Stochastic Machine Learning
Chapter 04 - Filtering III

Thorsten Schmidt

Abteilung für Mathematische Stochastik

www.stochastik.uni-freiburg.de
thorsten.schmidt@stochastik.uni-freiburg.de

SS 2024

Filtering and estimation

- In general the parameters are not known and have to be estimated
- We assume that for all $n \geq 1$

$$\begin{aligned} X_n &= a_\theta(X_{n-1}, \xi_n) \\ Y_n &= A_\theta(X_n, Y_{n-1}, \eta_n). \end{aligned} \tag{1}$$

  where ξ_n, η_n are i.i.d. with a given (and known) distribution.
- The goal is to estimate the parameter $\theta \in \Theta$.

Example (Kalman filter)

- The unobserved signal $X = (X_n)_{n\geq 1}$ is a Gaussian process given by

$$X_n = aX_{n-1} + b\epsilon_n,$$

  where (ϵ_n) are i.i.d. $\mathcal{N}(0, 1)$.
- The observation is given by

$$Y_n = AX_n + B\eta_n,$$

  where also (η_n) are also i.i.d., independent of (ϵ_n).
- Hence, the parameter vector θ is given by $(a, b, A, B, \mu_0, \Sigma_0)$.

Maximum Likelihood

- Let us first look at an example.
- We are interested in maximizing
$$f_\theta(x, y),$$
over $\theta \in \Theta$, where - however, x is not observable.
- We thus need to maximize
$$f_\theta(y) = \int_{\mathcal{X}} f_\theta(x, y) dx$$
which is often difficult to compute.
- The general theory is as follows: Assume we are given a family
$$P_\theta : \theta \in \Theta$$
where $P_\theta \sim R$, this means P_θ and R have the same nullsets for all $\theta \in \Theta$ (i.e. $P_\theta(A) = 0$ if and only if $R(A) = 0$).
- By the Radon-Nikodym theorem there exists a density
$$\frac{dP_\theta}{dR}$$
and hence the general maximum likelihood problem (under full information) is to find
$$\hat{\theta} \in \arg\max_{\theta \in \Theta} \log \frac{dP_\theta}{dR}.$$

Example

- If we look at the normal distribution we often take as reference the Lesbesgue-measure

$$\log \frac{dP_\theta}{d\lambda} \propto -\log \theta_2 - \frac{(x-\theta_1)^2}{2\theta_2^2}.$$

- We could also choose $R \sim \mathcal{N}(0,1)$. Then,

$$\begin{aligned}
\frac{dP_\theta}{dP_{(0,1)}} &= \frac{\frac{1}{\sqrt{2\pi\sigma^2}}}{\frac{1}{\sqrt{2\pi}}} \cdot \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{e^{-\frac{x^2}{2}}} \\
&= \frac{1}{\sigma} \cdot \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2} + \frac{x^2}{2} \Big) \\
&\propto \frac{1}{\sigma} \cdot \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2} \Big)
\end{aligned}$$

ML under incomplete information

- Consider the complete system $Z = (X, Y)$ with values in some state space G and the observation Y with values in some state space F. We assume that Z has density

$$f(z, \theta) : \quad \theta \in \Theta,$$

when the unknown parameter vector θ lies in the parameter vector space Θ. Moreover, the observation Y has density

$$g(y, \theta) : \quad \theta \in \Theta.$$

- Then

$$g(y, \theta) = \int_E f(x, y, \theta) dx.$$

- The log-likelihood is given by

$$\begin{aligned} L(\theta) &:= \log g(y, \theta) \\ &= E_{\theta_{n-1}} [\log g(Y, \theta) | Y = y] \\ &= E_{\theta_{n-1}} \left[\log f(X, y, \theta) - \log \frac{f(X, y, \theta)}{g(y, \theta)} | Y = y \right] \\ &=: Q(\theta, \theta_{n-1}) - H(\theta, \theta_{n-1}). \end{aligned}$$

- The **EM-algorithm** is an iterative procedure where the iteration $\theta_{n-1} \to \theta_n$ is given in two steps:
  - **E-Step:** Compute $Q(\theta, \theta_{n-1}) = \int_E \log f(x, Y, \theta) \pi_n(dx, \theta_{n-1})$,
  - **M-Step:** Choose θ_n as any maximizer of $\theta \mapsto Q(\theta, \theta_{n-1})$.
- Sometimes it is numerically difficult to solve the M-step. The **generalized EM**-algorithm (GEM) just chooses any θ_{n+1} such that $Q(\theta_n, \theta_{n-1}) \geq Q(\theta_{n-1}, \theta_{n-1})$. Convergence will be slower in this case but can be obtained in a similar way.

- Recall $L(\theta) = Q(\theta, \theta_{n-1}) - H(\theta, \theta_{n-1})$
- Now we obtain that

$$L(\theta_n) \geq L(\theta_{n+1}) \tag{2}$$

for each iteration of the GEM-algorithm by the simple fact

$$H(\theta, \theta_{n-1}) \leq H(\theta_{n-1}, \theta_{n-1})$$

which follows from Jensen's inequality.

Indeed,

$$\begin{aligned}
H(\theta', \theta) &= E_{\theta'}\Big[-\log \frac{f(X, y, \theta)}{g(y, \theta)} | Y = y \Big] \\
&= \int -\log \frac{f(x, y, \theta)}{g(y, \theta)} \frac{f(x, y, \theta')}{g(y, \theta')} dx \\
&=: \int -F'(x) \log F(x) dx \\
&= \int F'(x) \log \frac{F'(x)}{F(x)} dx - \int F'(x) \log F'(x) dx \\
&= \int \frac{F'(x)}{F(x)} \log \frac{F(x)}{F'(x)} F(x) dx + \int F'(x) \log F'(x) dx
\end{aligned}$$

Now note that $G(x) = x \log x$ is convex, and hence Jensen's inequality yields that

$$H(\theta', \theta) \geq C + G\Big(\int F'(x) dx \Big) = C + G(1) = C$$

with equality for $\theta = \theta'$.

Example (Exponential family)

We say that $\{f(\cdot,\theta) : \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^K$ is called a K-dimensional **exponential family**, if for all $\theta \in \Theta$

$$f(z,\theta) = \exp\Big(c(\theta)^\top T(z) + S(z) + d(\theta)\Big)\mathbb{1}_A(z).$$

T is called sufficient statistic. The maximum-likelihood estimator $\hat{\theta}$ is given by the solution of the equations $\mathbb{E}_\theta[T(Z)] = T(z)$ if c is one-to-one. This is easy to see if $c(\theta) = \theta$: consider $K = 1$ and note that the normalizing constant d is given by

$$d(\theta) = -\log \int \exp(c(\theta)^\top T(z) + S(z))dz.$$

Derivation with respect to θ gives that $d'(\theta) = -\mathbb{E}[T(Z)]$. On the other side, the maximizer of the log-likelihood needs to satisfy

$$0 = -\partial_\theta \log f(z,\theta) = T(z) + d'(\theta)$$

and we obtain $\mathbb{E}_\theta[T(Z)] = T(z)$. Because of this, when considering exponential families, the EM-algorithm takes on the following, simpler form:

- **E-Step:** Compute $t^n(y) := E_{\theta_{n-1}}[T(X,Y)|Y = y]$,
- **M-Step:** Choose θ_n as a solution of $E_\theta[T(X,Y)] = t^n(y)$.

Convergence

In this regard, assume that

$$\Theta \subset \mathbb{R}^K \tag{3}$$

$$\Theta_0 = \{\theta \in \Theta : L(\theta) \geq L(\theta_0)\} \quad \text{is compact for any } \theta_0 \text{ such that } L(\theta_0) > -\infty, \tag{4}$$

$$L \text{ is continuous and differentiable in the interior of } \Theta. \tag{5}$$

Let M be a point-to-set map generating $(\theta_n)_{n\geq 0}$, i.e. $\theta_n \in M(\theta_{n-1})$ for all $n \geq 1$. The map M from points in Θ to subsets of Θ is called a **point-to-set map** in Θ. It is called **closed** at x if

$$x_k \to x, \quad M(x_k) \ni y_k \to y \Rightarrow y \in M(y).$$

Note that continuity for a point-to-point map implies closedness.
We assume

(i) M is a closed over the complement of Γ,

(ii) $L(\theta_{n+1}) > L(\theta_n)$ for all $\theta_n \notin \Gamma$.

If we are interested in convergence of (θ_n) rather that of $(L(\theta_n))$ we can use the following result. We fix the point-to-set map M and the sequence $(\theta_n)_{n\geq 0}$ generated by M.

Theorem

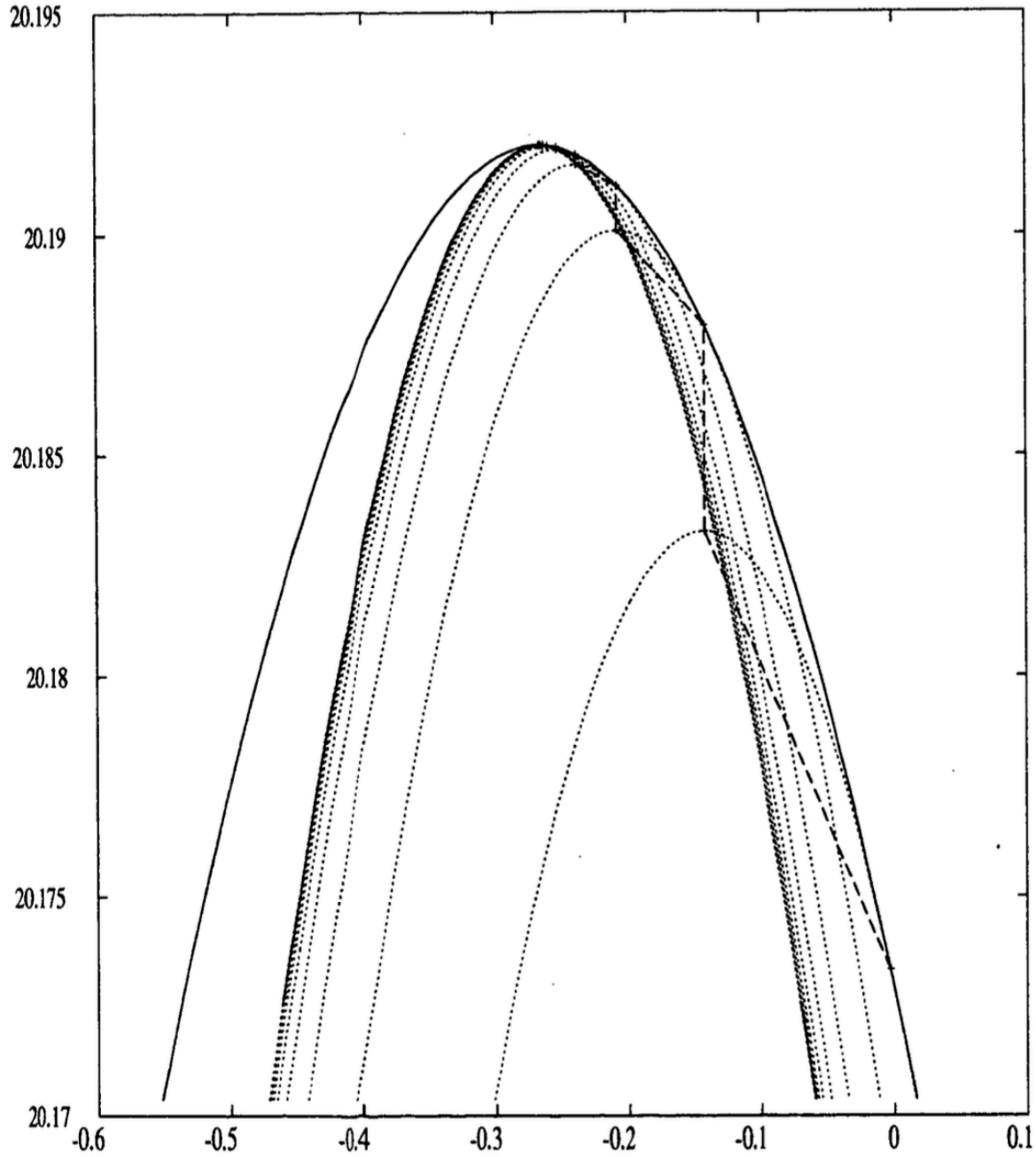
Suppose that (3)-(5) and (i) and (ii) hold. Denote by L^* the limit of $(L(\theta_n))$ and by Γ^* the set of local maxima of L^*. If

$$\| \theta_n - \theta_{n-1} \| \xrightarrow[n \to \infty]{} 0$$

then all limit points of $(\theta_n)_{n \geq 0}$ are in a connected and compact subset of Γ^*.

In particular, if Γ^* is discrete, i.e. its only connected components are singletons, then $(\theta_n)_{n \geq 0}$ converges to some $\theta^* \in \Gamma^*$. If Γ^* has only one maximum then convergence to this point holds. The proof uses general results on limit points of point-to-set-maps and is skipped (Zangwill 1969). For exponential families, the assumptions hold true.

- It is on discussion if maximum-likelihood is not better than the EM, see Campillo and Le Gland (1989).
- In particular if the likelihood is steep (some components are well observable), direct maximum likelihood seems to be better.


20.195
20.19
20.185
20.18
20.175
20.17
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1

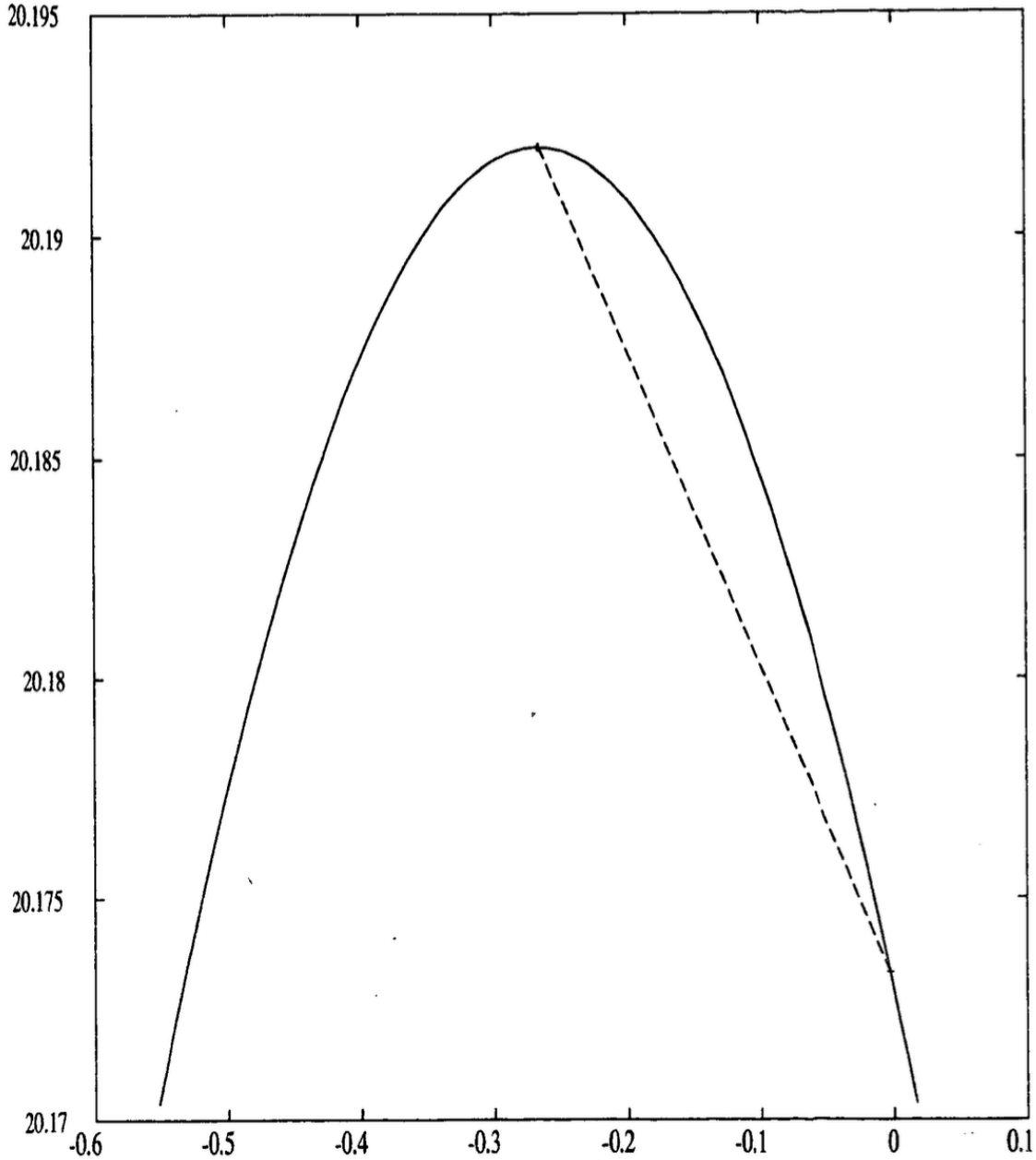


Figure 1: Case I – Direct maximization

 Campillo, Fabien and François Le Gland (1989). „MLE for partially observed diffusions: direct maximization vs. the em algorithm“. In: **Stochastic Processes and their Applications** 33.2, pp. 245–274.