

Deterministic algorithms for computing/approximating Hidden Markov Process entropy rates

Erik Ordentlich, HP Labs

BIRS Workshop, Banff
October 2, 2007

Notation and Setting

- $x^n \triangleq x_1, \dots, x_n$

- Finite valued stationary underlying Markov process X

$$\pi_{ij} = Pr(X_t = j | X_{t-1} = i)$$

$$\lambda_i = Pr(X_t = i)$$

- Finite valued memoryless observations Y

$$\delta_{ij} = Pr(Y_t = j | X_t = i)$$

- HMP distribution

$$P_{Y^n}(y^n) = \sum_{x^n} \lambda_{x_1} \delta_{x_1 y_1} \prod_{t=2}^n \pi_{x_{t-1} x_t} \delta_{x_t y_t}$$

- Entropy rate

$$\mathcal{H}(Y) = \lim_{n \rightarrow \infty} H(Y_n | Y^{n-1})$$

Why compute?

- As already mentioned: Estimate information rates of finite state channels:
 $Y = X + Z$ where
 - signal X is i.i.d. and noise Z is Markov - e.g. Gilbert-Elliott channel.
 - X is Markov and Z is i.i.d.

Why compute?

- As already mentioned: Estimate information rates of finite state channels:
 $Y = X + Z$ where
 - signal X is i.i.d. and noise Z is Markov - e.g. Gilbert-Elliott channel.
 - X is Markov and Z is i.i.d.
- Recent computation for a paper:
 - n samples of a binary Markov source X are transmitted to a receiver over a memoryless binary symmetric channel $Y = X \oplus Z$
 - How many additional (coded) bits $m(n)$ must be transmitted to allow receiver to reconstruct X^n with probability going to 1?

Why compute?

- As already mentioned: Estimate information rates of finite state channels:
 $Y = X + Z$ where
 - signal X is i.i.d. and noise Z is Markov - e.g. Gilbert-Elliott channel.
 - X is Markov and Z is i.i.d.
- Recent computation for a paper:
 - n samples of a binary Markov source X are transmitted to a receiver over a memoryless binary symmetric channel $Y = X \oplus Z$
 - How many additional (coded) bits $m(n)$ must be transmitted to allow receiver to reconstruct X^n with probability going to 1?
 - Theorem [Shamai, Verdú (2003)]:

$$m(n) = n \cdot \frac{\mathcal{H}(X|Y)}{C}$$

where $\mathcal{H}(X|Y) = \mathcal{H}(X) + \mathcal{H}(Z) - \mathcal{H}(Y)$.

Monte Carlo approximation

- Pfister-Siegel (2001), Arnold-Loeliger (2001), Sharma-Singh (2001):

Monte Carlo approximation

- Pfister-Siegel (2001), Arnold-Loeliger (2001), Sharma-Singh (2001): Simulate Y^n for n very large and set

$$\hat{\mathcal{H}}(Y) = \frac{1}{n} \log(P(Y^n))$$

- By Shannon-McMillan-Breiman Theorem $\hat{\mathcal{H}}(Y) \rightarrow \mathcal{H}(Y)$ with probability one.
- Focus of this talk: deterministic algorithms with provable error bounds.

Deterministic approximation

- Approximation algorithm \mathcal{A} :
 - Inputs: HMP parameters $\{\pi_{ij}\}, \{\delta_{ij}\}$, absolute precision ϵ
 - Output: $\hat{\mathcal{H}}(Y)$ satisfying $|\hat{\mathcal{H}}(Y) - \mathcal{H}(Y)| \leq \epsilon$

Deterministic approximation

- Approximation algorithm \mathcal{A} :
 - Inputs: HMP parameters $\{\pi_{ij}\}, \{\delta_{ij}\}$, absolute precision ϵ
 - Output: $\hat{\mathcal{H}}(Y)$ satisfying $|\hat{\mathcal{H}}(Y) - \mathcal{H}(Y)| \leq \epsilon$
- Complexity vs. precision:
 - $N(\epsilon) \triangleq$ number of “operations” required by $\mathcal{A}(\{\pi_{ij}\}, \{\delta_{ij}\}, \epsilon)$

Deterministic approximation

- Approximation algorithm \mathcal{A} :
 - Inputs: HMP parameters $\{\pi_{ij}\}, \{\delta_{ij}\}$, absolute precision ϵ
 - Output: $\hat{\mathcal{H}}(Y)$ satisfying $|\hat{\mathcal{H}}(Y) - \mathcal{H}(Y)| \leq \epsilon$
- Complexity vs. precision:
 - $N(\epsilon) \triangleq$ number of “operations” required by $\mathcal{A}(\{\pi_{ij}\}, \{\delta_{ij}\}, \epsilon)$
 - For \mathcal{A} to be useful, $N(\epsilon)$ should be (low degree) polynomial in ϵ^{-1} .

Deterministic approximation

- Approximation algorithm \mathcal{A} :
 - Inputs: HMP parameters $\{\pi_{ij}\}, \{\delta_{ij}\}$, absolute precision ϵ
 - Output: $\hat{\mathcal{H}}(Y)$ satisfying $|\hat{\mathcal{H}}(Y) - \mathcal{H}(Y)| \leq \epsilon$
- Complexity vs. precision:
 - $N(\epsilon) \triangleq$ number of “operations” required by $\mathcal{A}(\{\pi_{ij}\}, \{\delta_{ij}\}, \epsilon)$
 - For \mathcal{A} to be useful, $N(\epsilon)$ should be (low degree) polynomial in ϵ^{-1} .
- Key problems:
 - Characterize $N(\epsilon)$ for known approximation algorithms.

Deterministic approximation

- Approximation algorithm \mathcal{A} :
 - Inputs: HMP parameters $\{\pi_{ij}\}, \{\delta_{ij}\}$, absolute precision ϵ
 - Output: $\hat{\mathcal{H}}(Y)$ satisfying $|\hat{\mathcal{H}}(Y) - \mathcal{H}(Y)| \leq \epsilon$
- Complexity vs. precision:
 - $N(\epsilon) \triangleq$ number of “operations” required by $\mathcal{A}(\{\pi_{ij}\}, \{\delta_{ij}\}, \epsilon)$
 - For \mathcal{A} to be useful, $N(\epsilon)$ should be (low degree) polynomial in ϵ^{-1} .
- Key problems:
 - Characterize $N(\epsilon)$ for known approximation algorithms.
 - Find algorithms/improvements with smaller $N(\epsilon)$.

Running example

- Binary HMP $Y_t \in \{0, 1\}$ with underlying **stationary** Markov process $X_t \in \{0, 1\}$:

$$\{\pi_{ij}\} = \begin{bmatrix} 1 - \pi_{01} & \pi_{01} \\ \pi_{10} & 1 - \pi_{10} \end{bmatrix}, \quad \{\delta_{ij}\} = \begin{bmatrix} 1 - \delta_{01} & \delta_{01} \\ \delta_{10} & 1 - \delta_{10} \end{bmatrix}$$

- For simplicity, mostly we take $\delta_{01} = \delta_{10} = \delta$:

$$Y = \begin{array}{ccc} X & \oplus & Z \\ \uparrow & & \uparrow \\ \text{Markov} & & \text{i.i.d. Bernoulli } \delta \end{array}$$

References

- Cover and Thomas (Elements of Information Theory)
- Hochwald and Jelenković (State learning and mixing in entropy ... 1999)
- Pfister (Chapter 4, PhD Thesis, 2003)
- Le Gland and Mevel (Basic properties of the projective product ... 2000)
- Ordentlich and Weissman (Bounds on the entropy rate ... in prep. and ISIT 2005)
- Egner et. al. (On the entropy rate ... ISIT 2004)

Method 1: Birch bounds

- $$H(Y_n|Y^{n-1}, X_0) \leq \mathcal{H}(Y) \leq H(Y_n|Y^{n-1})$$

where

$$H(Y_n|Y^{n-1}, X_0) = -E \log P(Y_n|Y^{n-1}, X_0)$$

$$H(Y_n|Y^{n-1}) = -E \log P(Y_n|Y^{n-1})$$

- Compute, for fixed n , by brute force evaluation of expectations:
 - eg. for upper bound determine $P(y^n) \log P(y_n|y^{n-1})$ and sum over 2^n sequences y^n .

Method 1: Birch bounds

- $$H(Y_n|Y^{n-1}, X_0) \leq \mathcal{H}(Y) \leq H(Y_n|Y^{n-1})$$

where

$$H(Y_n|Y^{n-1}, X_0) = -E \log P(Y_n|Y^{n-1}, X_0)$$

$$H(Y_n|Y^{n-1}) = -E \log P(Y_n|Y^{n-1})$$

- Compute, for fixed n , by brute force evaluation of expectations:
 - eg. for upper bound determine $P(y^n) \log P(y_n|y^{n-1})$ and sum over 2^n sequences y^n .
- Approximation algorithm: Given ϵ evaluate upper and lower bounds for increasing n until $H(Y_n|Y^{n-1}) - H(Y_n|Y^{n-1}, X_0) \leq \epsilon$
- What is $N(\epsilon)$?

Simple precision bound

- Stationarity and convexity imply that $H(Y_n|Y^{n-1})$ is decreasing in n (more conditioning) while $H(Y_n|Y^{n-1}, X_0)$ is increasing in n (less conditioning - known state further in the past).

Simple precision bound

- Stationarity and convexity imply that $H(Y_n|Y^{n-1})$ is decreasing in n (more conditioning) while $H(Y_n|Y^{n-1}, X_0)$ is increasing in n (less conditioning - known state further in the past).
- Therefore

$$\begin{aligned} H(Y_n|Y^{n-1}) - H(Y_n|Y^{n-1}, X_0) &\leq \frac{1}{n} \sum_{j=1}^n H(Y_j|Y^{j-1}) - H(Y_j|Y^{j-1}, X_0) \\ &= \frac{1}{n} (H(X_0) - H(X_0|Y^n)) \\ &\leq \frac{1}{n} \end{aligned}$$

Simple precision bound

- Stationarity and convexity imply that $H(Y_n|Y^{n-1})$ is decreasing in n (more conditioning) while $H(Y_n|Y^{n-1}, X_0)$ is increasing in n (less conditioning - known state further in the past).
- Therefore

$$\begin{aligned} H(Y_n|Y^{n-1}) - H(Y_n|Y^{n-1}, X_0) &\leq \frac{1}{n} \sum_{j=1}^n H(Y_j|Y^{j-1}) - H(Y_j|Y^{j-1}, X_0) \\ &= \frac{1}{n} (H(X_0) - H(X_0|Y^n)) \\ &\leq \frac{1}{n} \end{aligned}$$

- This is bad: To get precision below ϵ requires $n = \epsilon^{-1}$ implying $N(\epsilon) = O(\epsilon^{-1} 2^{\epsilon^{-1}})$

Exponential convergence

Theorem [Birch 1962, and others]:

$$H(Y_n|Y^{n-1}) - H(Y_n|Y^{n-1}, X_0) \leq B\eta^{n-1}$$

where B, η depend on $\delta, \{\pi_{ij}\}$.

- To get precision below ϵ requires $n \approx \log_2(B\epsilon^{-1}) / \log_2(\eta^{-1})$.
- Implies $N(\epsilon) (\approx 2^n)$ is polynomial in ϵ^{-1} with degree $1 / \log_2(\eta^{-1})$.

What is known about η ?

True η unknown.

What is known about η ?

True η unknown.

- Birch (1962):

$$\text{For } \pi_{ij} > 0, \eta = 1 - \frac{1}{4} \left(\frac{\delta}{1-\delta} \right)^4 \min_{i,j,k,l,m} \left[\frac{\pi_{ij}\pi_{jk}}{\pi_{il}\pi_{lm}} \right]^2$$

- η always greater than $1/4$, even for X i.i.d.
- $\eta \rightarrow 1$ as any $\pi_{ij} \rightarrow 0$

Improvements using contraction

- Hochwald-Jelenković (1999):

For $\pi_{ij} > 0$,
$$\eta = \frac{|1 - \pi_{01} - \pi_{10}|}{[\sqrt{(1 - \pi_{01})(1 - \pi_{10})} + \sqrt{\pi_{01}\pi_{10}}]^2}$$

- η equals Birkhoff's contraction coefficient of $\{\pi_{ij}\}$ (extends to greater than 2 states)
- $\eta \rightarrow 0$ as X approaches i.i.d.
- For $\pi_{01} = \pi_{10} = \pi$, $\eta = (1 - 2\pi)$.
- Still $\eta \rightarrow 1$ as any $\pi_{ij} \rightarrow 0$

Improvements using contraction

- Hochwald-Jelenković (1999):

For $\pi_{ij} > 0$,
$$\eta = \frac{|1 - \pi_{01} - \pi_{10}|}{[\sqrt{(1 - \pi_{01})(1 - \pi_{10})} + \sqrt{\pi_{01}\pi_{10}}]^2}$$

- η equals Birkhoff's contraction coefficient of $\{\pi_{ij}\}$ (extends to greater than 2 states)
 - $\eta \rightarrow 0$ as X approaches i.i.d.
 - For $\pi_{01} = \pi_{10} = \pi$, $\eta = (1 - 2\pi)$.
 - Still $\eta \rightarrow 1$ as any $\pi_{ij} \rightarrow 0$
- LeGland-Mevel (2000), Pfister (2003):
Even with some $\pi_{ij} = 0$ obtain $\eta < 1$. Still, $\eta \rightarrow 1$ as any non-zero $\pi_{ij} \rightarrow 0$.

Further improvements: likelihood ratio process

(as in O-W 2005 and in prep.)

- Define likelihood ratio processes:

$$\ell_n(y^n) = \log \frac{P(X_n=1|y^n)}{P(X_n=0|y^n)} \quad \text{and} \quad \ell'_n(y^n, x_0) = \log \frac{P(X_n=1|y^n, x_0)}{P(X_n=0|y^n, x_0)} .$$

Further improvements: likelihood ratio process

(as in O-W 2005 and in prep.)

- Define likelihood ratio processes:

$$\ell_n(y^n) = \log \frac{P(X_n=1|y^n)}{P(X_n=0|y^n)} \quad \text{and} \quad \ell'_n(y^n, x_0) = \log \frac{P(X_n=1|y^n, x_0)}{P(X_n=0|y^n, x_0)} .$$

- Observation:

$$H(Y_{n+1}|Y^n) = E \log P(Y_{n+1}|Y^n) = E g(\ell_n(Y^n))$$

$$H(Y_{n+1}|Y^n, X_0) = E \log P(Y_{n+1}|Y^n, X_0) = E g(\ell'_n(Y^n, X_0))$$

for a “nice” $g(x)$ (with bounded derivatives when $\delta \in (0, 1)$, and other cases).

Further improvements: refined use of contraction

- Define $F(x, y) = (2y - 1) \log \frac{1-\delta}{\delta} + f(x)$ where $f(x) = \log \frac{\pi_{01} + e^x(1-\pi_{10})}{(1-\pi_{01}) + e^x \pi_{10}}$.

Further improvements: refined use of contraction

- Define $F(x, y) = (2y - 1) \log \frac{1-\delta}{\delta} + f(x)$ where $f(x) = \log \frac{\pi_{01} + e^x(1-\pi_{10})}{(1-\pi_{01}) + e^x\pi_{10}}$.
- $\ell_n(y^n) = F(\ell_{n-1}, y_n)$ with $\ell_0 = \log \frac{\pi_{01}}{\pi_{10}}$
 $\ell'_n(y^n, x_0) = F(\ell'_{n-1}, y_n)$ with $\ell'_0 = (2x_0 - 1)\infty$

Further improvements: refined use of contraction

- Define $F(x, y) = (2y - 1) \log \frac{1-\delta}{\delta} + f(x)$ where $f(x) = \log \frac{\pi_{01} + e^x(1-\pi_{10})}{(1-\pi_{01}) + e^x \pi_{10}}$.
- $\ell_n(y^n) = F(\ell_{n-1}, y_n)$ with $\ell_0 = \log \frac{\pi_{01}}{\pi_{10}}$
 $\ell'_n(y^n, x_0) = F(\ell'_{n-1}, y_n)$ with $\ell'_0 = (2x_0 - 1)\infty$
- So $|\ell_n - \ell'_n| = |f(\ell_{n-1}) - f(\ell'_{n-1})|$

Further improvements: refined use of contraction

- Define $F(x, y) = (2y - 1) \log \frac{1-\delta}{\delta} + f(x)$ where $f(x) = \log \frac{\pi_{01} + e^x(1-\pi_{10})}{(1-\pi_{01}) + e^x \pi_{10}}$.
- $\ell_n(y^n) = F(\ell_{n-1}, y_n)$ with $\ell_0 = \log \frac{\pi_{01}}{\pi_{10}}$
 $\ell'_n(y^n, x_0) = F(\ell'_{n-1}, y_n)$ with $\ell'_0 = (2x_0 - 1)\infty$
- So $|\ell_n - \ell'_n| = |f(\ell_{n-1}) - f(\ell'_{n-1})| \rightarrow |\ell_n - \ell'_n| \leq \eta_n |\ell_{n-1} - \ell'_{n-1}|$

Further improvements: refined use of contraction

- Define $F(x, y) = (2y - 1) \log \frac{1-\delta}{\delta} + f(x)$ where $f(x) = \log \frac{\pi_{01} + e^x(1-\pi_{10})}{(1-\pi_{01}) + e^x\pi_{10}}$.
- $\ell_n(y^n) = F(\ell_{n-1}, y_n)$ with $\ell_0 = \log \frac{\pi_{01}}{\pi_{10}}$
 $\ell'_n(y^n, x_0) = F(\ell'_{n-1}, y_n)$ with $\ell'_0 = (2x_0 - 1)\infty$
- So $|\ell_n - \ell'_n| = |f(\ell_{n-1}) - f(\ell'_{n-1})| \rightarrow |\ell_n - \ell'_n| \leq \eta_n |\ell_{n-1} - \ell'_{n-1}|$
where $\eta_n = \sup_{x \in \mathcal{S}_n} |f'(x)|$ and \mathcal{S}_n is a finite union of intervals such that $[\ell_{n-1}, \ell'_{n-1}] \subseteq \mathcal{S}_n$ for all y^{n-1}, x_0

Further improvements: refined use of contraction

- Define $F(x, y) = (2y - 1) \log \frac{1-\delta}{\delta} + f(x)$ where $f(x) = \log \frac{\pi_{01} + e^x(1-\pi_{10})}{(1-\pi_{01}) + e^x \pi_{10}}$.
- $\ell_n(y^n) = F(\ell_{n-1}, y_n)$ with $\ell_0 = \log \frac{\pi_{01}}{\pi_{10}}$
 $\ell'_n(y^n, x_0) = F(\ell'_{n-1}, y_n)$ with $\ell'_0 = (2x_0 - 1)\infty$
- So $|\ell_n - \ell'_n| = |f(\ell_{n-1}) - f(\ell'_{n-1})| \rightarrow |\ell_n - \ell'_n| \leq \eta_n |\ell_{n-1} - \ell'_{n-1}|$
where $\eta_n = \sup_{x \in \mathcal{S}_n} |f'(x)|$ and \mathcal{S}_n is a finite union of intervals such that $[\ell_{n-1}, \ell'_{n-1}] \subseteq \mathcal{S}_n$ for all y^{n-1}, x_0
- Let $\eta \geq \sup_{n > N} \eta_n$.

Further improvements: refined use of contraction

- Define $F(x, y) = (2y - 1) \log \frac{1-\delta}{\delta} + f(x)$ where $f(x) = \log \frac{\pi_{01} + e^x(1-\pi_{10})}{(1-\pi_{01}) + e^x\pi_{10}}$.
- $\ell_n(y^n) = F(\ell_{n-1}, y_n)$ with $\ell_0 = \log \frac{\pi_{01}}{\pi_{10}}$
 $\ell'_n(y^n, x_0) = F(\ell'_{n-1}, y_n)$ with $\ell'_0 = (2x_0 - 1)\infty$
- So $|\ell_n - \ell'_n| = |f(\ell_{n-1}) - f(\ell'_{n-1})| \rightarrow |\ell_n - \ell'_n| \leq \eta_n |\ell_{n-1} - \ell'_{n-1}|$
where $\eta_n = \sup_{x \in \mathcal{S}_n} |f'(x)|$ and \mathcal{S}_n is a finite union of intervals such that $[\ell_{n-1}, \ell'_{n-1}] \subseteq \mathcal{S}_n$ for all y^{n-1}, x_0
- Let $\eta \geq \sup_{n > N} \eta_n$.
- Then $|\ell_n - \ell'_n| = C\eta^{n-1}$

Further improvements: refined use of contraction

- Define $F(x, y) = (2y - 1) \log \frac{1-\delta}{\delta} + f(x)$ where $f(x) = \log \frac{\pi_{01} + e^x(1-\pi_{10})}{(1-\pi_{01}) + e^x\pi_{10}}$.
- $\ell_n(y^n) = F(\ell_{n-1}, y_n)$ with $\ell_0 = \log \frac{\pi_{01}}{\pi_{10}}$
 $\ell'_n(y^n, x_0) = F(\ell'_{n-1}, y_n)$ with $\ell'_0 = (2x_0 - 1)\infty$
- So $|\ell_n - \ell'_n| = |f(\ell_{n-1}) - f(\ell'_{n-1})| \rightarrow |\ell_n - \ell'_n| \leq \eta_n |\ell_{n-1} - \ell'_{n-1}|$
where $\eta_n = \sup_{x \in \mathcal{S}_n} |f'(x)|$ and \mathcal{S}_n is a finite union of intervals such that $[\ell_{n-1}, \ell'_{n-1}] \subseteq \mathcal{S}_n$ for all y^{n-1}, x_0
- Let $\eta \geq \sup_{n > N} \eta_n$.
- Then $|\ell_n - \ell'_n| = C\eta^{n-1} \rightarrow |g(\ell_n) - g(\ell'_n)| = B\eta^{n-1}$

Further improvements: refined use of contraction

- Define $F(x, y) = (2y - 1) \log \frac{1-\delta}{\delta} + f(x)$ where $f(x) = \log \frac{\pi_{01} + e^x(1-\pi_{10})}{(1-\pi_{01}) + e^x\pi_{10}}$.
- $\ell_n(y^n) = F(\ell_{n-1}, y_n)$ with $\ell_0 = \log \frac{\pi_{01}}{\pi_{10}}$
 $\ell'_n(y^n, x_0) = F(\ell'_{n-1}, y_n)$ with $\ell'_0 = (2x_0 - 1)\infty$
- So $|\ell_n - \ell'_n| = |f(\ell_{n-1}) - f(\ell'_{n-1})| \rightarrow |\ell_n - \ell'_n| \leq \eta_n |\ell_{n-1} - \ell'_{n-1}|$
where $\eta_n = \sup_{x \in \mathcal{S}_n} |f'(x)|$ and \mathcal{S}_n is a finite union of intervals such that $[\ell_{n-1}, \ell'_{n-1}] \subseteq \mathcal{S}_n$ for all y^{n-1}, x_0
- Let $\eta \geq \sup_{n > N} \eta_n$.
- Then $|\ell_n - \ell'_n| = C\eta^{n-1} \rightarrow |g(\ell_n) - g(\ell'_n)| = B\eta^{n-1} \rightarrow H(Y_{n+1}|Y^n) - H(Y_{n+1}, X_0) = B\eta^{n-1}$

Further improvements: refined use of contraction

- Define $F(x, y) = (2y - 1) \log \frac{1-\delta}{\delta} + f(x)$ where $f(x) = \log \frac{\pi_{01} + e^x(1-\pi_{10})}{(1-\pi_{01}) + e^x\pi_{10}}$.
- $\ell_n(y^n) = F(\ell_{n-1}, y_n)$ with $\ell_0 = \log \frac{\pi_{01}}{\pi_{10}}$
 $\ell'_n(y^n, x_0) = F(\ell'_{n-1}, y_n)$ with $\ell'_0 = (2x_0 - 1)\infty$
- So $|\ell_n - \ell'_n| = |f(\ell_{n-1}) - f(\ell'_{n-1})| \rightarrow |\ell_n - \ell'_n| \leq \eta_n |\ell_{n-1} - \ell'_{n-1}|$
where $\eta_n = \sup_{x \in \mathcal{S}_n} |f'(x)|$ and \mathcal{S}_n is a finite union of intervals such that $[\ell_{n-1}, \ell'_{n-1}] \subseteq \mathcal{S}_n$ for all y^{n-1}, x_0
- Let $\eta \geq \sup_{n > N} \eta_n$.
- Then $|\ell_n - \ell'_n| = C\eta^{n-1} \rightarrow |g(\ell_n) - g(\ell'_n)| = B\eta^{n-1} \rightarrow H(Y_{n+1}|Y^n) - H(Y_{n+1}, X_0) = B\eta^{n-1}$
- This η improves on previously obtained η for some corner cases like $\pi_{01} > 0, \pi_{10} \rightarrow 1$.

Open issues with Method 1

- Universal factor η for decay of $H(Y_n|Y^{n-1}) - H(Y_n|Y^{n-1}, X_0)$ even as Markov chain becomes less mixing?

Open issues with Method 1

- Universal factor η for decay of $H(Y_n|Y^{n-1}) - H(Y_n|Y^{n-1}, X_0)$ even as Markov chain becomes less mixing?
 - Hochwald-Jelenković (1999) show this in special cases.

Open issues with Method 1

- Universal factor η for decay of $H(Y_n|Y^{n-1}) - H(Y_n|Y^{n-1}, X_0)$ even as Markov chain becomes less mixing?
 - Hochwald-Jelenković (1999) show this in special cases.
- What is the true η for a given set of parameters?

Method 2: Variable length conditioning improvement of Birch

- Egner, et al. (2004):

$$-E \log P(Y_0 | Y_\tau^{-1}, X_{\tau-1}) \leq \mathcal{H}(Y) \leq -E \log P(Y_0 | Y_\tau^{-1})$$

where τ is a bounded reverse stopping time starting at -1 and extending towards the negative.

Method 2: Variable length conditioning improvement of Birch

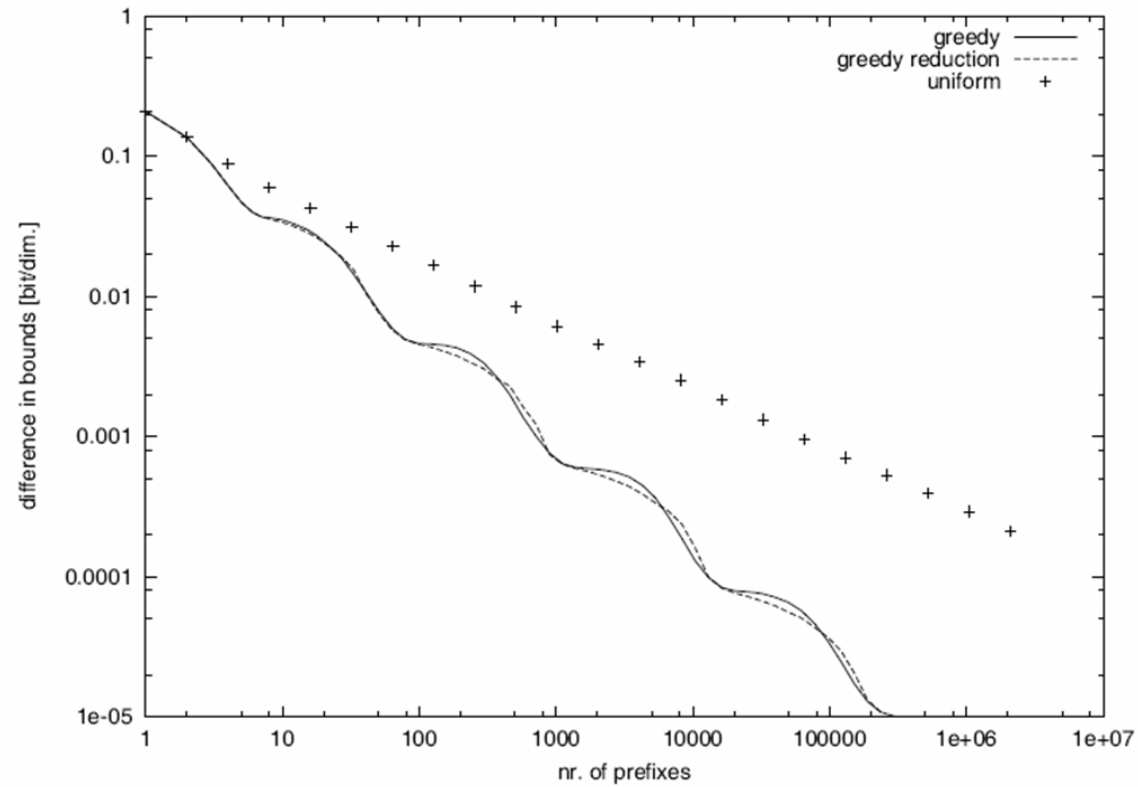
- Egner, et al. (2004):

$$-E \log P(Y_0 | Y_\tau^{-1}, X_{\tau-1}) \leq \mathcal{H}(Y) \leq -E \log P(Y_0 | Y_\tau^{-1})$$

where τ is a bounded reverse stopping time starting at -1 and extending towards the negative.

- $\tau = \text{constant}$ \rightarrow Birch bounds
- Egner, et al. give heuristics for finding good τ : close bounds, with few stopped sequences to sum over.

Method 2 vs. Birch



Gap between bounds versus number of terms summed over in approximation

$$(\pi_{01} = 10^{-5}, \pi_{10} = 5 \times 10^{-5}, \delta_{01} = 10^{-2}, \delta_{10} = 1/2).$$

Method 3: Quantization of likelihood ratio process

(as in O-W (2005) and in prep.)

- See also Pfister (2003).

Method 3: Quantization of likelihood ratio process

(as in O-W (2005) and in prep.)

- See also Pfister (2003).

- Recall:

$$H(Y_{n+1}|Y^n) = E \log P(Y_{n+1}|Y^n) = E g(\ell_n(Y^n))$$

Method 3: Quantization of likelihood ratio process

(as in O-W (2005) and in prep.)

- See also Pfister (2003).

- Recall:

$$H(Y_{n+1}|Y^n) = E \log P(Y_{n+1}|Y^n) = E g(\ell_n(Y^n))$$

- Still, determining distribution of ℓ_n is hard (Blackwell 1957).

Method 3: Quantization of likelihood ratio process

(as in O-W (2005) and in prep.)

- See also Pfister (2003).

- Recall:

$$H(Y_{n+1}|Y^n) = E \log P(Y_{n+1}|Y^n) = Eg(\ell_n(Y^n))$$

- Still, determining distribution of ℓ_n is hard (Blackwell 1957).
- Idea: Define a finite valued (quantized) process $\hat{\ell}_n$ satisfying

- Property 1: For all n

$$|\ell_n(y^n) - \hat{\ell}_n(y^n)| \leq c\epsilon$$

and, therefore,

$$|Eg(\ell_n(Y^n)) - Eg(\hat{\ell}_n(Y^n))| \leq \epsilon$$

Method 3: Quantization of likelihood ratio process

(as in O-W (2005) and in prep.)

- See also Pfister (2003).

- Recall:

$$H(Y_{n+1}|Y^n) = E \log P(Y_{n+1}|Y^n) = Eg(\ell_n(Y^n))$$

- Still, determining distribution of ℓ_n is hard (Blackwell 1957).
- Idea: Define a finite valued (quantized) process $\hat{\ell}_n$ satisfying

- Property 1: For all n

$$|\ell_n(y^n) - \hat{\ell}_n(y^n)| \leq c\epsilon$$

and, therefore,

$$|Eg(\ell_n(Y^n)) - Eg(\hat{\ell}_n(Y^n))| \leq \epsilon$$

- Property 2: $\hat{\mathcal{H}} = \lim_{n \rightarrow \infty} Eg(\hat{\ell}_n(Y^n))$ is “easily” computed.

Method 3: Quantization of likelihood ratio process

(as in O-W (2005) and in prep.)

- See also Pfister (2003).

- Recall:

$$H(Y_{n+1}|Y^n) = E \log P(Y_{n+1}|Y^n) = Eg(\ell_n(Y^n))$$

- Still, determining distribution of ℓ_n is hard (Blackwell 1957).
- Idea: Define a finite valued (quantized) process $\hat{\ell}_n$ satisfying

- Property 1: For all n

$$|\ell_n(y^n) - \hat{\ell}_n(y^n)| \leq c\epsilon$$

and, therefore,

$$|Eg(\ell_n(Y^n)) - Eg(\hat{\ell}_n(Y^n))| \leq \epsilon$$

- Property 2: $\hat{\mathcal{H}} = \lim_{n \rightarrow \infty} Eg(\hat{\ell}_n(Y^n))$ is “easily” computed.

- Then $|\hat{\mathcal{H}} - \mathcal{H}| \leq \epsilon$.

Definition of $\hat{\ell}_n$

- Recall for $F(x, y) = (2y - 1) \log \frac{1-\delta}{\delta} + f(x)$ where $f(x) = \log \frac{\pi_{01} + e^x(1-\pi_{10})}{(1-\pi_{01}) + e^x \pi_{10}}$, that

$$\ell_n = F(\ell_{n-1}, y_n), \quad \ell_0 = \log \frac{\pi_{01}}{\pi_{10}}$$

Definition of $\hat{\ell}_n$

- Recall for $F(x, y) = (2y - 1) \log \frac{1-\delta}{\delta} + f(x)$ where $f(x) = \log \frac{\pi_{01} + e^x(1-\pi_{10})}{(1-\pi_{01}) + e^x \pi_{10}}$, that

$$\ell_n = F(\ell_{n-1}, y_n), \quad \ell_0 = \log \frac{\pi_{01}}{\pi_{10}}$$

- Let \mathcal{S} be an interval containing supports of ℓ_n for all n .

Definition of $\hat{\ell}_n$

- Recall for $F(x, y) = (2y - 1) \log \frac{1-\delta}{\delta} + f(x)$ where $f(x) = \log \frac{\pi_{01} + e^x(1-\pi_{10})}{(1-\pi_{01}) + e^x \pi_{10}}$, that

$$\ell_n = F(\ell_{n-1}, y_n), \quad \ell_0 = \log \frac{\pi_{01}}{\pi_{10}}$$

- Let \mathcal{S} be an interval containing supports of ℓ_n for all n .
- Let $Q_\nu : \mathcal{S} \rightarrow \mathcal{S}$ be finite valued quantizer satisfying $|Q_\nu(x) - x| \leq \nu$ for all $x \in \mathcal{S}$
 - $|Q_\nu| \leq \lceil \mu(\mathcal{S}) / (2\nu) \rceil$.

Definition of $\hat{\ell}_n$

- Recall for $F(x, y) = (2y - 1) \log \frac{1-\delta}{\delta} + f(x)$ where $f(x) = \log \frac{\pi_{01} + e^x(1-\pi_{10})}{(1-\pi_{01}) + e^x \pi_{10}}$, that

$$\ell_n = F(\ell_{n-1}, y_n), \quad \ell_0 = \log \frac{\pi_{01}}{\pi_{10}}$$

- Let \mathcal{S} be an interval containing supports of ℓ_n for all n .
- Let $Q_\nu : \mathcal{S} \rightarrow \mathcal{S}$ be finite valued quantizer satisfying $|Q_\nu(x) - x| \leq \nu$ for all $x \in \mathcal{S}$
 - $|Q_\nu| \leq \lceil \mu(\mathcal{S}) / (2\nu) \rceil$.
- Define $\hat{\ell}_n$ as

$$\hat{\ell}_n = Q_\nu(F(\hat{\ell}_{n-1}, y_n)), \quad \hat{\ell}_0 = Q_\nu\left(\log \frac{\pi_{01}}{\pi_{10}}\right)$$

Checking property 1: \hat{l}_n close to l_n

- Accumulation of quantization errors offset by contraction of mapping F .

Checking property 1: $\hat{\ell}_n$ close to ℓ_n

- Accumulation of quantization errors offset by contraction of mapping F .



$$|\ell_n - \hat{\ell}_n| = |F(\ell_{n-1}, y_n) - Q_\nu(F(\hat{\ell}_{n-1}, y_n))|$$

Checking property 1: $\hat{\ell}_n$ close to ℓ_n

- Accumulation of quantization errors offset by contraction of mapping F .



$$\begin{aligned} |\ell_n - \hat{\ell}_n| &= |F(\ell_{n-1}, y_n) - Q_\nu(F(\hat{\ell}_{n-1}, y_n))| \\ &\leq \nu + |F(\ell_{n-1}, y_n) - F(\hat{\ell}_{n-1}, y_n)| \end{aligned}$$

Checking property 1: $\hat{\ell}_n$ close to ℓ_n

- Accumulation of quantization errors offset by contraction of mapping F .

-

$$\begin{aligned} |\ell_n - \hat{\ell}_n| &= |F(\ell_{n-1}, y_n) - Q_\nu(F(\hat{\ell}_{n-1}, y_n))| \\ &\leq \nu + |F(\ell_{n-1}, y_n) - F(\hat{\ell}_{n-1}, y_n)| \\ &\leq \nu + \eta |\ell_{n-1} - \hat{\ell}_{n-1}| \end{aligned}$$

where $\eta = \sup_{x \in \mathcal{S}} |f'(x)| < 1$.

Checking property 1: $\hat{\ell}_n$ close to ℓ_n

- Accumulation of quantization errors offset by contraction of mapping F .

-

$$\begin{aligned} |\ell_n - \hat{\ell}_n| &= |F(\ell_{n-1}, y_n) - Q_\nu(F(\hat{\ell}_{n-1}, y_n))| \\ &\leq \nu + |F(\ell_{n-1}, y_n) - F(\hat{\ell}_{n-1}, y_n)| \\ &\leq \nu + \eta |\ell_{n-1} - \hat{\ell}_{n-1}| \end{aligned}$$

where $\eta = \sup_{x \in \mathcal{S}} |f'(x)| < 1$.

- This and $|\ell_0 - \hat{\ell}_0| \leq \nu$ imply

$$|\ell_n - \hat{\ell}_n| \leq \nu / (1 - \eta)$$

- Property 1 established by taking $\nu = c\epsilon(1 - \eta)$.

Checking property 2: $\lim_{n \rightarrow \infty} Eg(\hat{\ell}_n(Y^n))$ is “easily” computed

- Although ℓ_n is a Markov process (Blackwell 1957), $\hat{\ell}_n$ is not.

Checking property 2: $\lim_{n \rightarrow \infty} Eg(\hat{\ell}_n(Y^n))$ is “easily” computed

- Although ℓ_n is a Markov process (Blackwell 1957), $\hat{\ell}_n$ is not.
- Key idea: Easier to find conditional distributions: $P(\hat{\ell}_n | X_n = 0)$ and $P(\hat{\ell}_n | X_n = 1)$.

Checking property 2: Joint Markov process

- Define finite valued joint Markov process (U_n, V_n) as

$$U_0 = V_0 = \log \frac{\pi_{01}}{\pi_{10}}$$

$$U_n = \begin{cases} Q_\nu(F(Z_n, U_{n-1})) & \text{with prob. } 1 - \pi_{01} \\ Q_\nu(F(Z_n, V_{n-1})) & \text{with prob. } \pi_{01} \end{cases}$$

$$V_n = \begin{cases} Q_\nu(F(1 - Z_n, V_{n-1})) & \text{with prob. } 1 - \pi_{10} \\ Q_\nu(F(1 - Z_n, U_{n-1})) & \text{with prob. } \pi_{10} \end{cases}$$

where Z_n is i.i.d. Bernoulli δ .

Checking property 2: Joint Markov process

- Define finite valued joint Markov process (U_n, V_n) as

$$U_0 = V_0 = \log \frac{\pi_{01}}{\pi_{10}}$$

$$U_n = \begin{cases} Q_\nu(F(Z_n, U_{n-1})) & \text{with prob. } 1 - \pi_{01} \\ Q_\nu(F(Z_n, V_{n-1})) & \text{with prob. } \pi_{01} \end{cases}$$

$$V_n = \begin{cases} Q_\nu(F(1 - Z_n, V_{n-1})) & \text{with prob. } 1 - \pi_{10} \\ Q_\nu(F(1 - Z_n, U_{n-1})) & \text{with prob. } \pi_{10} \end{cases}$$

where Z_n is i.i.d. Bernoulli δ .

- Lemma:

$$P(\hat{\ell}_n | X_n = 0) = P(U_n), \quad P(\hat{\ell}_n | X_n = 1) = P(V_n)$$

Checking property 2: Joint Markov process

- Define finite valued joint Markov process (U_n, V_n) as

$$U_0 = V_0 = \log \frac{\pi_{01}}{\pi_{10}}$$

$$U_n = \begin{cases} Q_\nu(F(Z_n, U_{n-1})) & \text{with prob. } 1 - \pi_{01} \\ Q_\nu(F(Z_n, V_{n-1})) & \text{with prob. } \pi_{01} \end{cases}$$

$$V_n = \begin{cases} Q_\nu(F(1 - Z_n, V_{n-1})) & \text{with prob. } 1 - \pi_{10} \\ Q_\nu(F(1 - Z_n, U_{n-1})) & \text{with prob. } \pi_{10} \end{cases}$$

where Z_n is i.i.d. Bernoulli δ .

- Lemma:

$$P(\hat{\ell}_n | X_n = 0) = P(U_n), \quad P(\hat{\ell}_n | X_n = 1) = P(V_n)$$

- $Eg(\hat{\ell}_n(Y^n)) = \lambda_0 Eg(U_n) + \lambda_1 Eg(V_n)$

Checking property 2: continued

- $\lim_{n \rightarrow \infty} \lambda_0 Eg(U_n) + \lambda_1 Eg(V_n) = \lambda_0 Eg(U) + \lambda_1 Eg(V)$ where U and V have the stationary (marginal) distributions of U_n and V_n (reachable from V_0 and U_0).

Checking property 2: continued

- $\lim_{n \rightarrow \infty} \lambda_0 Eg(U_n) + \lambda_1 Eg(V_n) = \lambda_0 Eg(U) + \lambda_1 Eg(V)$ where U and V have the stationary (marginal) distributions of U_n and V_n (reachable from V_0 and U_0).
- Stationary marginals can be obtained by solving $2|Q_\nu| \times 2|Q_\nu|$ system of equations: can be done in $O(|Q_\nu|^3)$ operations.

Checking property 2: continued

- $\lim_{n \rightarrow \infty} \lambda_0 Eg(U_n) + \lambda_1 Eg(V_n) = \lambda_0 Eg(U) + \lambda_1 Eg(V)$ where U and V have the stationary (marginal) distributions of U_n and V_n (reachable from V_0 and U_0).
- Stationary marginals can be obtained by solving $2|Q_\nu| \times 2|Q_\nu|$ system of equations: can be done in $O(|Q_\nu|^3)$ operations.
- Precision vs. complexity:
 - Recall we need $\nu = c\epsilon(1 - \eta) \rightarrow |Q_\nu| = O(\mu(\mathcal{S}) / (2c\epsilon(1 - \eta)))$.

Checking property 2: continued

- $\lim_{n \rightarrow \infty} \lambda_0 Eg(U_n) + \lambda_1 Eg(V_n) = \lambda_0 Eg(U) + \lambda_1 Eg(V)$ where U and V have the stationary (marginal) distributions of U_n and V_n (reachable from V_0 and U_0).
- Stationary marginals can be obtained by solving $2|Q_\nu| \times 2|Q_\nu|$ system of equations: can be done in $O(|Q_\nu|^3)$ operations.
- Precision vs. complexity:
 - Recall we need $\nu = c\epsilon(1 - \eta) \rightarrow |Q_\nu| = O(\mu(\mathcal{S}) / (2c\epsilon(1 - \eta)))$.
 - $N(\epsilon) = O(\epsilon^{-3})$

Observations about Method 3

- $N(\epsilon)$ is polynomial in ϵ^{-1} with degree 3, independent of process parameters (but constant factors may be arbitrarily large).

Observations about Method 3

- $N(\epsilon)$ is polynomial in ϵ^{-1} with degree 3, independent of process parameters (but constant factors may be arbitrarily large).
- Constant factors can be improved through more refined contraction analysis.

Observations about Method 3

- $N(\epsilon)$ is polynomial in ϵ^{-1} with degree 3, independent of process parameters (but constant factors may be arbitrarily large).
- Constant factors can be improved through more refined contraction analysis.
- Larger Markov chain state spaces ($|\mathcal{X}|$): problematic since analogue of ℓ_n is $|\mathcal{X}| - 1$ dimensional and requires $O(M^{|\mathcal{X}| - 1})$ quantization points to cover space well.

Observations about Method 3

- $N(\epsilon)$ is polynomial in ϵ^{-1} with degree 3, independent of process parameters (but constant factors may be arbitrarily large).
- Constant factors can be improved through more refined contraction analysis.
- Larger Markov chain state spaces ($|\mathcal{X}|$): problematic since analogue of ℓ_n is $|\mathcal{X}| - 1$ dimensional and requires $O(M^{|\mathcal{X}| - 1})$ quantization points to cover space well.
 - Birch vs. Quantization:
Birch may be better for large state space $|\mathcal{X}|$ and small observation space $|\mathcal{Y}|$.
Quantization other way around.

Method 4: Truncated Taylor expansion

- Han, Marcus (2005); Jacquet, et. al (2004); Zuk, et. al (2005)

Method 4: Truncated Taylor expansion

- Han, Marcus (2005); Jacquet, et. al (2004); Zuk, et. al (2005)
- If $\mathcal{H}(Y)$ is computable and analytic (c.f. Han, Marcus (2005)) at $\{\pi_{ij}\}, \{\delta_{ij}\}$:
 - Approximate $\mathcal{H}(Y)$ in a neighborhood of $\{\pi_{ij}\}, \{\delta_{ij}\}$ via truncated Taylor series.

Method 4: Truncated Taylor expansion

- Han, Marcus (2005); Jacquet, et. al (2004); Zuk, et. al (2005)
- If $\mathcal{H}(Y)$ is computable and analytic (c.f. Han, Marcus (2005)) at $\{\pi_{ij}\}, \{\delta_{ij}\}$:
 - Approximate $\mathcal{H}(Y)$ in a neighborhood of $\{\pi_{ij}\}, \{\delta_{ij}\}$ via truncated Taylor series.
 - Running binary example:
 - ◇ Analytic in δ at $\delta = 0$ for fixed $\pi_{01} > 0, \pi_{10} > 0$.
 - ◇ $\mathcal{H}(Y)$ at $\delta = 0$ is simply $\mathcal{H}(X)$.

Method 4: Truncated Taylor expansion

- Han, Marcus (2005); Jacquet, et. al (2004); Zuk, et. al (2005)
- If $\mathcal{H}(Y)$ is computable and analytic (c.f. Han, Marcus (2005)) at $\{\pi_{ij}\}, \{\delta_{ij}\}$:
 - Approximate $\mathcal{H}(Y)$ in a neighborhood of $\{\pi_{ij}\}, \{\delta_{ij}\}$ via truncated Taylor series.
 - Running binary example:
 - ◇ Analytic in δ at $\delta = 0$ for fixed $\pi_{01} > 0, \pi_{10} > 0$.
 - ◇ $\mathcal{H}(Y)$ at $\delta = 0$ is simply $\mathcal{H}(X)$.
- Precision/complexity:
 - Complexity of computing higher order derivatives?
 - Bounds on higher order derivatives over neighborhoods?

Open problems

- Full characterization of convergence of Birch bounds.

Open problems

- Full characterization of convergence of Birch bounds.
 - Universal exponential rate irrespective of mixing of Markov chain?

Open problems

- Full characterization of convergence of Birch bounds.
 - Universal exponential rate irrespective of mixing of Markov chain?
- Precision vs. complexity for variable length conditioning improvement of Birch (Method 2).

Open problems

- Full characterization of convergence of Birch bounds.
 - Universal exponential rate irrespective of mixing of Markov chain?
- Precision vs. complexity for variable length conditioning improvement of Birch (Method 2).
- Practical extension of quantized likelihood process of Method 3 to larger Markov chain state spaces?

Open problems

- Full characterization of convergence of Birch bounds.
 - Universal exponential rate irrespective of mixing of Markov chain?
- Precision vs. complexity for variable length conditioning improvement of Birch (Method 2).
- Practical extension of quantized likelihood process of Method 3 to larger Markov chain state spaces?
- Precision vs. complexity for truncated Taylor series of Method 4, when applicable.