Stochastic Filtering

Salvador Ortiz-Latorre

University of Oslo

MAT4790/9790
Spring 2026

Introduction I

- Consider a random walk on $\mathbb{Z}$, i.e.,

$$X_j = X_{j-1} + \varepsilon_j, \qquad j \in \mathbb{N}, \qquad X_0 = 0,$$

where ε_j is a sequence of i.i.d. random signs $P(\varepsilon_j = \pm 1) = 1/2$.

- Suppose the position of the particle at time j is to be guessed on the basis of noisy observations

$$Y_i = X_i + \xi_i, \qquad i = 1, ..., j,$$

where ξ_i is a sequence of i.i.d. random variables (white noise) with Gaussian distributions

$$P(\xi_i \in [a, b]) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{u^2}{2}} du, \qquad i \in \mathbb{N}.$$

Introduction II

- An estimate is a rule assigning a real number to any outcome of

$$Y_{[1,j]} := Y_{1:j} := (Y_1, ..., Y_j),$$

that is, a map $\varphi_j(y) : \mathbb{R}^j \to \mathbb{R}$.

- **How guesses are compared?** φ_j is better than ψ_j if

$$\mathbb{E}\left[(X_j - \varphi_j(Y_{1:j}))^2\right] \leq \mathbb{E}\left[(X_j - \psi_j(Y_{1:j}))^2\right], \qquad (0.1)$$

that is, $\varphi_j(Y_{1:j})$ has smaller square error on average. Note that, even if (0.1) holds, one may have

$$(X_j(\omega) - \varphi_j(Y_{1:j}(\omega)))^2 > (X_j(\omega) - \psi_j(Y_{1:j}(\omega)))^2,$$

for some experiment/realization $\omega \in \Omega$.

- Once the criterion (0.1) is accepted, we want to find the best (optimal) estimate.

Introduction III

- Simple guess

$$\tilde{X}_j := \tilde{\varphi}_j\left(Y_{1:j}\right) = Y_j, \qquad \textbf{(last observation)}.$$

The corresponding mean square error (MSE) is

$$\tilde{P}_j = \mathbb{E}\left[\left(X_j - Y_j\right)^2\right] = \mathbb{E}\left[\left(X_j - X_j - \xi_j\right)^2\right] = \mathbb{E}\left[\xi_j^2\right] = 1.$$

This estimate does not take into account past observations.

- **Second order moving average**

$$\tilde{\tilde{X}}_j = \frac{Y_j + Y_{j-1}}{2}.$$

Its MSE is given by

$$\tilde{\tilde{P}}_j = \mathbb{E}\left[\left(X_j - \frac{Y_j + Y_{j-1}}{2}\right)^2\right] = \mathbb{E}\left[\left(X_j - \frac{X_j + X_{j-1} + \xi_j + \xi_{j-1}}{2}\right)^2\right]$$

Introduction IV

$$= \mathbb{E}\left[\left(\frac{X_j - X_{j-1}}{2} - \frac{\xi_j + \xi_{j-1}}{2}\right)^2\right] = \mathbb{E}\left[\left(\frac{\varepsilon_j}{2} - \frac{\xi_j + \xi_{j-1}}{2}\right)^2\right]$$
$$= \frac{1}{4}\left\{\mathbb{E}\left[(\varepsilon_j)^2\right] + \mathbb{E}\left[(\xi_j)^2\right] + \mathbb{E}\left[(\xi_{j-1})^2\right]\right\} = \frac{3}{4} = 0.75,$$

so there is an improvement of 25% with respect to $\tilde{X}_j$.

- Let us increase the "**memory**" of the estimate by considering the **third order moving average.** Then, we get that

$$\mathbb{E}\left[\left(X_j - \frac{Y_j + Y_{j-1} + Y_{j-2}}{3}\right)^2\right] = \ldots = \mathbb{E}\left[\left(\frac{2\varepsilon_j + \varepsilon_{j-1}}{3} - \frac{\xi_j + \xi_{j-1} + \xi_{j-2}}{3}\right)^2\right]$$
$$= \frac{4}{9} + \frac{1}{9} + \frac{3}{9} = \frac{8}{9} = 0.89$$

The MSE has increased!!! The problem is that we give the same weights to the "new" and "old" measurements.

Introduction V

- **How can we optimally weight the data?** One can prove that the optimal estimate can be generated "very efficiently" by the difference equation

$$\hat{X}_j = \hat{X}_{j-1} + P_j \left(Y_j - \hat{X}_{j-1} \right), \qquad X_0 = 0, \tag{0.2}$$

where P_j is the sequence of numbers

$$P_j = \frac{P_{j-1} + 1}{P_{j-1} + 2}, \qquad P_0 = 0. \tag{0.3}$$

- Let us calculate the MSE for $\hat{X}_j$. The sequence $\Delta_j := X_j - \hat{X}_j$ satisfies

$$\begin{aligned}
\Delta_j &= \Delta_{j-1} - \left(X_{j-1} - \hat{X}_{j-1} \right) + X_j - \hat{X}_j \\
&= \Delta_{j-1} + \varepsilon_j - P_j \left(Y_j - \hat{X}_{j-1} \right) \\
&= \Delta_{j-1} + \varepsilon_j - P_j \left(X_j - \hat{X}_{j-1} + \xi_j \right) \\
&= \Delta_{j-1} + \varepsilon_j - P_j \left(\Delta_{j-1} + \xi_j + \varepsilon_j \right) \\
&= \left(1 - P_j \right) \Delta_{j-1} + \left(1 - P_j \right) \varepsilon_j - P_j \xi_j .
\end{aligned}$$

Introduction VI

- Therefore, using the independence between ε_j, ζ_j and Δ_{j-1}, we get

$$\hat{P}_j := \mathbb{E}\left[\Delta_j^2\right] = (1-P_j)^2 \hat{P}_{j-1} + (1-P_j)^2 + P_j^2, \qquad \hat{P}_0 = 0.$$

- Note that the sequence P_j satisfies (manipulating equation $(0.3)$and multiplying the equation by $(1-P_j)$)

$$P_j = (1-P_j)^2 P_{j-1} + (1-P_j)^2 + P_j^2, \qquad P_0 = 0.$$

- Hence, the difference $\hat{P}_j - P_j$ satisfies the equation

$$\hat{P}_j - P_j = (1-P_j)^2 \left(\hat{P}_{j-1} - P_{j-1}\right), \qquad j \in \mathbb{N},$$

and, since $\hat{P}_0 - P_0 = 0$, it follows that $\hat{P}_j - P_j \equiv 0, j \in \mathbb{N}$, and we have just proved that the MSE corresponding to $\hat{X}_j$ is given by P_j.

Introduction VII

- Numerically we get

j	1	2	3	4	5
P_j	0.5	0.6	0.6154	0.6176	0.618

- Moreover, P_j converges to the limit $P_\infty = (\sqrt{5}-1)/2 \simeq 0.618$, which is the unique positive root of

$$P = \frac{P+1}{P+2}.$$

- It is close to a 40% improvement in the MSE compared to the MSE of $\tilde{X}_j$.
- No further improvement is possible among linear estimates.
- **What about nonlinear estimates?**

Introduction VIII

- The "simplest" nonlinear estimate could be

$$\dot{X}_1 := \text{sign}\left(Y_1\right) = \begin{cases} 1 & \text{if} \quad Y_1 \geq 0 \\ -1 & \text{if} \quad Y_1 < 0 \end{cases}.$$

The MSE is computed as follows

$$\begin{aligned} \dot{P}_1 &= \mathbb{E}\left[\left(X_1 - \dot{X}_1\right)^2\right] = \mathbb{E}\left[\left(X_1 - \text{sign}\left(X_1 + \xi_1\right)\right)^2\right] \\ &= \frac{1}{2}\mathbb{E}\left[\left(1 - \text{sign}\left(1 + \xi_1\right)\right)^2\right] + \frac{1}{2}\mathbb{E}\left[\left(-1 - \text{sign}\left(-1 + \xi_1\right)\right)^2\right] \\ &= \frac{1}{2}\left(2\right)^2 P\left(\xi_1 \leq -1\right) + \frac{1}{2}\left(2\right)^2 P\left(\xi_1 \geq 1\right) = 4P\left(\xi_1 \leq -1\right) \\ &= 4\Phi\left(-1\right) \simeq 0.6346. \end{aligned}$$

It is worse than the MSE for $\hat{X}_1$!!!

Introduction IX

- Let us try the estimate $\bar{X}_1 := \tanh(Y_1) = \frac{e^{2Y_1}-1}{e^{2Y_1}+1}$.
  The MSE is given by

$$\begin{aligned}\bar{P}_1 &= \mathbb{E}\left[(X_1 - \bar{X}_1)^2\right] \\ &= \frac{1}{2}\int_{\mathbb{R}} \left\{(1-\tanh(u+1))^2 + (1+\tanh(u-1))^2\right\} \frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}} du \\ &\simeq 0.4496,\end{aligned}$$

  which is the best so far (indeed it is the best possible)

- **Can we compute the best nonlinear estimate of X_j efficiently (recursively)?**

Introduction X

- It turns out that it is possible. Let $\rho_j(i), i \in \mathbb{Z}, j \in \mathbb{Z}_+$ be generated by the nonlinear recursion

$$\rho_j(i) = \exp\left(iY_j - \frac{i^2}{2}\right)\{\rho_{j-1}(i-1) + \rho_{j-1}(i+1)\}, \qquad j \geq 1,$$

subject to $\rho_0(0) = 1$ and $\rho_0(i) = 0, i \neq 0$. Then, the best estimate of X_j from $Y_{1:j}$ is given by

$$\bar{X}_j = \frac{\sum_{i=-\infty}^{+\infty} i\rho_j(i)}{\sum_{i=-\infty}^{+\infty} \rho_j(i)}.$$

One can check that $\bar{P}_{100}$ is around 0.54, which is better than $\hat{P}_\infty$.

- We can think about the analogous problem in continuous time.

Introduction XI

- Intuitively, $X_t, t \geq 0$ is a continuous time Markov chain satisfying

$$P\left(X_{t+\varepsilon} = i \middle| X_s, 0 \leq s \leq t\right) = \begin{cases} 1 - 2\varepsilon & i = X_t \\ \varepsilon & i = X_t \pm 1 \\ 0 & \text{otherwise} \end{cases},$$

  for sufficiently small $\varepsilon > 0$. One way of building a process X satisfying the previous condition is to consider two independent Poisson processes Π^+ and Π^- and define

$$X_t = \Pi_t^+ - \Pi_t^-, \qquad t \geq 0.$$

- To define Y_t, the analogue of Y_j in continuous time, is much more delicate. The reason is that then concept of "white noise" needs to be introduced in continuous time.

Introduction XII

- The term "white noise" comes from the fact that the spectral density of i.i.d. random variables $\{\xi_j\}_{j\geq 1}$ is flat, i.e„

$$S_\xi(\lambda) = \sum_{j\in\mathbb{Z}} \mathbb{E}[\xi_0\xi_j]\, e^{-i\lambda j} = \sum_{j\in\mathbb{Z}} \delta_0(j)\, e^{-i\lambda j} = 1,$$

  for all $\lambda \in (-\pi, \pi]$.

- Any random sequence with flat spectral density is called (discrete time) white noise and its variance is recovered by integration of the spectral density

$$\mathbb{E}[\xi_j^2] = \frac{1}{2\pi}\int_{-\pi}^{\pi} 1 d\lambda = 1.$$

- In continuous time this definition leads to

$$\mathbb{E}[\xi_t^2] = \frac{1}{2\pi}\int_{-\infty}^{+\infty} 1 d\lambda = +\infty.$$

Introduction XIII

- We can solve this problem by defining the observation process as

$$Y_t = \int_0^t X_s ds + W_s, \tag{0.4}$$

where W is a Wiener process.

- **Why the model (0.4) is compatible with the notion of "white noise"?**
- Define the family of processes

$$\nu_t^\Delta = \frac{W_t - W_{t-\Delta}}{\Delta}, \qquad \Delta > 0.$$

Then, $\mathbb{E}\left[\nu_t^\Delta\right] = 0$ and

$$\mathbb{E}\left[\nu_t^\Delta \nu_s^\Delta\right] = \frac{1}{\Delta^2}\mathbb{E}\left[\left(W_t - W_{t-\Delta}\right)\left(W_s - W_{s-\Delta}\right)\right]$$

Introduction XIV

$$= \frac{1}{\Delta^2} \begin{cases} \Delta - |t-s| & \text{if} \quad |t-s| \leq \Delta \\ 0 & \text{if} \quad |t-s| > \Delta \end{cases}.$$

Hence, the process ν_t^Δ is stationary with autocorrelation function

$$R_\nu^\Delta(\tau) = \frac{1}{\Delta^2} \begin{cases} \Delta - |\tau| & \text{if} \quad |\tau| \leq \Delta \\ 0 & \text{if} \quad |\tau| > \Delta \end{cases}.$$

For small $\Delta > 0$, $R_\nu^\Delta(\tau)$ approximates the Dirac $\delta(\tau)$ function in the sense that for any $\varphi \in C_0(\mathbb{R})$ we have

$$\int_{-\infty}^{+\infty} \varphi(\tau) R_\nu^\Delta(\tau) \, d\tau \underset{\Delta \to 0}{\longrightarrow} \varphi(0),$$

and if the limit process $\nu := \lim_{\Delta \to 0} \nu$ existed, it would have flat spectral density. Then, the observation process in (0.4) would contain the same information as

$$\dot{Y}_t = X_t + \nu_t,$$

Introduction XV

with ν being the derived noise. This is only an intuition because ν does not exist as a limit in any reasonable sense (its variance explodes at any time t).

- **What would be the analogues of the filtering equations** (0.3)**?**
- Instead of difference equations we get differential equations, e.g.,

$$\dot{\hat{X}} = P_t \left(\dot{Y}_t - \hat{X}_t \right), \qquad \hat{X}_0 = 0.$$

Since $\dot{Y}_t$ does not exist we need to consider its integral version

$$\hat{X}_t = \int_0^t P_s dY_s - \int_0^t P_s \hat{X}_s ds,$$

where the first integral can be defined as a Stieltjes integral or, alternatively,

$$\int_0^t P_s dY_s := Y_t P_t - \int_0^t Y_s \dot{P}_s ds,$$

Introduction XVI

since P is differentiable. The optimal filter then can be derived

$$\hat{X}_t = \int_0^t P_s \left\{ dY_s - \hat{X}_s ds \right\},$$
$$\dot{P}_t = 2 - P_t^2, \qquad P_0 = 0.$$

- **What about nonlinear filtering?**
- In this case the equations require integration of some stochastic processes with respect to Y and we need to use Itô integration. In terms of this integral the nonlinear filtering formula is given by

$$\rho_t(i) = \delta_0(i) + \int_0^t \{\rho_s(i+1) + \rho_s(i-1) - 2\rho_s(i)\} ds + \int_0^t i\rho_s(i)\, dY_s,$$

and

$$\bar{X}_t = \frac{\sum_{i=-\infty}^{+\infty} i\rho_t(i)}{\sum_{i=-\infty}^{+\infty} \rho_t(i)}.$$

Introduction XVII

- The previous example is a particular case of the general filtering problem.
- Let $(X_t, Y_t)_{t\in\Pi}$, where $\Pi = \{\mathbb{N}, \mathbb{R}_+, ..., \}$, be a pair of random processes defined on a probability space $(\Omega, \mathcal{F}, P)$.
- We call X the signal process. X takes values in some space S and we cannot observe it directly. S can be $\mathbb{R}, \mathbb{R}^d$, or some metric space.
- We can observe the process Y and we call it the observation process. Y is usually assumed to be a functional of X perturbed by some noise.
- Denote by $\mathcal{Y}_t = \mathcal{F}_t^Y$ the minimal augmented filtration generated by Y.
- The filtering problem consists in computing the conditional distribution of X_t given $\mathcal{Y}_t$, that is, to compute

$$\pi_t(\varphi) := \mathbb{E}[\varphi(X_t)|\mathcal{Y}_t] = \int_S \varphi(x)\,\pi_t(dx), \qquad \varphi \in \mathcal{H}, t \in \Pi,$$

for $\mathcal{H}$ a, suffciently large, space of test functions.

Introduction XVIII

- π_t is a (probability) measure valued function of the path of Y up to time t.
- To find this functional exactly is impossible in general.
- We aim at finding approximations, which, to be efficient, need to be recursive.
- In order to find those recursive approximation we need first to find the equation satisfied by π_t. This is known as the Kushner-Stratonovich equation, which is a stochastic partial differential equation (SPDE).
- To solve the stochastic filtering problem was one of the first main motivations to study SPDEs.

History

- **Kolmogorov** and **Wiener**: Stationary processes.
- **Kalman** and **Bucy**: Linear stochastic differential equations.
- **Kushner**, **Stratonovich**, **Fujisaki**, **Kallianpur** and **Kunita**: Dynamic equations for π_t.
- **Kallianpur** and **Striebel**: Bayes formula for π_t, i.e.,

$$\pi_t(\varphi) = \frac{\rho_t(\varphi)}{\rho_t(\mathbf{1})},$$

  where ρ_t is the "unnormalised" conditional distribution and satisfies a linear SPDE.
- **Zakai**: Reference measure approach to nonlinear filtering and dynamic equation for ρ_t.
- Numerical methods:
  - **Pardoux**: Extended Kalman filter.
  - **Lototsky**, **Mikulevivius** and **Rozovskii**: Wiener chaos (spectral) approach.
  - **Gyöngy** and **Krylov**: PDE method.
  - **Doucet** and **Crisan**: Particle filters.

Applications I

- The range of applications of nonlinear filtering is huge.
- "Statistical description of a partially and noisy observed dynamical system" is a basic and ubiquitous problem.
- Engineering: satellite tracking, rocket guidance,...
- Finance: credit risk, stochastic volatility,...
- Biology: Gene prediction.
- Geoscience: In this field, "stochastic filtering" is known as "data assimilation". Weather prediction.

Applications II

Example 0.1 (Bearings only tracking)

- We take $S = \mathbb{R}^2$.
- X_t is the $2d$-position of a ship at time t, assumed to follow the following dynamics

$$X_t = X_{t-1} + U_t, \qquad U_t \sim \mathcal{N}_2\left(0, \sigma_X^2 I_2\right).$$

- Y_t is a (noisy) angular measurement obtained by radar

$$Y_t = \arctan\left(\frac{X_t^2}{X_t^1}\right) + V_t, \qquad V_t \sim N\left(0, \sigma_Y^2\right).$$

- The problem is to estimate the position of the ship X_t given the past measurements $Y_s, 0 \leq s \leq t$.

Applications III

Example 0.2 (Stochastic volatility)

- Consider the SDE
$$dX_t = \sqrt{\nu_t} dW_t^X,$$
where W^X is a Wiener process.
- ν is some continuous time nonnegative Markov process (can be the solution of another SDE, possibly correated with X).
- $S_t = e^{X_t}$ is the price process of a stock.
- In this model ν is actually observable because we have that $d[X, X]_t = \nu_t dt$, for the quadratic variation of X.
- However, high-frequency data tends to be subject to the so-called microstructure noise, that is, we observe
$$dY_t = X_t dt + \sigma_Y dW_t^Y,$$
rather than X itself, where W^Y is a standard Wiener process and $\sigma_Y > 0$.
- The problem is to estimate ν based on the past observations $Y_s, 0 \leq s \leq t$.

References

- The main reference will be the notes: "Introduction to Nonlinear Filtering" by Pavel **Chigansky**. You can find it online in the webpage of the author.
- I will follow mainly these notes, but sometimes I will depart from them. For example, for the nonlinear filtering in continous time I will use Bain & Crisan's book below.

Books

1. *Statistics of Random Processes I and II*. Springer. **Lipster** & **Shiryaev**
2. *Fundamentals of Stochastic Filtering*. Springer. **Bain** & **Crisan**
3. *An Introduction to Stochastic Filtering Theory*. OUP. **Xiong**
4. *An Introduction to Sequential Monte Carlo*. Springer. **Chopin** & **Papaspiliopoulos**
5. *Bayesian Filtering and Smoothing*. Second Edition. CUP. **Särkkä** & **Svensson**
6. *Data Assimilation Fundamentals*. Springer. **Evensen**, **Vossepoel** & **van Leeuwen**

Outline of the course I

1. Linear Filtering in Discrete Time
   - Introduction
   - Orthogonal projection and estimation
   - Recursive orthogonal projection
   - The Kalman-Bucy filter in discrete time
   - Properties of the Kalman-Bucy filter

2. Nonlinear Filtering in Discrete Time
   - Introduction
   - The conditional expectation: a closer look
   - The Bayes formula: an abstract formulation
   - The nonlinear filter via the Bayes formula
   - The nonlinear filter via the reference measure approach
   - The linear Gaussian case: Kalman-Bucy filter revisited

3. Review on Stochastic Analysis

Outline of the course II

4. Linear Filtering in Continuous Time
   - The Kalman-Bucy filter: scalar case
   - The Kalman-Bucy filter: general case
   - Linear filtering beyond linear diffusions

5. Nonlinear Filtering in Continuous Time
   - General framework
   - The change of reference probability measure method
   - Unnormalised conditional distribution
   - The Zakai equation
   - The Kushner-Stratonovich equation
   - The innovation process approach

6. Particle Methods