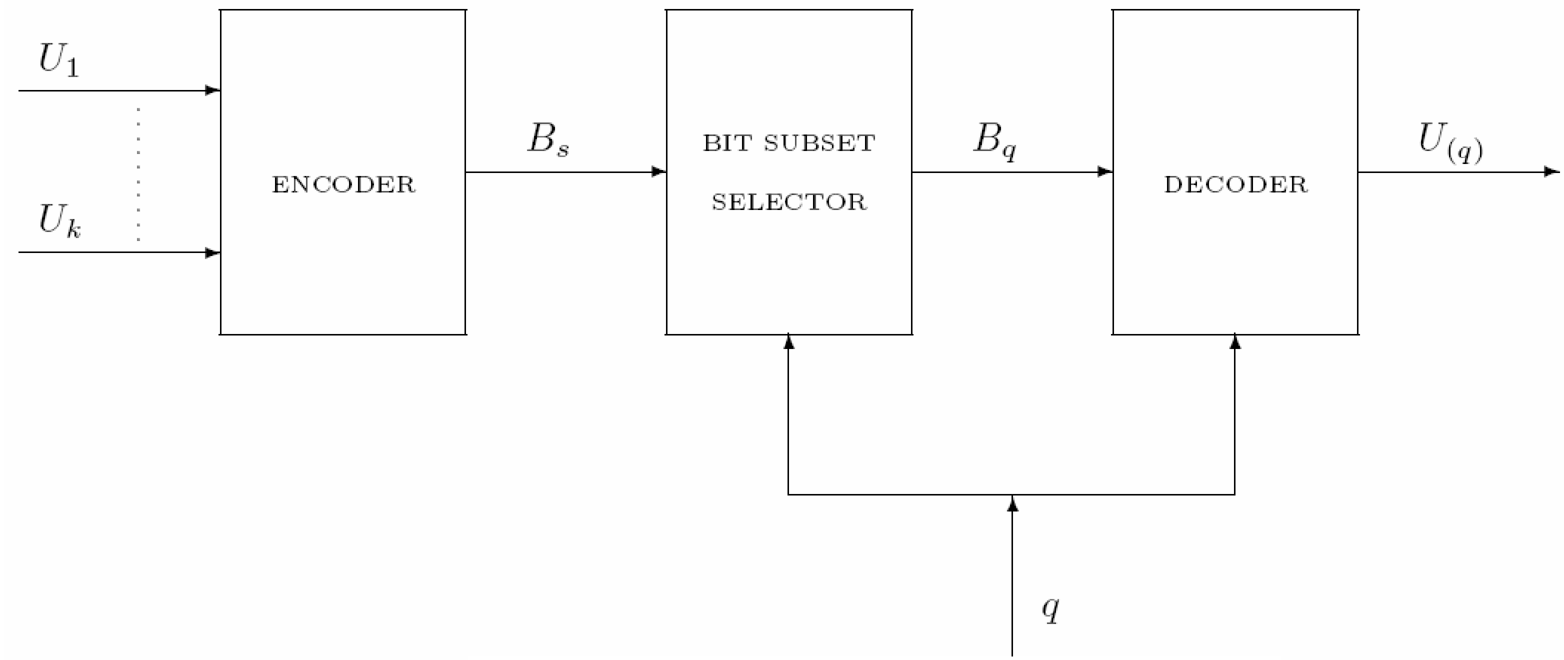
Correlated Source Coding for Fusion Storage and Selective Retrieval

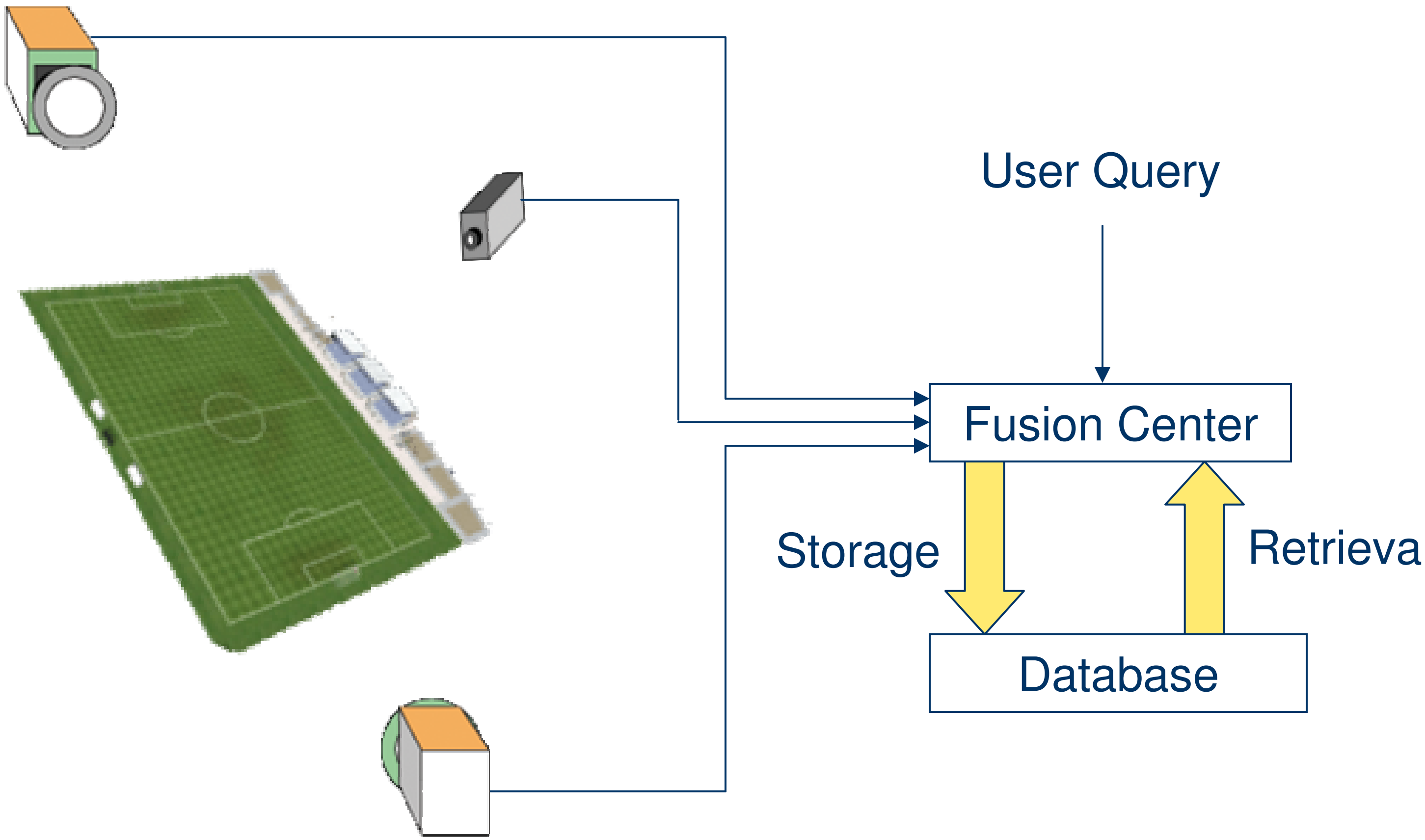
J. Nayak, S. Ramaswamy and K. Rose
University of California, Santa Barbara

Problem Setup



- Datastreams: $\{U_{ki}\}_{i=1}^{\infty}, k = 1, \ldots, K$
- Query: $q = \{k_1, \ldots, k_q\} \subset \{1, \ldots, K\}$

Motivation



Storage vs. Retrieval Tradeoff

- Possibility 1: Compress all streams together
  - Minimal storage cost
  - High retrieval cost
- Possibility 2: Store descriptions of every subset of streams separately
  - Minimal retrieval cost
  - High storage cost

Problem Statement

- R_s bits per instant
- Query distribution given

$$P(q), q \subset \{1, \ldots, K\}$$

- Minimize average retrieval rate

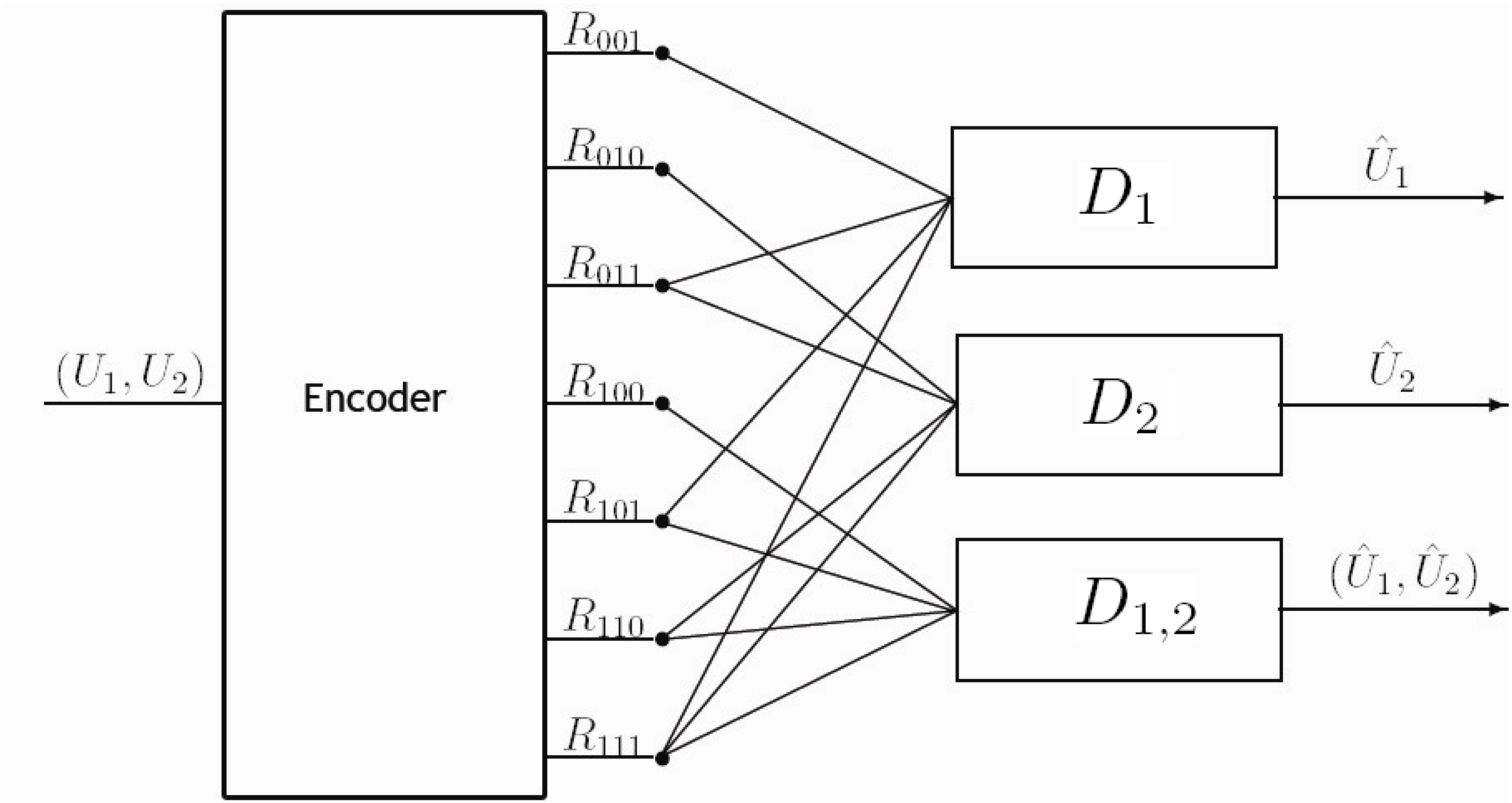
$$\bar{R}_r = \sum_q P(q) R_r(q)$$

such that $P[U_{(q)} \neq \hat{U}_{(q)}] \rightarrow 0, \forall q$

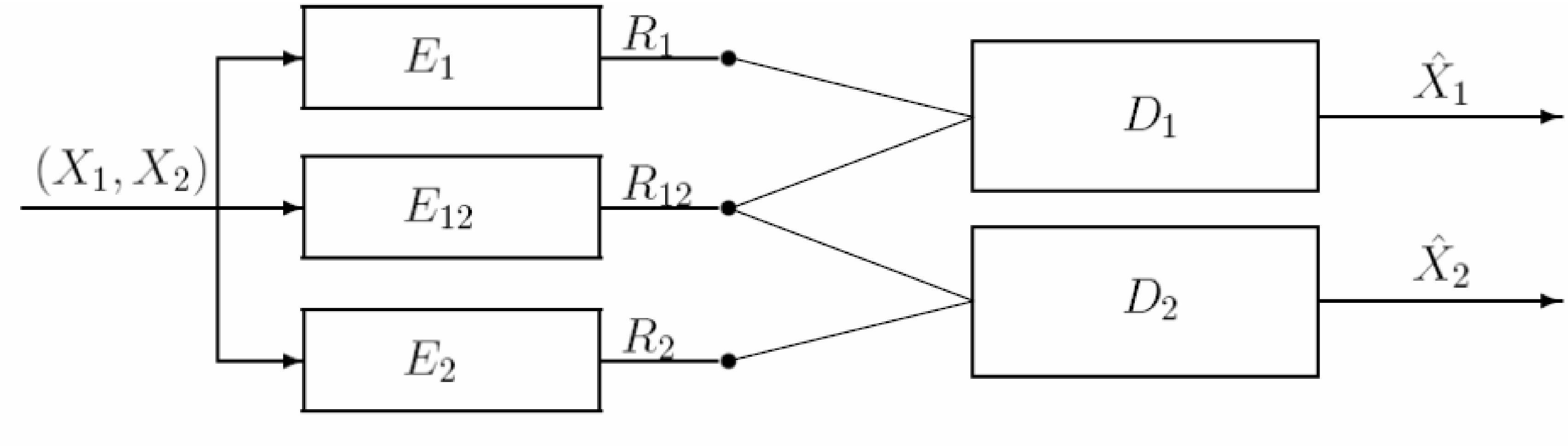
General framework

- Every bit associated with some subset of the queries
- Group together bits associated with same set of queries
- Each group corresponds to a constituent encoder

General Framework, K = 2

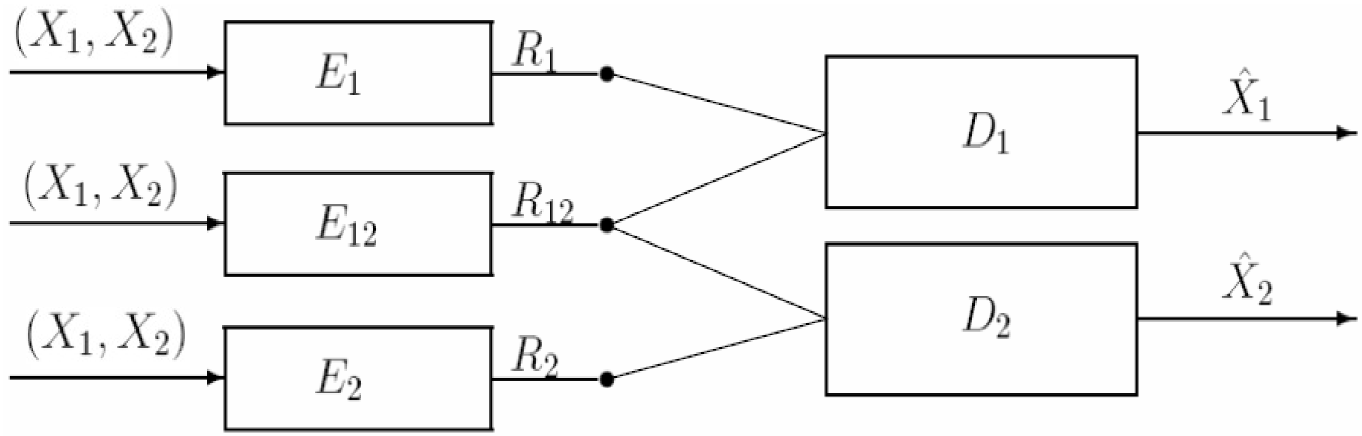


Shared Descriptions



- Scenario considered by Gray & Wyner
- In general, D decoders, 2^D-1 encoders

Shared Descriptions: Rate Region



- Asymptotic rate does not change if we assume encoders are independent.
  - Non-single letter characterization of rate region using results from multiterminal source coding (Han and Kobayashi '80).

Shared Descriptions: Rate Region

- Equivalent to multiple descriptions with only a subset of decoders active
  - Known achievable rate regions for MD can be extended
    - Retain only a subset of the auxiliary variables
    - Binning schemes used to enlarge rate region
  - Methods can be used to analyze cases where distortion is allowed

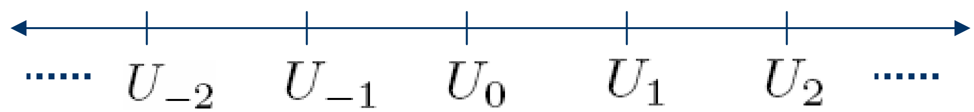
Storage vs. Retrieval Tradeoff

- Storage-retrieval tradeoff completely characterized by rate region

$$\bar{R}_r^*(R_s) = \min_{(R_{(\Sigma)})\in\mathcal{R}^*}\{\sum_{q\subset\mathcal{K}} P(q)\sum_{i\in\Sigma_q} R_i : \sum_{i\in\Sigma} R_i \leq R_s\}$$

- Rate region characterizes achievable tradeoff even when number of encoders is constrained
- Single letter achievable rate region gives bounds on minimum retrieval rate

A Toy Example



$$U_k = \{\tilde{U}_{k',d}, d \in [0, D], |k - k'| \le d\}$$

- Information centered at k' and spread to a distance d : $\tilde{U}_{k',d}$
- All $\tilde{U}_{k',d}$ are mutually independent
- $H(\tilde{U}_{k,d}) = \xi(d), \forall k$
- Query model
  1. l well separated streams : P_{w}
  2. l consecutive streams : $P_{\mathrm{c}} = 1 - P_{\mathrm{w}}$

$|k - k'| \ge 2D + 1$

A Toy Example (contd.)

- Trivial strategy: store each stream separately
- "Optimal" strategy: store each $\tilde{U}_{k',d}$ once
  - Optimal storage cost
  - Optimal retrieval cost
- If "correlation distance" $\gamma \triangleq \frac{\sum_{d=0}^{D} d\xi(d)}{\sum_{d=0}^{D} \xi(d)}$

$$\frac{R_{r,\mathrm{triv}}}{R_{r,\mathrm{min}}} = \frac{2\gamma + 1}{P_{\mathrm{w}}(2\gamma + 1) + P_{\mathrm{c}}(\frac{2\gamma}{l} + 1)}$$

$$\frac{R_{s,\mathrm{triv}}}{R_{s,\mathrm{min}}} = 2\gamma + 1$$

Constrained Encoder: Example

- K = 2, L = 2, $R_s = H(U_1, U_2)$
- Optimal encoding

$$\bar{R}(H(U_1, U_2), 2) = \bar{R}_{\min} + \min[P(\{1\})H(U_2|U_1), P(\{2\})H(U_1|U_2)]$$

$$\bar{R}_{r\min} \triangleq \sum_{q \subset \mathcal{K}} P(q)H(\mathbf{U}_q)$$

To sum up…

- Problem: tradeoff between storage and retrieval costs in correlated source coding
  - Developed a non-single letter characterization of the rate region that determines the tradeoff
  - Also developed a single letter achievable rate region
  - Observed that there can be significant gains if there is enough correlation between the sources