Lec 11 : Universal compression and redundancy

Yanjun Han

Recall from Lec 1: If $X^n \sim P$, then $\exists$ a uniquely decodable code $f: \mathcal{X}^n \to \{0,1\}^*$ s.t.

$$\mathbb{E}[\ell(f(X^n))] < H(P) + 1 \text{ bit}$$

- Examples: Shannon / Huffman / Arithmetic codes
- Issue: All these codes require the perfect knowledge of P

Question: Is there a universal code that:

① is uniquely decodable;

② does not require the knowledge of P;

③ achieve an expected codelength close to $H(P)$ for every P in a given class $\mathcal{P}$?

A motivating example: i.i.d. Bernoulli. Suppose $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(p)$, with unknown $p \in [0,1]$.

The Shannon limit is $nH(\mathrm{Bern}(p)) = n\left(p \log \frac{1}{p} + (1-p)\log\frac{1}{1-p}\right)$.

A simple universal code:

- compute $n_1 = \sum_{i=1}^n \mathbb{1}(X_i = 1)$, the number of 1's;
- express n_1 using $\log(n+1)$ bits; (as $n_1 \in \{0, 1, \cdots, n\}$)
- condition on n_1, there are $\binom{n}{n_1}$ possibilities for $(X_1, \cdots, X_n)$, so we can encode the final sequence using $\log\binom{n}{n_1}$ bits.

Clearly this is uniquely decodable: the decoder first decodes n_1 from the $\log(n+1)$ bits, and then decodes $(X_1, \cdots, X_n)$ from n_1 and the last $\log\binom{n}{n_1}$ bits.

If $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(p)$, the expected codelength is

$$\begin{aligned}
\mathbb{E}[\ell(f(X^n))] &= \log(n+1) + \mathbb{E}\left[\log\binom{n}{n_1}\right] \\
&\le \log(n+1) + n\mathbb{E}\left[H\left(\mathrm{Bern}\left(\tfrac{n_1}{n}\right)\right)\right] \quad \left(\text{Using } H(X^n) \le \textstyle\sum_i H(X_i)\right) \\
&\le \log(n+1) + nH\left(\mathbb{E}\,\mathrm{Bern}\left(\tfrac{n_1}{n}\right)\right) \quad (P \mapsto H(P) \text{ is concave}) \\
&= \log(n+1) + nH(\mathrm{Bern}(p)).
\end{aligned}$$

In other words, compared with the Shannon limit with the knowledge of p, there is a universal code with an extra overhead of only $O(\log n)$ bits!

<u>General scenario</u>: $X^n \sim P$, where $P \in \mathcal{P}$ is an <u>unknown</u> source distribution.

- Shannon limit with the knowledge of P: $H(P)$
- Universal code: by Kraft inequality, any uniquely decodable code f can be equivalently represented by a probability distribution Q via

$$Q(x^n) = 2^{-\ell(f(x^n))}.$$

Therefore, the expected codelength of the code (represented by Q) under $X^n \sim P$ is

$$\mathbb{E}_P[\ell(f(X^n))] = \mathbb{E}_P[\log \frac{1}{Q(X^n)}].$$

- The "overhead" of a universal code is

$$\mathbb{E}_P[\log \frac{1}{Q(X^n)}] - H(P) = \mathbb{E}_P[\log \frac{P(X^n)}{Q(X^n)}] = D_{KL}(P \| Q).$$

<u>Defn (minimax redundancy)</u> The minimax redundancy of a distribution class $\mathcal{P}$ on X^n is defined as

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

- Q_{X^n} corresponds to a universal code that can achieve an overhead at most $\mathrm{Red}(\mathcal{P})$ bits with respect to every $P_{X^n} \in \mathcal{P}$
- In most cases, $\mathrm{Red}(\mathcal{P}) = o(n)$, and even $\mathrm{Red}(\mathcal{P}) = O(\log n)$.

<u>Bernoulli example continued</u>. How to find good Q_{X^n} when $\mathcal{P} = \{\mathrm{Bern}(p)^{\otimes n} : p \in [0,1]\}$?

Try the following conditional distributions:

- Q_{X_1}: $\mathrm{Unif}(\{0,1\})$
- $Q_{X_{t+1}|X^t}$: let $n_1(x^t)$, $n_0(x^t)$ be the number of 1's and 0's in x^t, respectively, set

$$Q_{X_{t+1}|X^t}(1) = \frac{n_1(x^t)+1}{t+2}, \quad Q_{X_{t+1}|X^t}(0) = \frac{n_0(x^t)+1}{t+2}.$$

This is called the <u>add-1 estimator</u> or the <u>Laplace estimator</u>.

Then $$Q_{X^n}(x^n) = \prod_{t=0}^{n-1} Q_{X_{t+1}|X^t}(x_n|x^{t-1}) = \frac{(1\cdot 2 \cdots n_1(x^n))(1\cdot 2 \cdots n_0(x^n))}{2\cdot 3 \cdot \cdots \cdot (n+1)} = \frac{n_1(x^n)!\, n_0(x^n)!}{(n+1)!}.$$

On the other hand, for any $p \in [0,1]$,

$$P_{X^n}(x^n) = p^{n_1(x^n)}(1-p)^{n_0(x^n)} \leq \left(\frac{n_1(x^n)}{n}\right)^{n_1(x^n)}\left(\frac{n_0(x^n)}{n}\right)^{n_0(x^n)}.$$

$$\Longrightarrow \frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)} = (n+1)\cdot \frac{\frac{n_1(x^n)^{n_1(x^n)}}{n_1(x^n)!}\frac{n_0(x^n)^{n_0(x^n)}}{n_0(x^n)!}}{\frac{n^n}{n!}} = O(n) \quad \text{(by Stirling)}$$

This means that

$$\mathrm{Red}(Q_{X^n};\mathcal{P}) = \sup_{P_{X^n}\in\mathcal{P}} D_{KL}(P_{X^n}\|Q_{X^n}) = \sup_{P_{X^n}} \mathbb{E}_{P_{X^n}}\left[\log\frac{P_{X^n}}{Q_{X^n}}\right] \leq \log n + O(1).$$

<u>Bernoulli example continued, again</u>. Now let's use

$$Q_{X_{t+1}|X^t}(1) = \frac{n_1(x^t)+\frac{1}{2}}{t+1}, \quad Q_{X_{t+1}|X^t}(0) = \frac{n_0(x^t)+\frac{1}{2}}{t+1}$$

This is called the <u>add-$\frac{1}{2}$ estimator</u>, or <u>Krichevsky-Trofimov estimator</u>.

In this case,

$$\begin{aligned} Q_{X^n}(x^n) &= \prod_{t=0}^{n-1} Q_{X_{t+1}|X^t}(x_{t+1}|x^t) \\ &= \frac{1}{n!}\left(\frac{1}{2}\cdot\frac{3}{2}\cdots\left(n_1(x^n)-\frac{1}{2}\right)\right)\left(\frac{1}{2}\cdot\frac{3}{2}\cdots\left(n_0(x^n)-\frac{1}{2}\right)\right) \\ &= \frac{(2n_1(x^n)-1)!!\,(2n_0(x^n)-1)!!}{2^n\, n!}, \end{aligned}$$

and $$\frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)} \leq \frac{2^n n! \cdot \frac{n_1(x^n)^{n_1(x^n)}\, n_0(x^n)^{n_0(x^n)}}{n^n}}{(2n_1(x^n)-1)!!\,(2n_0(x^n)-1)!!} = O(\sqrt{n}). \quad \text{(Stirling)}$$

Therefore, $$\mathrm{Red}(Q_{X^n};\mathcal{P}) \leq \log O(\sqrt{n}) = \frac{1}{2}\log n + O(1).$$

This constant $\frac{1}{2}$ turns out to be tight: $\mathrm{Red}(\mathcal{P}) = \frac{1}{2}\log n + \Theta(1)$.

In the previous examples, what we're using is actually $\mathrm{Red}(\mathcal{P}) \le R^*(\mathcal{P})$.

Defn (worst-case/pointwise redundancy)

$$R^*(\mathcal{P}) = \inf_{Q_{X^n}} \sup_{P_{X^n}\in\mathcal{P}} \sup_{x^n} \log \frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)}.$$

- It's clear that $\mathrm{Red}(\mathcal{P}) \le R^*(\mathcal{P})$.
- $R^*(\mathcal{P})$ treats x^n as an individual sequence, instead of drawing from a probability distribution.
- In online learning, $R^*(\mathcal{P})$ is also the minimax regret under the log loss $\ell_{\log}(p,x) = \log\frac{1}{p(x)}$

$$R^*(\mathcal{P}) = \inf_{q^n} \sup_{x^n} \left(\sum_{t=1}^n \ell_{\log}(q_t(\cdot|x^{t-1}), x_t) - \inf_{p^n\in\mathcal{P}} \sum_{t=1}^n \ell_{\log}(p_t(\cdot|x^{t-1}), x_t)\right).$$

Unlike $\mathrm{Red}(\mathcal{P})$ which can be hard to characterize, $R^*(\mathcal{P})$ has a combinatorial characterization.

Thm.

$$R^*(\mathcal{P}) = \log\Big(\underbrace{\sum_{x^n} \sup_{P_{X^n}\in\mathcal{P}} P_{X^n}(x^n)}_{\text{"Shtarkov sum"}} \Big).$$

Pf. (Upper bound) Let $Z = \sum_{x^n} \sup_{P_{X^n}\in\mathcal{P}} P_{X^n}(x^n)$, and

$$Q^*_{X^n}(x^n) = \frac{1}{Z} \sup_{P_{X^n}\in\mathcal{P}} P_{X^n}(x^n). \quad \text{(normalized maximum likelihood distribution)}$$

Then

$$\sup_{P_{X^n}\in\mathcal{P}} \sup_{x^n} \log \frac{P_{X^n}(x^n)}{Q^*_{X^n}(x^n)} = \log Z \implies R^*(\mathcal{P}) \le \log Z.$$

(Lower bound) For any Q_{X^n},

$$\sup_{P_{X^n}\in\mathcal{P}} \sup_{x^n} \log\frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)} = \sup_{P_{X^n}\in\mathcal{P}} \sup_{x^n} \left(\log\frac{P_{X^n}(x^n)}{Q^*_{X^n}(x^n)} + \log\frac{Q^*_{X^n}(x^n)}{Q_{X^n}(x^n)}\right)$$

$$= \log Z + \underbrace{\sup_{x^n} \log\frac{Q^*_{X^n}(x^n)}{Q_{X^n}(x^n)}}_{\ge \sum_{x^n} Q^*_{X^n}(x^n)\log\frac{Q^*_{X^n}(x^n)}{Q_{X^n}(x^n)} = D_{KL}(Q^*_{X^n}\|Q_{X^n}) \ge 0.} \ge \log Z. \qquad \square$$

This combinatorial nature of $R^*(\mathcal{P})$ makes it easy to upper bound $\mathrm{Red}(\mathcal{P})$ for non-i.i.d. families $\mathcal{P}$.

Example (Markov chain) Let $\mathcal{P} = \{ P_{X^n} = p(x_1) \prod_{t=1}^{n-1} M(x_{t+1} | x_t) \}$ be the class of all time-homogeneous (first-order) Markov chains on state space $[k]$.

Claim:
$$\mathrm{Red}(\mathcal{P}) \le \frac{k(k-1)}{2} \log n + O_k(1).$$

Pf. Apply add-$\frac{1}{2}$ estimator to all transition probabilities:
$$Q_{X_{t+1}|X^t}(i) = \frac{n_{j\to i}(x^t) + \frac{1}{2}}{n_j(x^t) + \frac{k}{2}} \quad \text{if } x_t = j.$$

where
$$n_{j\to i}(x^t) = \sum_{s=1}^{t-1} \mathbb{1}(x_s = j, x_{s+1} = i),$$
$$n_j(x^t) = \sum_{s=1}^{t-1} \mathbb{1}(x_s = j).$$

Then for any $x^n \in [k]^n$,
$$\frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)} = \frac{p(x_1)}{1/k} \cdot \prod_{j=1}^{k} \underbrace{\prod_{\substack{t\in[n-1]:\\ x_t = j}} \frac{M(x_{t+1}|j)}{Q_{X_{t+1}|X^t}(x_{t+1}|x^t)}}_{= O(\sqrt{n})^{k-1} \text{ by the analysis in the i.i.d. model}}$$

$$\le k \cdot (C\sqrt{n})^{k(k-1)}$$

$$\Rightarrow \mathrm{Red}(\mathcal{P}) \le R^*(\mathcal{P}) \le \log\left(k \cdot (C\sqrt{n})^{k(k-1)}\right) = \frac{k(k-1)}{2}\log n + O(k^2). \quad \boxed{\square}$$

The same programs can be extended to other processes such as the hidden Markov models; we refer to the book [Gassiat, 2018] for more examples.

Redundancy bounds for i.i.d. families.

Entropic upper bound. By the global Fano proof in Lec 9, we have

> **Thm.** $\mathrm{Red}(\mathcal{P}^{\otimes n}) \le \inf_{\varepsilon>0} \left(n\varepsilon^2 + \log N_{KL}(\mathcal{P}, \varepsilon) \right)$.

Example. When $\mathcal{P} = (P_\theta)_{\theta\in\mathbb{R}^d}$ is a parametric family with d parameters, usually $\log N_{KL}(\mathcal{P}, \varepsilon) \sim d \log \frac{1}{\varepsilon}$. Choosing $\varepsilon \sim \sqrt{\frac{d}{n}}$ gives the upper bound
$\mathrm{Red}(\mathcal{P}^{\otimes n}) \le \frac{d}{2} \log \frac{n}{d} + O(d)$.

Lower bound by Rissanen. We begin with a variational representation of $\mathrm{Red}(\mathcal{P})$.

> **Redundancy-capacity Thm.** For $\mathcal{P} = (P_\theta)_{\theta\in\Theta}$,
> $$\mathrm{Red}(\mathcal{P}) = \sup_{\rho\in\Delta(\Theta)} I(\theta; X) \quad \text{where } \theta\sim\rho,\ X|\theta\sim P_\theta.$$

(The quantity $\sup_\rho I(\theta; X)$ is the capacity of the channel $\mathcal{P} = (P_\theta)_{\theta\in\Theta}$.)

Pf. The "golden formula" for mutual info (Lec 7) says

$$I(\theta; X) = \inf_{Q_X} \mathbb{E}_{\theta\sim\rho}[D_{KL}(P_\theta \| Q_X)].$$

Then

$$\sup_\rho I(\theta; X) = \sup_\rho \inf_{Q_X} \mathbb{E}_{\theta\sim\rho}[D_{KL}(P_\theta \| Q_X)]$$

$$= \inf_{Q_X} \sup_\rho \mathbb{E}_{\theta\sim\rho}[D_{KL}(P_\theta \| Q_X)] \quad \text{(minimax thm.)}$$

$$= \inf_{Q_X} \sup_\theta D_{KL}(P_\theta \| Q_X) = \mathrm{Red}(\mathcal{P}). \qquad \square$$

Implication: to lower bound $\mathrm{Red}(\mathcal{P})$, can find a proper prior distribution ρ such that $I(\theta; X)$ is large when $\theta\sim\rho$.

Rissanen's program: Find an estimator $\hat{\theta}(X^n)$ s.t. $\sup_{\theta\in\Theta} \mathbb{E}_\theta[\|\theta-\hat{\theta}(X^n)\|^2] \le \varepsilon_n^2$.

Thm. Let $\Theta \subseteq \mathbb{R}^d$ have a non-empty interior. Then

$$\mathrm{Red}(\mathcal{P}^{\otimes n}) \ge \log \mathrm{Vol}_d(\Theta) - \frac{d}{2}\log\left(\frac{2\pi e\,\varepsilon_n^2}{d}\right).$$

Pf. Let $\theta \sim \rho = \mathrm{Unif}(\Theta)$, and $h(\cdot)$ denote the differential entropy on $\mathbb{R}^d$.

Then

$$\begin{aligned} I(\theta; X^n) &= h(\theta) - h(\theta \mid X^n) \\ &= \log \mathrm{Vol}_d(\Theta) - h(\theta\mid X^n), \end{aligned}$$

and

$$\begin{aligned} h(\theta\mid X^n) &= h(\theta - \hat{\theta}(X^n) \mid X^n) \\ &\le h(\theta - \hat{\theta}(X^n)) \qquad \text{(conditioning reduces entropy)} \\ &\le \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det\left(\mathbb{E}[(\theta-\hat{\theta}(X^n))(\theta-\hat{\theta}(X^n))^T]\right) \\ &\qquad \text{(maximum entropy principle)} \\ &\le \frac{d}{2}\log(2\pi e) + \frac{d}{2}\log\left(\frac{\mathbb{E}\|\theta-\hat{\theta}(X^n)\|^2}{d}\right) \qquad \left(\det(A) = \prod_i \lambda_i \le \left(\frac{\sum_i \lambda_i}{d}\right)^d = \left(\frac{\mathrm{Tr}(A)}{d}\right)^d \text{ for PSD } A\right) \\ &= \frac{d}{2}\log\left(\frac{2\pi e\,\varepsilon_n^2}{d}\right). \end{aligned}$$

$\boxed{\checkmark}$

Example. In parametric families, usually $\mathrm{Vol}_d(\Theta) = \Omega\left(\frac{1}{d}\right)^{d/2}$ and $\varepsilon_n^2 = O\left(\frac{d}{n}\right)$. Therefore, Rissanen's lower bound gives $\mathrm{Red}(\mathcal{P}^{\otimes n}) \ge \frac{d}{2}\log\frac{n}{d} - O(d)$.

Lower bounds by Haussler & Opper

The argument of Haussler & Opper (1997) chooses ρ to be a uniform mixture

$$\rho = \frac{1}{M}\sum_{i=1}^M \delta_{\theta_i}.$$

Lemma. For $X \mid \theta \sim P_\theta$ and $0 < \lambda \le 1$:

$$I(\theta; X) \ge -\mathbb{E}_{\theta, X}\left[\log \mathbb{E}_{\theta'}\left(\frac{P_{\theta'}(X)}{P_\theta(X)}\right)^\lambda\right],$$

where θ' is an independent copy of θ. $(\theta' \perp\!\!\!\perp (\theta, X))$

Pf. Let $f(\lambda) = \text{RHS}$, then $f(1) = I(\theta; X)$. Since CGF is convex, f is concave.

Finally,

$$f'(\lambda) = \mathbb{E}_{\theta, X}[\log p_\theta(X)] - \underbrace{\mathbb{E}_{\theta, X} \frac{\mathbb{E}_{\theta'}[p_{\theta'}(X)^\lambda \log p_{\theta'}(X)]}{\mathbb{E}_{\theta'}[p_{\theta'}(X)^\lambda]}}.$$

when $\lambda = 1$: $(\cdot) = \int_\theta \int_x p(\theta|x) \int_{\theta'} p(\theta') p_{\theta'}(x) \log p_{\theta'}(x) \, d\theta \, d\theta' \, dx$

$= \int_x \int_{\theta'} p(\theta') p_{\theta'}(x) \log p_{\theta'}(x)$

$= \mathbb{E}_{\theta, X}[\log p_\theta(X)]$.

Therefore, $f'(1) = 0 \Rightarrow f'(\lambda) \ge 0$ by concavity of f

$\Rightarrow f(\lambda) \le f(1) = I(\theta; X)$ $\boxed{\Sigma}$

Thm. $\mathrm{Red}(\mathcal{P}^{\otimes n}) \ge \sup_{\varepsilon > 0} \min\left\{\frac{n\varepsilon^2}{2}, \log M_H(\mathcal{P}, \varepsilon)\right\} - \log 2.$

Pf. Let $P_{\theta_1}, \cdots, P_{\theta_M}$ be an ε-packing of $\mathcal{P}$ under Hellinger, and $\rho = \frac{1}{M}\sum_{i=1}^M \delta_{\theta_i}$.

Then for $\theta \sim \rho$,

$$I(\theta; X^n) \ge -\frac{1}{M}\sum_{i=1}^M \mathbb{E}_{P_{\theta_i}^{\otimes n}}\left[\log\left(\frac{1}{M}\sum_{j=1}^M \sqrt{\frac{P_{\theta_j}^{\otimes n}(X^n)}{P_{\theta_i}^{\otimes n}(X^n)}}\right)\right]$$

$$\ge -\frac{1}{M}\sum_{i=1}^M \log\left(\frac{1}{M}\sum_{j=1}^M \mathbb{E}_{P_{\theta_i}^{\otimes n}} \sqrt{\frac{P_{\theta_j}^{\otimes n}(X^n)}{P_{\theta_i}^{\otimes n}(X^n)}}\right) \quad (x \mapsto -\log x \text{ is convex})$$

$$= -\frac{1}{M}\sum_{i=1}^M \log\left(\frac{1}{M}\sum_{j=1}^M \left(1 - \tfrac{1}{2}H^2(P_{\theta_i}, P_{\theta_j})\right)^n\right) \quad (\text{tensorization of } H^2)$$

$$\ge -\frac{1}{M}\sum_{i=1}^M \log\left(\frac{1}{M} + e^{-\frac{n}{2}\varepsilon^2}\right) \quad (H^2(P_{\theta_i}, P_{\theta_j}) \ge \varepsilon^2 \text{ for } i \ne j)$$

$$\ge \min\left\{M, \tfrac{n}{2}\varepsilon^2\right\} - \log 2 \quad \left(\frac{1}{a} + \frac{1}{b} \le \frac{2}{\min\{a, b\}}\right).$$

$\boxed{\Sigma}$

Example. In parametric families, usually $\log M_H(\mathcal{P}, \varepsilon) \sim d \log \frac{1}{\varepsilon}$, so the Haussler-Opper lower bound is also $\mathrm{Red}(\mathcal{P}^{\otimes n}) \geq \frac{d}{2} \log \frac{n}{C d \log n}$.

Relationship between redundancy & prediction risk.

Defn (prediction risk)

$$\mathrm{Risk}_n(\mathcal{P}) = \inf_{Q_{X_{n+1}|X^n}} \sup_{P_{X^{n+1}} \in \mathcal{P}} \mathbb{E}_{P_{X^n}}[D_{KL}(P_{X_{n+1}|X^n} \| Q_{X_{n+1}|X^n})]$$

("KL risk" for next-symbol prediction)

Mutual info representation. If $\mathcal{P} = (P_\theta)_{\theta \in \Theta}$,

$$\mathrm{Risk}_n(\mathcal{P}) = \sup_{\rho \in \Delta(\Theta)} I(\theta; X_{n+1} | X^n).$$

Pf.
$$\begin{aligned}\mathrm{Risk}_n(\mathcal{P}) &= \inf_{Q_{X_{n+1}|X^n}} \sup_{\rho} \mathbb{E}_{\theta\sim\rho}[D_{KL}(P_{X_{n+1}|X^n,\theta} \| Q_{X_{n+1}|X^n})] \\ &= \sup_{\rho} \inf_{Q_{X_{n+1}|X^n}} \mathbb{E}_{\theta\sim\rho}[D_{KL}(P_{X_{n+1}|X^n,\theta} \| Q_{X_{n+1}|X^n})] \quad \text{(minimax thm.)} \\ &= \sup_{\rho} I(\theta; X_{n+1} | X^n).\end{aligned}$$
$\square$

Redundancy - risk inequality: $\mathrm{Red}_n(\mathcal{P}) \leq \sum_{t=0}^{n-1} \mathrm{Risk}_t(\mathcal{P})$.

Pf. Chain rule: $I(\theta; X^n) = \sum_{t=0}^{n-1} I(\theta; X_t | X^{t-1})$

$$\Longrightarrow \sup_\rho I(\theta; X^n) \leq \sum_{t=0}^{n-1} \sup_\rho I(\theta; X_t | X^{t-1}).$$
$\square$

Tightness: For i.i.d. $\mathcal{P}^{\otimes n}$ with $\mathcal{P} = (P_\theta)_{\theta\in\Theta\subseteq\mathbb{R}^d}$, the MLE $\hat{\theta}_t$ based on X^t asymptotically achieves $\mathbb{E}_\theta[D_{KL}(P_\theta \| P_{\hat\theta})] \sim \frac{d}{2t}$ by Wilk's Thm.
So $\mathrm{Risk}_t \sim \frac{d}{2t}$, and $\mathrm{Red}_n \sim \frac{d}{2}\log n \sim \sum_{t=1}^{n-1} \mathrm{Risk}_t$.

"Online-to-batch" conversion: if each $P_{X^{n+1}} \in \mathcal{P}$ is stationary, i.e. $P_{X_{t_1},\dots,X_{t_k}} = P_{X_{t_1+\tau},\dots,X_{t_k+\tau}}$

then

$$\mathrm{Risk}_n(\mathcal{P}) \le \frac{1}{n}\mathrm{Red}(\mathcal{P}) + \mathrm{Mem}(\mathcal{P}).$$

where the "memory term" is

$$\mathrm{Mem}(\mathcal{P}) = \sup_{P_{X^{n+1}} \in \mathcal{P}} \frac{1}{n}\sum_{t=1}^{n} I(X_{n+1}; X^{n-t} \mid X_{n-t+1}^{n}).$$

Pf. Let $Q_{X^{n+1}} = \prod_{t=1}^{n+1} Q_{X_t \mid X^{t-1}}$ attain the minimax redundancy $\mathrm{Red}(\mathcal{P})$.

Now choose the Yang-Barron type predictor:

$$\widetilde{Q}_{X_{n+1}\mid X^n} = \frac{1}{n}\sum_{t=1}^{n} Q_{X_{t+1}\mid X^t}(\cdot \mid X_{n-t+1}^{n}).$$

Then

$$\mathbb{E}_{P_{X^n}}\left[D_{KL}(P_{X_{n+1}\mid X^n} \,\|\, \widetilde{Q}_{X_{n+1}\mid X^n})\right]$$

$$\le \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{P_{X^{n+1}}}\left[\log \frac{P_{X_{n+1}\mid X^n}(X_{n+1}\mid X^n)}{Q_{X_{t+1}\mid X^t}(X_{n+1}\mid X_{n-t+1}^{n})}\right] \quad \text{(convexity of KL)}$$

$$= \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{P_{X^{n+1}}}\left[\log \frac{P_{X_{n+1}\mid X_{n-t+1}^{n}}(X_{n+1}\mid X_{n-t+1}^{n})}{Q_{X_{t+1}\mid X^t}(X_{n+1}\mid X_{n-t+1}^{n})} + \log \frac{P_{X_{n+1}\mid X^n}(X_{n+1}\mid X^n)}{P_{X_{n+1}\mid X_{n-t+1}^{n}}(X_{n+1}\mid X_{n-t+1}^{n})}\right]$$

$$= \frac{1}{n}\sum_{t=1}^{n}\left(\mathbb{E}_{P_{X^{t+1}}}\left[\log \frac{P_{X_{t+1}\mid X^t}(X_{t+1}\mid X^t)}{Q_{X_{t+1}\mid X^t}(X_{t+1}\mid X^t)}\right] + I(X_{n+1}; X^{n-t} \mid X_{n-t+1}^{n})\right) \quad \text{(stationarity)}$$

$$\le \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{P_{X^t}}\left[D_{KL}(P_{X_{t+1}\mid X^t} \,\|\, Q_{X_{t+1}\mid X^t})\right] + \mathrm{Mem}(\mathcal{P})$$

$$\le \frac{1}{n} D_{KL}(P_{X^{n+1}} \,\|\, Q_{X^{n+1}}) + \mathrm{Mem}(\mathcal{P}) \le \frac{1}{n}\mathrm{Red}(\mathcal{P}) + \mathrm{Mem}(\mathcal{P})$$

(chain rule) $\square$

Example (Markov chain prediction) $\mathcal{P} = \{$stationary Markov chains on $[k]$ of length $n+1\}$

Then

$$\mathrm{Red}(\mathcal{P}) = O(k^2 \log n)$$

$$\mathrm{Mem}(\mathcal{P}) = \sup_{P_{X^{n+1}}} \frac{1}{n} I(X_{n+1}; X^n) \le \frac{\log k}{n}$$

$$\Longrightarrow \mathrm{Risk}_n(\mathcal{P}) = O\left(\frac{k^2 \log n}{n}\right).$$

- The main surprising feature is that this upper bound on $\mathrm{Risk}_n(P)$ does not depend on the mixing property of the Markov chain
- A pure statistical proof of this upper bound is unknown without mixing conditions
- This bound is tight for $3 \le k \ll \sqrt{n}$.

<u>Special topic</u>: characterization of R^* in Gaussian models (Mourtada. 2023)

Gaussian family: $\mathcal{P}_A = \{ N(\theta, I_n) : \theta \in A \}$, with $A \subseteq \mathbb{R}^n$.

By the entropic upper bound and Haussler-Opper lower bound, with

$$D_{KL}(N(\theta, I_n) \| N(\theta', I_n)) = \frac{1}{2}\|\theta - \theta'\|^2$$

$$\int \sqrt{dN(\theta, I_n) dN(\theta', I_n)} = \exp\left(-\frac{1}{8}\|\theta - \theta'\|^2\right),$$

we get the following characterization of $\mathrm{Red}(\mathcal{P}_A)$:

$$\boxed{\mathrm{Red}(\mathcal{P}_A) \asymp \inf_{r>0} \log N(A, \|\cdot\|_2, r) + r^2}$$

The main result of this section is a similar characterization of $R^*(\mathcal{P}_A)$.

$$\boxed{R^*(\mathcal{P}_A) \asymp \inf_{r>0} \log N(A, \|\cdot\|_2, r) + w_A(r)}$$

where $w_A(r)$ is the <u>local Gaussian width</u>:

$$w_A(r) = \sup_{\theta \in A} w(A \cap B(\theta, r)) = \sup_{\theta \in A} \mathbb{E}\left[\sup_{w \in A \cap B(\theta, r)} \langle w, Z \rangle\right], \quad Z \sim N(0, I_n)$$

(Alternative representation: let $r_N = \sup\{ r > 0 : \log N(A, r, \|\cdot\|_2) \ge r^2 \}$

$r_W = \sup\{ r > 0 : w_A(r) \ge r^2 \}$

then $\mathrm{Red}(\mathcal{P}_A) \asymp r_N^2$

$R^*(\mathcal{P}_A) \asymp r_N^2 + r_W^2$.)

Example (ellipsoids) If $A = \{\theta \in \mathbb{R}^n : \sum_{i=1}^n \frac{\theta_i^2}{a_i^2} \le 1\}$, then

$$\mathrm{Red}(\mathcal{P}_A) \asymp \inf_{r>0} \Big(\sum_{i: a_i > 2r} \log\big(\tfrac{a_i}{r}\big) + r^2 \Big),$$

$$R^*(\mathcal{P}_A) \asymp \inf_{r>0} \Big(\sum_{i=1}^n \log\big(1 + \tfrac{a_i^2}{r^2}\big) + r^2 \Big).$$

The key result in the proof lies in the following lemma.

Lemma.
$$w(A) - \sup_{\theta \in A} \frac{\|\theta\|^2}{2} \le R^*(\mathcal{P}_A) \le w(A).$$

Pf of thm assuming the lemma.

(Upper bound of R^*) First, observe a simple inequality:

$$R^*\Big(\bigcup_{i=1}^N \mathcal{P}_i\Big) \le \max_{i \in [N]} R^*(\mathcal{P}_i) + \log N.$$

The proof is easy: let Q_i attain $R^*(\mathcal{P}_i)$, then $\bar{Q} = \frac{1}{N}\sum_{i=1}^N Q_i$ attains the claimed upper bound of $R^*\big(\bigcup_{i=1}^N \mathcal{P}_i\big)$.

To apply this inequality, consider a r-covering $\theta_1, \cdots, \theta_N$ of A under $\|\cdot\|_2$. Then

$$\begin{aligned} R^*(\mathcal{P}_A) &\le \max_{i \in [N]} R^*(\mathcal{P}_{A \cap B(\theta_i, r)}) + \log N \\ &\le \max_{i \in [N]} w(A \cap B(\theta_i, r)) + \log N \quad \text{(by Lemma)} \\ &\le w_A(r) + \log N. \end{aligned}$$

(Lower bound of R^*) First, $R^*(\mathcal{P}_A) \ge \mathrm{Red}(\mathcal{P}_A) \gtrsim r_N^2$ by Haussler-Opper.

Second, for $r = r_w$ and $\theta \in A$,

$$\begin{aligned} R^*(\mathcal{P}_A) &\ge R^*(\mathcal{P}_{A \cap B(\theta, r)}) \\ &\ge w(A \cap B(\theta, r)) - \frac{r^2}{2} \quad \text{(by Lemma \& translation invariance)} \end{aligned}$$

$$\Longrightarrow R^*(\mathcal{P}_A) \ge w_A(r) - \frac{r^2}{2} \gtrsim \frac{r^2}{2} \quad \text{for } r = r_w \text{ and defn. of } r_w. \quad \square$$

Next we prove the lemma. First we write out the Shtarkov sum.

<u>Lemma.</u> $$R^*(\mathcal{P}_A) = \log \int_{\mathbb{R}^n} \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}d(x,A)^2} dx,$$
where $d(x,A) = \inf_{y \in A} \|x-y\|_2$.

<u>Pf.</u> Trivially follow from Shtarkov.

Using auxillary $Z \sim N(0, I_n)$,

$$\int_{\mathbb{R}^n} \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}d(x,A)^2} dx = \mathbb{E}_Z\left[e^{\frac{1}{2}(Z^2 - \mathrm{dist}(Z,A)^2)}\right]$$
$$= \mathbb{E}_Z\left[\exp\left(\sup_{w \in A}\left(\frac{Z^2}{2} - \frac{(Z-w)^2}{2}\right)\right)\right]$$
$$= \mathbb{E}_Z\left[\exp\left(\sup_{w \in A} \langle w, Z\rangle - \frac{\|w\|^2}{2}\right)\right].$$

(Pf of lower bound)

$$\log \mathbb{E}_Z\left[\exp\left(\sup_{w \in A} \langle w, Z\rangle - \frac{\|w\|^2}{2}\right)\right] \geq \log \mathbb{E}_Z\left[\exp\left(\sup_{w \in A} \langle w, Z\rangle\right)\right] - \frac{1}{2}\sup_{w \in A}\|w\|^2$$
$$\geq \mathbb{E}_Z\left[\sup_{w \in A} \langle w, Z\rangle\right] - \frac{1}{2}\sup_{w \in A}\|w\|^2 \quad \text{(Jensen)}$$
$$= w(A) - \frac{1}{2}\sup_{w \in A}\|w\|^2.$$

(Pf of upper bound) Let $\nu = N(0, I_n)$, and $f(z) = \sup_{w \in A} \langle w, z\rangle - \frac{\|w\|^2}{2}$.

By Gibbs' variational principle,

$$\log \mathbb{E}_Z[\exp(f(Z))] = \sup_{\mu} \mathbb{E}_{Z\sim\mu}[f(Z)] - D_{KL}(\mu \| \nu)$$
$$\leq \sup_{\mu} \mathbb{E}_{Z\sim\mu}[f(Z)] - \frac{1}{2}W_2^2(\mu, \nu) \quad \text{(Talagrand } T_2 \text{ ineq.)}$$
$$= \sup_{\mu} \mathbb{E}_{Z\sim\mu}[f(Z)] - \sup_{g}\left(\mathbb{E}_\mu g + \mathbb{E}_\nu g^c\right)$$
$$\left(g^c(z) = \inf_x \frac{\|x-z\|^2}{2} - g(x)\right)$$
$$\leq \mathbb{E}_{Z\sim\mu}\left[\sup_x f(x) - \frac{\|x-z\|^2}{2}\right].$$

On the other hand,

$$\sup_x \left(f(x) - \frac{\|x-z\|^2}{2} \right) = \sup_{w \in A} \sup_x \left(\langle w, x \rangle - \frac{\|w\|^2}{2} - \frac{\|x-z\|^2}{2} \right)$$

$$= \sup_{w\in A} \langle w, z \rangle$$

$$\Rightarrow \log \mathbb{E}_z e^{f(z)} \leq \mathbb{E}_z \left[\sup_{w \in A} \langle w, z \rangle \right] = w(A).$$

(An alternative proof in (Mourtada, 2023), using convex geometry)

<u>Defn (mixed volume)</u> Let $K_1, \cdots, K_r$ be convex bodies in $\mathbb{R}^n$. Write

$$\mathrm{Vol}_n(\lambda_1 K_1 + \cdots + \lambda_r K_r) = \sum_{j_1, \cdots, j_n = 1}^{r} V(K_{j_1}, \cdots, K_{j_n}) \lambda_{j_1} \cdots \lambda_{j_n},$$

the quantity $V(K_{j_1}, \cdots, K_{j_n})$ is called the <u>mixed volume</u>.

<u>Defn (intrinsic volume)</u> For $j \in [0, n]$,

$$V_j(K) = \binom{n}{j} \frac{V(\overbrace{K, \cdots, K}^{j}, \overbrace{B, \cdots, B}^{n-j})}{\kappa_{n-j}},$$

where $\kappa_j = \dfrac{\pi^{j/2}}{\Gamma(\frac{j}{2}+1)}$ is the volume of the unit ball in $\mathbb{R}^j$.

<u>Thm (Steiner formula)</u> $\mathrm{Vol}_n(K + tB) = \sum_{j=0}^{n} V_{n-j}(K) \kappa_j t^j$.

<u>Thm (Alexandrov-Fenchel)</u>

$$V(K_1, K_2, \cdots, K_n)^2 \geq V(K_1, K_1, K_3, \cdots, K_n) V(K_2, K_2, K_3, \cdots, K_n).$$

<u>Corollary</u>. By choosing $(K_1, K_2, \cdots, K_n) = (K, B, \overbrace{K, \cdots, K}^{j-1}, \overbrace{B, \cdots, B}^{n-j-1})$, we get

$$j V_j(K)^2 \geq (j+1) V_{j+1}(K) V_{j-1}(K).$$

In particular,

$$V_j(K) \leq \frac{V_1(K)^j}{j!}.$$

Back to the proof of upper bound: since $R^*(P_A) \le R^*(P_{conv(A)})$ and $w(A) = w(conv(A))$, WLOG we may assume $A = K$ is convex. Then

$$\int_{\mathbb{R}^n} \exp\left(-\tfrac{1}{2} d(x,K)^2\right) dx = \int_0^\infty \mathrm{Vol}_n\left(\{x \in \mathbb{R}^n : e^{-\frac{1}{2}d(x,K)^2} \ge t\}\right) dt$$
$$= \int_0^\infty \mathrm{Vol}_n\left(\{x \in \mathbb{R}^n : d(x,K) \le r\}\right) r e^{-r^2/2}\, dr$$
$$= \int_0^\infty \mathrm{Vol}_n(K + rB)\, r e^{-r^2/2}\, dr$$
$$= \int_0^\infty \sum_{j=0}^n V_{n-j}(K)\, \kappa_j r^j \cdot r e^{-r^2/2}\, dr$$
$$= \sum_{j=0}^n V_{n-j}(K)(2\pi)^{j/2}. \quad \left(\int_0^\infty r^{j+1} e^{-r^2/2}\, dr = 2^{j/2}\Gamma\left(\tfrac{j}{2}+1\right)\right)$$

$$\Longrightarrow \quad R^*(P_K) = \log \sum_{j=0}^n V_j(K)(2\pi)^{-j/2} = \log \underbrace{\sum_{j=0}^n V_j\left(\frac{K}{\sqrt{2\pi}}\right)}_{\text{called the Wills functional}}.$$

Using the corollary,

$$R^*(P_K) \le \log \sum_{j=0}^n \frac{V_1(K/\sqrt{2\pi})^j}{j!} < \log \exp\left(V_1\left(\frac{K}{\sqrt{2\pi}}\right)\right) = V_1\left(\frac{K}{\sqrt{2\pi}}\right)$$
$$= w(K).$$