Probabilistic and Combinatorial Methods

Error-Correcting Codes: Theory and Practice Boot Camp

Jonathan Mosheiff
Ben-Gurion University

What these talks are about

What these talks are about

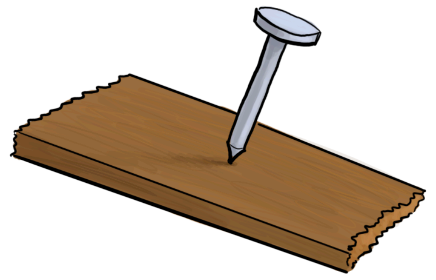
- Combinatorial questions (and some answers)

What these talks are about

- Combinatorial questions (and some answers)
- No algorithmic results! (But some algorithmic motivation)

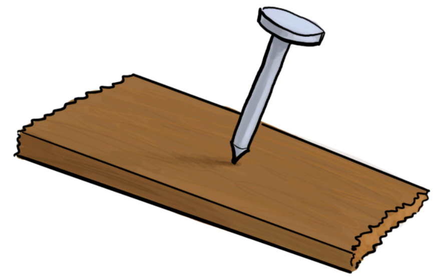
What these talks are about

- Combinatorial questions (and some answers)
- No algorithmic results! (But some algorithmic motivation)
- **Example motivation:** how List-decodable and list-recoverable are Reed-Solomon codes?



What these talks are about

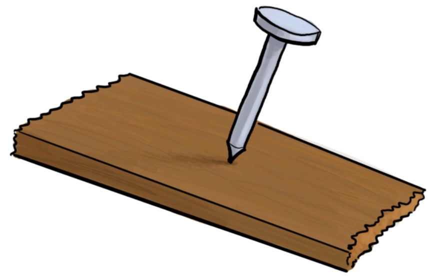
- Combinatorial questions (and some answers)
- No algorithmic results! (But some algorithmic motivation)
- **Example motivation:** how List-decodable and list-recoverable are Reed-Solomon codes?



- **A star player:** The Random Linear Code (RLC)

What these talks are about

- Combinatorial questions (and some answers)
- No algorithmic results! (But some algorithmic motivation)
- **Example motivation:** how List-decodable and list-recoverable are Reed-Solomon codes?



- **A star player:** The Random Linear Code (RLC)
- **Technique:** We **reduce** from RLC to more structured codes.

List-Decoding

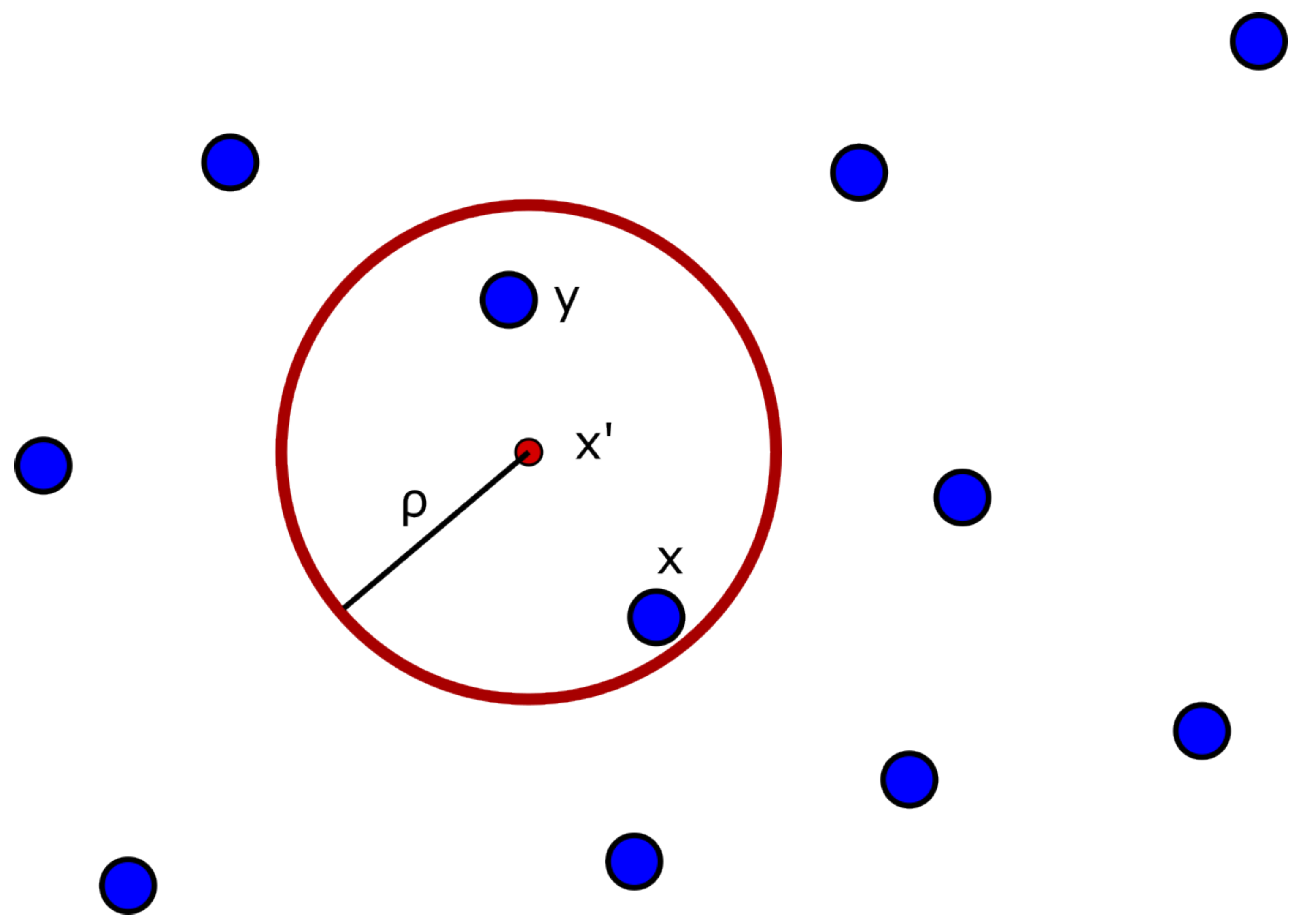
List-Decoding

- A code $C \subseteq \mathbb{F}_q^n$ is **ρ-uniquely-decodable** if the receiver can always **uniquely** recover a codeword $x \in C$ given ρn errors.

List-Decoding

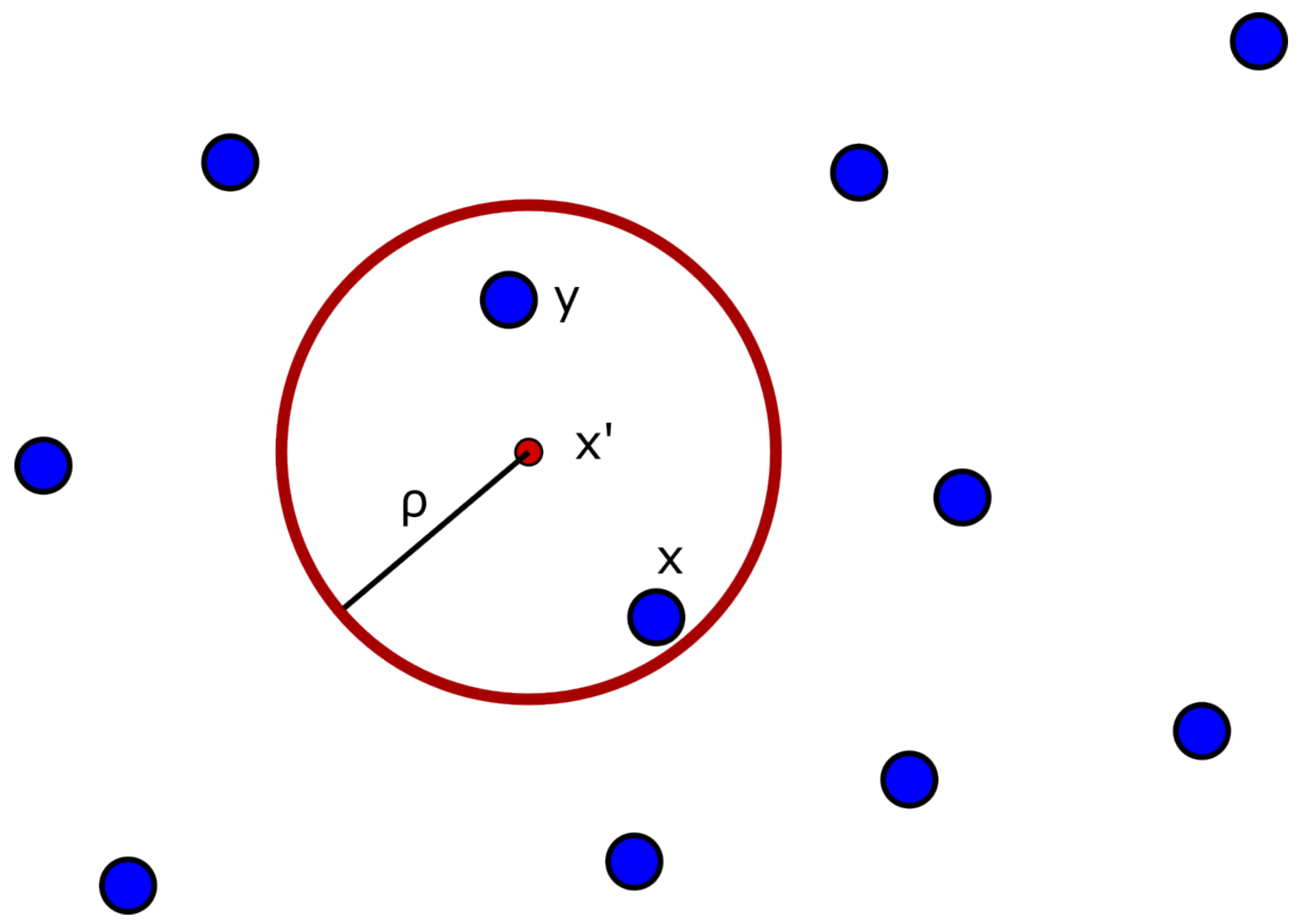
- A code $C \subseteq \mathbb{F}_q^n$ is **ρ-uniquely-decodable** if the receiver can always **uniquely** recover a codeword $x \in C$ given ρn errors.

- Namely, need to avoid this:



List-Decoding

- A code $C \subseteq \mathbb{F}_q^n$ is **ρ-uniquely-decodable** if the receiver can always **uniquely** recover a codeword $x \in C$ given ρn errors.
- Namely, need to avoid this:



- C is **(ρ,L)-list-decodable** if the receiver can always recover a **list** of at most L codewords, such that the list contains x.

List-Recovery

List-Recovery

- In **List-Decoding** we want every Hamming ball to contain a small number of codewords.

- In **List-Recovery** we care about **combinatorial rectangles** instead of **balls.**

List-Recovery

We say that $C \subseteq \mathbb{F}_q^n$ is **(ℓ, L)-list-recoverable** if:

For every $S_1, \ldots, S_n \subseteq \mathbb{F}_q$ with $|S_i| \leq \ell$ we have

$$|C \cap (S_1 \times S_2 \times \ldots \times S_n)| \leq L.$$

$S_1 \times S_2 \times \ldots \times S_n$ is called a **combinatorial rectangle**

Random Linear Codes (RLCs)

Random Linear Codes (RLCs)

- An **RLC** of **length** n and **rate** R over **alphabet** $\mathbb{F}_q$ is a uniformly-sampled Rn-dimensional linear subspace of $\mathbb{F}_q^n$.

Random Linear Codes (RLCs)

- An **RLC** of **length** n and **rate** R over **alphabet** $\mathbb{F}_q$ is a uniformly-sampled Rn-dimensional linear subspace of $\mathbb{F}_q^n$.
- The go-to code for **existence proofs**!

Random Linear Codes (RLCs)

Random Linear Codes (RLCs)

- **Achieves with high probability:**
  - The **Gilbert-Varshamov Bound** *
    $R \approx 1 - h_q(\delta)$
  - The **"List-decoding GV-bound":**
    $$R = 1 - h_q(\delta) - O\left(\frac{1}{L}\right)$$
  - **List-recovery** results as well.

* $H_q(\rho) = \rho \log_q(q-1) - \rho \log_q \rho - (1-\rho)\log_q(1-\rho)$

Random Linear Codes (RLCs)

- **Achieves with high probability:**
  - The **Gilbert-Varshamov Bound** *
    $R \approx 1 - h_q(\delta)$
  - The **“List-decoding GV-bound”:**
    $R = 1 - h_q(\delta) - O\left(\frac{1}{L}\right)$
  - **List-recovery** results as well.

- **However:**
  - Decoding is **probably hard**
  - Certifying is **probably hard**
  - Construction requires $\Theta\left(n^2\right)$ **random bits**.

* $H_q(\rho) = \rho \log_q(q-1) - \rho \log_q \rho - (1-\rho)\log_q(1-\rho)$

The only thing you need to know about RLCs

Let C be an **RLC** of rate R. Fix $v_1, \ldots, v_k \in \mathbb{F}_2^n$.

Then:

$$\Pr\left[\{v_1, \ldots, v_k\} \subseteq C\right] = 2^{-(1-R)\cdot n \cdot \dim\{v_1,\ldots,v_k\}}$$

List-Decodability of an RLC

List-Decodability of an RLC

- **Motivation:** Show that a binary **RLC** achieves the **list-decoding GV-bound**.

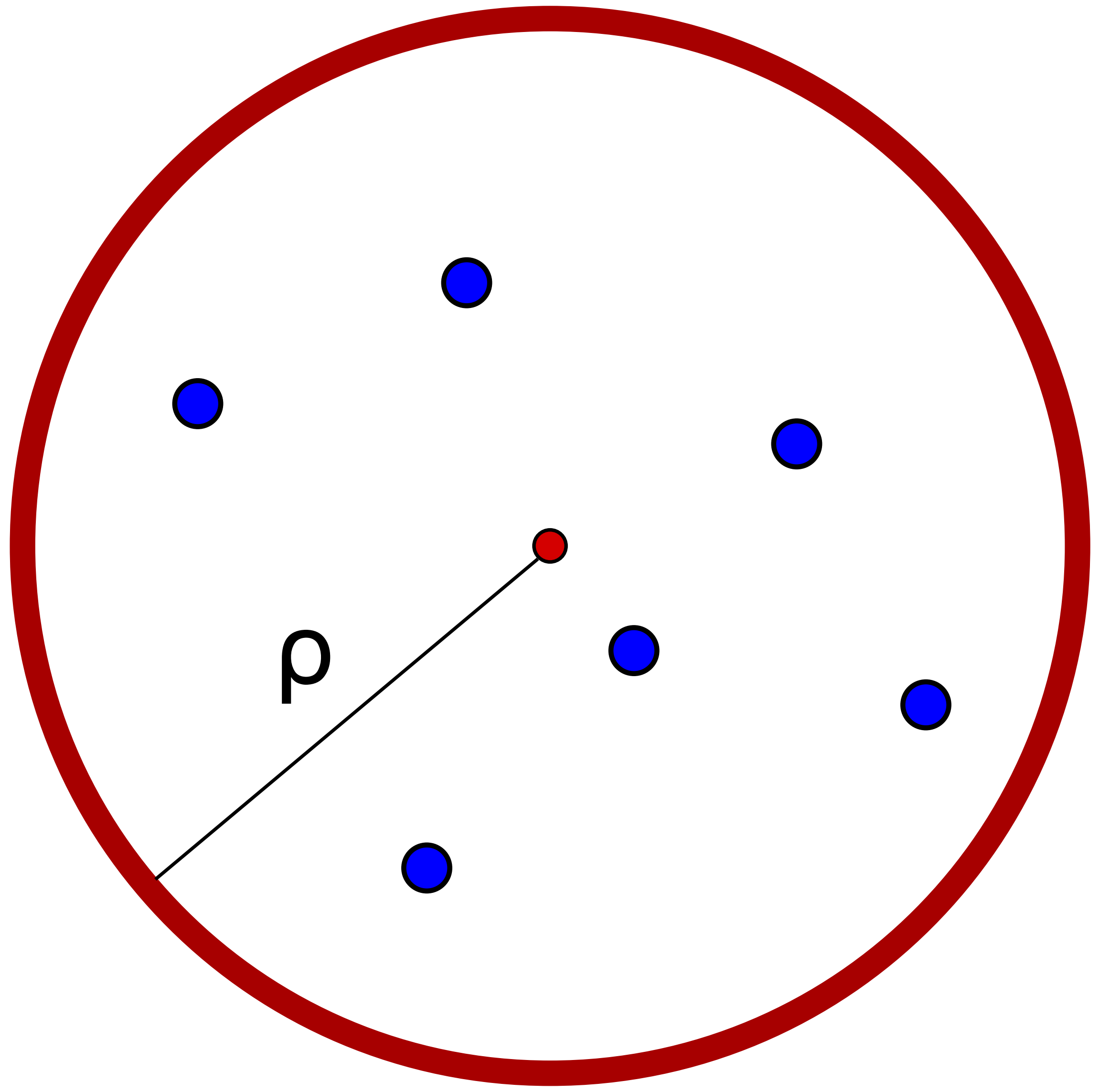
List-Decodability of an RLC

- **Motivation:** Show that a binary **RLC** achieves the **list-decoding GV-bound**.
- **More precisely**: Show that an **RLC** with $R = 1 - h(\rho) - \epsilon$ is $(\rho, O(1/\epsilon))$**-list-decodable** with high probability.

List-Decodability of an RLC

- Say that the vectors $x_1, \ldots, x_{L+1}$ are **ρ-clustered** if they are **distinct** and **contained in some radius ρ ball**.

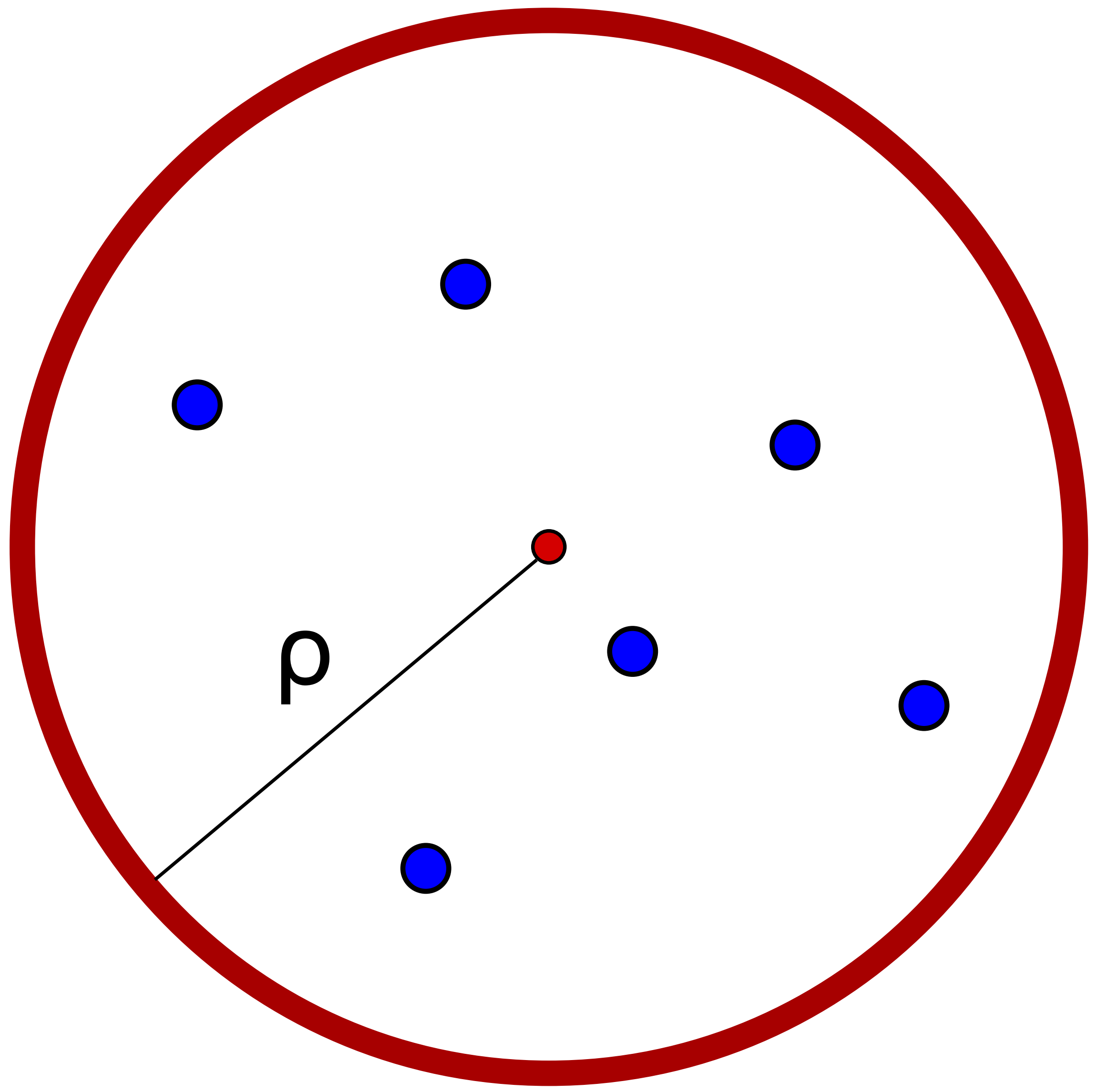
- The tuple $(x_1, \ldots, x_{L+1})$ is a **witness** to C **not being (ρ, L)-list-decodable.**



List-Decodability of an RLC

Let's try an expectation approach:

Try to Prove that the expected number of **clustered tuples** in an **RLC** is $o(1)$.



Is the expectation method tight?

Is the expectation method tight?

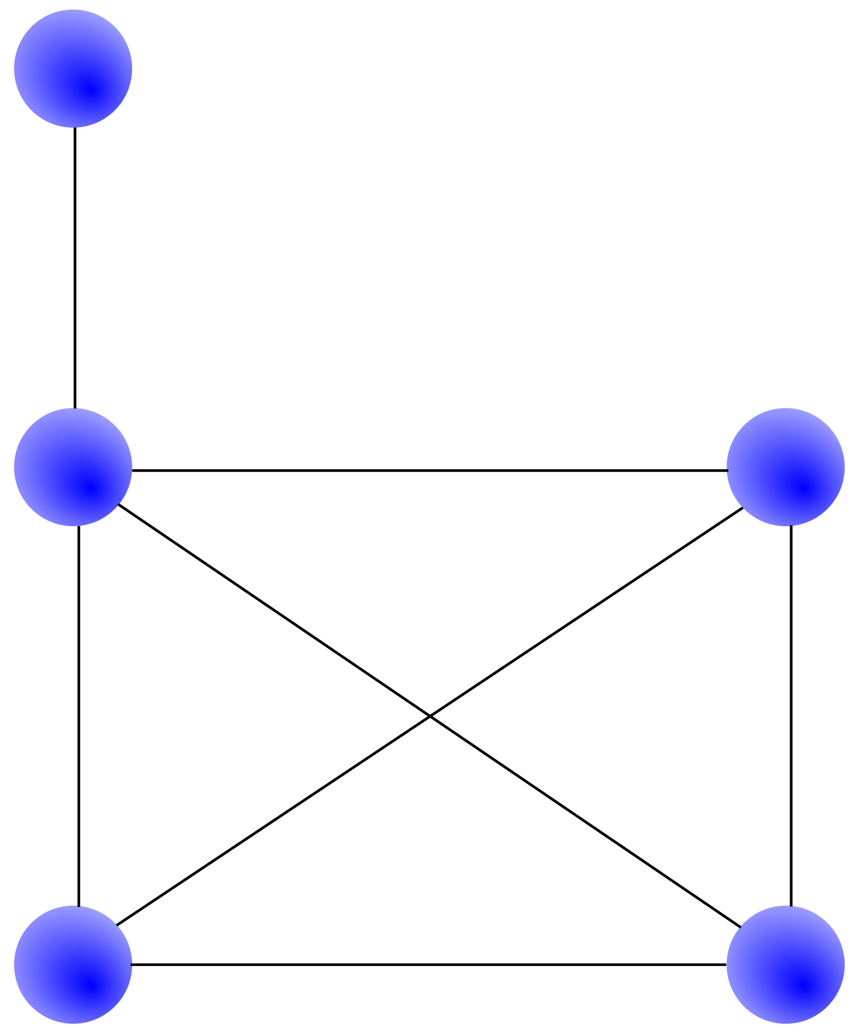
- Maybe not!

Is the expectation method tight?

- Maybe not!
- Analogy from **random $G(n, p)$ graphs**.

Is the expectation method tight?

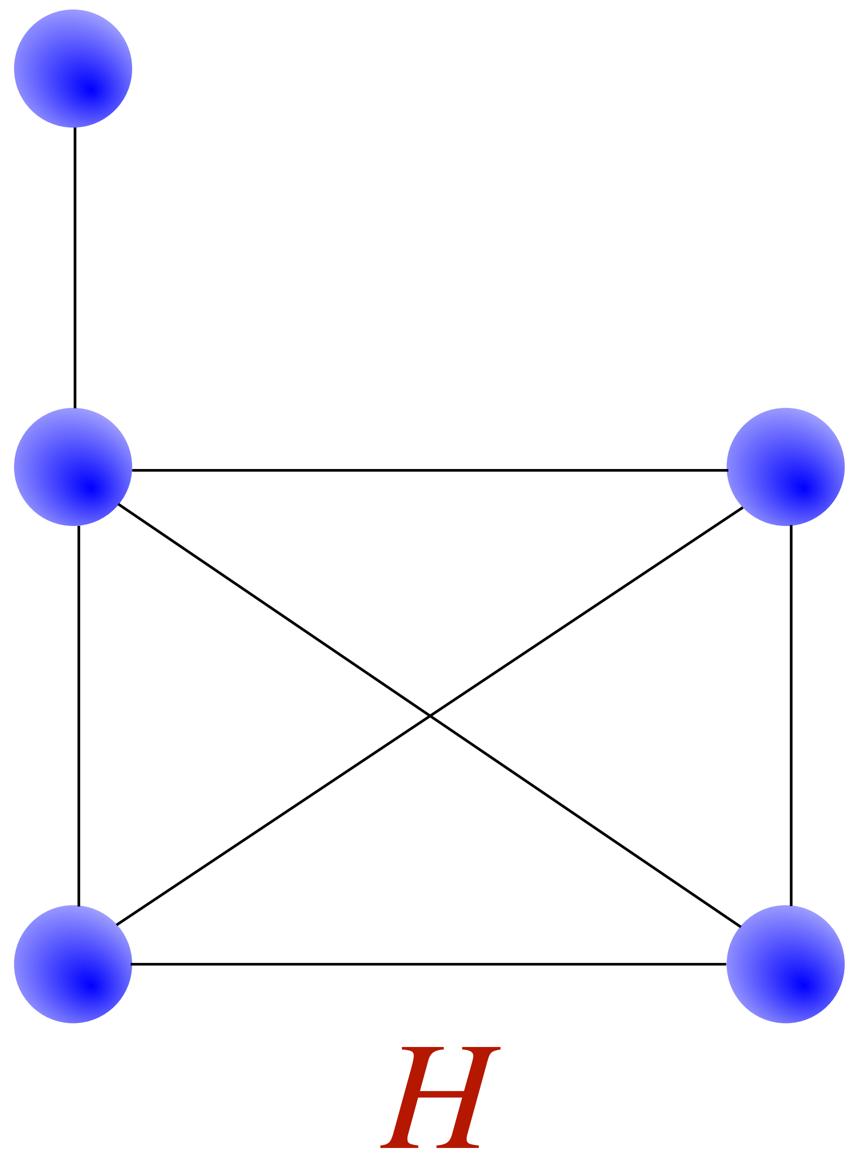
- Maybe not!
- Analogy from **random $G(n, p)$ graphs**.
- What is the probability that G **contains an H subgraph**?



H

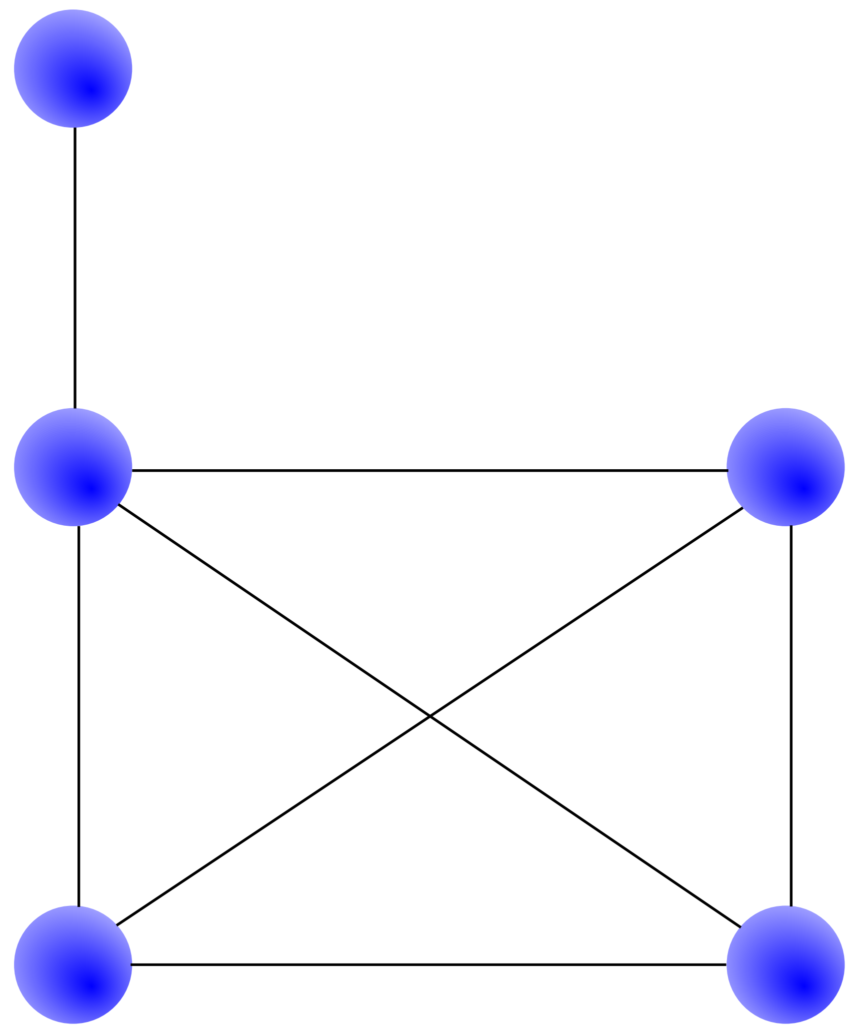
Is the expectation method tight?

- Maybe not!
- Analogy from **random $G(n, p)$ graphs**.
- What is the probability that G **contains an H subgraph**?
  - $\mathbb{E}(\#H \text{ in } G) \approx n^5 \cdot p^7$

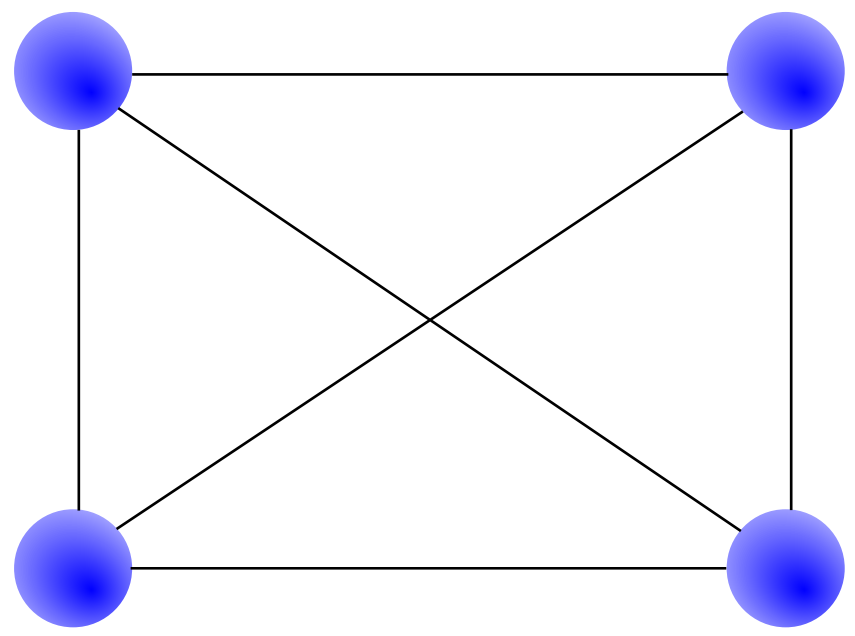


Is the expectation method tight?

- Maybe not!
- Analogy from **random $G(n, p)$ graphs**.
- What is the probability that G **contains an H subgraph**?
  - $\mathbb{E}(\#H \text{ in } G) \approx n^5 \cdot p^7$
  - $\mathbb{E}(\#S \text{ in } G) \approx n^4 \cdot p^6$



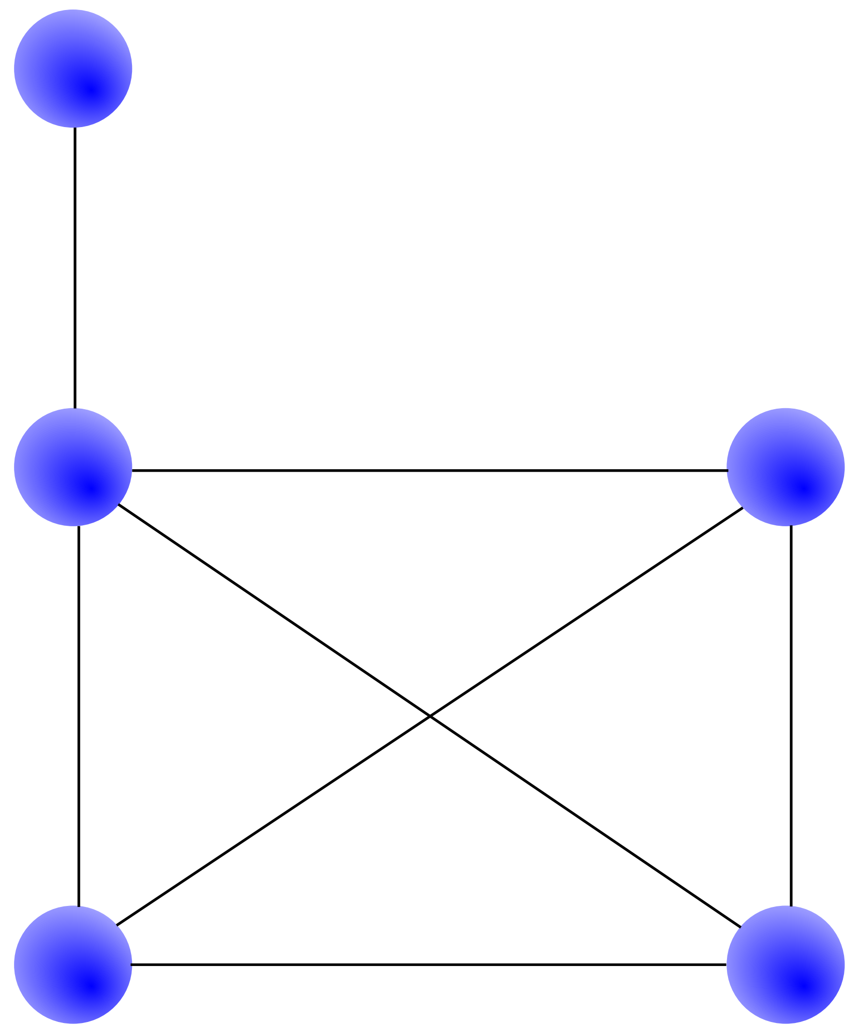
H



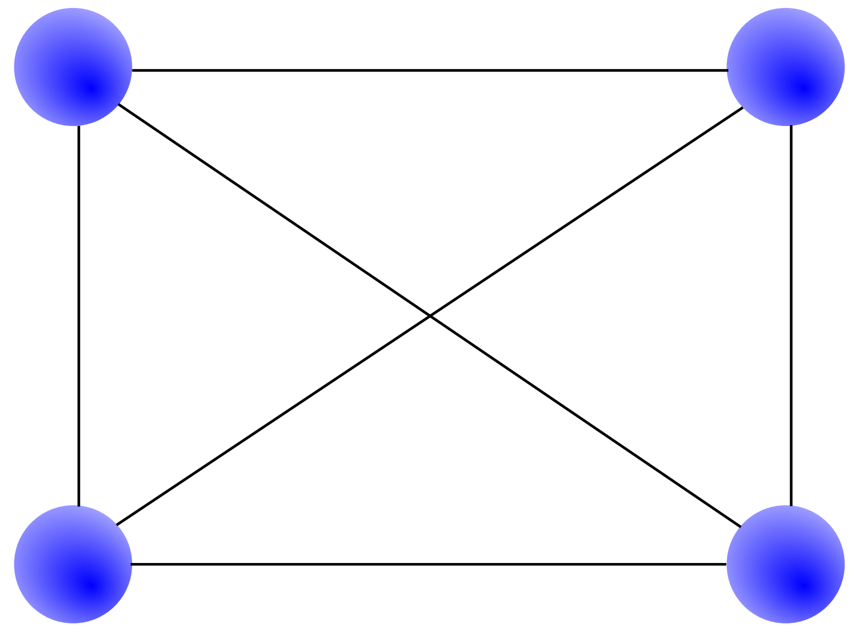
S

Is the expectation method tight?

- Maybe not!
- Analogy from **random $G(n, p)$ graphs**.
- What is the probability that G **contains an H subgraph**?
  - $\mathbb{E}(\#H \text{ in } G) \approx n^5 \cdot p^7$
  - $\mathbb{E}(\#S \text{ in } G) \approx n^4 \cdot p^6$
- Let $p = n^\alpha$, with $-5/7 < \alpha < -2/3$.
  - Then $\mathbb{E}(\#H \text{ in } G) \to \infty$ but $\mathbb{E}(\#S \text{ in } G) \to 0$.



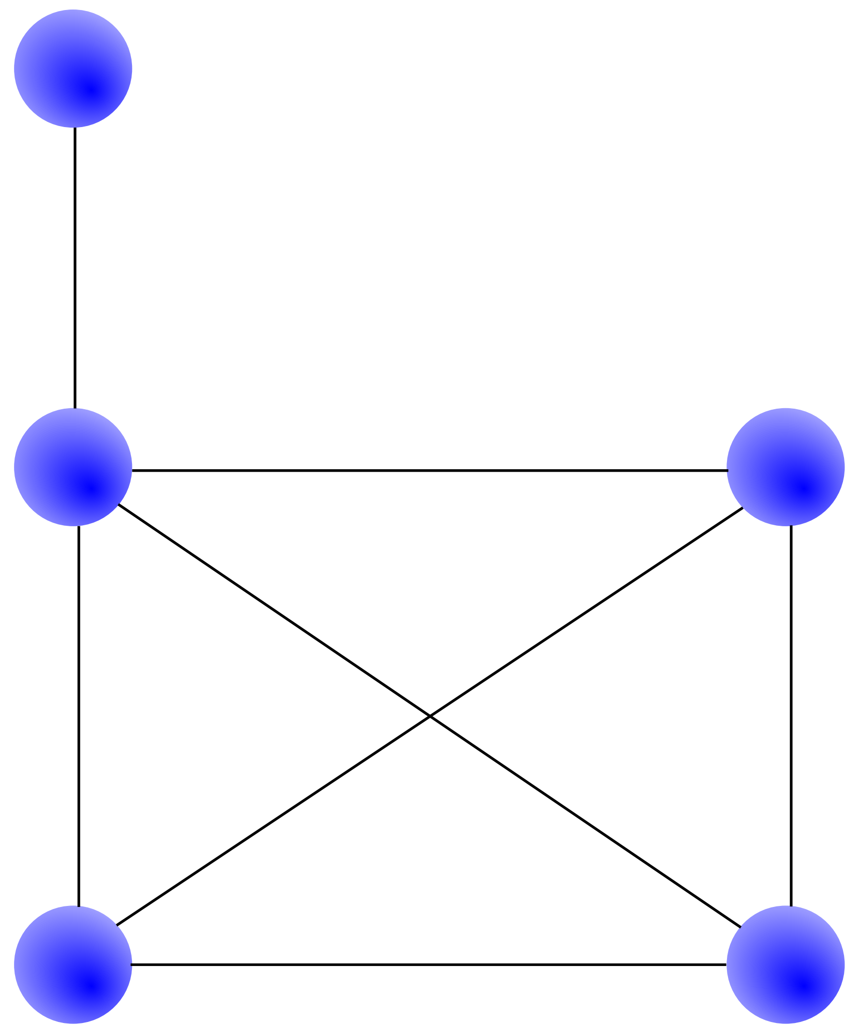
H



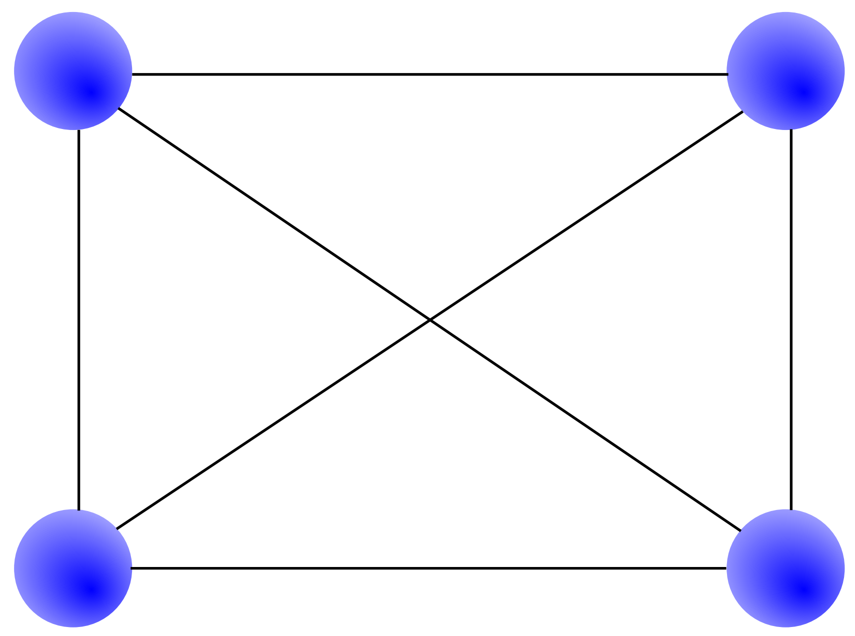
S

Is the expectation method tight?

- Maybe not!
- Analogy from **random $G(n, p)$ graphs**.
- What is the probability that G **contains an H subgraph**?
  - $\mathbb{E}(\#H \text{ in } G) \approx n^5 \cdot p^7$
  - $\mathbb{E}(\#S \text{ in } G) \approx n^4 \cdot p^6$
- Let $p = n^\alpha$, with $-5/7 < \alpha < -2/3$.
  - Then $\mathbb{E}(\#H \text{ in } G) \to \infty$ but $\mathbb{E}(\#S \text{ in } G) \to 0$.
  - So almost surely **not a single H can be found in G** even though many such subgraphs appear **in expectation**.



H

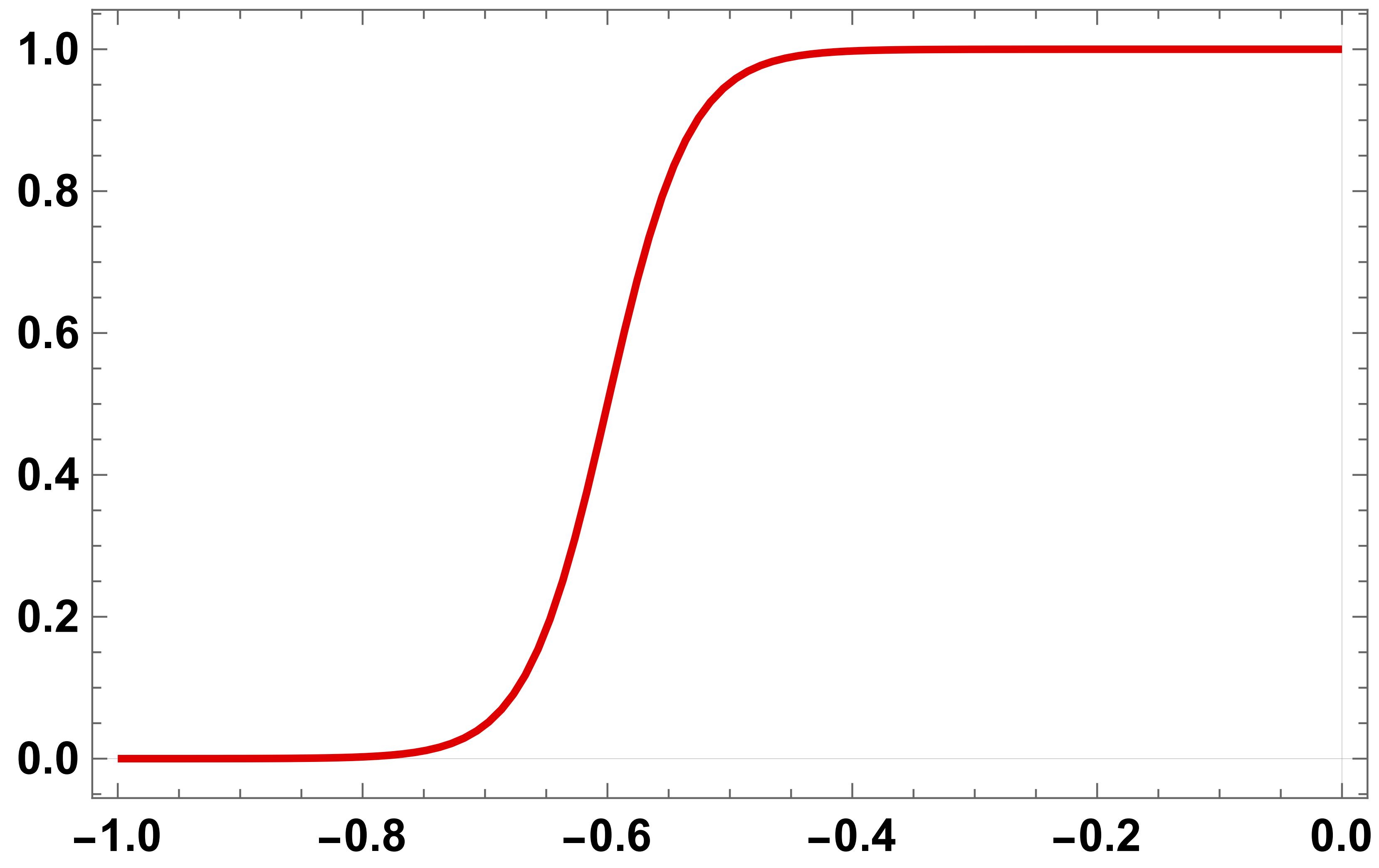


S

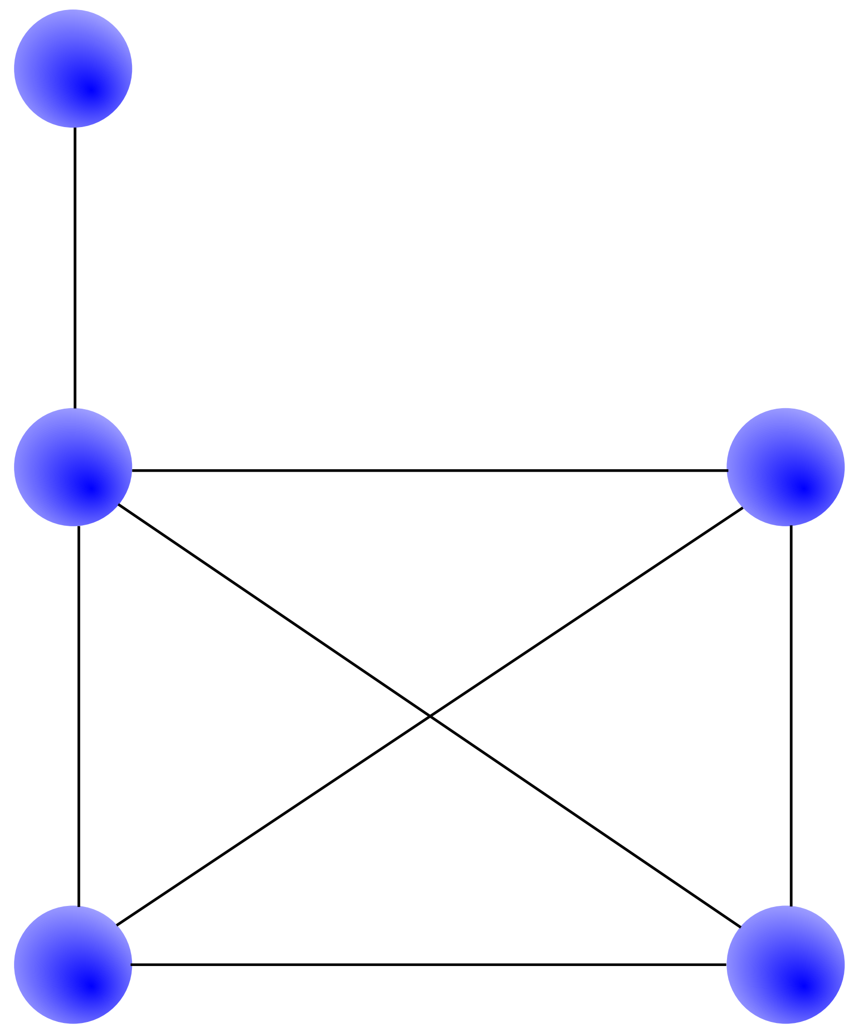
Threshold for random graphs

- **Theorem (Bollobás 1981):** A subgraph H is likely found in G if and only if $\mathbb{E}(\#S \text{ in } G) \to \infty$ for all $S \subseteq H$.

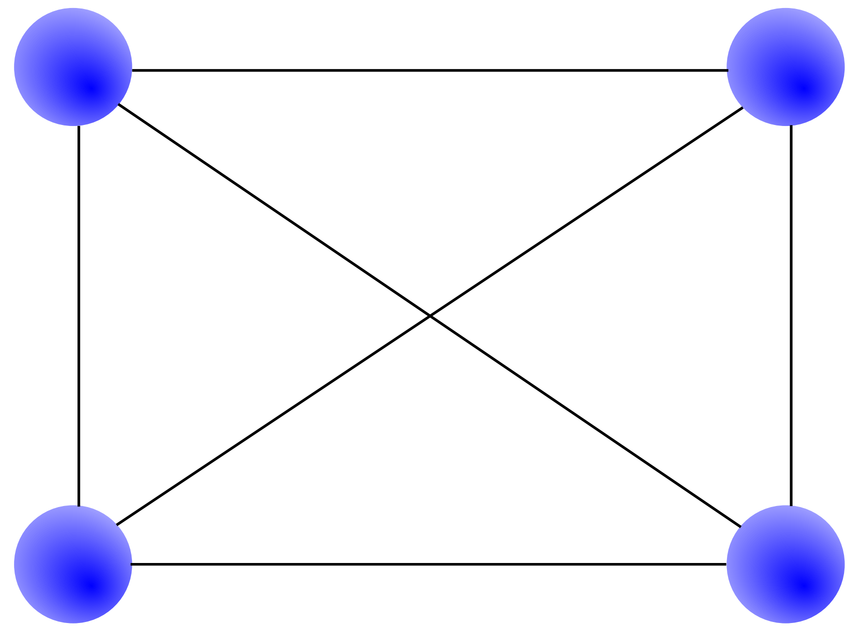
$\Pr(G(n, p^{\alpha})$ **contains** $H)$



α



H

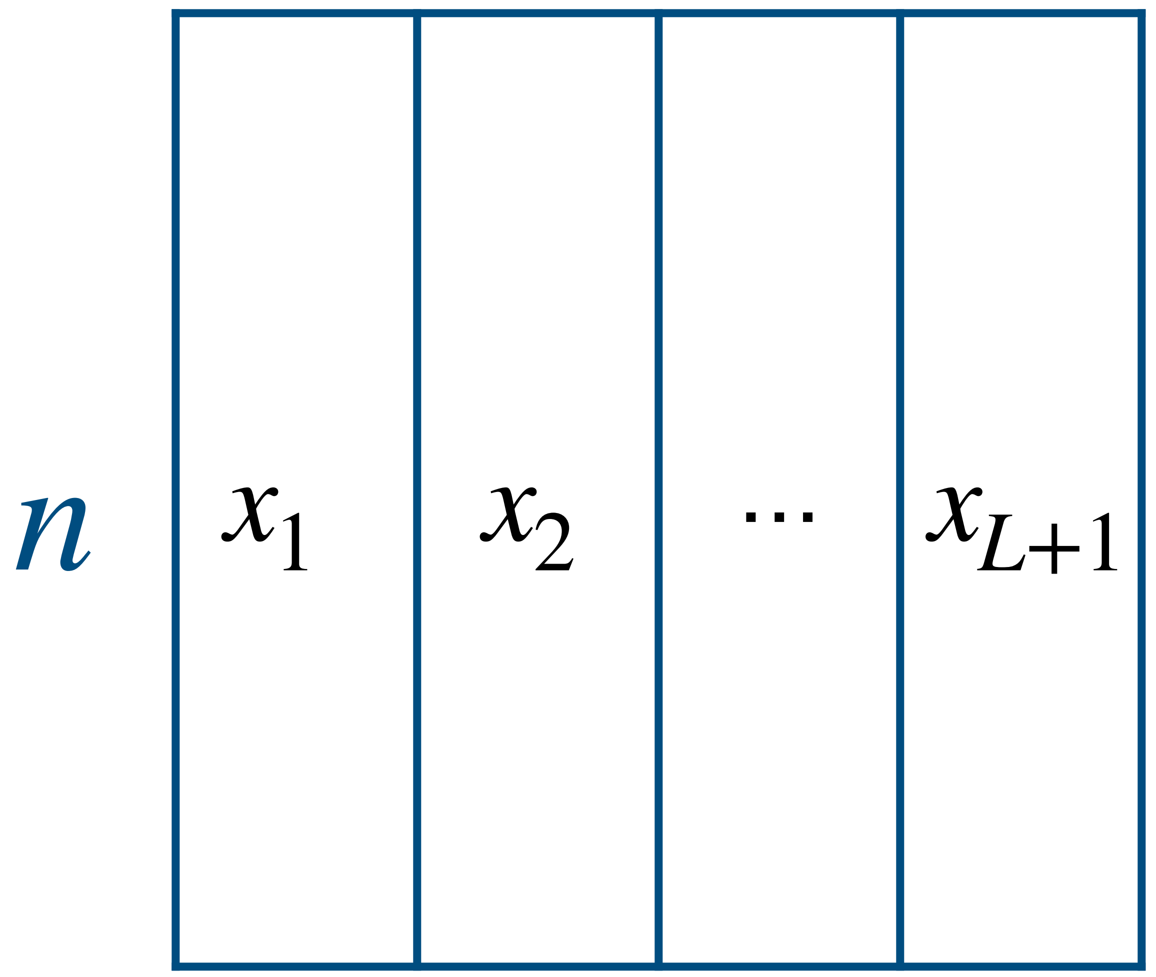


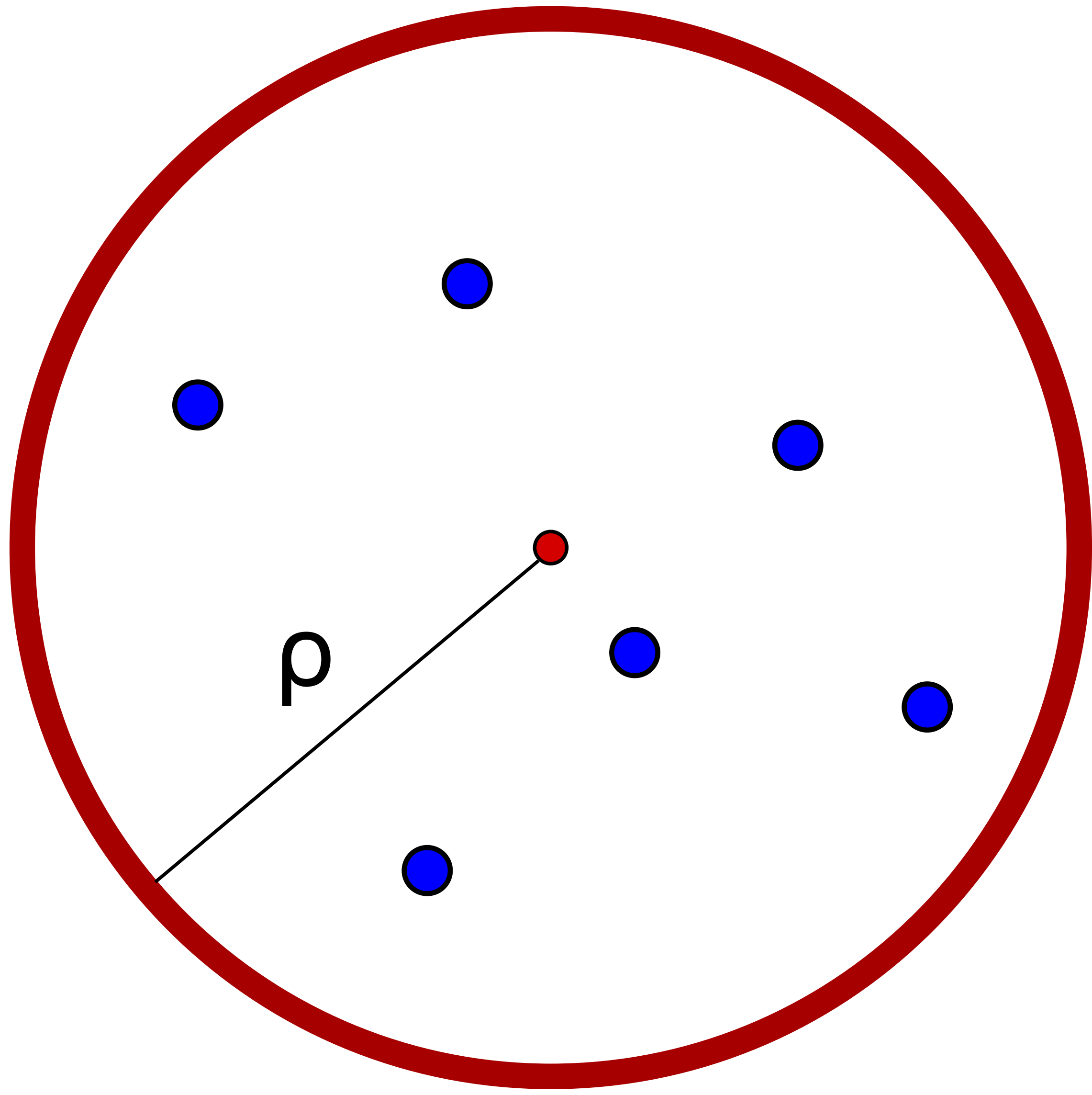
S

Back to list-decodability of an RLC

Back to list-decodability of an RLC

- **Notation:** write a ρ**-clustered set** $\{x_1, \ldots, x_{L+1}\} \subseteq \mathbb{F}_2^n$ as a matrix A.

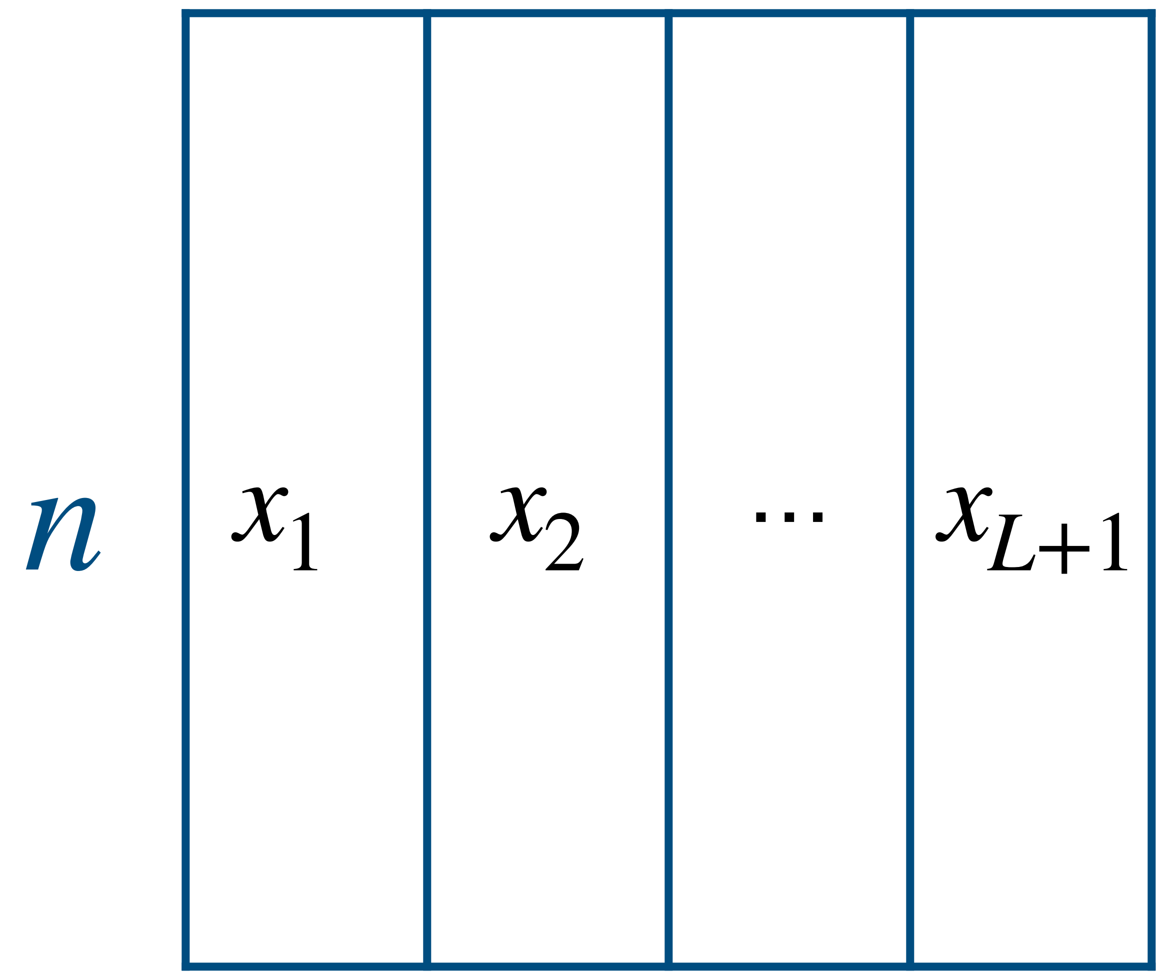


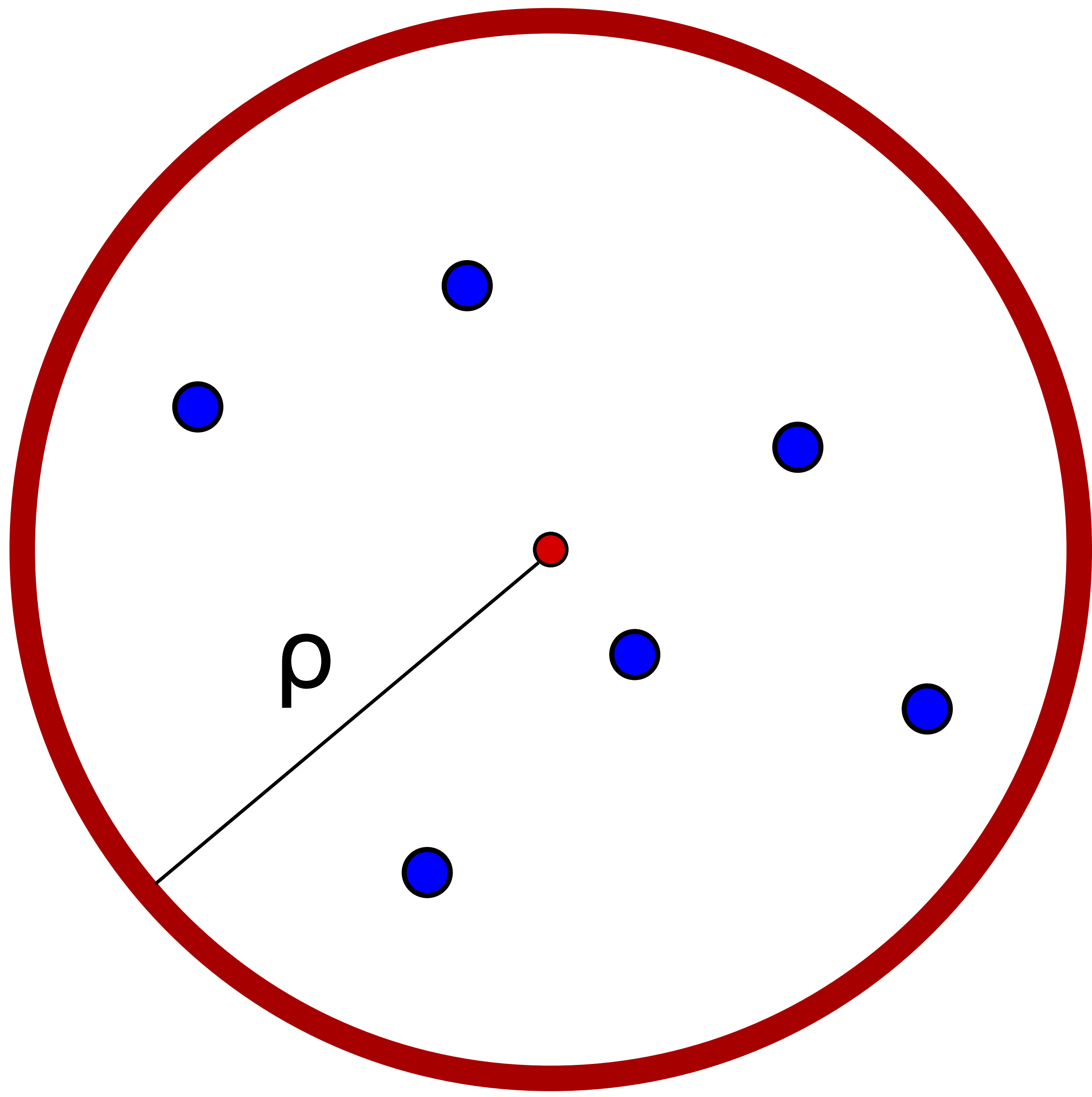


Back to list-decodability of an RLC

- **Notation:** write a ρ**-clustered set** $\{x_1, \ldots, x_{L+1}\} \subseteq \mathbb{F}_2^n$ as a matrix A.

- **Observation:** the family of ρ**-clustered matrices** is **closed to row permutations**.



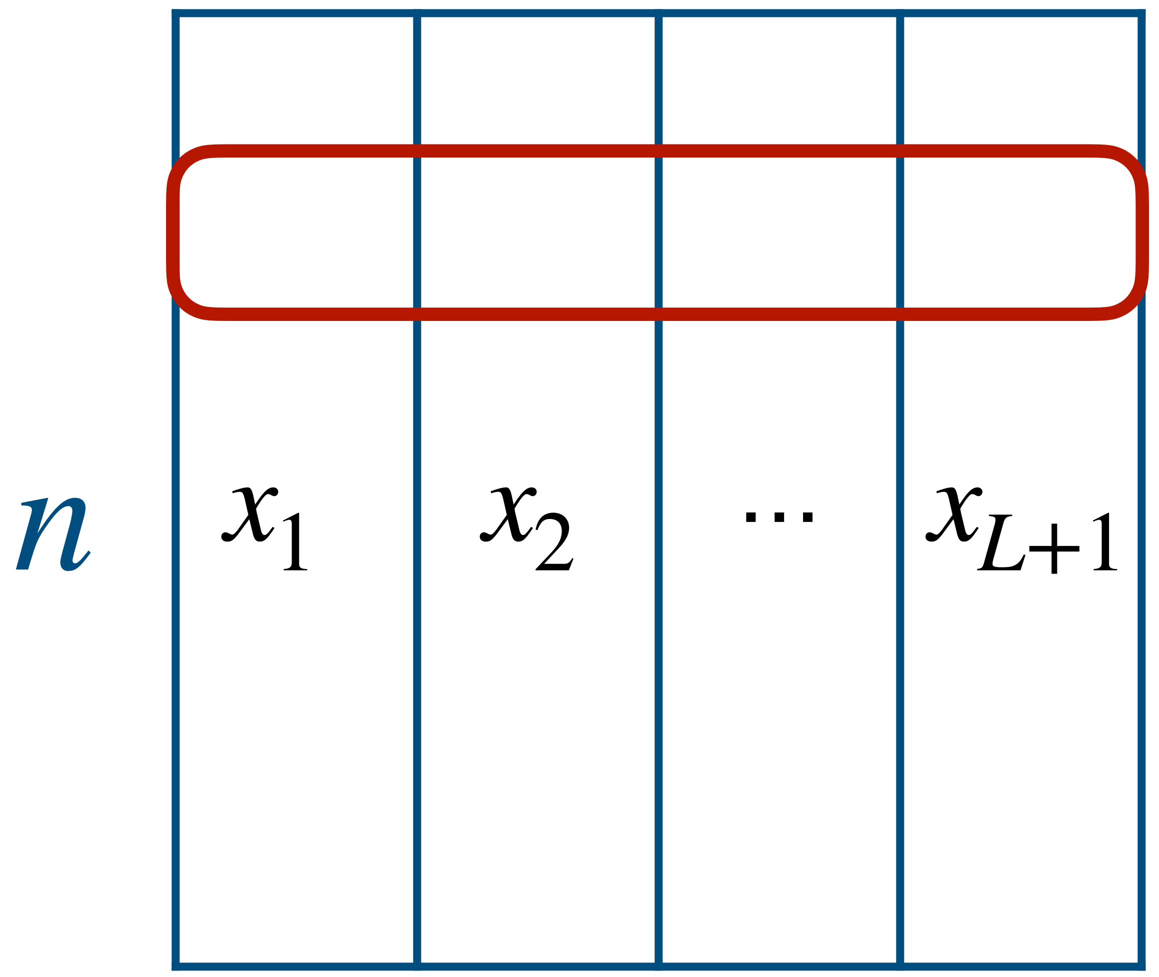


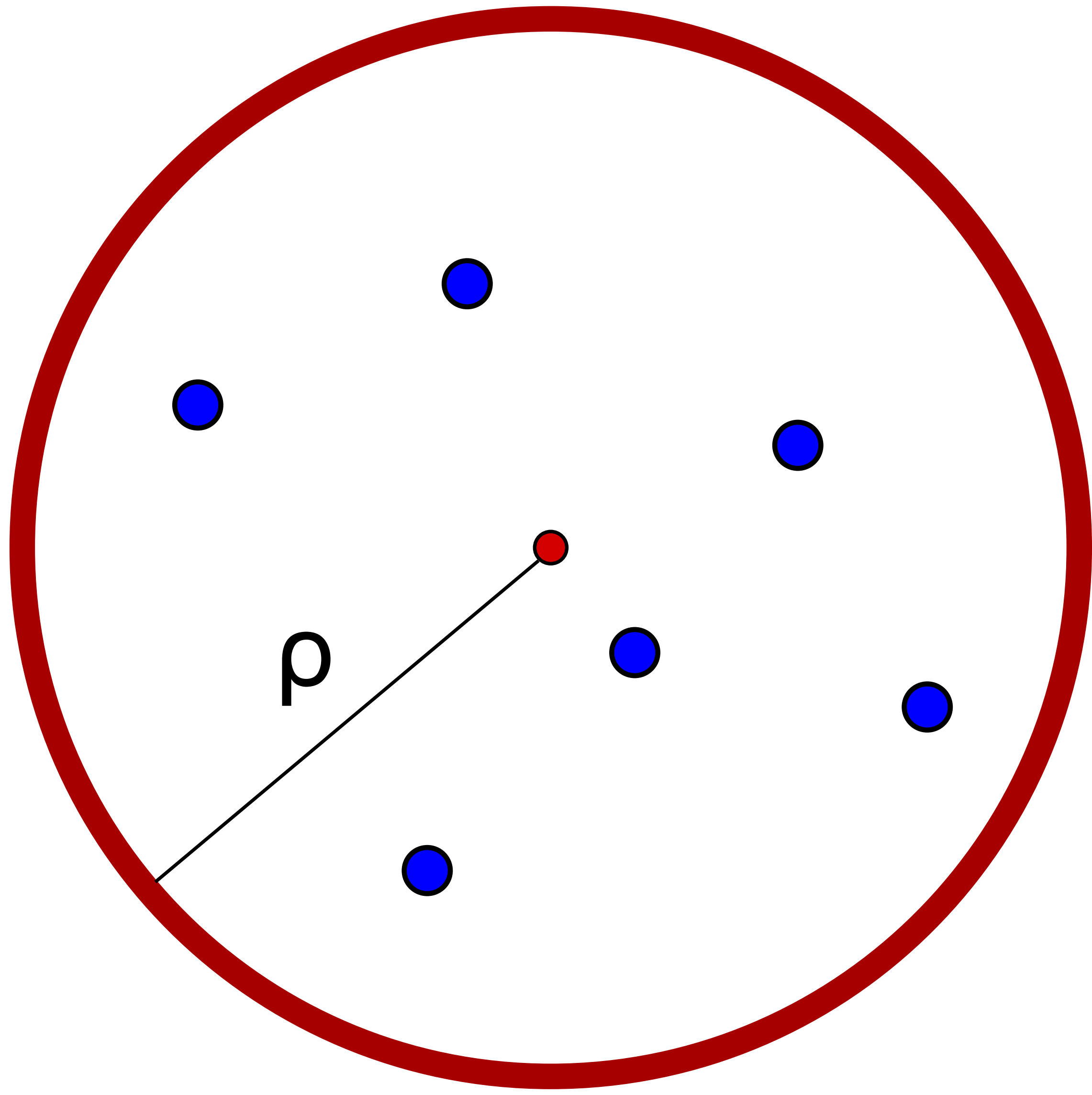
Back to list-decodability of an RLC

- **Notation:** write a ρ**-clustered set** $\{x_1, \ldots, x_{L+1}\} \subseteq \mathbb{F}_2^n$ as a matrix A.

- **Observation:** the family of ρ**-clustered matrices** is **closed to row permutations**.

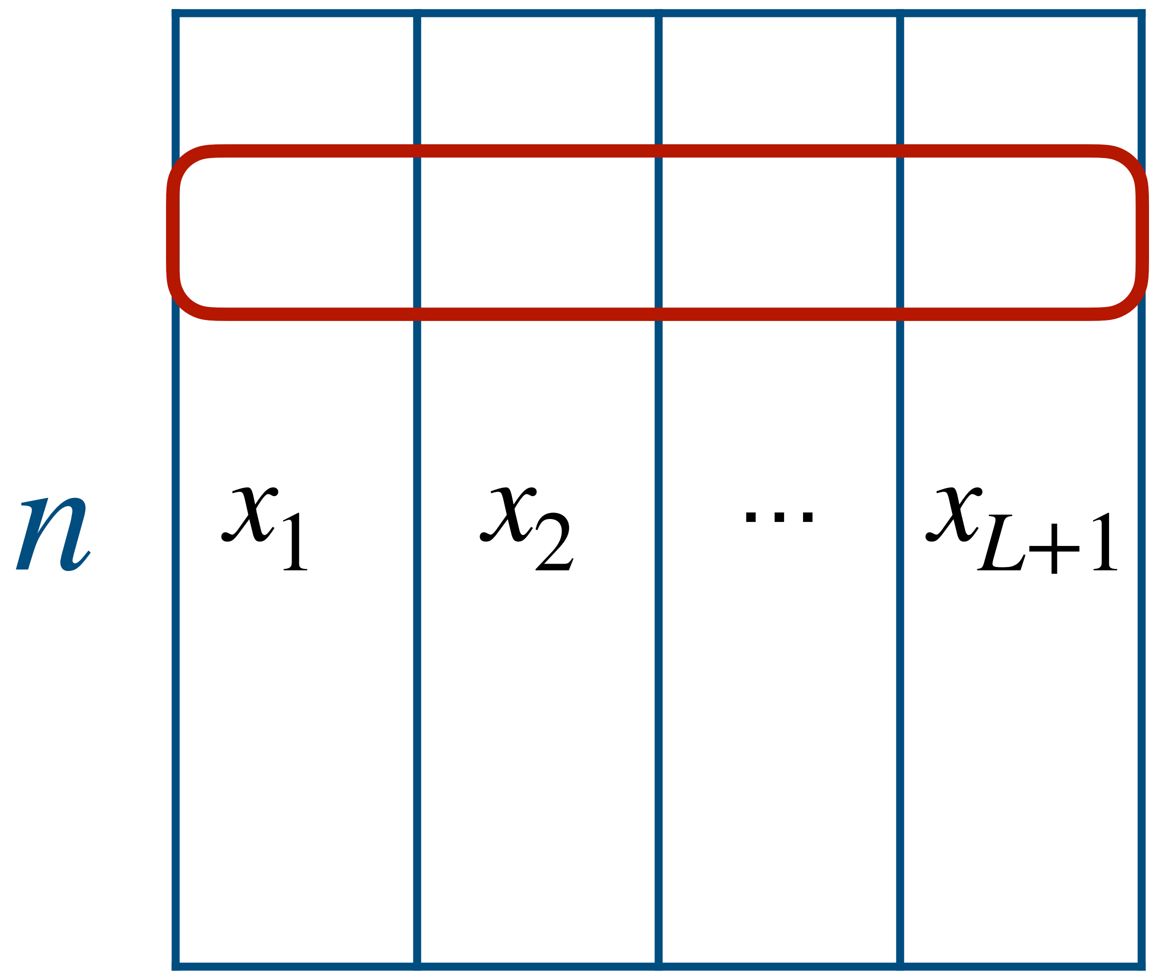
- To determine if A is ρ**-clustered** we only need to know its **row distribution**. That is, how many times each vector in $\mathbb{F}_2^n$ appears in A.





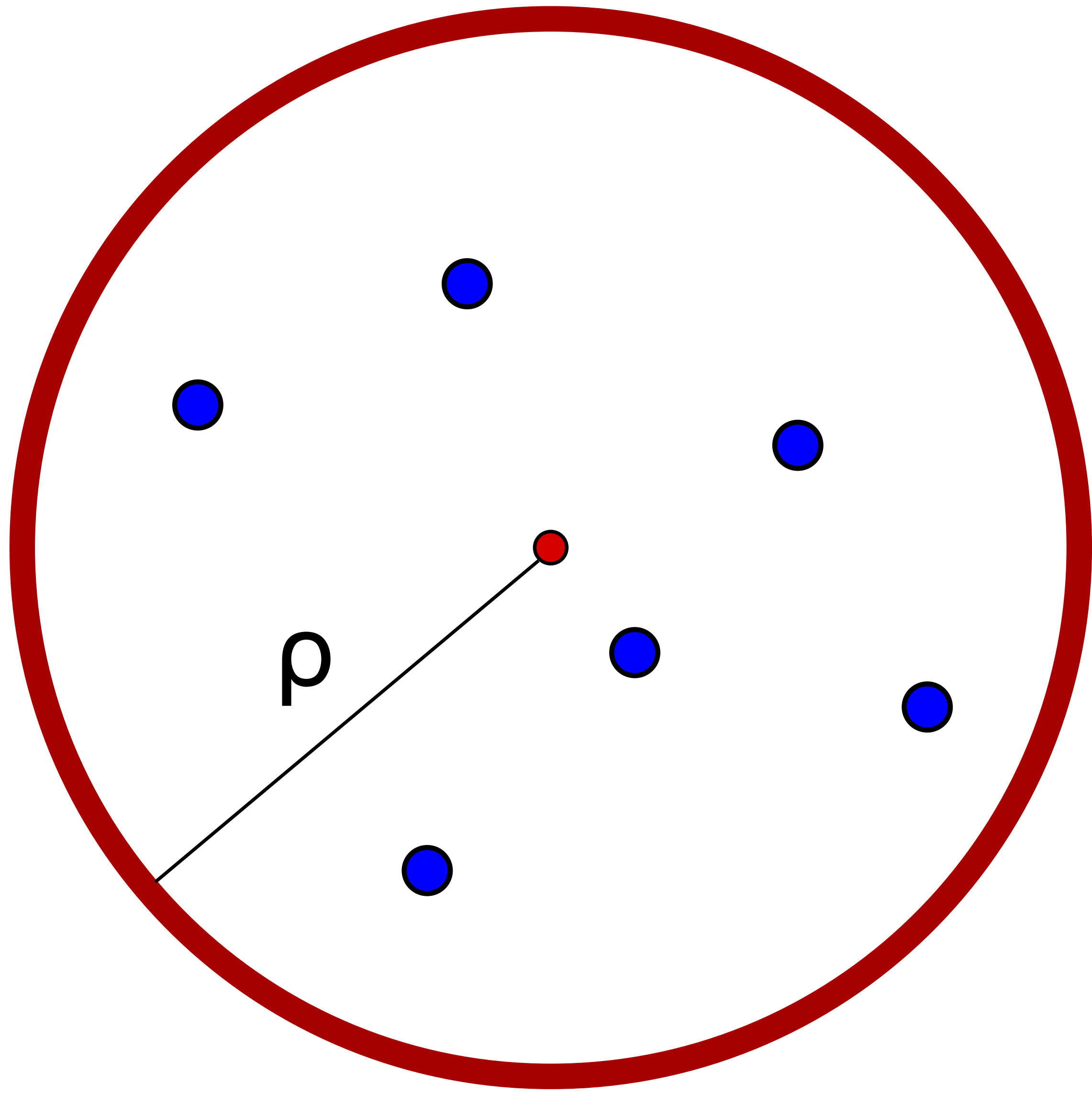
Back to list-decodability of an RLC

- **Notation:** write a **ρ-clustered set** $\{x_1, \ldots, x_{L+1}\} \subseteq \mathbb{F}_2^n$ as a matrix A.



- **Observation:** the family of **ρ-clustered matrices** is **closed to row permutations**.

- To determine if A is **ρ-clustered** we only need to know its **row distribution**. That is, how many times each vector in $\mathbb{F}_2^n$ appears in A.

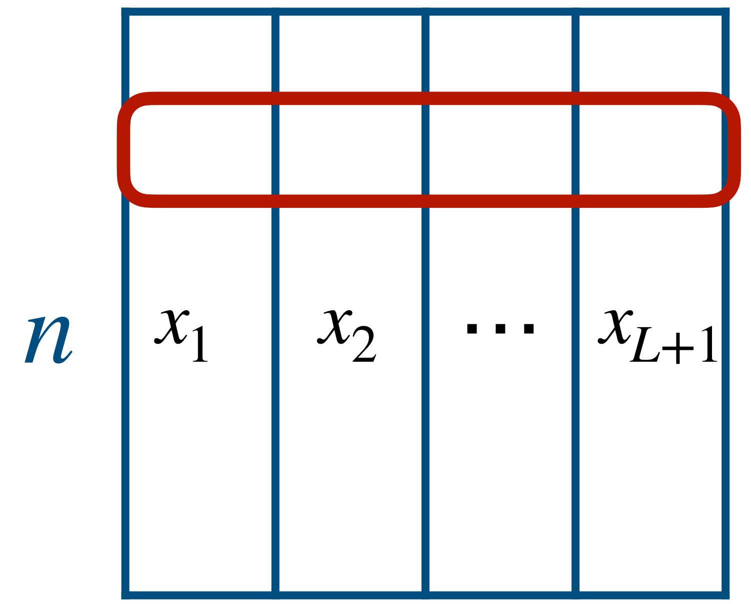


- There are at most $n^{2^{L+1}}$ **ρ-clustered distributions**. This is a **tiny** number so we can **treat each clustered distribution separately**.

Expectations in an RLC

- Let τ be a distribution over $\mathbb{F}_2^{L+1}$.
- How many τ-distributed matrices do we expect in an **RLC**?

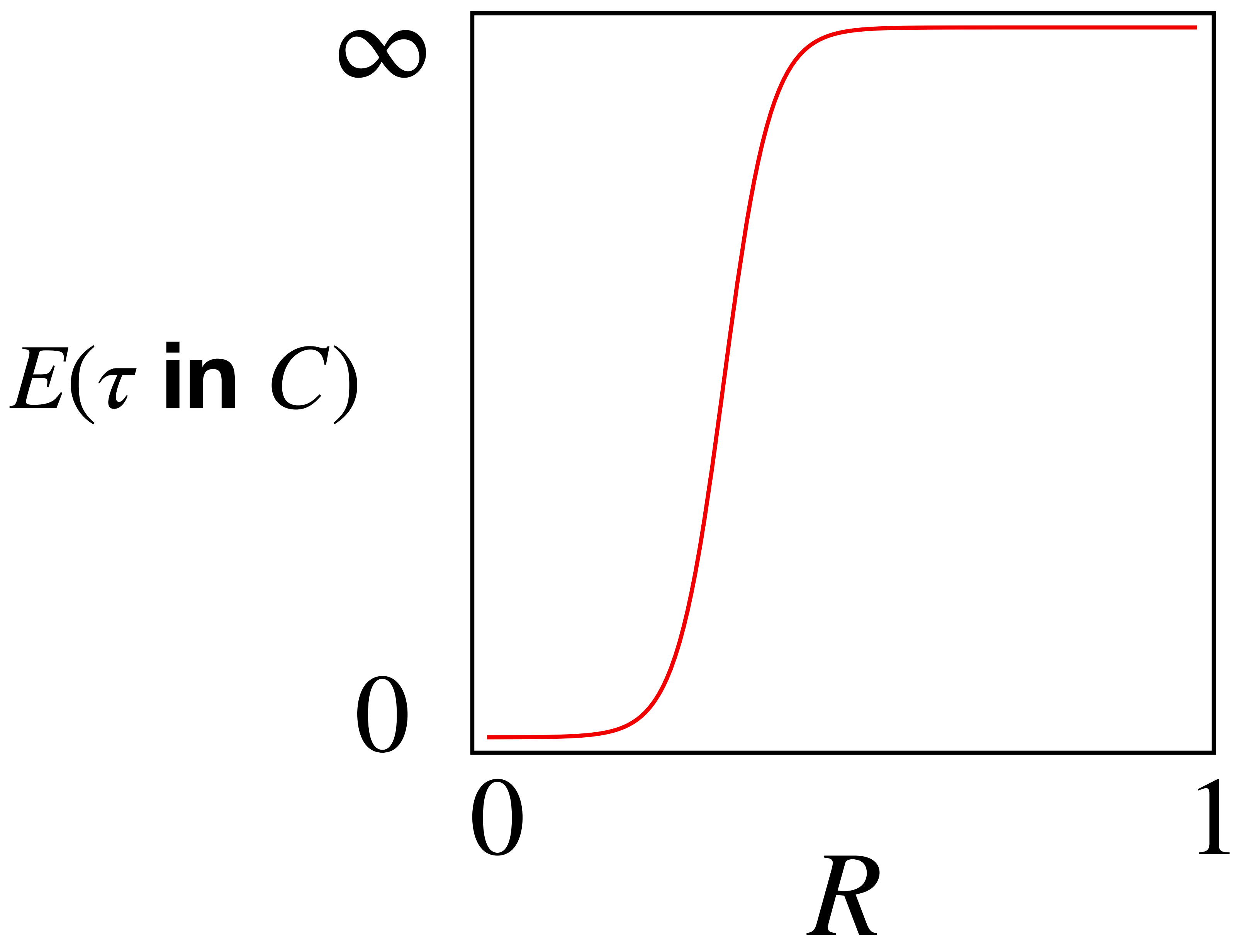
$$\mathbb{E}(\tau\text{-distributed matrices in } C) = \#\tau\text{-distributed matrices} \cdot \Pr_{A\sim\tau}(A \subseteq C)$$
$$\approx 2^{nH(\tau)} \cdot 2^{-n(1-R)\cdot \dim\{x_1,\ldots,x_{L+1}\}}$$
$$= 2^{n\left(H(\tau) - (1-R)\cdot \dim(\mathrm{supp}(\tau))\right)}$$

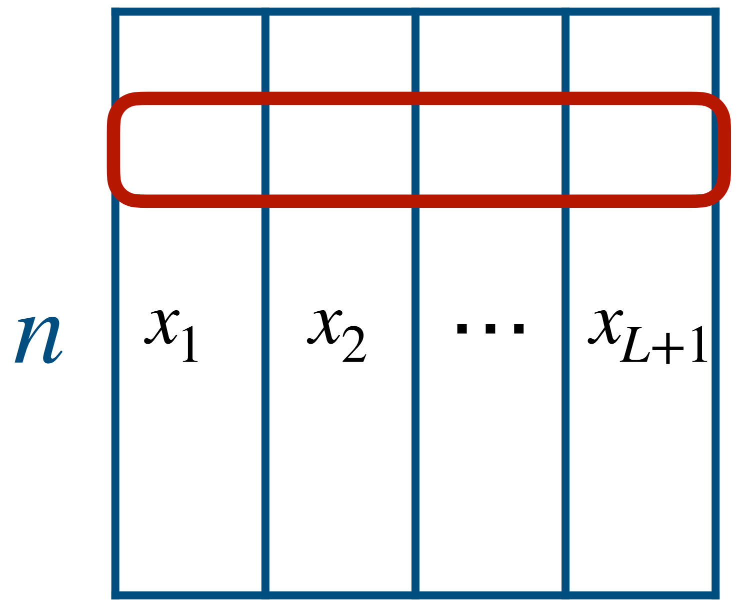


Expectations in an RLC

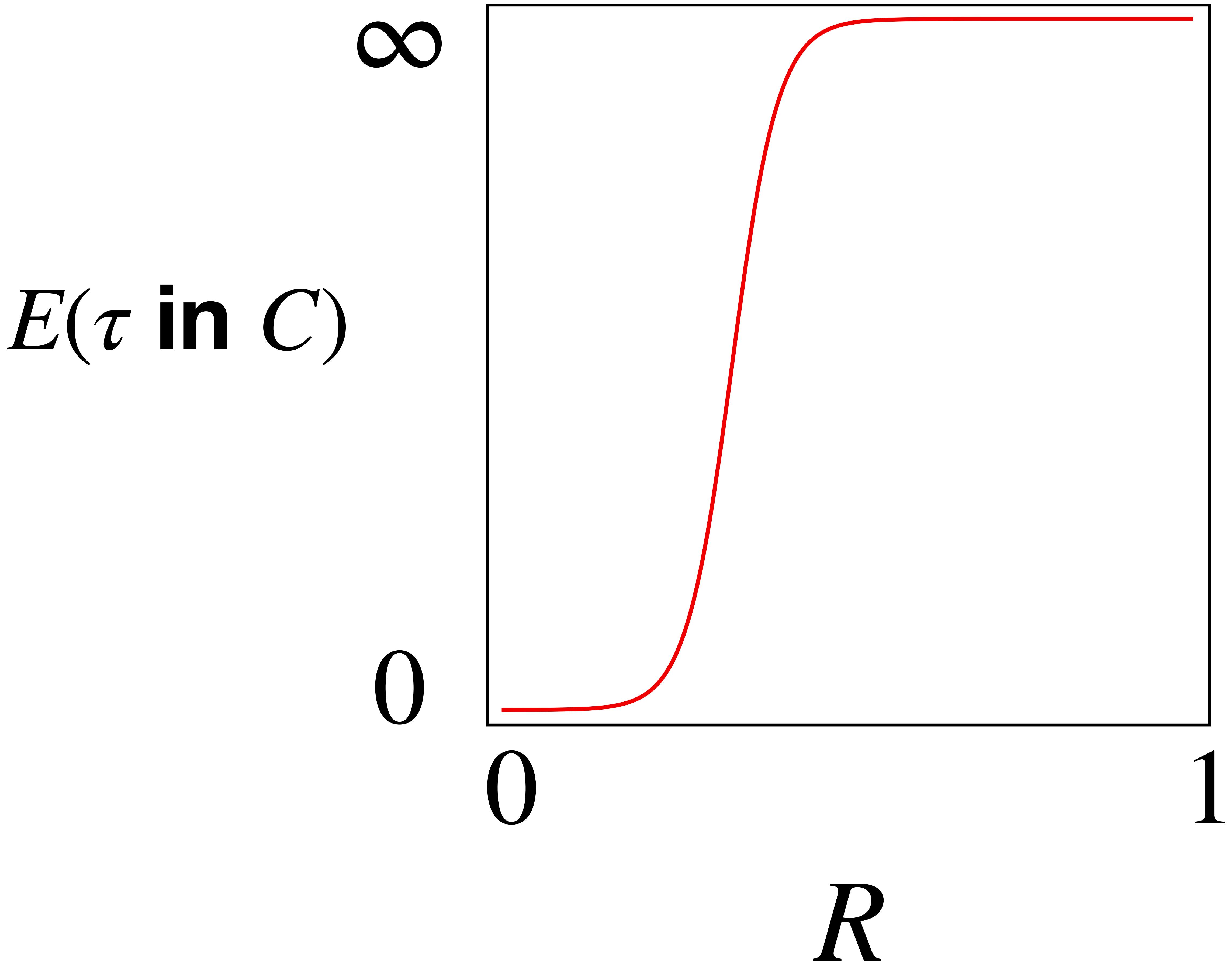
- Let τ be a distribution over $\mathbb{F}_2^{L+1}$.
- How many τ-distributed matrices do we expect in an **RLC**?

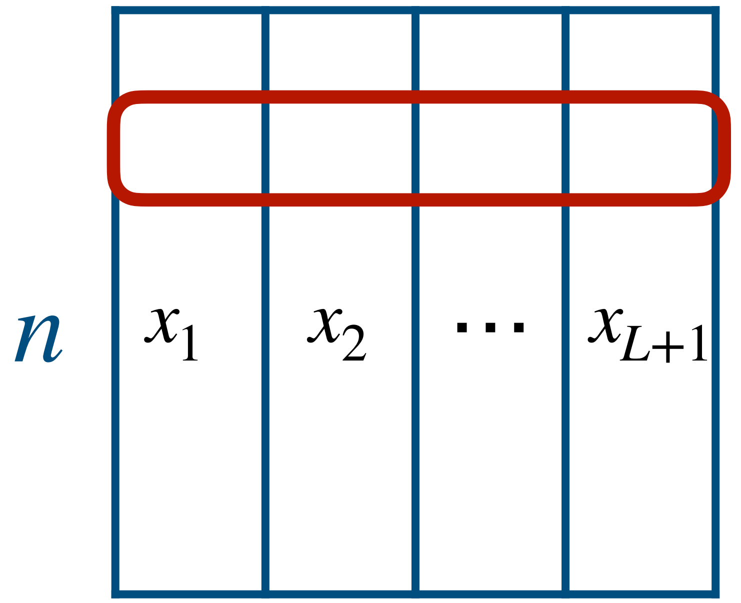
$$\mathbb{E}(\tau\text{-distributed matrices in } C) = \#\tau\text{-distributed matrices} \cdot \Pr_{A\sim\tau}(A \subseteq C)$$
$$\approx 2^{nH(\tau)} \cdot 2^{-n(1-R)\cdot\dim\{x_1,\ldots,x_{L+1}\}}$$
$$= 2^{n\left(H(\tau) - (1-R)\cdot\dim(\mathrm{supp}(\tau))\right)}$$



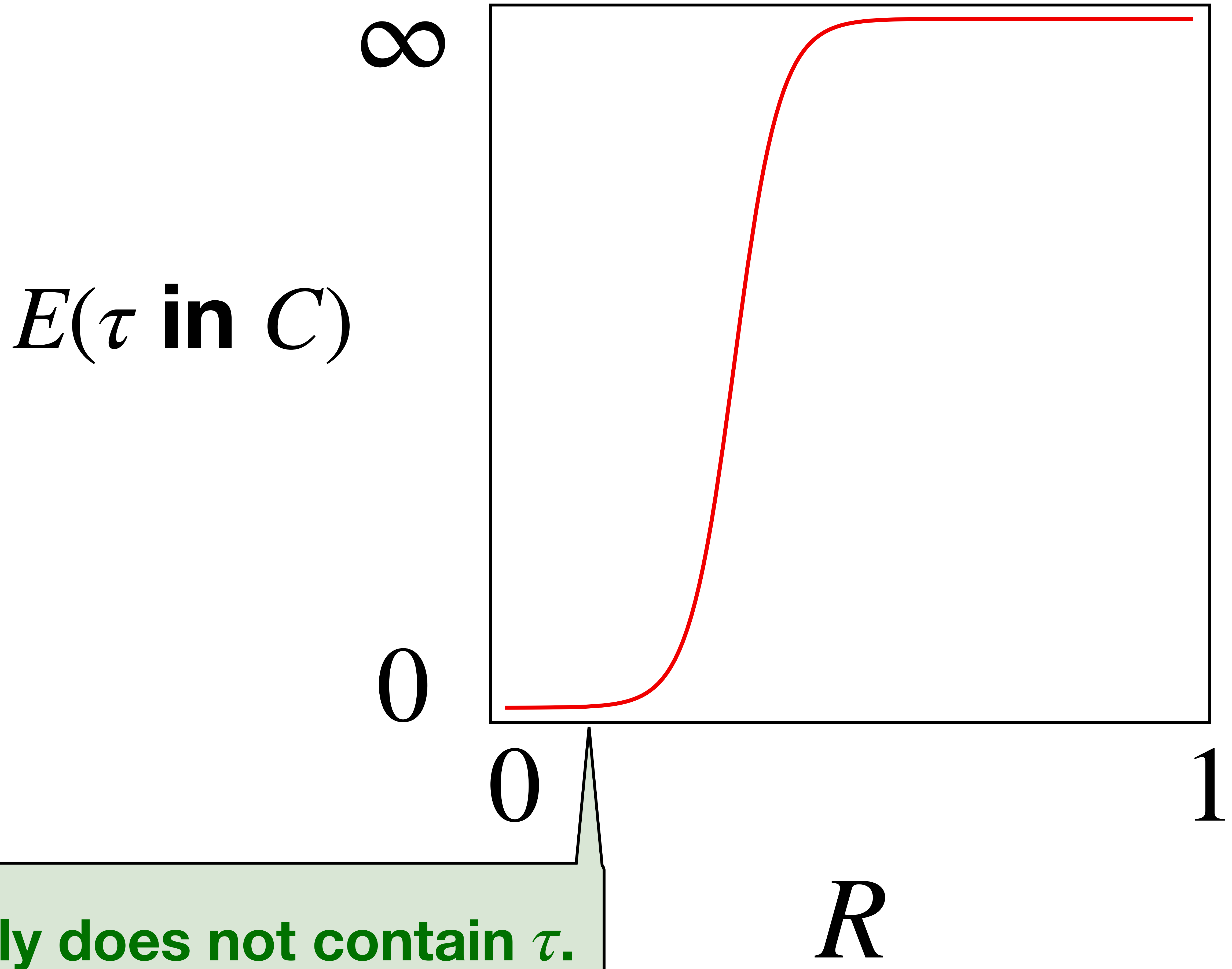


Expectations in an RLC

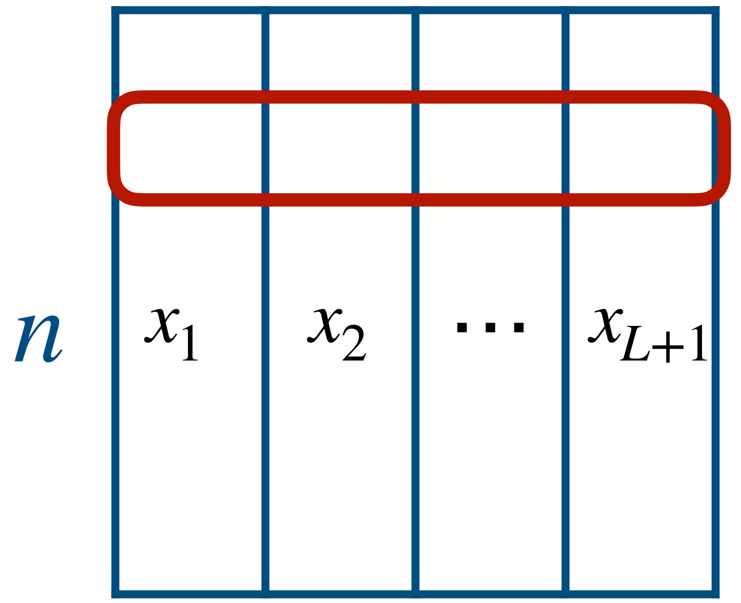




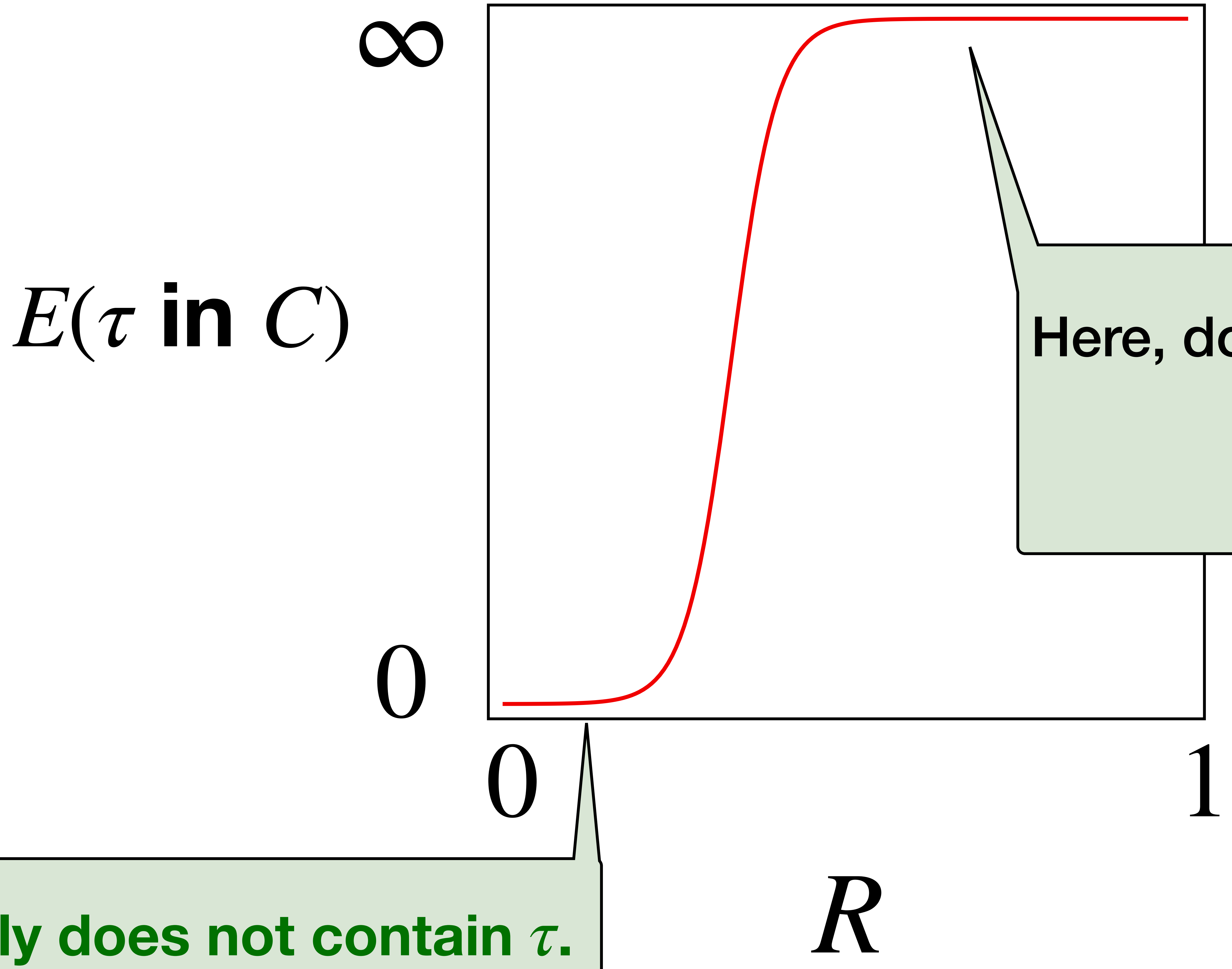
Expectations in an RLC



Here, C almost surely does not contain τ.



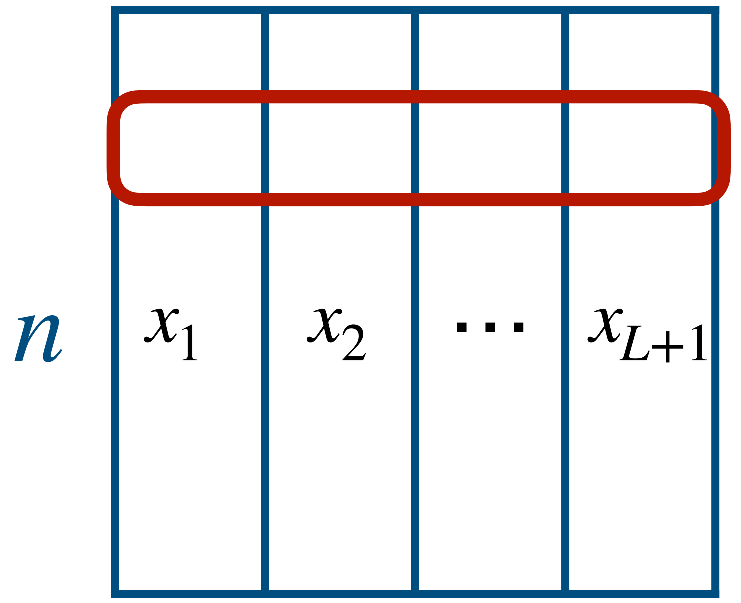
Expectations in an RLC

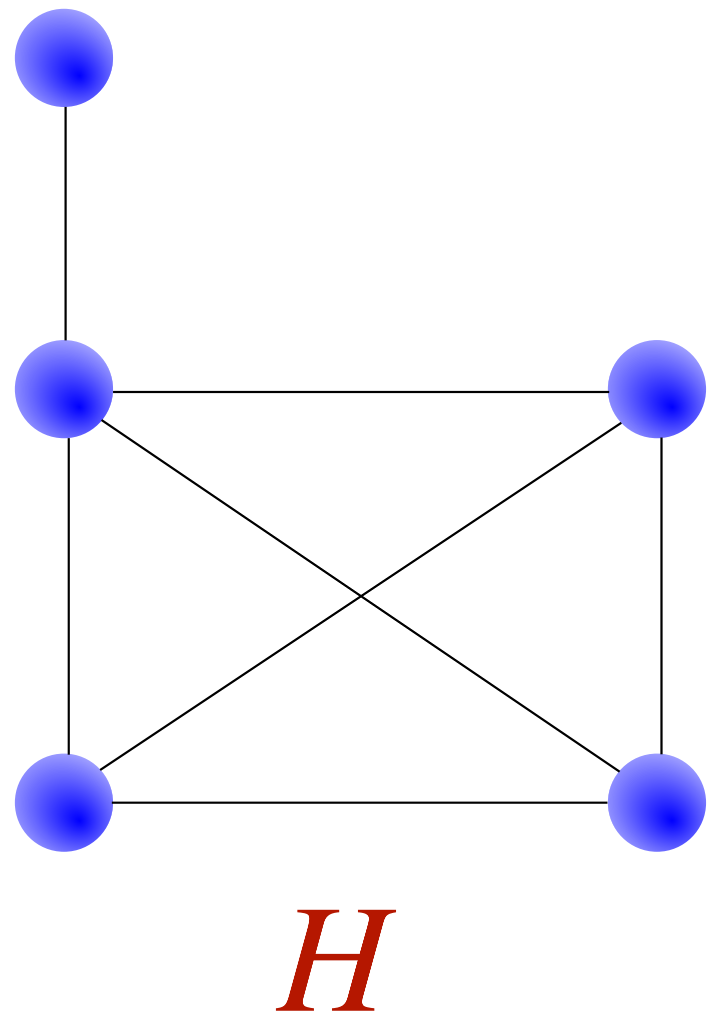


Here, does C **almost surely contain** τ?

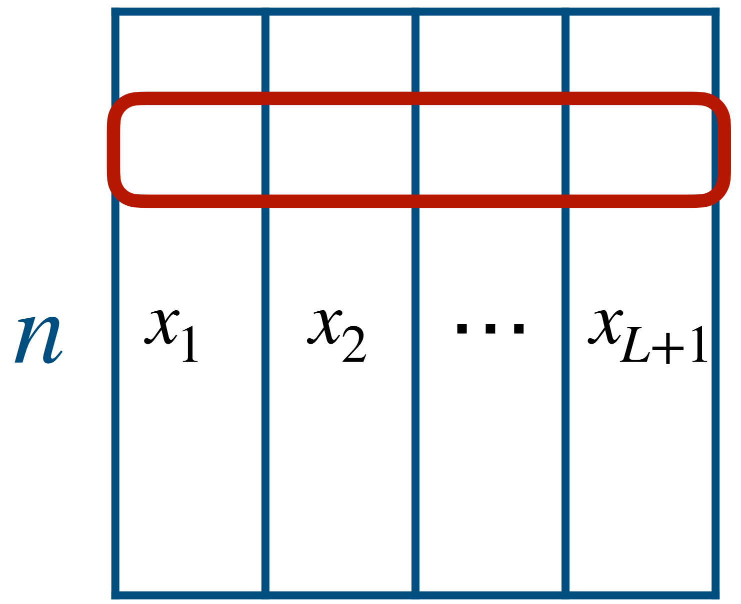
Not necessarily!

Here, C **almost surely does not contain** τ.



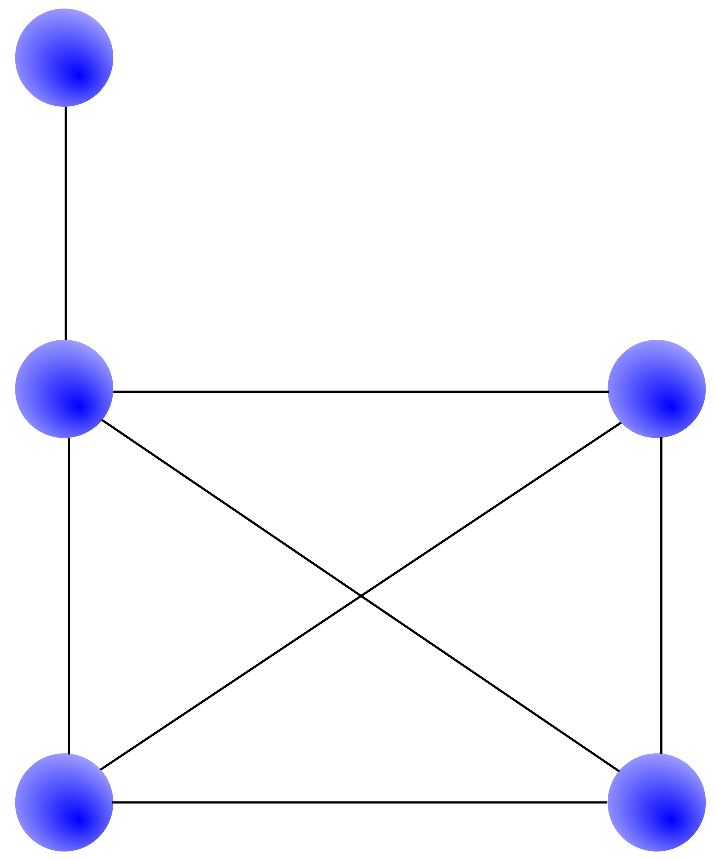


- The **distribution** τ is analogous to a **subgraph** H.

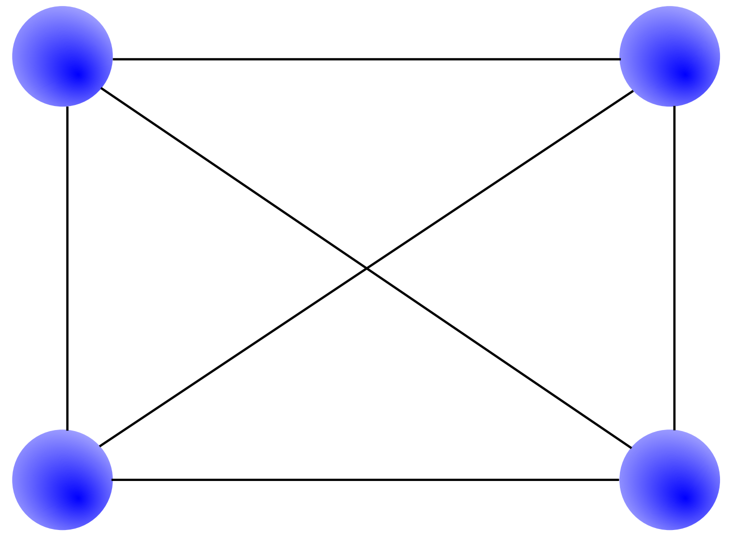


- The **distribution** τ is analogous to a **subgraph** H.

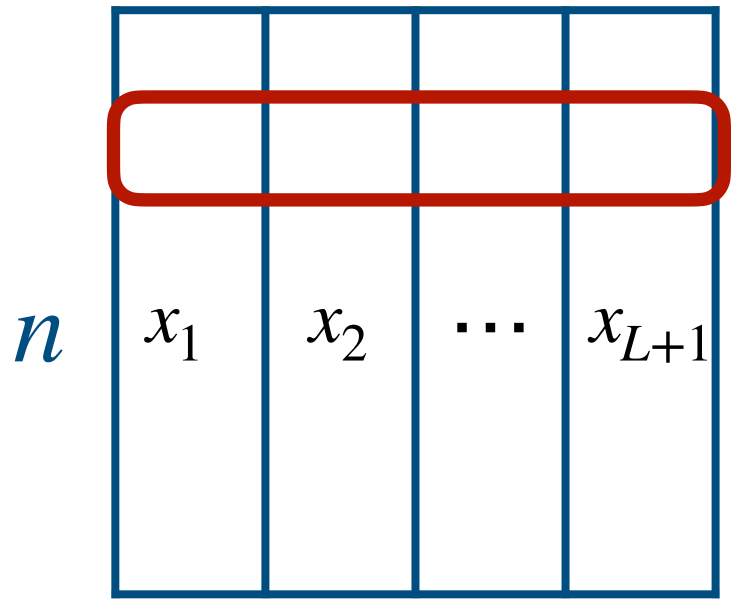
- What about **subgraphs** of H?

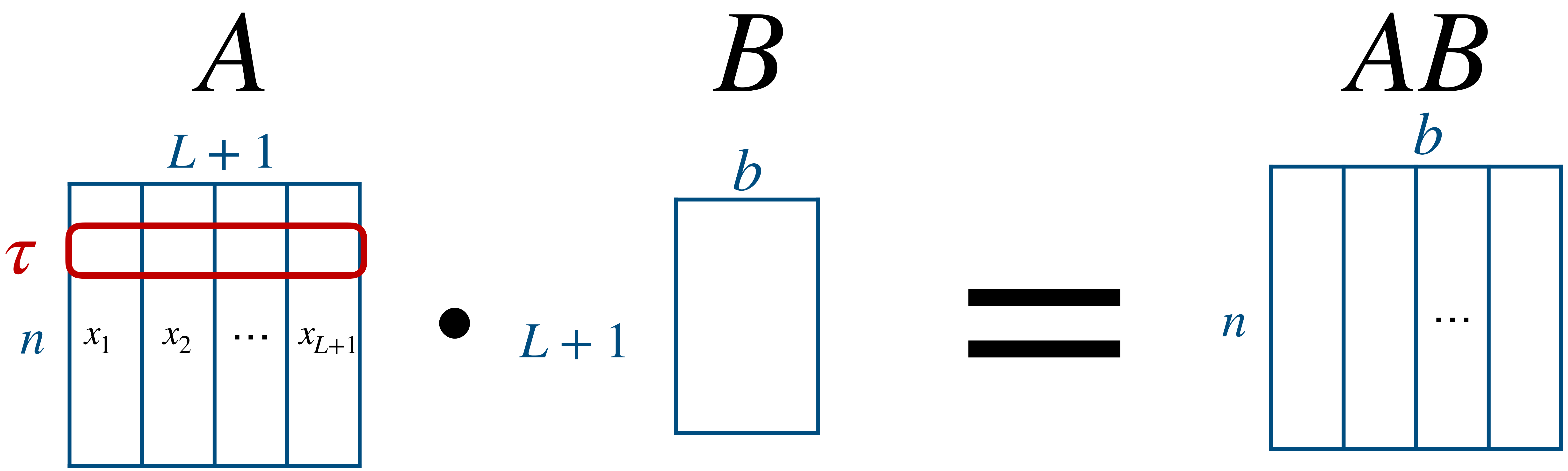


H



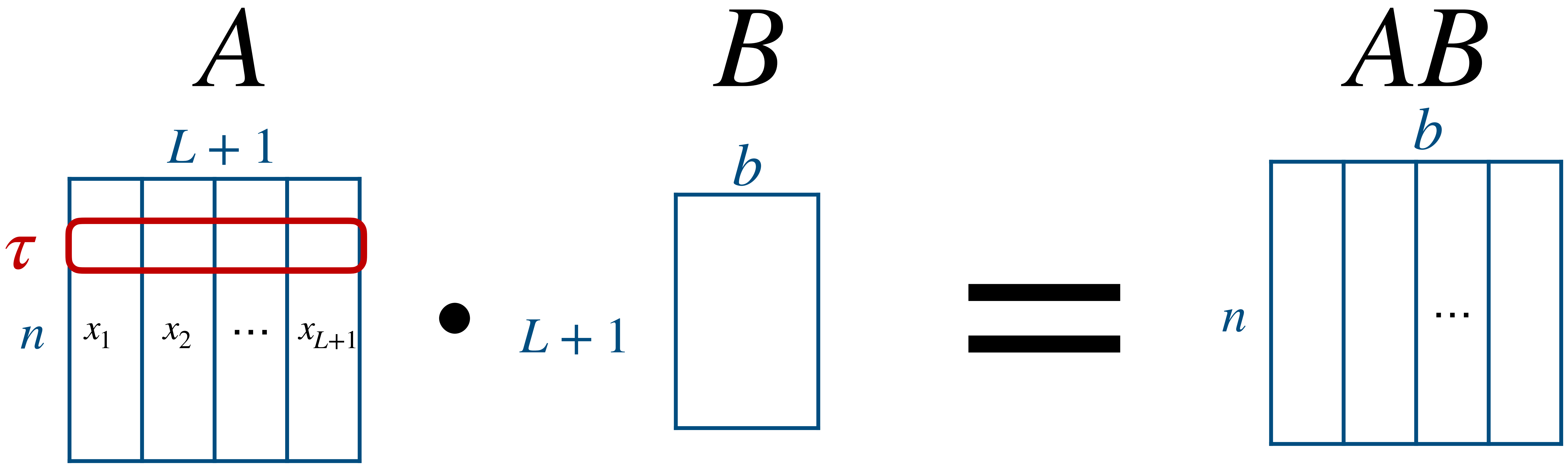
S





A
B
AB
L + 1
b
b
τ
n
x1
x2
⋯
xL+1
L + 1
n
⋯

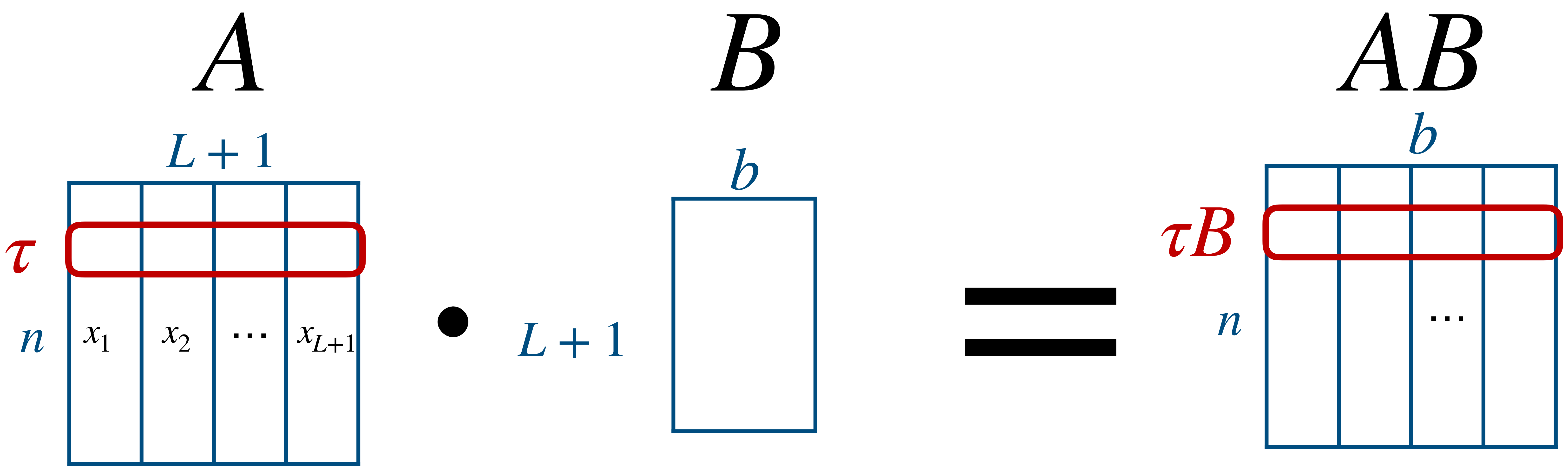
- Suppose $A \subseteq C$. Then **C also contains AB** whenever $B \in \mathbb{F}_2^{(L+1)\times b}$ $(b \leq L+1)$.



- Suppose $A \subseteq C$. Then C **also contains AB** whenever $B \in \mathbb{F}_2^{(L+1)\times b}$ ($b \leq L+1$).

- A uniformly random row of AB is distributed like zB where $z \sim \tau$.

- We denote this distribution τB

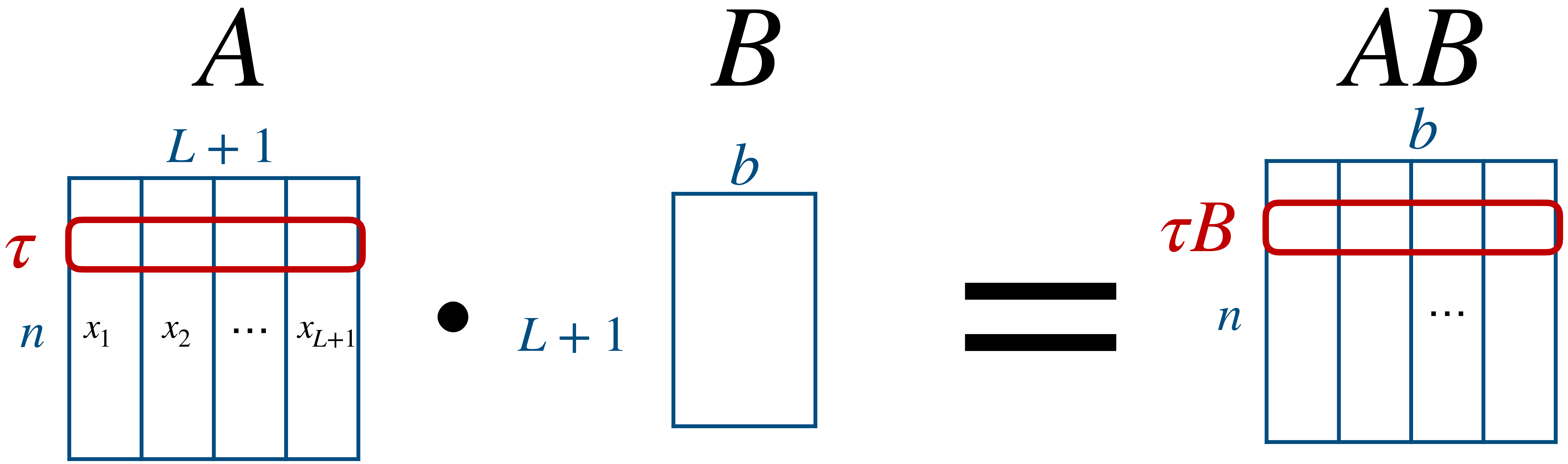


- Suppose $A \subseteq C$. Then **C also contains AB** whenever $B \in \mathbb{F}_2^{(L+1)\times b}$ $(b \leq L+1)$.

- A uniformly random row of AB is distributed like zB where $z \sim \tau$.

- We denote this distribution τB

- In order to **contain τ**, a **linear code** must **contain τB**.



Theorem (thresholds for RLCs):

An **RLC** of rate R is likely to **contain a τ distributed matrix** if and only if

$$\mathbb{E}(\#\tau B \text{ distributed matrices in } C) \to \infty$$

for all $B \in \mathbb{F}_2^{(L+1)\times b}$.

Theorem (thresholds for RLCs):

An **RLC** of rate R is likely to **contain a τ distributed matrix** if and only if

$$\mathbb{E}(\#\tau B \text{ distributed matrices in } C) \to \infty$$

for all $B \in \mathbb{F}_2^{(L+1)\times b}$.

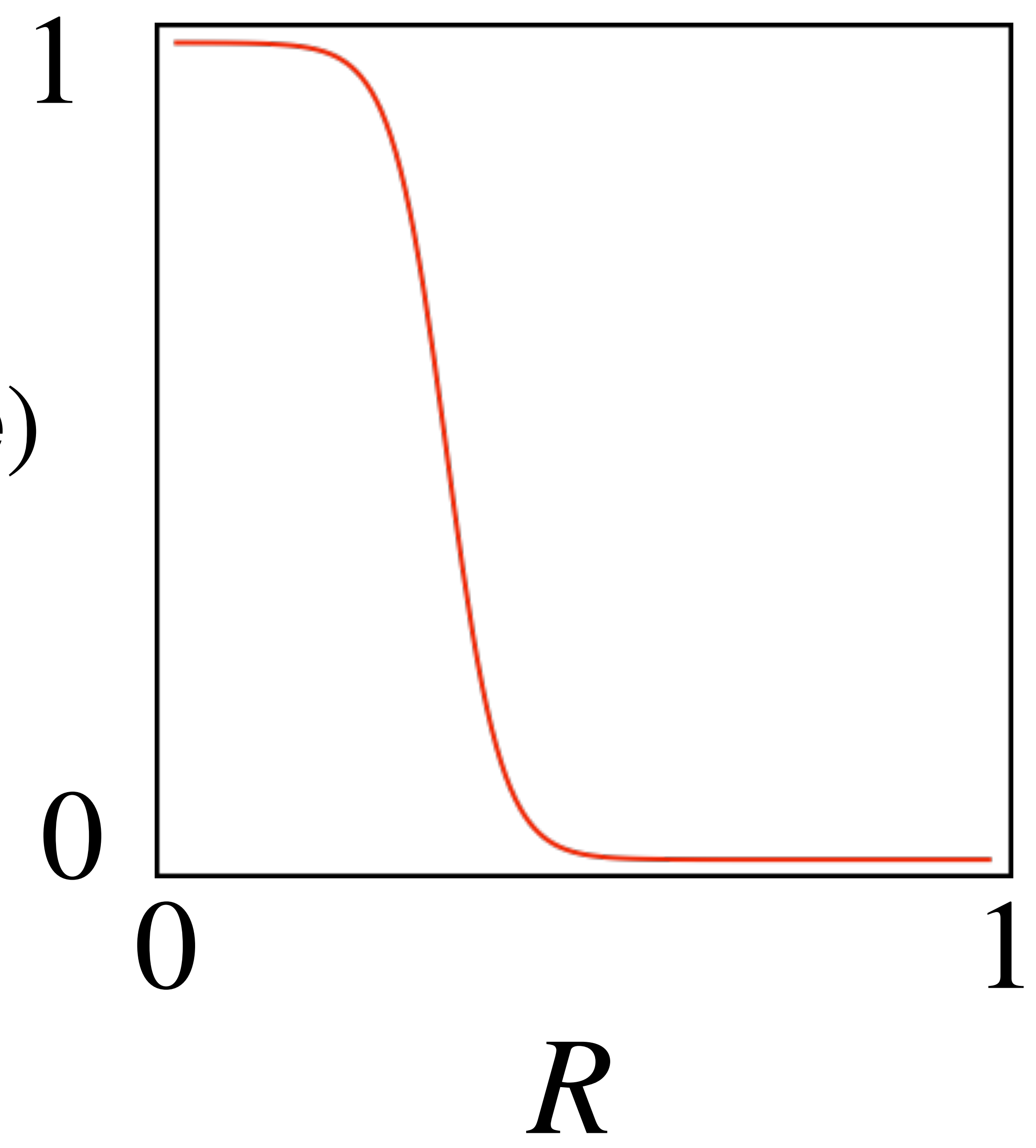
Corollary (list-decodability of RLCs):

An **RLC** of rate R is likely **(ρ, L)-list-decodable** if and only if

every **ρ-clustered distribution** τ over $\mathbb{F}_2^{L+1}$ has some $B \in \mathbb{F}_2^{(L+1)\times b}$ such that

$$\mathbb{E}(\#\tau B \text{ distributed matrices in } C) \to 0$$

$\Pr(C$ **is** (ρ, L)**-list-decodable**$)$



Corollary (list-decodability of RLCs):

An **RLC** of rate R is likely (ρ, L)**-list-decodable** if and only if

every ρ**-clustered distribution** τ over $\mathbb{F}_2^{L+1}$ has some $B \in \mathbb{F}_2^{(L+1)\times b}$ such that

$\mathbb{E}(\#\tau B$ distributed matrices in $C) \to 0$

Take aways from the threshold theorem

Take aways from the threshold theorem

- The **list-decodability** of an **RLC** can be explained by **expectations**.
    - Namely, we only care about certain terms of the form

$$2^{n\left(H(\tau) - (1-R)\cdot \dim(\mathrm{supp}(\tau))\right)}$$

Take aways from the threshold theorem

- The **list-decodability** of an **RLC** can be explained by **expectations**.
  - Namely, we only care about certain terms of the form

$$2^{n\left(H(\tau) - (1-R)\cdot \dim(\mathrm{supp}(\tau))\right)}$$

- This holds for more than just **list-decodability**.
  - Any **property** characterized by **“foribdden distributions”** has such a characterization.
  - For example, **list-recoverability**!
  - In general, any **monotone, local and symmetric property**.

But what did we gain?

But what did we gain?

- Reasoning about **list-decodability** of **RLCs** via expectations is **complete**.

But what did we gain?

- Reasoning about **list-decodability** of **RLCs** via expectations is **complete**.
- But what is this good for? we already know (through a long line of works) that **RLCs achieve the list-decoding GV-bound**.

But what did we gain?

- Reasoning about **list-decodability** of **RLCs** via expectations is **complete**.
- But what is this good for? we already know (through a long line of works) that **RLCs achieve the list-decoding GV-bound**.
- But now these results tell us something about **expectations**!

Definition: A random code ensemble $C \subseteq \mathbb{F}_q^n$ is **locally-similar** to an RLC of rate R if

$$\Pr\left[\{v_1, \ldots, v_k\} \subseteq C\right] \approx 2^{-(1-R)\cdot n \cdot \dim\{v_1,\ldots,v_k\}}$$

for all $v_1, \ldots, v_k \in \mathbb{F}_q^n$.

Theorem: If C is **locally-similar** to an **RLC of rate** R then it **achieves the list-decoding GV-bound** with high probability.

Theorem: If C is **locally-similar** to an **RLC of rate R** then it **achieves the list-decoding GV-bound** with high probability.

Proof:

Theorem: If C is **locally-similar** to an **RLC of rate R** then it **achieves the list-decoding GV-bound** with high probability.

Proof:

Let D be an **RLC of rate R**. We know from previous works that an **D almost surely achieves the list-decoding GV-bound**.

Let ρ,L such that D **is likely (ρ, L)-list-decodable**. It suffices to show that **the same holds for C**.

Theorem: If C is **locally-similar** to an **RLC of rate R** then it **achieves the list-decoding GV-bound** with high probability.

Proof:

Let D be an **RLC of rate R**. We know from previous works that an **D almost surely achieves the list-decoding GV-bound**.

Let ρ,L such that D **is likely (ρ, L)-list-decodable**. It suffices to show that **the same holds for C**.

Let τ be a **ρ-clustered distribution** over $\mathbb{F}_q^{L+1}$. Then D is **unlikely to contain a τ-distributed matrix**. By the **threshold theorem**, there is some B such that

$$\mathbb{E}\left[\#\tau B\text{-distributed matrices in } D\right] \leq o(1).$$

Theorem: If C is **locally-similar** to an **RLC of rate R** then it **achieves the list-decoding GV-bound** with high probability.

Proof:

Let D be an **RLC of rate R**. We know from previous works that an **D almost surely achieves the list-decoding GV-bound**.

Let ρ,L such that D **is likely (ρ, L)-list-decodable**. It suffices to show that **the same holds for C**.

Let τ be a **ρ-clustered distribution** over $\mathbb{F}_q^{L+1}$. Then D is **unlikely to contain a τ-distributed matrix**. By the **threshold theorem**, there is some B such that

$$\mathbb{E}\left[\#\tau B\text{-distributed matrices in } D\right] \le o(1).$$

But

$$\begin{aligned}\mathbb{E}\left[\#\tau B\text{-distributed matrices in } C\right] &\approx \#\tau B\text{-distributed matrices} \cdot 2^{-(1-R)n\cdot\dim(\mathrm{supp}(\tau))} \\ &= \mathbb{E}\left[\#\tau B\text{-distributed matrices in } D\right] \le o(1)\end{aligned}$$

So C **is unlikely to contain τB and thus unlikely to contain τ**.

Theorem: If C is **locally-similar** to an **RLC of rate R** then it **achieves the list-decoding GV-bound** with high probability.

The same argument works for **list-recovery** or any other **local symmetric property:**

Theorem: If C is **locally-similar** to an **RLC of rate R** then it **achieves the same list-recovery parameters** as an RLC.

The reduction paradigm

1. Choose a **random code ensemble** C.

2. Show that C is **locally-similar** to an **RLC**.

3. Conclude that C has all the local symmetric properties of an RLC, including **achieving the list-decoding GV-bound**.

The reduction paradigm

1. Choose a **random code ensemble** C.
2. Show that C is **locally-similar** to an **RLC**.
3. Conclude that C has all the local symmetric properties of an RLC, including **achieving the list-decoding GV-bound**.

The reduction paradigm

1. Choose a **random code ensemble** C.

2. Show that C is **locally-similar** to an **RLC**.

3. Conclude that C has all the local symmetric properties of an RLC, including **achieving the list-decoding GV-bound**.

Done successfully for:

- **Random LDPC codes (Gallagher's Ensemble)** [M-Resch-(Ron-Zewi)-Silas,Wootters]
- **Randomly punctured low-bias codes** [Guruswami-M]

The reduction paradigm

1. Choose a **random code ensemble** C.

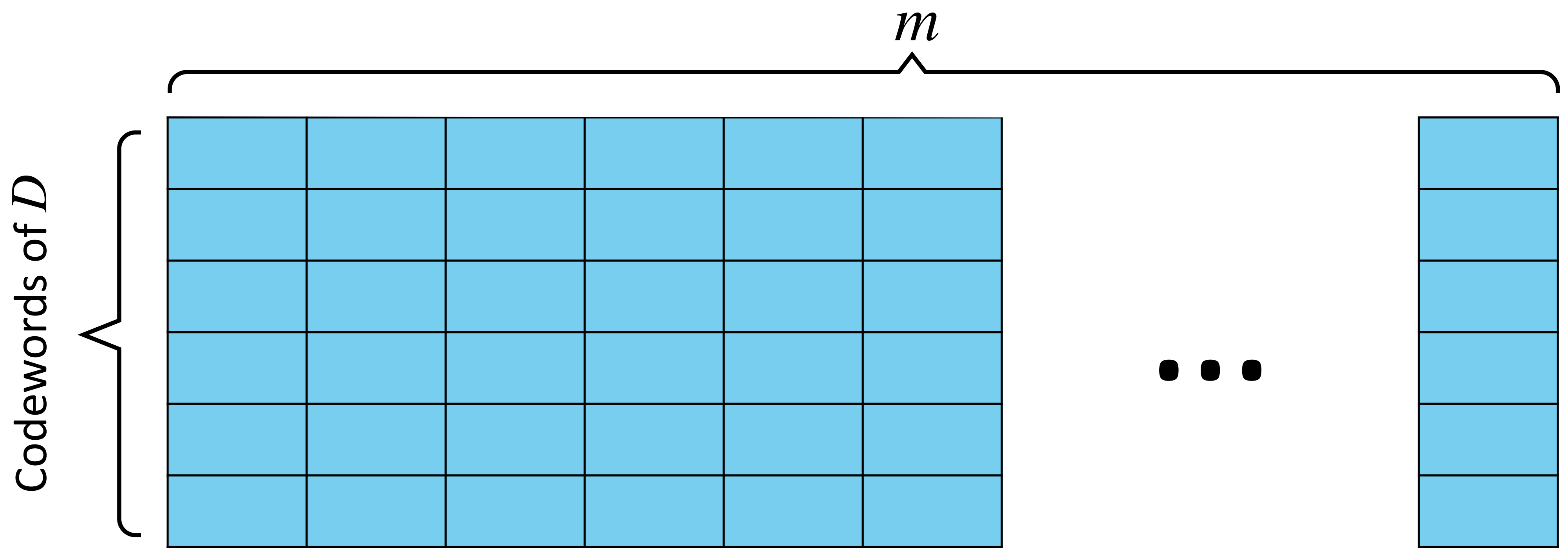
2. Show that C is **locally-similar** to an **RLC**.

3. Conclude that C has all the local symmetric properties of an RLC, including **achieving the list-decoding GV-bound**.

Done successfully for:

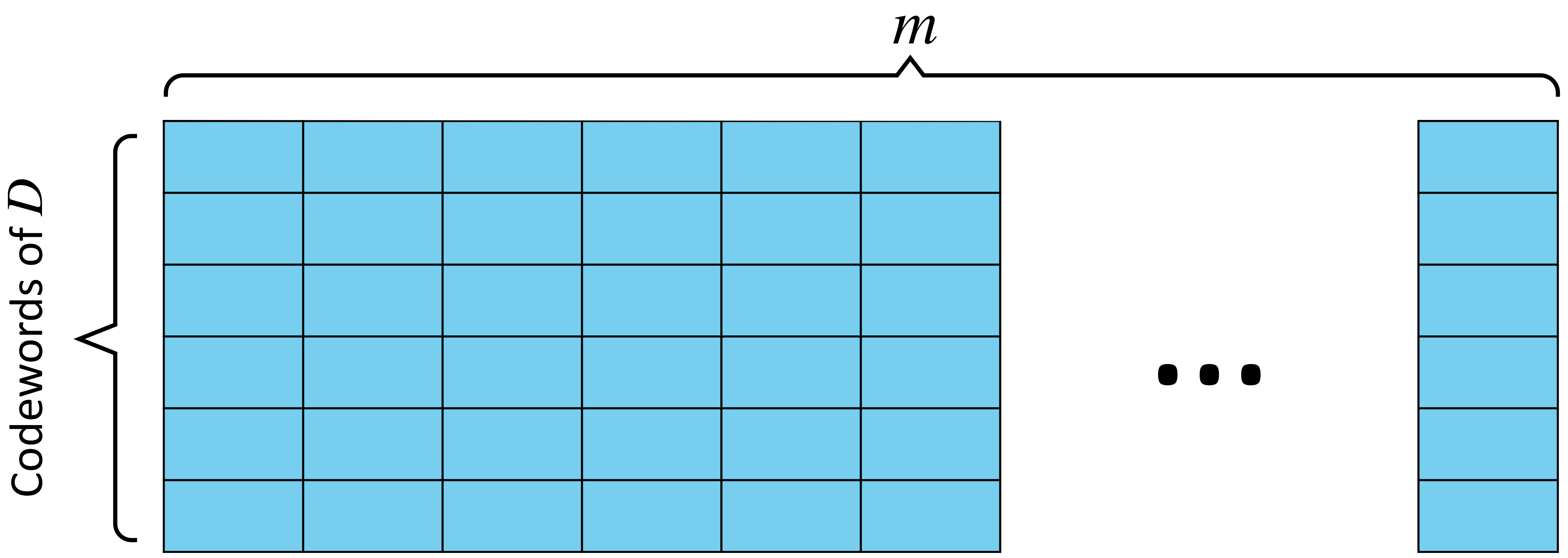
- **Random LDPC codes (Gallagher's Ensemble) [M-Resch-(Ron-Zewi)-Silas,Wootters]**
- **Randomly punctured low-bias codes [Guruswami-M]**

Puncturing of Codes



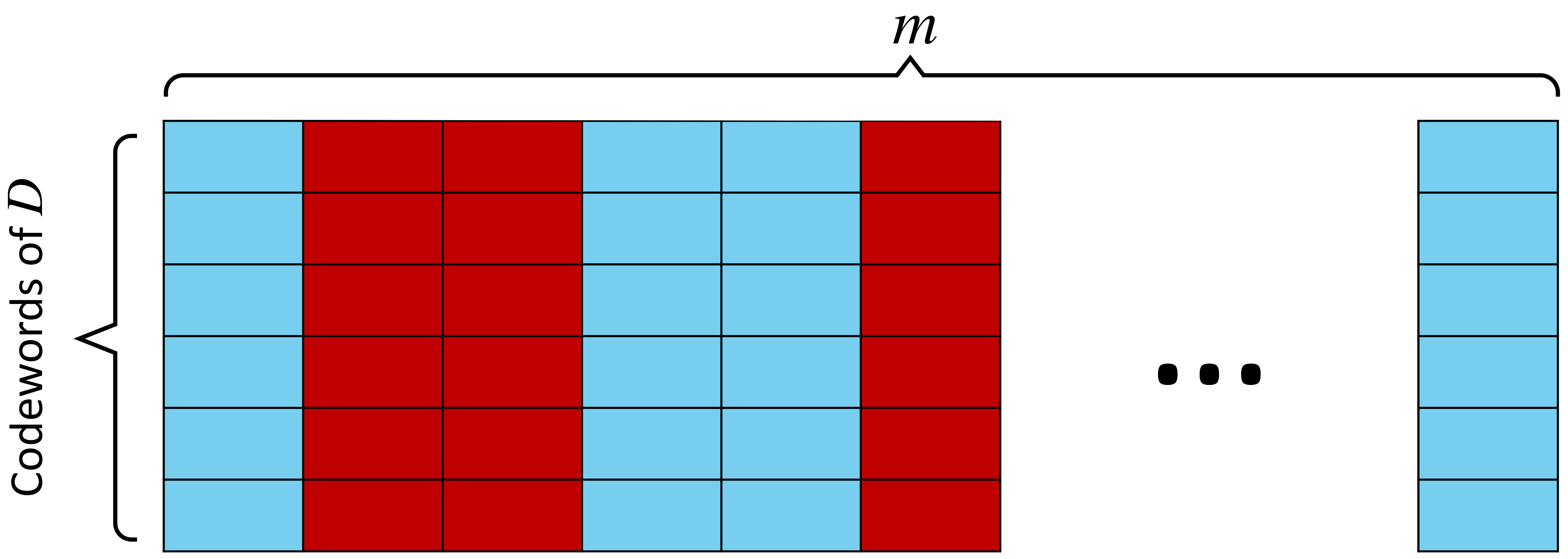
Puncturing of Codes

- From a code $D \subseteq \mathbb{F}_q^m$ to $C \subseteq \mathbb{F}_q^n$. Usually $n \ll m$.



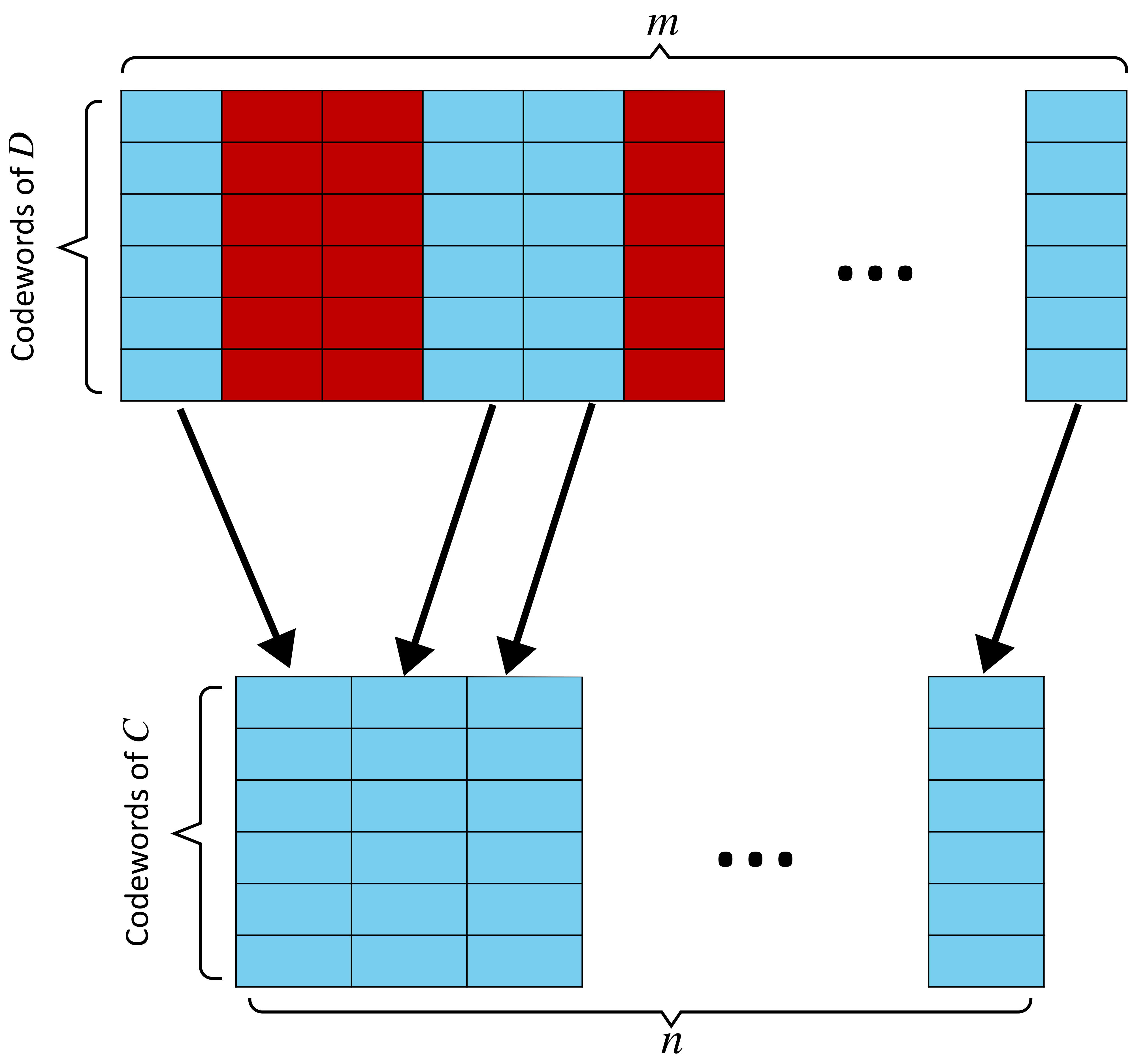
Puncturing of Codes

- From a code $D \subseteq \mathbb{F}_q^m$ to $C \subseteq \mathbb{F}_q^n$. Usually $n \ll m$.



Puncturing of Codes

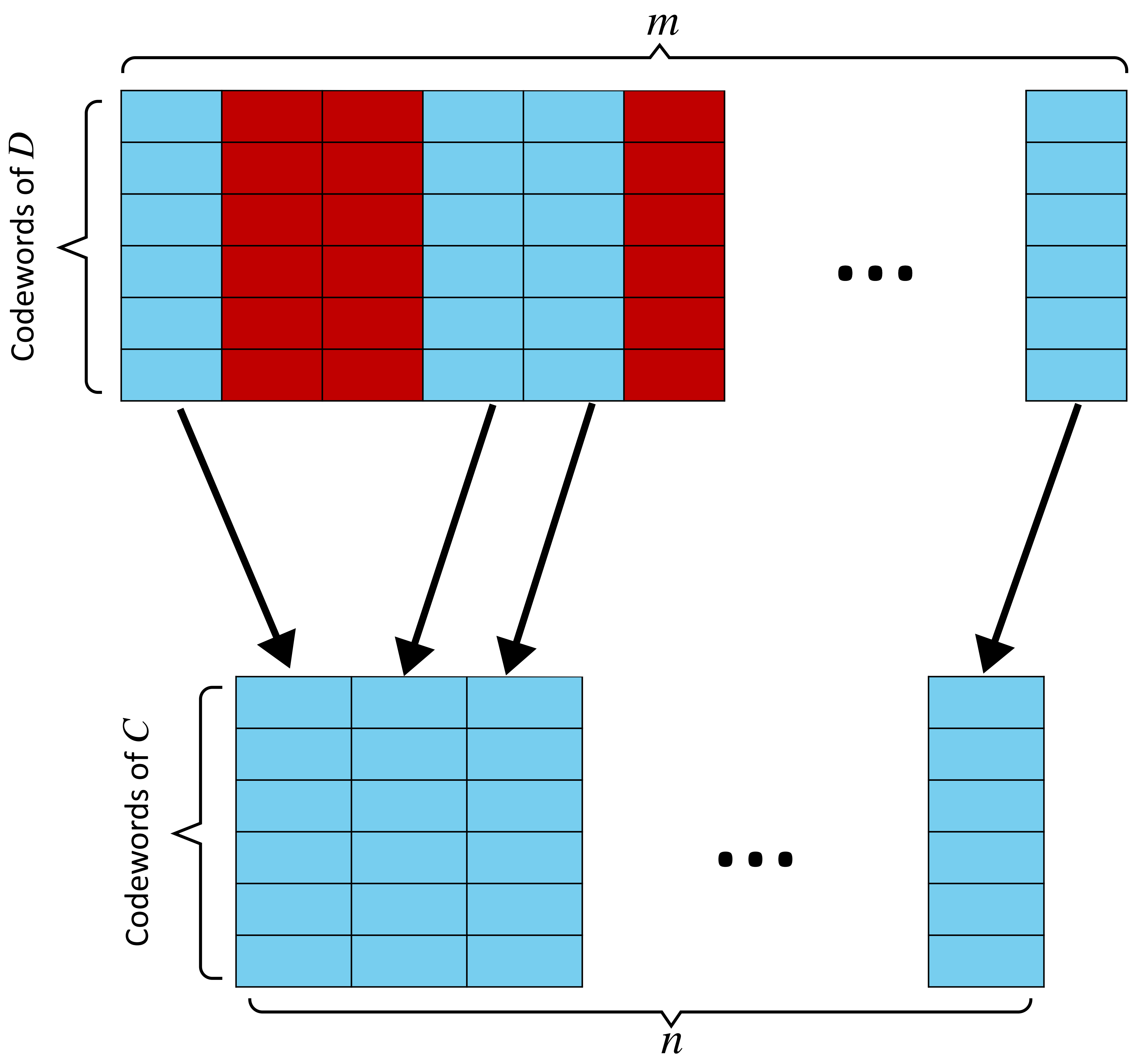
- From a code $D \subseteq \mathbb{F}_q^m$ to $C \subseteq \mathbb{F}_q^n$. Usually $n \ll m$.



Puncturing of Codes

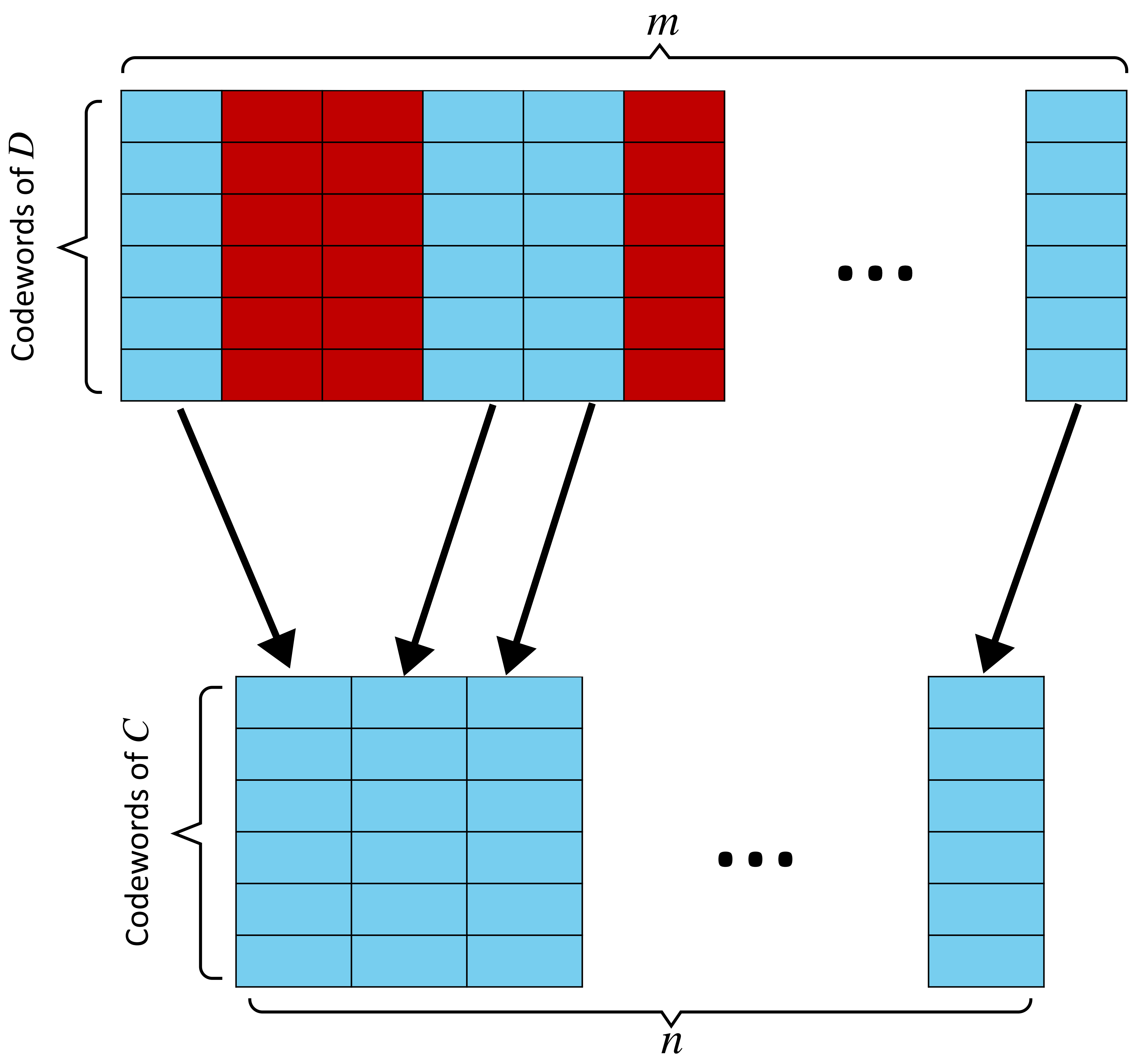
- From a code $D \subseteq \mathbb{F}_q^m$ to $C \subseteq \mathbb{F}_q^n$. Usually $n \ll m$.

- If the punctured columns are chosen at random, C is said to be a **random n-puncturing of D**.



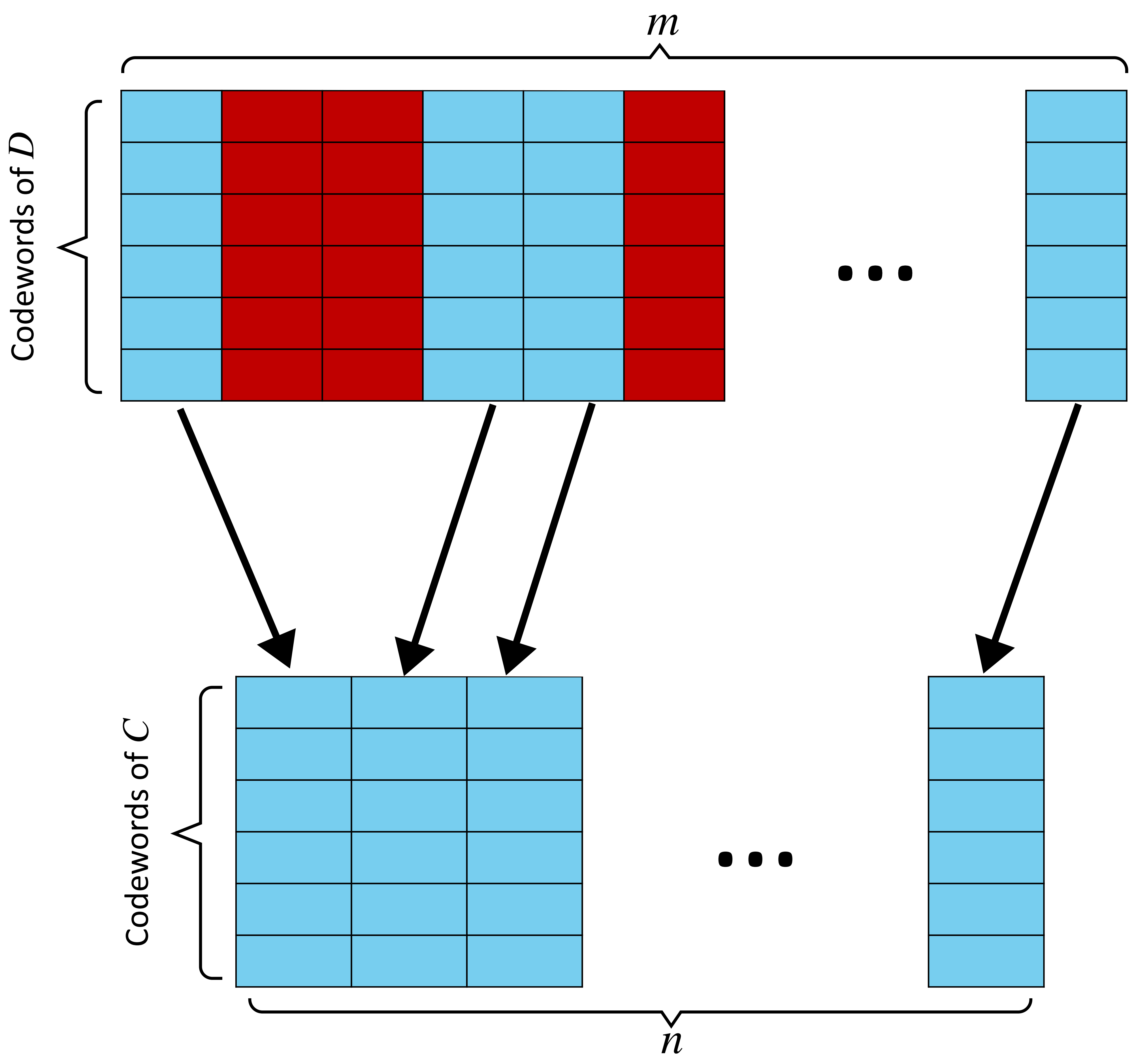
Puncturing of Codes

- From a code $D \subseteq \mathbb{F}_q^m$ to $C \subseteq \mathbb{F}_q^n$. Usually $n \ll m$.
- If the punctured columns are chosen at random, C is said to be a **random n-puncturing of D**.
- **Example:** An **RLC** of rate R in $\mathbb{F}_q^n$ is a **random puncturing** of the **Hadamard code** $H \subseteq \mathbb{F}_q^{q^{Rn}}$.

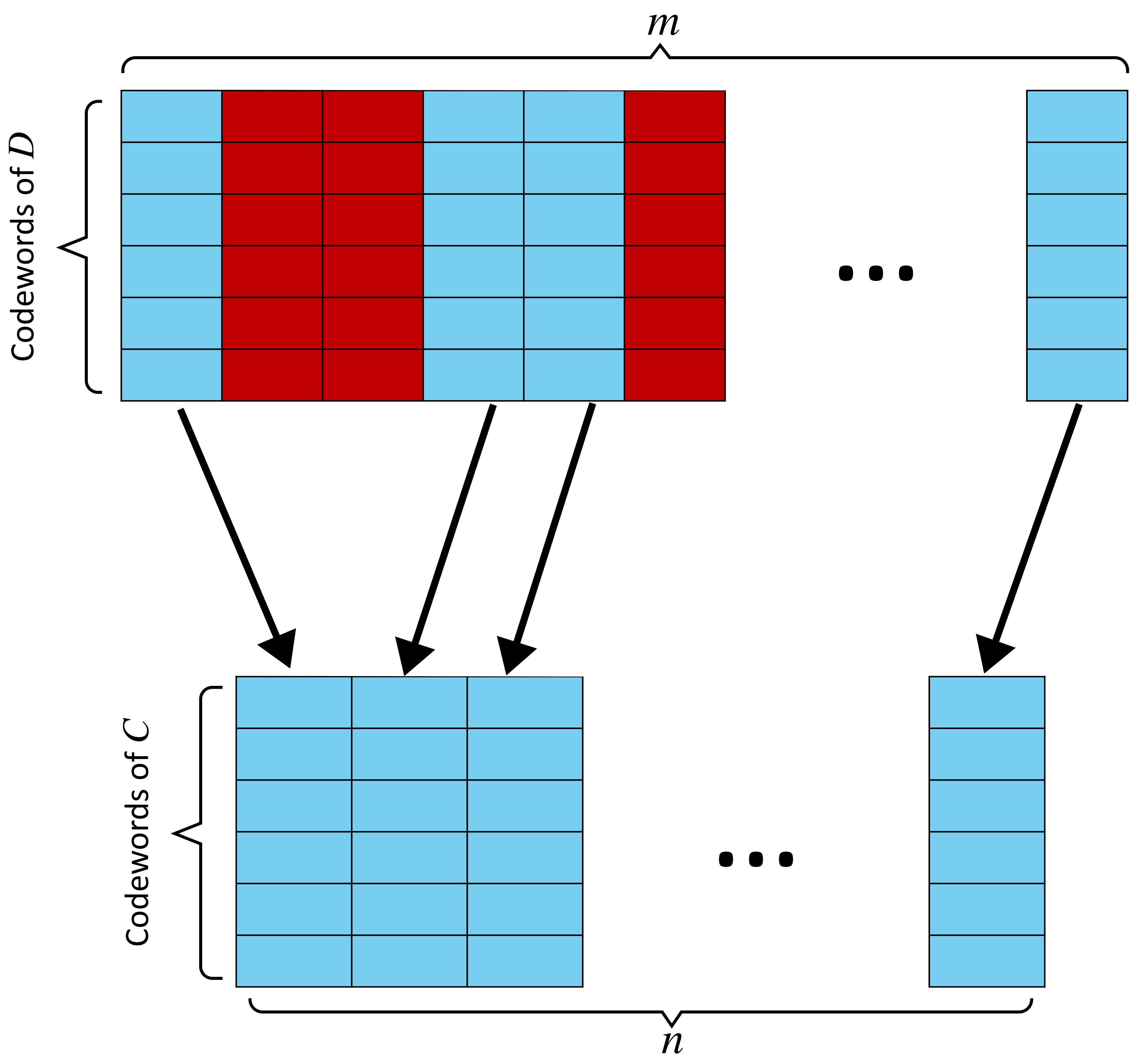


Puncturing of Codes

- From a code $D \subseteq \mathbb{F}_q^m$ to $C \subseteq \mathbb{F}_q^n$. Usually $n \ll m$.
- If the punctured columns are chosen at random, C is said to be a **random n-puncturing of D**.
- **Example:** An **RLC** of rate R in $\mathbb{F}_q^n$ is a **random puncturing** of the **Hadamard code** $H \subseteq \mathbb{F}_q^{q^{Rn}}$.
- A **Reed-Solomon code over a random evaluation set** is a random puncturing of the **full Reed-Solomon code.**

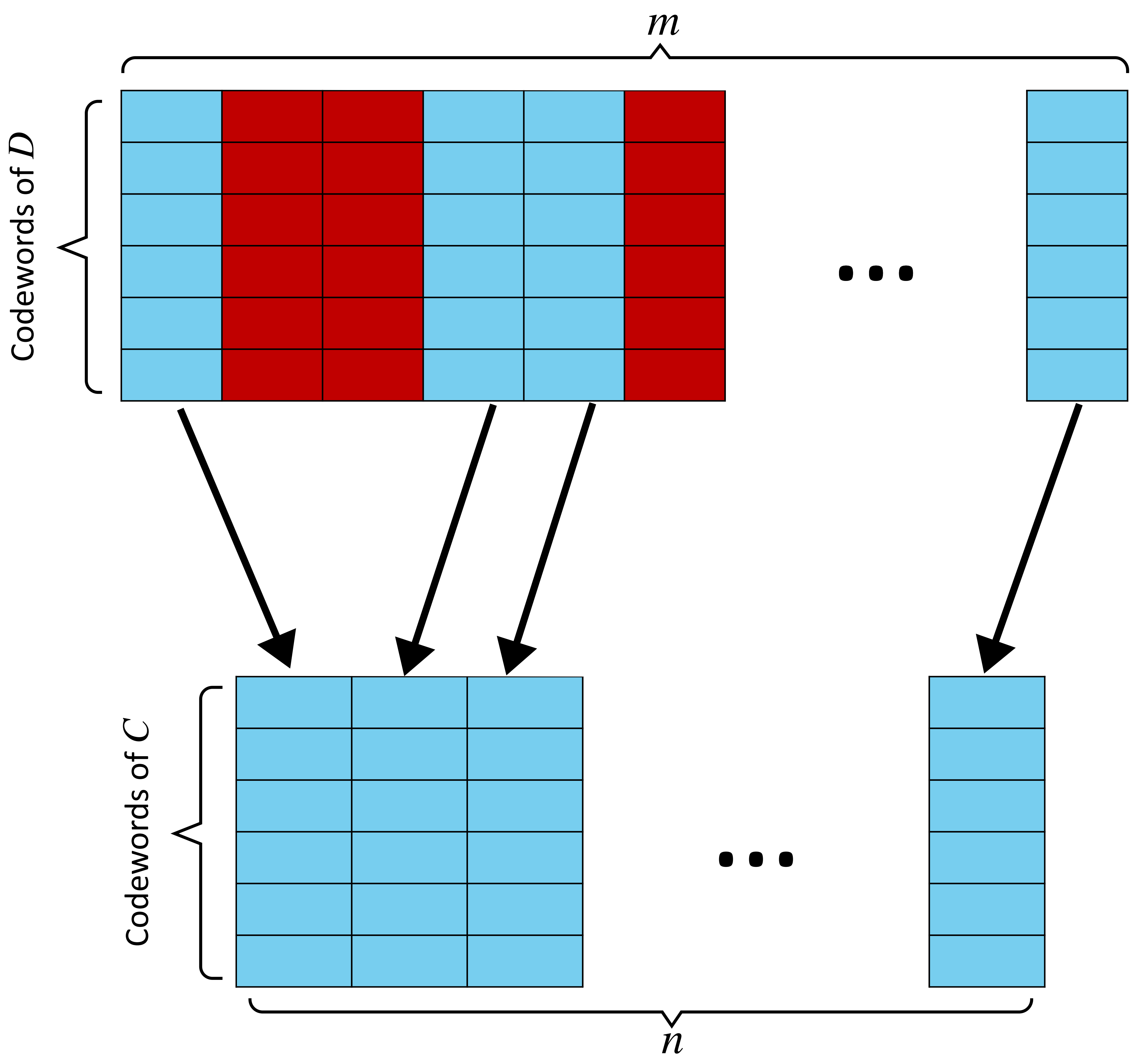


Puncturing of low-bias codes



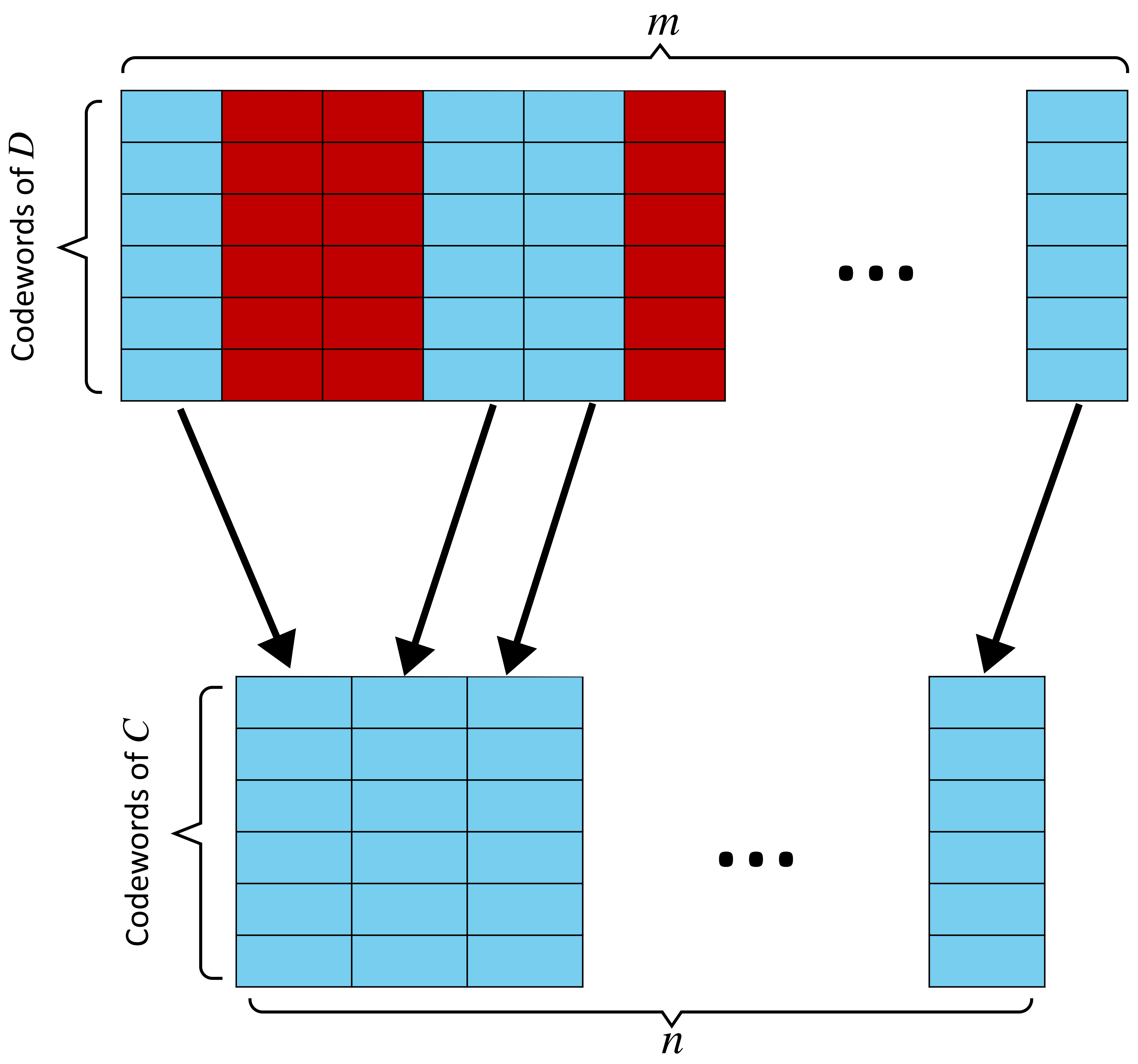
Puncturing of low-bias codes

- Let's focus on $q = 2$



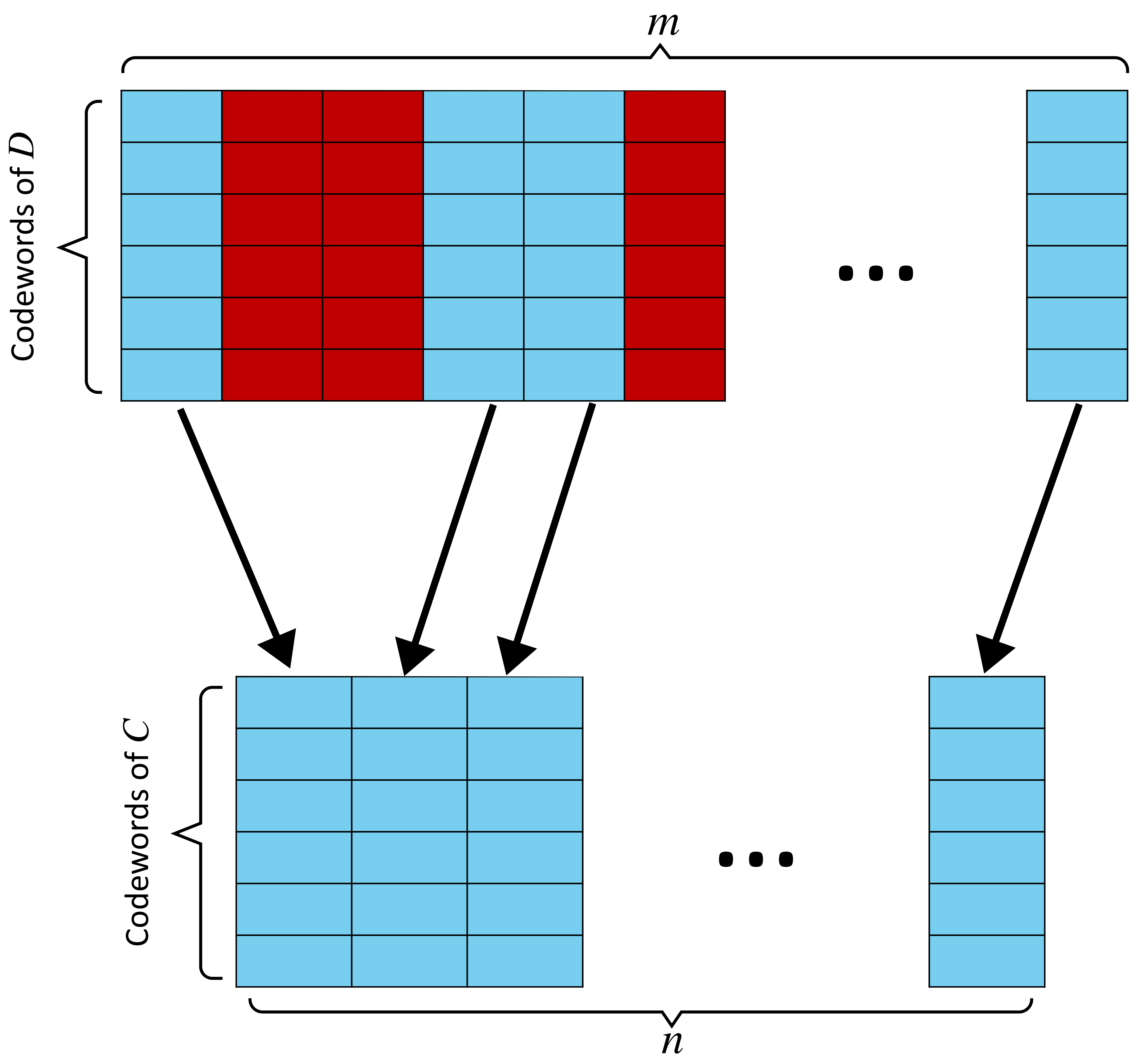
Puncturing of low-bias codes

- Let's focus on $q = 2$
- Suppose every $u \in D$ has weight close to $\frac{m}{2}$ (low-bias).



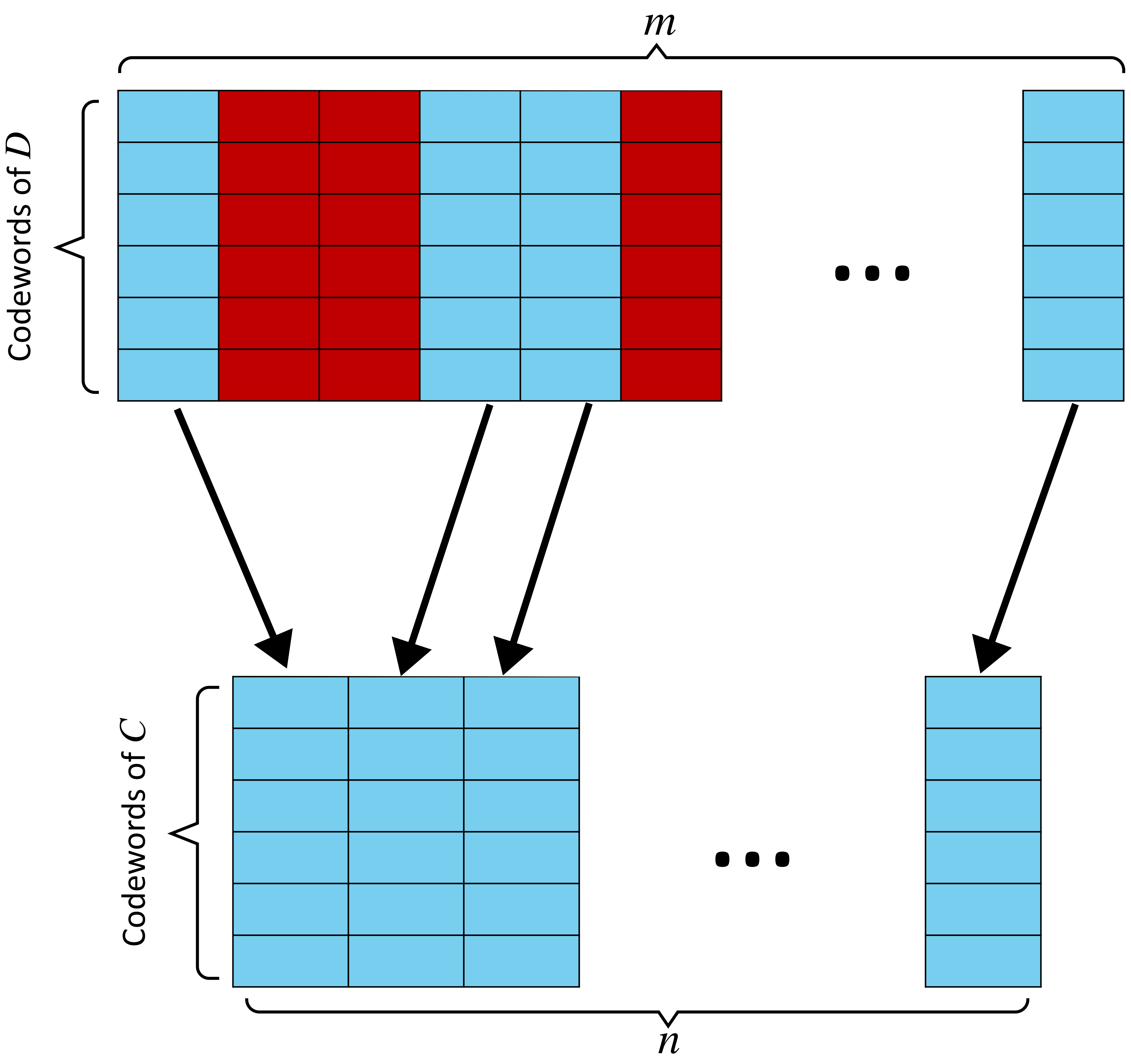
Puncturing of low-bias codes

- Let's focus on $q = 2$
- Suppose every $u \in D$ has weight close to $\frac{m}{2}$ (low-bias).
- **Claim:** C locally-similar to an RLC.



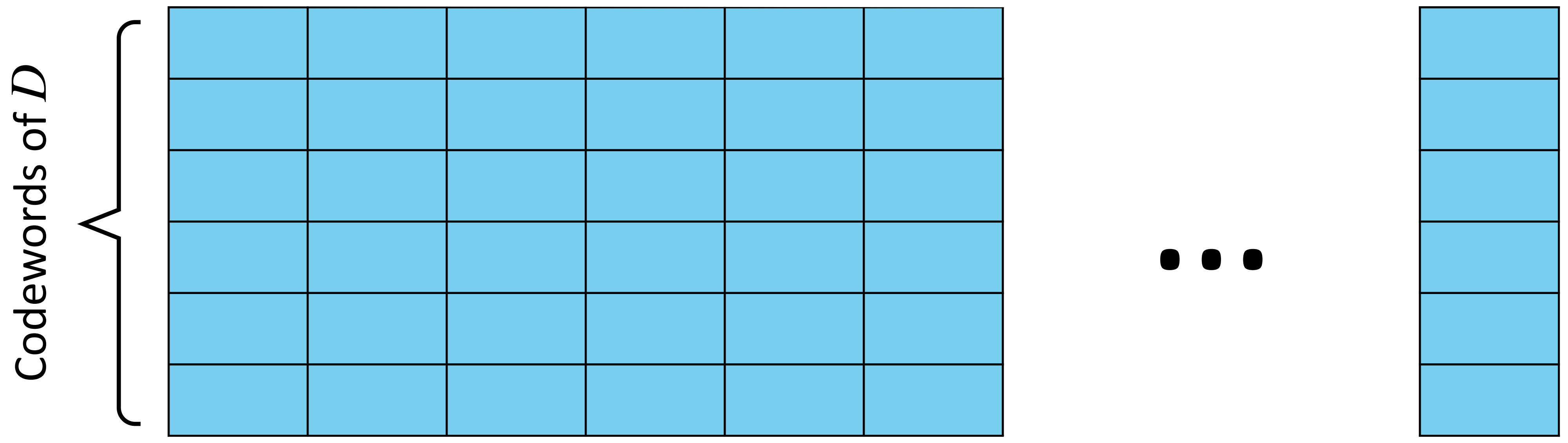
Puncturing of low-bias codes

- Let's focus on $q = 2$
- Suppose every $u \in D$ has weight close to $\frac{m}{2}$ (low-bias).
- **Claim:** C locally-similar to an RLC.
- **Conclusion:** C is as list-decodable and list-recoverable as an RLC.

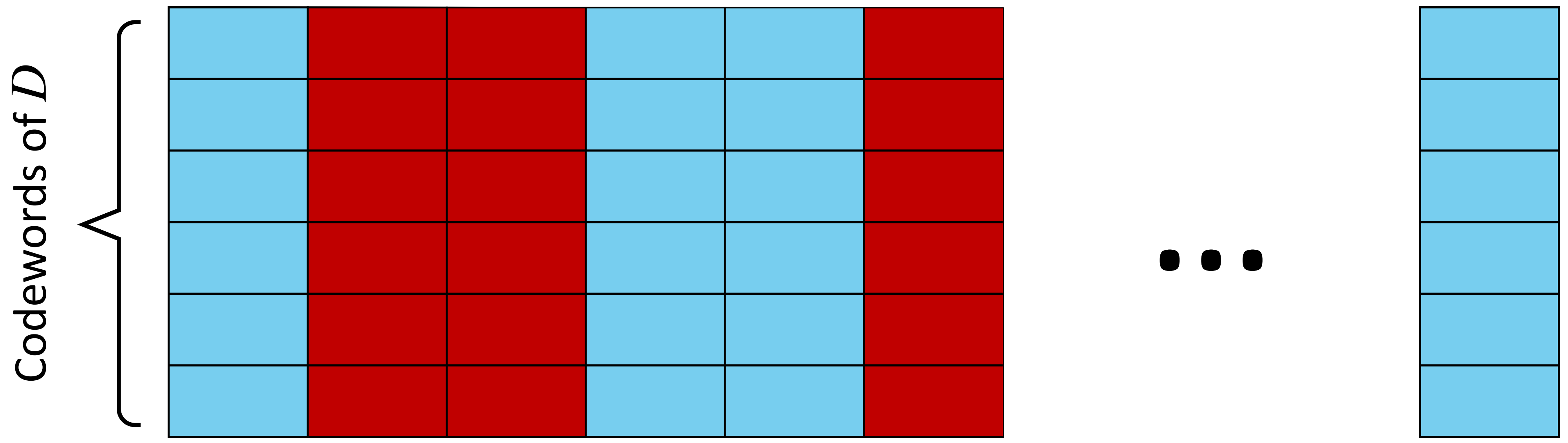


Proof sketch: C locally-similar to an RLC.

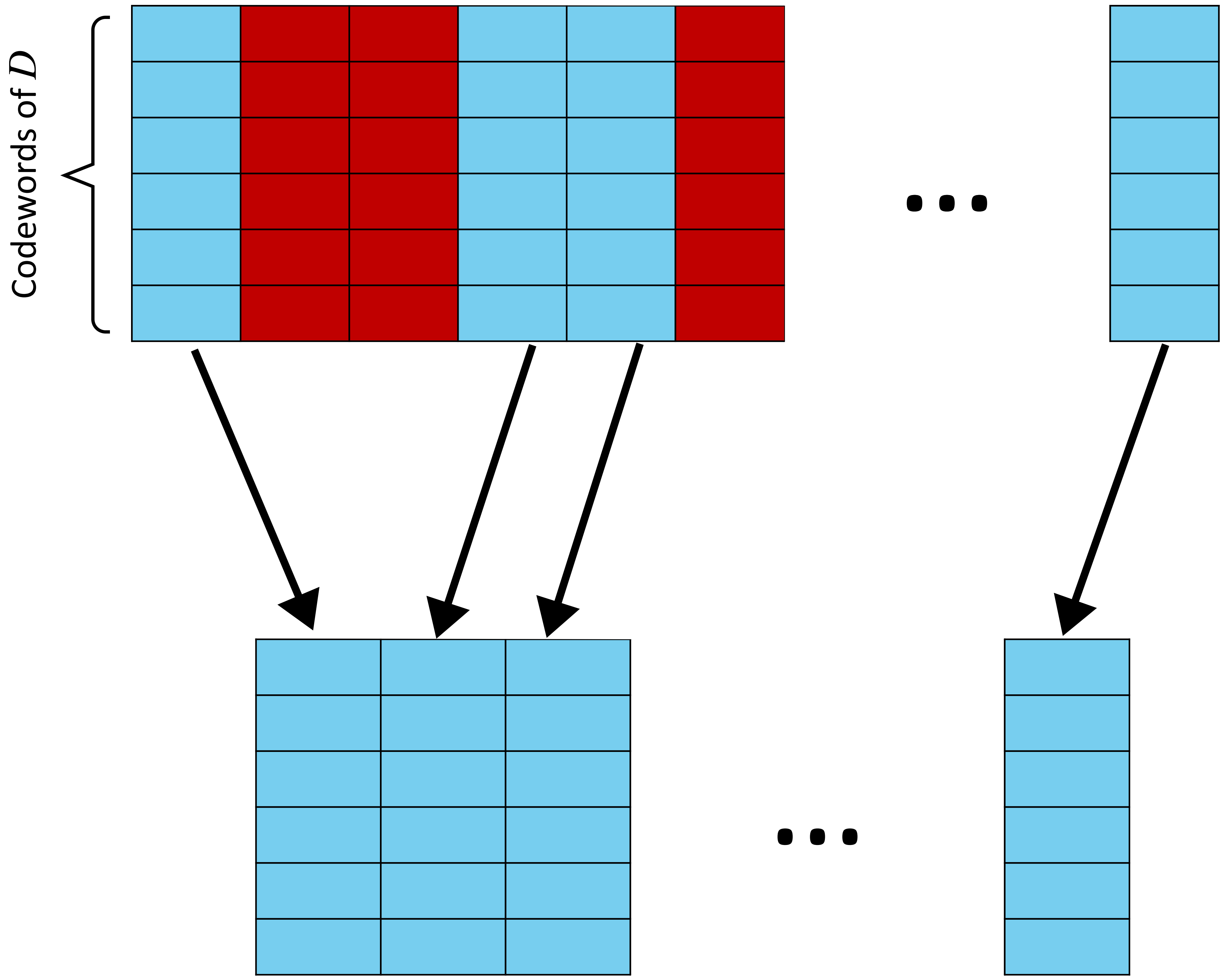
Proof sketch: C locally-similar to an RLC.



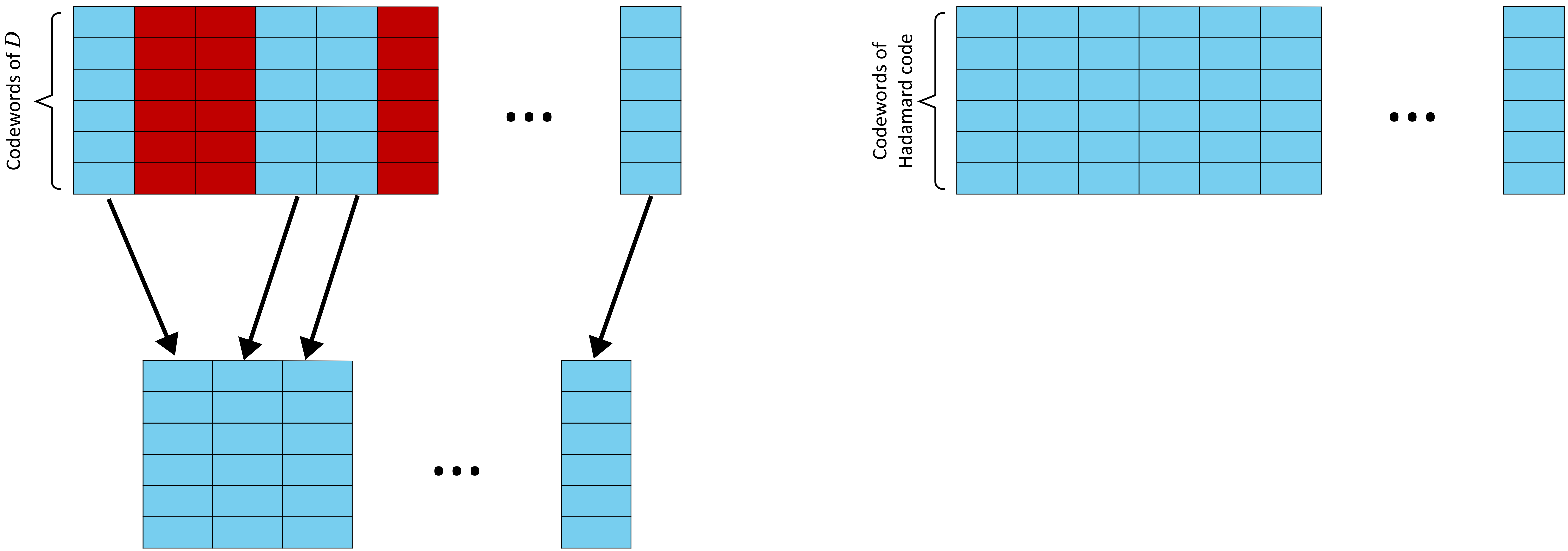
Proof sketch: C locally-similar to an RLC.



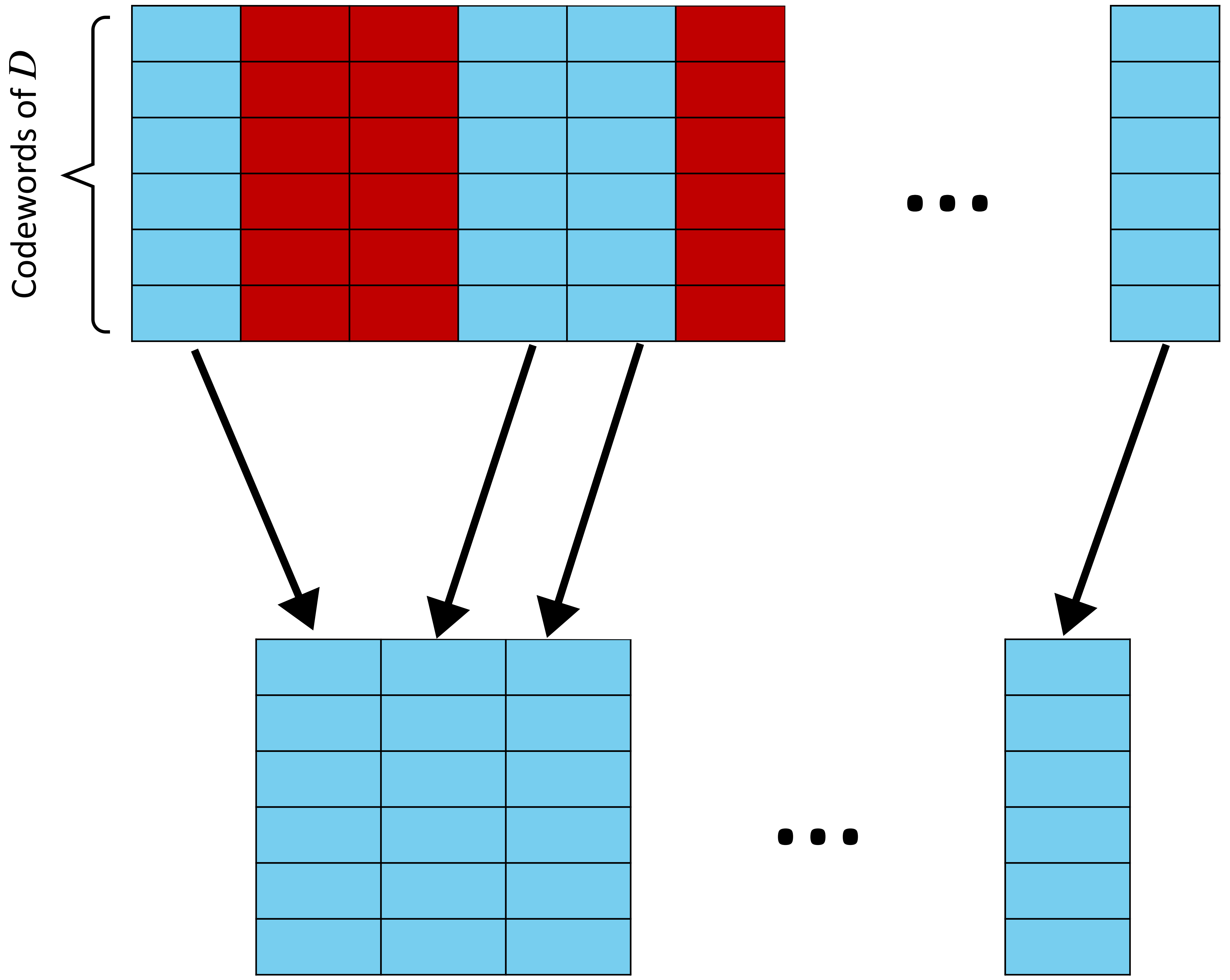
Proof sketch: C locally-similar to an RLC.

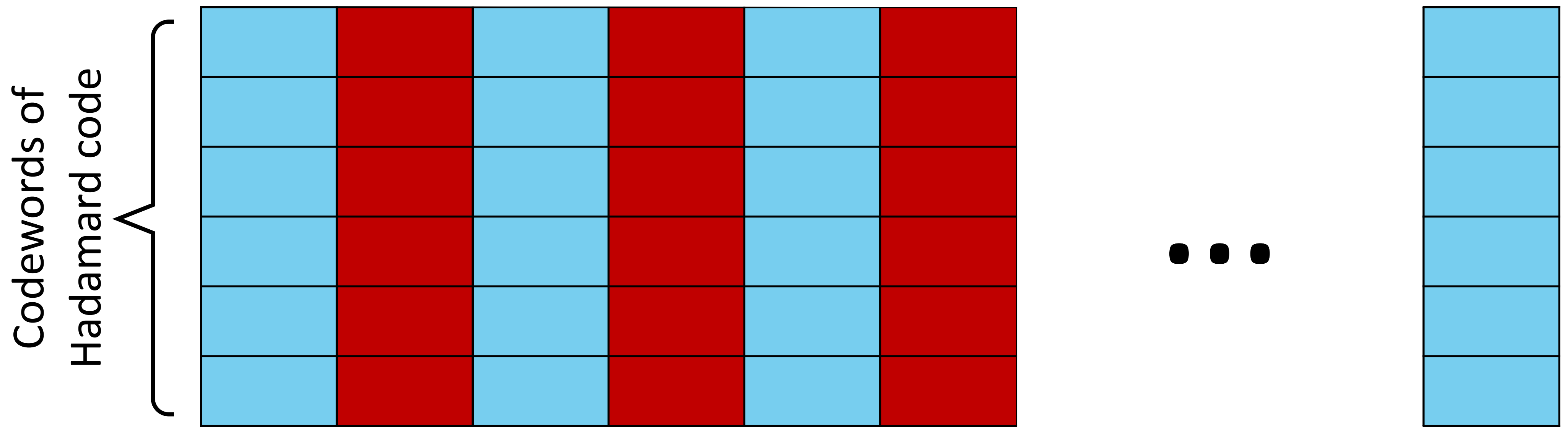


Proof sketch: C locally-similar to an RLC.

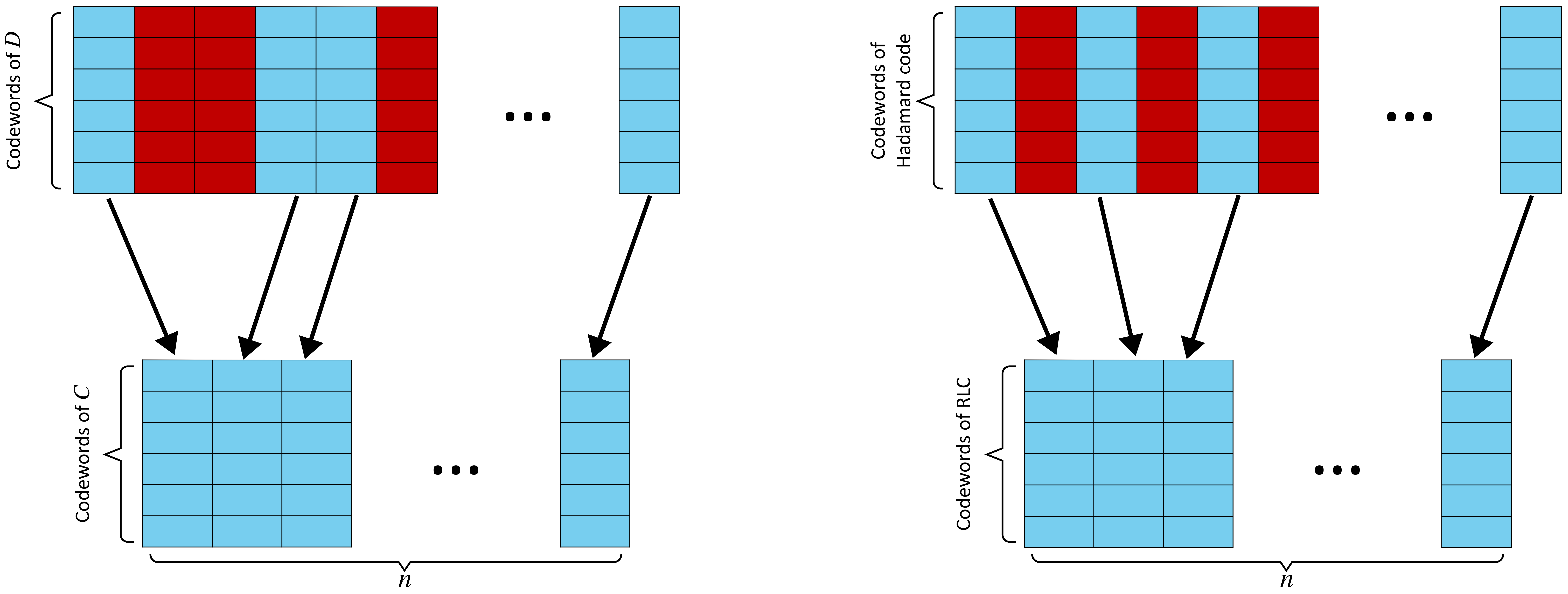


Proof sketch: C locally-similar to an RLC.

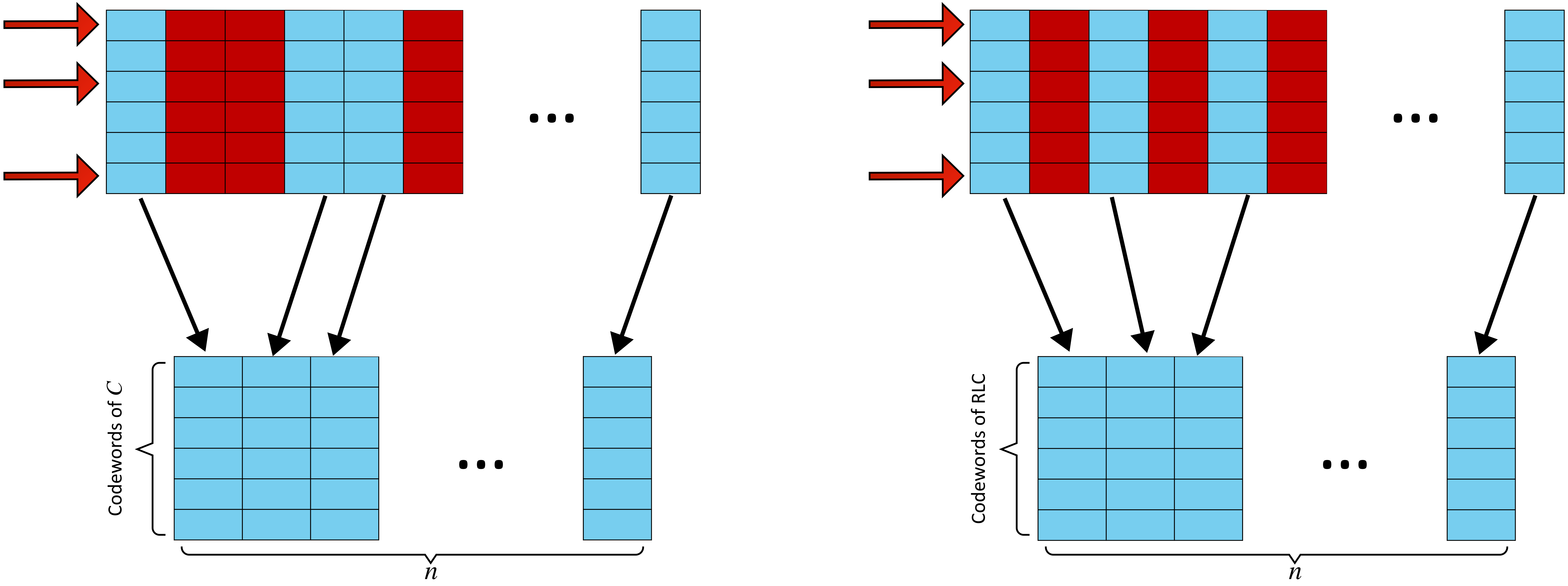




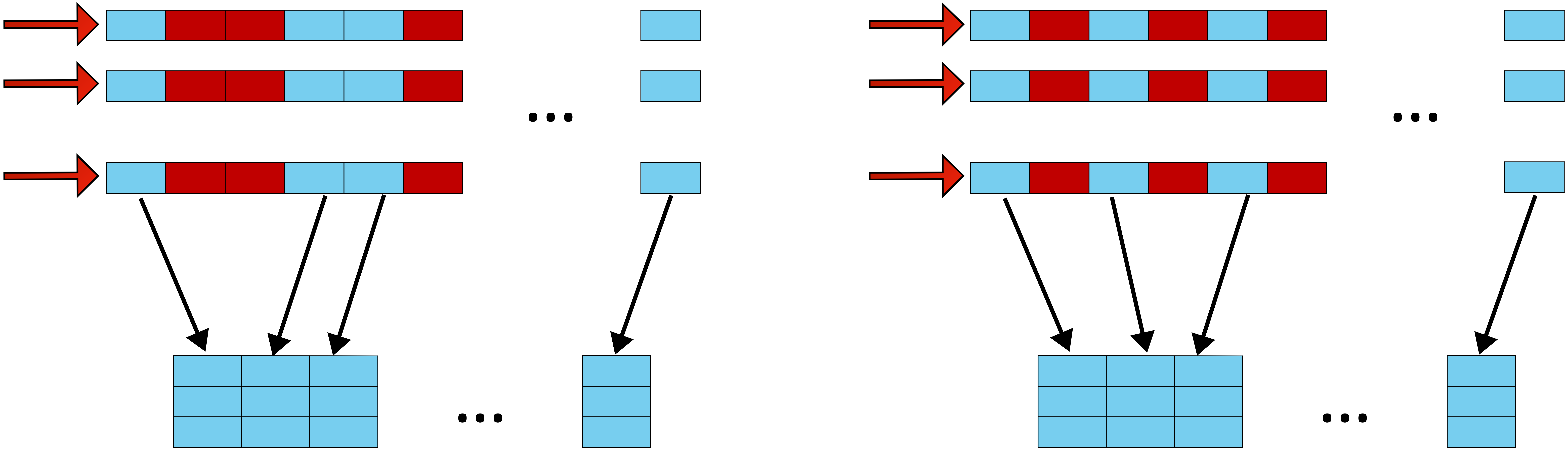
Proof sketch: C locally-similar to an RLC.



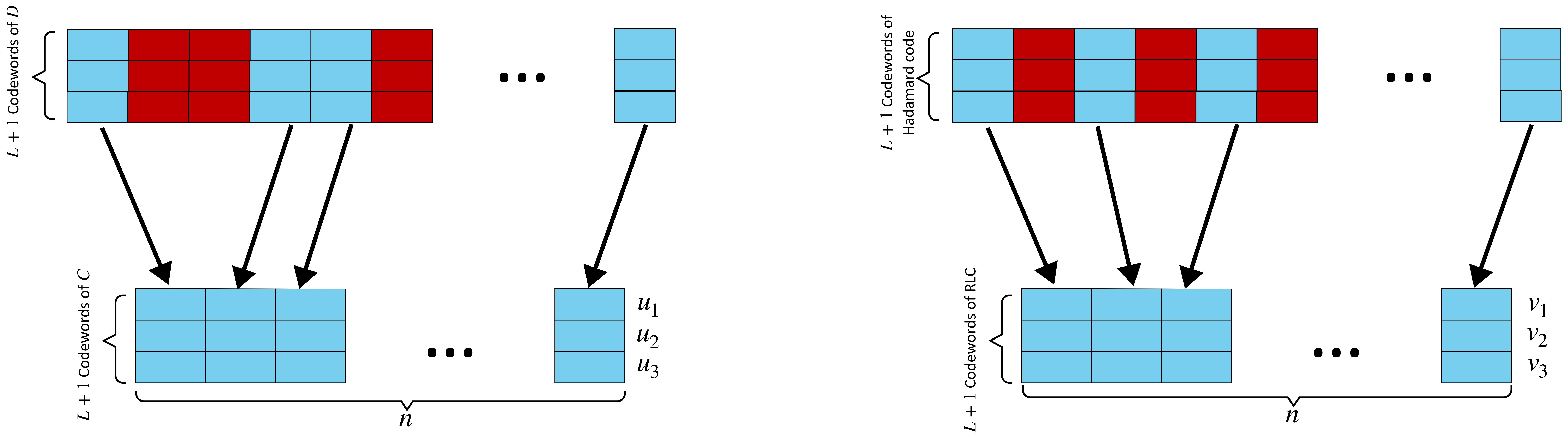
Proof sketch: C locally-similar to an RLC.



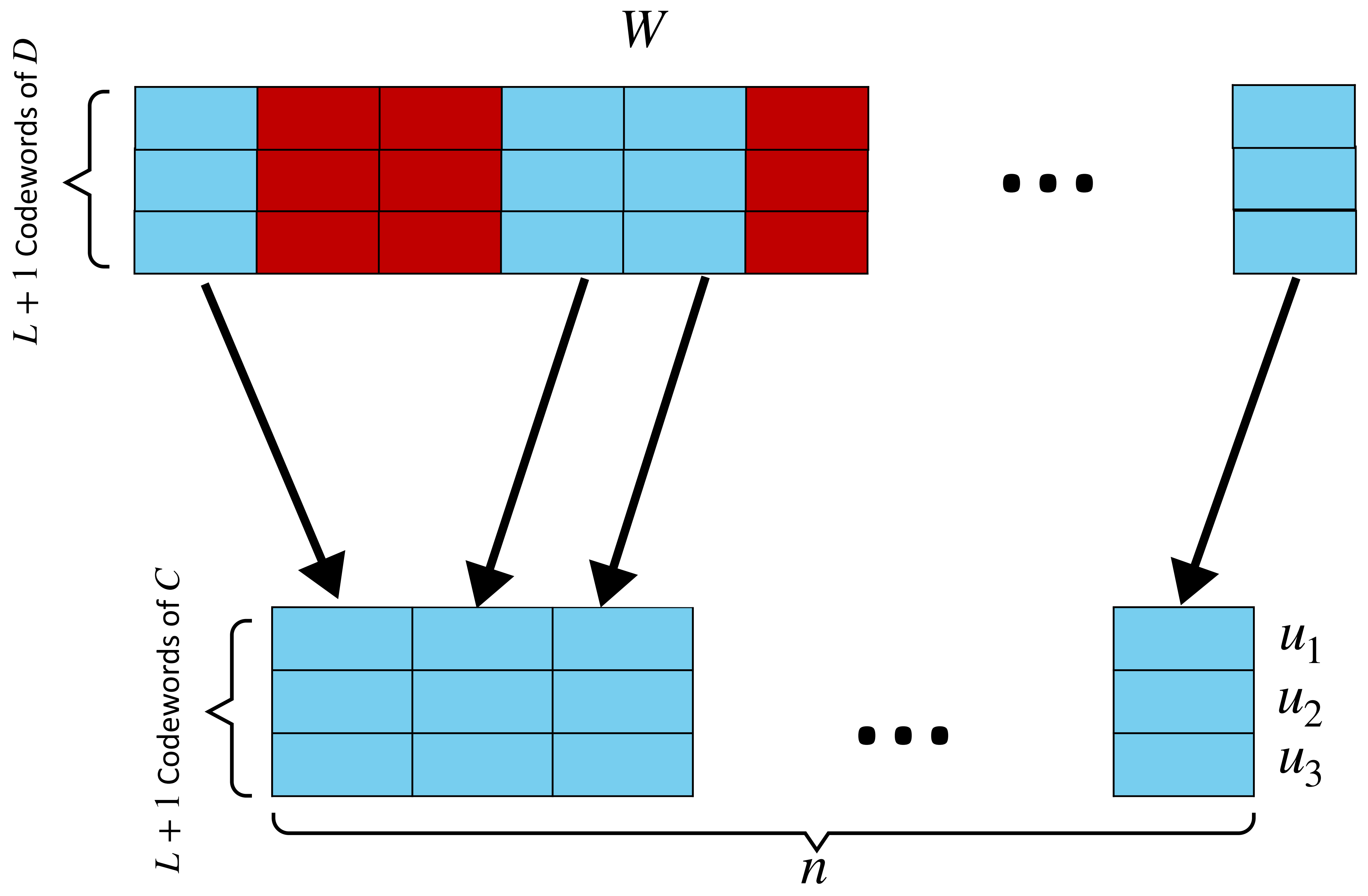
Proof sketch: C locally-similar to an RLC.

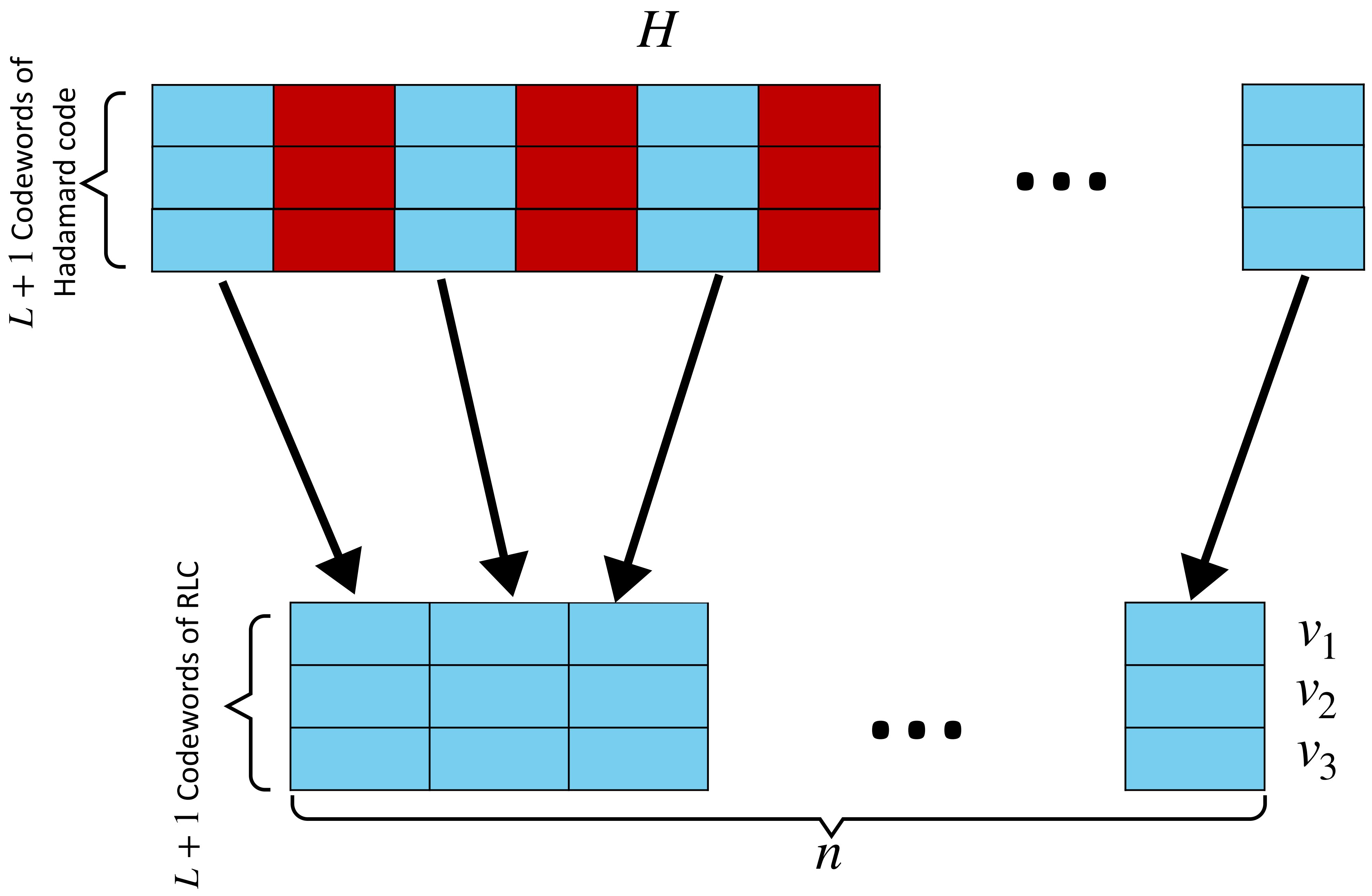


Proof sketch: C locally-similar to an RLC.

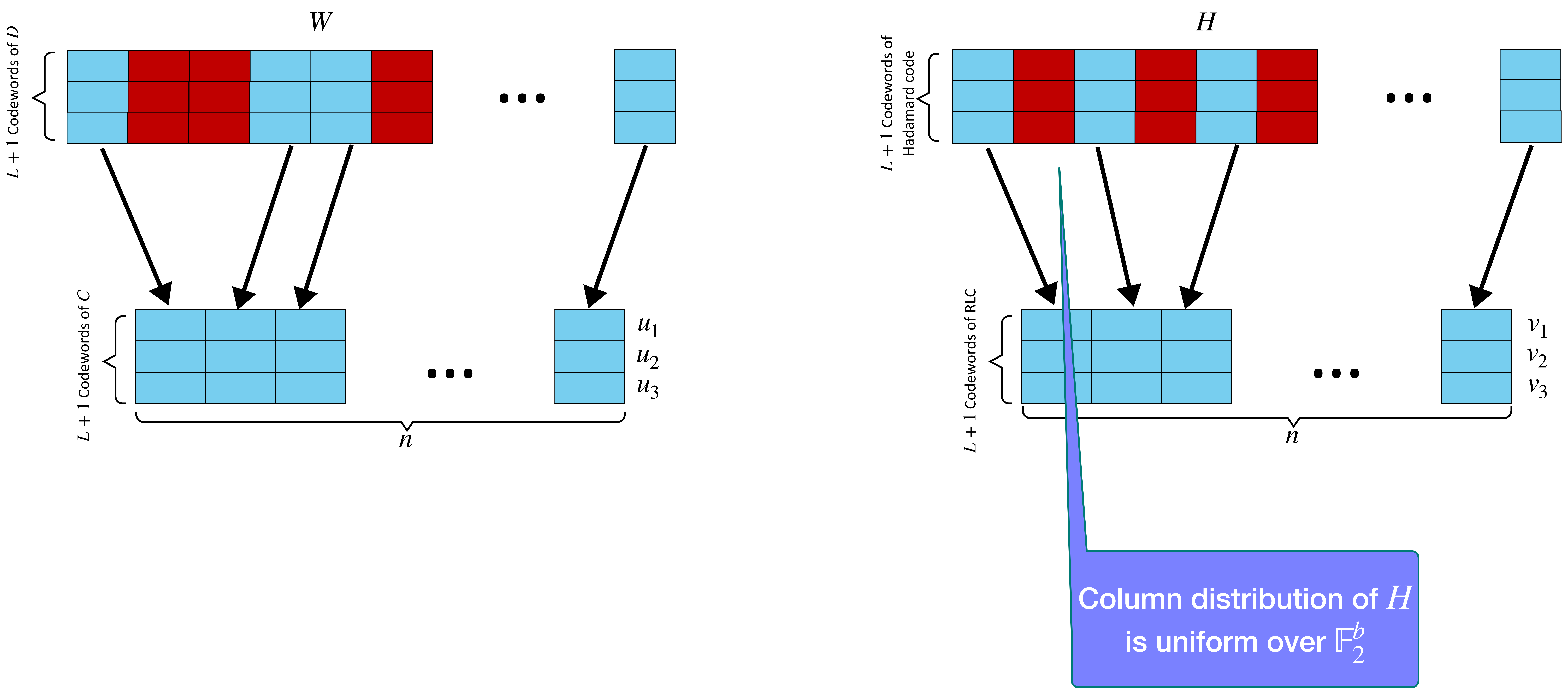


Proof sketch: C locally-similar to an RLC.





Proof sketch: C locally-similar to an RLC.



Proof sketch: C locally-similar to an RLC.