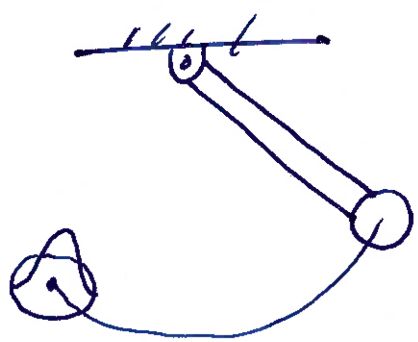
Today

- Integrating probability into learned models
- Stochastic dynamics
- Maximum likelihood
- Markov processes
- Euler – Maruyama

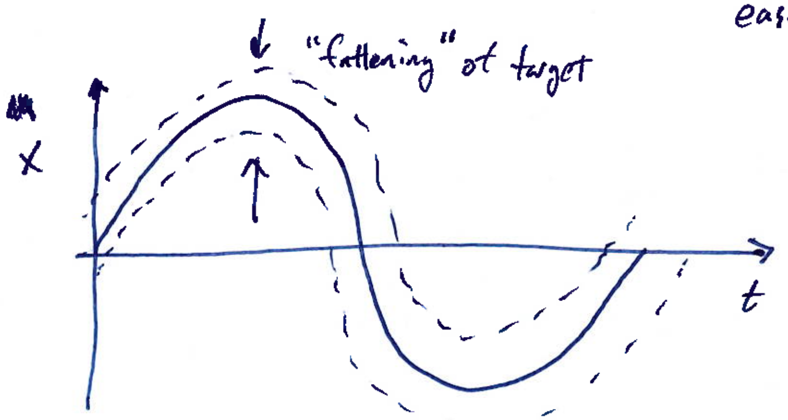
Motivation



- Tracking a point vs tracking a "blob" of probability
- Sensor noise, may have either noisy observations or noisy physics } aleatoric uncertainty
- or an incomplete description of physics (epistemic uncertainty)
- Even w/o noise, training a model in a probabilistic context is often easier
  → probabilistic regularization

$$\dot{x} = f(x) + g(x;\theta)$$

random variable



References for probability

- 'Probability Essentials' Jacod + Protter
  - → Short paperback, rigorous but quick definitions
- 'Probability: theory + examples' Durrett
  - → Measure theoretic → gnarly to get started from
- 'Machine learning, a probabilistic perspective' Kevin Murphy
  - → Accessible, lots of background and follow-on refs
  - like wikipedia for ML (no ODEs/PDEs though)
- 'Introduction to stochastic integration' Kuo
  - → where I'll pull refs from

Overall → I wouldn't recommend following a text for this course, because there isn't a good one for scientific computing + ML

Probability review

def a continuous random variable $\mathbb{X}$ takes a random value $x \in \mathbb{R}$

def a cumulative distribution function (CDF) defines probability over a range of values

$$F_{\mathbb{X}}(x) = \mathbb{P}(\mathbb{X} \le x)$$

def if CDF is differentiable, define probability density function

$$f_{\mathbb{X}}(x) = \frac{d}{dx} F_{\mathbb{X}}(x)$$

so $$\mathbb{P}(a \le X \le b) = \int_a^b f(x)\, dx = F(b) - F(a)$$

Drop subscripts unless necessary

We'll talk about different RVs + events by different notions of pdfs:

Joint distribution $$f(x_1, \dots x_N) = \mathbb{P}(\mathbb{X}_1 = x_1, \dots \mathbb{X}_N = x_N)$$

Marginalization $$f(\mathbb{X} = x) = \sum_y f(\mathbb{X} = x, \mathbb{Y} = y)$$

Rule of total probability

conditional dist $$f(\mathbb{Y} = y \mid \mathbb{X} = x) = \frac{f(\mathbb{X} = x, \mathbb{Y} = y)}{f(\mathbb{X} = x)}$$

product rule $$f(x, y) = f(x \mid y)\, f(y)$$

prob. chain rule

$$f(x_1, \dots x_N) = f(x_2, \dots x_N \mid x_1)\, f(x_1)$$
$$= f(x_3, \dots x_N \mid x_1, x_2)\, f(x_2 \mid x_1)\, f(x_1)$$
$$\vdots$$
$$= f(x_N \mid x_1, \dots x_{N-1}) \cdots f(x_2 \mid x_1)\, f(x_1)$$

Marginal indep

$x \perp y$ if $f(x,y) = f(x)\, f(y)$

or in general $f(x_1, \dots x_N) = \prod_{i=1}^{N} f(x_i)$

conditional indep

$x \mid z \perp y \mid z$ if $f(x, y \mid z) = f(x \mid z)\, f(y \mid z)$

Fitting distributions to data w/ MLE

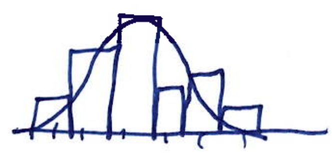
Given a dataset consisting of observations of a RV X

$$\mathcal{D} = \{x_i\}_{i=1}^{N_{data}}$$

Define the likelihood by evaluating the parameterized joint distribution $\cancel{f(x,\theta)}$ $\mathcal{L} = f(X_1, \dots X_{N_{data}} | \theta)$

Fit the distribution to data by solving

$$\theta^* = \underbrace{- \log \mathcal{L}(x_1, \dots x_{N_{data}} | \theta)}_{\text{NLL (negative log likelihood)}}$$



Example

- Assume independent data

$$f(x_1, \dots x_N | \theta) = \prod_{i=1}^{N} f(x_i | \theta)$$

- Choose marginal distribution

$$f(x_i | \theta) = N(x; \mu, \sigma^2) \quad \rightarrow \theta = \{\mu, \sigma^2\}$$

$$N(x; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

$$\log N = -C - \frac{1}{2}\log\sigma^2 - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$$

$$NLL = \sum_i \frac{1}{2}\log\sigma^2 + \frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2$$

- Solve for θ^*

$$0 = \nabla_\mu NLL = \sum_i \nabla_\mu \frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2$$

$$0 = -\sum_i (x_i - \mu)$$

$$\sum_i \mu = \sum_i x_i$$

$$N\mu = \sum_i x_i$$

$$\mu = \frac{\sum_i x_i}{N}$$

similarly $\sigma^2 = \dfrac{\sum_i (x_i-\mu)^2}{N}$

Stochastic differential equations

To account for random forcing/physics we'll expand our notion of ODEs

First, write

$$\frac{dx}{dt} = f(x,t)$$

Instead as

$$\int_0^t dx = \int_0^t f(x,t)\,dt$$

For short hand people omit the limits of integration

$$dx_t = f(x_t)\,dt \quad , \; x_t = x(t)$$

We will see that random forcing leads to solutions which aren't differentiable, so its necessary to intepret in integral sense.

To account for stochastic terms, we will consider a SDE

$$dx_t = \underbrace{f(x_t, t)\,dt}_{\text{drift}} + \underbrace{g(x_t, t)\,dW_t}_{\text{diffusion}}$$

stochastic process

Need to define dW_t and how to integrate it.
In general, we could spend a whole course studying stochastic processes – we'll give an accelerated version here.

We'll consider the Wiener process $W_t = \int_0^t dW_t$

defined via ① $W_0 = 0$

② W has independent increments for $t > 0, u \geq 0$

$W_{t+u} - W_t$ are independent of W_s, for any $s < t$

③ W has Gaussian increments w/ variance equal to time increment

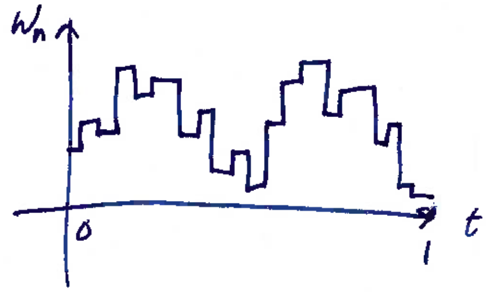
$$W_{t+u} - W_t \sim N(0, u)$$

④ W is continuous

Many different constructions of W_t that satisfy these processes, e.g.

Let $\xi_1, \xi_2, \ldots$ be IID RVs w/ $\mathbb{E}[\xi_i] = 0$, $\text{var}[\xi_i] = 1$

$$W_n(t) = \frac{1}{\sqrt{n}} \sum_{1 \leq k \leq \lfloor nt \rfloor} \xi_k$$



Then $\lim_{n \to \infty} W_n(t) = W_t$

To integrate against something like this, consider the Riemann construction

(*) $$\int_0^t g(x_t,t)\,dW_t = \lim_{\Delta t \to 0} \sum_{i=1}^{n-1} g(x_{t_i},t_i)\left(dW_{t_{i+1}} - dW_{t_i}\right)$$

It turns out that this leads to many pathological issues that violate the usual assumptions from standard calculus.

1. We can prove that w/ probability 1, W_t is continuous
2. Also w/ prob 1, W_t is <u>nowhere</u> differentiable

In a probability class we would get into details – but this is enough to pose a simple scheme for solving SDEs in our learning problems by discretizing the above eq (*)

Euler-Maruyama Method

Given SDE

$$dX_t = f(X_t,t)\,dt + g(X_t,t)\,dW_t$$

Solve, for $X_{t_n} = X(t=nk)$

for $n = 0, 1, \ldots$

$$X_{n+1} = X_n + k\,f(X_{t_n},t_n) + \xi_n\,g(X_{t_n},t_n)$$
$$\xi_n \sim N(0,k)$$

Some better integrators (see Milstein's method) but require a deeper dive

Using maximum likelihood we can use a <u>Markov process</u> to fit an SDE to data

<u>def</u> For a K^{th}-order Markov process, given a discrete time series $\vec{y} = \langle y_0, y_1, \dots y_N \rangle$

$$P(y_i \mid y_0, \dots y_{i-1}) = P(y_i \mid y_{i-1}, \dots y_{i-K})$$

This means you only need to model the last K time steps → this is like a multi-step integrator w/ K steps.

Euler-Maruyama is Markov w/ $K=1$

$$P(X_{n+1} \mid X_n) = N\left(X_n + K f(X_n, t_n),\ K g(X_n, t_n)^2\right)$$

<u>Pf</u> If $x \sim N(\mu, \sigma^2)$

$Ax + b \sim N(A\mu + b,\ A^2\sigma^2)$

Identify $b = x_n + K f_n$

$A = g$

$\sigma^2 = K$

Now we can go back and derive a NLL

swap y's w/ x's on board...

$$-\log p(y_1, \dots y_N) = -\log p(y_N | y_{N-1})\, p(y_{N-1} | y_{N-2}) \cdots$$

$$= -\log \prod_{i=1}^{N} p(y_i | y_{i-1})$$

$$= -\sum_{i=1}^{N} \log p(y_i | y_{i-1})$$

$$= -\sum_{i=1}^{N} \log N(x_i + \kappa f_i,\ \kappa g_i^2)$$

$$= C + \sum_{i=1}^{N} \frac{1}{2} \log \kappa g_i^2 + \frac{1}{2} \frac{(x_{i+1} - x_i - \kappa f_i)^2}{\kappa g_i^2}$$

If we replace $f(x,t)$ w/ $f(x,t;\theta)$ i.e. make the SDE trainable
$g(x,t)$ $g(x,t;\theta)$

We can solve

$$\min_{\theta} \sum_{i=1}^{N} \log \kappa g_{i,\theta}^2 + \frac{(x_{i+1} - x_i - \kappa f_{i,\theta})^2}{\kappa g_{i,\theta}^2}$$

Example of structure preserving SDEs

Metriplectic / GENERIC formalism

$$dx_t = \left(L\,\partial_x E + M\,\partial_x S + k_B\,\partial_x \cdot M \right) dt + \sqrt{2 k_B M}\, dW_t$$

where
- $E \to$ energy
- $S \to$ entropy
- $L = -L^T$
- $M = M^T$ SPD
- $k_B \to$ Boltzmann constant

\+ Degeneracy Condition

$^*L\,\partial_x S = M\,\partial_x E = 0$

- When "coarse-graining" a reversible system, the loss of information manifests as a entropy increase, or introduction of dissipation
- To counter this, stochastic forcing does work on the system.
- Exact fluctuation dissipation theorem guarantees this holds rigorously

Deterministic limit ($k_B \to 0$)

Thm Energy is conserved

Pf $$\frac{dE}{dt} = \partial_x E^T \frac{dx}{dt} \quad \text{(chain rule)}$$

$$= \partial_x E^T (L \partial_x E + M \partial_x S)$$

$$= \cancel{\partial_x E^T L \partial_x E} + \cancel{\partial_x S^T M \partial_x E}$$

By skew-symmetry   By degen condition

Remark Metriplectic is therefore a generalization of Hamiltonian mechanics, with degen condition killing off cross terms between rev and irrev parts.

Thm Entropy is non-decreasing

Pf $$\cancel{\frac{\partial S}{\partial t}} \frac{dS}{dt} = \partial_x S^T \frac{dx}{dt}$$

$$\cancel{= \partial_x S^T M \partial_x E + \partial}$$

$$= \underbrace{\partial_x E^T L \partial_x S}_{=0} + \underbrace{\partial_x S^T M \partial_x S}_{\geq 0} \geq 0$$