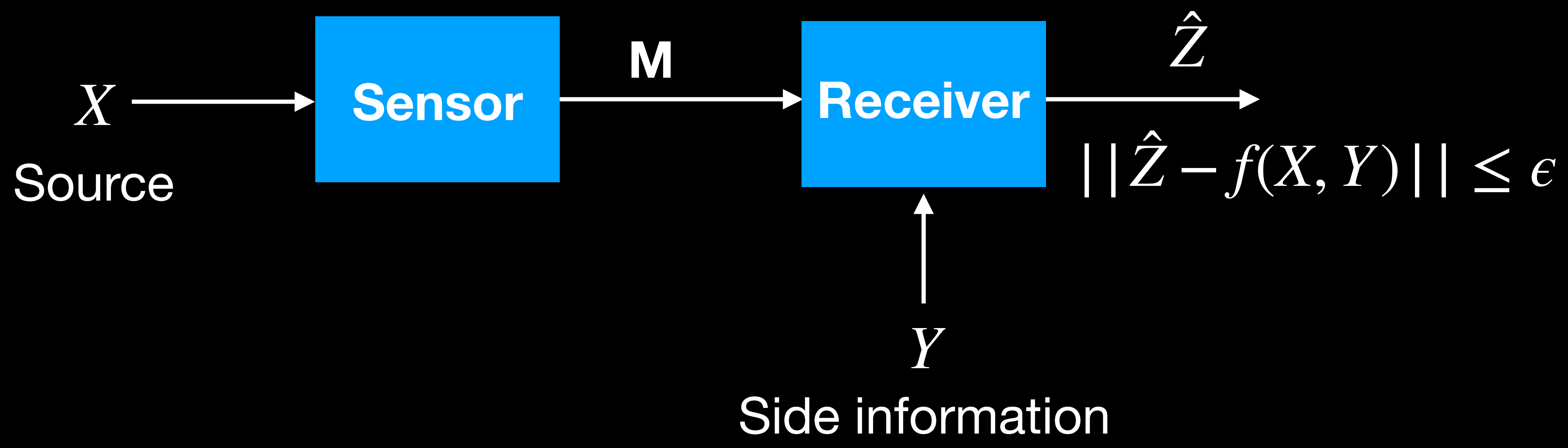
Functional Epsilon Entropy

Sourya Basu: CSL, ECE Dept., University of Illinois at Urbana-Champaign
Daewon Seo: ECE Dept., University of Wisconsin-Madison
Lav Varshney: CSL, ECE Dept., University of Illinois at Urbana-Champaign; Salesforce Research

Motivation

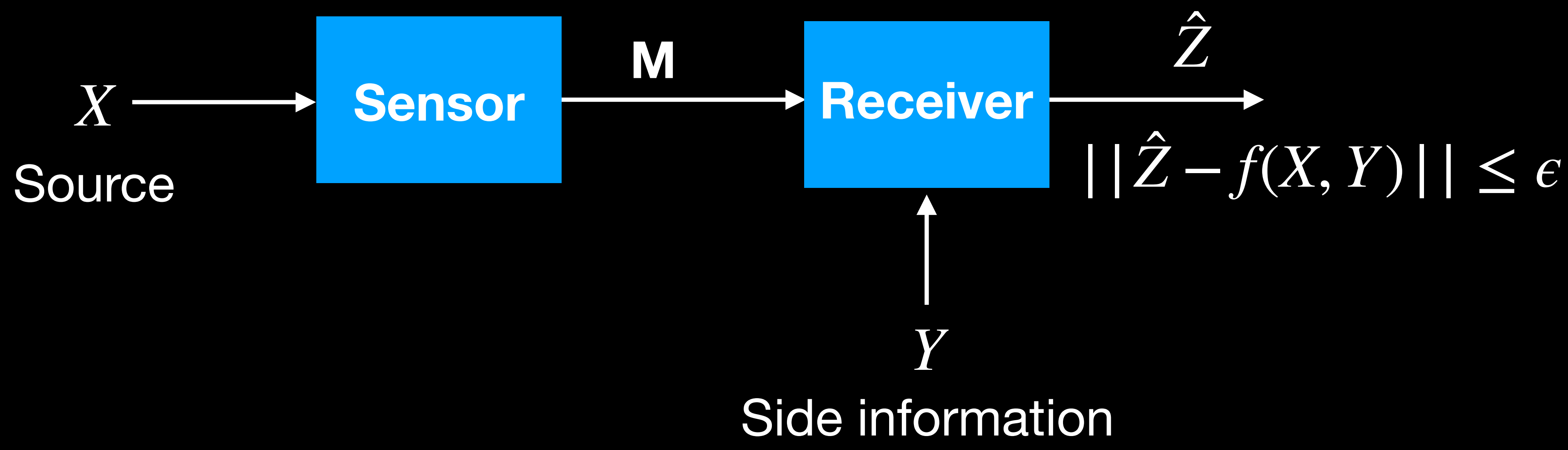
- Lossy functional source coding



- Sensor observes X
- Receiver observes Y
- Both agents want the receiver to compute $f(X, Y)$ as $\hat{Z}$ such that $||\hat{Z} - f(X, Y)|| \leq \epsilon$

Motivation

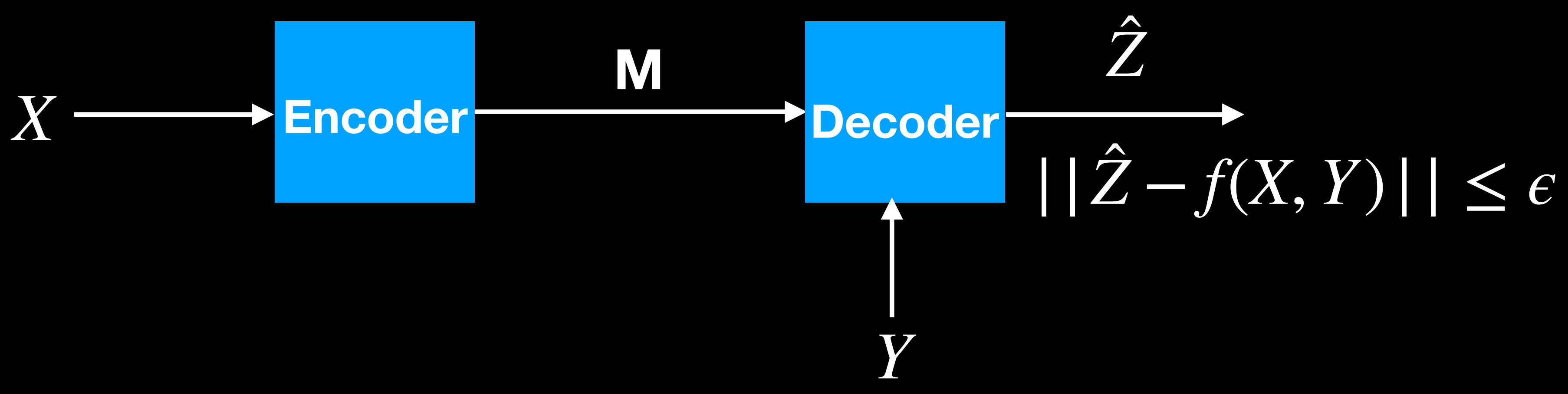
- Lossy functional source coding



Applications:

- *Sensor networks*: The receiver only needs to compute some function of the received data
- *Big data and bioinformatics*: Often only a known function of data is of interest and the observed data is not of primary importance. Observed data is often huge

Previous work: optimal codes



Yamamoto's solution [1]

$$R_{min} = \min_{\substack{U - X - Y \\ p \in \mathscr{P}(0)}} I(U;X|Y),$$

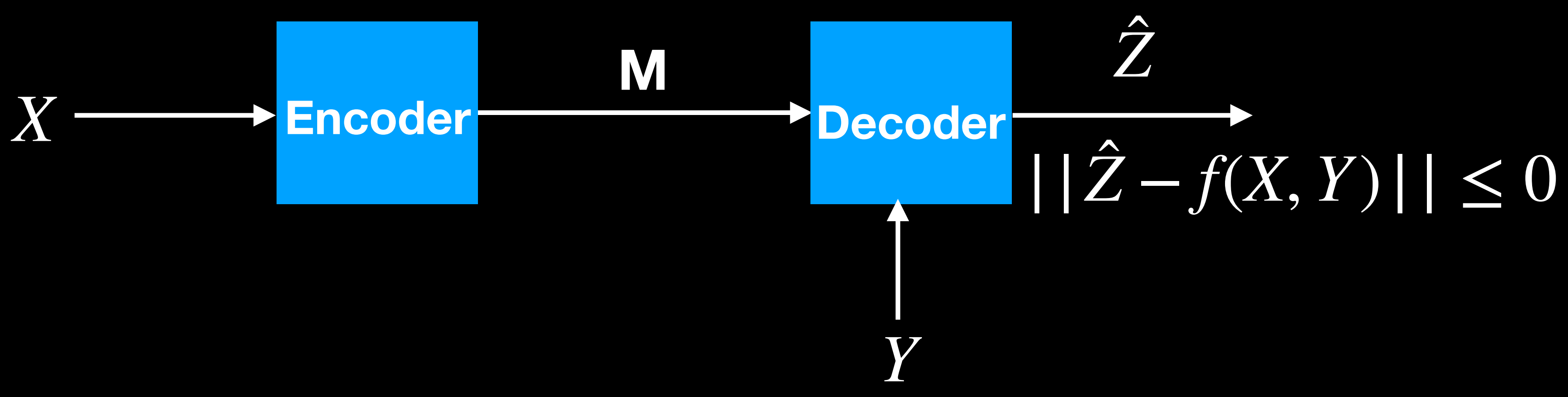
where $\mathscr{P}(0)$ is the set of all $p(u|x)$, such that there exists a $g : \mathscr{U} \times \mathscr{Y} \mapsto \mathscr{Z}$ satisfying $E[1_{||f(X,Y)-g(U,Y)||>D}] \leq 0$.

Salient features

- Optimal rates for *both lossless and lossy functional source coding*
- Depends on the existence of g and W which makes it *difficult to design practical codes* that achieve this rate

1. H.Yamamoto,“Wyner-Ziv theory for a general function of the correlated sources,” *IEEE Trans. Inf. Theory*, vol. 28, no. 5, pp. 803–807, Sep. 1982.

Previous work: lossless compression



Lossless case [2]

$$R_{min} = \min_{\substack{W - X - Y \\ X \in W \in \Gamma(G)}} I(W; X | Y)$$

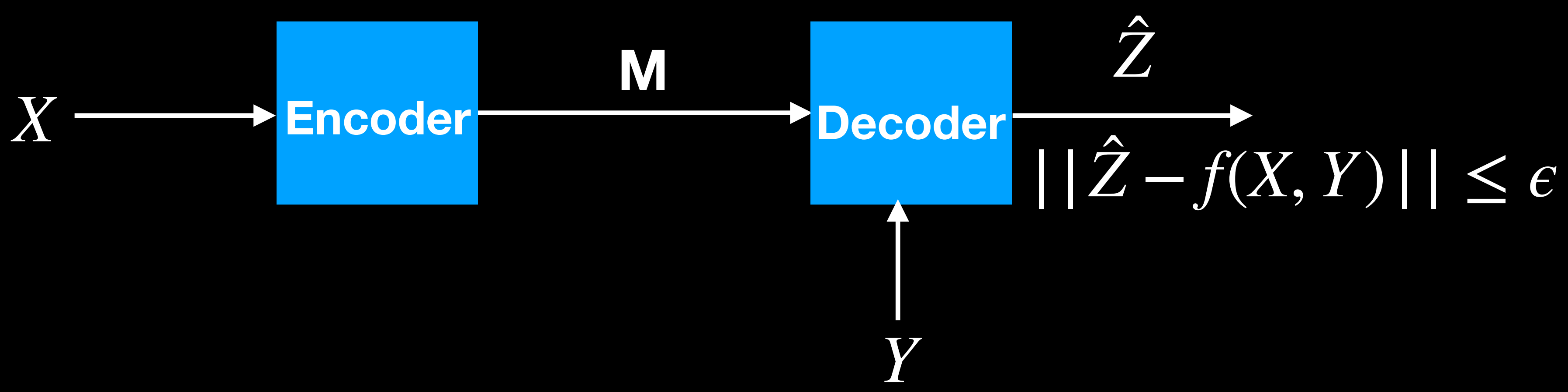
where $\Gamma(G)$ is the set of all hyperedges of a ***characteristic hypergraph*** which can be constructed based on f, X, Y as a part of coding scheme [2].

Salient features

- Optimal codes
- Practical codes which can be implemented
- Works *only for lossless functional compression*

2. A. Orlitsky and J. R. Roche, “Coding for computing,” IEEE Trans. Inf. Theory, vol. 47, no. 3, pp. 903–917, Mar. 2001.

Previous work: maximal distortion



Efficient lossy codes [3]

$$R_{min} \leq \min_{\substack{W - X - Y \\ X \in W \in \Gamma(G_D)}} I(W; X | Y)$$

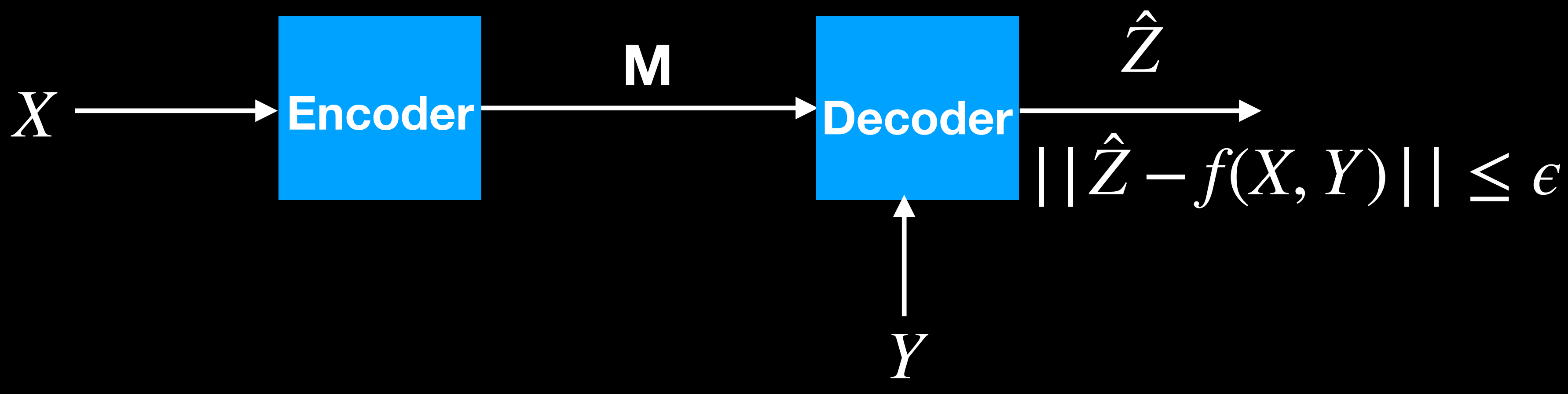
where $\Gamma(G_D)$ is the set of all hyperedges of a ***D-characteristic hypergraph*** which can be constructed based on f, X, Y as a part of coding scheme [1].

Salient features

- Efficient *practical codes*
- Works for *lossy* functional compression under *maximal distortion*
- However, these codes are *suboptimal*

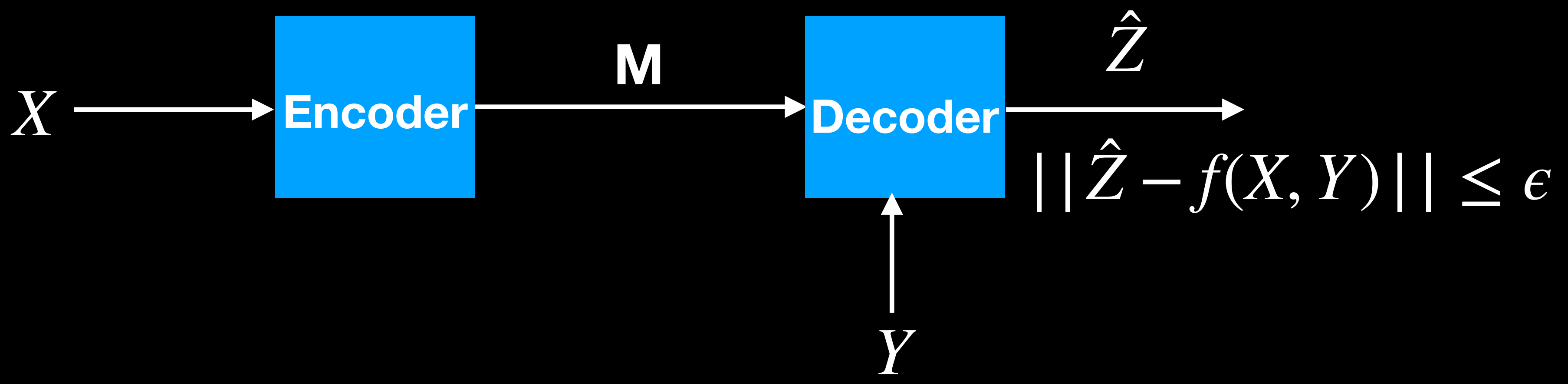
3. V. Doshi, D. Shah, M. Médard, and M. Effros, "Functional compression through graph coloring." *IEEE Transactions on Information Theory* 56.8 (2010): 3901-3917.

Unresolved/open problem



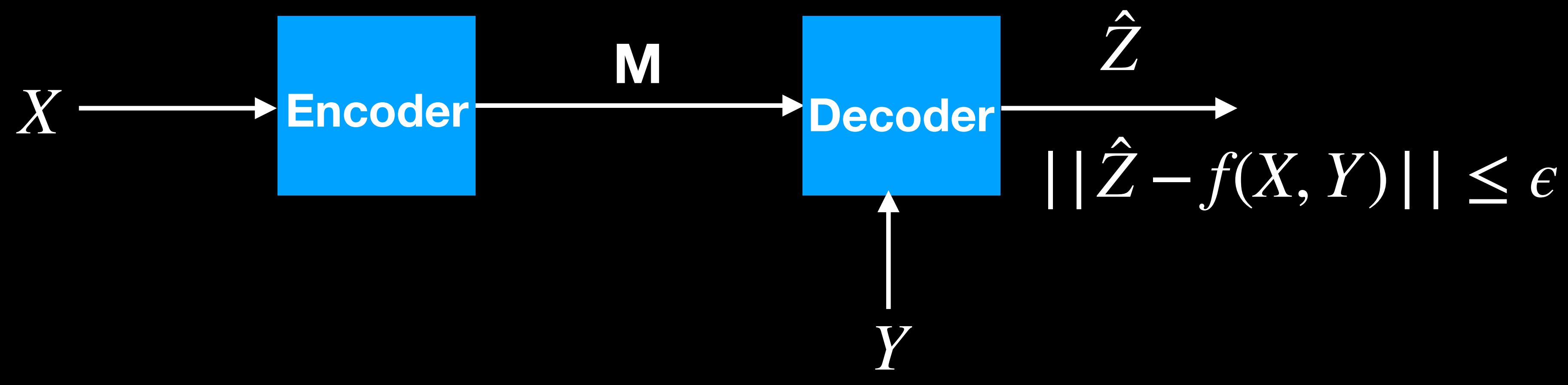
- **No optimal practical code known for lossy functional source coding**

Our contribution: optimal practical codes



- **We close the gap between optimal codes and known practical codes for maximal distortion**

Our contribution: optimal practical codes



Optimal practical lossy codes

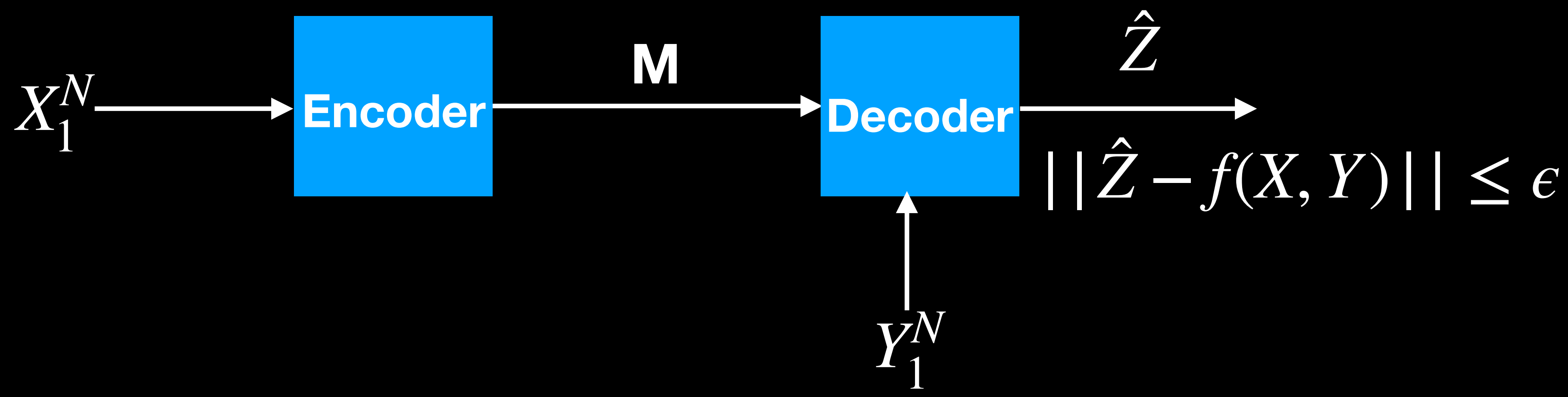
$R_{min} = H_{G_\epsilon}(X|Y)$,
where $H_{G_\epsilon}(X|Y)$ is the *functional epsilon entropy*, defined later

Salient features

- Practical codes
- *Optimal* for *lossy* functional compression under *maximal distortion*
- Based on *better geometric construction* of hypergraphs compared to [3].

3. V. Doshi, D. Shah, M. Médard, and M. Effros, "Functional compression through graph coloring." *IEEE Transactions on Information Theory* 56.8 (2010): 3901-3917.

Problem setting



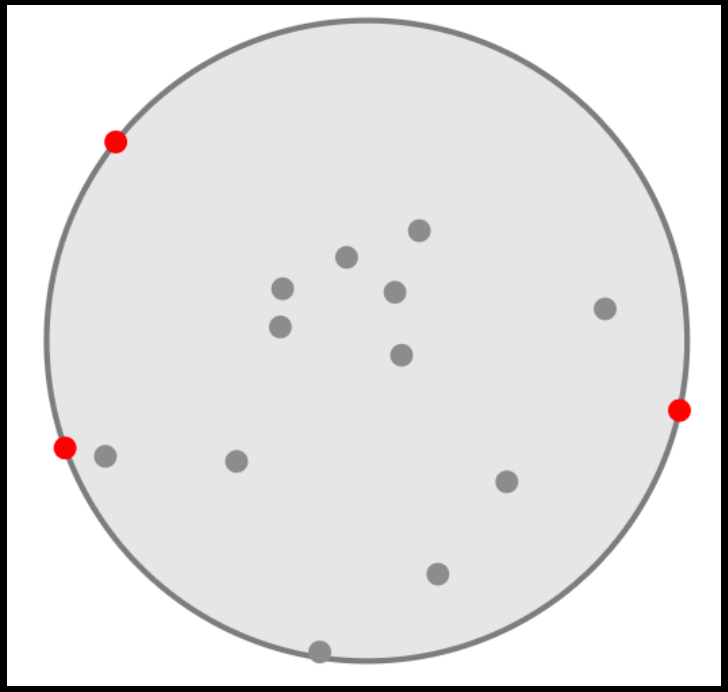
- (X, Y) are distributed as $P_{X,Y}$
- (X_1^N, Y_1^N) are N iid random variables, where $X_i \in \mathscr{X}, Y_i \in \mathscr{Y}$ for $i \in \{1, \ldots, N\}$, and $\mathscr{X}, \mathscr{Y}$ are finite sets.
- Reconstruct $f(X, Y)_1^N$ as $\hat{Z}_1^N$ such that $P_{avg}(X, Y, \hat{Z}) \rightarrow 0$ as $N \rightarrow \infty$, where

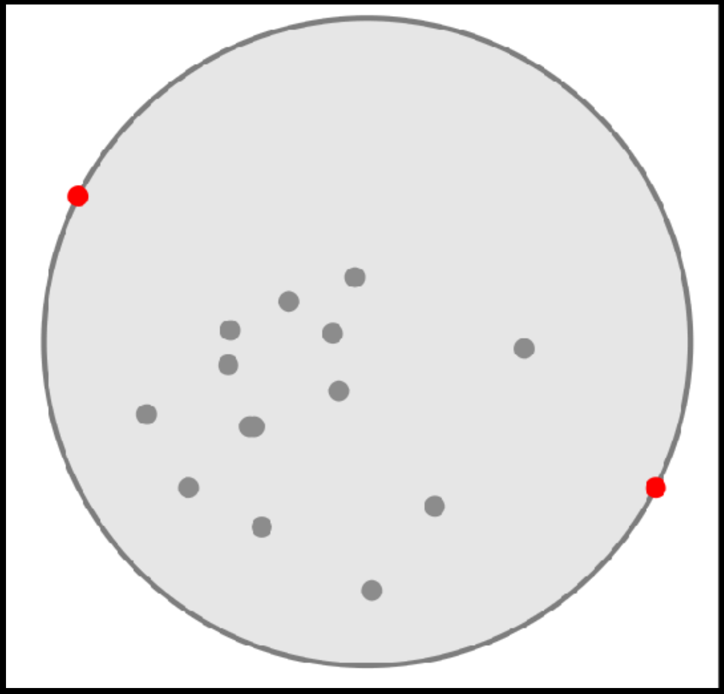
$$P_\epsilon^{avg}(\hat{Z}^N, X^N, Y^N) = \frac{1}{N} \sum_{i=1}^{N} \Pr\left[||\hat{Z}_i - f(X_i, Y_i)|| > \epsilon \right]$$

Smallest enclosing circles

Smallest enclosing circles:

- **For a set of point S, the circle with smallest radius covering all the points in S is called the smallest enclosing circle of S**





Images constructed using https://www.nayuki.io/page/smallest-enclosing-circle

Epsilon characteristic hypergraphs

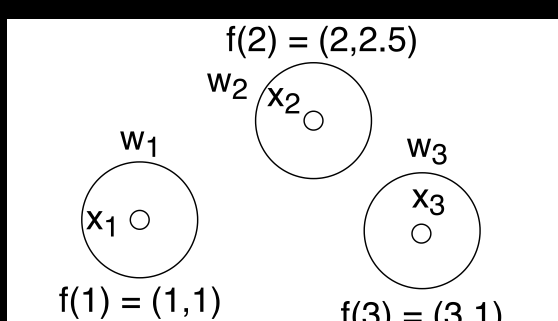
ϵ-characteristic hypergraph:

- **Denoted by G_ϵ when the function f, and random variables X, Y are clear from context**
- **Vertex set of G_ϵ is $\mathcal{X}$**
- **Let $S \subseteq \mathcal{X}$ and $y \in \mathcal{Y}$, let $S_y = \{x : x \in S$ and $p(x, y) > 0\}$**
- **S is a hyperedge in G_ϵ if and only if the radius of the smallest enclosing circle containing the set of points $\{f(x, y) : x \in S_y\}$ is less than or equal to ϵ for all $y \in \mathcal{Y}$**

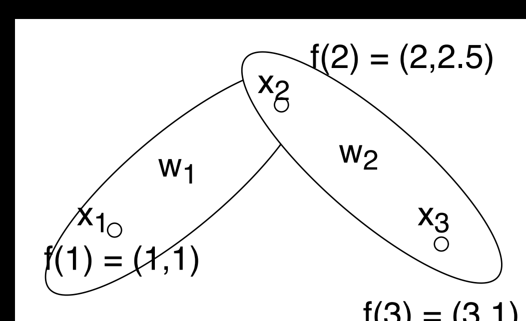
Epsilon characteristic hypergraphs

Example:

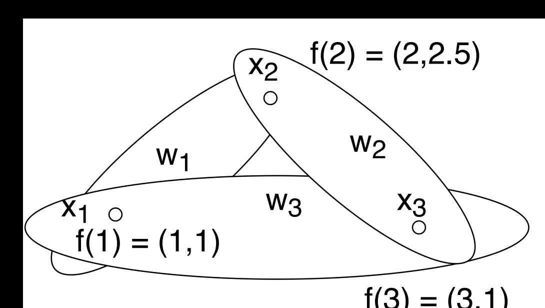
- **Let** $f : \mathscr{X} \mapsto \mathscr{Z}$ **be defined as** $f(1) = (1,1), f(2) = (2,2.5), f(3) = (3,1)$, **where** $\mathscr{X} = \{1,2,3\}$ **and** X **is uniformly distributed**
- **No side information** Y **in this example**



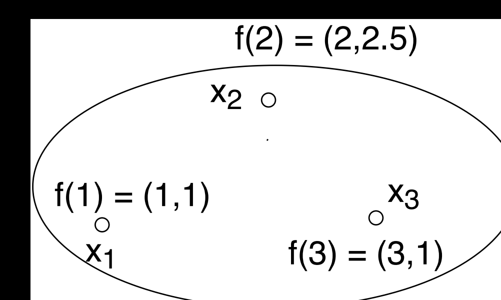
$$0 \leq \epsilon < \frac{\sqrt{13}}{4}$$



$$\frac{\sqrt{13}}{4} \leq \epsilon < 1$$



$$1 \leq \epsilon < \frac{13}{12}$$



$$\frac{13}{12} \leq \epsilon$$

Functional epsilon entropy

Functional ϵ-entropy:

- **Denoted by** $H_{G_\epsilon}(X \mid Y)$

$$H_{G_\epsilon}(X \mid Y) = \min_{\substack{W - X - Y \\ X \in W \in \Gamma(G_\epsilon)}} I(W; X \mid Y),$$

where X **induces a probability distribution over the vertices of the hypergraph** G_ϵ**. The random variable** W **is obtained by defining transition probabilities** $p(w \mid x)$ **over all hyperedges** w **that contain** x**, i.e.** $p(w \mid x) \geq 0$ **for all** $x \in w \in \Gamma(G_\epsilon)$ **and** $\sum_{w \ni x} p(w \mid x) = 1$**.**

Main result: theorem

Theorem:

$$R_{min} = H_{G_\epsilon}(X|Y),$$

where R_{min} is obtained from Yamamoto's codes under maximal distortion

Yamamoto's solution [1]

$$R_{min} = \min_{\substack{U-X-Y \\ p \in \mathcal{P}(0)}} I(U;X|Y),$$

where $\mathcal{P}(0)$ is the set of all $p(u|x)$, such that there exists a $g : \mathcal{U} \times \mathcal{Y} \mapsto \mathcal{Z}$ satisfying $E[1_{||f(X,Y)-g(U,Y)||>D}] \leq 0.$

1. H.Yamamoto,"Wyner-Ziv theory for a general function of the correlated sources," *IEEE Trans. Inf. Theory*, vol. 28, no. 5, pp. 803–807, Sep. 1982.

Main result: outline of the proof

Theorem:

$$R_{min} = H_{G_\epsilon}(X|Y),$$

where R_{min} is obtained from Yamamoto's codes under maximal distortion

- **Proof:** $R_{min} \leq H_{G_\epsilon}(X|Y)$ is trivial since Yamamoto's codes are optimal. We need to show that $R_{min} \geq H_{G_\epsilon}(X|Y)$.
- Idea: For every U, g satisfying conditions in R_{min}, we need to find corresponding W that satisfy conditions in $H_{G_\epsilon}(X|Y)$.

Main result: outline of the proof

Theorem:

$$R_{min} = H_{G_\epsilon}(X|Y),$$

where R_{min} is obtained from Yamamoto's codes under maximal distortion

- **Proof:** $R_{min} \leq H_{G_\epsilon}(X|Y)$ is trivial since Yamamoto's codes are optimal. We need to show that $R_{min} \geq H_{G_\epsilon}(X|Y)$.
- Idea: For every U, g satisfying conditions in R_{min}, we need to find corresponding W that satisfy conditions in $H_{G_\epsilon}(X|Y)$.
- Define $\hat{w}(u) = \{x : p(u,x) > 0\}$.

Main result: outline of the proof

Theorem:

$$R_{min} = H_{G_\epsilon}(X \mid Y),$$

where R_{min} is obtained from Yamamoto's codes under maximal distortion

- **Proof:** $R_{min} \leq H_{G_\epsilon}(X \mid Y)$ is trivial since Yamamoto's codes are optimal. We need to show that $R_{min} \geq H_{G_\epsilon}(X \mid Y)$.
- Idea: For every U, g satisfying conditions in R_{min}, we need to find corresponding W that satisfy conditions in $H_{G_\epsilon}(X \mid Y)$.
- Define $\hat{w}(u) = \{x : p(u, x) > 0\}$.
- Define W as $p(w \mid u, x) = \begin{cases} 1, \text{ if } w = \hat{w}(u) \\ 0, \text{ otherwise.} \end{cases}$

Main result: outline of the proof

Theorem:

$$R_{min} = H_{G_\epsilon}(X \mid Y),$$

where R_{min} is obtained from Yamamoto's codes under maximal distortion

- **Proof:** $R_{min} \leq H_{G_\epsilon}(X \mid Y)$ is trivial since Yamamoto's codes are optimal. We need to show that $R_{min} \geq H_{G_\epsilon}(X \mid Y)$.
- Idea: For every U, g satisfying conditions in R_{min}, we need to find corresponding W that satisfy conditions in $H_{G_\epsilon}(X \mid Y)$.
- Define $\hat{w}(u) = \{x : p(u, x) > 0\}$.
- Define W as $p(w \mid u, x) = \begin{cases} 1, \text{ if } w = \hat{w}(u) \\ 0, \text{ otherwise.} \end{cases}$
- Rest of the proof is showing that the above definitions of W satisfy the required conditions.

Main result: outline of the proof

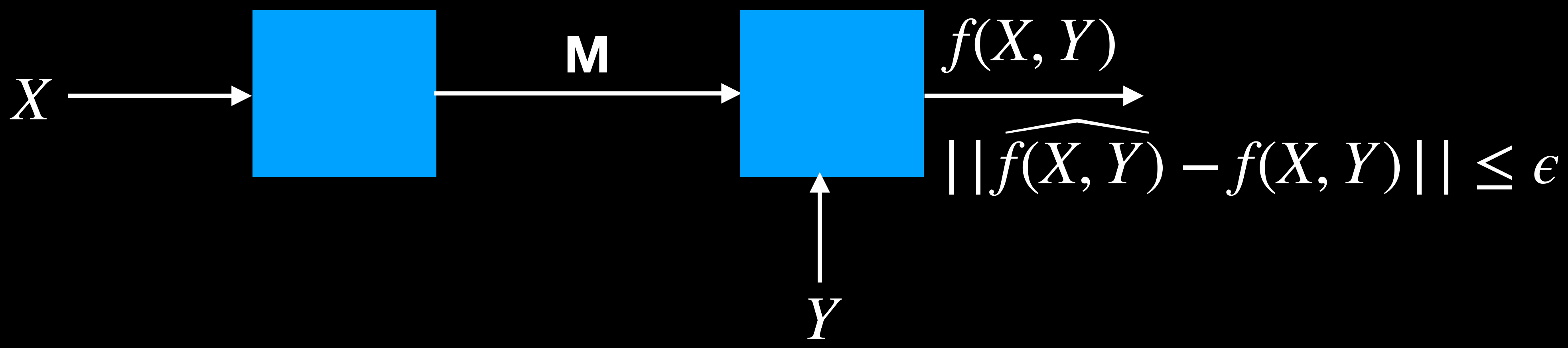
Theorem:

$$R_{min} = H_{G_\epsilon}(X \mid Y),$$

where R_{min} is obtained from Yamamoto's codes under maximal distortion

- **Proof:** $R_{min} \leq H_{G_\epsilon}(X \mid Y)$ is trivial since Yamamoto's codes are optimal. We need to show that $R_{min} \geq H_{G_\epsilon}(X \mid Y)$.
- Idea: For every U, g satisfying conditions in R_{min}, we need to find corresponding W that satisfy conditions in $H_{G_\epsilon}(X \mid Y)$.
- Define $\hat{w}(u) = \{x : p(u, x) > 0\}$.
- Define W as $p(w \mid u, x) = \begin{cases} 1, \text{ if } w = \hat{w}(u) \\ 0, \text{ otherwise.} \end{cases}$
- Rest of the proof is showing that the above definitions of W satisfy the required conditions.
- In particular, smallest enclosing circles play crucial role in the proof.

Main result: key insight



- Our coding scheme uses better geometric methods for constructing hypergraphs.
- In particular, [3] used point-to-point comparison for constructing hypergraphs.
- We use smallest enclosing circles to form hypergraphs, which we prove is optimal.

3. V. Doshi, D. Shah, M. Médard, and M. Effros, "Functional compression through graph coloring." *IEEE Transactions on Information Theory* 56.8 (2010): 3901-3917.

Consequences: constructing optimal practical codes in $O(N \log N)$ time

Optimal code construction:

$R_{min} = H_{G_\epsilon}(X \mid Y)$ implies optimal code construction reduces to:

- **Step 1:** Construct G_ϵ
- **Step 2:** Use polar coding technique

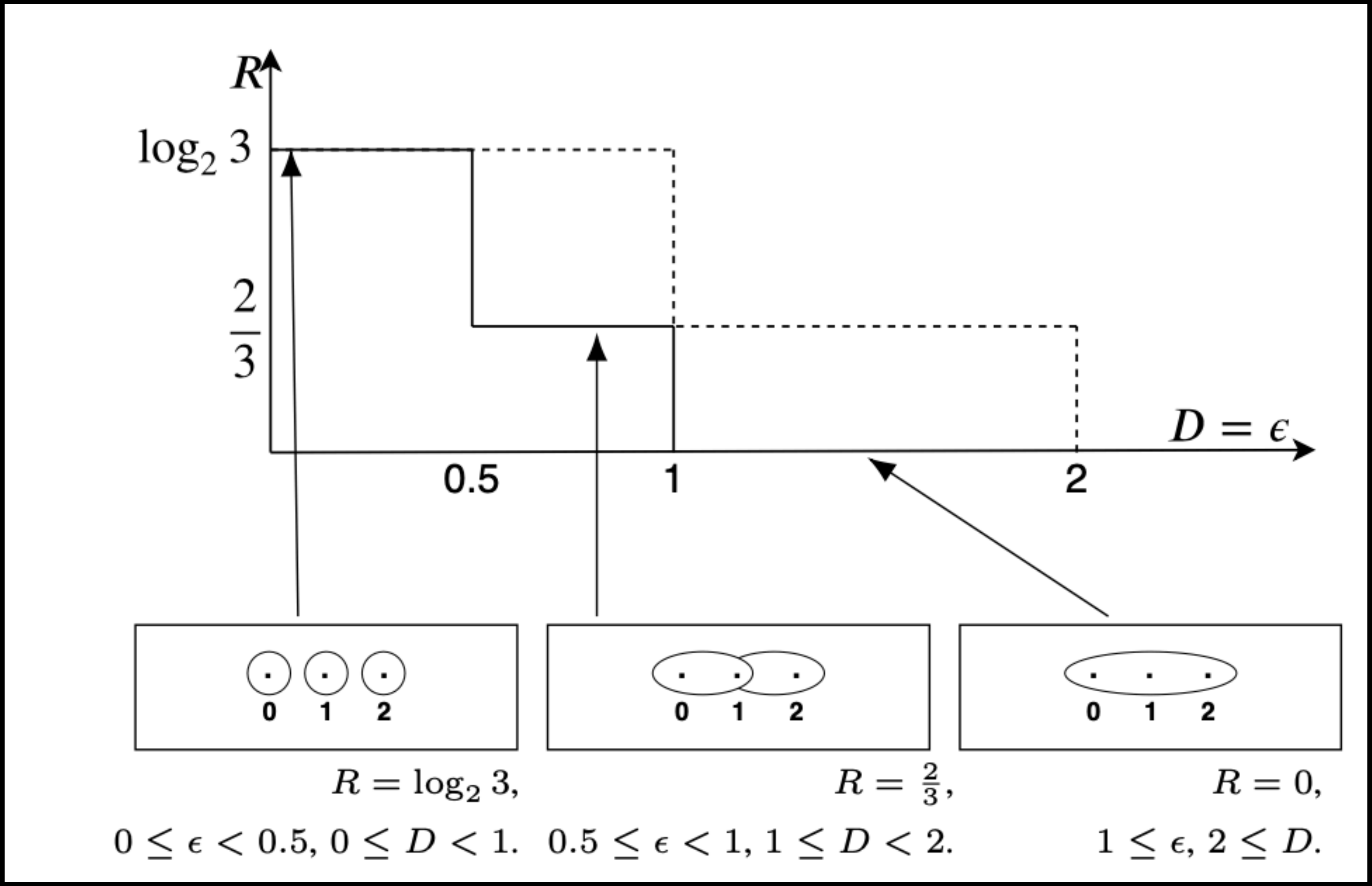
Time complexity:

Time complexity of the coding scheme = $O(N \log N)$ because:

- G_ϵ construction takes finite time
- Polar coding takes $O(N \log N)$ time

Consequences: Improvement in practical rate

Example: Consider the point to point source coding problem with no side information, X uniformly distributed over $\{0,1,2\}$ and take the function f as the identity function.

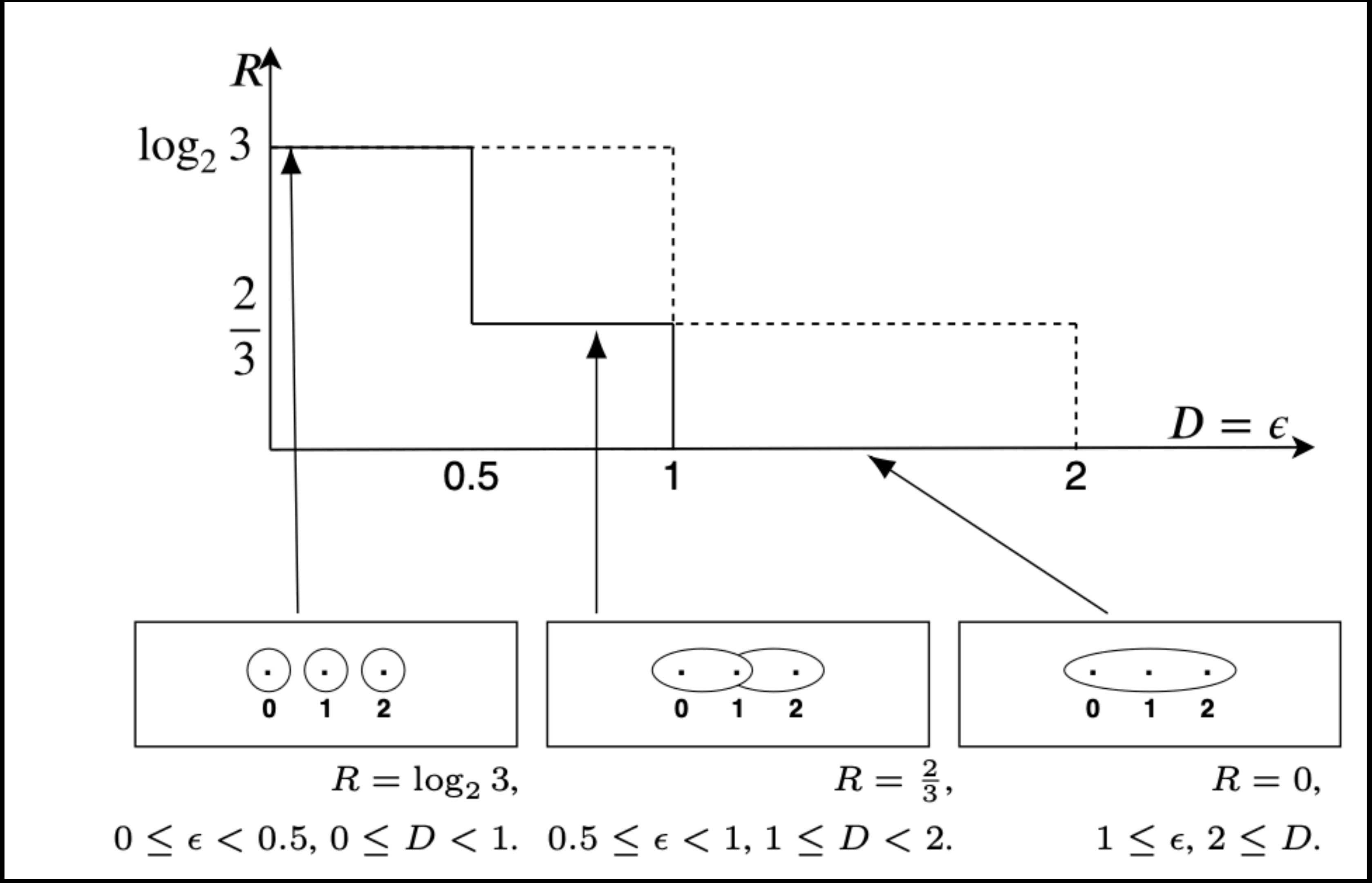


- **Solid line: our rate**
- **Dashed line: existing practical codes**

Other important findings

Discontinuity of functional ϵ-entropy

- **Discontinuity of $H_{G_\epsilon}(X \mid Y)$ as a function of ϵ**
- **Fits our intuition of maximal distortion, but previously unknown**

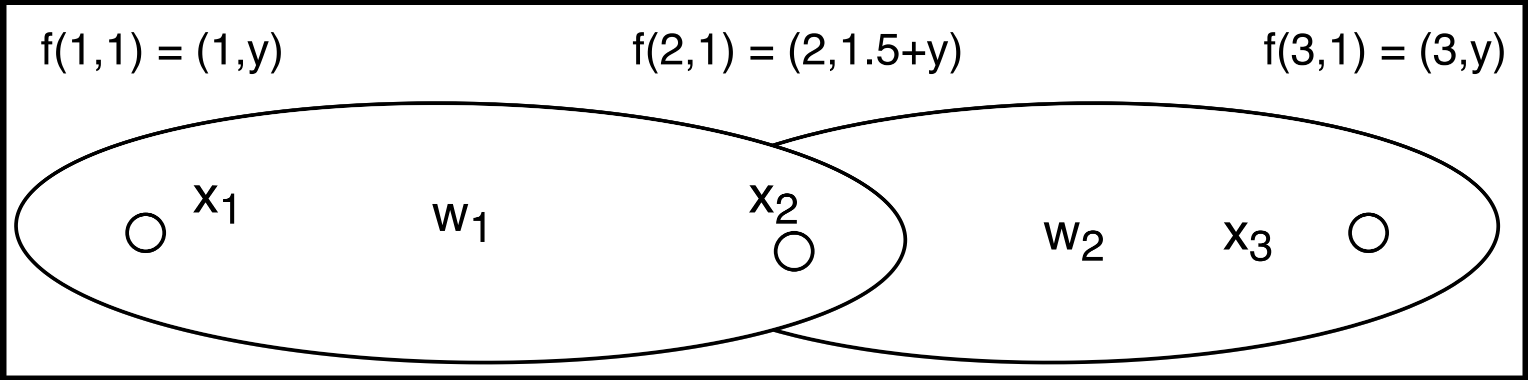


Other important findings

Overlapping hyperedges in G_ϵ

- **In [2] where $\epsilon = 0$, $p(x, y) > 0$ implied non-overlapping hyperedges**
- **Using counterexample, we show that for $\epsilon > 0$, G_ϵ can have overlapping hyperedges**

Example: Let X and Y be independent uniform random variables defined on the support set $\mathcal{X} = \{1,2,3\}$ and $\mathcal{Y} = \{1,2\}$ respectively. Let $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$, where $\mathcal{Z} \subset \mathbb{R}^2$, and f be defined as $f(1,y) = (1,y)$, $f(2,y) = (2,1.5 + y)$, $f(3,y) = (3,y)$, and let $\epsilon = \frac{\sqrt{13}}{4}$. Then the characteristic hypergraph G_ϵ has overlapping hypergraphs.



2. A. Orlitsky and J. R. Roche, "Coding for computing," IEEE Trans. Inf. Theory, vol. 47, no. 3, pp. 903–917, Mar. 2001.

Other important findings

When f is unknown

- **When is a well-behaved function (e.g. L-Lipschitz continious), we provide efficient coding schemes even when f is unknown**

Corollary:

Let $f : \mathscr{X} \mapsto \mathscr{Z}$ be a L-Lipschitz continuous function. Then $R(\epsilon)$ can be upper-bounded as

$$R(\epsilon) \leq H_{G_{\epsilon/L}}(X),$$

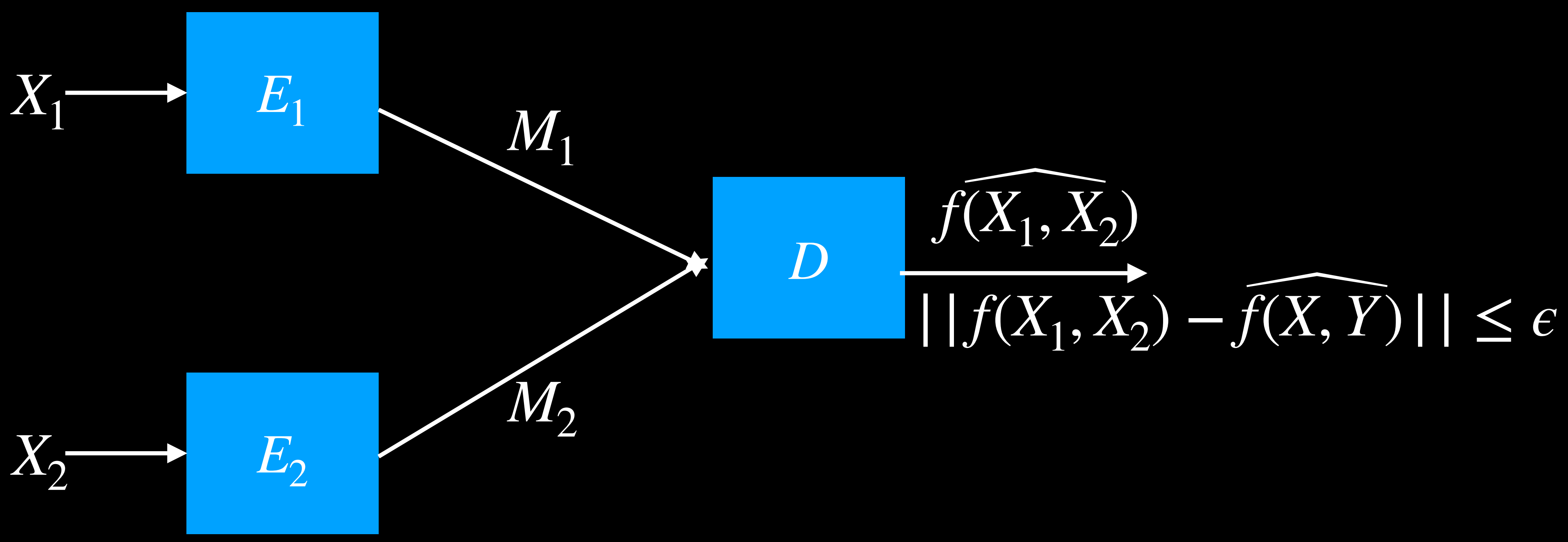
where $G_{\epsilon/L}$ is constructed with respect to the random variable X and the identity function and hence the upper-bound is achievable by the encoder even when f is unknown.

Summary

- First optimal practical codes for lossy functional source coding
- We close the gap in rates between practical and optimal codes under maximal distortion.
- We introduce better geometrical methods in our coding schemes improving on existing techniques
- Rate region is discontinuous as a function of fidelity parameter
- Counterintuitive overlapping hyperedges even for all positive probabilities
- Efficient coding schemes for unknown but well-behaved functions

Future work

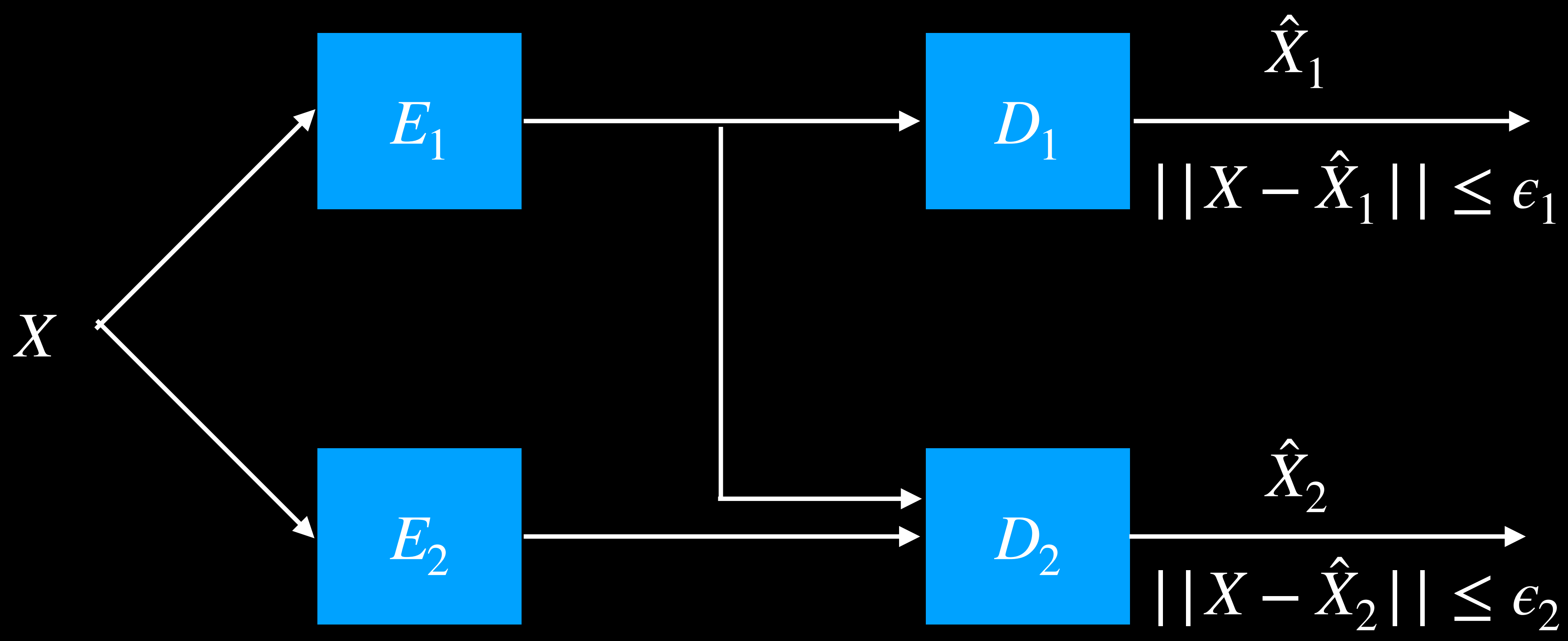
- Using similar geometric techniques progress are being made on practical codes for distributed source coding and successive refinement problem



Distributed source coding

Future work

- **Using similar geometric techniques progress are being made on practical codes for distributed source coding and successive refinement problem**



Successive refinement coding