributions is then given by

$$\frac{d\pi}{d\pi'}(x) = \frac{Z'}{Z} e^{\beta V(x)} = \frac{e^{\beta V(x)}}{\mathbb{E}_{X'\sim\pi'}[e^{\beta V(X')}]},$$

and hence

$$\mathbb{E}_{X\sim\pi}[f(X)] = \frac{\mathbb{E}_{X'\sim\pi'}\left[e^{\beta V(X')} f(X')\right]}{\mathbb{E}_{X'\sim\pi'}\left[e^{\beta V(X')}\right]}.$$

Thus, the samples are generated according to π', while the importance weights

$$w(x) \coloneqq e^{\beta V(x)}$$

allow us to reconstruct quantities defined with respect to the target distribution π.

The distribution π' is the invariant distribution of the biased overdamped Langevin dynamics

$$dX'_t = -\nabla U'(X'_t)\, dt + \sqrt{2\beta^{-1}}\, dW_t,$$

or, equivalently,

$$dX'_t = -\nabla\big(U + V\big)(X'_t)\, dt + \sqrt{2\beta^{-1}}\, dW_t.$$

Its infinitesimal generator is

$$\mathcal{L}'f = -\langle \nabla(U+V), \nabla f\rangle + \beta^{-1}\Delta f.$$

Recalling that the target generator is

$$\mathcal{L}f = -\langle \nabla U, \nabla f\rangle + \beta^{-1}\Delta f,$$

we obtain

$$\boxed{\mathcal{L}' = \mathcal{L} - \langle \nabla V, \nabla(\,\cdot\,)\rangle.}$$

Therefore, choosing a different sampling distribution corresponds, at the dynamical level, to adding a bias potential V to the original potential. The biased dynamics is used only to generate data more efficiently, while the importance weights are used to recover the geometry associated with the original distribution π and, consequently, to learn the spectral properties of the original generator $\mathcal{L}$.

To recover the target dynamics from samples generated by the biased process, we replace the expectations with respect to π by weighted expectations with respect to π'. Let $z = (z_1, \dots, z_m)^\top$ be a prescribed dictionary of functions and let

$$w(x) = e^{\beta V(x)}.$$

Given biased samples $x'_1, \dots, x'_n$, define

$$\widehat{C}_{ij} = \frac{1}{n}\sum_{k=1}^{n} w(x'_k) z_i(x'_k) z_j(x'_k)$$

and

$$\widehat{W}_{ij} = \widehat{\mathcal{E}}^{\eta}_{\pi'}\left[\sqrt{w}\, z_i, \sqrt{w}\, z_j\right].$$

For the overdamped Langevin dynamics, the latter is given explicitly by

$$\widehat{W}_{ij} = \frac{1}{n}\sum_{k=1}^{n}\left[\eta w(x'_k) z_i(x'_k) z_j(x'_k) + \beta^{-1}\nabla\left(\sqrt{w}\, z_i\right)(x'_k)^\top \nabla\left(\sqrt{w}\, z_j\right)(x'_k)\right].$$

These matrices approximate the covariance and energy matrices associated with the original distribution π, although the samples were generated from π'. The resolvent of the target generator is then estimated by

$$\widehat{G}_{\eta,\gamma} = \left(\widehat{W} + \eta\gamma I\right)^{-1}\widehat{C}.$$

Finally, if

$$\widehat{G}_{\eta,\gamma}\widehat{v}_i = \widehat{\nu}_i \widehat{v}_i,$$

the eigenpairs of the original generator are estimated as

$$\widehat{\lambda}_i = \eta - \frac{1}{\widehat{\nu}_i}, \qquad \widehat{f}_i(x) = z(x)^\top \widehat{v}_i.$$

Thus, the biased dynamics is used to explore the state space, while importance weighting allows us to recover the spectral information of the unbiased generator.

Bibliography

[Bakry et al., 2014] Bakry, D., Gentil, I., and Ledoux, M. (2014). *Analysis and Geometry of Markov Diffusion Operators*, volume 348 of *Grundlehren der mathematischen Wissenschaften*. Springer International Publishing, Cham.

[Chewi, 2024] Chewi, S. (2024+). *Log-Concave Sampling*. `https://chewisinho.github.io/main.pdf`. Book draft.

[Davies, 1995] Davies, E. B. (1995). *Spectral Theory and Differential Operators*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge.

[Devergne et al., 2024] Devergne, T., Kostic, V., Parrinello, M., and Pontil, M. (2024). From Biased to Unbiased Dynamics: An Infinitesimal Generator Approach. arXiv:2406.09028 [cs.LG].

[Kostic et al.,] Kostic, V. R., Lounici, K., Halconruy, H., Devergne, T., Parrinello, M., and Pontil, M. Langevin-Informed Transfer Learning: Replacing Target Samples by Black-Box Feedback.

[Kostic et al., 2024] Kostic, V. R., Lounici, K., Halconruy, H., Devergne, T., and Pontil, M. (2024). Learning the Infinitesimal Generator of Stochastic Diffusion Processes. *Advances in Neural Information Processing Systems*, 37:137806–137846.

[Mohri et al., 2012] Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. The MIT Press.

[Särkkä and Solin, 2019] Särkkä, S. and Solin, A. (2019). *Applied Stochastic Differential Equations*. Cambridge University Press, 1 edition.