Information bottlenecks in emerging distributed systems

Victoria Kostina
Caltech, USA

2022 North American School of Information Theory | UCLA | Aug. 19, 2022

1948

A Mathematical Theory of Communication

By C. E. SHANNON

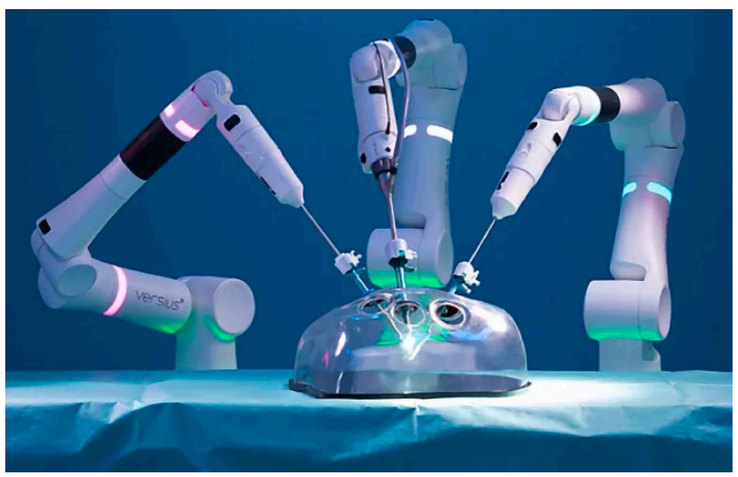
INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

Emerging distributed systems

- Classical information theory lacks ready answers for emerging distributed systems where delays, feedback and context are important



Emerging distributed systems

- Delay-sensitive
  - coding over long blocks of observed data is not feasible
- Interactive
  - feedback is important
- Context-dependent
  - the part of data to be transmitted depends on the goal of communication, so coding and control/computation algorithms should be designed jointly

The plan

- **Part I**: Non-asymptotic rate-distortion theory
  - bounds to the non-asymptotic operational limit;
- **Part II**: coding for control
  - causality, feedback and memory of the past are important
- **Part III**: coding for computation
  - causality, feedback and memory of the past are important.

Part I:

Non-asymptotic rate-distortion theory

Lossless data compression

Definition: A lossless data compression code is a pair of mappings:

$$\text{Compressor:} \qquad \mathsf{f} : \mathcal{X} \mapsto \{0,1\}^*$$

$$\text{Decompressor:} \qquad \mathsf{g} : \{0,1\}^* \mapsto \mathcal{X}$$

$$\{0,1\}^* = \{\emptyset, 0, 1, 00, 01, 10, 11, 000, 001, \ldots\}$$

$$\mathsf{g}(\mathsf{f}(a)) = a, \ \forall a \in \mathcal{X}$$

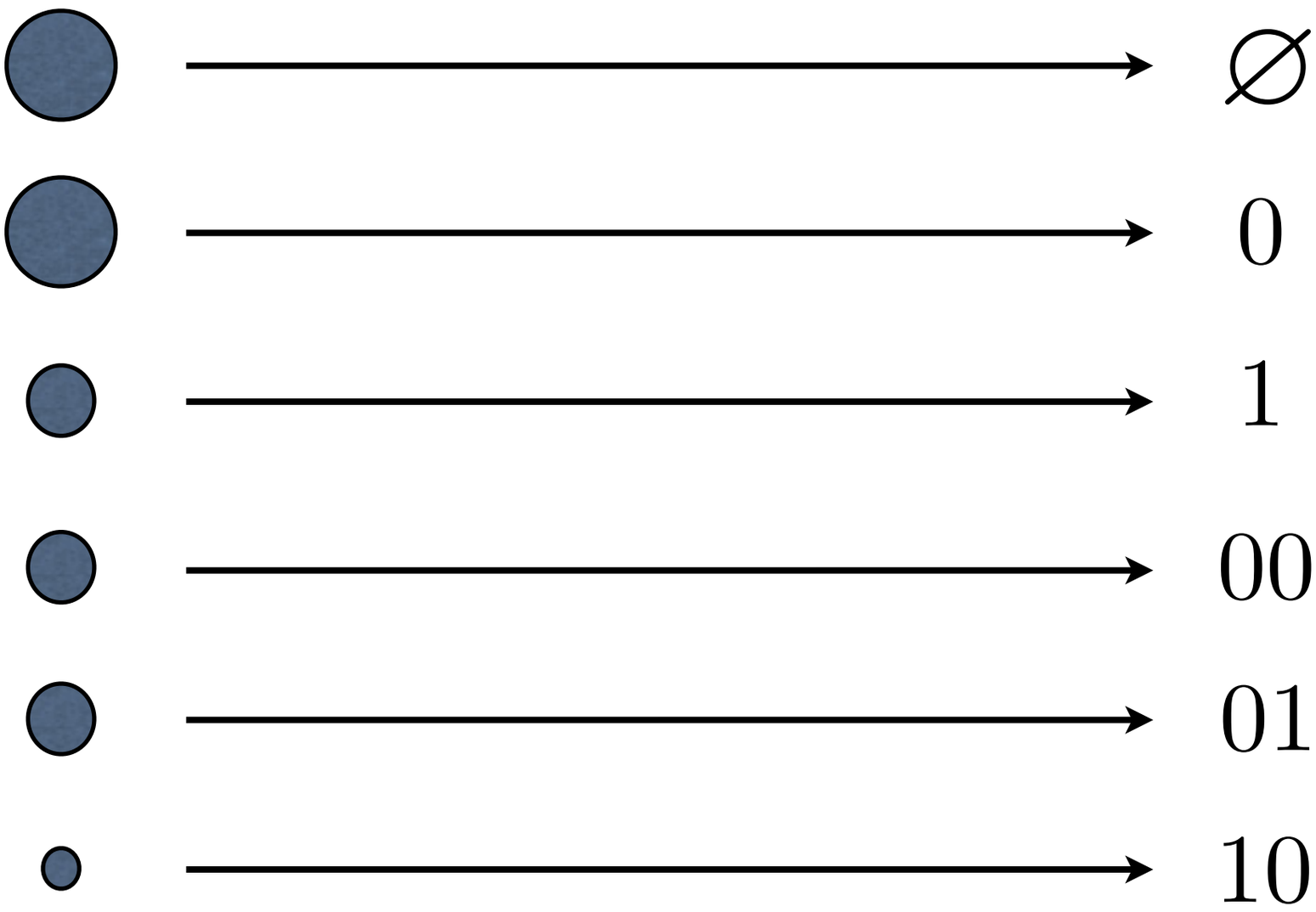
Objective of lossless data compression

- Find the best compressor to minimize $\mathbb{E}\left[\ell(f(X))\right]$, sup, median, $\mathbb{P}\left[\ell(x) > k\right]$, ...

One minimizes all!

Idea of lossless data compression

- Longer codewords to less likely symbols



Optimum lossless source code

Without loss of generality, label the elements of $\mathcal{X}$, in decreasing probability: $P_X[1] \geq P_X[2] \geq \ldots\ldots$

$$
\begin{aligned}
\mathsf{f}^*(1) &= \emptyset \\
\mathsf{f}^*(2) &= 0 \\
\mathsf{f}^*(3) &= 1 \\
\mathsf{f}^*(4) &= 00 \\
\mathsf{f}^*(5) &= 01 \\
\mathsf{f}^*(6) &= 10 \\
\mathsf{f}^*(7) &= 11 \\
\mathsf{f}^*(8) &= 000 \\
\mathsf{f}^*(9) &= 001 \\
\mathsf{f}^*(10) &= 010 \\
&\vdots
\end{aligned}
$$

Length of the optimum lossless source code

$$\ell(\mathsf{f}^*(x)) = \lfloor \log_2 x \rfloor$$

Are we done?

No, because it's hard to get insight into the behavior of $\ell(\mathsf{f}^\star(x))$.

Optimal average length vs. entropy

Theorem (average length)

$$H(X) - \log_2(H(X)+1) - \log_2 e \leq \mathbb{E}\left[\ell(\mathsf{f}^\star(X))\right] \leq H(X)$$

Alon-Orlitsky'94 Wyner'72

Intuition:

$H(X)$ captures storage requirements for X:

$$\mathbb{E}\left[\ell(\mathsf{f}^\star(X))\right] = H(X) + O\left(\log H(X)\right)$$

Memoryless sources:

$$\mathcal{X} = \mathcal{A}^n,\ X = (X_1, \ldots, X_n),\ P_{X_1 \ldots X_n} = P_{\mathsf{X}}^{\otimes n}$$

$$\mathbb{E}\left[\ell(\mathsf{f}^\star(X_1, \ldots, X_n))\right] = nH(\mathsf{X}) + O\left(\log n\right)$$

Optimal average length vs. entropy

Proof (upper bound)

$$\begin{aligned}\mathbb{E}\left[\ell(\mathsf{f}^\star(X))\right] &= \mathbb{E}\left[\lfloor \log_2 X \rfloor\right] \\ &\leq \mathbb{E}\left[\log_2 X\right] \\ &\leq \mathbb{E}\left[\log_2 \frac{1}{P_X(X)}\right] \\ &= H(X)\end{aligned}$$

$$P_X[a] \leq \frac{1}{a}$$

This is because the elements of $\mathcal{X} = \{1, 2, \ldots\}$ are ordered in decreasing probability order. Assume the opposite:

$$1 < iP_X(i) \leq \sum_{j=1}^{i} P_X(j).$$

Impossible!

Lossless data compression: research directions

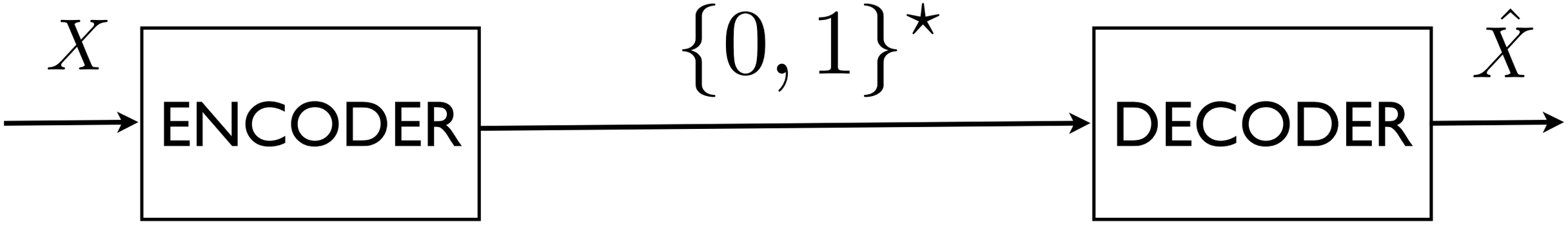
- Refined nonasymptotic upper and lower bounds to various operational quantities of interest - e.g. quantiles of $\ell(\mathsf{f}^\star(X))$, its moments, etc.

- Refined asymptotic expansions for memoryless sources using large deviations, moderate deviations and central limit theorem results from probability theory, e.g. refining the $O(\log n)$ term in $\mathbb{E}\left[\ell(\mathsf{f}^\star(X_1,\ldots,X_n))\right] = nH(\mathsf{X}) + O(\log n)$

- Doing the above for an unknown source distribution (universal compression)

- Doing the above for separate encoding of multiple sources (Slepian-Wolf problem), including a large (massive) number of sources

Lossy data compression



distortion measure: $\mathsf{d}\colon \mathcal{X} \times \hat{\mathcal{X}} \mapsto [0, +\infty)$

$$\mathbb{E}\left[\mathsf{d}(X, \hat{X})\right] \leq d$$

Compressor: $\quad \mathsf{f} : \mathcal{X} \mapsto \{0,1\}^*$

Decompressor: $\quad \mathsf{g} : \{0,1\}^* \mapsto \mathcal{X}$

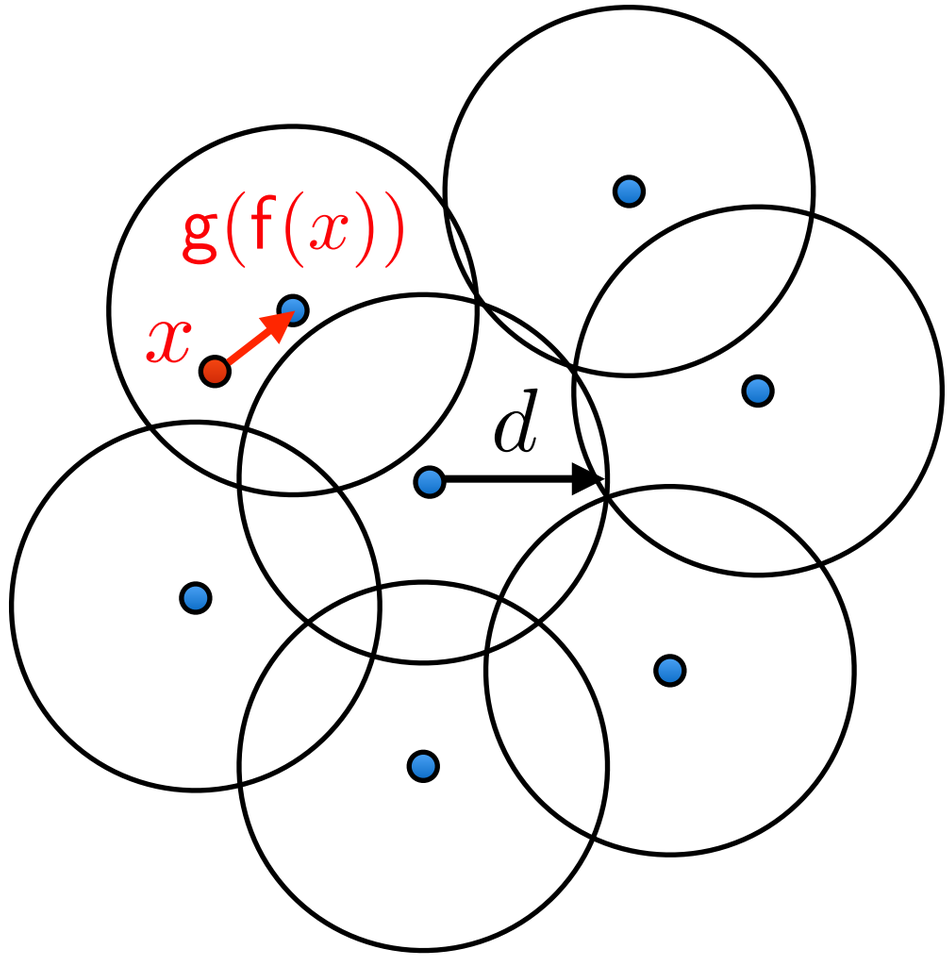
$$\{0,1\}^* = \{\emptyset, 0, 1, 00, 01, 10, 11, 000, 001, \ldots\}$$

Objective of lossy data compression

- Find the best compressor to minimize $\mathbb{E}\left[\ell(\mathsf{f}(X))\right]$ subject to

$$\mathsf{d}(X, \mathsf{g}(\mathsf{f}(X))) \leq d \text{ a.s.}$$

- Optimal code is unknown because the optimal placement of quantization points is unknown



d-ball entropy

$$R_d^+(X) \triangleq \inf_{P_{\hat{X}}} \mathbb{E}\left[\log_2 \frac{1}{P_{\hat{X}}(\mathcal{B}_d(X))}\right]$$

where $\mathcal{B}_d(x) \triangleq \left\{\hat{x} \in \widehat{\mathcal{X}} \colon \mathsf{d}(x,\hat{x}) \leq d\right\}$

Optimal average length vs. d-ball entropy

$$R_d^+(X) - \log_2(R_d^+(X)+1) - \log_2 e \leq \min_{\substack{\mathsf{f},\mathsf{g}:\\ \mathsf{d}(X,\mathsf{g}(\mathsf{f}(X)))\leq d}} \mathbb{E}\left[\ell(\mathsf{f}(X))\right] \leq R_d^+(X)$$

V. Kostina, Y. Polyanskiy and S. Verdú, "Variable-length compression allowing errors", IEEE Transactions on Information Theory, vol. 61, no. 9, pp. 4316-4330, Aug. 2015

d-ball entropy vs. rate-distortion function

$$R_d(X) \triangleq \min_{\substack{P_{\hat{X}|X}:\, \mathcal{X} \mapsto \hat{\mathcal{X}}: \\ \mathbb{E}\left[\mathsf{d}(X,\hat{X})\right] \le d}} I(X;\hat{X}) \qquad \text{rate-distortion function}$$

$$\le$$

$$R_d^+(X) \triangleq \inf_{P_{\hat{X}}} \mathbb{E}\left[\log_2 \frac{1}{P_{\hat{X}}(\mathcal{B}_d(X))}\right] \qquad \text{d-ball entropy}$$

$$\le$$

$$R_d(X) + O\left(\log_2 R_d(X)\right) \qquad \text{under regularity conditions}$$

V. Kostina, Y. Polyanskiy and S. Verdú, "Variable-length compression allowing errors", IEEE Transactions on Information Theory, vol. 61, no. 9, pp. 4316-4330, Aug. 2015

Optimal average length vs. rate-distortion function

$$\min_{\substack{\mathsf{f},\mathsf{g}:\\ \mathsf{d}(X,\mathsf{g}(\mathsf{f}(X)))\le d}} \mathbb{E}\left[\ell(\mathsf{f}(X))\right] = R_d(X) + O\left(\log R_d(X)\right)$$

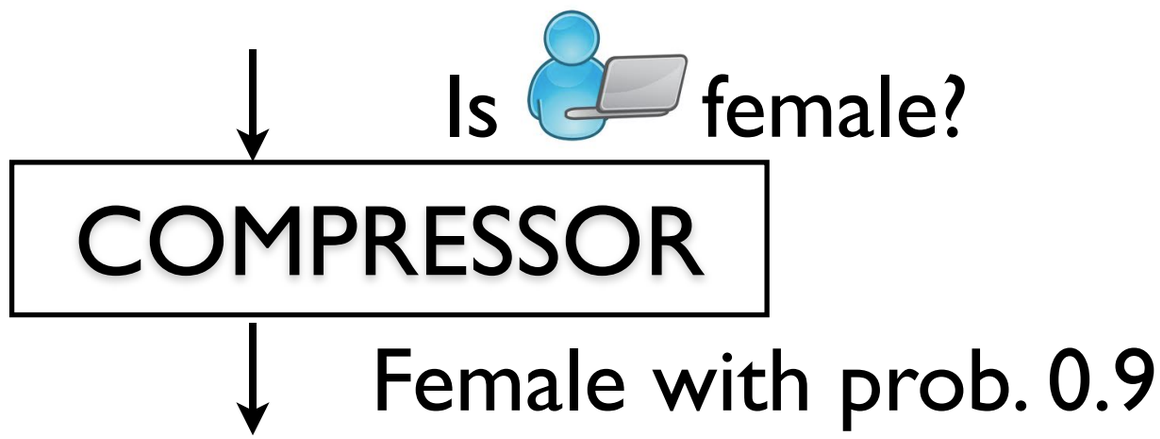
V. Kostina, Y. Polyanskiy and S. Verdú, "Variable-length compression allowing errors", IEEE Transactions on Information Theory, vol. 61, no. 9, pp. 4316-4330, Aug. 2015

Lossy data compression: research directions

- Perceptually meaningful distortion measures
  - empirical mapping techniques because eye/ear not well understood

$$d\left(\quad , \quad \right) \mapsto \mathbb{R}^{+}$$

- Rate-distortion theory for data inference tasks
  - The input is a query
  - The output is a belief



Jerry D. Gibson, Jing Hu. Rate-Distortion Bounds for Voice and Video, Foundations and Trends in Communications and Information Theory, Vol. 10, No. 4, Feb. 2014.

T. A. Courtade, T. Weissman, 2014, "Multiterminal source coding under logarithmic loss," IEEE Transactions on Information Theory, vol. 60, no. 1, pp. 740–761, Jan. 2014"

Part I: Takeaways

- Information theory provides tools to tightly sandwich *nonasymptotic* operational limits
- Probability theory provides tools to *approximate* those operational limits, even if blocklength is finite
- This fundamental area of research still has many open problems
- This fundamental area of research can inform practical code designs

Part II:

Coding for control

Scalar stochastic linear control

system gain

$$X_{i+1} = aX_i + U_i + V_i$$

system state

control action

random disturbance (mean 0)

Fully observed vs. partially observed

In a *fully observed* system, the controller observes the system state directly.

In a *partially observed* system, the controller has only partial information about the system state.

Control objective

find a control strategy to achieve stability:

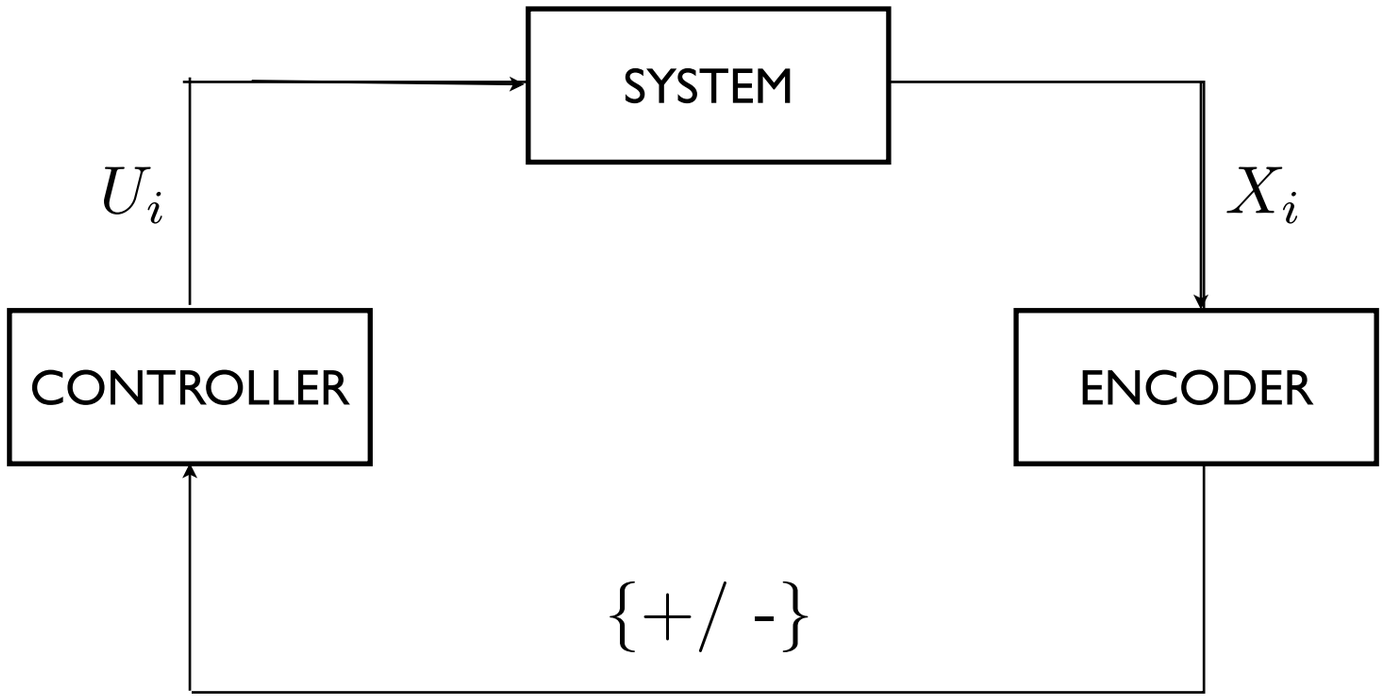
$$\lim_{i \to \infty} \mathbb{E}\left[|X_i|^2\right] < \infty$$

In a *fully observed* system,

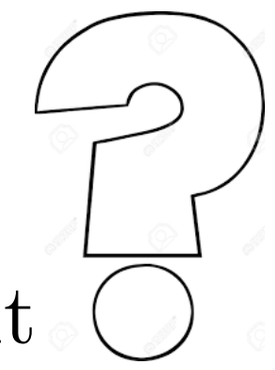
solution: $U_i = -aX_i$ achieves $\mathbb{E}\left[|X_{i+1}|^2\right] = \mathrm{Var}\left[V_i\right]$

Simple question

$$X_{i+1} = aX_i + U_i + V_i$$



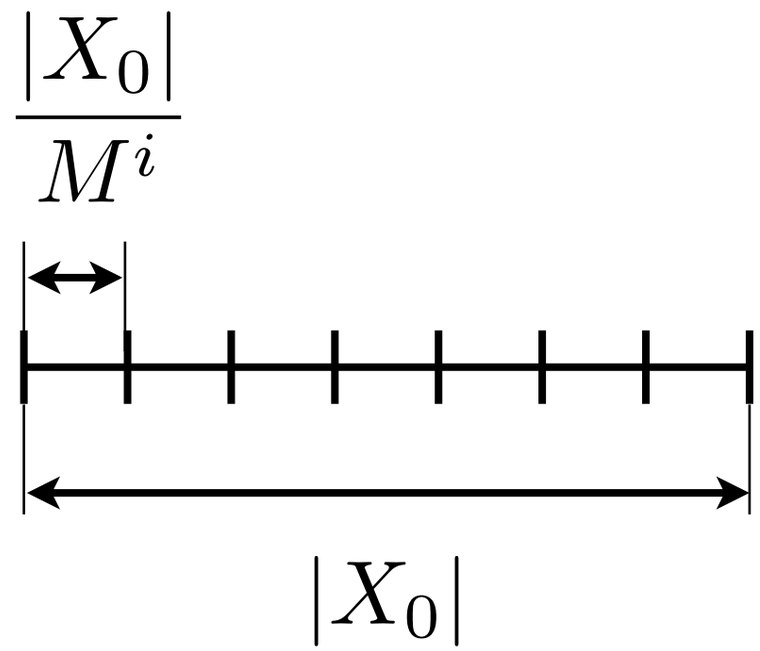
If $a \in [1, 2)$, can we achieve $\lim_{i\to\infty} \mathbb{E}\left[|X_i|^2\right] < \infty$ with 1 bit



Bounded disturbances: converse

$$X_{i+1} = aX_i + U_i$$

- $a > 1$: unstable
- $X_i = a^i \underbrace{\underbrace{\left(X_0 + \sum_{j=0}^{i-1} a^{-j-1} U_j\right)}_{\tilde{U}_i}}_{\text{for } X_i \le X_0, \text{ should be } \le a^{-i} X_0}$

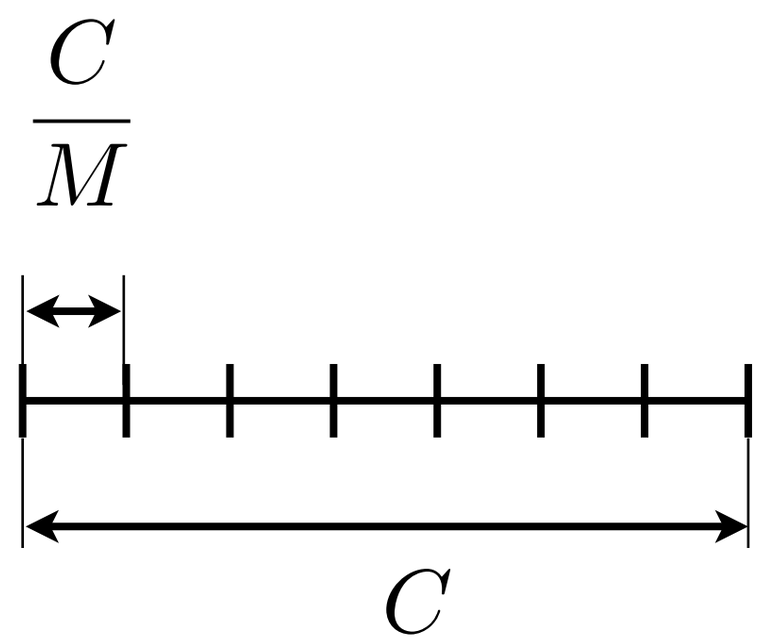


- Quantization bin size $= \frac{|X_0|}{M^i}$. So, $M \ge a$
- (actually, $M > a$ if $V_i \neq 0$).

Bounded disturbances: achievable scheme

$$X_{i+1} = aX_i + U_i + V_i$$

- $a > 1$: unstable
- $|V_i| \leq B$



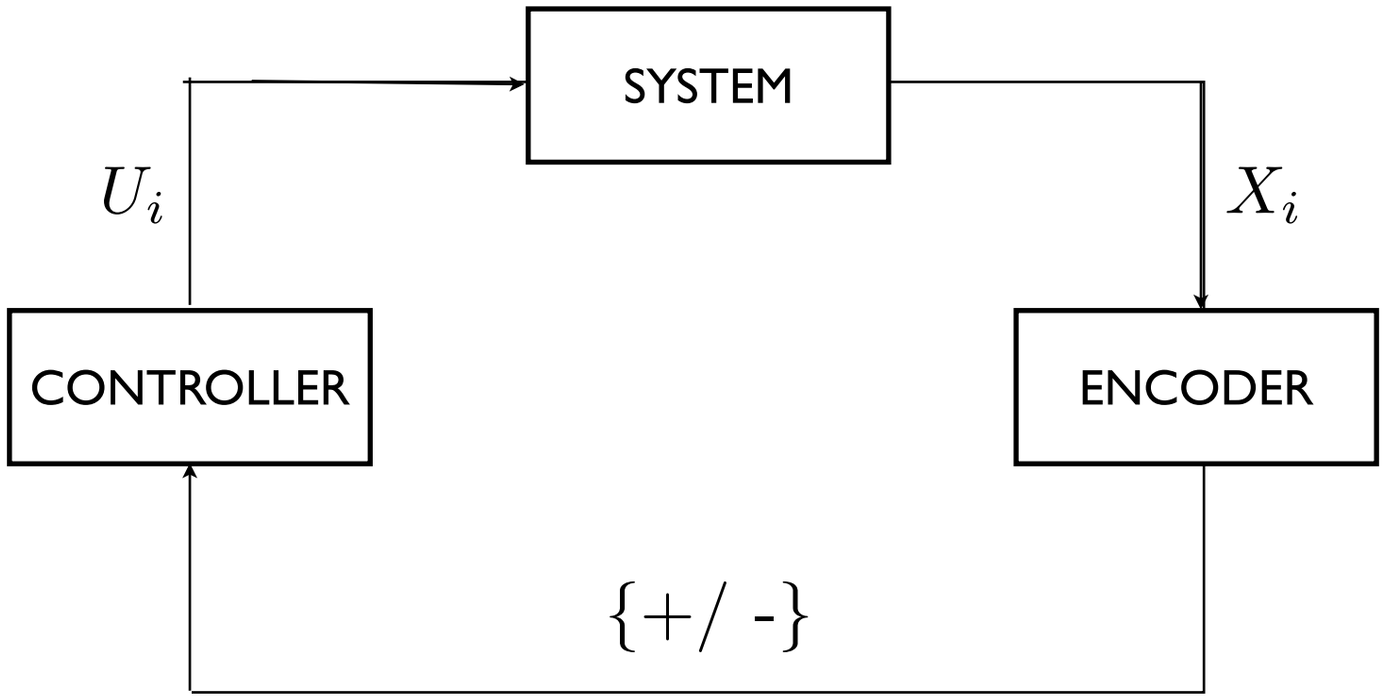
A simple time-invariant scheme:

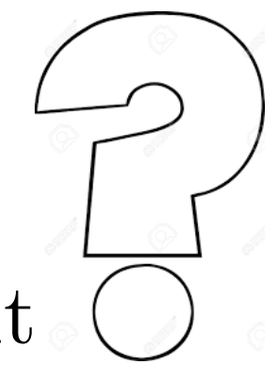
- $U_i = -a\hat{X}_i$, where $\hat{X}_i$ is a quantized version of X_i.
- If $|X_i| \leq C_i$, then $|X_{i+1}| \leq C_{i+1} \triangleq \frac{aC_i}{M} + B$.
- Actually, $C_i \downarrow \frac{B}{1-a/M}$ and diverges if $M \leq a$.
- So, $M^\star = \lfloor a + 1 \rfloor$ is the minimum possible.

Baillieul'99, Wong-Brockett'99

Bounded disturbances - simple answer

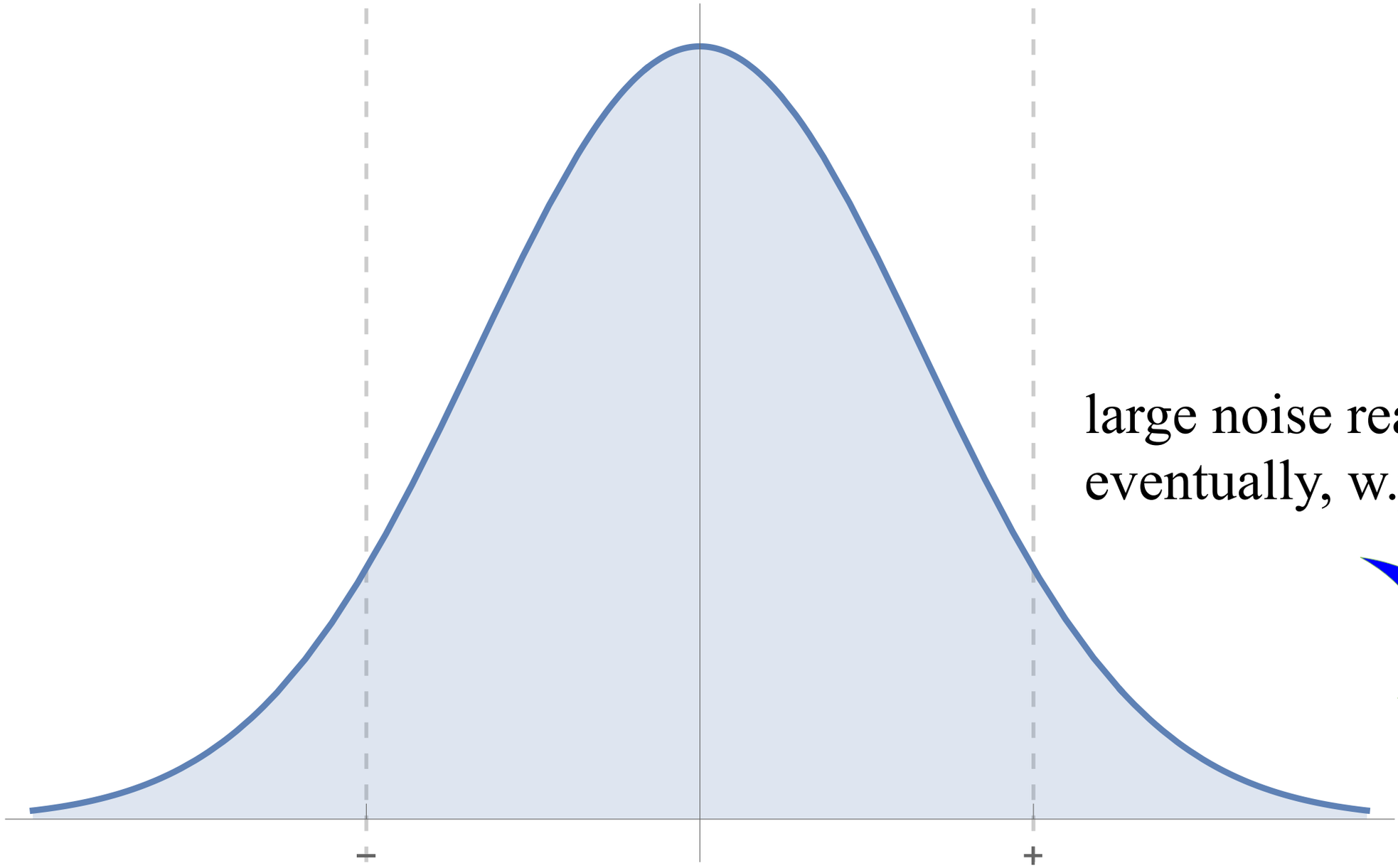
$$X_{i+1} = aX_i + U_i + V_i$$



If $a \in [1, 2)$, can we achieve $\lim_{i\to\infty} \mathbb{E}\left[|X_i|^2\right] < \infty$ with 1 bit 

Yes, by keeping track of the state uncertainty at the controller and using the sign of the X_i to select one of the two quantization bins

Unbounded disturbances



Unbounded disturbances

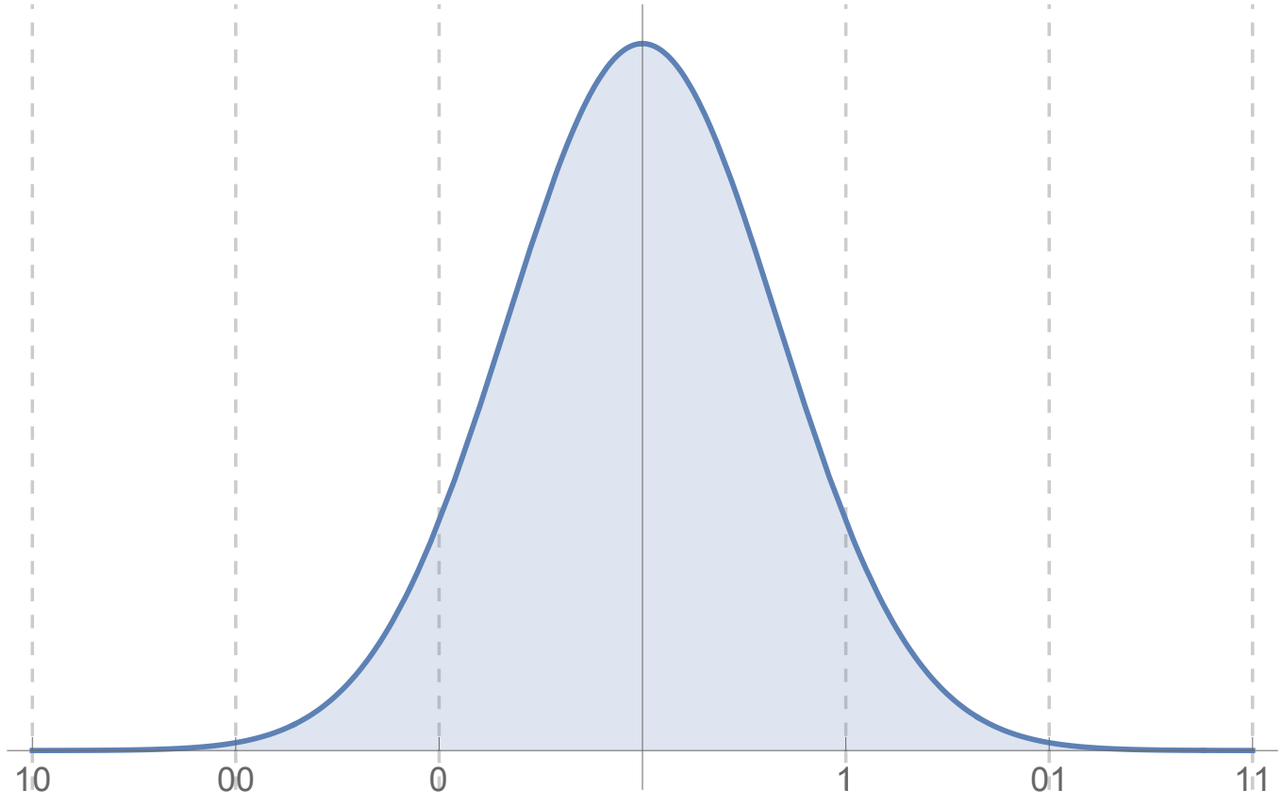
Proposition (Nair-Evans'04)

For unstable linear systems with unbounded disturbances, any time-invariant quantizer with finitely many levels cannot achieve bounded $\lim_{i\to\infty} \mathbb{E}\left[|X_i|^\beta\right]$, for any $\beta > 0$.

So, the only option is a **zooming adaptive quantizer** (introduced by Brockett-Liberzon'00)

Variable-length quantization is easier!

Longer bit strings are transmitted to encode rarer (larger) noise realizations



Kostina-Hassibi’16
Silva-Derpich-Ostergaard-Encina’16

Fundamental limit

Definition

$$M_\beta^\star \triangleq \min\left\{M : \exists\, M\text{-bin causal quantizer-controller s.t. } \limsup_i \mathbb{E}\left[|X_i|^\beta\right] < \infty\right\}$$

An *M-bin causal quantizer-controller* for $X_1, X_2, \ldots$ is a sequence $\{\mathsf{f}_n, \mathsf{g}_n\}_{n=1}^\infty$, where

- $\mathsf{f}_n : \mathbb{R}^n \mapsto \{1, 2, \ldots, M\}$ is the encoding (quantizing) function, and

- $\mathsf{g}_n : \{1, 2, \ldots, M\}^n \mapsto \mathbb{R}$ is the decoding (controlling) function.

At time i, the controller outputs

$$U_n = \mathsf{g}_n(\mathsf{f}_1(X_1), \mathsf{f}_2(X^2), \ldots, \mathsf{f}_n(X^n)).$$

A "data rate theorem"

Theorem

Let V_i be independent, with bounded α-moments, and V_i, X_0 have a density. Then for any $0 < \beta < \alpha$, the minimum number of quantization points to achieve β-moment stability is

$$M_\beta^\star = \lfloor a + 1 \rfloor.$$

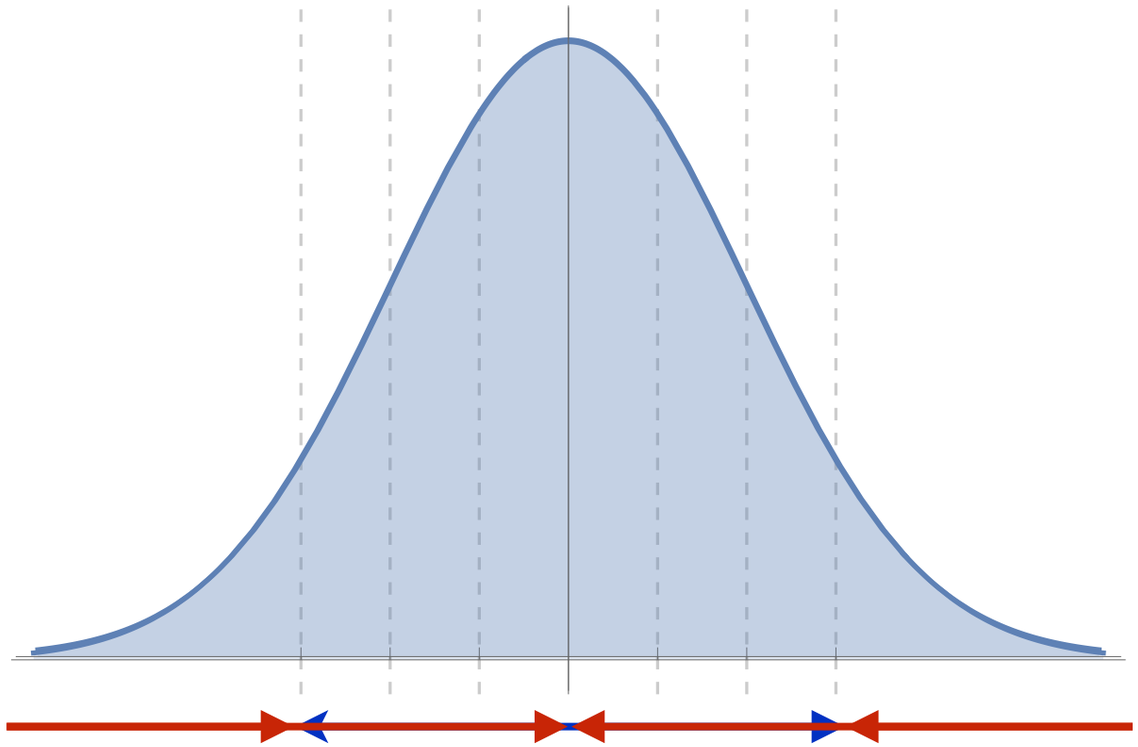
- $a < 1$ (stable system) $\rightarrow M_\beta^\star = 1.$
- $a \in [1, 2) \rightarrow M_\beta^\star = 2.$
- $a \in [2, 3) \rightarrow M_\beta^\star = 3.$

V. Kostina, Y. Peres, G. Ranade, and M. Sellke, "Exact minimum number of bits to stabilize a linear system ," IEEE Transactions on Automatic Control, Nov. 2021.

Data rate theorems in literature

- Bounded disturbances: Baillieul'99, Wong-Brockett'99, Tatikonda-Mitter'04.
- Linear systems: Brockett-Liberzon'00, Nair-Evans'04, Yüksel'10, Johnston-Yüksel'14, You-Xie'11.
- Nonlinear systems: Yüksel-Meyn'13, Yüksel-Basar'13, Yüksel'16.
- Converse: Nair-Evans'04, Martins-Dahlia-Elia'06, Matveev'08, Matveev-Savkin'08, Colonius-Kawan'08, Minero-Franceschetti-Dey-Nair'09, Colonius-Kawan-Nair'13, Yüksel-Basar'13, Yüksel'16.

Zooming adaptive quantizers: prior work

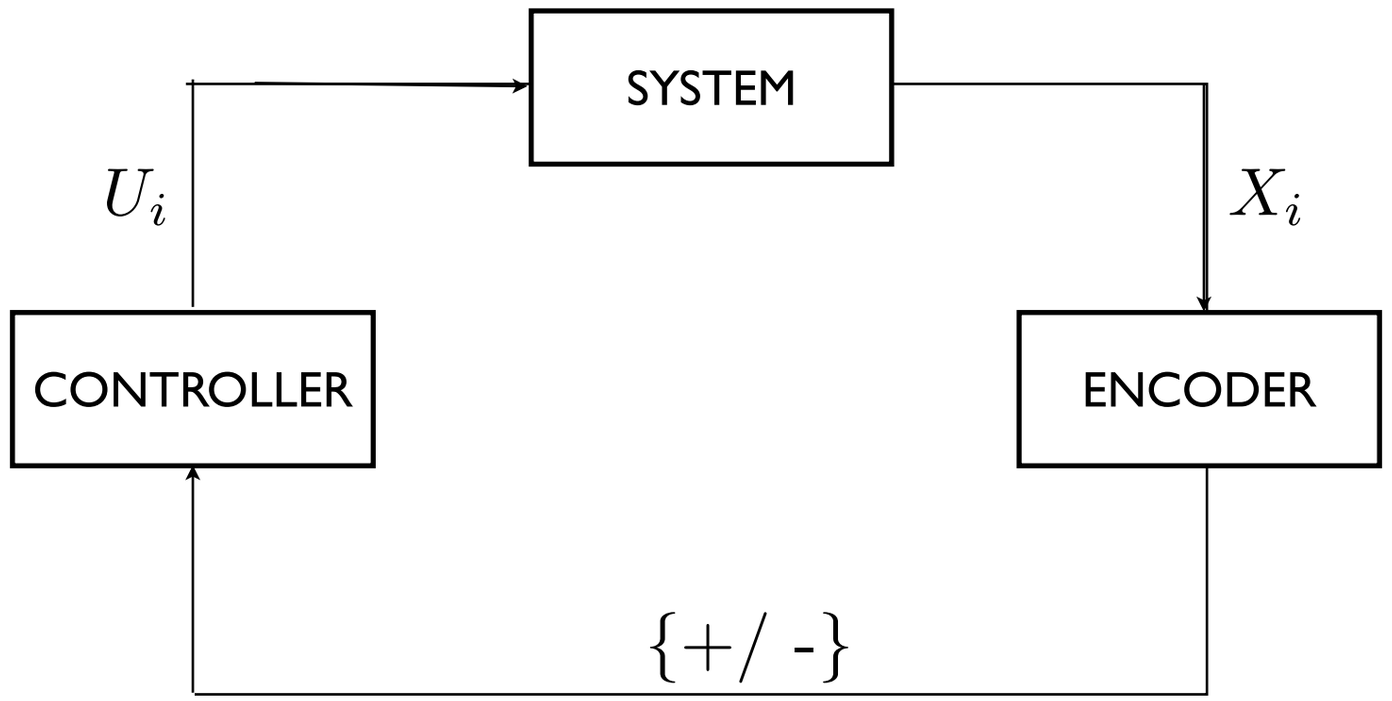


Zoom out if in overload region

Nair-Evans'04, Yüksel'10

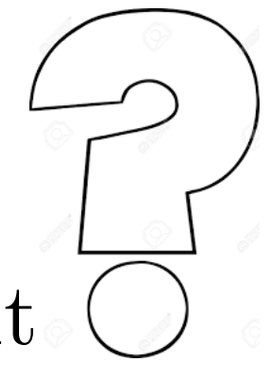
Unbounded disturbances - achievable scheme

$$X_{i+1} = aX_i + U_i + V_i$$



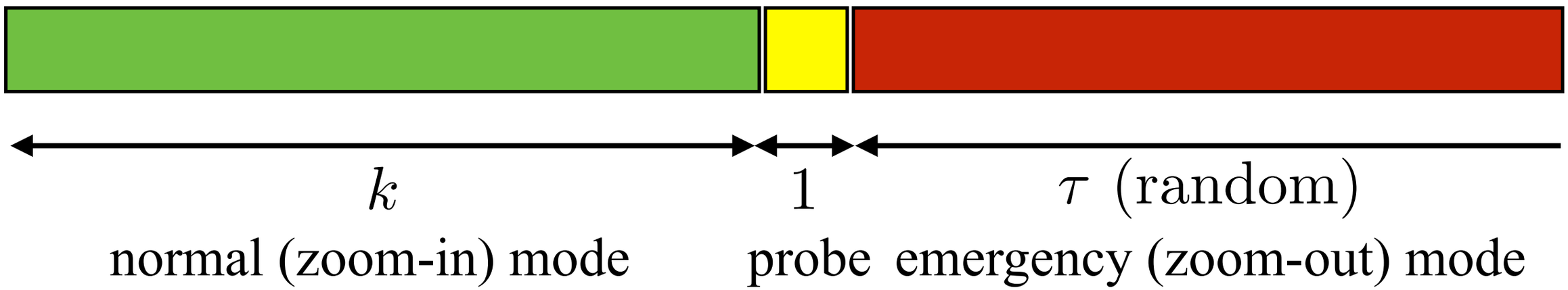
Let $a \in [1, 2)$.

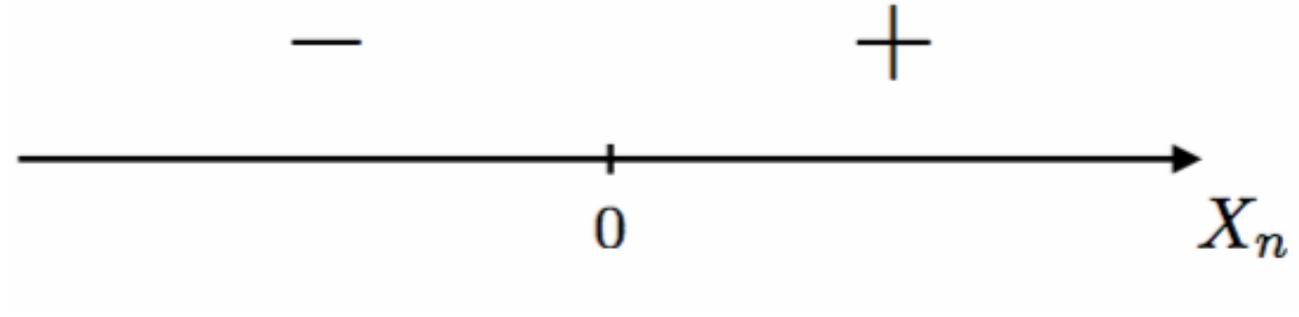
How to stabilize with 1 bit



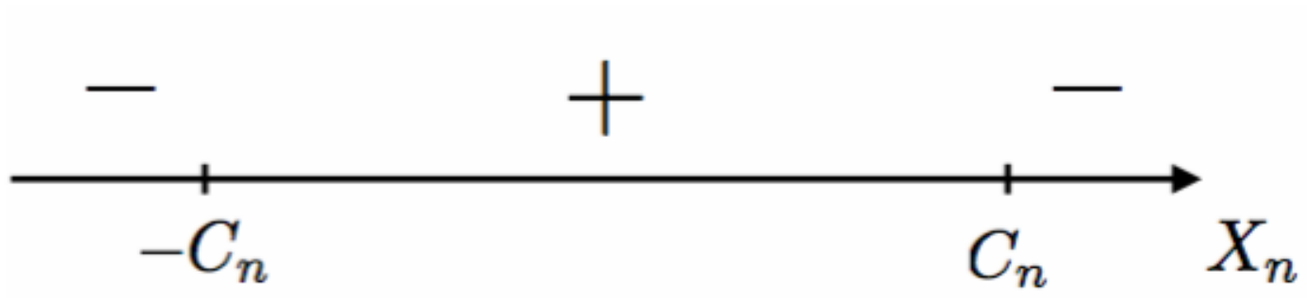
Achievable scheme

- Proceed in *rounds* of at least $k+1$ moves:





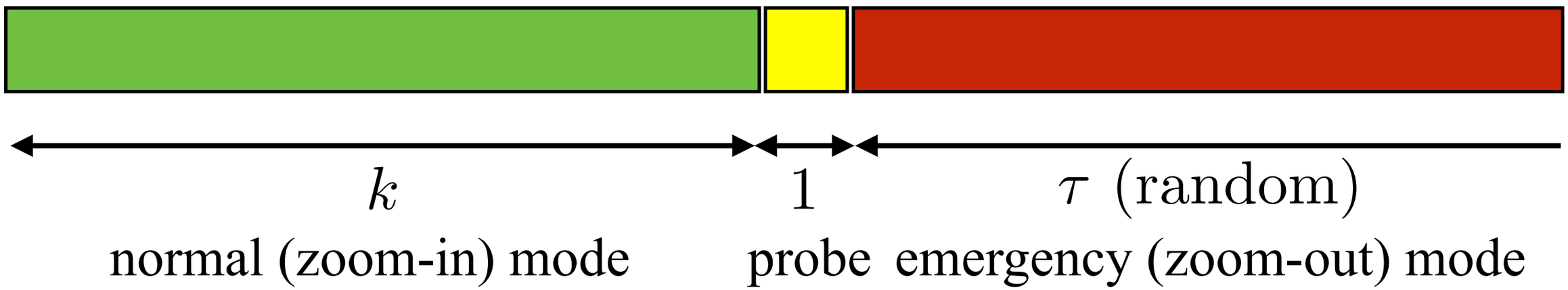
(a) Sign test



(b) Magnitude test

Achievable scheme

- Proceed in *rounds* of at least $k+1$ moves:

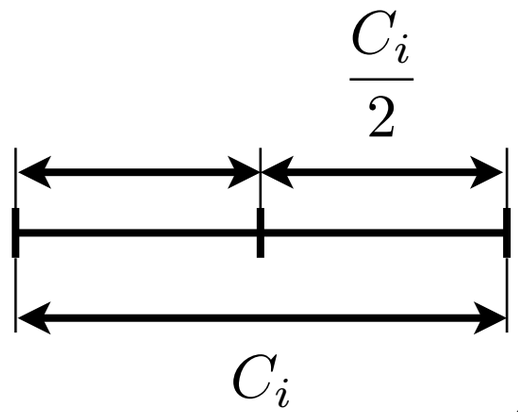


In *normal*, or *zoom-in*, mode, **pretend** that $|V_i| \le B$:

- Recall: in the bounded case,
  if $|X_i| \le C_i$, then $|X_{i+1}| \le \frac{aC_i}{2} + B$.

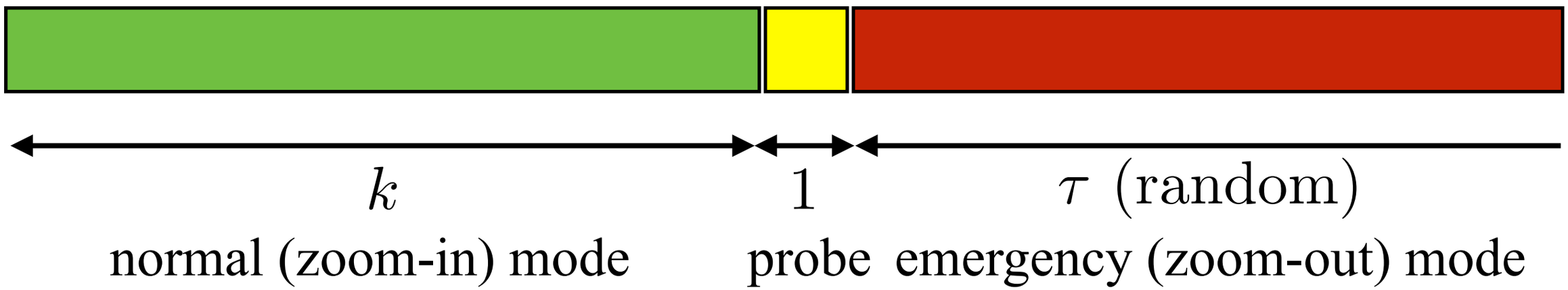
- So, update rule for C_i: $\quad C_{i+1} = \dfrac{aC_i}{2} + B.$

- Recall: $C_i \downarrow \frac{B}{1-a/2}$.



Achievable scheme

- Proceed in *rounds* of at least $k+1$ moves:

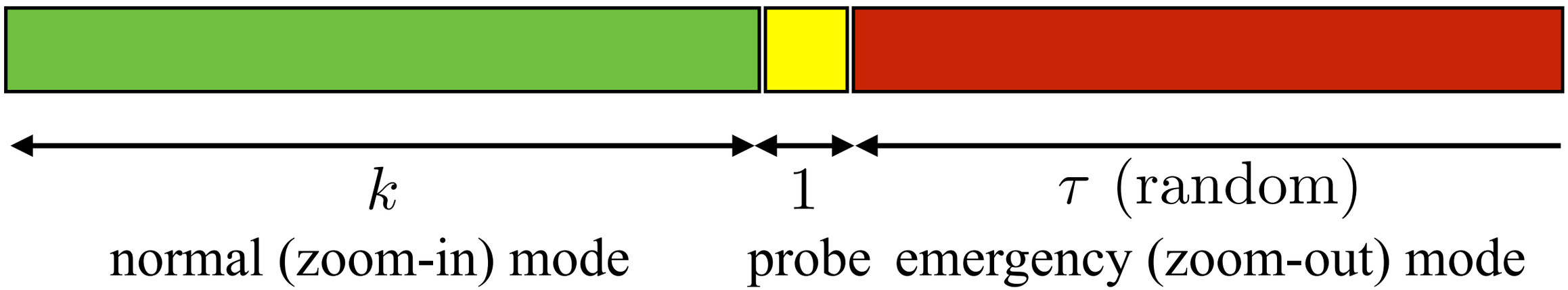


The *probe*: test whether the X_i is staying within desired bounds:

- The quantizer applies the magnitude test to check whether $|X_{m+k}| \leq C_{m+k}$. The controller is silent ($U_{m+k} = 0$).
- If $|X_{m+k}| \leq C_{m+k}$, we return to normal mode.
- Otherwise, we enter the *emergency*, or *zoom-out*, mode.

Achievable scheme

- Proceed in *rounds* of at least $k+1$ moves:



Emergency, or *zoom-out*, mode:

- Repeatedly perform more magnitude tests via $C_{m+k+j} = P\, C_{m+k+j-1}$.
- Return to normal mode the first time the test is passed.

Analysis

After a round is completed, $\mathbb{E}\left[|X_i|^\beta\right]$ tends to decrease compared to the start of a round.

Remark

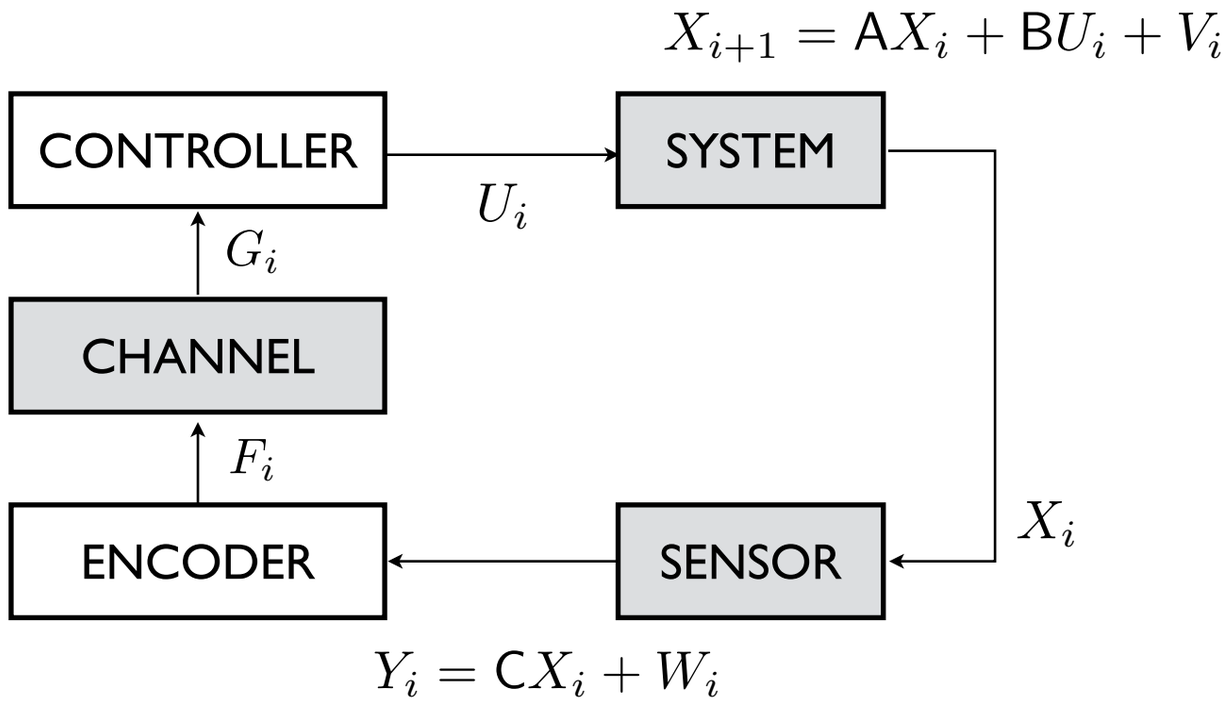
- In the achievability, we do not need the assumption that X_0 has a density.
- However, for the converse the assumption is not superficial.
- For example, consider $V_i \equiv 0$ and X_0 uniformly distributed on the Cantor set, and $a = 2.9$. This system can be stabilized with 1 bit, by telling the controller at each step the undeleted third of the interval the state is at.

Control with fixed rate: research directions

- Control over noisy channels
  - only special cases (packet drop channel with feedback, scalar Gaussian channel) are solved; achievability bounds for the BSC
  - This is a joint source-channel coding problem

- A refined control objective: achieving a specific bound on some moment of the system state instead of only requiring its boundedness
  - wide open!

Coding for control



We would like to choose

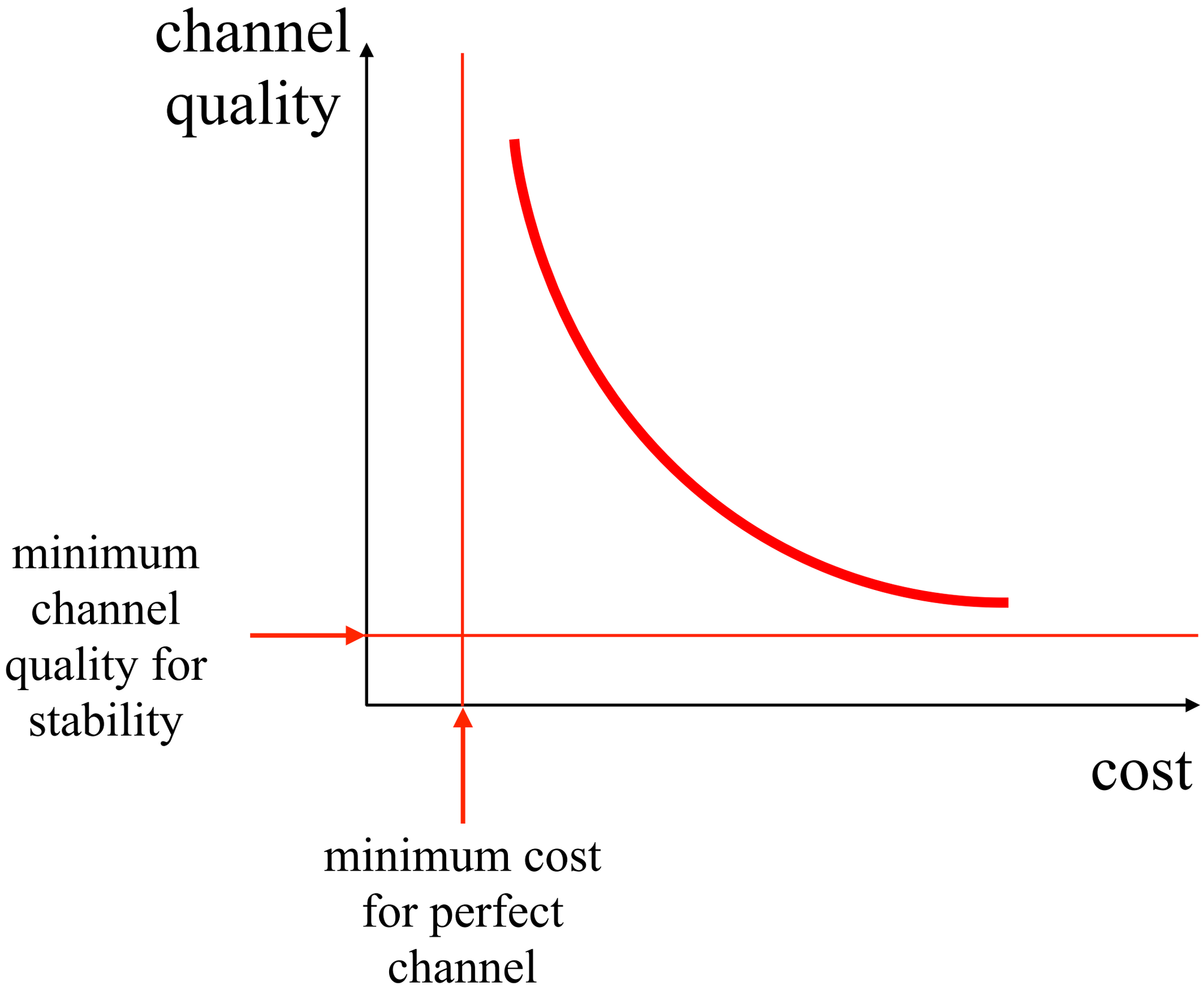
- the encoding sequence $F_1, \ldots, F_t$,
- the control sequence $U_1, \ldots, U_t$.

to minimize

Linear quadratic regulator

$$\mathrm{LQR}(X^t, U^{t-1}) \triangleq \mathbb{E}\left[\sum_{i=1}^{t-1}\left(X_i^T \mathsf{Q} X_i + U_i^T \mathsf{R} U_i\right) + X_t^T \mathsf{S}_t X_t\right]$$

Goal: information-theoretic tradeoffs



Separation between control and communication

Assuming past controls U_i are available at the encoder, control-communication *separation* holds:

$$\text{total LQR cost} = \{\text{control cost assuming controller observes } X_i\}$$
$$+ \{\text{distortion between } X_i \text{ and } \bar{X}_i = \mathbb{E}\left[X_i | Y^i, U^{i-1}\right]\}$$
$$+ \{\text{distortion between } X_i \text{ and its estimate at the controller}\}$$

Since the controls are additive and the distortion is shift-invariant, encoding-decoding policy does not affect the first two terms!

Thus, the problem reduces to *tracking* the source $\bar{X}_i$ under distortion (a causal rate-distortion problem)

V. Kostina, B. Hassibi. Rate-cost tradeoffs in control. IEEE Transactions on Automatic Control, 2019

Gauss-Markov source: a simple source with memory

$$X_{i+1} = aX_i + V_i$$

$$X_1, \{V_i\}_{i=1}^{\infty} \sim \mathcal{N}\left(0, \sigma^2\right) \text{ i.i.d.}$$

- $|a| < 1$: asymptotically stationary source
- $|a| \geq 1$: nonstationary source
- $|a| = 1$: the Wiener process

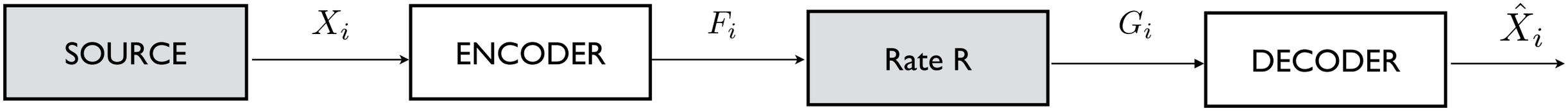
Noiseless variable-rate channel

The channel is *noiseless* and satisfies the *average* rate constraint:

$$\frac{1}{t}\sum_{i=1}^{t}\mathbb{E}\left[\ell_i\right] \leq R$$

length of data packet transmitted at time i

Zero delay causal tracking



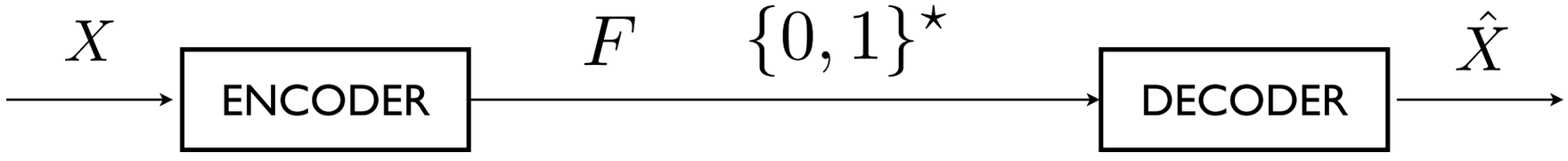
Goal: design (encoder, decoder) to minimize the MSE cost

$$\sum_{i=1}^{t} \mathbb{E}\left[(X_i - \hat{X}_i)^2\right]$$

$\hat{X}_i$ is the *real-time* estimate of X_i given everything the decoder knows up to time i

Encoder(s) and decoder have *memory* of the past.

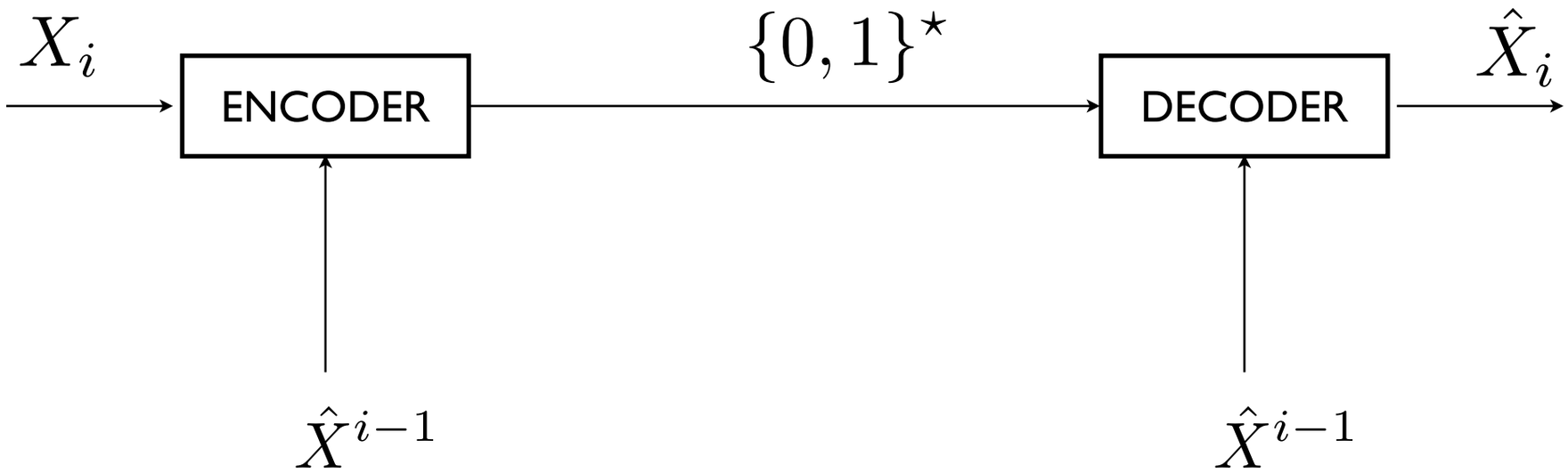
The classical rate-distortion function



Recall from Part I:

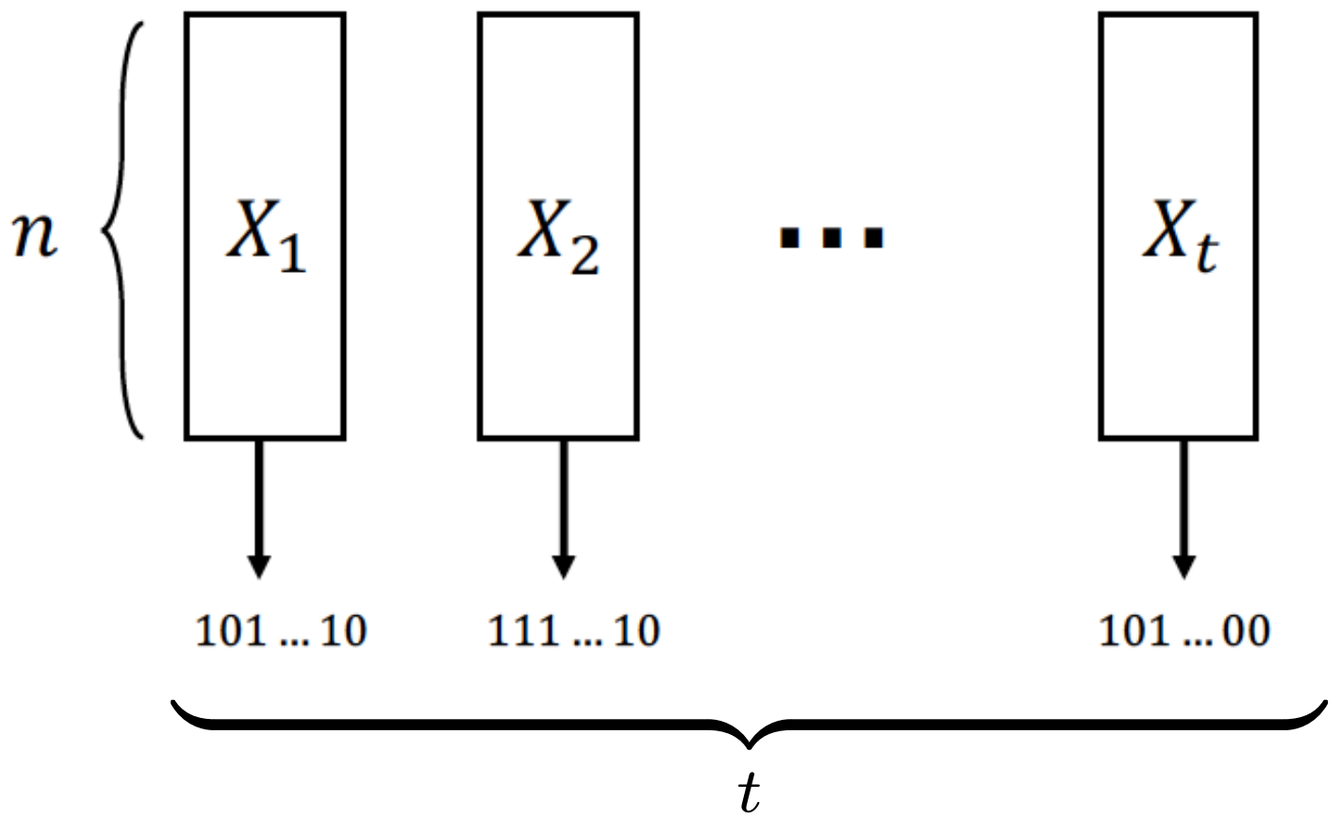
$$\min_{\substack{\mathsf{f},\mathsf{g}:\\ \mathsf{d}(X,\mathsf{g}(\mathsf{f}(X)))\leq d}} \mathbb{E}\left[\ell(\mathsf{f}(X))\right] = R_d(X) + O\left(\log R_d(X)\right)$$

Apply rate-distortion theorem at each step?



Does not quite work because of memory of the past:
$\hat{X}^{i-1}$ is the result of coding at previous steps and needs to be optimized over

Instead, a large time horizon



What is the total average minimum number of bits compatible with $\frac{1}{t}\sum_{i=1}^{t}\mathbb{E}\left[(X_i-\hat{X}_i)^2\right]\leq d$?

Causal conditioning (Kramer'98) and Bayes' rule

Any joint distribution $P_{X^tY^t}$ can be written in two ways:

- $P_{X^t}P_{Y^t|X^t}$
- $P_{X^t||Y^{t-1}}P_{Y^t||X^t}$ $\longleftrightarrow$ $X \rightleftarrows Y$ (feedback $\mathcal{D}$)

The two factorizations are the same iff there is no feedback

$$P_{Y^t||X^t} \triangleq \prod_{i=1}^{t} P_{Y_i|Y^{i-1},X^i}$$

$$P_{X^t||Y^{t-1}} \triangleq \prod_{i=1}^{t} P_{X_i|Y^{i-1},X^{i-1}}$$

Directed information (Massey, 1990)

Given a joint $P_{X^tY^t}$, we can remove feedback and create another joint $P_{X^t}P_{Y^t\|X^t}$ while preserving the marginals P_{X^t} and P_{Y^t}. The mutual information between X^t and Y^t according to the second joint is the *directed information*:

$$I(X^t \to Y^t) \triangleq \mathbb{E}\left[\log \frac{P_{Y^t||X^t}(Y^t||X^t)}{P_{Y^t}(Y^t)}\right]$$

$$I(X \to Y)$$

$$X \xrightarrow[\mathcal{D}]{} Y$$

$$I(Y \to \mathcal{D}X)$$

(Informational) causal rate-distortion function

Definition The causal rate-distortion function is defined by

$$\mathbb{R}_t(b) \triangleq \frac{1}{t} \inf_{\substack{P_{\hat{X}^t \| X^t}: \\ \frac{1}{t}\mathbb{E}\left[(\hat{X}_i - X_i)^2\right] \le d}} I(X^t \to \hat{X}^t)$$

$$\mathbb{R}(b) \triangleq \limsup_{t \to \infty} \mathbb{R}_t(b)$$

We will see that this function has an operational meaning.

Causal rate-distortion function: Markov sources

Theorem For the source $X_{i+1} = aX_i + V_i$, where $V_i \sim P_V$ are i.i.d. (not necessarily Gaussian),

$$\mathbb{R}(b) \geq \frac{1}{2} \log \left(a + \frac{N(V)}{d} \right)$$

$$N(V) \triangleq \frac{1}{2\pi e} e^{2h(V)} \qquad \text{entropy power}$$

$$h(V) \triangleq -\int_{\mathbb{R}^n} f_V(v) \log f_V(v) dx \qquad \text{differential entropy}$$

For Gaussian V: $N(V) = \mathrm{Var}\,[V]$. [Gorbunov, Pinsker '74] [Tatikonda, Sahai, Mitter '04]

V. Kostina, B. Hassibi. Rate-cost tradeoffs in control. IEEE Transactions on Automatic Control, 2019

A matching upper bound

Theorem There exists a variable-length quantizer achieving MSE error d with output entropy:

$$\mathbb{H}(d) \leq \frac{1}{2}\log\left(a + \frac{N(V)}{d}\right) + O_1 + O_2\left(d^{\frac{1}{2}}\right).$$

O_1: space-filling loss ($\downarrow 0$ as $n \uparrow \infty$)

O_2: high definition ($\downarrow 0$ as $d \downarrow 0$)

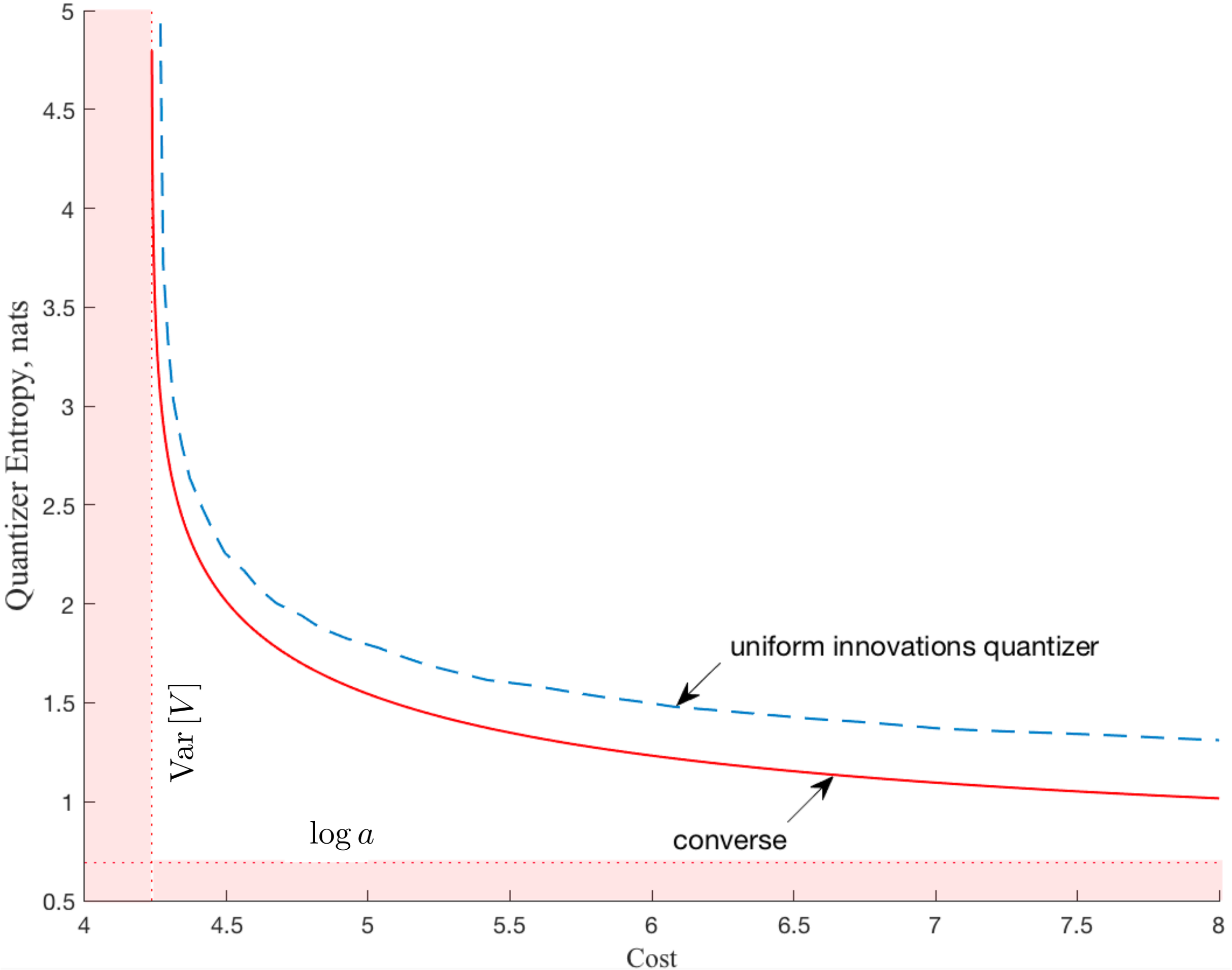
V. Kostina, B. Hassibi. Rate-cost tradeoffs in control. IEEE Transactions on Automatic Control, 2019

Achievable scheme: DPCM

- Encoder's state estimate before transmission: $\hat{X}_i = a\hat{X}_{i-1}$.
- Encoder sends quantized value of *innovation*, $\mathsf{q}(X_i - \hat{X}_i)$.

- Decoder's state estimate after transmission:

$$\hat{X}_{i+1} = \hat{X}_i + \mathsf{q}(X_i - \hat{X}_i)$$


Quantizer Entropy, nats
Var [V]
log a
uniform innovations quantizer
converse
Cost
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
4
4.5
5
5.5
6
6.5
7
7.5
8

Rate-limited control: research directions

- Learning for control
  - Suppose that the system gain is fixed but unknown, how should we code for such a system?
  - This is a universal rate-distortion problem
- For the Gauss-Markov source and quadratic distortion, the informational causal rate-distortion function can be found exactly. Are there other examples of sources/distortion measures for which we can compute it?
  - Without causality constraints, the informational rate-distortion function is known in closed form for the Bernoulli source with Hamming distortion, any discrete source with logloss.
- Control with multiple observers
  - Causal coding counterparts of tractable multiuser IT problems

Part II: Takeaways

- Coding for control presents a set of challenges not found in traditional block coding
  - even rare error events are catastrophic if not taken care of
  - causality constraints render existing block codes not applicable
  - due to the delay constraint, source-channel separation does not hold
- Information theory provides tools to tackle these *nonasymptotic* problems
- This rich area of research on the intersection of information theory and control has a lot to gain from information theorists
- Industry interest in ultra-reliable real-time codes for streaming

Part III:

Coding for computation

Federated learning

- Edge devices each have access to local data
- The goal is to train a machine learning model
- The server coordinates informational exchanges
- The communication bottleneck slows down the convergence process

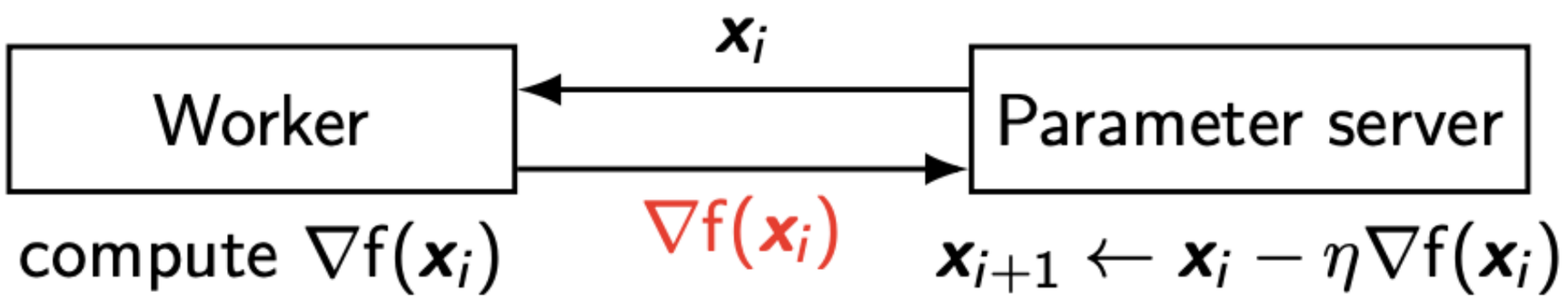
Unquantized gradient descent (GD)

$$\boxed{\text{Solve } \boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \mathsf{f}(\boldsymbol{x})}$$

- with GD:

$$\boldsymbol{x}_{i+1} \leftarrow \boldsymbol{x}_i - \eta\nabla\mathsf{f}(\boldsymbol{x}_i)$$

  - $\eta > 0$ is a constant stepsize.
- in distributed training



Active research area

Gradient quantization

- Stochastic gradient descent
  - Seide et al. 2014
  - Wen et al. 2017
  - Alistarh et al. 2017
  - Bernstein et al. 2018
  - Wu et al. 2018
  - Gandikota et al. 2019
  - Ramezani-Kebrya et al. 2019
  - Mayekar & Tyangi 2019, 2020
- GD
  - Luo & Tseng 1993
  - Friedlander & Schmidt 2012
  - Alistarh et al. 2016

Gradient sparsification

- Strom et al. 2015
- Aji & Heafield 2017
- Lin et al. 2018
- Wangni et al. 2018
- Wang et al. 2018
- Stich et al. 2018

Convergence lower bounds are scarse.

We are interested in the necessary and sufficient *bit rate* to attain a target *convergence rate*.

Class of functions

$$\mathcal{F}_n(\mu, L, r) \triangleq \{\mathsf{f} : \mathbb{R}^n \to \mathbb{R} \mid \mathsf{f} \text{ satisfies the following.}\}$$

- f is L-smooth: $\|\nabla \mathsf{f}(\boldsymbol{v}) - \nabla \mathsf{f}(\boldsymbol{w})\| \leq L \|\boldsymbol{v} - \boldsymbol{w}\|$
- f is μ-strongly convex: $(\nabla \mathsf{f}(\boldsymbol{v}) - \nabla \mathsf{f}(\boldsymbol{w}))^{\mathrm{T}} (\boldsymbol{v} - \boldsymbol{w}) \geq \mu \|\boldsymbol{v} - \boldsymbol{w}\|^2$
- The minimizer $\|\boldsymbol{x}^*(\mathsf{f})\| \leq r$ for some $r > 0$

Unquantized gradient descent

For *any* L-smooth and μ-strongly convex f, unquantized GD satisfies [Polyak, 1987]:

$$\|\boldsymbol{x}_T - \boldsymbol{x}^*(\mathrm{f})\| \leq \sigma^T \|\boldsymbol{x}_0 - \boldsymbol{x}^*(\mathrm{f})\|$$

- $\sigma \triangleq \frac{L-\mu}{L+\mu}$: contraction factor of GD.
- The bound is tight: $\exists$f s.t. "=" holds.

Quantizers for worst-case distortion

A *quantizer of dimension n and rate R* is a function $\mathsf{q}\colon \mathcal{D} \to \mathbb{R}^n$, where $\mathcal{D} \subseteq \mathbb{R}^n$ is the domain, such that the image of q satisfies

$$|\mathrm{Im}(\mathsf{q})| = 2^{nR}.$$

Quantizers for worst-case distortion

For a bounded-domain quantizer $\mathsf{q}\colon \mathcal{D} \to \mathbb{R}^n$, we refer to

$$r(\mathsf{q}) \triangleq \max\{\delta\colon \mathcal{B}(\delta) \subseteq \mathcal{D}\}$$

as the *dynamic range* of q, to

$$\mathsf{d}(\mathsf{q}) \triangleq \min\{d\colon \forall \boldsymbol{x} \in \mathcal{D},\ \|\boldsymbol{x} - \mathsf{q}(\boldsymbol{x})\| \leq d\}$$

as its *covering radius*, and to

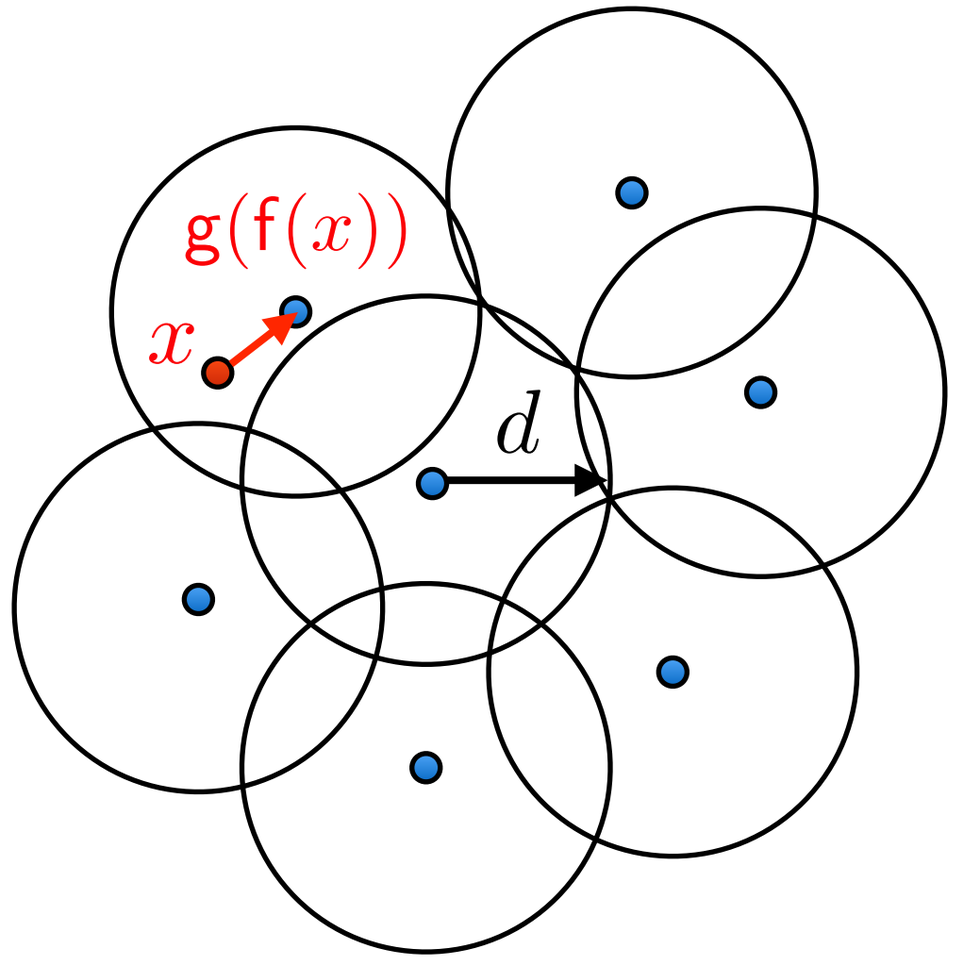
$$\rho\,(\mathsf{q}) \triangleq |\mathrm{Im}(\mathsf{q})|^{1/n}\ \frac{\mathsf{d}(\mathsf{q})}{r(\mathsf{q})}$$

as its *covering efficiency*.

Quantizers for worst-case distortion

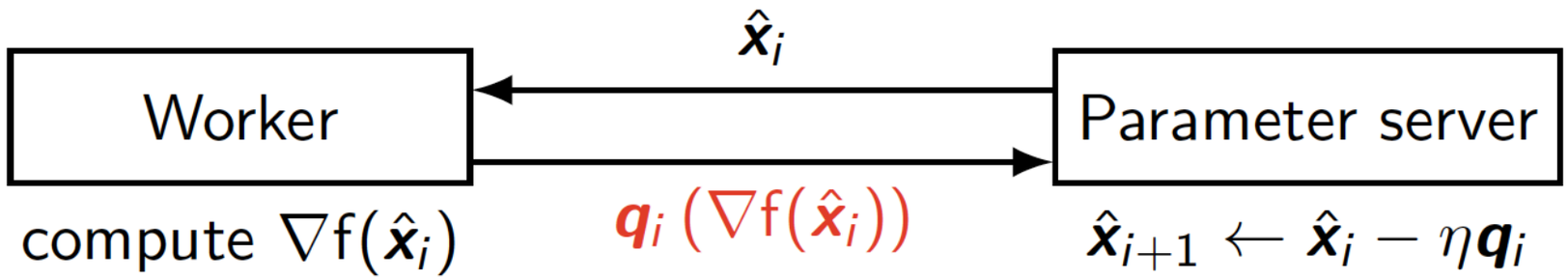
ρ_n: covering efficiency of the quantizer

- uniform scalar quantizer: $\rho_n = \sqrt{n}$.
- $\rho_n \geq 1$.
- $\rho_n = 1 + o_n(1)$ is achievable with lattice quantizers [Rogers 1963].



Naively quantized gradient descent

NQ-GD [Friedlander & Schmidt 2012, Alistarh et al. 2016] directly quantizes the gradient at $\hat{\boldsymbol{x}}_i$:

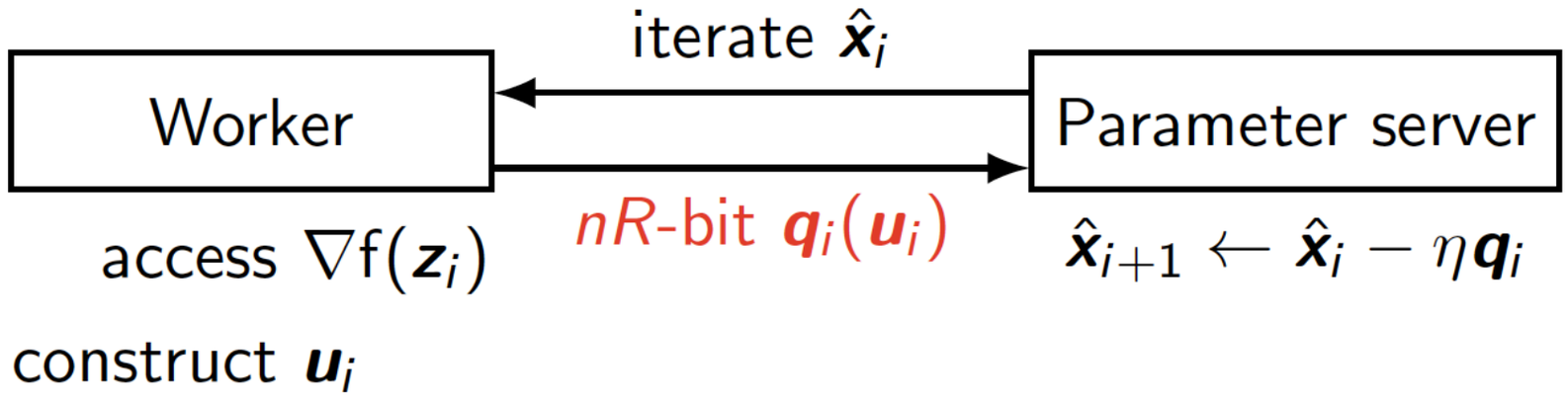


Theorem: NQ-GD

NQ-GD achieves the following contraction factor over $\mathcal{F}_n$

$$\sigma_{\text{NQ-GD}}(n, R) \leq \sigma + \rho_n 2^{-R}$$

Quantized gradient descent: class of algorithms



The worker, based on $\hat{\boldsymbol{x}}_i$ and $\boldsymbol{e}_0, \ldots, \boldsymbol{e}_{i-1}$ ($\boldsymbol{e}_\ell \triangleq \boldsymbol{q}_\ell - \boldsymbol{u}_\ell$), decides:

- gradient query point $\boldsymbol{z}_i$
- quantizer's input $\boldsymbol{u}_i$.

Goals:

- characterize the tradeoff between how fast any QGD algorithm converges and R.
- propose an algorithm that achieves it.

Quantized gradient descent: operational tradeoff

For a QGD algorithm A operating at R bits per problem dimension, worst-case (over $\mathsf{f} \in \mathcal{F}_n(\mu, L, r)$) contraction factor:

$$\sigma_{\mathrm{A}}(n, R) \triangleq \sup_{\mathsf{f} \in \mathcal{F}_n} \limsup_{T \to \infty} \left\| \hat{\mathbf{x}}_T(R) - \mathbf{x}^*(\mathsf{f}) \right\|^{\frac{1}{T}}$$

- unquantized GD: $\sigma_{\mathrm{GD}}(n, \infty) = \sigma = \frac{L-\mu}{L+\mu}$.

Quantized gradient descent: converse

Theorem: converse

The contraction factor of any QGD algorithm A operating at R bits per problem dimension satisfies

$$\sigma_{\mathrm{A}}(n, R) \geq \max\left\{\sigma,\, 2^{-R}\right\}$$

Proof combines two converses:

- Reduction to unquantized GD: $\sigma_{\mathrm{A}}(n, R) \geq \sigma$;
- Volume division argument: $\sigma_{\mathrm{A}}(n, R) \geq 2^{-R}$.

C. Lin, V. Kostina, B. Hassibi, "Differentially quantized gradient methods", IEEE Transactions on Information Theory, Apr. 2022

Quantized gradient descent: achievability

Algorithm 1: DQ-GD

Initialize $\boldsymbol{e}_{-1} \leftarrow \mathbf{0}$

for $i = 0$ **to** $T-1$ **do**

  Worker:

    $\boldsymbol{z}_i \leftarrow \hat{\boldsymbol{x}}_i + \eta \boldsymbol{e}_{i-1}$

    $\boldsymbol{u}_i \leftarrow \nabla \mathrm{f}(\boldsymbol{z}_i) - \boldsymbol{e}_{i-1}$

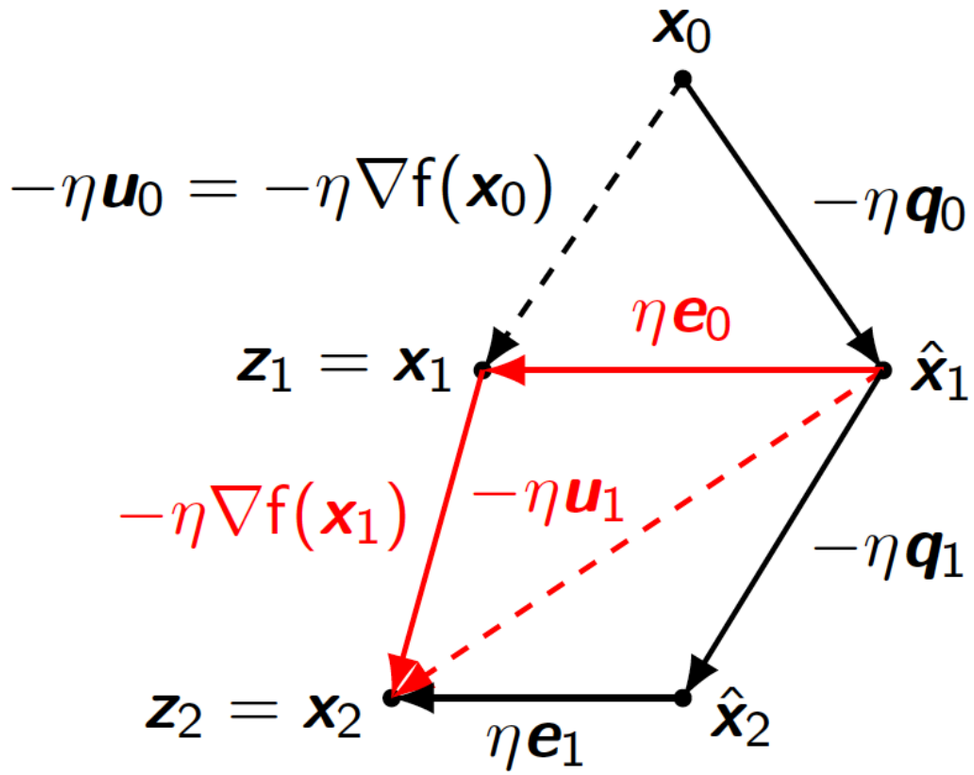
    $\boldsymbol{q}_i = \mathrm{q}_i(\boldsymbol{u}_i)$

    $\boldsymbol{e}_i \leftarrow \boldsymbol{q}_i - \boldsymbol{u}_i$

  Parameter server:

    $\hat{\boldsymbol{x}}_{i+1} \leftarrow \hat{\boldsymbol{x}}_i - \eta \boldsymbol{q}_i$

end



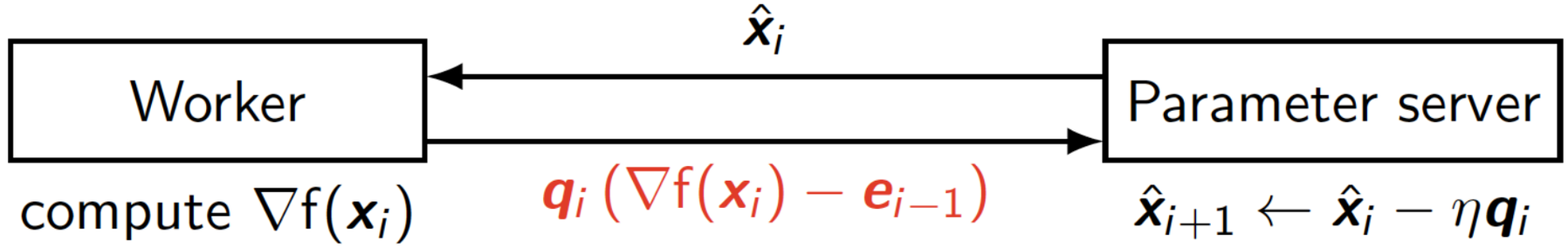
Differential quantization[1] directs the quantized trajectory to the unquantized trajectory.

[1] The idea of error compensation dates back to $\Sigma\Delta$ modulation [Gray, 1989].

C. Lin, V. Kostina, B. Hassibi, "Differentially quantized gradient methods", IEEE Transactions on Information Theory, Apr. 2022

Quantized gradient descent: achievability

DQ-GD computes the gradient at $\boldsymbol{x}_i$ and compensates the previous quantization error:



Theorem: DQ-GD

DQ-GD achieves the following contraction factor over $\mathcal{F}_n$

$$\sigma_{\text{DQ-GD}}(n, R) \leq \max\left\{\sigma,\ \rho_n 2^{-R}\right\}$$

- Since $\rho_n \to 1$ is achievable [Rogers, 1963], DQ-GD attains the converse as $n \to \infty$
- $R \geq \log_2 \rho_n/\sigma$: achieves the contraction factor σ of unquantized GD
- $R < \log_2 \rho_n/\sigma$: achieved contraction factor is only $\rho_n 2^{-R}$

C. Lin, V. Kostina, B. Hassibi, "Differentially quantized gradient methods", IEEE Transactions on Information Theory, Apr. 2022

Achievability and converse together

Optimal contraction factor over $\mathsf{f} \in \mathcal{F}_n$ and $\mathrm{A} \in \mathrm{QGD}$

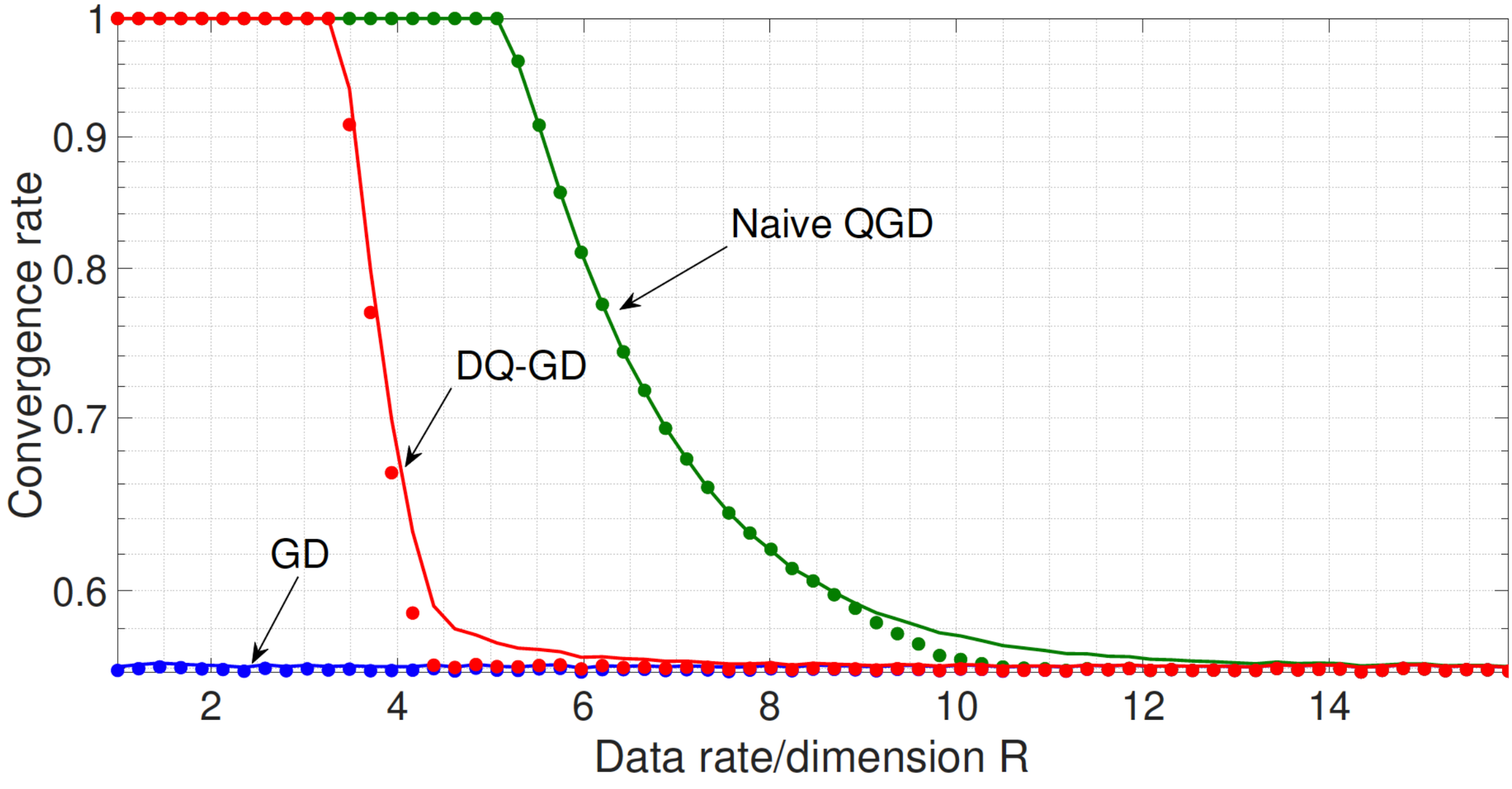
$$\lim_{n\to\infty} \inf_{\mathrm{QGD}\ \mathrm{A}} \sigma_{\mathrm{A}}(n, R) = \max\left\{\sigma, 2^{-R}\right\}.$$

Phase transition

- $R \geq \log_2 1/\sigma$: contraction factor σ of unquantized GD is achievable
- $R < \log_2 1/\sigma$: only 2^{-R} is achievable

C. Lin, V. Kostina, B. Hassibi, "Differentially quantized gradient methods", IEEE Transactions on Information Theory, Apr. 2022

Least-squares problems: Gaussian ensemble



$f(\boldsymbol{x}) = \|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|^2/2$. $\mathbf{A} \in \mathbb{R}^{1000\times 100}$ and $\boldsymbol{y} \in \mathbb{R}^{1000}$ iid standard normal entries. $\kappa(\mathbf{A}) \approx 1.8862$ on average.

C. Lin, V. Kostina, B. Hassibi, "Differentially quantized gradient methods", IEEE Transactions on Information Theory, Apr. 2022

Coding for computation: research directions

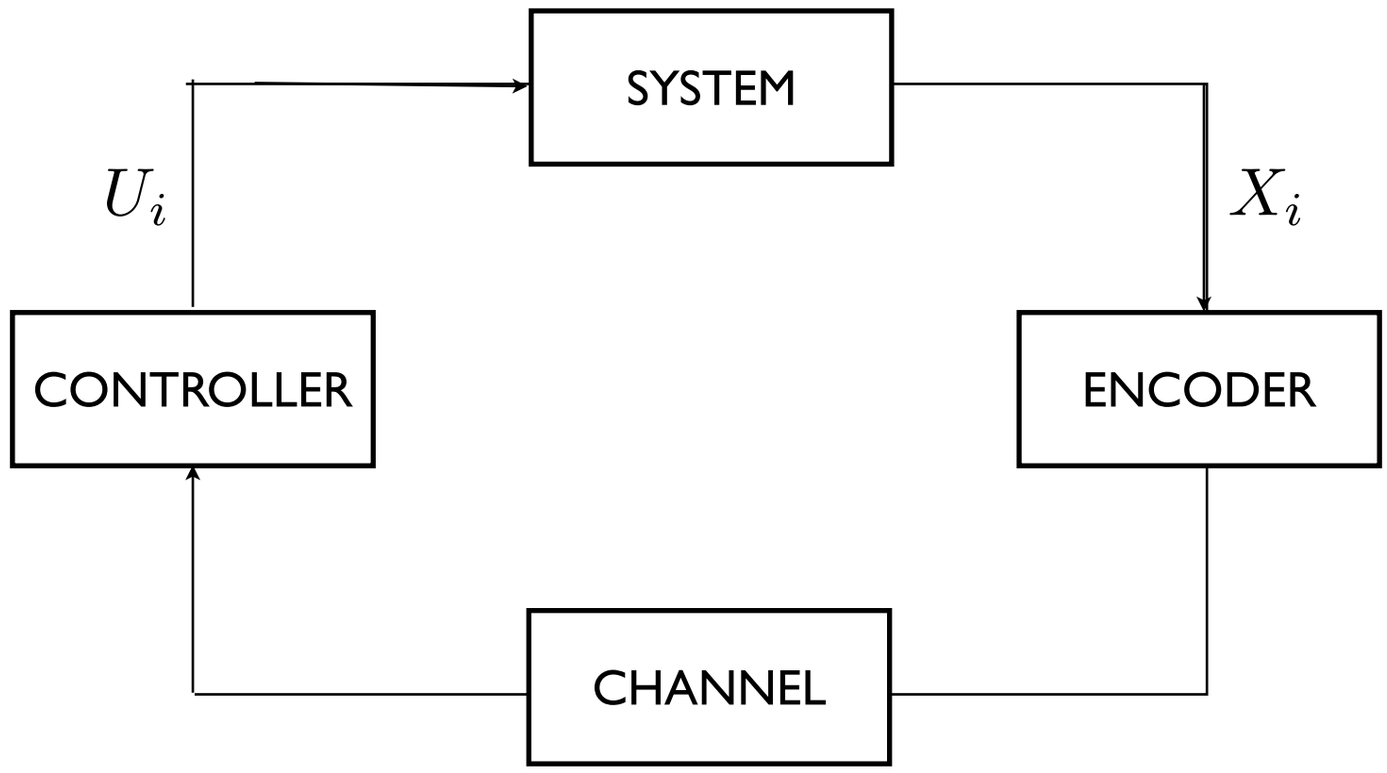
- Multiuser rate-convergence tradeoff
  - Considerably more challenging (no wonder since even classical multi-compressor rate-distortion problems are challenging)
  - High rate asymptotics?
- Rate-convergence tradeoff for a wider class of iterative algorithms
  - Differential quantization extends to accelerated GD algorithms
  - What about stochastic GD?
- Rate-convergence tradeoff for a wider class of functions
  - Smoothness and strong convexity are strong assumptions. But, tools exist to study convergence without these assumptions.
- Rate-convergence tradeoff over noisy channels

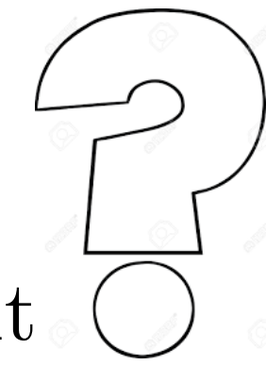
Part III: Takeaways

- Coding for computation presents a set of challenges similar to that in control
  - communication is interactive
  - delay-sensitive
  - The objective of communication is important for system design
- Like coding for control, it is a source coding problem
- Communication is a real bottleneck in large-scale optimization
- This hot area of research has a lot to gain from information theorists (many algorithms in existing literature but few converses)

Open problem A: control over a channel

$$X_{i+1} = aX_i + U_i + V_i$$

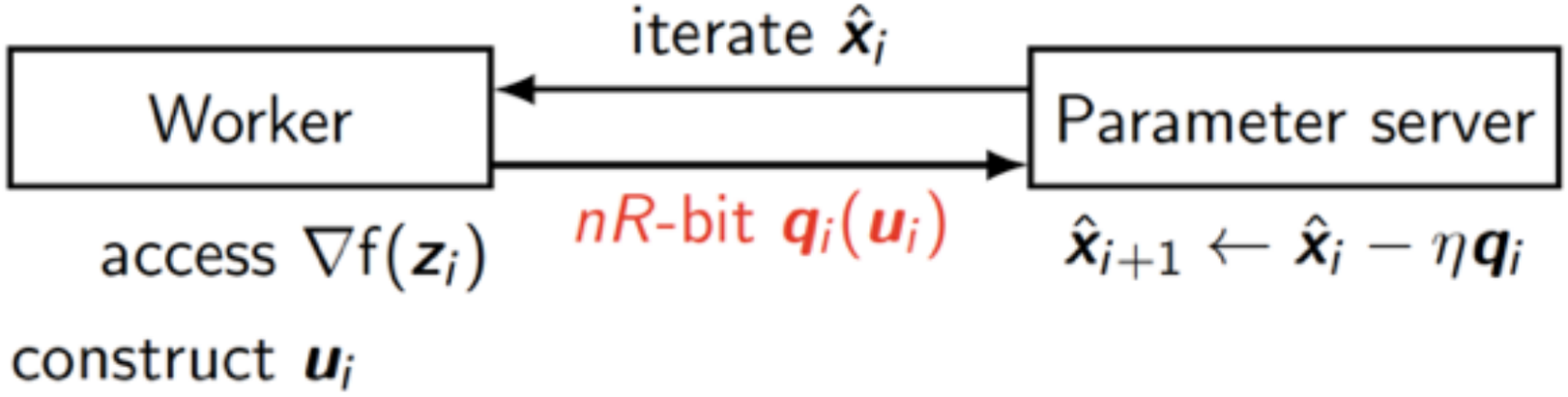




If $a \in [1, 2)$, can we achieve $\lim_{i\to\infty} \mathbb{E}\left[|X_i|^2\right] < \infty$ with 1 bit

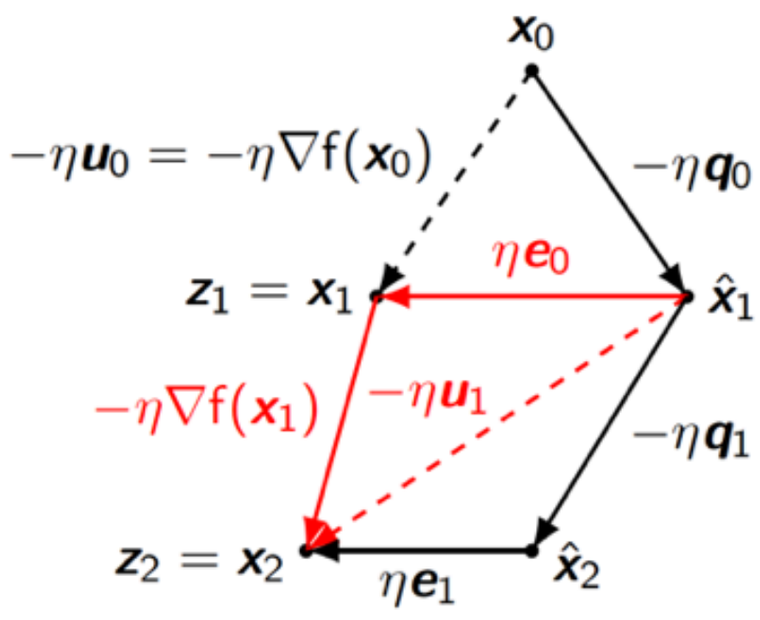
Binary-input binary-output channel: consider a simple channel (choose binary erasure, binary symmetric, or a channel that does not introduce more than pn errors out of n channel uses; assume perfect feedback from the controller to the encoder).

Open problem B: optimization over a channel



Algorithm 1: DQ-GD

Initialize $e_{-1} \leftarrow 0$
for $i = 0$ **to** $T-1$ **do**
  Worker:
    $z_i \leftarrow \hat{x}_i + \eta e_{i-1}$
    $u_i \leftarrow \nabla f(z_i) - e_{i-1}$
    $q_i = q_i(u_i)$
    $e_i \leftarrow q_i - u_i$
  Parameter server:
    $\hat{x}_{i+1} \leftarrow \hat{x}_i - \eta q_i$
end



Suppose the nR-bit codeword from the worker to the server can be corrupted by a noisy channel. Consider a simple channel. Is there an extension of the DQ algorithm to handle this?

Acknowledgements

- Collaborators: S. Verdú, Y. Polyankiy, Y. Peres, G. Ranade, M. Sellke, B. Hassibi, C.Y. Lin
- NSF support: CCF-1751356, CCF-1817241, and CCF-1956386