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

March, 17th, 2015

Lossless compression over a countable alphabet

Lossless compression

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

Non-ambiguous codes satisfy Kraft-McMillan inequality

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

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

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 $-\log_2 Q(\omega) + 1$.

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 $P^n(\mathbf{x})$

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

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

Definition (Minimax redundancy)

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

Definition (Maximin redundancy)

π : prior distribution on sources

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

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

Redundancies: alphabet size matters

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..

Krichevsky-Trofimov coding is asymptotically maximin and approximately minimax

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

Countable alphabets

Negative results

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

iff

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

J. Kieffer (1993), Gyorfi, Pali van der Meulen (1993)

Envelop classes

For stationary ergodic sources over a countable alphabet, no analogue of Lempel-Ziv coding.
To obtain positive results... it is necessary to impose constraints on source classes

Envelop function

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

Envelop 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 $k \geq l_f := \max\{k :\ \sum_{j \geq k} f(j) \geq 1\}$ envelope distribution
- $\overline{F} = 1 - F$ tail envelope function
- $U(t) = \inf\{x : F(x) \geq 1 - 1/t\}$ tail quantile (envelope) function

Envelopes

Sub-exponential classes

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

Example

- 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 $\gamma > 1$
- ...

Regularly varying envelops

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

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

$$U_c(t) = t^{\gamma} \ell(t)$$

where ℓ is slowly varying

Example

- Power-law envelopes: $U_c(t) = \kappa t^{\gamma}$
- Heavy-Tailed envelopes $U_c(t) = \kappa t^{\gamma} \ell(t)$

Bounds 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 $\{1, \ldots, \tau\}$.

If the envelop is not known, look for a data-driven threshold

▸ Lower bounds

Flavors of adaptivity

For collections of small classes

Definition (Asymptotic adaptivity)

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

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

For collections of massive envelop classes

Definition (Weak asymptotic adaptivity)

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

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

Censuring codes: sketch

AC-code : Thresholding above last record

$m_i = \max_{1 \le j \le i} x_j$.
The j^{th} record is denoted by $\widetilde{m}_j$ ($\widetilde{m}_0 = 0$)
Let $\widetilde{\mathbf{m}} = (\widetilde{m}_i - \widetilde{m}_{i-1} + 1)1$.
Symbols from $\widetilde{\mathbf{m}}$ encoded using Elias penultimate code.

Progressive KT coding below the last record

$\widetilde{x}_i = x_i \mathbb{I}_{x_i \le m_{i-1}}$.
C_M : progressive KT- encoding of $\widetilde{x}_{1:n}0$

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

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

where n_i^j is the number of occurrences of symbol j in $x_{1:i}$, $n_i^0 = 0$.

Example

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

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

▸ Details

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

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

Proxy threshold: m_c solution of

$t\overline{F}_c(u) = u$ 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 $\gamma/(\gamma+1)$-regularly varying.

Empirical theshold

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

Weak adaptivity of ETAC encoding

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

$\forall \epsilon > 0$, for sufficiently large n, $\mathbb{E}X_{M_n,n} \leq m_n(1+\epsilon)$ $\qquad R^+(\Lambda_f^n) \geq \frac{m_n}{2}$.

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

For power law envelopes $U_c(t) = \kappa t^\gamma$ (Acharya et al. 2014)

$$R^+(\Lambda_f^n) \sim \left(\frac{\kappa^{1/\gamma}}{\gamma} n\right)^{\frac{\gamma}{\gamma+1}} \left(\frac{1}{\gamma} + \gamma \log \mathrm{e} + c\right)$$

Details

Thanks

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 Alphabet
  s 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 â€Ş 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). 1-12

Envelop classes

Smoothed distribution function

- F_c has piecewise constant hazard rate,
- $\overline{F}_c(n) = \overline{F}(n)$
- $U_c(t) = \inf\{x : 1/\overline{F}_c(x) \geq t\}$.

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

Bounds on minimax redundancy

Redundancy-Capacity theorem

For any prior μ on $\Lambda^1(f)$

$$R^+(\Lambda^n) = I(\theta; X_{1:n})$$

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 $X_{1:n}$

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

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

Made in California

For light-tailed envelopes

$$R^+(\Lambda^n_f) \sim \log(e) \int_1^n \frac{U_c(x)}{2x} \mathrm{d}x \, (1+o(1))$$

Bontemps, B. & Gassiat, 2014 using Haussler & Opper, AoS, 1997

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

$$R^+(\Lambda^n_f) \sim \left(\frac{\kappa^{1/\gamma}}{\gamma} n\right)^{\frac{\gamma}{\gamma+1}} \left(\frac{1}{\gamma} + \gamma \log \mathrm{e} + c\right)$$

Acharya, J., Jafarpour, A., Orlitsky, A., & Suresh, A. T. (2014)

Return

Censuring codes: sketch

$x_{1:n}$

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

$\tilde{x}_{1:n} \curvearrowright$ progressive KT encoding

0 0 8 1 0 7 1 2 1 8 4 7 15 1 5 17 13 4 12 12

$\tilde{m} \curvearrowright$ Elias encoding

6 11 16

Return

Light-tailed envelopes

Decomposing redundancy of AC-code

Decomposing pointwise redundancy

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

Establishing main theorem in (BBG, 2014)

$\hookrightarrow$

- (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 $R^+(\Lambda_f^n)$.

Return

Light-tailed envelopes

Stochastic behavior of M_n

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

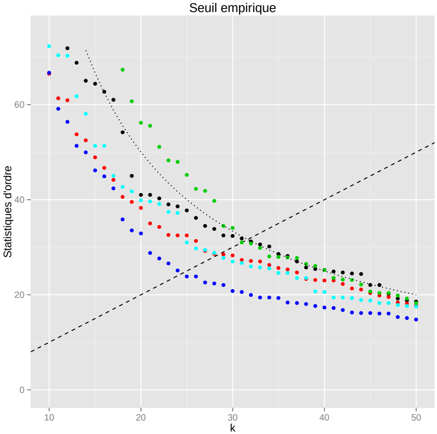
$$\begin{aligned} \mathbb{E}M_n &\leq U_c(en) + 1 \\ \mathbb{E}[M_n \log M_n] &\leq [U_c(en) + 1] \log[U_c(en) + 1] + 2/b^2. \end{aligned}$$

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 → sharp tail and moment bounds for order statistics (B. & Thomas, 2012)

Return

Weak adaptivity of ETAC encoding



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

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

- $\frac{M_n}{m_n} \xrightarrow{P} 1$.
- $\frac{X_{M_n,n}}{m_C(n)} \xrightarrow{P} 1$.

M_n is self-bounded

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

Return