

Analysis: Inference in Hidden Markov Models

Cappé, Moulines & Rydén (2005)

David Chan, Claude Sonnet 4.6

AI-Symbiosis Research · April 2026

Page 1 — Chapter 2 & 5 Deep Dive

Chapter 2: Filtering and Smoothing Recursions

The chapter opens with a clean separation of three problems:

Definition 18 — The Three Inference Tasks:

- **Smoothing:** $\phi_{v,k|n}$ — distribution of X_k given *all* $Y_0, \dots, Y_n$. Best estimate, uses future data.
- **Filtering:** $\phi_{v,k|k}$ — distribution of X_k given *past* $Y_0, \dots, Y_k$. Real-time estimate.
- **Prediction:** $\phi_{v,k+p|k}$ — distribution of X_{k+p} given data up to k . Forward-looking.

The book focuses on **fixed-interval smoothing** — n is fixed, and you want all $\phi_{v,k|n}$ simultaneously. This is the offline training case.

The Forward-Backward Decomposition:

Any smoothed distribution factors as:

$$\phi_{v,k|n}(f) = L^{-1} \alpha_{v,n} \beta_{v,k|n}(f)$$

The forward measure $\alpha_{v,k}$ accumulates evidence left-to-right. The backward function $\beta_{v,k|n}$ accumulates evidence right-to-left. Multiply them and normalize — you get the full posterior.

Proposition 25 — Normalized Recursion (the one you implement):

$$c_{v,k} = \int \int \phi_{v,k-1}(dx) Q(x, dx') g_k(x')$$

$$\phi_{v,k}(f) = c_{v,k}^{-1} \int f(x') \phi_{v,k-1}(dx) Q(x, dx') g_k(x')$$

Note: The scalars $c_{v,k}$ normalize at each timestep, preventing floating point underflow on long sequences. The total log-likelihood is simply $\sum_k \log c_{v,k}$ — it falls out of the forward pass for free.

Forward Smoothing Kernels (Definition 30):

$$F_{k|n}(x, A) = [\beta_{k|n}(x)]^{-1} \int A Q(x, dx') g_{k+1}(x') \beta_{k+1|n}(x')$$

This defines how probability mass flows *forward through the smoother*. The key structural insight: conditionally on $Y_{0:n}$, the time-reversed sequence $X_{n-k}^- = X_{n-k}$ is itself a non-homogeneous Markov chain with these kernels as transitions. **The backward pass is a time-reversed Markov chain.**

Note: The book explicitly notes that filtering theory (continuous-time, Shiryaev 1966, Wonham 1965) and discrete-time HMMs evolved as two mostly independent fields. The continuous-time analogues involve stochastic differential equations — deliberately deferred.

Chapter 5: Maximum Likelihood Inference

The Setup: Parameters $\theta = (v, Q, g)$ are unknown. You observe $Y_{0:n}$, want θ^{MLE} . Direct maximization of $\ell_n(\theta) = \log L_{v,n}(\theta)$ is hard — the hidden states make it a missing data problem.

Proposition 98 — The Fundamental EM Inequality:

$$\ell(\theta) - \ell(\theta') \geq Q(\theta; \theta') - Q(\theta'; \theta')$$

where

$$Q(\theta; \theta') = E_{\theta'}[\log f(X_{0:n}, Y_{0:n}; \theta) | Y_{0:n}]$$

Note: Maximizing Q over θ is guaranteed to increase ℓ . Every single EM iteration improves the model — you cannot go backward. This is Baum's 1970 proof.

Fisher's Identity (eq. 5.28):

$$\nabla \theta \ell_n(\theta) = E[\nabla \theta \log v(X_0) | Y_{0:n}] + \sum_{k=0}^{n-1} E[\nabla \theta \log g_k(X_k) | Y_{0:n}] \\ + \sum_{k=0}^{n-1} E[\nabla \theta \log q(X_k, X_{k+1}) | Y_{0:n}]$$

Note: The gradient of the log-likelihood decomposes into three smoothed expectations — initial state, emissions, transitions. Each is computable via forward-backward without automatic differentiation. Use this if you want gradient ascent instead of EM.

Louis' Identity — Observed Information:

$$-\nabla^2 \theta \ell = -\nabla^2 \theta Q(\theta; \theta) + \text{Var}_{\theta}[\nabla \theta \log f(X_{0:n}, Y_{0:n}; \theta) | Y_{0:n}]$$

Note: Observed information = curvature of Q minus conditional variance of complete score. Gives you uncertainty estimates on parameters — useful for knowing how confident your regime estimates are.

Normal HMM — The Explicit EM Updates (Sec. 5.3):

$$\mu^*_{\langle \text{em} \rangle i} = \sum_k \phi_{\langle \text{em} \rangle k | n(i)} Y_k \sum_k \phi_{\langle \text{em} \rangle k | n(i)}$$

$$v_{\langle \text{em} \rangle \langle \text{em} \rangle i} = \sum_k \phi_{\langle \text{em} \rangle k | n(i)} Y_k^2 \sum_k \phi_{\langle \text{em} \rangle k | n(i)} - (\mu^*_{\langle \text{em} \rangle i})^2$$

$$q^*_{\langle \text{em} \rangle ij} = \sum_k \phi_{\langle \text{em} \rangle k-1:k | n(i,j)} \sum_k \phi_{\langle \text{em} \rangle k | n(i)}$$

Note: Every update is a weighted average, weighted by $\phi_{k|n(i)}$ — your current posterior belief about which regime was active at time k . If you were 90% sure you were in the "bull" regime on day k , that day's return gets 90% weight in the bull mean estimate.

Gaussian Linear State-Space Extension (Sec. 5.4):

For continuous hidden states (e.g. hidden volatility), Kalman smoother replaces forward-backward:

$$A^* = [\sum_{k=0}^{n-1} C_{k,k+1} | n + X^{\langle em \rangle k} | n X^{\langle T \rangle k+1} | n] T [\sum_{k=0}^{n-1} \Sigma_k | n + X^{\langle em \rangle k} | n X^{\langle T \rangle k+1} | n] - 1$$

$$Y^{\langle em \rangle} R = 1/n \sum_{k=0}^{n-1} \{ [\Sigma_{k+1} | n + X^{\langle em \rangle k+1} | n X^{\langle T \rangle k+1} | n] - A^{\langle em \rangle} [C_{k,k+1} | n + X^{\langle em \rangle k} | n X^{\langle T \rangle k+1} | n] \}$$

Note: Same EM structure, but now updating a state dynamics matrix A and process noise covariance YR . The book notes Gaussian linear state-space models are the only important HMM subclass with tractable non-iterative estimators.

MLE Asymptotics (Ch. 6) — Three Guarantees:

As $n \rightarrow \infty$:

1. $n^{-1} \ln(\theta) \rightarrow \ell(\theta)$ a.s. uniformly — log-likelihood converges to a continuous limit with unique maximum at θ^*
2. $n^{-1/2} \nabla \theta \ln(\theta^{\langle em \rangle}) \rightarrow \mathcal{N}(0, J(\theta^{\langle em \rangle}))$ weakly — score is asymptotically normal
3. $-n^{-1} \nabla^2 \theta \ln(\theta^{\langle em \rangle}) \rightarrow J(\theta^{\langle em \rangle})$ a.s. — observed information converges to Fisher information

Note: With enough data, Baum-Welch finds the true parameters. Your regime estimates are statistically consistent.

Page 2 — KEGA Gap Analysis

What is KEGA?

KEGA (Knowledge Extension via Gap Analysis) is a methodology for deriving implied knowledge from coherent systems. Applied here: given the book's internal structure, what theorems or extensions does it imply *must exist* — either explicitly deferred, structurally absent, or left as open problems?

Gap 1 — The Continuous-Time Bridge

What the book says: "Filtering theory and hidden Markov models evolved as two mostly independent fields." (Ch. 2, p.14)

The gap: The book proves forward-backward for discrete-time HMMs. Continuous-time filtering (Kushner-Stratonovich, Zakai equation) uses stochastic differential equations. These two frameworks solve the same problem using different mathematical tools and are not formally connected in this text.

Implied theorem: There exists a limiting procedure $\Delta t \rightarrow 0$ taking the discrete normalized recursion (Prop. 25) to the Zakai SDE for the unnormalized filter:

$$\dot{\sigma}(f) = \sigma(Lf)dt + \sigma(fh)dY_t$$

The normalization constants $c_{v,k}$ should converge to the innovation process. The forward smoothing kernels should converge to the Rauch-Tung-Striebel backward SDE.

Why it matters for markets: High-frequency trading operates in continuous time. A unified discrete-to-continuous HMM filter would allow regime detection at tick level without discretization error.

Gap 2 — The Natural Gradient (Baum-Sell Manifold)

What the book says: EM updates maximize $Q(\theta; \theta')$ over θ . The parameter space is treated as flat Euclidean throughout.

The gap: The transition matrix Q lives on a **simplex manifold** — each row sums to 1. The emission parameters (μ, σ) live on a curved statistical manifold with Fisher information metric $G(\theta)$. Standard EM ignores this curvature. Baum and Sell proved (in a paper referenced in the 1970 paper but never widely cited) that growth transformations generalize to functions on manifolds — but this book never uses it.

Implied algorithm: Replace the flat M-step with a natural gradient step:

$$\theta_{t+1} = \theta_t + \eta \cdot G(\theta_t)^{-1} \nabla_{\theta} Q(\theta_t; \theta_t)$$

where $G(\theta)$ is the Fisher information matrix. This respects the geometry of the parameter space and is known to converge in fewer iterations (Amari, 1998).

The missing bridge: Baum 1970 → Baum-Sell manifold paper → Amari information geometry (1985). None of these cite each other. The unified treatment does not exist.

Gap 3 — Online / Streaming EM

What the book says: Baum-Welch requires the full sequence $Y_{0:n}$ stored in memory — it's a batch algorithm. Online variants are mentioned but deferred.

The gap: For real-time trading you cannot rerun Baum-Welch on the full history every bar. You need θ_t to update as new data arrives at $O(s^2)$ cost per step.

Implied algorithm: A stochastic approximation version of the M-step:

$$\theta_{t+1} = \theta_t + \gamma_t \nabla_{\theta} Q(\theta_t; \theta_t) | \text{single observation}$$

where γ_t are Robbins-Monro step sizes satisfying $\sum \gamma_t = \infty$, $\sum \gamma_t^2 < \infty$.

Open problem: Proving convergence of this recursion under mild market non-stationarity — where the true θ^* drifts slowly over time. Standard convergence proofs assume stationarity. This is the theorem that makes Lab 4 deployable in live trading.

Gap 4 — Non-Stationary Transition Matrix

What the book says: The transition matrix $Q=(q_{ij})$ is fixed and time-homogeneous throughout.

The gap: Markets are non-stationary. The probability of transitioning from bull to bear in 2008 is not the same as in 2021. The book has no mechanism for time-varying Q_k .

Implied extension: Replace q_{ij} with a covariate-driven function:

$$q_{ij}(k) = \text{softmax}(W \cdot z_k)_{ij}$$

where z_k is a vector of observable macro covariates (VIX, yield curve slope, credit spreads). This is an **input-output HMM**. The EM updates become:

$$q^*_{ij}(k) = f_{ij}(k) / \sum_{j'} f_{ij'}(k) \quad f_{ij}(k) = \prod_{t=1}^n p(y_t = i | y_{t-1} = j, z_t)$$

Open problem: Asymptotic theory for non-stationary Q_k . The current Ch. 6 consistency results assume stationarity. They do not apply when Q_k varies. A new convergence theorem is needed.

Gap 5 — Model Selection: How Many Regimes?

What the book says: Throughout the book, the number of hidden states s is assumed known.

The gap: In practice you don't know if markets have 2, 3, or 4 regimes. Standard AIC/BIC criteria don't apply because the HMM likelihood is irregular at boundary cases — when two states merge, the Fisher information matrix is singular. The model is not identifiable at the boundary.

Implied theorem: A penalized likelihood criterion specific to HMMs:

$$IC(s) = -2\ln(\theta^s) + \text{pen}(s, n)$$

where $\text{pen}(s, n)$ must grow faster than $\log n$ to account for the irregular boundary. The correct rate is believed to be $O(s^2 \log n)$ but this has not been proved with full generality for continuous emission HMMs.

KEGA Summary Table

Gap	Type	Diffi- culty	Priority for Lab 4
1. Continuous-time bridge	Theoretical	Hard	Low
2. Natural gradient on manifold	Algorithmic	Medium	High
3. Online / streaming EM	Algorithmic	Medium	Critical
4. Non-stationary transition matrix	Model extension	Medium	High
5. Model selection (s unknown)	Statistical	Hard	Medium

The Next Theorem Worth Proving

Gap 3 — a convergent online EM for HMMs under mild non-stationarity.

This is the one that makes Lab 4 deployable in live trading. The other four gaps are intellectually interesting but not blocking. Gap 3 is blocking: without it, the regime detector must retrain from scratch on the full history at each step, which is $O(n)$ per bar and infeasible at scale.

The proof strategy: extend the Cappé (2011) online EM results to a slowly time-varying parameter setting using a tracking argument — show that if $|\theta_{\text{em}}^t - \theta^*| \leq \epsilon$ for small ϵ , the online EM estimate $\hat{\theta}^t$ stays within $O(\epsilon/\gamma t)$ of θ^* . This requires a uniform

forgetting rate (Ch. 3 results) plus a persistence-of-excitation condition on the observations.

That theorem, if proved, closes the gap between the theory in this book and a production-grade regime detector.

References

1. Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer.
2. Baum, L.E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41(1), 164–171.
3. Baum, L.E. and Sell, G.R. Growth transformation for functions on manifolds. *Pacific Journal of Mathematics*.
4. Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10(2), 251–276.
5. Cappé, O. (2011). Online EM algorithm for hidden Markov models. *Journal of Computational and Graphical Statistics*, 20(3), 728–749.
6. Kushner, H.J. (1964). On the differential equations satisfied by conditional probability densities of Markov processes. *SIAM Journal on Control*, 2, 106–119.
7. Chan, D. (2026). KEGA: Knowledge Extension via Gap Analysis. *AI-Symbiosis Research*.
8. Chan, D. (2026). Observing HMM Everywhere. *AI-Symbiosis Research*.