

# On the Entropy, Filtering and Smoothing of Hidden Markov Processes

Dongning Guo

Department of EECS, Northwestern University

October 3, 2007

# Outline

## Part I: Entropy of Hidden Markov Processes

- 1 HMP Observed via Arbitrary Channels
- 2 Result: A Fixed-point Equation for the CDF of the Log-Likelihood
- 3 Algorithm for Computation of Entropy Rate
- 4 Precision Analysis: BSC Case

## Part II: Joint Estimation/Detection on Hidden Markov Models

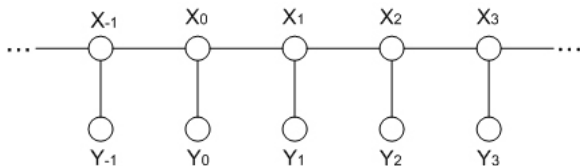
- 5 Channel Model
- 6 Inference on Graph
- 7 Numerical Results
- 8 Conclusion

# Part I

## Entropy of HMPs

With Jun Luo.

# A Hidden Markov Process (HMP)



- $\{X_i\}$  a binary Markov chain.
- Random transformation  $P_{Y|X} : X_i \mapsto Y_i$ .
- What is  $P_{X_i|Y_1^n}$ ?

$$\lim_{n \rightarrow \infty} \frac{1}{n} H(Y_1, \dots, Y_n) = ?$$

## Related Work

- Blackwell, “The entropy of functions of finite-state Markov chains”, 1957.
- Wonham, “Some applications of stochastic differential equations to optimal non-linear filtering”, 1965.
- Ordentlich-Weissman, “New bounds on the entropy rate of hidden Markov processes”, 2004. “Approximations ...”, 2005.
- Nair-Ordentlich-Weissman, “Asymptotic filtering and entropy rate of a hidden Markov process in the rare transitions regime”, 2005.
- Han-Marcus, “Analyticity of entropy rate ...,” 2006.
- Zuk-Kanter-Domany, “The entropy of a binary hidden Markov Process,” 2005.
- Jacquet-Seroussi-Szpankowski, 2004.
- Pfister, 2003, Chigansky, 2006, Holliday-Goldsmith-Glynn, 2006 ...

# Entropy Rate and Filtering

- Output entropy:

$$\begin{aligned} H(Y_1^n) &= H(Y_1^n | X_1^n) + I(X_1^n; Y_1^n) \\ &= nH(P_{Y|X}) + H(X_1^n) - H(X_1^n | Y_1^n) \end{aligned}$$

- By sandwiching (cf. Birch '62),

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} H(Y_1^n) &= nH(P_{Y|X}) + H_2(\epsilon) - \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1^n | Y_1^n) \\ &= H(P_{Y|X}) + H_2(\epsilon) - H(X_1 | X_0, Y_1^\infty). \end{aligned}$$

## Filtering Posterior

- The statistics of  $P_{X_1|X_0, Y_1^\infty}$  is enough because

$$H(X_1|X_0, Y_1^\infty) = E \left\{ H_2 \left( P_{X_1|X_0, Y_1^\infty}(\cdot|X_0, Y_1^\infty) \right) \right\}$$

- The key is  $P_{X_i|Y_{i+1}^\infty}$  since

$$P_{X_1|X_0, Y_1^\infty} \propto P_{X_1|X_0} P_{X_1|Y_1} P_{X_1|Y_2^\infty}$$

- Assume binary Markov chain. Define log-likelihood ratio

$$L_i = \frac{P_{X_i|Y_{i+1}^\infty}(+1|Y_{i+1}^\infty)}{P_{X_i|Y_{i+1}^\infty}(-1|Y_{i+1}^\infty)}$$

- $L_i$  is well-defined on the filtration  $Y_{i+1}^\infty$ . It is also stationary.
- Let  $F(\cdot)$  denote the cdf of  $L_i$  conditioned on  $X_i = +1$ .

# Theorem: A Fixed-point (Functional) Equation for CDF

## Theorem (Luo & Guo)

Consider symmetric Markov chain and symmetric channel:

$$P_{Y|X}(y|x) = P_{Y|X}(-y|-x)$$

The conditional cdf  $F$  is the unique solution to

$$F \left( \log \frac{\epsilon + (1 - \epsilon)e^x}{\epsilon e^x + (1 - \epsilon)} \right) \\ = \epsilon + E \{ (1 - \epsilon)F(x - r(W)) - \epsilon F(-x - r(W)) \}$$

$\forall x \in \mathbb{R}$ , where  $W \sim P_{Y|X}(\cdot|+1)$  and

$$r(y) = \log \frac{P_{Y|X}(y|+1)}{P_{Y|X}(y|-1)}.$$

## Proof: Fixed-point Equation

- $L_i$  is an alternative Markov process (cf. Ordentlich-Weissman)

$$\begin{aligned}L_{i-1} &= \log \frac{P_{X_{i-1}|Y_i^\infty}(+1|Y_i^\infty)}{P_{X_{i-1}|Y_i^\infty}(-1|Y_i^\infty)} \\&= \log \frac{P_{Y_i^\infty|X_{i-1}}(Y_i^\infty|+1)}{P_{Y_i^\infty|X_{i-1}}(Y_i^\infty|-1)} \\&= \log \frac{(1-\epsilon)P_{Y_i^\infty|X_i}(Y_i^\infty|+1) + \epsilon P_{Y_i^\infty|X_i}(Y_i^\infty|-1)}{\epsilon P_{Y_i^\infty|X_i}(Y_i^\infty|+1) + (1-\epsilon)P_{Y_i^\infty|X_i}(Y_i^\infty|-1)} \\&= \log \frac{e^{\alpha+r(Y_i)+L_i} + 1}{e^{r(Y_i)+L_i} + e^\alpha}\end{aligned}$$

where  $\alpha = \log[(1-\epsilon)/\epsilon]$ .

- Density evolution yields the fixed-point equation.

## Proof: Uniqueness

- Want to show the uniqueness of the solution to

$$F(q_\epsilon(x)) = \epsilon + E\{(1 - \epsilon)F(x - r(W)) - \epsilon F(-x - r(W))\}$$

where  $q_\epsilon(x) = \log \frac{\epsilon + (1 - \epsilon)e^x}{\epsilon e^x + (1 - \epsilon)}$ .

- Define mapping  $\Psi$  from {cdf} to {cdf}

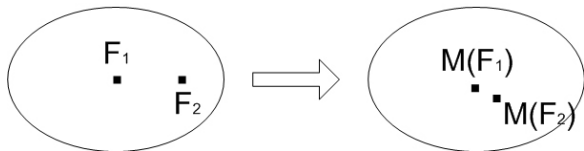
$$\Psi(F)(q_\epsilon(x)) = \epsilon + E\{(1 - \epsilon)F(x - r(W)) - \epsilon F(-x - r(W))\}$$

- Claim:  $\Psi$  is a contraction mapping under  $L^2$ .

# Contraction Mapping $\Rightarrow$ Unique Solution

- $\Psi$  is a contraction mapping if  $\forall F_1, F_2$

$$|\Psi(F_1) - \Psi(F_2)| < \theta |F_1 - F_2|, \quad 0 < \theta < 1.$$



- Denote the inverse of  $q_\epsilon$  by  $f_\epsilon(z) = \log \frac{(1-\epsilon)e^z - \epsilon}{(1-\epsilon) - \epsilon e^z}$ .

## Proof: Contraction

$$\begin{aligned} & |\Psi(F_1) - \Psi(F_2)| \\ &= \int \left| \mathbb{E} \left\{ (1 - \epsilon) \{F_1 - F_2\} (f_\epsilon(u) - r(W)) + \epsilon \{F_2 - F_1\} (-f_\epsilon(u) - r(W)) \right\} \right| du \\ &\leq \mathbb{E} \int (1 - \epsilon) |\{F_1 - F_2\} (f_\epsilon(u) - r(W))| + \epsilon |\{F_1 - F_2\} (f_\epsilon(u) + r(W))| du \\ &= (1 - \epsilon) \mathbb{E} \left\{ \int |F_1(t) - F_2(t)| q'_\epsilon(t + r(W)) dt \right\} \\ &\quad + \epsilon \mathbb{E} \left\{ \int |F_1(t) - F_2(t)| q'_\epsilon(t - r(W)) dt \right\} \\ &\leq (1 - 2\epsilon) |F_1 - F_2|. \end{aligned}$$

# Non-Symmetric Channel

- Channel is characterized by  $P_{Y|X}(\cdot|+1)$  and  $P_{Y|X}(\cdot|-1)$ .
- Let  $F_+(\cdot)$  (resp.  $F_-(\cdot)$ ) denote the cdf of  $L_i$  conditioned on  $X_{i-1} = +1$  (resp.  $X_{i-1} = -1$ ).

## Theorem

The conditional cdfs  $F_+$  and  $F_-$  satisfy

$$\begin{bmatrix} F_+(q_\epsilon(x)) \\ F_-(q_\epsilon(x)) \end{bmatrix} = \begin{bmatrix} 1 - \epsilon & \epsilon \\ \epsilon & 1 - \epsilon \end{bmatrix} \begin{bmatrix} \mathbb{E}\{F_+(x - r(U))\} \\ \mathbb{E}\{F_-(x - r(V))\} \end{bmatrix}$$

$\forall x \in \mathbb{R}$ , where  $U \sim P_{Y|X}(\cdot|+1)$  and  $V \sim P_{Y|X}(\cdot|-1)$  are independent.

# Computation of Entropy Rate

- Recall

$$\lim_{n \rightarrow \infty} \frac{1}{n} H(Y_1^n) = H(P_{Y|X}) + H_2(\epsilon) - H(X_1|X_0, Y_1^\infty).$$

- Therefore

$$\begin{aligned} H(X_1|X_0, Y_1^\infty) &= \mathbb{E} \left\{ H_2 \left( P_{X_1|X_0, Y_1^\infty}(+1|X_0, Y_1^\infty) \right) \middle| X_0, Y_1 \right\} \\ &= \mathbb{E} \left\{ H_2 \left( \frac{1}{1 + \exp[-\alpha X_0 - r(Y_1) - L]} \right) \middle| X_0, Y_1 \right\} \\ &= \int g(z) dF(z) \end{aligned}$$

because  $P_{L|X_0, Y_1}$  is given by  $F$ .

## Special Case: BSC

- Fixed-point equation:

$$F(q_\epsilon(x)) = (1 - \epsilon)(1 - \delta)F(x - \beta) + (1 - \epsilon)\delta F(x + \beta) \\ + \epsilon(1 - \delta)(1 - F(-x - \beta)) + \epsilon\delta(1 - F(-x + \beta))$$

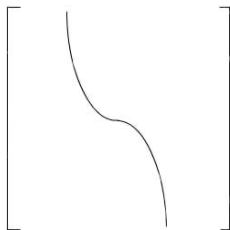
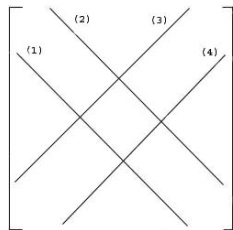
where  $\beta = \log(1/\delta - 1)$ .

- $L$  has finite support.
- It suffices to quantize  $F$  in finite interval.

$$F(q_\epsilon(x)) - (1 - \epsilon)(1 - \delta)F(x - \beta) - (1 - \epsilon)\delta F(x + \beta) + \epsilon(1 - \delta)F(-x - \beta) + \epsilon\delta F(-x + \beta) = \Phi F = \epsilon$$

- Discretize  $F$  and the kernel  $\Phi$ :

$$\mathbf{P}\mathbf{F} - \mathbf{K}\mathbf{F} = \epsilon\mathbf{1}$$

Structure of Matrix  $\mathbf{P}$ Structure of Matrix  $\mathbf{K}$

# Solving the Linear System

- Standard numerical methods.
- Alternatively, quadratic programming:

$$\begin{aligned} \min_{\mathbf{F}} \quad & \|(\mathbf{K} - \mathbf{P})\mathbf{F} + \epsilon\mathbf{1}\|^2 \\ \text{s.t.} \quad & F_i = 0, \quad 1 \leq i \leq M_1; \\ & F_i - F_{i-1} \geq 0, \quad M_1 + 1 \leq i \leq M_1 + M_2 + 1; \\ & F_i = 1, \quad M_1 + M_2 + 1 \leq i \leq M_1 + M_2 + M_3 \end{aligned}$$

where  $F_i$  denotes the  $i$ th element of  $\mathbf{F}$ .

- Can use active set method.
- The complexity is  $O(M^3)$ , where  $M$  is the number of samples of  $F(\cdot)$ .

## Special Case: Gaussian Channel

- Consider

$$Y = \sqrt{\gamma}X + N$$

where  $N \sim \mathcal{N}(0, 1)$ .

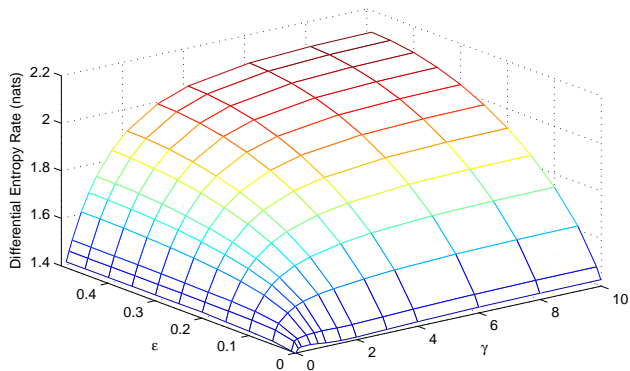
- Fixed-point equation:

$$F(q_\epsilon(x)) = \epsilon + E \{(1 - \epsilon)F(x - 2W) - \epsilon F(-x - 2W)\}$$

with  $W \sim \mathcal{N}(\sqrt{\gamma}, 1)$ .

- Need to quantize the support of distribution of  $W$ ;
- The RHS is the superposition of shifted copies of  $F(\cdot)$ .

# Plot for AWGN Channel



Entropy rate as function of the transition probability and the SNR

## Precision: BSC Case

- Difficulty: The cdf  $F$  is not analytic. Numerically it shows “jumps”.
- Sampling  $F$  may suffer non-vanishing error at some points.

## Precision: BSC Case

- Difficulty: The cdf  $F$  is not analytic. Numerically it shows “jumps”.
- Sampling  $F$  may suffer non-vanishing error at some points.
- Intuition:

$$\int g(z)(dF(z) - d\hat{F}(z)) \rightarrow 0$$

for bounded smooth  $g()$ .

# Quantization Error

- Recall

$$F = \Psi F = \epsilon + E \{ (1 - \epsilon) F(f_\epsilon(\cdot) - r(W)) - \epsilon F(-f_\epsilon(\cdot) - r(W)) \}$$

- Define

$$\hat{F} = \hat{\Psi} \hat{F} = \epsilon + E \{ (1 - \epsilon) \hat{F}(Q(f_\epsilon(\cdot) - r(W))) - \epsilon \hat{F}(Q(-f_\epsilon(\cdot) - r(W))) \}$$

- Two sources of error:

- ▶ Discretizing the kernel  $\Psi$  to obtain  $\hat{F} = \hat{\Psi} \hat{F}$ .
- ▶ Computing the entropy using  $\hat{F}$ .

# Upperbounding the Error

- Let  $\tilde{F}$  be the step function extension of  $\hat{F}$ .

$$\begin{aligned}|F - \tilde{F}| &= |\Psi F - \Psi \tilde{F} + \Psi \tilde{F} - \tilde{F}| \\ &\leq (1 - 2\epsilon)|F - \tilde{F}| + |\tilde{F}' - \hat{F}| \\ &\leq (1 - 2\epsilon)|F - \tilde{F}| + \frac{1}{M}\end{aligned}$$

- Thus

$$|F - \tilde{F}| \leq \frac{1}{2\epsilon M}$$

- This implies

$$|\hat{H} - H| = O\left(\frac{\log M}{M}\right)$$

## Part II

# Joint Detection/Estimation on a HMM

With Yan Zhu and Michael L. Honig.

# Wireless Communication

- Wireless channels:

$$Y_t = H_t X_t + I_t + N_t$$

- Fundamental limiting factors:
  - ▶ Interference.
  - ▶ Channel uncertainty — noise and fading.
- Usual separation of channel estimation and interference cancellation.
- Turbo processing.

# Strong Interference of The Same Signaling Format

- Model

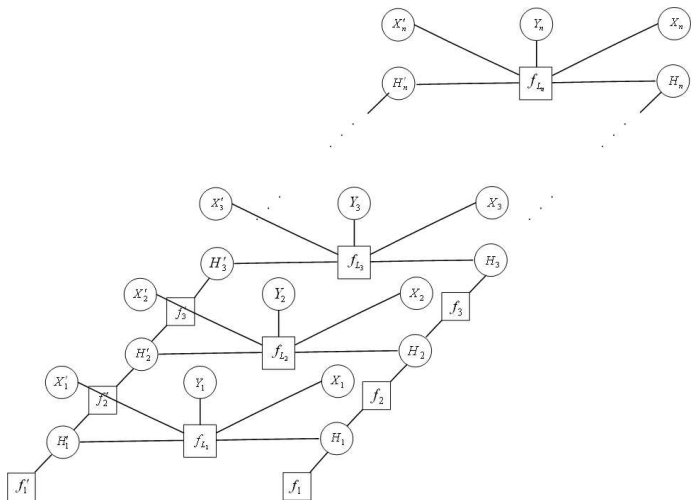
$$Y_i = H_i X_i + H'_i X'_i + N_i \quad i = 1 \dots n$$

- BPSK modulation:  $X_i, X'_i = \pm 1$ .
- Circularly-symmetric complex Gaussian noise.
- Gauss-Markov fading

$$H_i = \alpha H_{i-1} + \sqrt{1 - \alpha^2} W_i$$

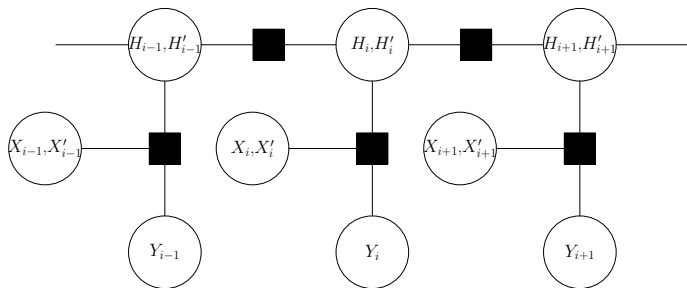
$$H'_i = \alpha H'_{i-1} + \sqrt{1 - \alpha^2} W'_i$$

# A Graphical Model



- Difficult to deal with because of cycles.

# The HMM



# Inference on Graph

- Goal:  $p(x_i|y_1^n)$ ,  $i = 1, \dots, n$ .
- Let  $\mathbf{H}_i = [H_i, H_i']^\dagger$ .

$$\begin{aligned} p(x_i|y_1^n) &= \sum_{x_i'} \int p(x_i, x_i', \mathbf{h}_i | y_1^n) d\mathbf{h}_i \\ &\propto \sum_{x_i'} \int p(y_i, x_i, x_i' | \mathbf{h}_i) p(y_1^{i-1} | \mathbf{h}_i) p(y_{i+1}^n | \mathbf{h}_i) p(\mathbf{h}_i) d\mathbf{h}_i \\ &\propto \sum_{x_i'} p(x_i, x_i') \int p(y_i | \mathbf{h}_i, x_i, x_i') p(\mathbf{h}_i | y_1^{i-1}) p(\mathbf{h}_i | y_{i+1}^n) / p(\mathbf{h}_i) d\mathbf{h}_i \end{aligned}$$

# Filtering of HMP

- Need  $p(\mathbf{h}_i|y_{i+1}^n)$ .
- Recursion:

$$\begin{aligned} p(\mathbf{h}_i|y_{i+1}^n) &= \int p(\mathbf{h}_i|\mathbf{h}_{i+1})p(\mathbf{h}_{i+1}|y_{i+1}^n)d\mathbf{h}_{i+1} \\ &\propto \int p(\mathbf{h}_i|\mathbf{h}_{i+1})p(y_{i+1}|\mathbf{h}_{i+1})p(\mathbf{h}_{i+1}|y_{i+2}^n)d\mathbf{h}_{i+1} \\ &\propto \sum_{x_i, x'_i} \int p(\mathbf{h}_i|\mathbf{h}_{i+1})p(\mathbf{h}_{i+1}|y_{i+2}^n) \\ &\quad p(y_{i+1}|\mathbf{h}_{i+1}, x_{i+1}, x'_{i+1})p(x_{i+1}, x'_{i+1}) d\mathbf{h}_{i+1} \end{aligned}$$

# Gaussian Mixture

- The posteriors  $p(\mathbf{h}_i|y_{i+1}^n)$  and  $p(\mathbf{h}_i|y_1^{i-1})$  are mixture Gaussian.
- Assuming that  $p(\mathbf{h}_{i+1}|y_{i+2}^n) = \sum_j \rho_j \mathcal{CN}(\mathbf{h}_{i+1}, m_j, K_j)$ ,

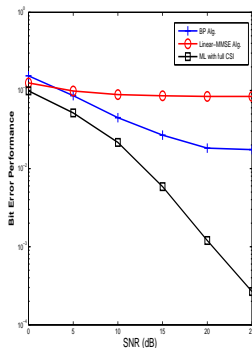
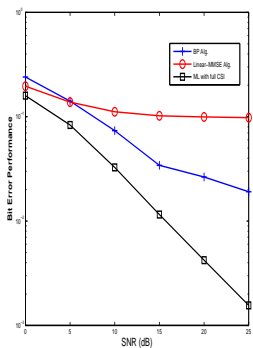
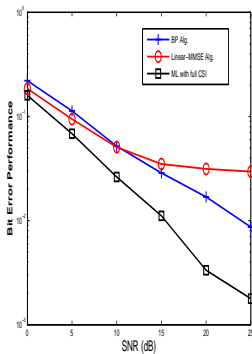
$$p(\mathbf{h}_i|y_{i+1}^n) \propto \sum_j \sum_{z_{i+1}} \rho_j p(z_{i+1}) \mathcal{CN}(\dots) \mathcal{CN}(\dots)$$

- Basically testing hypotheses as recursion proceeds.

# Representation of Gaussian Mixture

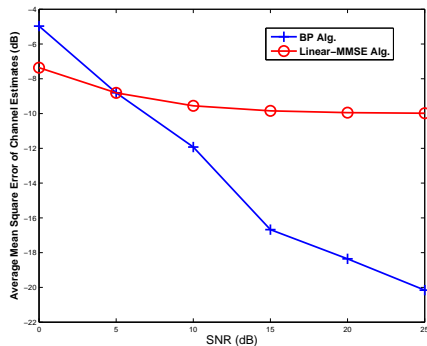
- The number of Gaussian components increases exponentially with recursion.
- Constrain to a fixed number of Gaussian components.
- Reduced representation
  - ▶ Keep those with larger amplitudes (stronger opinions).
  - ▶ Merge components.

# Comparison of BP and Linear Estimation



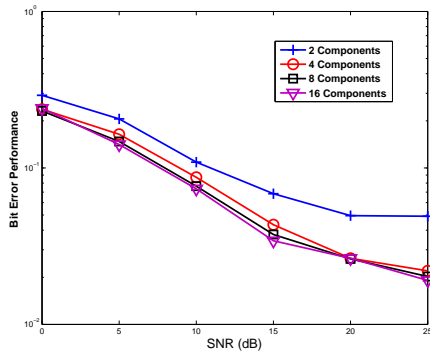
Interference 10 dB weaker / 3 dB weaker / equally strong.

# Channel Estimation Error



Interference is 3 dB weaker than the desired one.

# Performance vs. Number of Gaussian Components



# Conclusion

- Fixed-point functional equations characterize the stationary distribution of a filtering process.
- Numerical methods for approximating the entropy.
- A communication problem on a hidden Markov model.
- Challenges ...