

Introduction to Stochastic Filtering MAT 4990 / 9790

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

$$(1) \quad X_j = X_{j-1} + \varepsilon_j, \quad j \in \mathbb{Z}_+ \quad . \quad X_0 = 0$$

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

$$(2) \quad Y_i = X_i + \zeta_i, \quad i = 1, \dots, j$$

where ζ_j is a sequence of i.i.d. r.v. (white noise) with Gaussian distribution,

$$P(\zeta_i \in [a, b]) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-u^2/2} du \quad \forall i \geq 1$$

An estimate, is a rule assigning a real number to any outcome of $Y_{[1:j]} = Y_{1:j} = \{Y_1, \dots, Y_j\}$

That is, a map $\xi_j(y) : \mathbb{R}^j \rightarrow \mathbb{R}$.

• How guesses are compared?

ξ_j is better than ψ_j if

$$(3) \quad E[(X_j - \xi_j(Y_{1:j}))^2] \leq E[(X_j - \psi_j(Y_{1:j}))^2]$$

"minimal square error on average".

Note that even if (3) holds,

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

may happen for some experiment / realization.

Once the criterion (3) is accepted, we want to find the best (optimal) estimate.

Simple guess

$$\tilde{X}_j := \tilde{\xi}_j(Y_{1:j}) = X_j \quad (\text{last observation})$$

the corresponding mean square error is

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

This estimate does not take into account past observations.

Let's try

$$\tilde{\tilde{X}}_j = \frac{X_j + X_{j-1}}{2} \quad (\text{2 order moving average})$$

Its mean square error is

$$\begin{aligned} \hat{\tilde{P}}_j &= E[(X_j - \tilde{\tilde{X}}_j)^2] = E\left[\left(X_j - \frac{X_j + X_{j-1}}{2}\right)^2\right] \\ &= E\left[\left(X_j - \frac{X_j + X_{j-1} + \xi_j + \xi_{j-1}}{2}\right)^2\right] \\ &= E\left[\left(\frac{X_j - X_{j-1}}{2} - \frac{\xi_j + \xi_{j-1}}{2}\right)^2\right] = E\left[\left(\xi_j/2 + \frac{(\xi_j + \xi_{j-1})}{2}\right)^2\right] \\ &= \frac{E[\xi_j^2]}{4} + \frac{E[\xi_j^2] + E[\xi_{j-1}^2]}{4} = \frac{1}{4} + \frac{1}{2} = 0.75 \end{aligned}$$

Improvement of 25% with respect to $\tilde{X}_j$

Let's increase the "memory" of the estimate

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

$$= \frac{4}{9} + \frac{1}{9} + \frac{3}{9} = \frac{8}{9} \approx 0.89 \quad \text{Error increased!!!}$$

the problem here is that we give the same weights to the "new" and "old" measurements.

Have we optimally weight the data?

One can prove that the optimal estimate can be generated "very efficiently" by the difference equation

$$(4) \quad \hat{X}_j = \hat{X}_{j-1} + P_j (X_j - \hat{X}_{j-1}), \quad \hat{X}_0 = 0$$

where P_j is the sequence of numbers

$$(5) \quad P_j = \frac{P_{j-1} + 1}{P_{j-1} + 2}, \quad P_0 = 0$$

Let's calculate the mean square error.

The sequence $\Delta_j := X_j - \hat{X}_j$ satisfies

$$\begin{aligned} \Delta_j &= \Delta_{j-1} + \varepsilon_j - P_j(\Delta_{j-1} + \varepsilon_j + \xi_j) \\ &= (1 - P_j) \Delta_{j-1} + (1 - P_j) \varepsilon_j - P_j \xi_j \end{aligned}$$

Therefore, using the independence of ε_j , ξ_j and Δ_{j-1} , we get

$$\hat{P}_j = E[(\Delta_j)^2] = (1 - P_j)^2 \hat{P}_{j-1} + (1 - P_j)^2 + P_j^2, \quad \hat{P}_0 = 0$$

Note that the sequence P_j satisfies the following identity (expand the right hand side) using (5)

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

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

$$\hat{P}_j - P_j = (1 - P_j)^2 (\hat{P}_{j-1} - P_{j-1}), \quad j \geq 1$$

and since $\hat{P}_0 - P_0 = 0$, it follows that $\hat{P}_j = P_j$ for all $j \geq 0$.

We just have proved that the mean square error corresponding to $\hat{x}_j$ is given by P_j .

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_∞ ,

which is the unique positive root of

$$P = \frac{P+1}{P+2} \Rightarrow P_\infty = \sqrt{5}/2 - 1/2 \approx 0.618.$$

It's clear to a 40% improvement with respect to $\hat{X}_j$

No further improvement is possible among linear estimates

What about nonlinear estimates?

The "simplest" nonlinear estimate could be

$$\check{X}_1 = \text{sign}(Y_1) \quad \begin{array}{ll} \text{Guess } 1 & \text{if } Y_1 \geq 0 \\ \text{Guess } -1 & \text{if } Y_1 < 0 \end{array}$$

The error is computed as follows,

$$\begin{aligned} \check{P}_1 &= E[(X_1 - \check{X}_1)^2] \\ &= \frac{1}{2} E[(1 - \text{sign}(1 + \epsilon_1))^2] + \frac{1}{2} E[(-1 - \text{sign}(-1 + \epsilon_1))^2] \\ &= \frac{1}{2} 2^2 P(\epsilon_1 \leq -1) + \frac{1}{2} 2^2 P(\epsilon_1 \geq 1) = 4 P(\epsilon_1 \leq -1) \\ &= 4 \Phi(-1) \approx 0.6346 \\ &\quad \nwarrow \text{c.d.f. of a } N(0,1) \end{aligned}$$

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

Let's try the estimate $\bar{X}_1 = \tanh(Y_1) = \frac{e^{2Y_1} - 1}{e^{2Y_1} + 1}$

The corresponding mean square error is given by

$$\bar{P}_L = E[(X_L - \bar{X}_L)^2]$$

$$= \frac{1}{2} \int_{-\infty}^{\infty} (L - \tanh(u+1))^2 + (L + \tanh(u-1))^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$$

$$\approx 0.4496$$

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

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

It turns out that it is possible.

Let $\ell_j(i)$, $i \in \mathcal{H}$, $j \geq 0$ be generated by the nonlinear recursion

$$(6) \quad \ell_j(i) = \exp(Y_j i - i^2/2) (\ell_{j-1}(i-1) + \ell_{j-2}(i+1)) \quad j \geq 1$$

subject to $\ell_0(0) = 1$ and $\ell_0(i) = 0$, $i \neq 0$.

then the best estimate of X_j from $Y_{1:j}$ is given by

$$(7) \quad \bar{X}_j = \frac{\sum_{i=-\infty}^{+\infty} i \ell_j(i)}{\sum_{i=-\infty}^{+\infty} \ell_j(i)}$$

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

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

Intuitively, X is a continuous time Markov chain satisfying

$$(8) \quad P(X_{t+\varepsilon} = i \mid X_s, 0 \leq s < t) = \begin{cases} 1-2\varepsilon & i = X_t \\ \varepsilon & i = X_t \pm 1 \\ 0 & \text{otherwise} \end{cases}$$

for sufficiently small $\varepsilon > 0$.

(8) is sufficient to define uniquely a stochastic process.

It can be modelled using a pair of independent Poisson processes.

Let $(\tau_n)_{n \geq 1}$ be a sequence of i.i.d. r.v. with

$$(9) \quad P(\tau_n \leq t) = \begin{cases} 1 - e^{-t} & t \geq 0 \\ 0 & t < 0 \end{cases} \quad (\text{exponential})$$

then a standard Poisson process is defined as

$$\pi_t = \max \left\{ n : \sum_{k=1}^n \tau_k \leq t \right\}$$

Let π_t^+ and π_t^- be a pair of independent Poisson processes.

then the process

$$X_t = \pi_t^+ - \pi_t^- \quad t \geq 0 \quad \text{satisfies (8)}.$$

To define an analogue of X_t , the concept of "white noise" is to be introduced in continuous time. This is much more delicate.

The term "white noise" comes from the fact that the spectral density of a sequence of i.i.d. r.v. $(\xi_i)_{i \geq 1}$

is flat, i.e., autocorrelation function

$$S_{\xi}(\lambda) = \sum_{j=-\infty}^{+\infty} E[\xi_0 \xi_j] e^{-i\lambda j} = \sum_{j=-\infty}^{+\infty} \delta_0(j) e^{-i\lambda j} = 1 \quad \forall \lambda \in (-\pi, \pi)$$

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

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

In continuous time this definition leads to

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

↑ in continuous time we need to integrate $S_{\xi}(\lambda)$ over $\mathbb{R}$.

We can solve this problem by defining the observation process as

$$(10) \quad Y_t = \int_0^t X_s ds + W_t$$

where $W = (W_t)_{t \geq 0}$ is a Wiener process.

This process is characterized by:

- $W_0 = 0$
- W has continuous paths,
- W has independent increments,
- For all $0 \leq s \leq t$ $W_t - W_s \sim N(0, t-s)$

Why the model (10) is compatible with the notion of "white noise"?

Define the family of processes

$$V_t^\Delta = \frac{W_t - W_{t-\Delta}}{\Delta}, \quad \Delta > 0$$

Then, $E[V_t^\Delta] = 0$ and

$$\begin{aligned} E[V_t^\Delta V_s^\Delta] &= \frac{1}{\Delta^2} E[(W_t - W_{t-\Delta})(W_s - W_{s-\Delta})] \\ &= \frac{1}{\Delta^2} \begin{cases} \Delta - |t-s|, & |t-s| \leq \Delta \\ 0, & |t-s| \geq \Delta \end{cases} \end{aligned}$$

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

$$R_V^\Delta(\tau) = \frac{1}{\Delta^2} \begin{cases} \Delta - |\tau|, & |\tau| \leq \Delta \\ 0, & |\tau| \geq \Delta \end{cases}$$

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

$$\int_{-\infty}^{+\infty} \xi(\tau) R_V^\Delta(\tau) d\tau \xrightarrow{\Delta \rightarrow 0} \xi(0)$$

and if the limit process $V := \lim_{\Delta \rightarrow 0} V$ existed, it would have flat spectral density.

Then, the observation process in (10) would contain the same information as

$$\dot{Y}_t = X_t + V_t$$

with V_t being the derived noise.

This is only an intuition because V_t does not exist as a limit in any reasonable sense (its variance explodes at any time t).

Next, what would be the analogue of the filtering equations (4)-(5) and (6)-(A)?

Instead of difference equations we get differential equations, e.g.,

$$\dot{\hat{X}}_t = P_t (\dot{Y}_t - \hat{X}_t), \quad \hat{X}_0 = 0$$

However, the r.h.s. of this equation involves the time derivative of Y and, therefore, of W , which is impossible. Then, we need to consider its integral version

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

↑
Stieltjes integral

or, alternatively,

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

↙ P is differentiable.

The optimal linear filter then can be derived

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

$$\dot{P}_t = -P_t^2, \quad P_0 = 0$$

What about the nonlinear filtering?

In this case the equations would require integration of some stochastic processes with respect to Y .

Here, ordinary integration fails and we need to use Itô integration.

In terms of this integral the nonlinear filtering formula is given by

$$p_t(i) = \delta_0(i) + \int_0^t \{p_s(i+1) + p_s(i-1) - 2p_s(i)\} ds + \int_0^t i p_s(i) dY_s$$

and

$$\bar{X}_t = \frac{\sum_{m=-\infty}^{\infty} m p_t(m)}{\sum_{m=-\infty}^{\infty} p_t(m)}$$

this example is a particular case of the general filtering problem

Let $(X_t, Y_t)_{t \in \pi}$ be a pair of random processes defined on a probability space $(\Omega, \mathcal{F}, P)$. ($\pi = \{N, \mathbb{R}_+, \dots\}$)

We call X the signal process, X takes values in some space S and we cannot observe it directly. ($S = \mathbb{R}, \mathbb{R}^d$, or some metric space)

The process Y is called the observation process and it 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 .

Then the filtering problem consists in computing the conditional distribution of X_t given $\mathcal{Y}_t$, that is, to compute

$$(11) \quad \pi_t(\phi) := E[\phi(X_t) | \mathcal{Y}_t] = \int_S \phi(x) \pi_t(dx), \quad \begin{matrix} \phi \in \mathcal{H} \\ \mathcal{H} \subset \mathcal{C} \end{matrix}$$

for $\mathcal{H}$ a (sufficiently large) space of test functions.

π_t is a measure-valued functional of Y up to time t .

To find this functional exactly is impossible in general. In practice we find approximations. For these to be efficient, they need to be recursive.

In order to find recursive approximations, we need 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 nonlinear filtering problem was one of the first main motivations to study SPDEs.

History

- Kalman and Wiener: Stationary processes.
- Kalman - Bucy: linear stochastic differential equations.
- Kushner - Strogonovich (Fujisaki, Kallianpur and Kunita): dynamic equations for π_t^*
- Kallianpur - Striebel: Bayes formula for π_t^* , i.e.,
$$\pi_t^*(\xi) = \frac{\ell_t^*(\xi)}{\ell_t^*(1)}$$

ℓ_t^* is the "unnormalised" conditional distribution
- Zakai: reference measure approach to nonlinear filtering as a dynamic equation for ℓ_t^*
- Numerical methods:
 - Pardoux: Extended Kalman filter.
 - Lototsky, Mikulevicius and Rozenskiis: Spectral approach.
 - Gyöngy and Krylov: PDE method.
 - Doucet, Crisan: Particle filters.

The range of applications of nonlinear filtering is huge.

"Statistical description of a partially and noisily observed dynamical system" is a basic and ubiquitous problem.

- Engineering: satellite tracking, rocket guidance, etc.
- Finance: credit risk, stochastic volatility, etc.
- Biology: Gene prediction.
- Geoscience: In this field, "stochastic filtering" is known as "data assimilation".
It is used, for instance, in weather prediction.

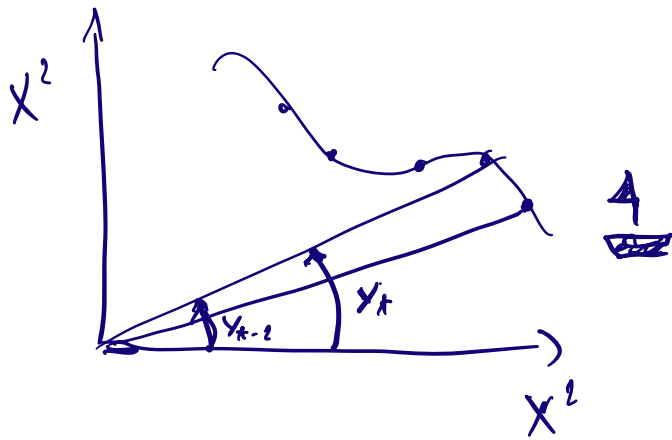
Example 1 (Bearings-only tracking)

$S = \mathbb{R}^2$. X_t is the 2d-position of a ship.

$$X_t = X_{t-1} + U_t \quad U_t \sim N_2(0, \sigma_x^2 I_2)$$

and Y_t is a (noisy) angular measurement obtained by radar or similar device.

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



Example 2 (Stochastic volatility)

$$dX_t = \sqrt{v(t)} dW_x(t)$$

W_x is a Wiener process

v is some continuous time Markov process.

$X_t = \log(S_t)$ S_t price process.

In this model v is actually observable because we have that $d[X, X]_t = v(t)dt$ for the quadratic variation of X .
However, high-frequency data tends to be subject to so-called microstructure noise, that is, we observe

$$dY_t = X(t)dt + \sigma_y dW_y(t)$$

rather than X itself, where W_y is a Wiener process and $\sigma_y > 0$.

We want to approximate $v(t)$ based on the past observations $Y(s), 0 \leq s \leq t$.

SYLLABUS (Tentative)

1. Linear filtering in discrete time
 - orthogonal projection, Kalman-Bucy filter
 2. Nonlinear filtering in discrete time
 - Bayes formula, reference measure approach.
 3. Particle filters in discrete time
 4. Linear filtering in continuous time
 - Kalman-Bucy filter
 5. Nonlinear filtering in continuous time
 - Innovations approach, reference measure approach, Kushner-Stratonovich equations, Zakai equations.
 - Finite dimensional filters.
-

6. Some numerical methods

7. Connection with stochastic control

8. Some detailed applications

Main reference:

"Introduction to Nonlinear Filtering" Pavel Chigansky.

You can find it online (Google search)

I will be following closely these notes.
At some points I may depart, but then
I will provide notes.

Some back:

Chigansky notes are based on:

- (1) Statistics of Random Processes I and II.
Springer. Lipster & Shiryaev.

More modern references:

- (2) Fundamentals of Stochastic Filtering. Springer
- Bain & Crisan (Theory and numerics)
- (3) An introduction to stochastic filtering theory.
Oxford university press. Xiong
(nice modern account).
- (4) An introduction to Sequential Monte Carlo.
Springer. Chopin & Papaspiliopoulos
(Statistical approach, particle filters...)

All the references, except (3), can be found
for free in Springer Link.