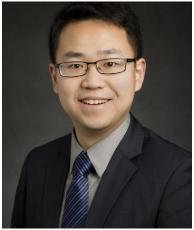
Next-symbol prediction without mixing: optimal rates, algorithms, and hardness

Yanjun Han (NYU Courant Math and CDS)

Joint work with:

Soham Jana
(Notre Dame)

Tianze Jiang
(Princeton)



Yihong Wu
(Yale)

Graduate Student/Postdoc Seminar, NYU Courant
February 28, 2025

A "ChatGPT-style" problem

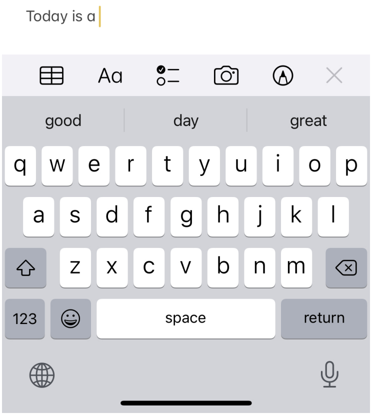
- Given data $X^n \equiv (X_1, \ldots, X_n)$, predict the next X_{n+1}.

A "ChatGPT-style" problem

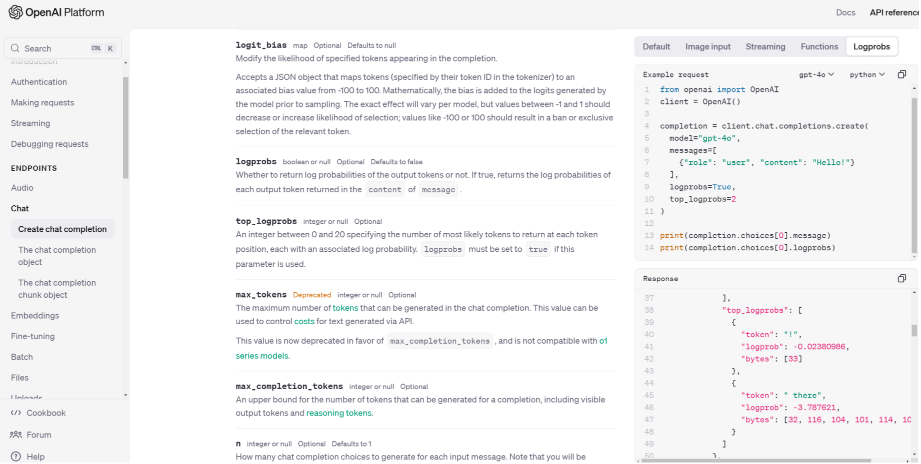
- Given data $X^n \equiv (X_1, \ldots, X_n)$, predict the next X_{n+1}.
- Allowing soft decisions, by prediction we meant estimating $P_{X_{n+1}|X^n}$

A "ChatGPT-style" problem

- Given data $X^n \equiv (X_1, \ldots, X_n)$, predict the next X_{n+1}.
- Allowing soft decisions, by prediction we meant estimating $P_{X_{n+1}|X^n}$
- Applications in NLP: autocomplete, text generation, LLM



Modern LLM



https://platform.openai.com/docs/api-reference/chat/create

Modeling dependent data

For these applications, iid model is clearly insufficient $\rightarrow$ Markov model [Shannon '48, '51]

III. The Series of Approximations to English

To give a visual idea of how this series of processes approaches a language, typical sequences in the approximations to English have been constructed and are given below. In all cases we have assumed a 27-symbol "alphabet," the 26 letters and a space.

1. Zero-order approximation (symbols independent and equiprobable).

   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. Second-order approximation (digram structure as in English).

   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order approximation (trigram structure as in English).

   IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

5. First-order word approximation. Rather than continue with tetragram, ..., n-gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.

   REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

6. Second-order word approximation. The word transition probabilities are correct but no further structure is included.

   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

The resemblance to ordinary English text increases quite noticeably at each of the above steps. Note that

Modeling dependent data

For these applications, iid model is clearly insufficient $\rightarrow$ Markov model [Shannon '48, '51]

III. THE SERIES OF APPROXIMATIONS TO ENGLISH

To give a visual idea of how this series of processes approaches a language, typical sequences in the approximations to English have been constructed and are given below. In all cases we have assumed a 27-symbol "alphabet," the 26 letters and a space.

1. Zero-order approximation (symbols independent and equiprobable).

   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. Second-order approximation (digram structure as in English).

   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order approximation (trigram structure as in English).

   IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

5. First-order word approximation. Rather than continue with tetragram, …, n-gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.

   REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

6. Second-order word approximation. The word transition probabilities are correct but no further structure is included.

   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

The resemblance to ordinary English text increases quite noticeably at each of the above steps. Note that

Challenges: (a) dependent data (b) large state space

Part I: Markov chains

Statistical inference for Markov chains

Parameter estimation:

- Transition matrix [Bartlett '51, Whittle '55, Anderson-Goodman '57, Billingsley '61, Wolfer-Kontorovich '19 ...]
- Properties
  - Order [Csiszár-Shields '00, van Handel '11]
  - Mixing time and spectral gap [Hsu et al '15, Levin-Peres '16]
  - Entropy rate [Kamath-Verdú '16, Han et al '18]
  - Property testing [Daskalakis et al '18, Cherapanamjeri-Bartlett '19 ...]
- Hidden Markov: [Douc-Moulines-Olsson-van Handel '11, Abraham-Naulet-Gassiat '21]

Parameter estimation:

- Transition matrix [Bartlett '51, Whittle '55, Anderson-Goodman '57, Billingsley '61, Wolfer-Kontorovich '19 ...]
- Properties
  - Order [Csiszár-Shields '00, van Handel '11]
  - Mixing time and spectral gap [Hsu et al '15, Levin-Peres '16]
  - Entropy rate [Kamath-Verdú '16, Han et al '18]
  - Property testing [Daskalakis et al '18, Cherapanamjeri-Bartlett '19 ...]
- Hidden Markov: [Douc-Moulines-Olsson-van Handel '11, Abraham-Naulet-Gassiat '21]

Prediction problem: a paradigm shift

- The quantity to be estimated (conditional distribution of the next state $P_{X_{n+1}|X^n}$) *depends on the sample path* itself; this is precisely why it is relevant for applications such as language models

Statistical inference for Markov chains

Parameter estimation:

- Transition matrix [Bartlett '51, Whittle '55, Anderson-Goodman '57, Billingsley '61, Wolfer-Kontorovich '19 ...]
- Properties
  - Order [Csiszár-Shields '00, van Handel '11]
  - Mixing time and spectral gap [Hsu et al '15, Levin-Peres '16]
  - Entropy rate [Kamath-Verdú '16, Han et al '18]
  - Property testing [Daskalakis et al '18, Cherapanamjeri-Bartlett '19 ...]
- Hidden Markov: [Douc-Moulines-Olsson-van Handel '11, Abraham-Naulet-Gassiat '21]

Prediction problem: a paradigm shift

- The quantity to be estimated (conditional distribution of the next state $P_{X_{n+1}|X^n}$) *depends on the sample path* itself; this is precisely why it is relevant for applications such as language models
- Estimation requires (strong) assumptions, prediction requires **none**

Prevailing assumptions

The chain mixes rapidly (large spectral gap) and stationary probabilities are not too small

Prevailing assumptions

The chain mixes rapidly (large spectral gap) and stationary probabilities are not too small

Both are necessary for estimation, but neither is needed for prediction

- If the chain moves at glacial speed, it is actually easy to predict
  - Observing aaaaaaaaaaaaa, predict a

Prevailing assumptions

The chain mixes rapidly (large spectral gap) and stationary probabilities are not too small

Both are necessary for estimation, but neither is needed for prediction

- If the chain moves at glacial speed, it is actually easy to predict
    - Observing aaaaaaaaaaaaa, predict a
- If a symbol is very rare, it is unlikely to appear next

Prevailing assumptions

The chain mixes rapidly (large spectral gap) and stationary probabilities are not too small

Both are necessary for estimation, but neither is needed for prediction

- If the chain moves at glacial speed, it is actually easy to predict
  - Observing aaaaaaaaaaaaa, predict a
- If a symbol is very rare, it is unlikely to appear next

Goal: understand optimal prediction of Markov chains in an *assumption-free* framework

- Challenge: lack of concentration results
- New idea: information-theoretic techniques

Example

Data: `abaaabbccaabcaba`

Example

Data: `abaaabbccaabcaba`?

Example

Data: `abaaabbccaabcaba`?

- Last symbol $X_{16} = \mathtt{a}$

Example

Data: `abaaabbccaabcaba`?

- Last symbol $X_{16} = \texttt{a}$
- Learn from historical examples: In the first 15 symbols,
  - `a` appeared 7 times

Example

Data: `abaaabbccaabcaba`?

- Last symbol $X_{16} =$ `a`
- Learn from historical examples: In the first 15 symbols,
  - `a` appeared 7 times
    - `aa`: 3 times

Example

Data: `abaaabbccaabcaba`?

- Last symbol $X_{16} = \texttt{a}$
- Learn from historical examples: In the first 15 symbols,
  - `a` appeared 7 times
    - `aa`: 3 times
    - `ab`: 4 times

Example

Data: abaaabbccaabcaba?

- Last symbol $X_{16} = \texttt{a}$
- Learn from historical examples: In the first 15 symbols,
  - a appeared 7 times
    - aa: 3 times
    - ab: 4 times
    - ac: 0 times

Example

Data: `abaaabbccaabcaba`?

- Last symbol $X_{16} = \texttt{a}$
- Learn from historical examples: In the first 15 symbols,
  - `a` appeared 7 times
    - `aa`: 3 times
    - `ab`: 4 times
    - `ac`: 0 times
- Predictor for X_{17}:

Example

Data: `abaaabbccaabcaba`?

- Last symbol $X_{16} = \mathtt{a}$
- Learn from historical examples: In the first 15 symbols,
  - `a` appeared 7 times
    - `aa`: 3 times
    - `ab`: 4 times
    - `ac`: 0 times
- Predictor for X_{17}:
  - Empirical transition frequencies $(\frac{3}{7}, \frac{4}{7}, \frac{0}{7})$

Example

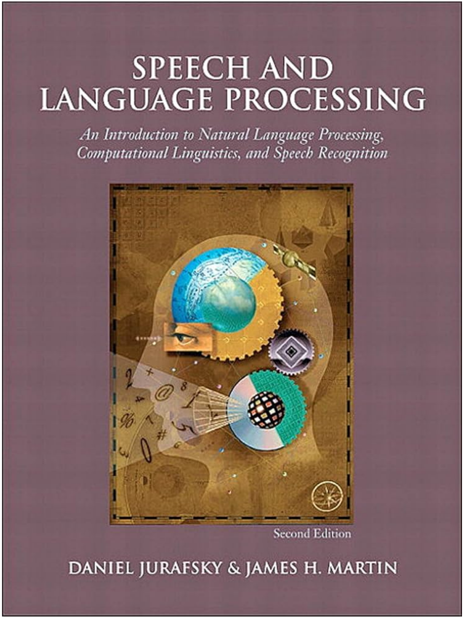
Data: `abaaabbccaabcaba`?

- Last symbol $X_{16} = \mathtt{a}$
- Learn from historical examples: In the first 15 symbols,
  - `a` appeared 7 times
    - `aa`: 3 times
    - `ab`: 4 times
    - `ac`: 0 times
- Predictor for X_{17}:
  - Empirical transition frequencies $(\frac{3}{7}, \frac{4}{7}, \frac{0}{7})$
  - Additively smoothed version:
    - Laplace's rule (add-1): $(\frac{4}{10}, \frac{5}{10}, \frac{1}{10})$

Example

Data: `abaaabbccaabcaba`?

- Last symbol $X_{16} = \texttt{a}$
- Learn from historical examples: In the first 15 symbols,
  - `a` appeared 7 times
    - `aa`: 3 times
    - `ab`: 4 times
    - `ac`: 0 times
- Predictor for X_{17}:
  - Empirical transition frequencies $(\frac{3}{7}, \frac{4}{7}, \frac{0}{7})$
  - Additively smoothed version:
    - Laplace's rule (add-1): $(\frac{4}{10}, \frac{5}{10}, \frac{1}{10})$
    - Krichevsky-Trofimov (add-$\frac{1}{2}$): $(\frac{7}{17}, \frac{9}{17}, \frac{1}{17})$


SPEECH AND
LANGUAGE PROCESSING
An Introduction to Natural Language Processing,
Computational Linguistics, and Speech Recognition
Second Edition
DANIEL JURAFSKY & JAMES H. MARTIN

Raw bigram counts:

	i	**want**	**to**	**eat**	**chinese**	**food**	**lunch**	**spend**
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

Figure 3.1 Bigram counts for eight of the words (out of $V = 1446$) in the Berkeley Restaurant Project corpus of 9332 sentences. Zero counts are in gray. Each cell shows the count of the column label word following the row label word. Thus the cell in row **i** and column **want** means that **want** followed **i** 827 times in the corpus.

Add-1 bigram counts:

	i	want	to	eat	chinese	food	lunch	spend
i	6	828	1	10	1	1	1	3
want	3	1	609	2	7	7	6	2
to	3	1	5	687	3	1	7	212
eat	1	1	3	1	17	3	43	1
chinese	2	1	1	1	1	83	2	1
food	16	1	16	1	2	5	1	1
lunch	3	1	1	1	1	2	1	1
spend	2	1	2	1	1	1	1	1

Figure 3.6 Add-one smoothed bigram counts for eight of the words (out of $V = 1446$) in the Berkeley Restaurant Project corpus of 9332 sentences. Previously-zero counts are in gray.

Smoothing in N-gram language models

Estimated bigram probabilities:

$$P_{\text{Laplace}}(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)+1}{\sum_w (C(w_{n-1}w)+1)} = \frac{C(w_{n-1}w_n)+1}{C(w_{n-1})+V} \qquad (3.27)$$

	i	**want**	**to**	**eat**	**chinese**	**food**	**lunch**	**spend**
i	0.0015	0.21	0.00025	0.0025	0.00025	0.00025	0.00025	0.00075
want	0.0013	0.00042	0.26	0.00084	0.0029	0.0029	0.0025	0.00084
to	0.00078	0.00026	0.0013	0.18	0.00078	0.00026	0.0018	0.055
eat	0.00046	0.00046	0.0014	0.00046	0.0078	0.0014	0.02	0.00046
chinese	0.0012	0.00062	0.00062	0.00062	0.00062	0.052	0.0012	0.00062
food	0.0063	0.00039	0.0063	0.00039	0.00079	0.002	0.00039	0.00039
lunch	0.0017	0.00056	0.00056	0.00056	0.00056	0.0011	0.00056	0.00056
spend	0.0012	0.00058	0.0012	0.00058	0.00058	0.00058	0.00058	0.00058

Figure 3.7 Add-one smoothed bigram probabilities for eight of the words (out of $V = 1446$) in the BeRP corpus of 9332 sentences. Previously-zero probabilities are in gray.

How to analyze these estimators without assumptions?

Wishful thinking (ignore smoothing for now):

Suppose the chain is stationary with stationary distribution $(\pi_{\texttt{a}}, \pi_{\texttt{b}}, \pi_{\texttt{c}})$ and transition matrix M.

- Number of occurrences of a: $N_{\texttt{a}} \approx n\pi_{\texttt{a}}$
- Number of occurrences of ab: $N_{\texttt{ab}} \approx n\pi_{\texttt{a}} M(\texttt{b}|\texttt{a})$
- So

$$\frac{N_{\texttt{ab}}}{N_{\texttt{a}}} \approx M(\texttt{b}|\texttt{a})$$

How to analyze these estimators without assumptions?

Wishful thinking (ignore smoothing for now):

Suppose the chain is stationary with stationary distribution $(\pi_{\mathtt{a}}, \pi_{\mathtt{b}}, \pi_{\mathtt{c}})$ and transition matrix M.

- Number of occurrences of a: $N_{\mathtt{a}} \approx n\pi_{\mathtt{a}}$
- Number of occurrences of ab: $N_{\mathtt{ab}} \approx n\pi_{\mathtt{a}} M(\mathtt{b}|\mathtt{a})$
- So

$$\frac{N_{\mathtt{ab}}}{N_{\mathtt{a}}} \approx M(\mathtt{b}|\mathtt{a})$$

- Let's attempt to analyze the denominator

Key difficulty

Suppose the chain is stationary with stationary distribution $(\pi_{\texttt{a}}, \pi_{\texttt{b}}, \pi_{\texttt{c}})$.

- Empirical frequency is unbiased: $\mathbb{E}[\hat{\pi}_{\texttt{a}}] = \mathbb{E}[\frac{N_{\texttt{a}}}{n}] = \pi_{\texttt{a}}$

Key difficulty

Suppose the chain is stationary with stationary distribution $(\pi_{\mathtt{a}}, \pi_{\mathtt{b}}, \pi_{\mathtt{c}})$.

- Empirical frequency is unbiased: $\mathbb{E}[\hat{\pi}_{\mathtt{a}}] = \mathbb{E}[\frac{N_{\mathtt{a}}}{n}] = \pi_{\mathtt{a}}$
- Concentration: [Lezaud '98, Pauline '15]

$$\mathrm{Var}(\hat{\pi}_{\mathtt{a}}) \lesssim \frac{1}{n \cdot \text{spectral gap}}$$

$$\mathbb{P}\left(|\hat{\pi}_{\mathtt{a}} - \pi_{\mathtt{a}}| > t\right) \le \exp\left(-\frac{cnt^2}{\pi_{\mathtt{a}} + t} \cdot \text{spectral gap}\right)$$

  This is tight in worst case; but spectral gap can be arbitrarily small
- So we need some new ideas other than applying concentration

Mathematical formulation

- Observe a single trajectory $X^n = (X_1, \ldots, X_n)$ of a random process taking values in a finite set $[k] \equiv \{1, \ldots, k\}$

Mathematical formulation

- Observe a single trajectory $X^n = (X_1, \ldots, X_n)$ of a random process taking values in a finite set $[k] \equiv \{1, \ldots, k\}$
- Consider Kullback-Leibler loss:

$$\mathrm{KL}(P\|Q) = \mathbb{E}_{X\sim P}\left[\log \frac{P}{Q}(X)\right] = \sum_{j=1}^{k} P(j) \log \frac{P(j)}{Q(j)}$$

Mathematical formulation

- Observe a single trajectory $X^n = (X_1, \dots, X_n)$ of a random process taking values in a finite set $[k] \equiv \{1, \dots, k\}$
- Consider Kullback-Leibler loss:

$$\mathrm{KL}(P\|Q) = \mathbb{E}_{X\sim P}\left[\log \frac{P}{Q}(X)\right] = \sum_{j=1}^{k} P(j)\log\frac{P(j)}{Q(j)}$$

- An estimate for $P_{X_{n+1}|X^n} \iff$ a conditional distribution $Q_{X_{n+1}|X^n}$

Mathematical formulation

- Observe a single trajectory $X^n = (X_1, \ldots, X_n)$ of a random process taking values in a finite set $[k] \equiv \{1, \ldots, k\}$
- Consider Kullback-Leibler loss:

$$\mathrm{KL}(P\|Q) = \mathbb{E}_{X\sim P}\left[\log \frac{P}{Q}(X)\right] = \sum_{j=1}^{k} P(j) \log \frac{P(j)}{Q(j)}$$

- An estimate for $P_{X_{n+1}|X^n} \iff$ a conditional distribution $Q_{X_{n+1}|X^n}$
- Average prediction risk:

$$\mathbb{E}[\mathrm{KL}(P_{X_{n+1}|X^n}\|Q_{X_{n+1}|X^n})]$$

Optimal (minimax) prediction risk

Model class $\mathcal{P}$ = collection of joint distributions of $(X_1, \ldots, X_{n+1})$

- iid data
- Markov model
- Hidden Markov model ...

the optimal prediction risk is:

$$\mathsf{Risk}_n \equiv \mathsf{Risk}_n(\mathcal{P}) \triangleq \inf_{Q_{X_{n+1}|X^n}} \sup_{P_{X^{n+1}} \in \mathcal{P}} \mathbb{E}_P[\mathsf{KL}(P_{X_{n+1}|X^n} \| Q_{X_{n+1}|X^n})]$$

Existing results: iid data

$X_1, X_2, \ldots \sim P$ on $[k]$: reduces to density estimation under KL loss

$$\mathrm{Risk}_n = \inf_Q \sup_P \mathbb{E}[\mathrm{KL}(P \| Q)]$$

Existing results: iid data

$X_1, X_2, \ldots \sim P$ on $[k]$: reduces to density estimation under KL loss

$$\mathrm{Risk}_n = \inf_Q \sup_P \mathbb{E}[\mathrm{KL}(P\|Q)]$$

Minimax rate is parametric:

$$\mathrm{Risk}_n \asymp \frac{k}{n}, \quad k \lesssim n$$

achieved by, e.g., add-one estimator (Laplace rule of succession)

$$Q(j) = \frac{N_j + 1}{n + k}, \quad N_j = \text{number of occurrences of } j$$

- Explicit computation with binomial: $\mathbb{E}[\mathrm{KL}(P\|Q)] \le \mathbb{E}[\chi^2(P\|Q)] \le \frac{k-1}{n+1}$

Existing results: iid data

$X_1, X_2, \ldots \sim P$ on $[k]$: reduces to density estimation under KL loss

$$\mathrm{Risk}_n = \inf_Q \sup_P \mathbb{E}[\mathrm{KL}(P\|Q)]$$

Minimax rate is parametric:

$$\mathrm{Risk}_n \asymp \frac{k}{n}, \quad k \lesssim n$$

achieved by, e.g., add-one estimator (Laplace rule of succession)

$$Q(j) = \frac{N_j + 1}{n + k}, \quad N_j = \text{number of occurrences of } j$$

- Explicit computation with binomial: $\mathbb{E}[\mathrm{KL}(P\|Q)] \le \mathbb{E}[\chi^2(P\|Q)] \le \frac{k-1}{n+1}$

Furthermore

- For fixed k: $\mathrm{Risk}_n = (1+o(1))\frac{k-1}{2n}$ [Braess et al '02]
- For $k \gg n$: $\mathrm{Risk}_n = (1+o(1))\log\frac{k}{n}$ [Paninski '04]

Existing results: Markov model

$X_1, X_2, \ldots$: stationary first-order Markov chain with k states

Optimal prediction risk: $\boxed{\mathsf{Risk}_{k,n} := \inf \sup \mathbb{E}[\mathsf{KL}(P_{X_{n+1}|X_n} \| Q_{X_{n+1}|X^n})]}$

Existing results: Markov model

$X_1, X_2, \ldots$: stationary first-order Markov chain with k states

Optimal prediction risk: $\boxed{\mathrm{Risk}_{k,n} := \inf \sup \mathbb{E}[\mathrm{KL}(P_{X_{n+1}|X_n} \| Q_{X_{n+1}|X^n})]}$

- Two states: [Falahatgar-Orlitsky-Pichapati-Suresh '16]

$$\mathrm{Risk}_{2,n} \asymp \frac{\log \log n}{n}$$

  - Slower than parametric (!)

Existing results: Markov model

$X_1, X_2, \ldots$: stationary first-order Markov chain with k states

Optimal prediction risk: $\boxed{\mathsf{Risk}_{k,n} := \inf \sup \mathbb{E}[\mathsf{KL}(P_{X_{n+1}|X_n} \| Q_{X_{n+1}|X^n})]}$

- Two states: [Falahatgar-Orlitsky-Pichapati-Suresh '16]

$$\mathsf{Risk}_{2,n} \asymp \frac{\log \log n}{n}$$

  - Slower than parametric (!)
- k states: [Hao-Orlitsky-Pichapati '18]

$$\mathsf{Risk}_{k,n} \gtrsim \frac{k \log \log n}{n}$$

Existing results: Markov model

$X_1, X_2, \ldots$: stationary first-order Markov chain with k states

Optimal prediction risk: $\boxed{\mathrm{Risk}_{k,n} := \inf \sup \mathbb{E}[\mathrm{KL}(P_{X_{n+1}|X_n} \| Q_{X_{n+1}|X^n})]}$

- Two states: [Falahatgar-Orlitsky-Pichapati-Suresh '16]

$$\mathrm{Risk}_{2,n} \asymp \frac{\log \log n}{n}$$

  - Slower than parametric (!)

- k states: [Hao-Orlitsky-Pichapati '18]

$$\mathrm{Risk}_{k,n} \gtrsim \frac{k \log \log n}{n}$$

Claimed $\mathrm{Risk}_{k,n} \lesssim \frac{k^2 \log \log n}{n}$, but implicitly assumed fast mixing

Main result

Theorem [H.-Jana-Wu '21]

For all $3 \le k \lesssim \sqrt{n}$,

$$\mathrm{Risk}_{k,n} \asymp \frac{k^2}{n} \log \frac{n}{k^2}$$

Remarks:

- Lower bound holds even for irreducible reversible chains
- Sample complexity (minimal sample size to achieve error ϵ) vs model complexity (number of parameters d)

$$n^*(d,\epsilon) \asymp \begin{cases} \frac{d}{\epsilon} & \text{iid} \\ \frac{d}{\epsilon} \log\log\frac{1}{\epsilon} & \text{Markov with 2 states} \\ \frac{d}{\epsilon} \log\frac{1}{\epsilon} & \text{Markov with } k \ge 3 \text{ states.} \end{cases}$$

- *Strict but only logarithmic* increase of sample complexity due to memory in the data

Higher-order Markov chains

- Optimal rate for mth-order Markov chains: $\frac{k^{m+1}}{n}\log\frac{n}{k^{m+1}}$ for $k \geq 2$
- The rate $\frac{\log\log n}{n}$ is highly special and only for binary 1st-order Markov chains

Next: only focus on 1st-order Markov chains.

An optimal estimator

Cesàro mean of add-1 estimators averaged over different sample sizes:

- Given trajectory $x^n = (x_1, \ldots, x_n)$, add-1 estimator for transition probability $M(j|i) \equiv P_{X_{n+1}|X_n}(j|i)$:

$$\hat{M}_{x^n}(j|i) \triangleq \frac{N_{ij} + 1}{N_i + k},$$

  where N_i = number of occurrences of i and N_{ij} = number of occurrences of consecutive ij

An optimal estimator

Cesàro mean of add-1 estimators averaged over different sample sizes:

- Given trajectory $x^n = (x_1, \ldots, x_n)$, add-1 estimator for transition probability $M(j|i) \equiv P_{X_{n+1}|X_n}(j|i)$:

$$\hat{M}_{x^n}(j|i) \triangleq \frac{N_{ij} + 1}{N_i + k},$$

  where N_i = number of occurrences of i and N_{ij} = number of occurrences of consecutive ij

- Final estimator:

$$Q(x_{n+1}|x^n) \triangleq \frac{1}{n} \sum_{t=1}^{n} \underbrace{\hat{M}_{x_{n-t+1}^n}(x_{n+1}|x_n)}_{\text{add-1 applied to most recent } t \text{ observations}}$$

An optimal estimator

Cesàro mean of add-1 estimators averaged over different sample sizes:

- Given trajectory $x^n = (x_1, \ldots, x_n)$, add-1 estimator for transition probability $M(j|i) \equiv P_{X_{n+1}|X_n}(j|i)$:

$$\hat{M}_{x^n}(j|i) \triangleq \frac{N_{ij} + 1}{N_i + k},$$

  where $N_i =$ number of occurrences of i and $N_{ij} =$ number of occurrences of consecutive ij
- Final estimator:

$$Q(x_{n+1}|x^n) \triangleq \frac{1}{n} \sum_{t=1}^{n} \underbrace{\hat{M}_{x_{n-t+1}^n}(x_{n+1}|x_n)}_{\text{add-1 applied to most recent } t \text{ observations}}$$

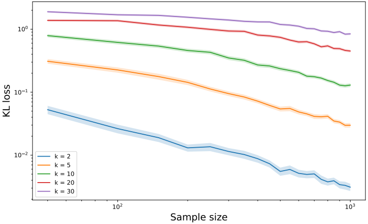
- Such Cesàro-mean-type strategy appeared before in density estimation literature [Yang-Barron '99]

Open question

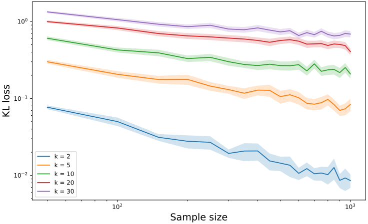
Open question: simple add-1 estimator with full data is optimal?

Open question: simple add-1 estimator with full data is optimal?

Numerical experiments suggest adaptivity to mixing time:



Large spectral gap $\gamma = 0.2$.



Small spectral gap $\gamma = 0.1$.

KL prediction loss: 95% confidence intervals over 500 independent trials.

Plan next

- Characterizing risk by redundancy
- Bounding redundancy
- Conclusions and discussions

Redundancy

Let $\mathcal{P}$ be a collection of joint distributions:

$$\boxed{\mathrm{Red}_n \triangleq \inf_{Q_{X^n}} \sup_{P_{X^n}\in\mathcal{P}} \mathrm{KL}(P_{X^n} \| Q_{X^n})}$$

- A key quantity in information theory (universal compression and prediction)
- Interpretation: best uniform approximation error of a class (not an estimation error!)
- Rule of thumb:

$$\mathrm{Red}_n \asymp \text{model complexity} \cdot \log n$$

- Redundancy-risk inequality:

$$\mathrm{Red}_n \le \sum_{m=1}^{n} \mathrm{Risk}_m$$

- We will show for Markov model: $\mathrm{Risk}_n \asymp \frac{\mathrm{Red}_n}{n}$.

Risk vs Redundancy: upper bound

$$\mathsf{Risk}_{k,n-1} \lesssim \frac{\mathsf{Red}_{k,n}}{n-1}$$

Risk vs Redundancy: upper bound

$$\mathsf{Risk}_{k,n-1} \lesssim \frac{\mathsf{Red}_{k,n}}{n-1}$$

Idea: "batch-to-online"

- Any joint distribution Q_{X^n} induces a Cesàro-mean style predictor:

$$\tilde{Q}_{X_n|X^{n-1}}(x_n|x^{n-1}) \triangleq \frac{1}{n-1}\sum_{t=2}^{n} Q_{X_t|X^{t-1}}(x_n|x_{n-t+1}^{n-1})$$

Risk vs Redundancy: upper bound

$$\mathsf{Risk}_{k,n-1} \lesssim \frac{\mathsf{Red}_{k,n}}{n-1}$$

Idea: "batch-to-online"

- Any joint distribution Q_{X^n} induces a Cesàro-mean style predictor:

$$\tilde{Q}_{X_n|X^{n-1}}(x_n|x^{n-1}) \triangleq \frac{1}{n-1}\sum_{t=2}^{n} Q_{X_t|X^{t-1}}(x_n|x_{n-t+1}^{n-1})$$

- Prediction risk:

$$\begin{aligned}
&\mathbb{E}[\mathrm{KL}(P_{X_n|X_{n-1}}\|\tilde{Q}_{X_n|X^{n-1}})] \\
\le{}& \frac{1}{n-1}\sum_{t=1}^{n}\mathbb{E}[\mathrm{KL}(P_{X_t|X^{t-1}}\|Q_{X^t|X^{t-1}})] && \text{convexity and stationarity} \\
={}& \frac{1}{n-1}\mathrm{KL}(P_{X^n}\|Q_{X^n}) && \text{chain rule}
\end{aligned}$$

Risk vs Redundancy: lower bound

$$\mathsf{Risk}_{k,n} \gtrsim \frac{1}{n}\mathsf{Red}^{\mathsf{sym}}_{k-1,n}$$

where

- $\mathsf{Red}^{\mathsf{sym}}_{k-1,n}$ = redundancy of Markov chain with $k-1$ states and symmetric transition matrix.
- We will show

$$\mathsf{Red}^{\mathsf{sym}}_{k-1,n} \asymp \underbrace{\text{model complexity}}_{\asymp k^2} \cdot \log n$$

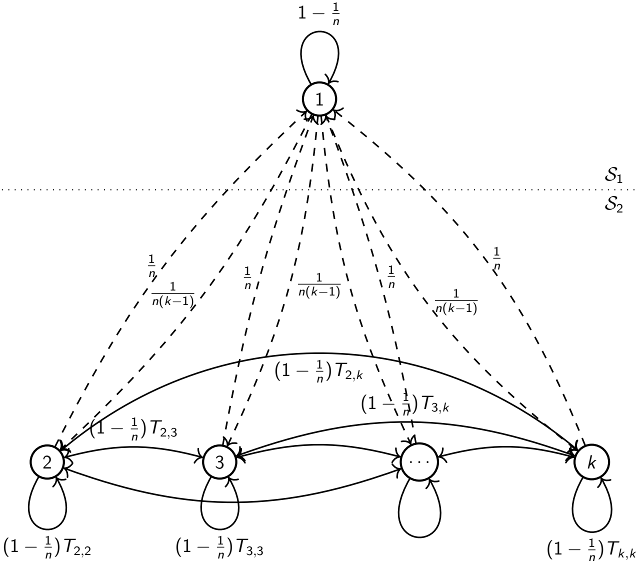
Sketch of reduction argument

Embed a $(k-1)$-state chain into a state space of size k:

$$M = \begin{bmatrix} 1-\frac{1}{n} & \frac{1}{n(k-1)} & \frac{1}{n(k-1)} & \cdots & \frac{1}{n(k-1)} \\ 1/n & & & & \\ 1/n & & & & \\ \vdots & & \left(1-\frac{1}{n}\right) T & & \\ 1/n & & & & \end{bmatrix}$$

Here T is a *symmetric* transition matrix for $k-1$ states to be optimized (randomized)

Sketch of reduction argument



Sketch of reduction argument

Stationary distribution $\pi = (\frac{1}{2}, \frac{1}{2(k-1)}, \cdots, \frac{1}{2(k-1)})$;

With constant probability, the chain starts from and stays at state 1 for a period of time, then enters $\mathcal{S}_2 = \{2, \ldots, k\}$ and never returns

Conditioned on this,

- Time t spent in $\mathcal{S}_2 \approx \text{Uniform}[n]$
- $(X_{n-t+1}, \ldots, X_n)$ is a Markov chain with $k-1$ states and transition matrix T

Sketch of reduction argument

Stationary distribution $\pi = (\frac{1}{2}, \frac{1}{2(k-1)}, \cdots, \frac{1}{2(k-1)})$;

With constant probability, the chain starts from and stays at state 1 for a period of time, then enters $\mathcal{S}_2 = \{2, \ldots, k\}$ and never returns

Conditioned on this,

- Time t spent in $\mathcal{S}_2 \approx \text{Uniform}[n]$
- $(X_{n-t+1}, \ldots, X_n)$ is a Markov chain with $k-1$ states and transition matrix T

$$\text{Overeall risk} \approx \underbrace{\frac{1}{n}\sum_{t=1}^{n} \text{Prediction risk for } T\text{-chain with sample size } t}_{\approx \text{Redundancy}}$$

Summary

For $k \geq 3$,

$$\frac{1}{n}\mathrm{Red}_{k,n}^{\mathrm{sym}} \lesssim \mathrm{Risk}_{k,n} \lesssim \frac{1}{n}\mathrm{Red}_{k,n}$$

- *Theoretical consequence*: it suffices to show both Red are $\Theta(k^2 \log n)$

Summary

For $k \geq 3$,

$$\frac{1}{n}\mathrm{Red}_{k,n}^{\mathrm{sym}} \lesssim \mathrm{Risk}_{k,n} \lesssim \frac{1}{n}\mathrm{Red}_{k,n}$$

- *Theoretical consequence*: it suffices to show both Red are $\Theta(k^2 \log n)$
- *Algorithmic consequence*: if $\mathrm{Red}_{k,n}$ is attained by a

$$Q_{X^n} = \prod_{t=1}^{n} Q_{X_t|X^{t-1}}$$

whose conditionals are fast to compute (sequential probability assignment), then we have an equally fast predictor

Summary

For $k \geq 3$,

$$\frac{1}{n}\mathsf{Red}^{\mathsf{sym}}_{k,n} \lesssim \mathsf{Risk}_{k,n} \lesssim \frac{1}{n}\mathsf{Red}_{k,n}$$

- *Theoretical consequence*: it suffices to show both Red are $\Theta(k^2 \log n)$
- *Algorithmic consequence*: if $\mathsf{Red}_{k,n}$ is attained by a

  $$Q_{X^n} = \prod_{t=1}^{n} Q_{X_t|X^{t-1}}$$

  whose conditionals are fast to compute (sequential probability assignment), then we have an equally fast predictor
- Bound redundancy from below: Bayesian argument and mutual information

Bounding redundancy

$$\mathrm{Red}_n = \inf_{Q_{X^n}} \sup_{P_{X^n} \in \mathcal{P}} \mathrm{KL}(P_{X^n} \| Q_{X^n})$$

Bounding redundancy

$$\mathrm{Red}_n = \inf_{Q_{X^n}} \sup_{P_{X^n} \in \mathcal{P}} \mathbb{E}_P \left[\log \frac{P_{X^n}}{Q_{X^n}}(X^n) \right]$$

Bounding redundancy

$$\mathrm{Red}_n = \inf_{Q_{X^n}} \sup_{P_{X^n} \in \mathcal{P}} \mathbb{E}_P \left[\log \frac{P_{X^n}}{Q_{X^n}}(X^n) \right]$$

Bounding redundancy

$$\begin{aligned}\mathrm{Red}_n &= \inf_{Q_{X^n}} \sup_{P_{X^n} \in \mathcal{P}} \mathbb{E}_P \left[\log \frac{P_{X^n}}{Q_{X^n}}(X^n) \right] \\ &\le \inf_{Q_{X^n}} \sup_{P_{X^n} \in \mathcal{P}} \max_{x^n} \log \frac{P_{X^n}}{Q_{X^n}}(x^n)\end{aligned}$$

Bounding redundancy

$$
\begin{aligned}
\mathrm{Red}_n &= \inf_{Q_{X^n}} \sup_{P_{X^n}\in\mathcal{P}} \mathbb{E}_P\left[\log \frac{P_{X^n}}{Q_{X^n}}(X^n)\right] \\
&\le \inf_{Q_{X^n}} \sup_{P_{X^n}\in\mathcal{P}} \max_{x^n} \log \frac{P_{X^n}}{Q_{X^n}}(x^n) \\
&= \log\left(\sum_{x^n} \max_{P_{X^n}\in\mathcal{P}} P_{X^n}(x^n)\right)
\end{aligned}
$$

Bounding redundancy

$$\begin{aligned}\mathrm{Red}_n &= \inf_{Q_{X^n}} \sup_{P_{X^n}\in\mathcal{P}} \mathbb{E}_P\left[\log \frac{P_{X^n}}{Q_{X^n}}(X^n)\right] \\ &\le \inf_{Q_{X^n}} \sup_{P_{X^n}\in\mathcal{P}} \max_{x^n} \log \frac{P_{X^n}}{Q_{X^n}}(x^n) \\ &= \log\left(\sum_{x^n} \max_{P_{X^n}\in\mathcal{P}} P_{X^n}(x^n)\right)\end{aligned}$$

attained by normalized maximum likelihood (NML) distribution [Shtarkov '87]

$$Q(x^n) \propto \max_{P_{X^n}\in\mathcal{P}} P_{X^n}(x^n) = \text{objective value of MLE}$$

Bounding redundancy

$$\begin{aligned}\mathrm{Red}_n &= \inf_{Q_{X^n}} \sup_{P_{X^n}\in\mathcal{P}} \mathbb{E}_P\left[\log \frac{P_{X^n}}{Q_{X^n}}(X^n)\right] \\ &\le \inf_{Q_{X^n}} \sup_{P_{X^n}\in\mathcal{P}} \max_{x^n} \log \frac{P_{X^n}}{Q_{X^n}}(x^n) \\ &= \log\left(\sum_{x^n} \max_{P_{X^n}\in\mathcal{P}} P_{X^n}(x^n)\right)\end{aligned}$$

attained by normalized maximum likelihood (NML) distribution [Shtarkov '87]

$$Q(x^n) \propto \max_{P_{X^n}\in\mathcal{P}} P_{X^n}(x^n) = \text{objective value of MLE}$$

- This quantity (Shtarkov sum) can be bounded by combinatorial methods, leading to $\mathrm{Red}_{k,n} \lesssim k^2 \log n$ for k-state Markov chains

Bounding redundancy

$$\begin{aligned}\mathrm{Red}_n &= \inf_{Q_{X^n}} \sup_{P_{X^n}\in\mathcal{P}} \mathbb{E}_P\left[\log \frac{P_{X^n}}{Q_{X^n}}(X^n)\right] \\ &\le \inf_{Q_{X^n}} \sup_{P_{X^n}\in\mathcal{P}} \max_{x^n} \log \frac{P_{X^n}}{Q_{X^n}}(x^n) \\ &= \log\left(\sum_{x^n} \max_{P_{X^n}\in\mathcal{P}} P_{X^n}(x^n)\right)\end{aligned}$$

attained by normalized maximum likelihood (NML) distribution [Shtarkov '87]

$$Q(x^n) \propto \max_{P_{X^n}\in\mathcal{P}} P_{X^n}(x^n) = \text{objective value of MLE}$$

- This quantity (Shtarkov sum) can be bounded by combinatorial methods, leading to $\mathrm{Red}_{k,n} \lesssim k^2 \log n$ for k-state Markov chains
- However, NML distribution is not sequentially defined through its conditionals

Bounding redundancy

For Markov chains, a simple sequential assignment is optimal up to constant factors [Davisson '83, Csiszár-Shields '04]

$$Q(x^n) = \frac{1}{k}\prod_{i=1}^{k}\frac{\prod_{j=1}^{k} N_{ij}!}{k\cdot(k+1)\cdot\cdots\cdot(N_i+k-1)}.$$

leading to add-1 estimators

$$Q(x_n|x^{n-1}) = \frac{N_{ij}+1}{N_i+k}$$

Bounding redundancy

For Markov chains, a simple sequential assignment is optimal up to constant factors
[Davisson '83, Csiszár-Shields '04]

$$Q(x^n) = \frac{1}{k}\prod_{i=1}^{k}\frac{\prod_{j=1}^{k} N_{ij}!}{k\cdot(k+1)\cdot\cdots\cdot(N_i+k-1)}.$$

leading to add-1 estimators

$$Q(x_n|x^{n-1}) = \frac{N_{ij}+1}{N_i+k}$$

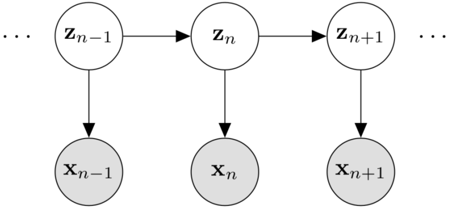
Comments:

- At the heart, replacing $\mathbb{E}$ by $\max_{x^n}$ is what allows risk bound *without mixing conditions*
- This information-theoretic technique departs from prevailing analysis based on concentration inequalities

Part II: Models with infinite memory

Hidden Markov Model (HMM)

HMM = Markov chain observed in iid noise

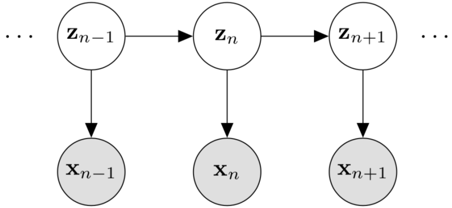


Parameters: Transition probabilities M and Emission probabilities T

- Latent states: $Z_t \xrightarrow{M} Z_{t+1}$; Observation: $Z_t \xrightarrow{T} X_t$
- Examples:
  - binary state binary observation (Gilbert-Elliot channel)
  - Gaussian emission (extension of Gaussian mixtures: iid states)

Hidden Markov Model (HMM)

HMM = Markov chain observed in iid noise



Parameters: Transition probabilities M and Emission probabilities T

- Latent states: $Z_t \xrightarrow{M} Z_{t+1}$; Observation: $Z_t \xrightarrow{T} X_t$
- Examples:
  - binary state binary observation (Gilbert-Elliot channel)
  - Gaussian emission (extension of Gaussian mixtures: iid states)

Long-range dependency: $X_{n+1} \not\perp X_1, \ldots, X_t | X_{t+1}, \ldots, X_n$

- Commonly used for modeling natural language and speech signals

Goal: $P_{X_{n+1}|X_1,\ldots,X_n}$

Main result

Theorem [H.-Jiang-Wu '24]

Consider HMM with $|\text{state space}| = k$ and $|\text{observation space}| = \ell$.

$$\text{Optimal prediction risk :} \quad \text{Risk}_n \asymp \frac{k\ell}{n} \log \frac{n}{k\ell} + \frac{k^2}{n} \log \frac{n}{k^2}$$

where

- the lower bound assumes sufficiently large n
- the upper bound is attained by an $n^{O(k^2+k\ell)}$-time dynamic programming algorithm

Remarks:

- Previous SOTA: $O(\frac{1}{\log n})$ based on Markov approximation [Sharan-Kakade-Liang-Valiant '18]
- Gaussian emissions in d dimensions: $\frac{k(k+d)\log n}{n}$, provided centers are in $[-1,1]^d$.
- Main idea: again redundancy

From finite to infinite memory

$\mathsf{Risk}_n \le \frac{\mathsf{Red}_n}{n}$ no longer holds. Instead,

$$\mathsf{Risk}_n \le \frac{\mathsf{Red}_n}{n} + \mathsf{Mem}_n$$

where Mem_n is a memory term (worst case over model class)

$$\frac{1}{n}\sum_{t=1}^{n} I(\underbrace{X_1, \dots, X_{n-t}}_{\text{past}}; \underbrace{X_{n+1}}_{\text{future}} \mid \underbrace{X_{n-t+1}, \dots, X_n}_{\text{recent}})$$

From finite to infinite memory

For HMM, one can show:

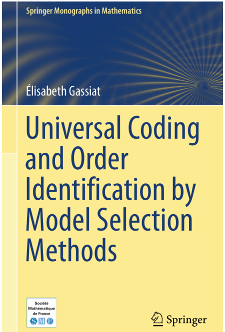
- Memory is weak: $\text{Mem}_n \le \frac{\log k}{n}$ [Birch '62]

Approximations for the Entropy for Functions of Markov Chains

John J. Birch

Ann. Math. Statist. 33(3): 930-938 (September, 1962). DOI: 10.1214/aoms/1177704462

- $\text{Red}_n \asymp \text{model complexity} \cdot \log n$ still holds [Gassiat '18]:
  - model complexity $\asymp k^2 + k\ell$ for discrete
  - model complexity $\asymp k^2 + kd$ for Gaussians



Algorithm

Joint state-observation likelihood:

$$P(x^{n+1}, z^{n+1}) = P(z^{n+1}) P(x^{n+1} | z^{n+1})$$

Probability assignment

$$Q(x^{n+1}, z^{n+1}) = \frac{1}{k} \prod_{t=1}^{n} M_t(z_{t+1}|z_t) \prod_{t=1}^{n} T_t(x_t|z_t)$$

where M_t and T_t are add-1 estimators for the transition and emission probabilities (applied to first $t-1$)

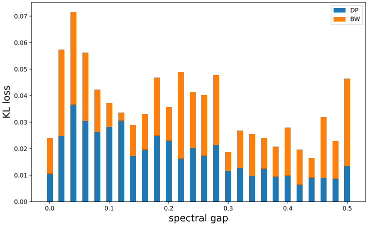
$$M_t(z'|z) = \frac{1 + \sum_{i=1}^{t-1} \mathbf{1}_{z_{i+1}=z' \text{ and } z_i=z}}{k + \sum_{i=1}^{t-1} \mathbf{1}_{z_i=z}}, \quad T_t(x|z) = \frac{1 + \sum_{i=1}^{t-1} \mathbf{1}_{z_i=z \text{ and } x_i=x}}{l + \sum_{i=1}^{t-1} \mathbf{1}_{z_i=z}}.$$

Marginalize out state sequences:

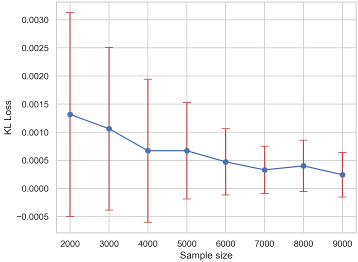
$$Q(x^{n+1}) = \sum_{z^{n+1}} Q(x^{n+1}, z^{n+1})$$

As before averaging its conditionals yields an optimal predictor

Experiment



KL loss vs spectral gap for DP and Baum-Welch ($n = 50$)



KL loss vs n for Baum-Welch.

HMM with binary symmetric Markov chain and emission.

- Baum-Welch: EM algorithm for HMM
- Question: Does Baum-Welch work for prediction without conditions assumed for parameter estimation [Yang-Balakrishnan-Wainwright '17]

Hardness

For large state space:

- Learning HMM is harder than certain hard problems such as Learning Parity in Noise [Mossel-Roch '06] and CSPs [Sharan-Kakade-Liang-Valiant '18]
- Prediction is also hard [H.-Jiang-Wu '24]: $k \geq \mathrm{polylog}(n)$, achieving optimal prediction is computationally hard based on these assumptions

Renewal Process: a Puzzle

Renewal process

Suppose for a given driver the time (in months) between consecutive traffic accidents are iid with finite mean. Observe the driving record (0 for safety or 1 for accident) for the past n months:

$$X^n = 0000\textcolor{red}{1}000000000000000\textcolor{red}{1}000\textcolor{red}{1}000000000\textcolor{red}{1}00000000\textcolor{red}{1}$$

Goal: Predict next month by estimating $P(X_{n+1} = 1 | X^n)$

This model class is

- Nonparametric: parametrized by interarrival time distribution
- Infinite memory: can be recast as an HMM with state space $\mathbb{N}$

Renewal process

Suppose for a given driver the time (in months) between consecutive traffic accidents are iid with finite mean. Observe the driving record (0 for safety or 1 for accident) for the past n months:

$$X^n = 00001000000000000000100010000000000100000000 1$$

Goal: Predict next month by estimating $P(X_{n+1} = 1 | X^n)$

This model class is

- Nonparametric: parametrized by interarrival time distribution
- Infinite memory: can be recast as an HMM with state space $\mathbb{N}$

Optimal prediction error [H.-Jiang-Wu '24]: $\Theta(n^{-\frac{1}{2}})$

- Based on $Red_n = \Theta(\sqrt{n})$ for renewal processes [Csiszár-Shields '96]
- Open problem: What's a simple algorithm?

Concluding remarks

Main result: Prediction risk via Redundancy

- Theoretical consequence: $\mathrm{Risk}_n \asymp \frac{\mathrm{Red}_n}{n}$ determines optimal prediction rate without mixing conditions
- Algorithmic consequence: sequential probability assignment $\Longrightarrow$ computationally efficient prediction algorithm

Concluding remarks

Many open problems

- Stationarity: Needed for reduction to Red, not for bounding Red
- How fast does the chain need to mix?
  - Spectral gap $\gtrsim \frac{(\log n)^2}{k} \implies$ Risk $\lesssim \frac{k^2}{n}$
- Practical prediction algorithm (Laplace smoothing or Baum–Welch?)

References

- Y. Han, S. Jana, and Y. Wu, *Optimal prediction of Markov chains with and without spectral gap*, NeuRIPS 2021 (Transactions on IT 2023), arxiv:2106.13947.
- Y. Han, T. Jiang, and Y. Wu, *Prediction from compression for models with infinite memory, with applications to hidden Markov and renewal processes*, COLT 2024, arxiv:2404.15454