On Large Deviations Tradeoffs Between Code–Length and Distortion in Certain Lossy Source Coding Problems

Neri Merhav, Technion

Based on joint works with Tsachy Weissman and with Ioannis Kontoyiannis.

MSRI Workshop on Information Theory
Berkeley, CA, February–March, 2002

Introduction and Problem Description

Consider the R-D problem for a DMS P, emitting $X_1, X_2, \ldots$ in a finite alphabet $\mathcal{X}$, with a reconstruction alphabet $\hat{\mathcal{X}}$, and a distortion measure ρ.

Marton (1974):

$$\min \Pr\{\rho(X^n, \hat{X}^n) > nD\} \;\; \text{s.t. } |\text{codebook}| \le 2^{nR}$$

Derived the fastest exponential decay rate:

$$F(D,R) = \min\{D(Q\|P) : \quad R_Q(D) \ge R\}.$$

Other work: Blahut (‘74,‘76,‘87), Omura (‘73,‘75), Csiszár (‘82), Kanlis & Narayan (‘96), Arikan & Merhav (‘98), Kontoyiannis (‘99), Haroutunian & Haroutunian (‘00), Tuncel & Rose (‘01).
Lossless case: Jelinek (‘68), Wyner (‘74), Humblet (‘81), Davisson, Longo & Sgarro (‘81), Anantharam (‘90), Merhav (‘91), Merhav & Neuhoff (‘92), Arikan (‘96), Han (‘00).

Purpose: Treat rate and distortion more symmetrically – best tradeoff between the exponents of

$$\Pr\{\rho(X^n, \hat{X}^n) > nD\} \text{ and } \Pr\{L(\hat{X}^n) > nR\}$$

in this and in other problems of lossy compression.

Introduction & Problem Description (Cont'd)

Specifically, minimize:

$$\Pr\{\rho(X^n, \hat{X}^n) > nD\} \text{ s.t. } \Pr\{L(\hat{X}^n) > nR\} \le e^{-\lambda n}.$$

Denote the best achievable exponent by $I(D, R, \lambda)$.

Optimal code (nonuniversal, as opposed to Marton):

$$L^*(\hat{X}^n) = \begin{cases} nR & D(Q||P) < \lambda \\ n \log|\mathcal{X}| & \text{otherwise} \end{cases}$$

Two cases:

1. $D(Q\|P) \le \lambda \Rightarrow R_Q(D) < R$, i.e., $\lambda < F(D, R)$.

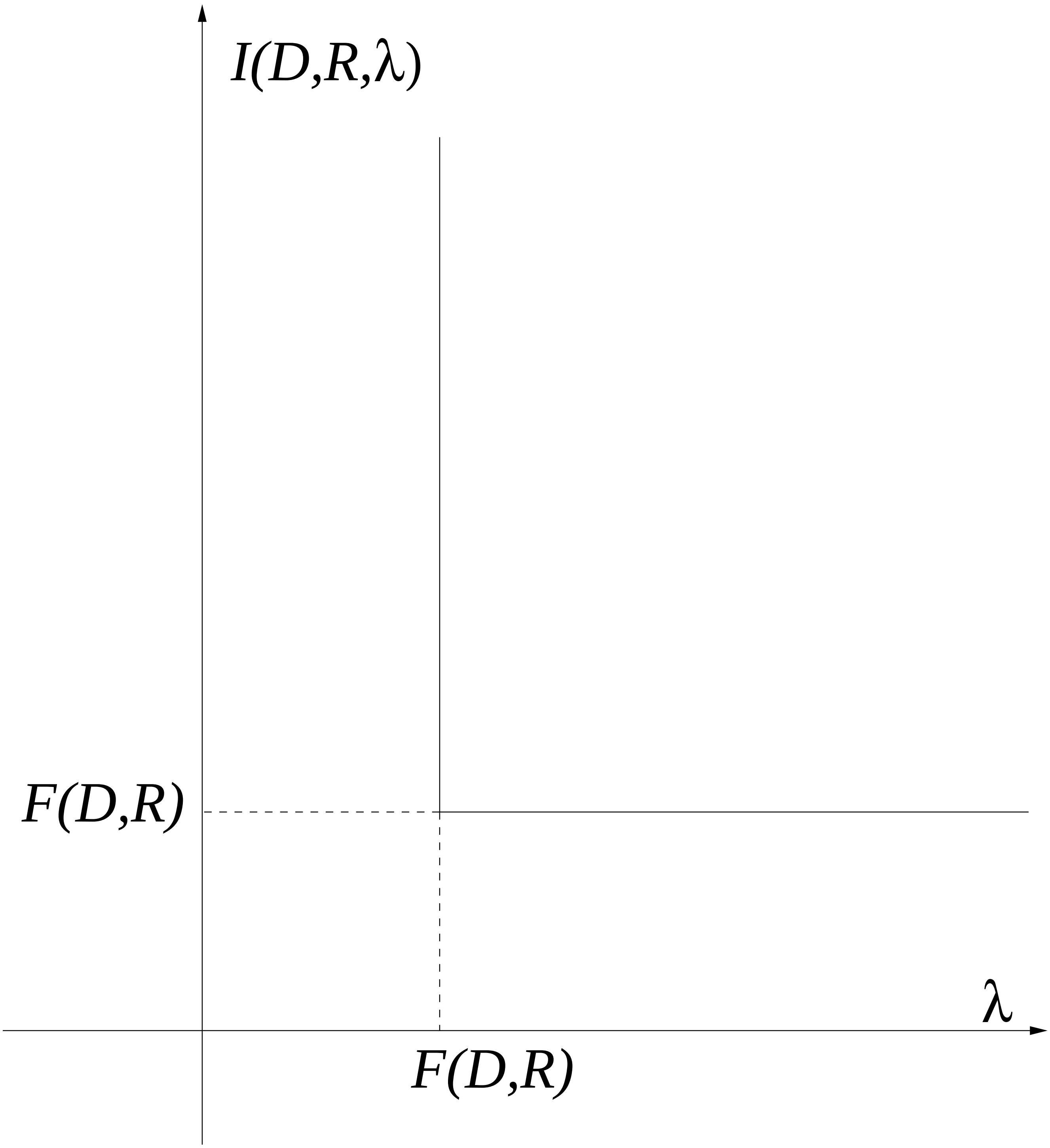
2. Complementary to 1.

In **Case 1**, all T_Q which don't allow $> nR$ bits are coverable by nD–spheres (type–covering). Others can be coded even losslessly $\Rightarrow I(D, R, \lambda) = \infty$.

In **Case 2**, all X^n with $R_Q(D) > R$ are distorted $> nD$, so $I(D, R, \lambda) = F(D, R)$.

Thus,

$$I(D, R, \lambda) = \begin{cases} \infty & \lambda < F(D, R) \\ F(D, R) & \lambda \ge F(D, R) \end{cases}$$



Abrupt transition in the tradeoff between exponents: No point is better than either fixed rate or fixed distortion.

Noisy Sources

P_{XY} – DMS of i.i.d. pairs $\{(X_i, Y_i)\}$.
$\{X_i\}$ – clean source, $\{Y_i\}$ noisy version fed to the encoder.

Problem:

$$\min \Pr\{\rho(X^n, \hat{X}^n) > nD\}$$

$$\text{s.t.} \;\; \Pr\{L(\hat{X}^n) > nR\} \le e^{-\lambda n}.$$

Denote the minimum by $G_n(D, R, \lambda)$.

Comments:

◇ We expect exponent $< \infty$ due to the noise.

◇ It is not clear that the NN encoding rule still applies.

Theorem

$$\begin{aligned} I(D, R-0, \lambda+0) &\le \liminf_{n\to\infty}\left[-\frac{1}{n}\log G_n(D,R,\lambda)\right] \\ &\le \limsup_{n\to\infty}\left[-\frac{1}{n}\log G_n(D,R,\lambda)\right] \\ &\le I(D, R+0, \lambda-0) \end{aligned}$$

where

$$I(D,R,\lambda) = \min\{\inf_{Q\in\mathcal{H}} A(Q,\infty,D), \inf_{Q\in\mathcal{H}^c} A(Q,R,D)\},$$

$$\mathcal{H} = \{Q: \ D(Q\|P_Y) \geq \lambda\},$$

$$A(Q,R,D) = D(Q\|P_Y) + \sup_{W:\mathcal{Y}\to\hat{\mathcal{X}}: \ I(Q,W)\leq R} F_0(Q\times W, D),$$

and

$$F_0(Q\times W, D) = \inf D(V\|P_{X|Y}|Q\times W),$$

the infimum being over $V : \hat{\mathcal{X}} \times \mathcal{Y} \to \mathcal{X}$ s.t.

$$E_{Q\times W\times V}\rho(X,\hat{X}) > D.$$

Optimal code: If $D(Q\|P_Y) \geq \lambda$, encode losslessly the optimal estimator of X^n. Otherwise, use a Q-covering code corresponding to $W^* = \text{argmax} F_0$.

Explanation: $I(R,D,\lambda)$ = the dominant between the exponents of the "unimportant" and the "important" types of Y^n. $A(Q,R,D)$ = contribution of T_Q of Y^n, where $D(Q\|P_Y)$ comes from $\Pr\{T_Q\}$ and the 2nd term is the best achievable distortion exponent given T_Q s.t. codelength $\leq nR$ bits.

Comments:

- $\diamond$ $I(D, R+0, \lambda-0) = I(D, R-0, \lambda+0)$ a.e.
- $\diamond$ The previous I is obtained as a special case of $Y = X$.
- $\diamond$ $I = 0$ for $R \leq R^*(D, P_{XY})$, the RDF of the noisy source, i.e., the ordinary RDF of P_Y w.r.t. $\rho'(y, \hat{x}) = E_{XY}\{\rho(X, \hat{X})|Y = y\}$.
- $\diamond$ Easy to extend to the case where correlated SI is available to both encoder and decoder.

Universal Coding

Returning to the noise–free case, suppose now that the DMS P_θ is unknown except for the fact that $\theta \in \Lambda$.

For $\lambda = \infty$, Marton's solution is already universal: use a type covering code for every T_Q. For $\lambda < \infty$, our above solution is not universal as it depends on $D(Q\|P)$.

Problem: Given a function $\lambda(\theta)$,
min $P_\theta\{\rho(X^n, \hat{X}^n) \geq nD\}$, uniformly over Λ,
s.t. $P_\theta\{L(\hat{X}^n) \geq nR\} \leq e^{-n\lambda(\theta)}$ $\forall \theta \in \Lambda$.

Questions:

- ◇ Best attainable distortion exponent =?
- ◇ What's the best coding strategy (independent of θ)?
- ◇ How to choose $\lambda(\cdot)$?
- ◇ How does the geometry of P_Λ and $\lambda(\cdot)$ affect the cost of universality?

Observation: If $D(Q\|P_\theta) \le \lambda(\theta)$ for *some* $\theta \in \Lambda$, one must use $\le nR$ bits, otherwise, the sky is the limit.

Defining $U(Q) = \inf_\theta[D(Q\|P_\theta) - \lambda(\theta)]$, let:

$$L(\hat{X}^n) = \begin{cases} nR & U(Q) \le 0 \quad \text{(distortion} = D_Q(R)) \\ n\log|\mathcal{X}| & U(Q) > 0 \quad \text{(distortion} = 0) \end{cases}$$

where for the 1st line, use a rate-R type–covering code for each T_Q.

Therefore, the best achievable exponent is

$$I^u(D, R, \lambda(\cdot)) = \inf D(Q\|P_\theta)$$

where the infimum is over

$$\{Q: \; U(Q) \le 0, D_Q(R) \ge D\},$$

or, equivalently,

$$\{Q: \; U(Q) \le 0, R_Q(D) \ge R\}.$$

Theorem: If I^u is continuous at D and R, then it is uniformly $\ge$ the distortion exponent of $\forall$ code that meets the rate constraint.

Discussion

If $\lambda(\theta) \geq F_\theta(D,R)$, the Q^* achieving

$$F_\theta(D,R) = \inf\{D(Q\|P_\theta) :\ R_Q(D) \geq R\}$$

gives $D(Q^*\|P_\theta) \leq \lambda(\theta)$, and hence, $U(Q^*) \leq 0$. Thus, $I^u(D,R,\lambda(\cdot)) = F_\theta(D,R)$ for all such θ.

Good news: No price of universality at those θ's.
Bad news: If $\lambda(\theta) = \infty\ \forall\theta$ (Marton's setting), then reducing $\lambda(\theta)$ to any value $> F_\theta(D,R)$ doesn't improve the distortion exponent.

For θ with $\lambda(\theta) < F_\theta(D,R)$, the price of universality $= \infty$: While $I(D,R,\lambda(\theta)) = \infty$, $I^u(D,R,\lambda(\cdot))$ can be $< \infty$. The former $= \min_\emptyset D(Q\|P_\theta)$, whereas

$$\{Q :\ U(Q) \leq 0,\ R_Q(D) \geq R\}$$

can be $\neq \emptyset$.

Choose $\lambda(\cdot)$ s.t. $I^u = \infty$ whenever possible. This happens if $U(Q) > 0\ \forall Q :\ R_Q(D) \geq R$, i.e.,

Discussion (Cont'd)

$$\lambda(\theta) < \lambda_0(\theta) \triangleq \inf_{Q:\ R_Q(D) \geq R} D(Q \| P_\theta).$$

But $\lambda_0 > 0$ if $\{Q :\ R_Q(D) \geq R\}$ is separated from P_Λ $\Rightarrow$ either $I^u = \infty\ \forall\theta$ or $I^u < \infty\ \forall\theta$. A reasonable choice:

$$\lambda(\theta) = \alpha\lambda_0(\theta) \quad \alpha \in (0, 1).$$

The dichotomy according to the sign of $U(Q)$ is intimately related to a universal decision rule for composite hypothesis testing (Levitan & Merhav, 2000).

Zero–Delay Finite–Memory (ZDFM) Codes

Consider now a ZDFM code, where each

$$\hat{X}_t = f_t(X_{t-k+1}^t), \quad \hat{X}_t \in \hat{\mathcal{X}}$$

is compressed individually within $L_t(\hat{X}_t|\hat{X}_{t-k+1}^{t-1})$ bits. $f_t(\cdot)$ is a T-V reproduction function and $k =$ the memory parameter.

We begin with fixed–rate codes, where

$$L_t(\hat{X}_t|\hat{X}_{t-k+1}^{t-1}) = \log|\hat{\mathcal{X}}_t| = R_t, \quad \hat{\mathcal{X}}_t \subseteq \hat{\mathcal{X}}$$

independently of $\hat{X}_{t-k+1}^t$, and where it is assumed that $|\hat{\mathcal{X}}_t|$ doesn't depend on the past, although $\hat{\mathcal{X}}_t$ itself may do.

Problem:

$$\min \Pr\{\sum_{t=1}^n \rho(X_t, \hat{X}_t) \geq nD\} \;\; \text{s.t.} \;\; \sum_{t=1}^n R_t \leq nR.$$

Earlier work on ZDFM (and related) codes: Gray ('75), Lloyd ('77), Berger & Lau ('77), Ericson ('79), Piret ('79), Gaarder & Slepian ('79,'82), Gilbert & Neuhoff ('79), Neuhoff & Gilbert ('82), Linder & Lugosi ('00), Linder & Zamir ('01), Weissman & Merhav ('01).

Let $\mathcal{G} = \{g_1, \ldots, g_r\}$, $g_i : \mathcal{X} \to \hat{\mathcal{X}}$, denote the set of all $r = |\hat{\mathcal{X}}|^{|\mathcal{X}|}$ memoryless reproduction functions $\mathcal{X} \to \hat{\mathcal{X}}$ and let

$$\Theta_R = \{\theta : \sum_{s=1}^{r} \theta_s \log \|g_s\| \leq R\}.$$

Define

$$\phi(D, \theta) = \sup_{\xi \geq 0} \left[\xi D - \sum_s \theta_s \ln E e^{\xi \rho(X, g_s(X))} \right],$$

and

$$F(D, R) = \sup_{\theta \in \Theta_R} \phi(D, \theta).$$

Theorem: Best distortion exponent $= F(D, R)$.

Discussion

- ◇ $F(D,R)$ – attained by time–sharing among the memoryless $\{g_s\}$ with relative frequencies according to θ^*.
- ◇ $\theta_s^* > 0$ on at most two $\{g_s\}$. Similar to Neuhoff & Gilbert ('82) (and Linder & Zamir ('01)) for general causal codes.
- ◇ The assumption of fixed k is crucial for an LDP (though not for the MGF).
- ◇ An alternative, "information-theoretic" expression:

$$F(D,R) = \sup_{\theta\in\Theta_R} \inf_{\{Q_s\}} \sum_s \theta_s D(Q_s\|P_s),$$

where P_s = the PMF of $Y_s \stackrel{\Delta}{=} \rho(X, g_s(X))$ and the inf is over all $\{Q_s\}$ s.t. $\sum_s \theta_s E_{Q_s} Y_s \geq D$ (in partial analogy to Marton's exponent).

- ◇ In complete duality, the fixed–distortion case gives: $G(D,R) = \sup_\theta \gamma(R,\theta)$, where

$$\gamma(R,\theta) = \sup_{\xi\geq 0}\left[\xi R - \sum_s \theta_s \ln E e^{\xi L_s(g_s(X))}\right],$$

where now s is an index of a combination (L,g).

Proof Idea – "Onion Peeling" (Stiglitz '67)

Divide the n–block into sub-blocks of length q (including gaps of k units). The cumulative distortion within a sub-block is an AVS.

The Chernoff bound of $\Pr\{\sum_t \rho(X_t, \hat{X}_t) \geq nD\}$ is based on the MGF:

$$\sum_{x_1} P(x_1) e^{\xi\rho(x_1, f_1(x_{2-k}^1))} \times$$
$$\sum_{x_2} P(x_2) e^{\xi\rho(x_2, f_2(x_{3-k}^2))} \times$$
$$\ldots \times$$
$$\sum_{x_q} P(x_q) e^{\xi\rho(x_q, f_q(x_{q-k+1}^q))}.$$

In the the last line, x_{q-k+1}^{q-1} just an "index" of a particular $f_q \Rightarrow$ cannot be

$$< m(R_q) \triangleq \min_{g:\ \log\|g\| \leq R_q} \sum_x P(x) e^{\xi\rho(x, g(x))}.$$

Having factored out the last line, we repeat this argument for the 2nd to the last line, and so on. Finally, we have a lower bound $\Pi_{t=1}^q m(R_t)$, achieved by a sequence of memoryless reproduction functions.

Comment: For Markov sources, the MGF is minimized by "Markov" encoders of the same order (as opposed to Neuhoff & Gilbert).

Rate–Distortion Lagrangian Criterion

Consider the minimization of

$$\Pr\left\{\sum_{t=1}^{n} L_t(\hat{X}_t|\hat{X}_{t-k+1}^{t-1}) + \lambda \sum_{t=1}^{n} \rho(X_t, \hat{X}_t) \geq nR_0\right\}.$$

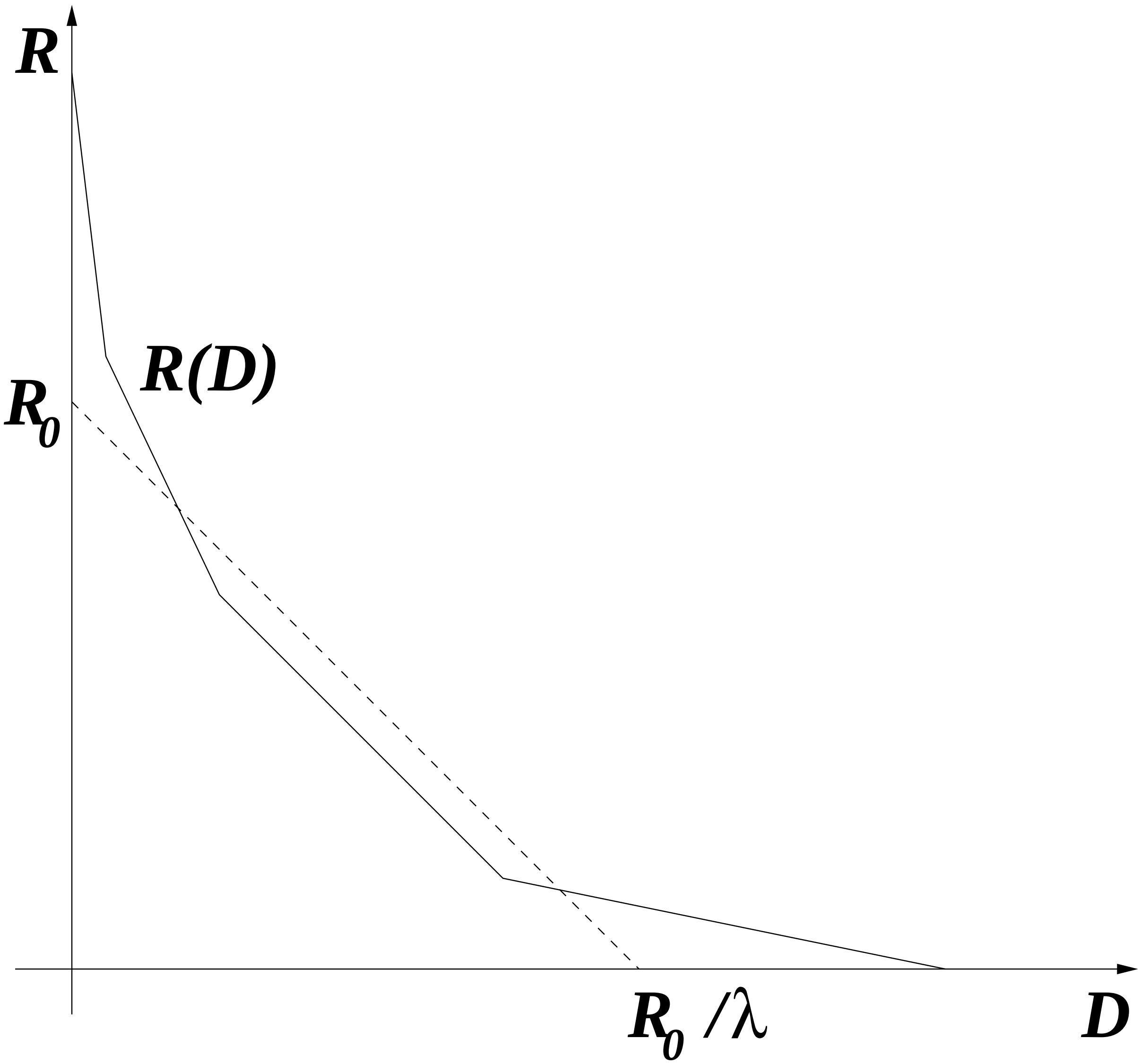
Motivation: This is the probability that the actual R–D working point falls above the line $R = R_0 - \lambda D$. Choose R_0 and λ s.t. this line is parallel and slightly above a certain linear segment of $R(D)$.

In other words, this is like

$$\Pr\left\{\sum_{t=1}^{n} L_t(\hat{X}_t|\hat{X}_{t-k+1}^{t-1}) > n\left[R\left(\frac{1}{n}\sum_{t=1}^{n} \rho(X_t, \hat{X}_t)\right) + \Delta\right]\right\}$$

in the region of a given slope.

In ordinary block codes, the best exponent is: $\inf D(Q\|P)$ over $\{Q : \inf_D[R(D,Q) + \lambda D] \geq R_0\}$.


R
R(D)
R_0
R_0/λ
D

Define

$$H(\lambda, R_0, \theta) = \sup_{\xi\geq 0}\Big[\xi R_0 - \sum_s \theta_s \ln E \exp\{\xi[L_s(g_s(X)) + \lambda\rho(X, g_s(X))]\}\Big].$$

Theorem: Best exponent $= H(\lambda, R_0) \stackrel{\Delta}{=} \sup_\theta H(\lambda, R_0, \theta)$.

Comment 1: As $H(\lambda, R_0, \theta)$ is affine in θ and there are no constraints on θ, the optimum θ^* puts all its mass on a single memoryless encoder (L_s, g_s), i.e., *no need for time–sharing.*

Comment 2: Easy to extend for the characterization of the probability of

$$\{L(\hat{X}^n)+\lambda\rho(X^n, \hat{X}^n) \geq nR_0,\ L(\hat{X}^n)+\lambda'\rho(X^n, \hat{X}^n) \geq nR_0'\},$$

corresponding, e.g., to two adjacent linear segments of $R(D)$.

Summary and Conclusion

◇ We have introduced new criteria for LD tradeoffs between rate and distortion: A Neyman–Pearson–like criterion (for block codes) and a Lagrange–type criterion (for ZDFM codes).

◇ We have characterized L-D tradeoffs of ordinary block codes, block codes for noisy sources (with SI), universal codes, ZDFM codes with fixed rate, fixed distortion, and fixed slope.

◇ For universal block codes, we have characterized the price of universality and pointed out the relationship with universal composite–hypothesis testing.

Summary and Conclusion (Cont'd)

◇ In all cases, exponents are characterized by single-letter expressions. In the ZDFM case, these stem from the fact that the best codes are memoryless ones.

◇ Techniques: For block codes – the type covering lemma; For ZDFM codes – "onion–peeling".

◇ "Onion–peeling" can be useful for other problems of causal systems, e.g., causal joint source–channel codes:

$$\sum_{u_t, x_t, y_t, v_t} P(u_t) P_t^e(x_t | u_{t-k+1}^t) P(y_t | x_t) P_t^d(v_t | y_{t-k+1}^t) e^{\xi \rho(u_t, v_t)}$$

is minimized by $P^e(x|u) = \delta(f - f(u))$ and $P^d(v|y) = \delta(v - g(y))$.

Future Research

Block Codes:

◇ Extension of the universal setting to the case of a noisy source. Difficulty: what is the best scheme within each type? In the non-universal noisy case, it depends on the active source. Universality is not always achievable even in the expectation sense (Dembo & Weissman, 2001).

◇ Error exponents for the Wyner–Ziv problem.

ZDFM Codes:

◇ ZD infinite–memory codes.

◇ Neyman–Pearson-like tradeoffs.

◇ Codes with finite anticipation (delay).

◇ More general sources: Markov sources (Sabbag, 2002).

◇ Universal coding.