

Adaptive compression over countable alphabets.

S. Boucheron,
joint work with D. Bontemps, A. Garivier, E. Gassiat & M. Ohanessian

Microsoft-INRIA, Paul Sabatier, Paris-Diderot, ENS, Paris-Sud

November 2014

Lossless compression 101

Lossless compression

Mapping messages (sequences of symbols from alphabet \mathcal{X}) to codewords (sequences of $\{0, 1\}$), so as to minimize the expected length of codewords in a one-to-one way.

Example

- i) Huffman coding
- ii) Lempel-Ziv, 1977-78 (zip, gzip, ...)
- iii) Burrows-Wheeler transform, move to front (bzip)

Non-ambiguous codes

A coding technique is **non-ambiguous**, iff any binary sequence can be parsed in at most one way into a sequence of

Non-ambiguous codes

Prefix codes are non-ambiguous

In a prefix code, no codeword is a strict prefix of another codeword.

A constraint on codeword lengths

Non-ambiguous codes satisfy a non-trivial constraint on codeword lengths.

Kraft-McMillan inequality

For $\lambda: A \rightarrow \mathbb{N}_+$,

$$\sum_{\omega \in A} 2^{-\lambda(\omega)} \leq 1, \text{ iff } \exists \text{ prefix code } f: A \rightarrow \{0, 1\}^* \text{ with } \ell[f(\omega)] = \lambda(\omega)$$

Corollary

No prefix code with codeword lengths $\ell(\omega) = \lfloor \log_2(p(\omega)) \rfloor + 1$ for

Coding probabilities

Kraft-Mac Millan inequality

provides a bridge between codes and probability distributions.

- ▶ Any non-ambiguous code defines a (sub)-probability distribution over the set of messages
- ▶ Any probability distribution Q over the set of messages defines a non-ambiguous encoding where codeword length is at most $-\log_2 Q(\omega) + 1$.
- ▶ This is not only an existential statement!

Example

Arithmetic coding w.r.t. Q^n encodes $x_{1:n} = (x_1, \dots, x_n) \in \Omega^n$ with
codeword length at most

Redundancy

Definition (Redundancy of coding probability Q^n with respect to source P^n)

Expected difference between codelengths obtained by feeding an arithmetic coder with $Q^n(\mathbf{x})$ rather than with the correct source statistics $P^n(\mathbf{x})$

$$D(P^n, Q^n) = \mathbb{E}_{P^n} \log \frac{P^n(X_{1:n})}{Q^n(X_{1:n})}$$

Redundancy corresponds to cumulative logarithmic/entropy/self-information loss in statistics and individual sequences analysis.

Cesa-Bianchi & Lugosi, Chap. 9 of Prediction, Learning and Games, 2006.

Minimax and Maximin

Λ^n is collection of probability distributions over messages of length n . Each probability distribution is called a source.

Definition (Minimax redundancy)

$$R^+(\mathcal{Q}^n, \Lambda^n) = \sup_{P \in \Lambda} D(P^n, \mathcal{Q}^n)$$

$$R^+(\Lambda^n) = \inf_{\mathcal{Q}} \sup_{P \in \Lambda} D(P^n, \mathcal{Q}^n)$$

Definition (Maximin redundancy)

π : prior distribution on sources

$$R_+(\Lambda^n) = \sup_{\pi} \inf_{\mathcal{Q}} \mathbb{E}_{\pi} D(P^n, \mathcal{Q}^n)$$

MinMax Theorem

$$R_+(\Lambda^n) = R^+(\Lambda^n)$$



Finite alphabet with cardinality k

Λ : memoryless sources over finite alphabet with cardinality k

Minimax redundancy

$$R^+(\Lambda^n) = \frac{k-1}{2} \log \frac{n}{2\pi e} + O(1)$$

Rissanen, Ryabko, Shtarkov, Krichevsky, Trofimov, Barron, Clarke, Xie et al..

Small alphabet setting, for $k = o(n)$

$$R^+(\Lambda^n) = \frac{k-1}{2} \log n - \frac{k}{2} \log \frac{k}{e} + o(k)$$

Szpankowski & Weinberger

Finite alphabet with cardinality k (cont'd)

Minimal Bayes redundancy is achieved by mixtures codes

$\inf_{\mathcal{Q}} \mathbb{E}_{\pi} D(P^n, \mathcal{Q}^n)$ is achieved by choosing $\mathcal{Q}^n(x_{1:n}) = \mathbb{E}_{\pi} P^n(x_{1:n})$

Krichevsky-Trofimov coding is asymptotically maximin and approximately minimax

$$\text{KT}(X_{n+1} = a | X_{1:n} = x_{1:n}) = \frac{n_a(x_{1:n}) + \frac{1}{2}}{n + \frac{k}{2}},$$

n_a = number of a 's in $x_{1:n}$.

KT-coding

KT-coding mixes coding probabilities using Jeffrey's (least favorable) prior.

↔ no need to explicitly estimate the source

Infinite alphabets

Negative result

$$\exists (Q^n)_n, \quad \forall P \in \Lambda, \lim_n \frac{1}{n} D(P^n, Q^n) = 0 \text{ iff}$$

$$\exists P^*, \quad \forall P \in \Lambda, \mathbb{E}_{P^1} - \log P^*(X) < \infty$$

J. Kieffer (1993)

\Leftrightarrow For stationary ergodic sources over a countable alphabet, no analogue of Lempel-Ziv coding.

Envelope classes

Envelope function

$f: \mathbb{N} \rightarrow \mathbb{R}_+$ with $1 < \sum_{j>0} f(j) < \infty$.

Envelope class

$\Lambda_f = \left\{ \mathbb{P} : \forall x \in \mathbb{N}, \mathbb{P}^1\{x\} \leq f(x) \text{ and } \mathbb{P} \text{ is stationary and memoryless.} \right\}$

Envelope distribution

- $F(k) = 1 - \sum_{j>k} f(j)$ For $k \geq l_f = \max\{k: \sum_{j \geq k} f(j) \geq 1\}$
- $\bar{F} = 1 - F$
- $U(t) = \inf\{x: F(x) \geq 1 - 1/t\}$

Light-tailed envelopes

Sub-exponential classes

- F_C has non-decreasing hazard rate (ako log-concavity assumption)
- $U_C \circ \exp$ is concave.

Examples

- ▶ Exponential envelopes. $f(k) = \gamma e^{-\left(\frac{k}{\beta}\right)^\alpha}$. with $\alpha \geq 1, \beta > 0$ and $\gamma > 1$
- ▶ Poisson envelopes $f(k) = \gamma e^{-\beta} \beta^k / k!$ with $\beta > 0$ and $\gamma > 1$
- ▶ ...

Fat-tailed envelopes

Regularly varying envelopes

F_C (resp. U_C) is regularly varying with index $-1/\gamma$ (resp. $\gamma > 0$)

$$\forall x > 0, \quad \lim_t \frac{F_C(tx)}{F_C(t)} = x^{-1/\gamma}.$$

$$U_C(t) = t^\gamma \ell(t)$$

where ℓ is slowly varying

Example

- ▶ Power-law envelopes: $U_C(t) = \kappa t^\gamma$
- ▶ Heavy-Tailed envelopes $U_C(t) = \kappa t^\gamma \ell(t)$

An upper-bound on minimax redundancy

Theorem (BGG, 2009)

If Λ is a class of memoryless sources, with the tail envelope distribution function $\bar{F}_{\Lambda^1}(u) = \sum_{k>u} \hat{p}(k)$, then:

$$R^*(\Lambda^n) \leq \inf_{u: u \leq n} \left[n \bar{F}_{\Lambda^1}(u) \log_2 e + \frac{u-1}{2} \log_2 n \right] + 2.$$

Suggestion

If the envelop is known, choose threshold τ as the solution of $\bar{F}_{\Lambda^1}(u) = \frac{u}{n}$.

- i) Encode symbols over threshold using Elias penultimate code
- ii) Encode other symbols using Krichevsky-Trofimov mixture over alphabet $\{1, \dots, \tau\}$.

Lower bounds

For any prior μ on $\Lambda^1(f)$

$$\begin{aligned} R^+(\Lambda^n) &\geq \inf_{\mathcal{Q}^n} \mathbb{E}_\mu D(P^n | \mathcal{Q}^n) \\ &= \mathbb{E}_\mu D(P^n | \mathbb{E}_\mu P^n) \\ &= I(\theta; X_{1:n}) \end{aligned}$$

Redundancy-Capacity Theorem

The Bayes risk coincides with the mutual information between the parameter θ and the observation $X_{1:n}$

Options

- ▶ Design of priors and ad hoc lower bounds on mutual information
- ▶ Design of analogs of Fano's lemma
- ▶ Connect mutual information and metric entropy of the envelope class

Cooking risk bounds

For an ad hoc prior

$$I(\theta; X_{1:n}) \geq \mathbb{E}Z_n$$

where Z_n is the number of distinct symbols in $X_{1:n}$

B., Garivier & Gassiat, 2009, B. Gassiat & Ohannessian 2014

For the same prior

$$\mathbb{E}Z_n \geq m_n$$

where m_n satisfies $\bar{F}_c(m_n) \approx \frac{m_n}{n}$

The number of distinct symbols for regularly varying sources reflects the regularly varying properties of the sources

Karlin 2007, Gnedin, Hansen & Pitman 2007

Cooking risk bounds

Message length N may be randomized by **Poissonization**

$N \sim \text{Poi}(n)$ and a message of length N has to be encoded.

$$R^+(\Lambda^n) \leq R^+(\Lambda^N) + 1$$

For envelope classes

$$R^+(\Lambda_f^N) \leq \sum_{i=1}^{\infty} R^+(\text{Poi}(nf_i)) \leq \sum_{i=1}^{\infty} \left(\log \left(\sqrt{\frac{2nf_i}{2\pi}} + 2 \right) \wedge nf_i \log e \right)$$

Acharya, J., Jafarpour, A., Orłitsky, A., & Suresh, A. T. (2014).
Poissonization and universal compression of envelope classes.

For power law envelopes $U_c(t) = \kappa t^\gamma$

Cumulative entropy risk and metric entropy

Hausser & Opper, AoS, 1997

Lower bound on minimax redundancy using **metric entropy** of Λ_f^1 under **Hellinger metric**.

Hellinger distance and metric entropy

$$H^2(P_1, P_2) = \sum_{k \in \mathbb{N}} \left(\sqrt{p_1(k)} - \sqrt{p_2(k)} \right)^2.$$

$$\mathcal{H}_\epsilon(\Lambda) = \ln \mathcal{D}_\epsilon(\Lambda)$$

Hausser, Opper, AoS 1997

For any prior π on Λ_1

$$R^+(\Lambda^n) \geq \mathbb{E}_\pi \left[-\log \mathbb{E}_\pi e^{-n \frac{H^2(P_1, P_2)}{2}} \right]$$

► Details

Consequence

$$R^+(\Lambda^n) \geq \log e \sup_{\epsilon} \min \left(H_\epsilon(\Lambda_1), \frac{n\epsilon^2}{2} \right) - 1$$

Flavors of adaptivity

For collections of small classes

Definition (Asymptotic adaptivity)

$(\mathcal{Q}^n)_n$ is **asymptotically adaptive** with respect to $(\Lambda_m)_{m \in \mathcal{M}}$ if

$$\forall m \in \mathcal{M}, \quad R^+(\mathcal{Q}^n, \Lambda_m^n) = \sup_{\mathbb{P} \in \Lambda_m} D(\mathbb{P}^n, \mathcal{Q}^n) \leq (1 + o(1))R^+(\Lambda_m^n)$$

For collections of massive envelop classes

Definition (Weak asymptotic adaptivity)

$(\mathcal{Q}^n)_n$ is **asymptotically weakly adaptive** with respect to $(\Lambda_m)_{m \in \mathcal{M}}$

$$\forall m \in \mathcal{M}, \quad R^+(\mathcal{Q}^n, \Lambda_m^n) \leq o(\log n)R^+(\Lambda_m^n).$$

Light-tailed envelopes

The AC-code is adaptive with respect to source classes defined by envelopes with finite and non-decreasing hazard rate.

Theorem (B., Bontemps, Gassiat, 2014)

Q^n : the coding probability associated with the AC-code,
 If f is an envelope with **non-decreasing hazard rate**,

$$R^+(Q^n; \Lambda_f^n) \leq (1 + o(1))R^+(\Lambda_f^n)$$

while

$$R^+(\Lambda_f^n) = (1 + o(1))(\log e) \int_1^n \frac{U_c(x)}{2x} dx$$

AC-code and sub-exponential envelope classes

Sub-exponential envelope class $\Lambda(\alpha, \beta, \gamma)$

with $\alpha \geq 1, \beta > 0$ and $\gamma > 1$, defined by envelope

$$f(k) = \gamma e^{-\left(\frac{k}{\beta}\right)^\alpha}.$$

Corollary

The AC-code is adaptive with respect to sub-exponential envelope classes.

For all $\alpha \geq 1, \beta > 0$ and $\gamma > 1$

$$R^+(\mathcal{Q}^n; \Lambda^n(\alpha, \beta, \gamma)) \leq (1 + o(1))R^+(\Lambda^n(\alpha, \beta, \gamma))$$

$$R^+(\Lambda^n(\alpha, \beta, \gamma)) = \frac{\alpha}{2(\alpha + 1)} \beta (\ln(2))^{1/\alpha} (\log n)^{1+1/\alpha} (1 + o(1)).$$

Censuring codes: sketch

AC-code : Thresholding above last record

$$m_i = \max_{1 \leq j \leq i} x_j.$$

The j^{th} record is denoted by \tilde{m}_j ($\tilde{m}_0 = 0$)

Let $\tilde{\mathbf{m}} = (\tilde{m}_i - \tilde{m}_{i-1} + 1)1$.

Symbols from $\tilde{\mathbf{m}}$ encoded using Elias penultimate code.

Progressive KT coding below the last record

$$\tilde{x}_i = x_i \mathbb{I}_{x_i \leq m_{i-1}}.$$

C_M : progressive **KT**- encoding of $\tilde{x}_{1:n}0$

$$Q_{i+1}(\tilde{X}_{i+1} = j | X_{1:i} = x_{1:i}) = \frac{n_i^j + \frac{1}{2}}{i + \frac{m_i + 1}{2}} \quad \text{if } 1 \leq j \leq m_i,$$

$$Q_{i+1}(\tilde{X}_{i+1} = 0 | X_{1:i} = x_{1:i}) = \frac{1/2}{i + \frac{m_i + 1}{2}},$$

Example

$x_{1:n}$ 5 15 8 1 30 7 1 2 1 8 4 7 15 1 5 17 13 4 12 12 $m_{1:n}$ 5 15 15 15 30
 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 $\check{x}_{1:n} \rightsquigarrow$ progressive KT
 encoding 0 0 8 1 0 7 1 2 1 8 4 7 15 1 5 17 13 4 12 12 $\check{m} \rightsquigarrow$ Elias
 encoding

6 11

16

Envelopes with heavier tails

If the tail envelope distribution is heavier than exponential, thresholding at maximum does not lead to (weakly) adaptive coding

Ideal threshold: solution of

$$t\bar{F}_c(u) = \frac{u}{2} \log t$$

Proxy threshold: m_c solution of

$$t\bar{F}_c(u) = u \text{ or } u = U_c\left(\frac{t}{u}\right)$$

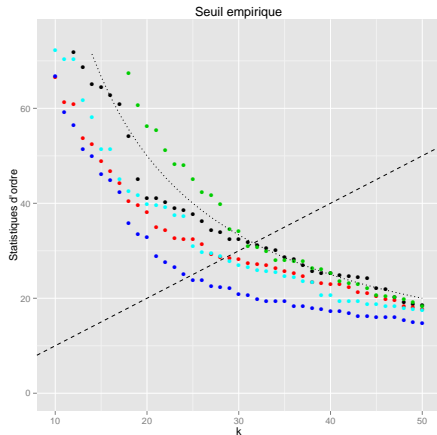
Properties

- ▶ m_c is non-decreasing.
- ▶ $m_c(t) \nearrow \infty$
- ▶ $m_c(t)/t \searrow 0$
- ▶ If U_c is γ -regularly varying, m_c is γ -regularly varying.

Empirical threshold

$$M_n = \min\left(n, \{k : X_{k,n} \leq k\}\right)$$

Adaptive thresholding



$$M_n = \min(n, \{k : X_{k,n} \leq t\})$$

$$F_C \in \text{MDA}(\gamma), \gamma > 0$$

$$\triangleright \frac{M_n}{m_n} \xrightarrow{P} 1.$$

$$\triangleright \frac{X_{M_n, n}}{m_c(n)} \xrightarrow{P} 1.$$

M_n is self-bounded

$$\begin{aligned} \mathbb{P}\{|M_n - \mathbb{E}M_n| \geq t\} \\ \leq 2e^{-\frac{t^2}{2(\mathbb{E}M_n + t)}}. \end{aligned}$$

Decomposing redundancy of AC-code

Decomposing pointwise redundancy

$$-\log Q^n(X_{1:n}) + \log P^n(X_{1:n}) = \underbrace{\ell(C_E)}_I + \underbrace{\ell(C_M) + \log P^n(X_{1:n})}_{II}.$$

Establishing main theorem in (BBG, 2014)

↪

- ▶ (I) (Elias encoding of increments between records) is negligible with respect to $R^+(\Lambda_f^n)$, uniformly for $\mathbb{P} \in \Lambda_f$,
- ▶ The expected value of (II) is upper bounded, uniformly for $\mathbb{P} \in \Lambda_f$, by a term which is equivalent to $R^+(\Lambda_f^n)$.

Bounding the length of Elias encoding

Proposition

Envelop f with non-decreasing hazard rate. Then, for all $\mathbb{P} \in \Lambda_f$,

$$\begin{aligned} \mathbb{E}[\ell(C_E)] &\leq \mathbb{E}\left[\sum_{i=1}^{n_n^0} (2\log(1 + \tilde{m}_i - \tilde{m}_{i-1}) + \rho)\right] \\ &\leq (2\log(e) + \rho)\mathbb{E}M_n \\ &\leq (2\log(e) + \rho)(U_C(\exp(H_n)) + 1) \end{aligned}$$

where $\rho \leq 2$.

Redundancy of progressive mixture coding

Proposition (pointwise bound)

Let $i_0 = 1 \vee \lfloor M_n/4 \rfloor$, then

$$-\ln Q^n(\tilde{X}_{1:n}) + \ln P^n(X_{1:n}) \leq \underbrace{\frac{M_n(\ln(M_n) + 10)}{2}}_{(A.I)} + \frac{\ln n}{2} + \underbrace{\sum_{i=i_0}^{n-1} \left(\frac{M_i}{2i+1} \right)}_{(A.II)}$$

$$\begin{aligned} & -\ln Q^n(\tilde{X}_{1:n}) + \ln P^n(X_{1:n}) \\ &= \underbrace{-\ln \text{KT}_{M_n+1}(\tilde{X}_{1:n}) + \ln P^n(X_{1:n})}_{(A) \leq \frac{M_n+1}{2} \ln(n) + 2 \ln(2)} - \underbrace{\ln Q^n(\tilde{X}_{1:n}) + \ln \text{KT}_{M_n+1}(\tilde{X}_{1:n})}_{(B) = -\sum_{i=1}^{n-1} \ln \left(\frac{2i+1+M_n}{2i+1+M_i} \right) \leq 0} \end{aligned}$$

Controlling (II), continued

Proposition

$f: \mathbb{N}_+ \rightarrow [0, 1]$: envelope with finite and non-decreasing hazard rate.

$$\mathbb{E}[\ell(C_M) + \log \mathbb{P}(X_{1:n})] \leq \log(e) \int_1^n \frac{U_C(x)}{2x} dx (1 + o(1))$$

as $n \nearrow \infty$.

► Nuts and bolts

Upper bounds on redundancy of ETAC-code

The length of the ETAC codeword is decomposed in the same way as for the AC codeword

$$\begin{aligned}
 & \ell(C_M) + \log_2 \mathbb{P}^n(X_{1:n}) \\
 &= \underbrace{-\log_2 \mathbb{KT}_{X_{M_n, n+1}}(\tilde{X}_{1:n}) + \log_2 \mathbb{P}^n(X_{1:n})}_{\leq \frac{X_{M_n, n+1}}{2} \log_2(n) + 2} \\
 & \quad \underbrace{-\log_2 Q^n(\tilde{X}_{1:n}) + \log_2 \mathbb{KT}_{X_{M_n, n+1}}(\tilde{X}_{1:n})}_{\leq 0} \\
 \\
 & \ell(C_E) \leq 2 \sum_{i=1}^{n-1} \mathbb{I}_{X_{i+1} > X_{M_i, i}} \left\{ \log_2(1 + X_{i+1} - X_{M_i, i}) + \rho \right\}.
 \end{aligned}$$

Bounding redundancy of ETAC encoding

If $\bar{F}_c \in MDA(-1/\gamma)$ with $\gamma > 0$,
 $\forall \epsilon > 0$, for sufficiently large n ,

$$\mathbb{E}X_{M_n, n} \leq m_n(1 + \epsilon).$$

Redundancy of ETAC code

If Q^n is the coding probability associated with the ETAC code

$$R^+(Q^n, \Lambda_n) \leq (5 + o_\Lambda(1)) \frac{m_n}{2} \log n + 2$$

B., Gassiat, Ohannessian, 2014

Bounding length of Elias encoding

If $\bar{F}_C \in MDA(-1/\gamma)$ with $\gamma > 0$,

$\forall \epsilon > 0$, for sufficiently large n ,

$$\mathbb{E} \ell(C_E) \leq 2(1+\epsilon) \sum_{i=1}^n \frac{m_i \log m_i}{i} \leq 2(1+\epsilon) \int_1^n \frac{m_t \log m_t}{t} dt \leq \frac{2\gamma}{\gamma+1} m_n \log$$

References

- ▶ S. Boucheron and E. Gassiat : A Bernstein-von Mises theorem for discrete probability distributions *Electronic Journal of Statistics*. **3** (2009) 114-148.
- ▶ S. Boucheron and A. Garivier and E. Gassiat : Coding over Infinite Alphabets *IEEE Trans. on Inform. Theory* **55** (2009) 358 - 373.
- ▶ D. Bontemps : Universal coding on infinite alphabets: exponentially decreasing envelopes. *IEEE Trans. Inform. Theory* 57 (2011), no. 3, 1466–1478.
- ▶ D. Bontemps, S. Boucheron and E. Gassiat : Adaptive compression against a countable alphabet. *IEEE Trans. Inform. Theory* 60 (2014), 808–821.
- ▶ S. Boucheron, E. Gassiat & M. Ohanessian : Weakly adaptive compression against a countable alphabet. 2014
- ▶ S. Boucheron, M. Thomas : Concentration inequalities for order statistics. *Electronic Communications in Probability*. **17** (2012).

Lossless compression 101

An ambiguous code

Binary expansion is one-to-one mapping of integers towards binary strings. It is not a non-ambiguous code.

Example

The string

1111111

may be parsed as

- ▶ the binary encoding of $127 = 2^7 - 1 = 2^6 + 2^5 + 2^4 + 2^3 + 2^2 + 2^1 + 2^0$ or
- ▶ the concatenation of the encodings of 3 (11), 7 (111), and 3 (11).

Envelop classes

Smoothed distribution function

- F_C has piecewise constant hazard rate,
- $\bar{F}_C(n) = \bar{F}(n)$
- $U_C(t) = \inf\{x: 1/\bar{F}_C(x) \geq t\}$.

If $X \sim F_C$ then $\lfloor X \rfloor + 1 \sim F$ and $U(t) = \lfloor U_C(t) \rfloor + 1$ for $t > 1$.

Lemma (Stochastic comparison by quantile coupling)

There exists a probability space where $X \sim G \in \Lambda_f$, $Y \sim F_C$ such that

$$\mathbb{P}\{X \leq Y\} = 1$$

◀ Return

Cumulative entropy risk and metric entropy

Repeatedly using Fubini's theorem and Jensen's inequality.
Fubini

$$\int_{\Theta} d\pi(\theta^*) \int_{X^n} dP_{\theta^*}^n(X_{1:n}) \frac{\int_{\Theta} d\pi(\tilde{\theta}) \frac{dP_{\tilde{\theta}}(X_{1:n})}{dP_{\theta^*}(X_{1:n})}}{\int_{\Theta} d\pi(\hat{\theta}) \sqrt{\frac{dP_{\hat{\theta}}(X_{1:n})}{dP_{\theta^*}(X_{1:n})}}} = 1$$

Cumulative entropy risk and metric entropy

Fubini

$$\int_{\Theta} d\pi(\theta^*) \int_{\mathcal{X}^n} dP_{\theta^*}^n(x_{1:n}) \frac{\int_{\Theta} d\pi(\tilde{\theta}) \frac{dP_{\tilde{\theta}}(X_{1:n})}{dP_{\theta^*}(X_{1:n})}}{\int_{\Theta} d\pi(\tilde{\theta}) \sqrt{\frac{dP_{\tilde{\theta}}(X_{1:n})}{dP_{\theta^*}(X_{1:n})}}} = 1$$

$$- \int_{\Theta} d\pi(\theta^*) \int_{\mathcal{X}^n} dP_{\theta^*}^n(x_{1:n}) \log \int_{\Theta} d\pi(\tilde{\theta}) \frac{dP_{\tilde{\theta}}(X_{1:n})}{dP_{\theta^*}(X_{1:n})}$$

Cumulative entropy risk and metric entropy

Fubini

$$\begin{aligned} & \int_{\Theta} d\pi(\theta^*) \int_{\mathcal{X}^n} dP_{\theta^*}^n(x_{1:n}) \frac{\int_{\Theta} d\pi(\tilde{\theta}) \frac{dP_{\tilde{\theta}}(X_{1:n})}{dP_{\theta^*}(X_{1:n})}}{\int_{\Theta} d\pi(\widehat{\theta}) \sqrt{\frac{dP_{\widehat{\theta}}(X_{1:n})}{dP_{\theta^*}(X_{1:n})}}} = 1 \\ & - \int_{\Theta} d\pi(\theta^*) \int_{\mathcal{X}^n} dP_{\theta^*}^n(x_{1:n}) \log \int_{\Theta} d\pi(\tilde{\theta}) \frac{dP_{\tilde{\theta}}(X_{1:n})}{dP_{\theta^*}(X_{1:n})} \\ & \geq - \int_{\Theta} d\pi(\theta^*) \int_{\mathcal{X}^n} dP_{\theta^*}^n(x_{1:n}) \log \int_{\Theta} d\pi(\widehat{\theta}) \sqrt{\frac{dP_{\widehat{\theta}}(X_{1:n})}{dP_{\theta^*}(X_{1:n})}} \end{aligned}$$

Cumulative entropy risk and metric entropy

$$\begin{aligned} & - \int_{\Theta} d\pi(\theta^*) \int_{\mathcal{X}^n} dP_{\theta^*}^n(x_{1:n}) \log \int_{\Theta} d\pi(\tilde{\theta}) \frac{dP_{\tilde{\theta}}(X_{1:n})}{dP_{\theta^*}(X_{1:n})} \\ & \geq - \int_{\Theta} d\pi(\theta^*) \int_{\mathcal{X}^n} dP_{\theta^*}^n(x_{1:n}) \log \int_{\Theta} d\pi(\widehat{\theta}) \sqrt{\frac{dP_{\widehat{\theta}}(X_{1:n})}{dP_{\theta^*}(X_{1:n})}} \\ & \geq - \int_{\Theta} d\pi(\theta^*) \log \int_{\mathcal{X}^n} dP_{\theta^*}^n(x_{1:n}) \int_{\Theta} d\pi(\widehat{\theta}) \sqrt{\frac{dP_{\widehat{\theta}}(X_{1:n})}{dP_{\theta^*}(X_{1:n})}} \end{aligned}$$

Cumulative entropy risk and metric entropy

$$\begin{aligned} & - \int_{\Theta} d\pi(\theta^*) \int_{\mathcal{X}^n} dP_{\theta^*}^n(x_{1:n}) \log \int_{\Theta} d\pi(\tilde{\theta}) \frac{dP_{\tilde{\theta}}(X_{1:n})}{dP_{\theta^*}(X_{1:n})} \\ & \geq - \int_{\Theta} d\pi(\theta^*) \int_{\mathcal{X}^n} dP_{\theta^*}^n(x_{1:n}) \log \int_{\Theta} d\pi(\hat{\theta}) \sqrt{\frac{dP_{\hat{\theta}}(X_{1:n})}{dP_{\theta^*}(X_{1:n})}} \\ & \geq - \int_{\Theta} d\pi(\theta^*) \log \int_{\mathcal{X}^n} dP_{\theta^*}^n(x_{1:n}) \int_{\Theta} d\pi(\hat{\theta}) \sqrt{\frac{dP_{\hat{\theta}}(X_{1:n})}{dP_{\theta^*}(X_{1:n})}} \\ & \geq - \int_{\Theta} d\pi(\theta^*) \log \int_{\Theta} d\pi(\hat{\theta}) \int_{\mathcal{X}^n} \sqrt{dP_{\hat{\theta}}(X_{1:n}) dP_{\theta^*}(X_{1:n})} \end{aligned}$$

Cumulative entropy risk and metric entropy

$$\begin{aligned} & - \int_{\Theta} d\pi(\theta^*) \int_{\mathcal{X}^n} dP_{\theta^*}^n(x_{1:n}) \log \int_{\Theta} d\pi(\tilde{\theta}) \frac{dP_{\tilde{\theta}}(X_{1:n})}{dP_{\theta^*}(X_{1:n})} \\ & \geq - \int_{\Theta} d\pi(\theta^*) \int_{\mathcal{X}^n} dP_{\theta^*}^n(x_{1:n}) \log \int_{\Theta} d\pi(\widehat{\theta}) \sqrt{\frac{dP_{\widehat{\theta}}(X_{1:n})}{dP_{\theta^*}(X_{1:n})}} \\ & \geq - \int_{\Theta} d\pi(\theta^*) \log \int_{\mathcal{X}^n} dP_{\theta^*}^n(x_{1:n}) \int_{\Theta} d\pi(\widehat{\theta}) \sqrt{\frac{dP_{\widehat{\theta}}(X_{1:n})}{dP_{\theta^*}(X_{1:n})}} \\ & \geq - \int_{\Theta} d\pi(\theta^*) \log \int_{\Theta} d\pi(\widehat{\theta}) \int_{\mathcal{X}^n} \sqrt{dP_{\widehat{\theta}}(X_{1:n}) dP_{\theta^*}(X_{1:n})} \\ & = - \int_{\Theta} d\pi(\theta^*) \log \int_{\Theta} d\pi(\widehat{\theta}) \alpha_H(P_{\widehat{\theta}}, P_{\theta^*})^n \end{aligned}$$

Cumulative entropy risk and metric entropy

$$\begin{aligned} & - \int_{\Theta} d\pi(\theta^*) \int_{\mathcal{X}^n} dP_{\theta^*}^n(x_{1:n}) \log \int_{\Theta} d\pi(\tilde{\theta}) \frac{dP_{\tilde{\theta}}(X_{1:n})}{dP_{\theta^*}(X_{1:n})} \\ & \geq - \int_{\Theta} d\pi(\theta^*) \int_{\mathcal{X}^n} dP_{\theta^*}^n(x_{1:n}) \log \int_{\Theta} d\pi(\widehat{\theta}) \sqrt{\frac{dP_{\widehat{\theta}}(X_{1:n})}{dP_{\theta^*}(X_{1:n})}} \\ & \geq - \int_{\Theta} d\pi(\theta^*) \log \int_{\mathcal{X}^n} dP_{\theta^*}^n(x_{1:n}) \int_{\Theta} d\pi(\widehat{\theta}) \sqrt{\frac{dP_{\widehat{\theta}}(X_{1:n})}{dP_{\theta^*}(X_{1:n})}} \\ & \geq - \int_{\Theta} d\pi(\theta^*) \log \int_{\Theta} d\pi(\widehat{\theta}) \int_{\mathcal{X}^n} \sqrt{dP_{\widehat{\theta}}(X_{1:n}) dP_{\theta^*}(X_{1:n})} \\ & = - \int_{\Theta} d\pi(\theta^*) \log \int_{\Theta} d\pi(\widehat{\theta}) \alpha_H(P_{\widehat{\theta}}, P_{\theta^*})^n \\ & \geq - \int_{\Theta} d\pi(\theta^*) \log \int_{\Theta} d\pi(\widehat{\theta}) \exp\left(-n \frac{H^2(P_{\widehat{\theta}}, P_{\theta^*})}{2}\right) \end{aligned}$$

Controlling (II), continued

Maximal inequalities

Let $Y_1, \dots, Y_n \sim_{i.i.d.} F$ with density $f = F'$ on $[1, \infty)$ and $f/\bar{F} \nearrow$.

Let b be the infimum of the hazard rate.

Let $U(t) = \inf\{x: F(x) \geq 1 - 1/t\} = F^{\leftarrow}(1 - 1/t)$.

Let $Y_{(1)} \geq \dots \geq Y_{(n)}$ be the order statistics.

$$\mathbb{E}[Y_{(1)}] \leq U(\exp(H_n))$$

$$\mathbb{E}[Y_{(1)} \ln(Y_{(1)})] \leq (\mathbb{E}Y_{(1)}) \ln(\mathbb{E}Y_{(1)}) + 2/b^2.$$

where $H_n = \sum_{i=1}^n 1/i$.

Controlling (II), continued

Corollary

Let $X_1, \dots, X_n \sim_{i.i.d.} P \in \Lambda_f^1$, let $M_n = \max(X_1, \dots, X_n)$, then,

$$\begin{aligned} \mathbb{E}M_n &\leq U_c(en) + 1 \\ \mathbb{E}[M_n \log M_n] &\leq [U_c(en) + 1] \log[U_c(en) + 1] + 2/b^2. \end{aligned}$$

Ingredients of proof

- ▶ Rényi's representation of order statistics & concavity of $U \circ \exp$
- ▶ Sub-additivity of relative entropy (see Ledoux, 2001, Massart, 2006)
- ▶ The entropy method \rightarrow sharp tail and moment bounds for order statistics (B. & Thomas, 2012)

Lower bounds : back to metric entropy

Corollary of Haussler & Opper, AoS, 1997

Assume there exists a (very) slowly varying function h such that:

$$\mathcal{H}_\epsilon(\Lambda) = h\left(\frac{1}{\epsilon}\right)(1 + o(1)) \quad \text{as } \epsilon \searrow 0.$$

Then

$$R^+(\Lambda^n) = (\log e)h(\sqrt{n})(1 + o(1)) \quad \text{as } n \nearrow +\infty.$$

Entropy of envelope classes with finite and non-decreasing hazard rate.

$$\mathcal{H}_\epsilon(\Lambda_f) = (1 + o(1)) \int_0^{1/\epsilon^2} \frac{U_c(x)}{2x} dx \quad \text{as } \epsilon \searrow 0.$$