

# Error exponents & non asymptotics

Stark Draper

European School on Information Theory  
Ohrid, Macedonia  
April 2013

Funding gratefully acknowledged:  
AFOSR “Complex Networks” program  
NSF CCF CAREER program



UNIVERSITY OF  
**TORONTO**



THE UNIVERSITY  
*of*  
**WISCONSIN**  
MADISON

# Objectives of talk

- Beyond typicality decoding
- As block length increases:
  - How quickly does error drop?
  - How quickly do you approach capacity?
- Intro to tools used to answer such questions:
  - Large deviations
  - Gaussian approximations
- When to use which type of tool & connections between
- Will try to illustrate general techniques and results in simplest illustrative context: BSC

# Agenda

**Analyzing decoding error for a bounded information decoder: regimes of interest**

Error exponents of ML decoders

Non-asymptotic analysis of ML decoder & Normal approximation

# Agenda

## Analyzing decoding error for a bounded information decoder: regimes of interest

### Error exponents of ML decoders

Today we follow a “recentered” analysis due to Forney (notes '01, see also Barg-Forney '02).

Also see classic texts: Gallager & Csiszár-Körner

### Non-asymptotic analysis of ML decoder & Normal approximation

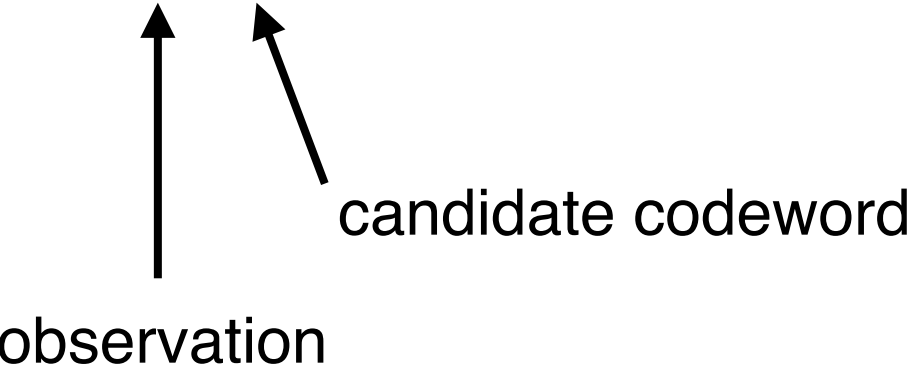
We follow development in Polyanskiy-Poor-Verdú (PPV '10). Extensive historic context is provided therein, particularly Strassen '62



# Motivating question

Want to analyze ML decoding

Given some codebook  $\mathcal{C}$  the ML decoding rule is

$$\hat{\mathbf{x}} = \operatorname{argmax}_{\mathbf{x} \in \mathcal{C}} p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$$


observation

candidate codeword

Minimizes average probability of error

# A few manipulations

$$\begin{aligned}\hat{\mathbf{x}} &= \operatorname{argmax}_{\mathbf{x} \in \mathcal{C}} p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) \\ &= \operatorname{argmax}_{\mathbf{x} \in \mathcal{C}} \log p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) \\ &= \operatorname{argmax}_{\mathbf{x} \in \mathcal{C}} \log \frac{p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})}{p_{\mathbf{Y}}(\mathbf{y})} \\ &= \operatorname{argmax}_{\mathbf{x} \in \mathcal{C}} i(\mathbf{x}; \mathbf{y})\end{aligned}$$

The “information density”,  
expectation = mutual information



# Standard random codebook ensemble analysis

- Expected average error over a random codebook ensemble
- $M$  codewords  $\mathbf{X}_0, \dots, \mathbf{X}_{M-1}$ , each length- $n$
- Each  $\mathbf{X}_m$  statistically independent of others
- Each  $\mathbf{X}_m$  generated in an i.i.d. manner  $\sim \text{Bern}(0.5)$
- Bound average probability of error

$$\begin{aligned}\Pr[\text{error}] &= \frac{1}{M} \sum_{m=0}^{M-1} \Pr[\text{error} | \mathbf{X} = \mathbf{X}_m] \\ &= \Pr[\text{error} | \mathbf{X} = \mathbf{X}_0] \\ &\leq \Pr \left[ \bigcup_{m=1}^{M-1} i(\mathbf{X}_m, \mathbf{Y}) \geq i(\mathbf{X}_0, \mathbf{Y}) \right]\end{aligned}$$

# Initial analysis: via “bounded information” decoder

Bounded information decoder with parameter  $\gamma$ : decode to  $\mathbf{X}_i$  if

- $i(\mathbf{X}_i, \mathbf{Y}) \geq \gamma$ , and
- $i(\mathbf{X}_j, \mathbf{Y}) < \gamma$  for all  $j \neq i$ .

Error bound:

$$\Pr[\text{error}] \leq \Pr \left[ (i(\mathbf{X}_0, \mathbf{Y}) < \gamma) \cup \left( \bigcup_{j=1}^{M-1} i(\mathbf{X}_j, \mathbf{Y}) \geq \gamma \right) \right]$$

$$\leq \Pr[i(\mathbf{X}_0, \mathbf{Y}) \leq \gamma] + (M - 1) \Pr[i(\mathbf{X}_1, \mathbf{Y}) \geq \gamma]$$

↑  
“outage” or  
“atypicality” event

↑  
“confusion” or  
“union bound” events

Will give us initial insight into error events and analysis regimes.

# Particularize expressions to BSCs

For a binary symmetric channel with crossover probability  $p$

$$\begin{aligned} i(\mathbf{X}, \mathbf{Y}) &= \log \frac{p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})}{p_{\mathbf{Y}}(\mathbf{y})} \\ &= \log \frac{p^{d_H(\mathbf{x}, \mathbf{y})} (1-p)^{n-d_H(\mathbf{x}, \mathbf{y})}}{2^{-n}} && \text{Common terms} \\ &= d_H(\mathbf{x}, \mathbf{y}) \log \frac{p}{1-p} + n \log \frac{1-p}{2} \end{aligned}$$

Hamming distance

and note:

- If  $p < 0.5$  then  $\log[p/(1-p)] < 0$ , so to maximize  $i(\mathbf{X}, \mathbf{Y})$  need to *minimize*  $d_H(\mathbf{X}, \mathbf{Y})$
- $d_H(\mathbf{X}, \mathbf{Y}) = \text{wt}_H(\mathbf{X} \oplus \mathbf{Y})$

# Bounded information = bounded distance decoder

For BSC simplifies:  $i(\mathbf{X}_0, \mathbf{Y}) \geq \gamma \Leftrightarrow \text{wt}_H(\mathbf{X}_0, \mathbf{Y}) \leq \Delta$

From before:

$$\begin{aligned} \Pr[\text{error}] &\leq \Pr[i(\mathbf{X}_0, \mathbf{Y}) \leq \gamma] + (M - 1) \Pr[i(\mathbf{X}_1, \mathbf{Y}) \geq \gamma] \\ &\leq \Pr[\text{wt}_H(\mathbf{X}_0 \oplus \mathbf{Y}) \geq \Delta] + (M - 1) \Pr[\text{wt}_H(\mathbf{X}_1 \oplus \mathbf{Y}) \leq \Delta] \end{aligned}$$

  
“re-center” analysis around observation

Statistical observations:

- $\mathbf{X}_0 \oplus \mathbf{Y} \sim \text{i.i.d. Bern}(p)$  by channel law & codeword dist.
- Since  $\mathbf{X}_1 \sim \text{Bern}(0.5)$  and  $\mathbf{X}_1 \perp\!\!\!\perp \{\mathbf{X}_0, \mathbf{Y}\}$  the sequence  $\mathbf{X}_1$  acts like a “one-time-pad” when added to  $\mathbf{Y}$ , meaning:
  - $(\mathbf{X}_1 \oplus \mathbf{Y}) \perp\!\!\!\perp \{\mathbf{X}_0, \mathbf{Y}\}$
  - $\mathbf{X}_1 \oplus \mathbf{Y}$  is uniformly distributed, i.e., it is i.i.d.  $\text{Bern}(0.5)$

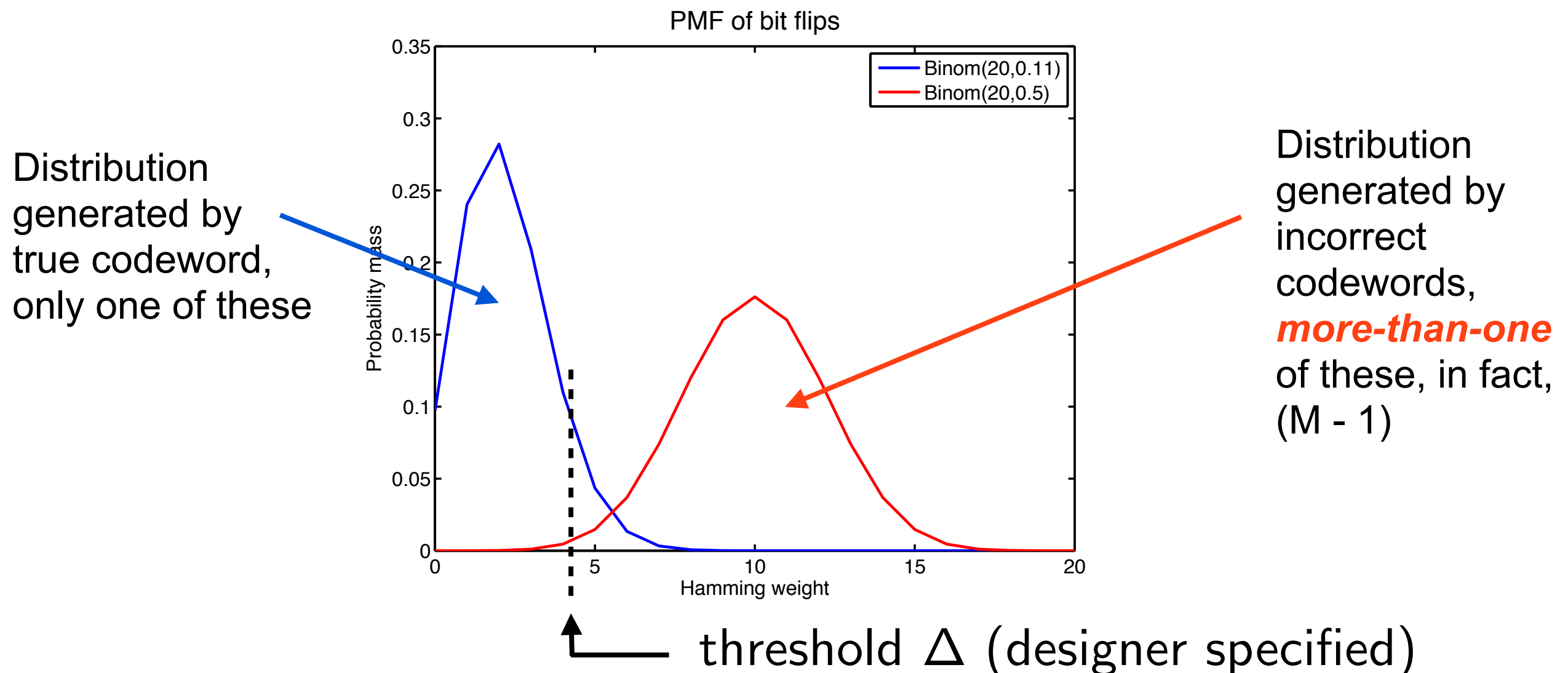
# Plot the distributions in play for (small) length: $n=20$

Simple expression for average probability of error:

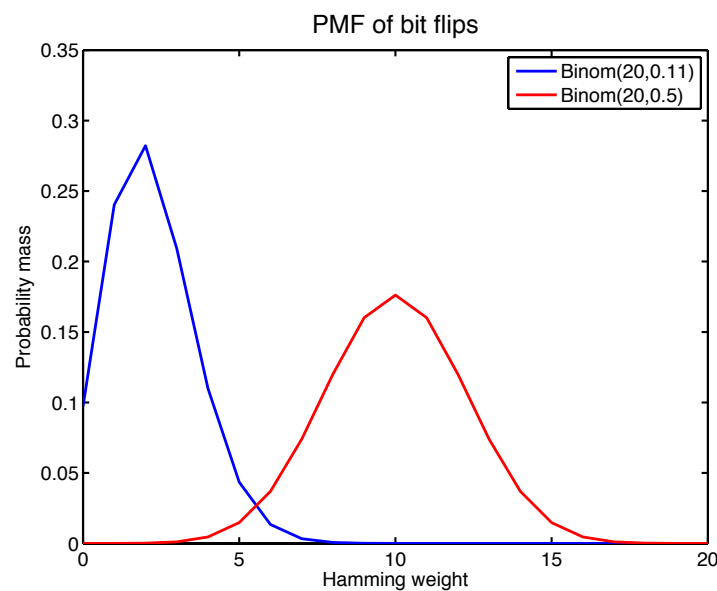
$$\Pr[\text{error}] \leq \Pr[\text{wt}_H(\tilde{\mathbf{X}}_0) \geq \Delta] + (M - 1) \Pr[\text{wt}_H(\tilde{\mathbf{X}}_1) \leq \Delta]$$

where we define the re-centered codewords as

- $\tilde{\mathbf{X}}_0 = \mathbf{X}_0 \oplus \mathbf{Y}$  and  $\tilde{\mathbf{X}}_1 = \mathbf{X}_1 \oplus \mathbf{Y}$ , so
- $\tilde{\mathbf{X}}_0 \sim \text{i.i.d. Bern}(p)$ ,  $\tilde{\mathbf{X}}_1 \sim \text{i.i.d. Bern}(0.5)$ , and  $\tilde{\mathbf{X}}_0 \perp\!\!\!\perp \tilde{\mathbf{X}}_1$
- So,  $\text{wt}_H(\tilde{\mathbf{X}}_0) \sim \text{Binom}(n, p)$  and  $\text{wt}_H(\tilde{\mathbf{X}}_1) \sim \text{Binom}(n, 0.5)$



# Replot as cumulative distribution functions (CDFs)



But, what we really care about is that  $\text{wt}_H(\tilde{\mathbf{X}}_0)$  is **too big**, or  $\text{wt}_H(\tilde{\mathbf{X}}_1)$  is **too small**. So, more useful to examine CDFs, rather than PMFs.

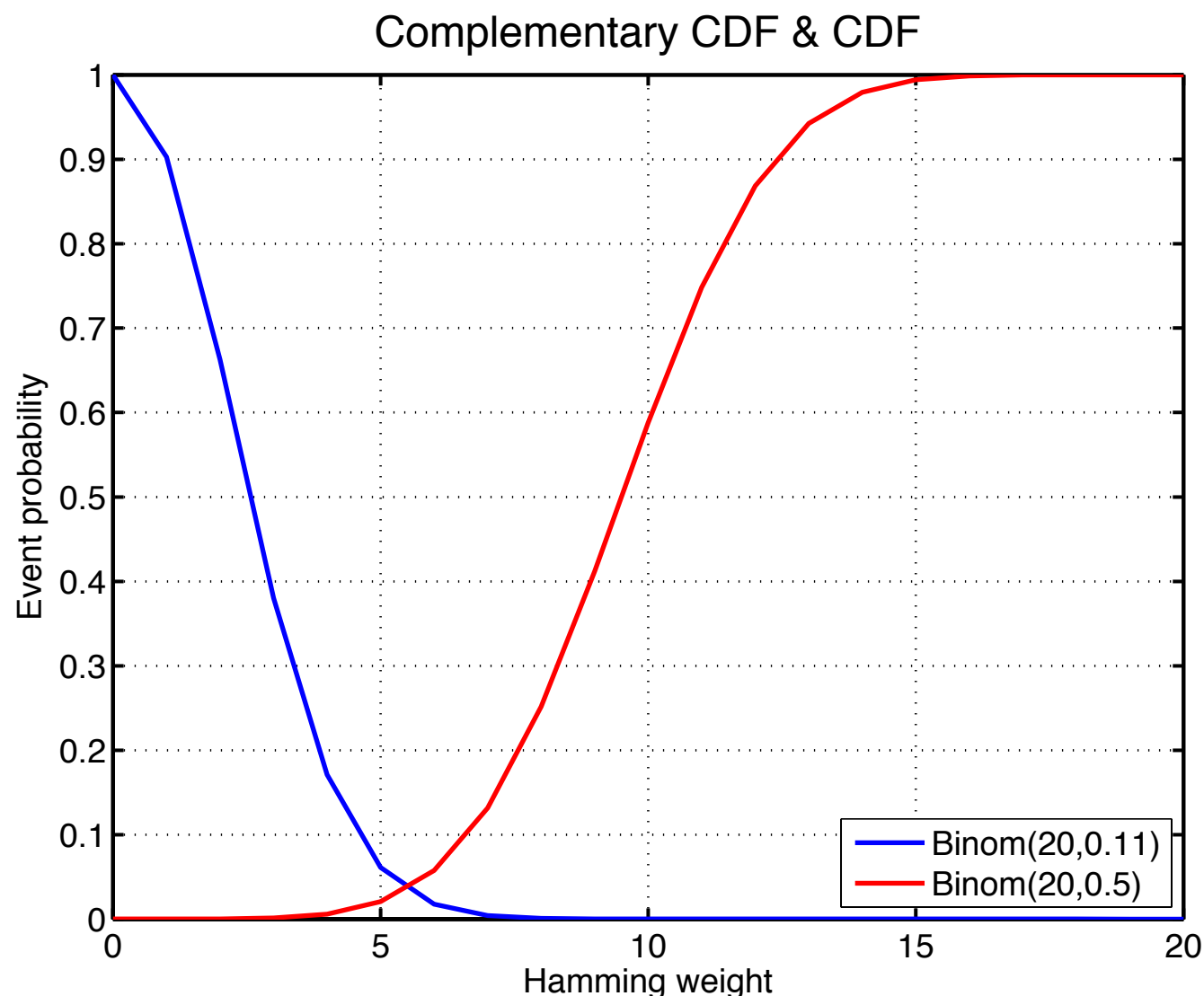
Complementary CDF of  $\text{wt}_H(\tilde{\mathbf{X}}_0)$ :

$$\Pr[\text{wt}_H(\tilde{\mathbf{X}}_0) \geq \Delta] = \sum_{t=\Delta}^n \Pr[\text{wt}_H(\tilde{\mathbf{X}}_0) = t]$$

CDF of  $\text{wt}_H(\tilde{\mathbf{X}}_1)$ :

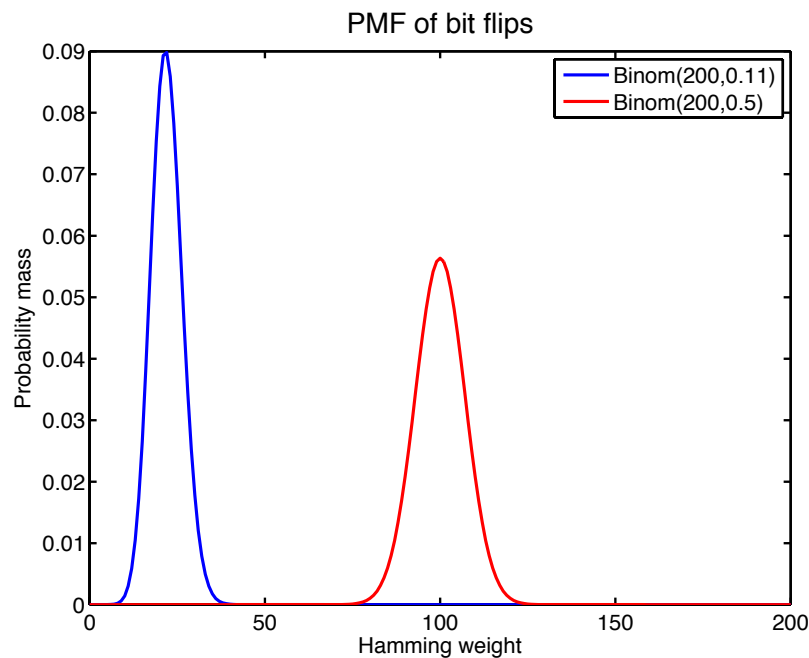
$$\Pr[\text{wt}_H(\tilde{\mathbf{X}}_1) \leq \Delta] = \sum_{t=0}^{\Delta} \Pr[\text{wt}_H(\tilde{\mathbf{X}}_1) = t]$$

**Tail behavior will be a topic of central importance today!**





# Let's think about an interesting block-length: $n=200$



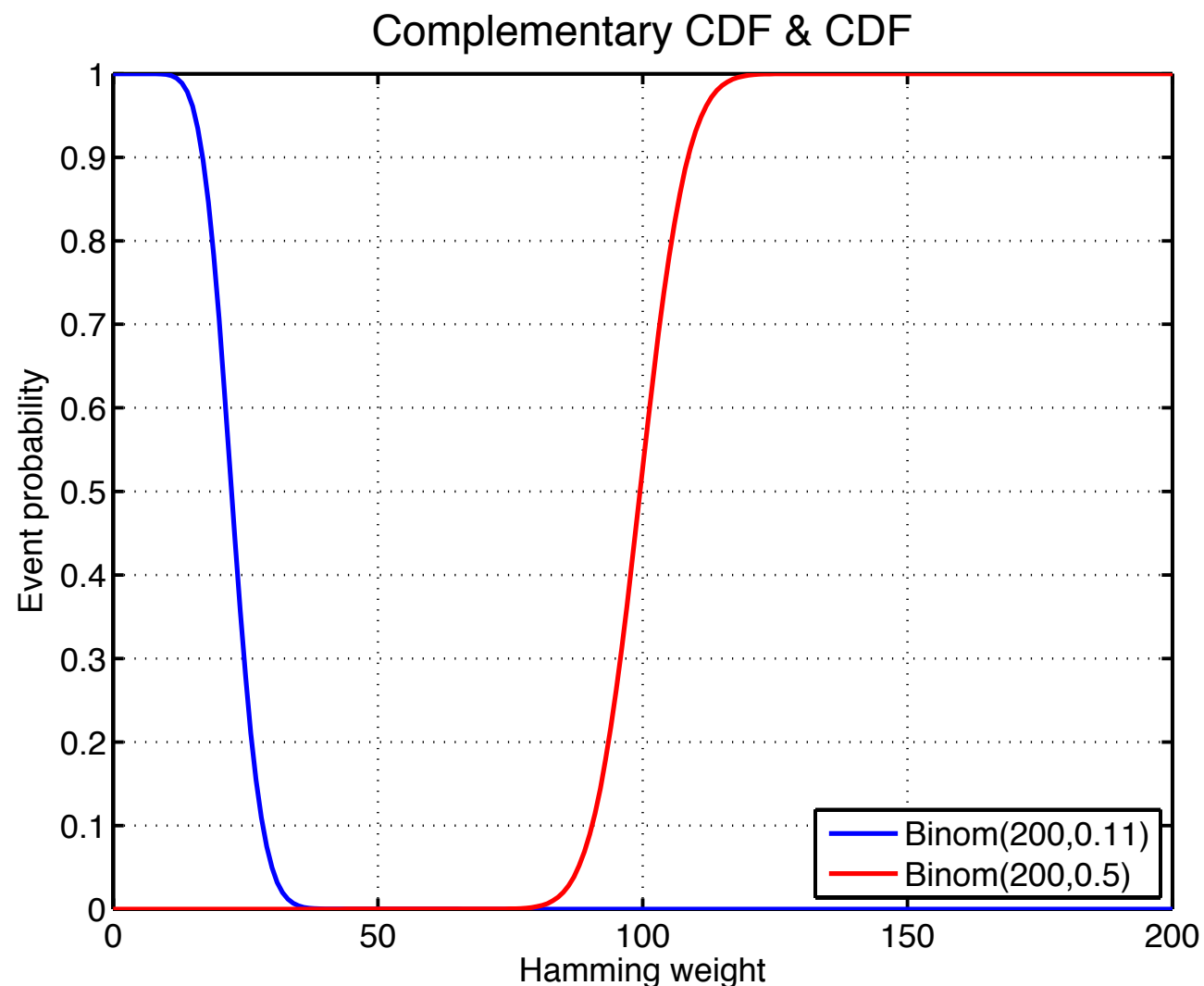
But, what we really care about is that  $\text{wt}_H(\tilde{\mathbf{X}}_0)$  is **too big**, or  $\text{wt}_H(\tilde{\mathbf{X}}_1)$  is **too small**. So, more useful to examine CDFs, rather than PMFs.

Complementary CDF of  $\text{wt}_H(\tilde{\mathbf{X}}_0)$ :

$$\Pr[\text{wt}_H(\tilde{\mathbf{X}}_0) \geq \Delta] = \sum_{t=\Delta}^n \Pr[\text{wt}_H(\tilde{\mathbf{X}}_0) = t]$$

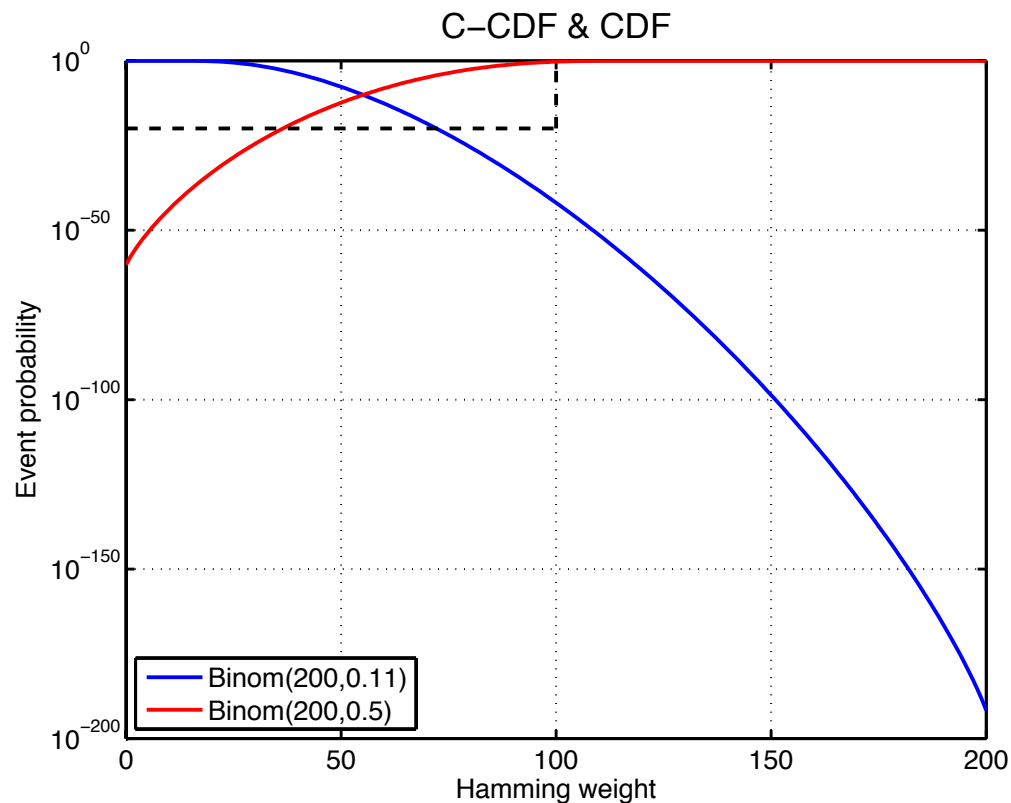
CDF of  $\text{wt}_H(\tilde{\mathbf{X}}_1)$ :

$$\Pr[\text{wt}_H(\tilde{\mathbf{X}}_1) \leq \Delta] = \sum_{t=0}^{\Delta} \Pr[\text{wt}_H(\tilde{\mathbf{X}}_1) = t]$$

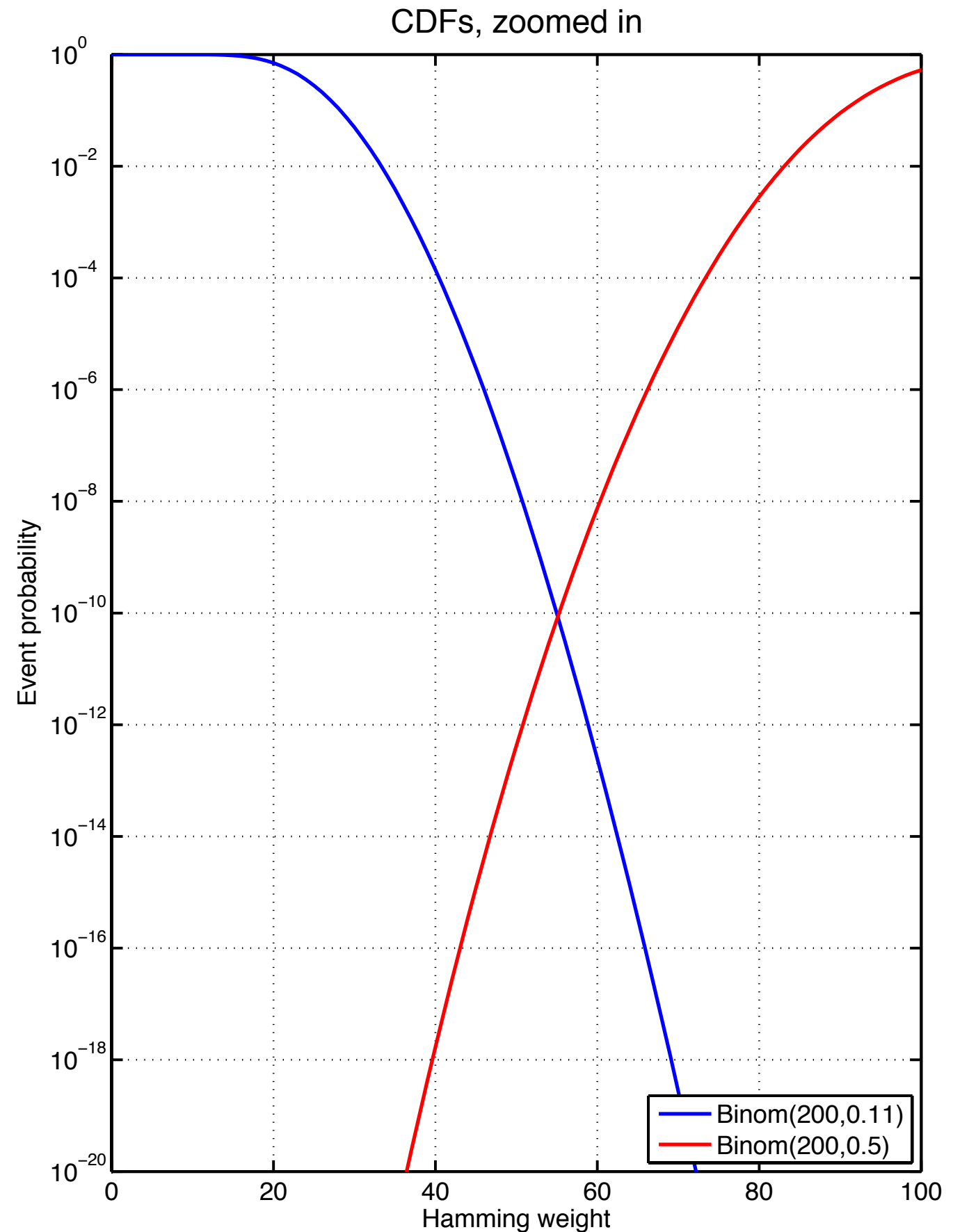


But, it's hard to see what's going on in the tails from this plot

# Plot on log-scale to see better



zoom in on upper left



**Question:** if you  
(i) use this decoder, and  
(ii) target an error rate of  $10^{-3}$ , then

**What code rate is possible?**

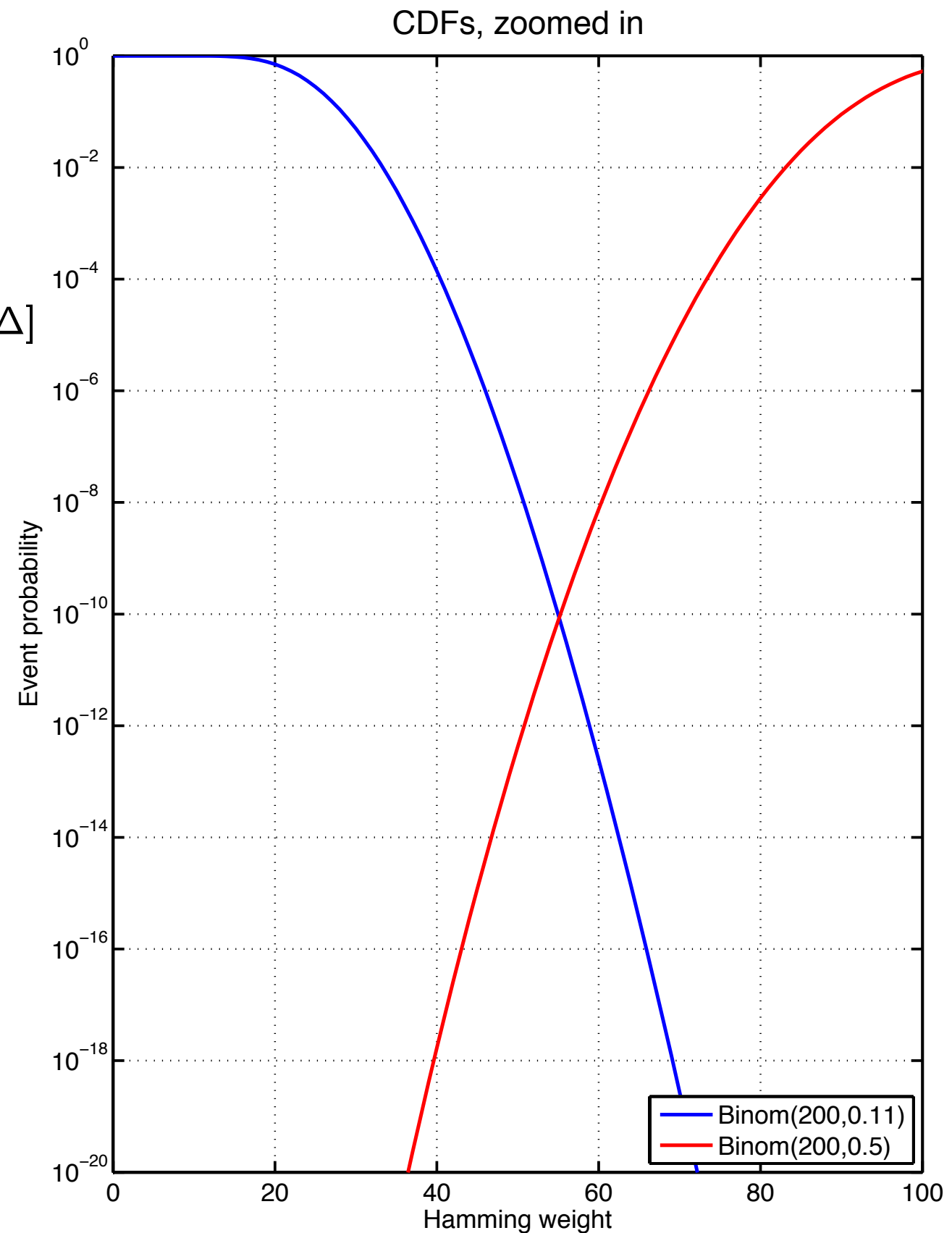
# Exercise: (roughly) what rate is possible at $\Pr[\text{err}] = 10^{-3}$

Necessary data:

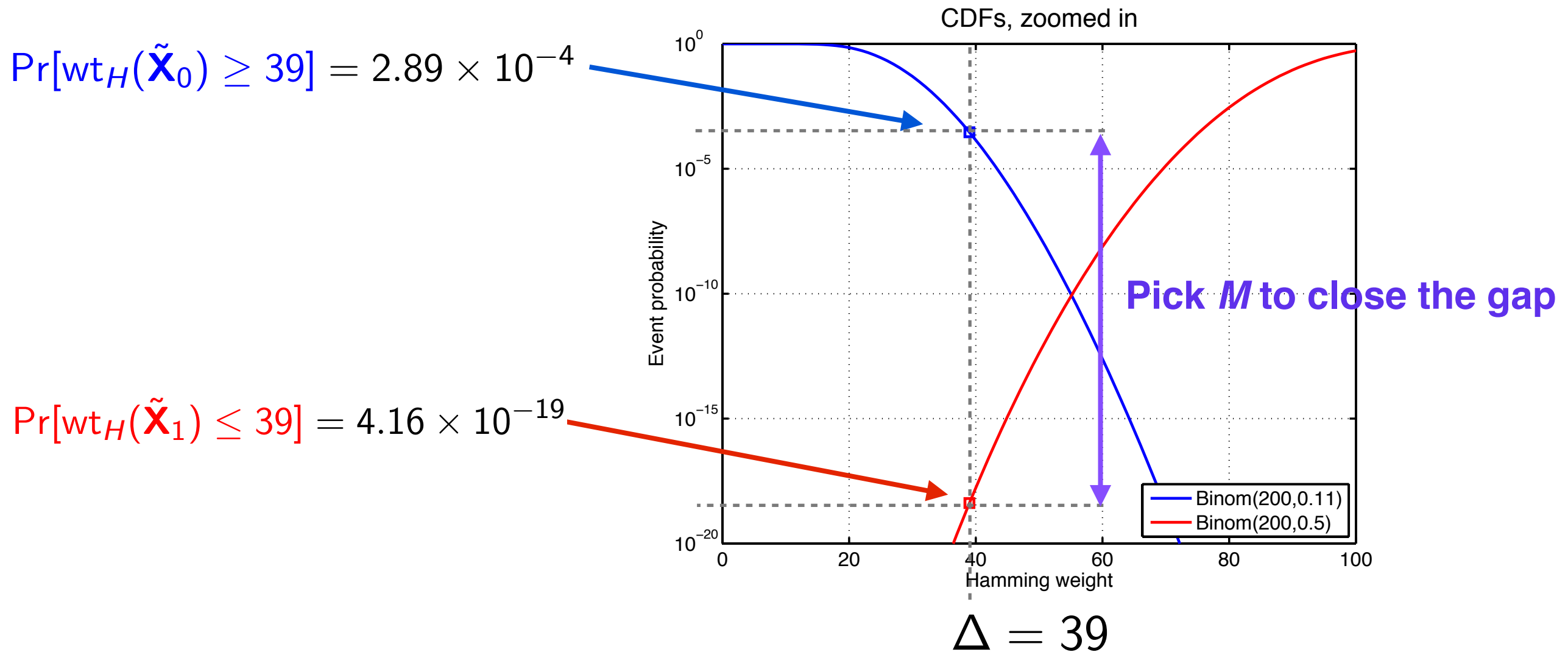
$$\Pr[\text{error}] \leq \Pr[\text{wt}_H(\tilde{\mathbf{X}}_0) \geq \Delta] + (M - 1) \Pr[\text{wt}_H(\tilde{\mathbf{X}}_1) \leq \Delta]$$

$$\Pr[\text{wt}_H(\tilde{\mathbf{X}}_0) \geq \Delta] = \sum_{t=\Delta}^n \Pr[\text{wt}_H(\tilde{\mathbf{X}}_0) = t]$$

$$\Pr[\text{wt}_H(\tilde{\mathbf{X}}_1) \leq \Delta] = \sum_{t=0}^{\Delta} \Pr[\text{wt}_H(\tilde{\mathbf{X}}_1) = t]$$



# Calculate achievable rate at target $\Pr[\text{err}] = 10^{-3}$



(One way) to solve for achievable  $M$ :

$$\Pr[\text{wt}_H(\tilde{\mathbf{X}}_0) \geq 39] + (M - 1) \Pr[\text{wt}_H(\tilde{\mathbf{X}}_1) \leq 39] = 10^{-3}$$

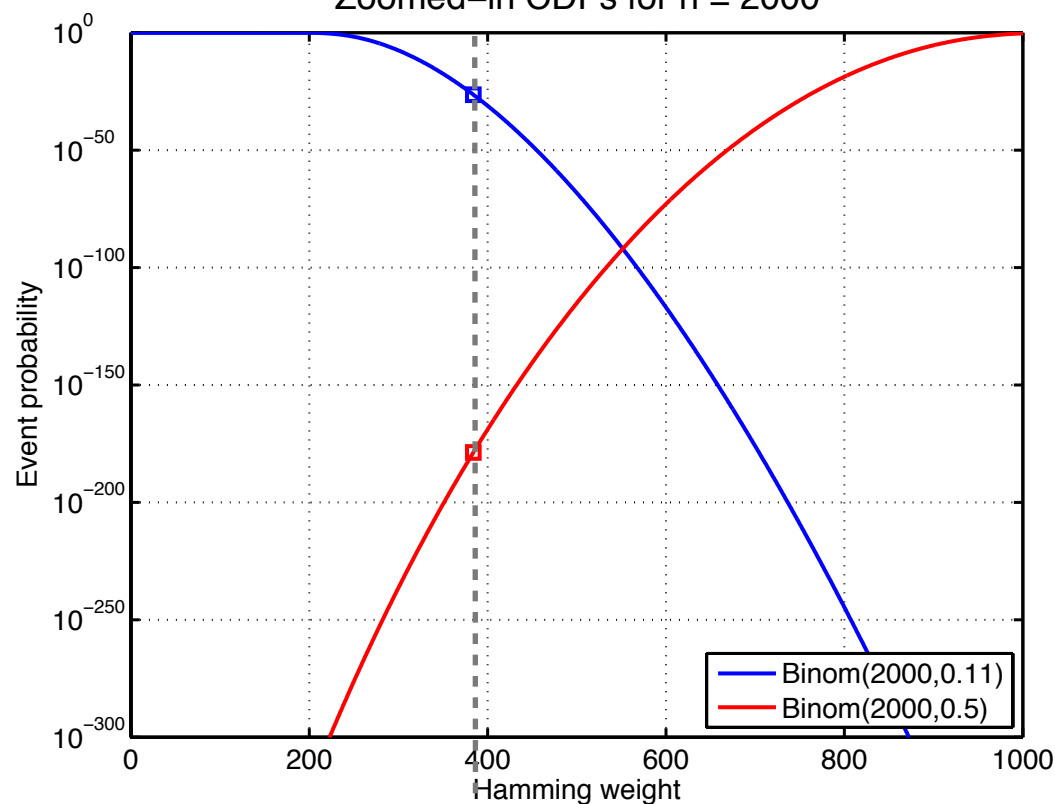
$$R = \frac{\log_2 M}{n} = \frac{\log_2 1.71 \times 10^{15}}{200} = 0.253 \frac{\text{bits}}{\text{channel use}}$$

# But this is Shannon theory, need something to get big

Let block length  $n$  get large, but what do we keep fixed? Rate?  $\Pr[\text{error}]$ ?  
The choice leads to two regimes of study:

Rate fixed = 0.253,  
 $n = 2000$

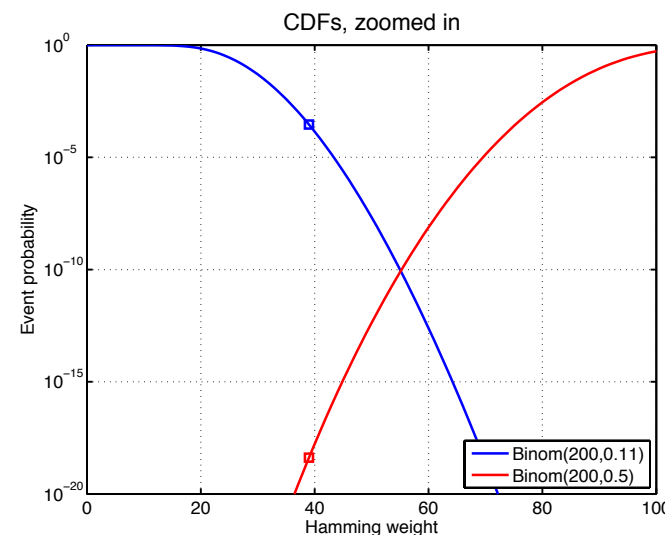
Zoomed-in CDFs for  $n = 2000$



$$\Delta = 384$$

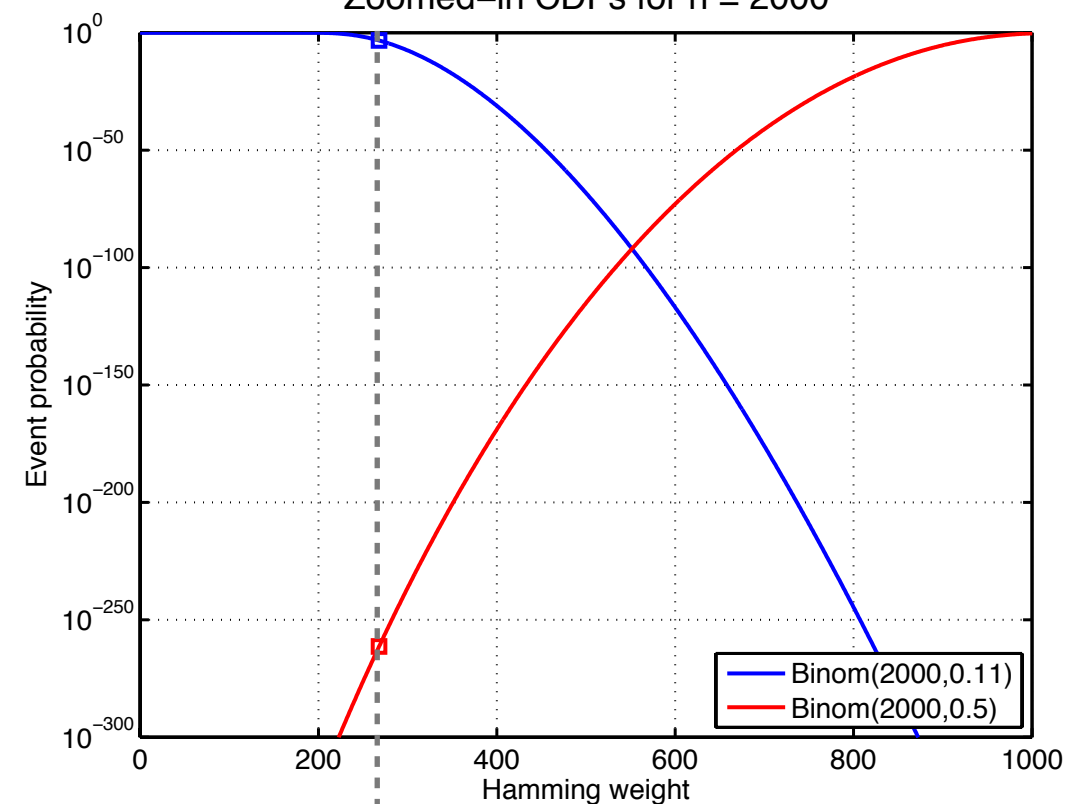
$$\text{rate} = 0.253, M = 2^{2000 \times 0.252} = 2^{506}$$

$$\Pr[\text{err}] = 7.9 \times 10^{-27} = 2^{-2000 \times 0.0434}$$



$\Pr[\text{error}]$  fixed =  $10^{-3}$ ,  
 $n = 2000$

Zoomed-in CDFs for  $n = 2000$



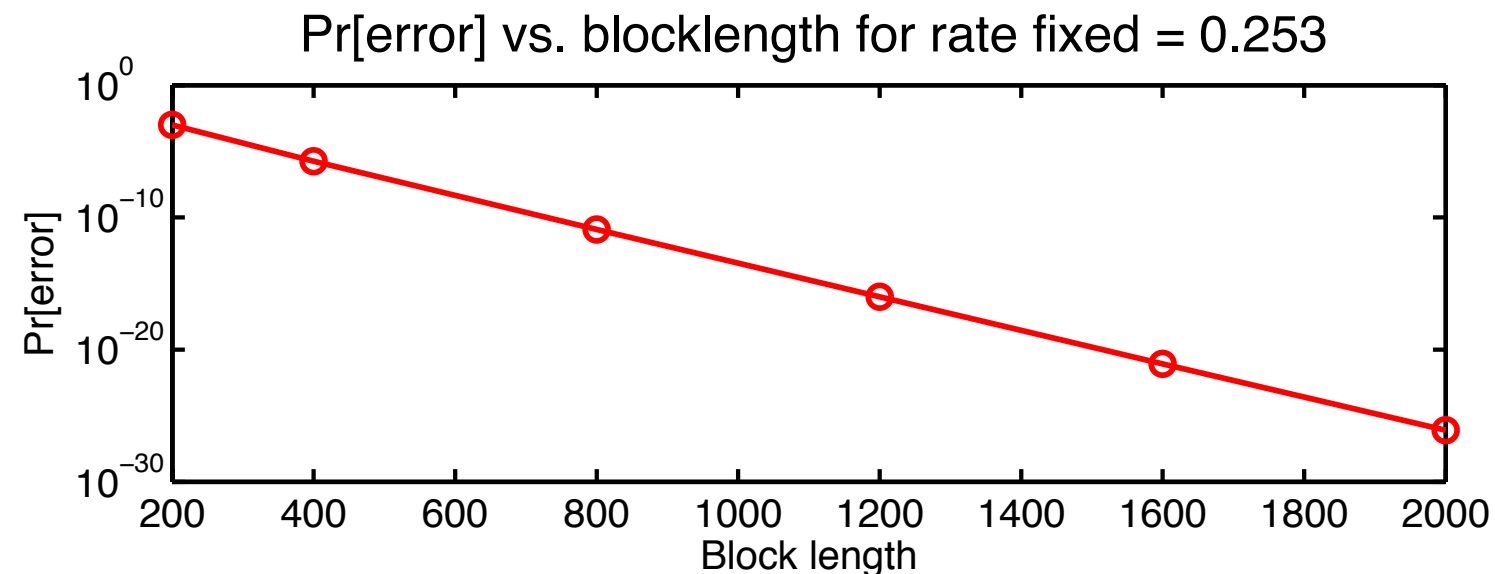
$$\Delta = 268$$

$$\text{rate} = 0.4288, M = 2^{2000 \times 0.4288} = 2^{858}$$

$$\Pr[\text{err}] = 10^{-3} = 2^{-2000 \times 0.0015}$$

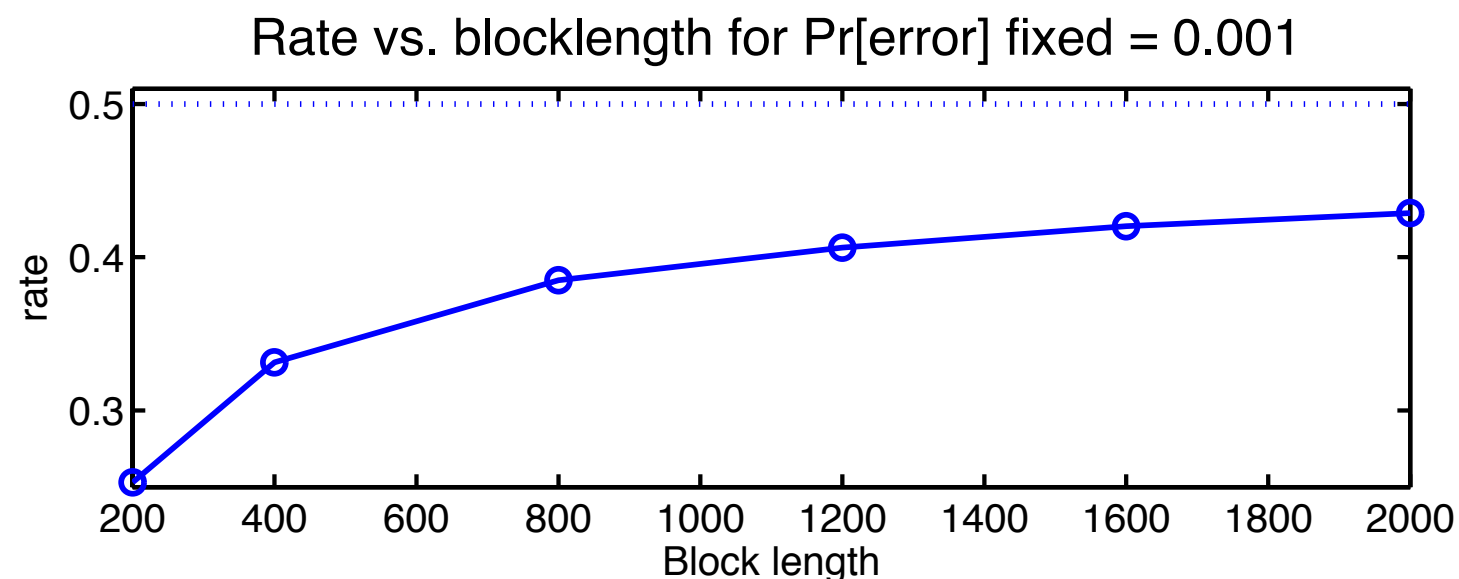
# Remainder of talk: two chunks

## Chunk 1: Error exponent analysis of ML decoding



Slope (magnitude) of error decay on a log plot is the “error exponent”. Here it is about 0.0434. Can it be improved?

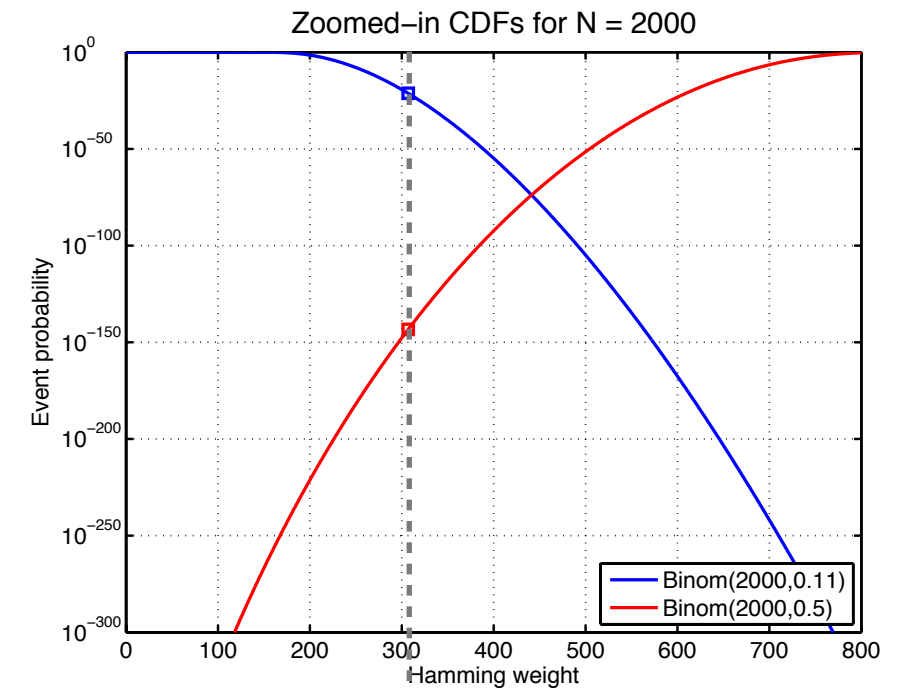
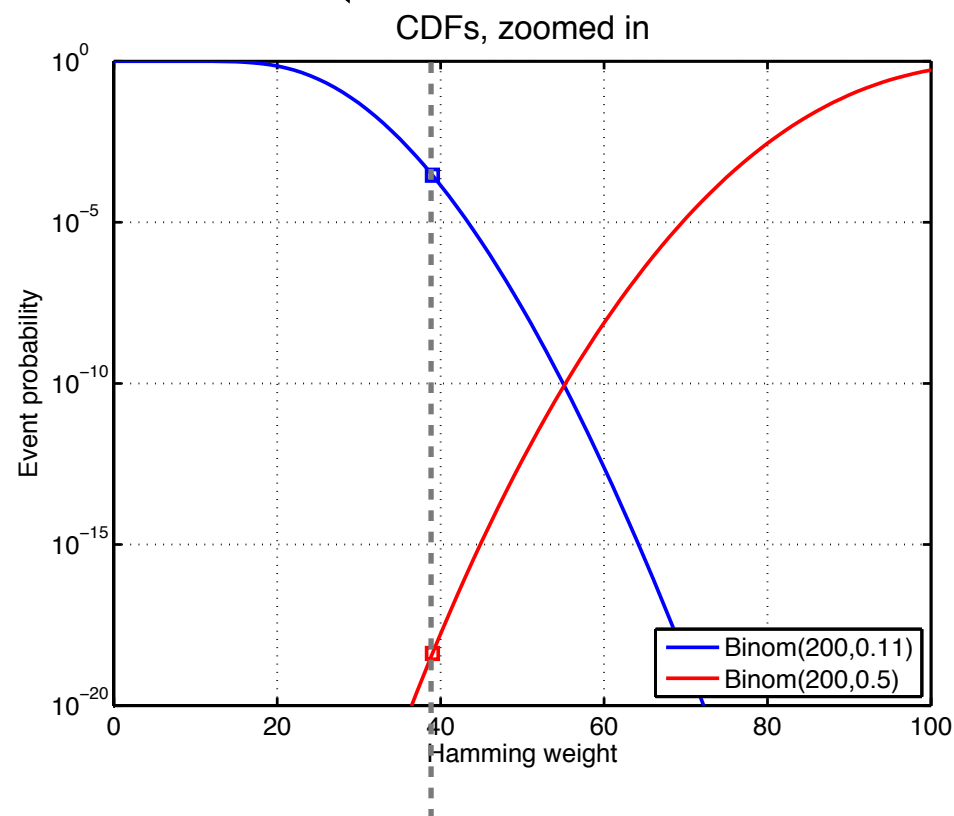
## Chunk 2: How fast approach capacity using ML decoding



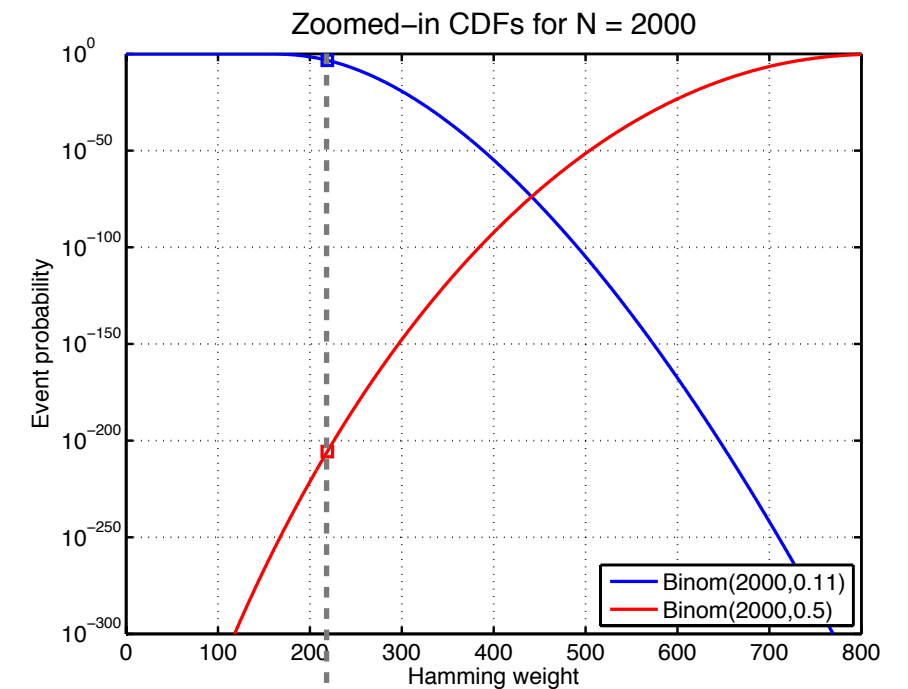
Asymptotes to capacity as  $n$  increases. But, how long to get close? How does the rate approach capacity? Can you approach faster than in plot to left?

# Key difference in analyses: how threshold increases with block length

Recall block length of:  
 $n = 200$  vs.  $n = 2000$



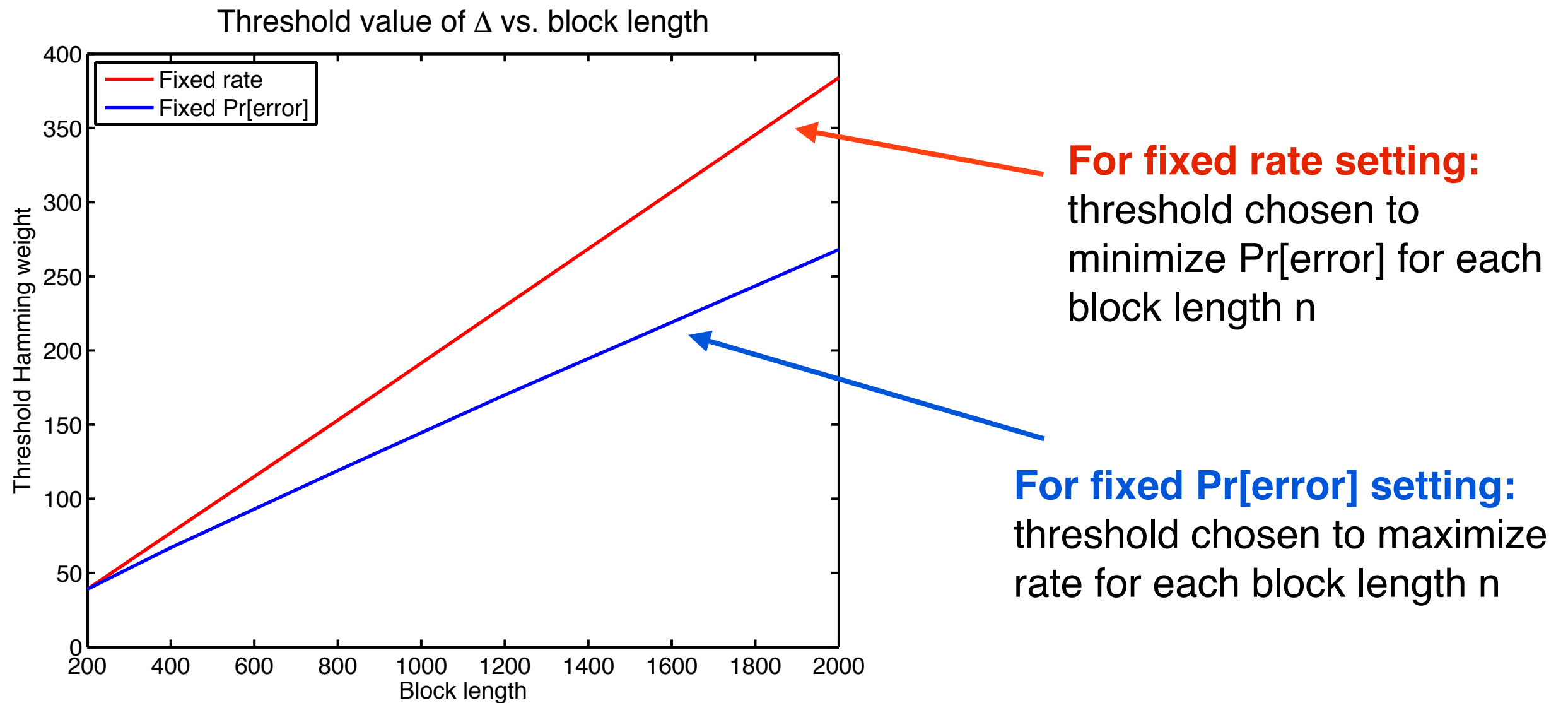
$$\Delta = 384$$



$$\Delta = 268$$

How does critical threshold value scale with block length for each objective (fixed rate or fixed error)?

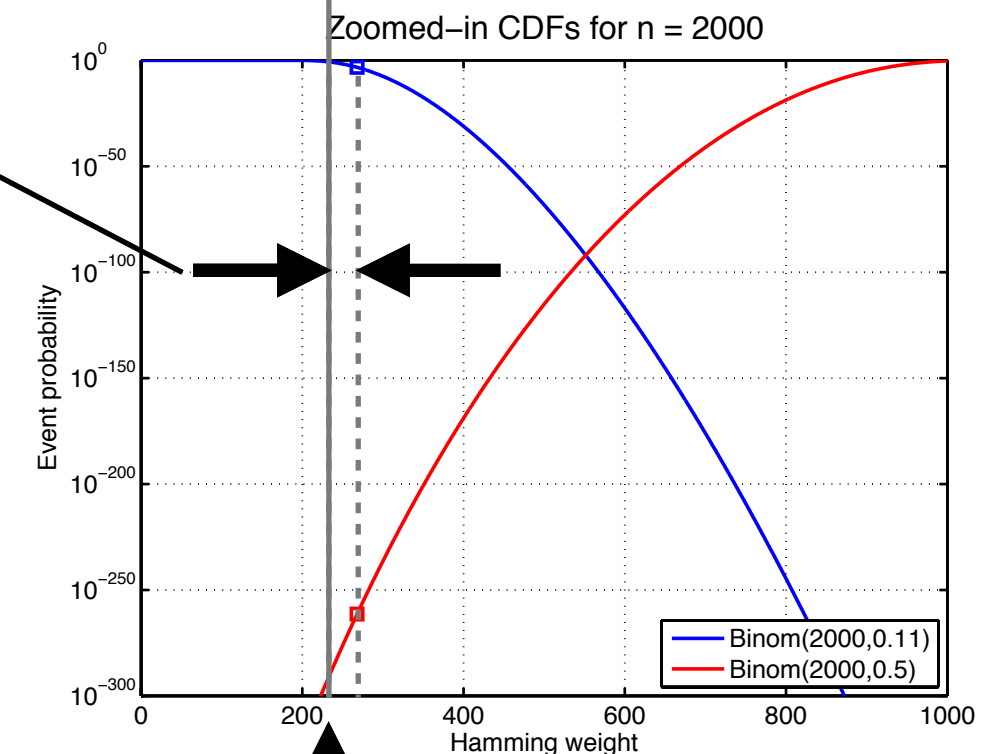
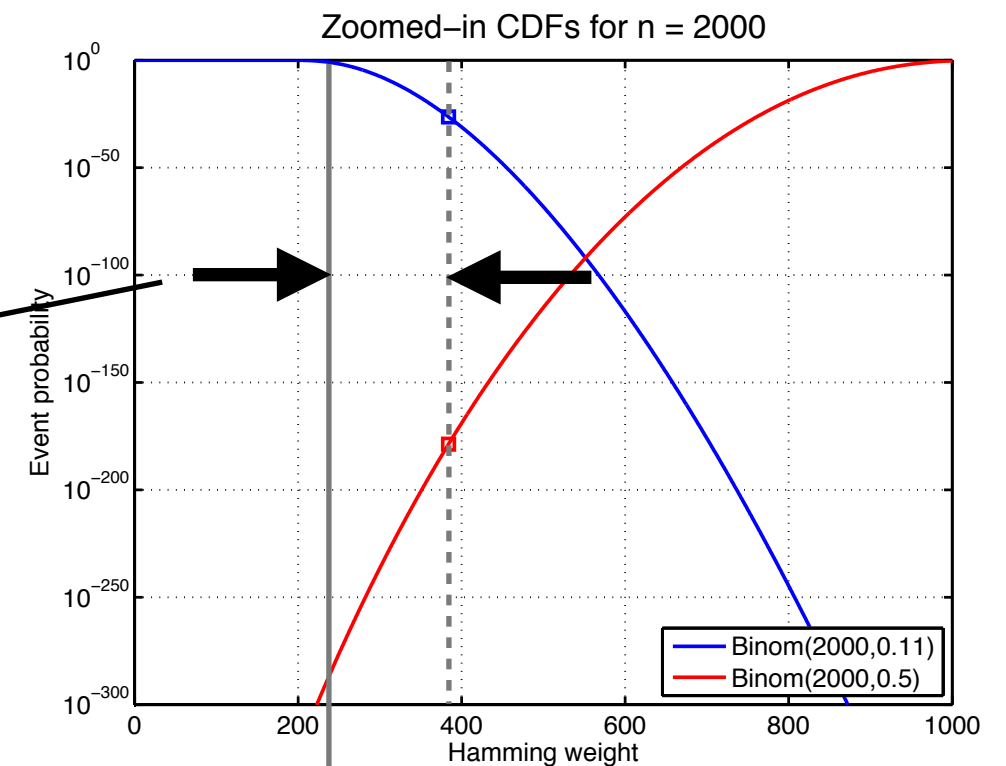
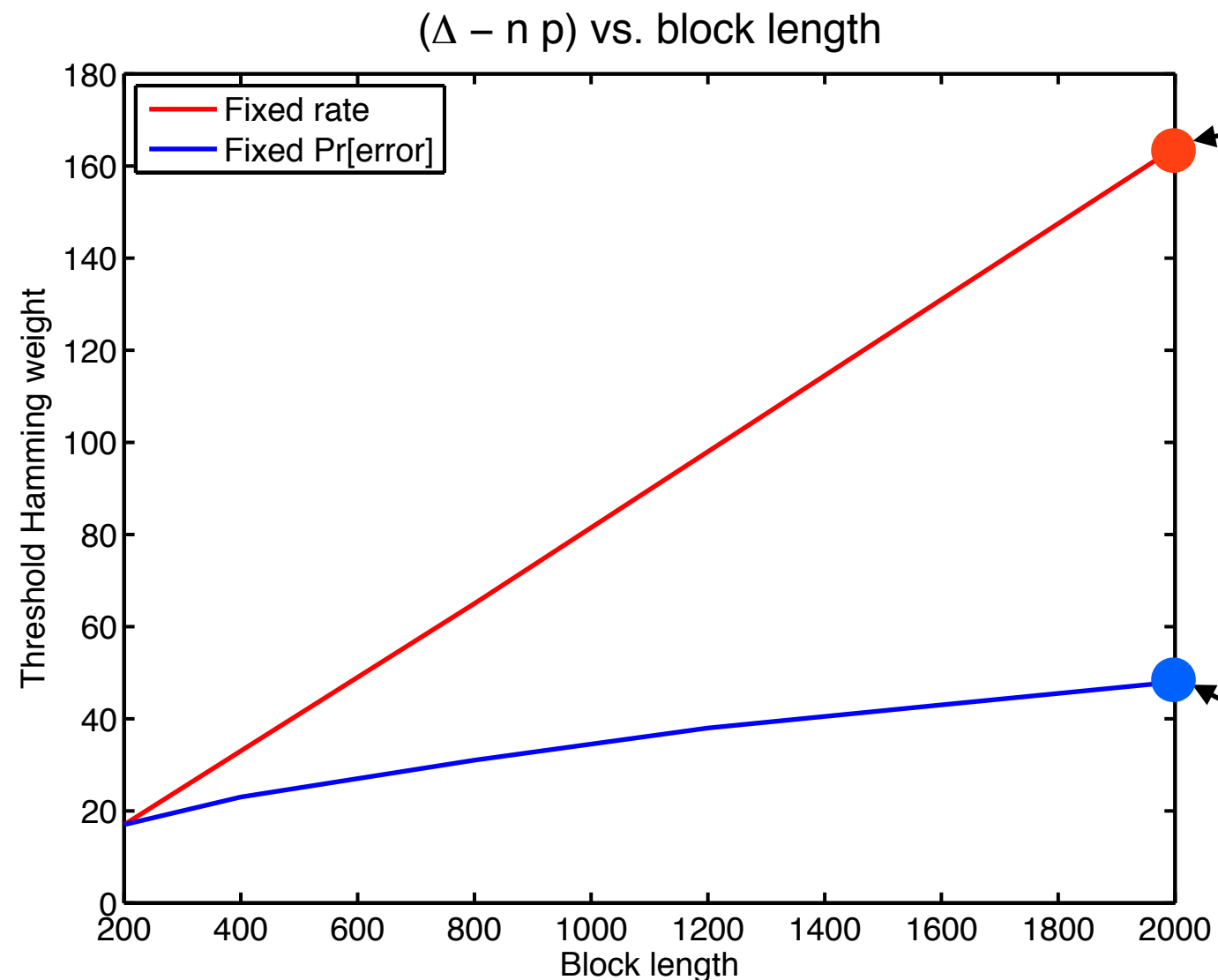
# Scaling of threshold with block length for each setting



Note: in **both** settings the threshold needs to be larger than the mean number of bit flips ( $= n p$ ), so let's subtract the mean from each and re-plot



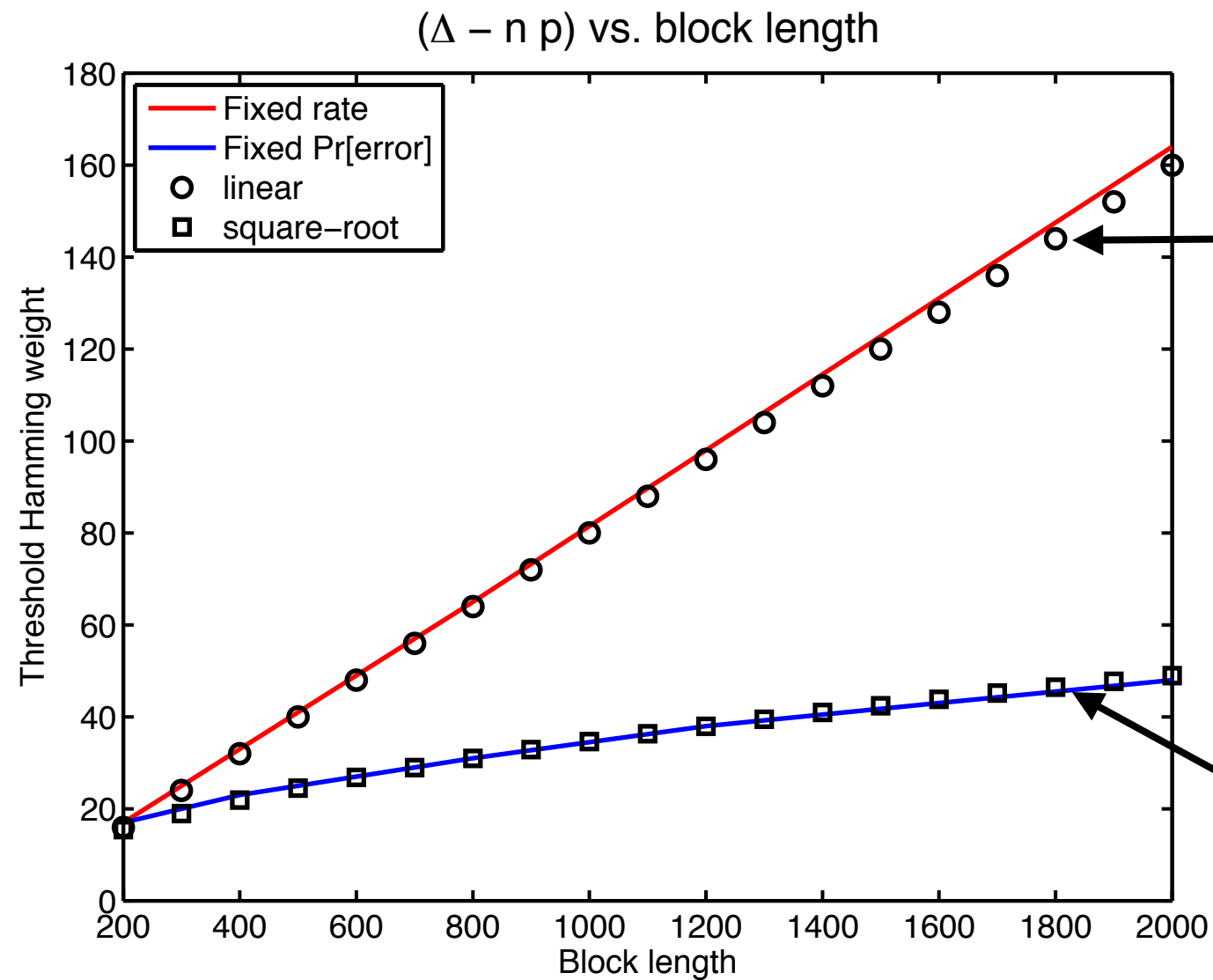
# Scaling of “excess” Hamming weight with block length



“**Excess**” weight is how far above the mean number of flips we set the threshold

mean Hamming weight =  $0.11 \times 2000 = 220$

# “Excess” Hamming weight: linear vs. square-root



**For fixed-rate setting:**

excess rate increases linearly in block length

$$\beta n$$

**For fixed Pr[error] setting:**

excess rate increases as square root of block length

$$\beta' \sqrt{n}$$

# Recap: what we have learned

Analyzed error in random code ensemble for BSC under a simple (and suboptimal) bounded-distance decoding

Observed behavior of tail of distribution is quite important

Characterized behavior of this decoder

Tradeoff between

- “outage”/ “atypicality” events of correct codeword, and
- “confusion” / “union bound” events of other (candidate, but incorrect) codewords

Two interesting regimes

- Fix rate, see how  $\Pr[\text{err}]$  drops with block length
- Fix  $\Pr[\text{err}]$ , see how rate increases with block length

# Also: will return to analyzing the ML decoder

To get a better grip want to analyze the (optimal) ML decoder. Why is bounded information (bounded distance for BSC) decoder suboptimal?

- Misses coupling between “outage” and “confusion” events
- Really relative events are what are important in ML
- E.g., you might be lucky,  $\text{wt}_H(\mathbf{X}_1, \mathbf{Y})$  might be much less than  $\Delta$ , making it harder for an incorrect codeword to be closer to  $\mathbf{Y}$

ML decoding

$$\Pr[\text{error}] \leq \Pr \left[ \bigcup_{m=1}^{M-1} i(\mathbf{X}_m, \mathbf{Y}) \geq i(\mathbf{X}_0, \mathbf{Y}) \right]$$

Bounded information decoding

$$\Pr[\text{error}] \leq \Pr \left[ (i(\mathbf{X}_0, \mathbf{Y}) < \gamma) \bigcup \left( \bigcup_{j=1}^{M-1} i(\mathbf{X}_j, \mathbf{Y}) \geq \gamma \right) \right]$$

$$\leq \Pr[i(\mathbf{X}_0, \mathbf{Y}) \leq \gamma] + (M - 1) \Pr[i(\mathbf{X}_1, \mathbf{Y}) \geq \gamma]$$

union removes coupling

# Agenda

Analyzing decoding error for a bounded information decoder: regimes of interest

## **Error exponents of ML decoders**

Non-asymptotic analysis of ML decoder & Normal approximation

# Refine the analysis of ML decoding

ML error:  $\Pr[\text{error}] \leq \Pr \left[ \bigcup_{m=1}^{M-1} i(\mathbf{X}_m, \mathbf{Y}) \geq i(\mathbf{X}_0, \mathbf{Y}) \right]$

Particularize to BSC(p):  $i(\mathbf{X}, \mathbf{Y}) = d_H(\mathbf{x}, \mathbf{y}) \log \frac{p}{1-p} + n \log \frac{1-p}{2}$

Substituting and “re-centering” we get:

$$\Pr[\text{error}] \leq \Pr \left[ \bigcup_{m=1}^{M-1} d_H(\mathbf{X}_m, \mathbf{Y}) \leq d_H(\mathbf{X}_0, \mathbf{Y}) \right]$$

$$= \Pr \left[ \bigcup_{m=1}^{M-1} \text{wt}_H(\tilde{\mathbf{X}}_m) \leq \text{wt}_H(\tilde{\mathbf{X}}_0) \right]$$

codewords “re-centered” about observation

**Recall:** recentered c.w. statistically indep.! (one-time-pad)

# Define a coupled event

Recall idea of bounded distance decoder: observation “too far” from true c.w. **or** “too close” to some other c.w.

Now we couple these: (i) “too far” from true codeword, **and** (ii) “too close” to some other codeword

two events  
of interest

$$\mathcal{A}_{n,\delta} := \{ \text{wt}_H(\tilde{\mathbf{X}}_0) \geq n\delta \}$$

$$\mathcal{B}_{n,\delta} := \{ \cup_{m=1}^{M-1} \text{wt}_H(\tilde{\mathbf{X}}_m) \leq n\delta \}$$

where  $\delta \in \left\{ 0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n}{n} \right\}$

fractional Hamming weight

coupled event

$$\mathcal{E}_{n,\delta} = \mathcal{A}_{n,\delta} \cap \mathcal{B}_{n,\delta}$$

# Analyze the coupled event

$$\begin{aligned}\Pr[\text{error}] &\leq \Pr \left[ \bigcup_{\delta \in \{0, \frac{1}{n}, \dots, 1\}} \mathcal{E}_{n,\delta} \right] \\ &= \sum_{\delta \in \{0, \frac{1}{n}, \dots, \frac{n}{n}\}} \Pr [\mathcal{E}_{n,\delta}] \\ &= \sum_{\delta \in \{0, \frac{1}{n}, \dots, \frac{n}{n}\}} \Pr [\mathcal{A}_{n,\delta} \cap \mathcal{B}_{n,\delta}] \\ &= \sum_{\delta \in \{0, \frac{1}{n}, \dots, \frac{n}{n}\}} \Pr [\mathcal{A}_{n,\delta}] \Pr [\mathcal{B}_{n,\delta}] \\ &\leq (n+1) \max_{\delta \in [0,1]} \Pr [\mathcal{A}_{n,\delta}] \Pr [\mathcal{B}_{n,\delta}]\end{aligned}$$

next: bound each (decoupled) probability





# Chernoff tail bounds

## Theorem

Let  $X_1, \dots, X_n \sim \text{i.i.d. Bern}(q)$ , then for any threshold  $1 \geq \tau \geq q$

$$\Pr[\text{wt}_H(\mathbf{X}) \geq n\tau] \doteq 2^{-nD(\tau\|q)}$$

and, if  $0 \leq \tau \leq q$ ,

$$\Pr[\text{wt}_H(\mathbf{X}) \leq n\tau] \doteq 2^{-nD(\tau\|q)}$$

where  $D(\tau\|q)$  is the (binary) KL divergence

$$D(\tau\|q) = \tau \log_2 \frac{\tau}{q} + (1 - \tau) \log_2 \frac{1 - \tau}{1 - q}$$

expected  
fraction ones

realized  
fraction ones

Where  $a_n \doteq b_n$  means we get the exponent correct:

$$\lim_{n \rightarrow \infty} \log \frac{a_n}{b_n} = 0$$

Note: KL divergence convex & increasing in separation between  $\tau$  and  $q$

# Tail bounds for each event

$$\Pr[\mathcal{A}_{n,\delta}] = \Pr [\text{wt}_H(\tilde{\mathbf{X}}_0) \geq n\delta] \doteq 2^{-nD(\delta\|p)}$$

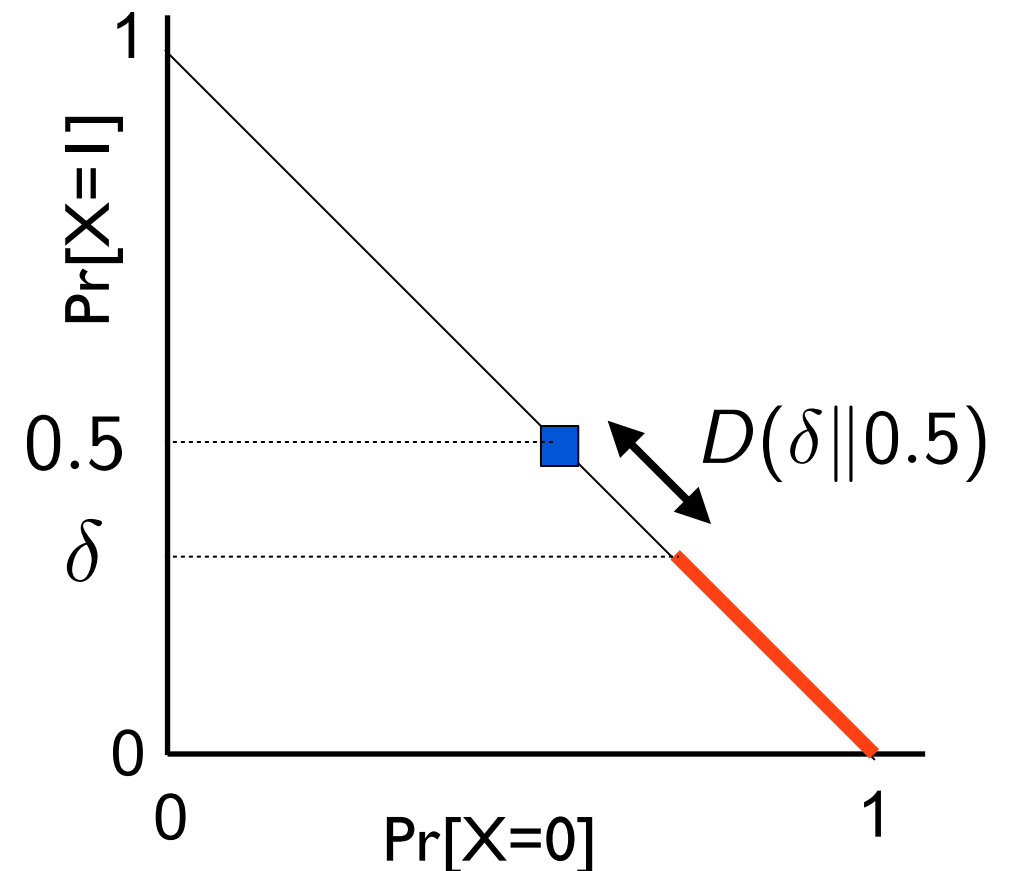
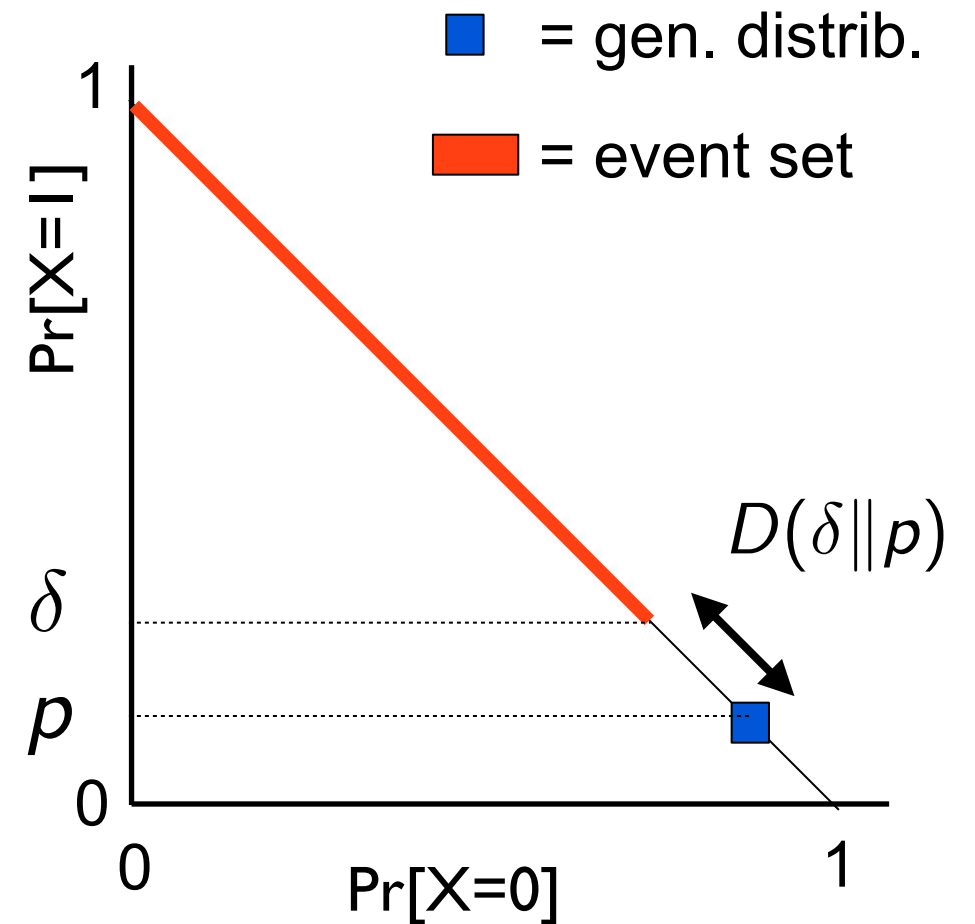
threshold

Generating distribution is i.i.d. Bern(p)

$$\begin{aligned} \Pr[\mathcal{B}_{n,\delta}] &= \Pr \left[ \bigcup_{m=1}^{M-1} \text{wt}_H(\tilde{\mathbf{X}}_m) \leq n\delta \right] \\ &\leq \min \left\{ 1, (M-1) \Pr [\text{wt}_H(\tilde{\mathbf{X}}_1) \leq n\delta] \right\} \\ &\doteq \min \{ 1, 2^{nR} 2^{-nD(\delta\|0.5)} \} \\ &= 2^{-n|D(\delta\|0.5)-R|^+} \end{aligned}$$

$|a|^+ = \max\{a, 0\}$

Generating distribution is now i.i.d. Bern(0.5)



# Combining and analyzing gives following result

## Theorem

*For ML decoding over the BSC the following “random coding” bound is achievable*

$$\Pr[\text{error}] \doteq 2^{-n} \left[ \min_{\delta \in [0,1]} D(\delta \| p) + |D(\delta \| 0.5) - R|^+ \right]$$

The  $\delta$  parameterized the likelihood of the two error events:

- The “atypical” probability that the noise level is too high
- The “confusion” probability that some spurious c.w. is too close

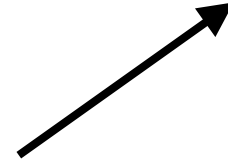
As you change  $\delta$  one term increases, the other decreases

The optimizing  $\delta$  strikes a balance between the two events

Proof: combine individual bounds in coupling event

# Analysis splits into high- and low-rate regimes

$$E_r(R) = \min_{\delta \in [0,1]} D(\delta \| p) + |D(\delta \| 0.5) - R|^+$$



Want to determine when positivity constraint is active. Define:

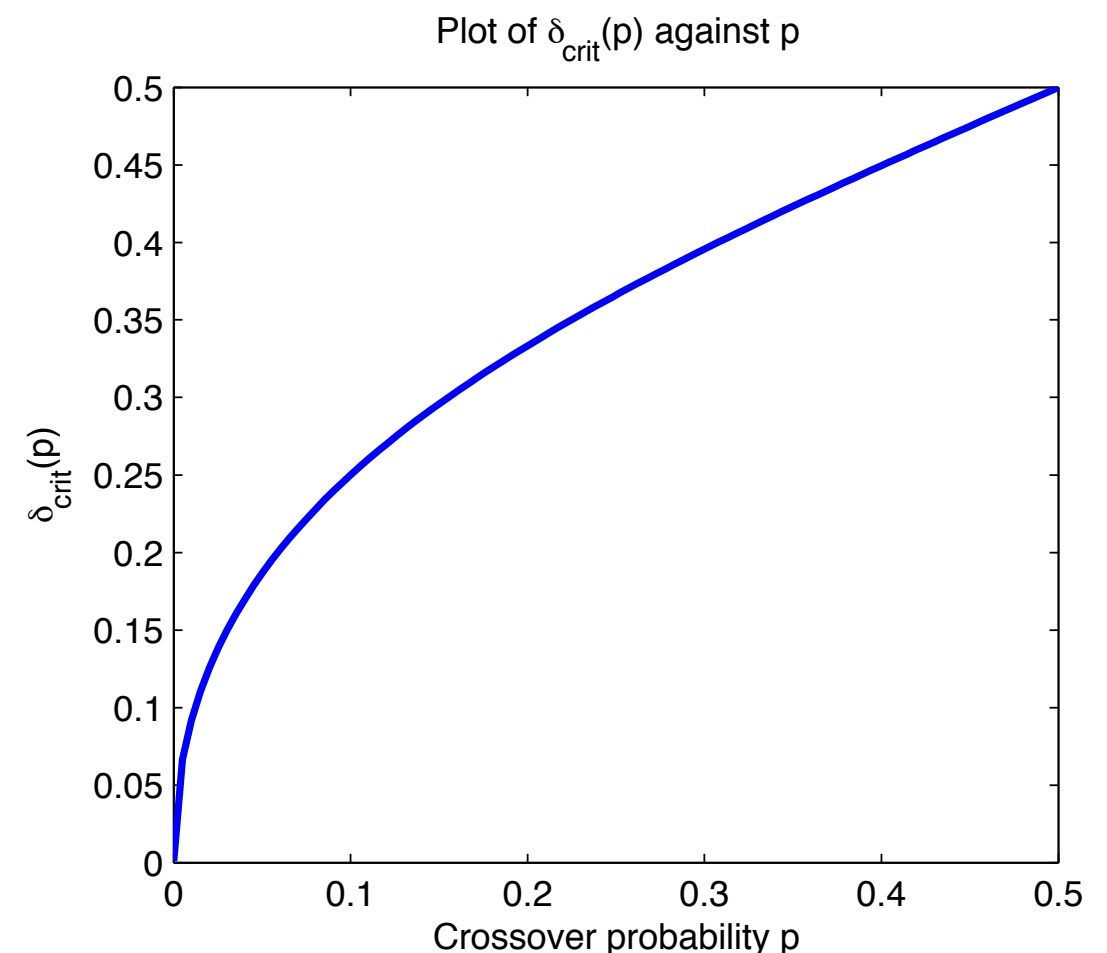
$$\delta_{crit} = \arg \min_{\delta \in [0,1]} D(\delta \| p) + D(\delta \| 0.5) - R$$

Differentiate and solve to find

$$\delta_{crit} = \frac{\sqrt{p}}{\sqrt{p} + \sqrt{1-p}}$$

Plotted on right.

Note that if  $p < 0.5$ , always above 45-degree line, why?

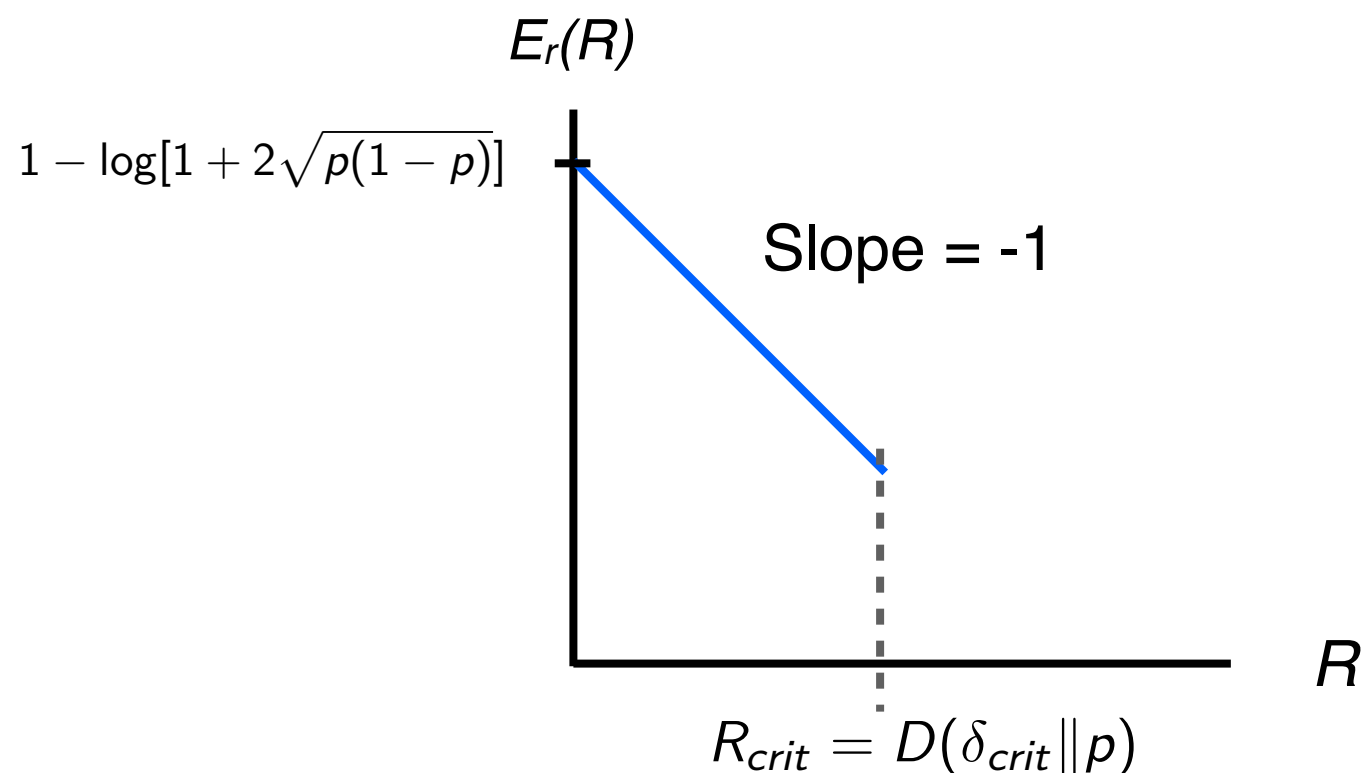


# Analysis of low rate regime

At sufficiently low rates, in particular if  $R < D(\delta_{crit}||0.5)$  then the  $|\cdot|^+$  is *not* active

This means that for these (low) rates:

$$\begin{aligned} E_r(R) &= \min_{\delta \in [0,1]} D(\delta||p) + |D(\delta||0.5) - R|^+ \\ &= \min_{\delta \in [0,1]} D(\delta||p) + D(\delta||0.5) - R \\ &= D(\delta_{crit}||p) + D(\delta_{crit}||0.5) - R \\ &= 1 - \log[1 + 2\sqrt{p(1-p)}] - R \end{aligned}$$




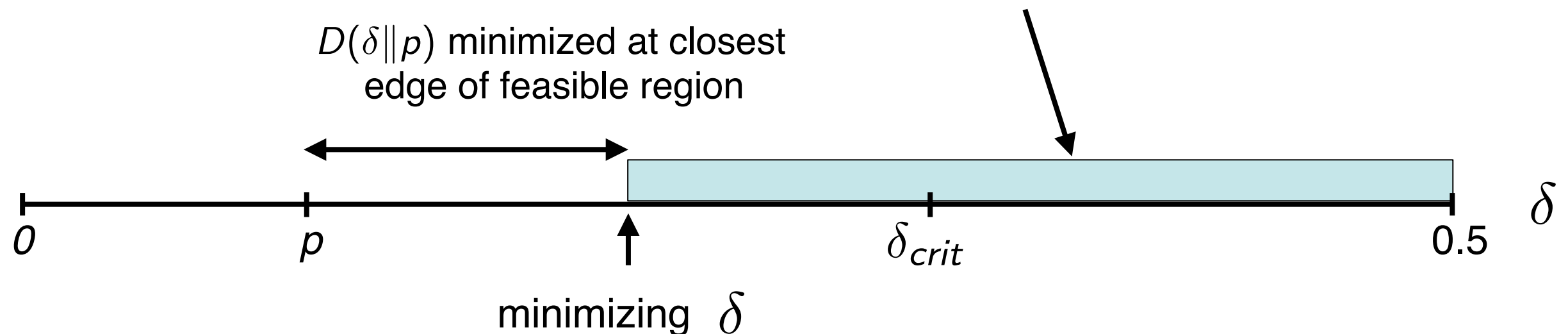
# Analysis of high rate regime

On the other hand, at high rates  $R > D(\delta_{crit}||0.5)$  the  $|\cdot|^+$  is active. It introduces an extra constraint into the optimization which we re-express as

$$E_r(R) = \min_{\delta \in [0,1] \text{ s.t. } D(\delta||0.5) \leq R} D(\delta||p)$$

We can picture the solution on the simplex:

Feasible region:  $\{\delta : D(\delta||0.5) \leq R\}$   
is . Note, extends to left of  $\delta_{crit}$



What happens as you let the feasible region extend all the way to  $p$ ?

- Optimizing  $\delta$  is  $\delta = p$ , so
- $D(p||0.5) = 1 - H_B(p) = C$
- $D(\delta||p) = D(p||p) = 0$  (the error exponent goes to 0 at capacity)

# Summary of results for high- and low-rate regimes

**Low Rate:**  $R < D(\delta_{crit}||0.5) \Rightarrow |\cdot|^+$  is *not* active

$$E_r(R) = D(\delta_{crit}||p) + D(\delta_{crit}||0.5) - R = 1 - \log[1 + 2\sqrt{p(1-p)}] - R$$

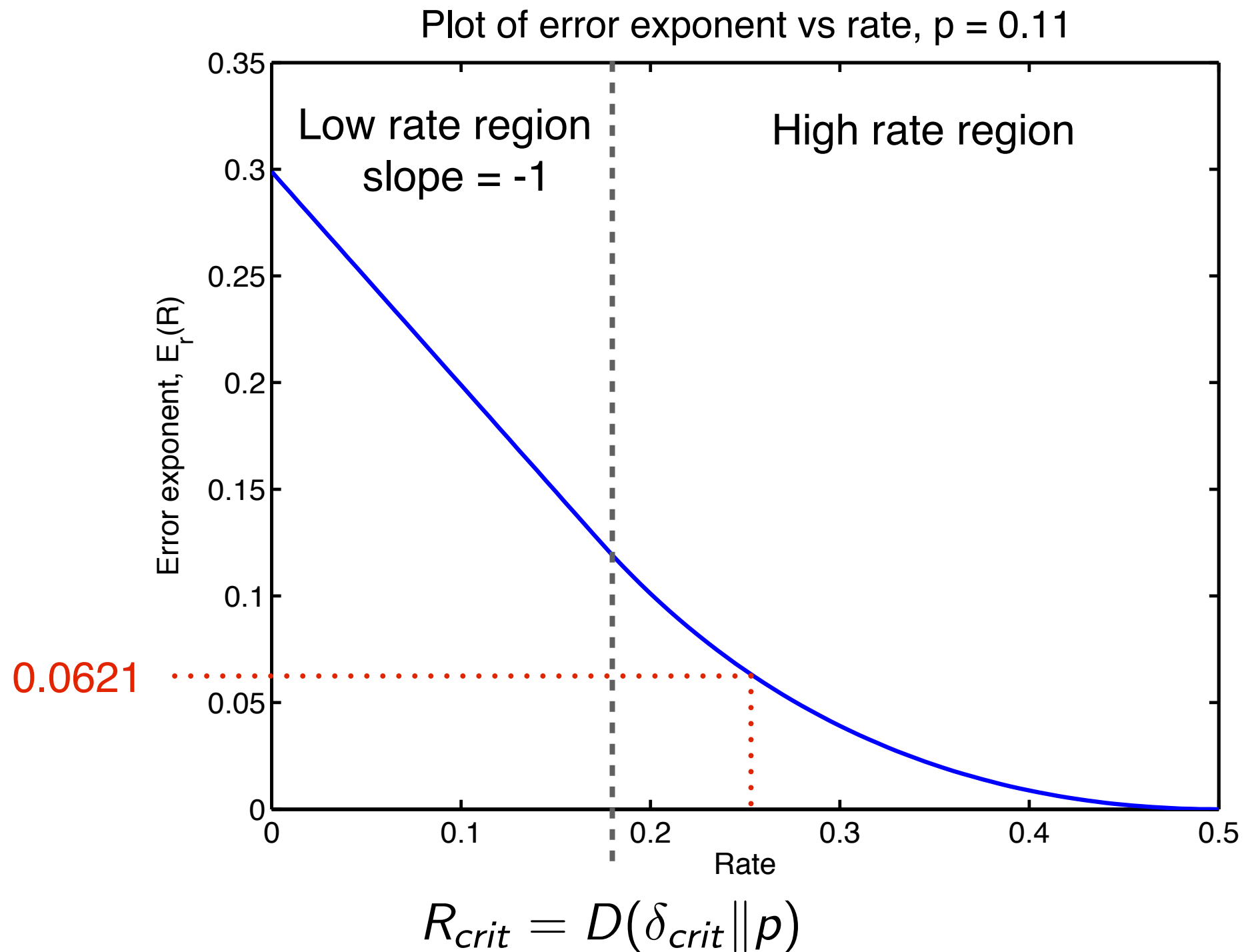
Exponent has a linear slope throughout this region

**High Rate:**  $R > D(\delta_{crit}||0.5) \Rightarrow |\cdot|^+$  is active

$$E_r(R) = \min_{\delta \in [0,1] \text{ s.t. } D(\delta||0.5) \leq R} D(\delta||p)$$

Exponent converges to zero as rate converges to capacity

# Plot of exponent as a function of rate



Can get a matching upper bound in high rate region, the “sphere-packing” bound



# Generalization to arbitrary DMCs

We've seen  $E_r(R)$  for the BSC is: (N.B. at start picked input distribution  $P_x$  to be  $Bern(0.5)$ )

$$\begin{aligned} E_r(R) &= \min_{\delta \in [0,1]} D(\delta \| p) + |D(\delta \| 0.5) - R|^+ \\ &= \min_{\delta \in [0,1]} D(\delta \| p) + |(1 - H_B(\delta)) - R|^+ \end{aligned}$$

worst case  
channel behavior

mutual info realized  
across channel in worst case

For general DMCs more complicated, but underlying idea is the same:

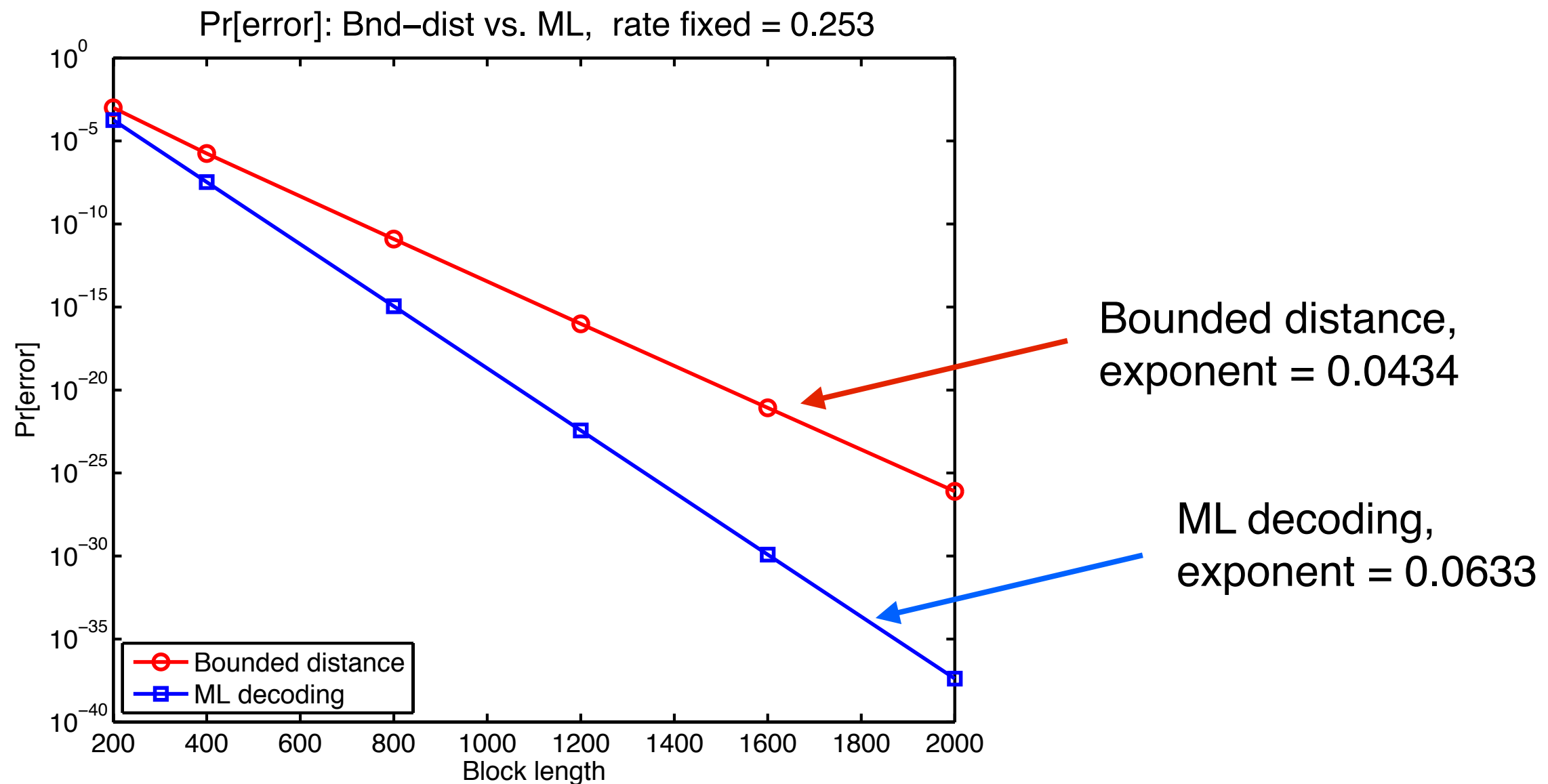
$$E_r(R) = \max_{P_X} \min_{V_{Y|X}} D(P_X V_{Y|X} \| P_X P_{Y|X}) + |I(P_X V_{Y|X}) - R|^+$$

choose best  
input distribution

worst case  
channel behavior

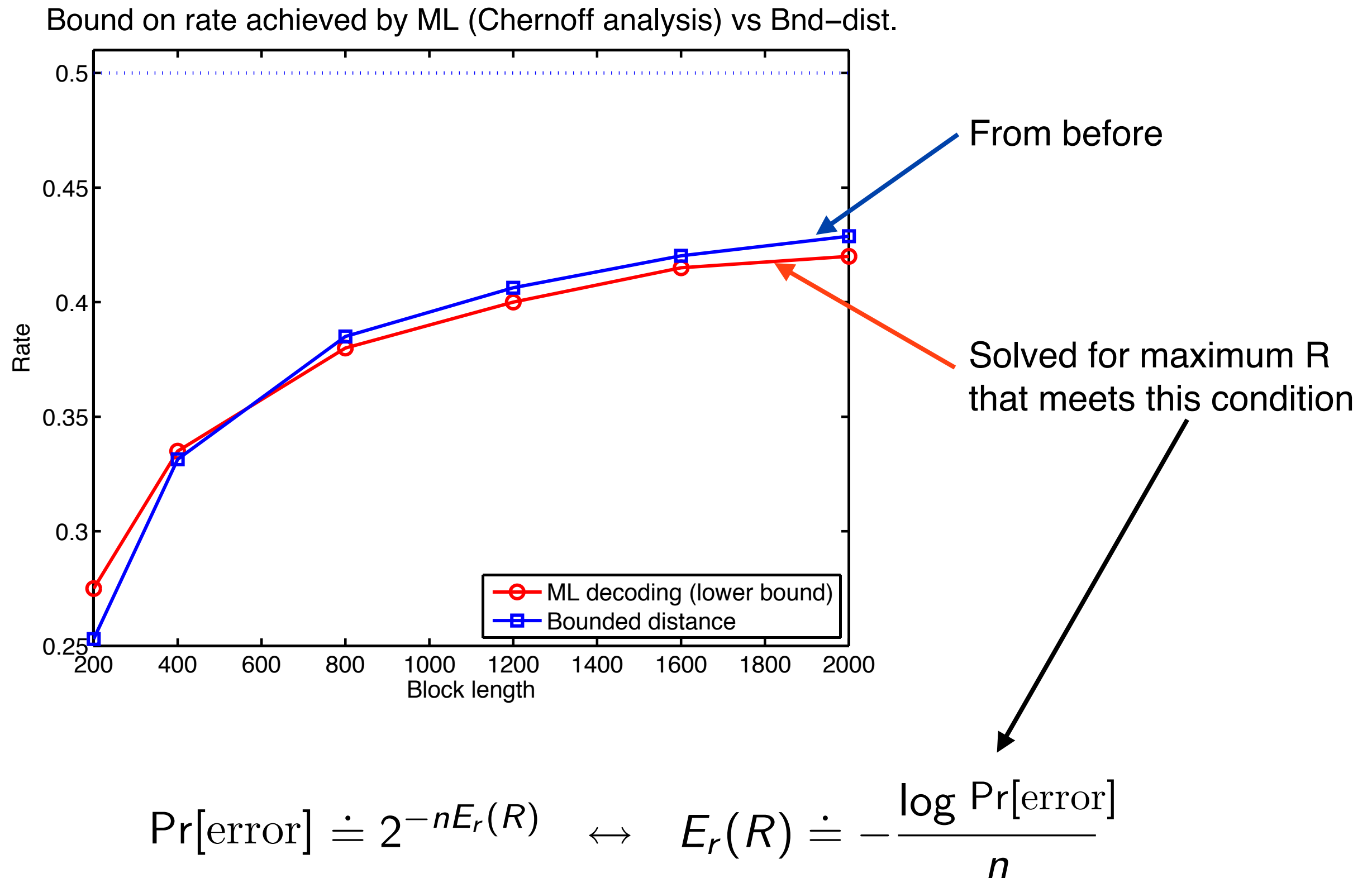
mutual info realized  
across channel in worst case

# How did we do? Compare to bounded-dist. decoder

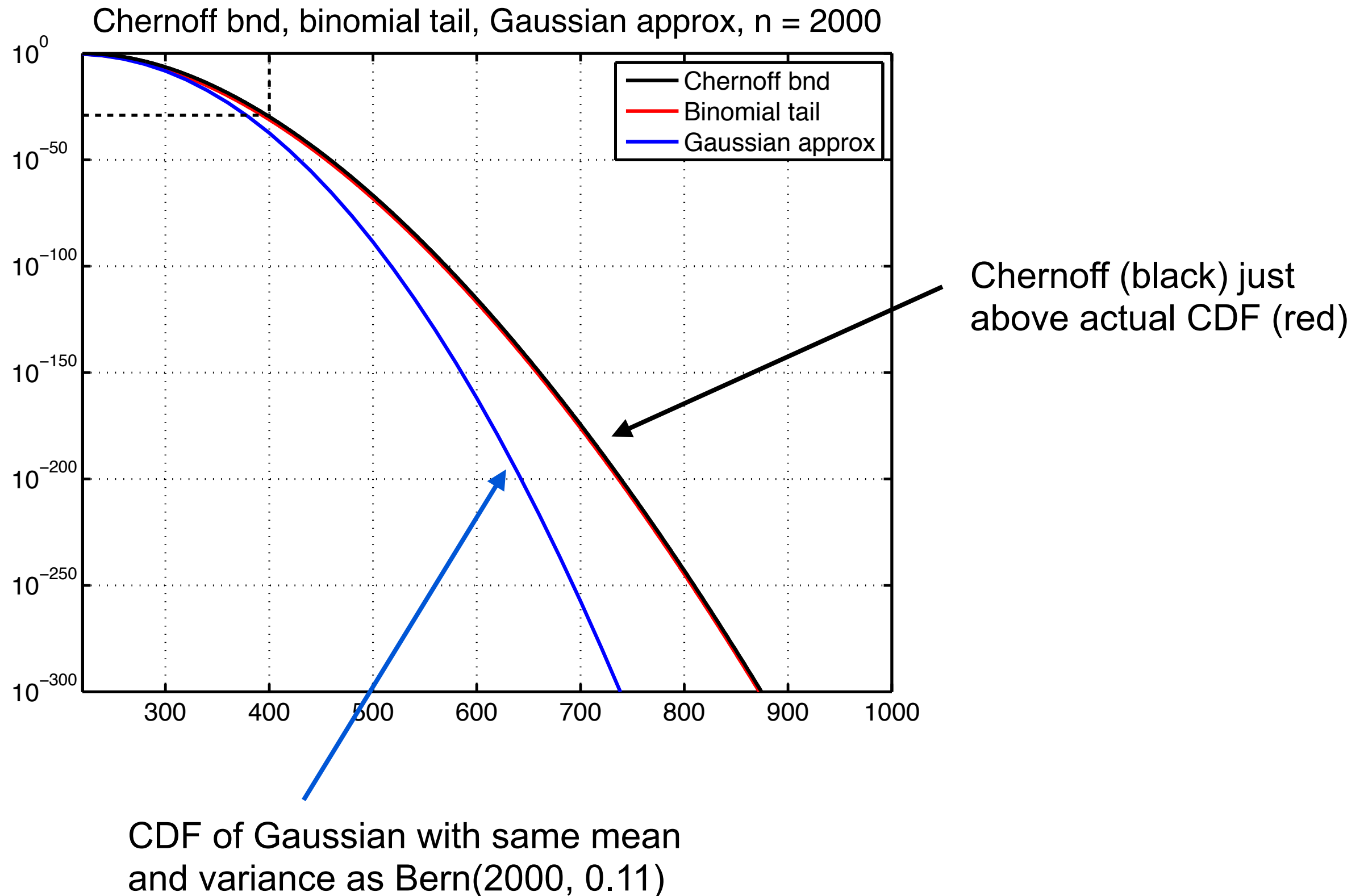


Improvement due to more accurate treatment and analysis of “coupled” events

# Rate analysis



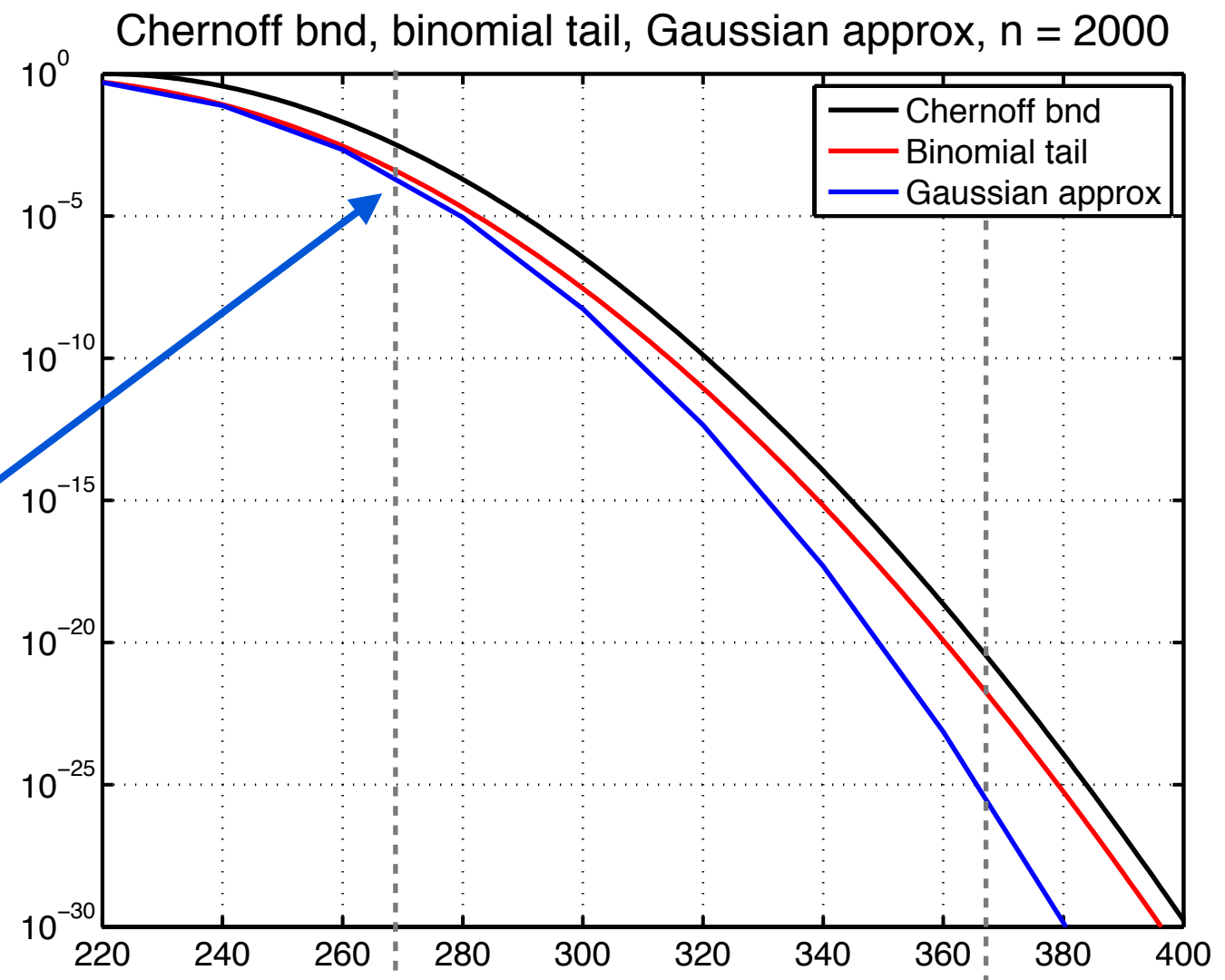
# Chernoff bounds are tight as you go out the tail...



# ...but aren't very good in the regime of interest for fixed $\Pr[\text{error}]$

Zoom in on upper left corner of plot on last slide

Gaussian approximation is actually much closer here (but it's only an approximation and not a bound)



## Remarks

- Our analysis of the ML decoder wasn't very good for the fixed  $\Pr[\text{err}]$  regime.
- This is why, when we compared the bounded distance to the ML decoder, the bounded distance decoder *appeared* to be superior to ML.
- ML will, of course, perform better.
- The weakness was in our analysis -- a good analysis of bounded distance vs. a loose analysis of ML. We improve the latter next.

# Agenda

Analyzing decoding error for a bounded information decoder: regimes of interest

Error exponents of ML decoders

**Non-asymptotic analysis of ML decoder & Normal approximation**

- Look at fixed-error regime

- Use an ML decoder (rather than bounded-dist)

# We branch the ML analysis down a different route

$$\begin{aligned}\Pr[\text{error}] &\leq \Pr \left[ \bigcup_{m=1}^{M-1} i(\mathbf{X}_m, \mathbf{Y}) \geq i(\mathbf{X}_0, \mathbf{Y}) \right] \\ &= \Pr \left[ \bigcup_{m=1}^{M-1} \text{wt}_H(\mathbf{X}_m, \mathbf{Y}) \leq \text{wt}_H(\mathbf{X}_0, \mathbf{Y}) \right] \\ &= \mathbb{E}_{\mathbf{X}_0, \mathbf{Y}} \left[ \Pr \left[ \bigcup_{m=1}^{M-1} \text{wt}_H(\mathbf{X}_m \oplus \mathbf{Y}) \leq \text{wt}_H(\mathbf{X}_0 \oplus \mathbf{Y}) \mid \mathbf{X}_0, \mathbf{Y} \right] \right]\end{aligned}$$

# We branch the ML analysis down a different route

$$\begin{aligned}\Pr[\text{error}] &\leq \Pr \left[ \bigcup_{m=1}^{M-1} i(\mathbf{X}_m, \mathbf{Y}) \geq i(\mathbf{X}_0, \mathbf{Y}) \right] \\ &= \Pr \left[ \bigcup_{m=1}^{M-1} \text{wt}_H(\mathbf{X}_m, \mathbf{Y}) \leq \text{wt}_H(\mathbf{X}_0, \mathbf{Y}) \right] \\ &= \mathbb{E}_{\mathbf{X}_0, \mathbf{Y}} \left[ \Pr \left[ \bigcup_{m=1}^{M-1} \text{wt}_H(\mathbf{X}_m \oplus \mathbf{Y}) \leq \text{wt}_H(\mathbf{X}_0 \oplus \mathbf{Y}) \mid \mathbf{X}_0, \mathbf{Y} \right] \right] \\ &= \mathbb{E}_{\mathbf{X}_0, \mathbf{Y}} \left[ \min \left\{ 1, \Pr \left[ \bigcup_{m=1}^{M-1} \text{wt}_H(\mathbf{X}_m \oplus \mathbf{Y}) \leq \text{wt}_H(\mathbf{X}_0 \oplus \mathbf{Y}) \mid \mathbf{X}_0, \mathbf{Y} \right] \right\} \right] \\ &\leq \mathbb{E}_{\mathbf{X}_0, \mathbf{Y}} \left[ \min \left\{ 1, (M-1) \Pr[\text{wt}_H(\mathbf{X}_1 \oplus \mathbf{Y}) \leq \text{wt}_H(\mathbf{X}_0 \oplus \mathbf{Y}) \mid \mathbf{X}_0, \mathbf{Y}] \right\} \right]\end{aligned}$$



# We branch the ML analysis down a different route

$$\begin{aligned}\Pr[\text{error}] &\leq \Pr \left[ \bigcup_{m=1}^{M-1} i(\mathbf{X}_m, \mathbf{Y}) \geq i(\mathbf{X}_0, \mathbf{Y}) \right] \\&= \Pr \left[ \bigcup_{m=1}^{M-1} \text{wt}_H(\mathbf{X}_m, \mathbf{Y}) \leq \text{wt}_H(\mathbf{X}_0, \mathbf{Y}) \right] \\&= \mathbb{E}_{\mathbf{X}_0, \mathbf{Y}} \left[ \Pr \left[ \bigcup_{m=1}^{M-1} \text{wt}_H(\mathbf{X}_m \oplus \mathbf{Y}) \leq \text{wt}_H(\mathbf{X}_0 \oplus \mathbf{Y}) \mid \mathbf{X}_0, \mathbf{Y} \right] \right] \\&= \mathbb{E}_{\mathbf{X}_0, \mathbf{Y}} \left[ \min \left\{ 1, \Pr \left[ \bigcup_{m=1}^{M-1} \text{wt}_H(\mathbf{X}_m \oplus \mathbf{Y}) \leq \text{wt}_H(\mathbf{X}_0 \oplus \mathbf{Y}) \mid \mathbf{X}_0, \mathbf{Y} \right] \right\} \right] \\&\leq \mathbb{E}_{\mathbf{X}_0, \mathbf{Y}} \left[ \min \left\{ 1, (M-1) \Pr[\text{wt}_H(\mathbf{X}_1 \oplus \mathbf{Y}) \leq \text{wt}_H(\mathbf{X}_0 \oplus \mathbf{Y}) \mid \mathbf{X}_0, \mathbf{Y}] \right\} \right] \\&= \mathbb{E}_{\text{wt}_H(\tilde{\mathbf{X}}_0)} \left[ \min \left\{ 1, (M-1) \Pr[\text{wt}_H(\tilde{\mathbf{X}}_1) \leq \text{wt}_H(\tilde{\mathbf{X}}_0) \mid \text{wt}_H(\tilde{\mathbf{X}}_0)] \right\} \right]\end{aligned}$$

# We branch the ML analysis down a different route

$$\begin{aligned}\Pr[\text{error}] &\leq \Pr \left[ \bigcup_{m=1}^{M-1} i(\mathbf{X}_m, \mathbf{Y}) \geq i(\mathbf{X}_0, \mathbf{Y}) \right] \\&= \Pr \left[ \bigcup_{m=1}^{M-1} \text{wt}_H(\mathbf{X}_m, \mathbf{Y}) \leq \text{wt}_H(\mathbf{X}_0, \mathbf{Y}) \right] \\&= \mathbb{E}_{\mathbf{X}_0, \mathbf{Y}} \left[ \Pr \left[ \bigcup_{m=1}^{M-1} \text{wt}_H(\mathbf{X}_m \oplus \mathbf{Y}) \leq \text{wt}_H(\mathbf{X}_0 \oplus \mathbf{Y}) \mid \mathbf{X}_0, \mathbf{Y} \right] \right] \\&= \mathbb{E}_{\mathbf{X}_0, \mathbf{Y}} \left[ \min \left\{ 1, \Pr \left[ \bigcup_{m=1}^{M-1} \text{wt}_H(\mathbf{X}_m \oplus \mathbf{Y}) \leq \text{wt}_H(\mathbf{X}_0 \oplus \mathbf{Y}) \mid \mathbf{X}_0, \mathbf{Y} \right] \right\} \right] \\&\leq \mathbb{E}_{\mathbf{X}_0, \mathbf{Y}} \left[ \min \left\{ 1, (M-1) \Pr[\text{wt}_H(\mathbf{X}_1 \oplus \mathbf{Y}) \leq \text{wt}_H(\mathbf{X}_0 \oplus \mathbf{Y}) \mid \mathbf{X}_0, \mathbf{Y}] \right\} \right] \\&= \mathbb{E}_{\text{wt}_H(\tilde{\mathbf{X}}_0)} \left[ \min \left\{ 1, (M-1) \Pr[\text{wt}_H(\tilde{\mathbf{X}}_1) \leq \text{wt}_H(\tilde{\mathbf{X}}_0) \mid \text{wt}_H(\tilde{\mathbf{X}}_0)] \right\} \right] \\&\leq \sum_{\Delta=0}^n \binom{n}{\Delta} p^\Delta (1-p)^{n-\Delta} \min \left\{ 1, (M-1) \sum_{s=0}^{\Delta} \binom{n}{s} 2^{-n} \right\}\end{aligned}$$

# Random coding union (RCU) bound

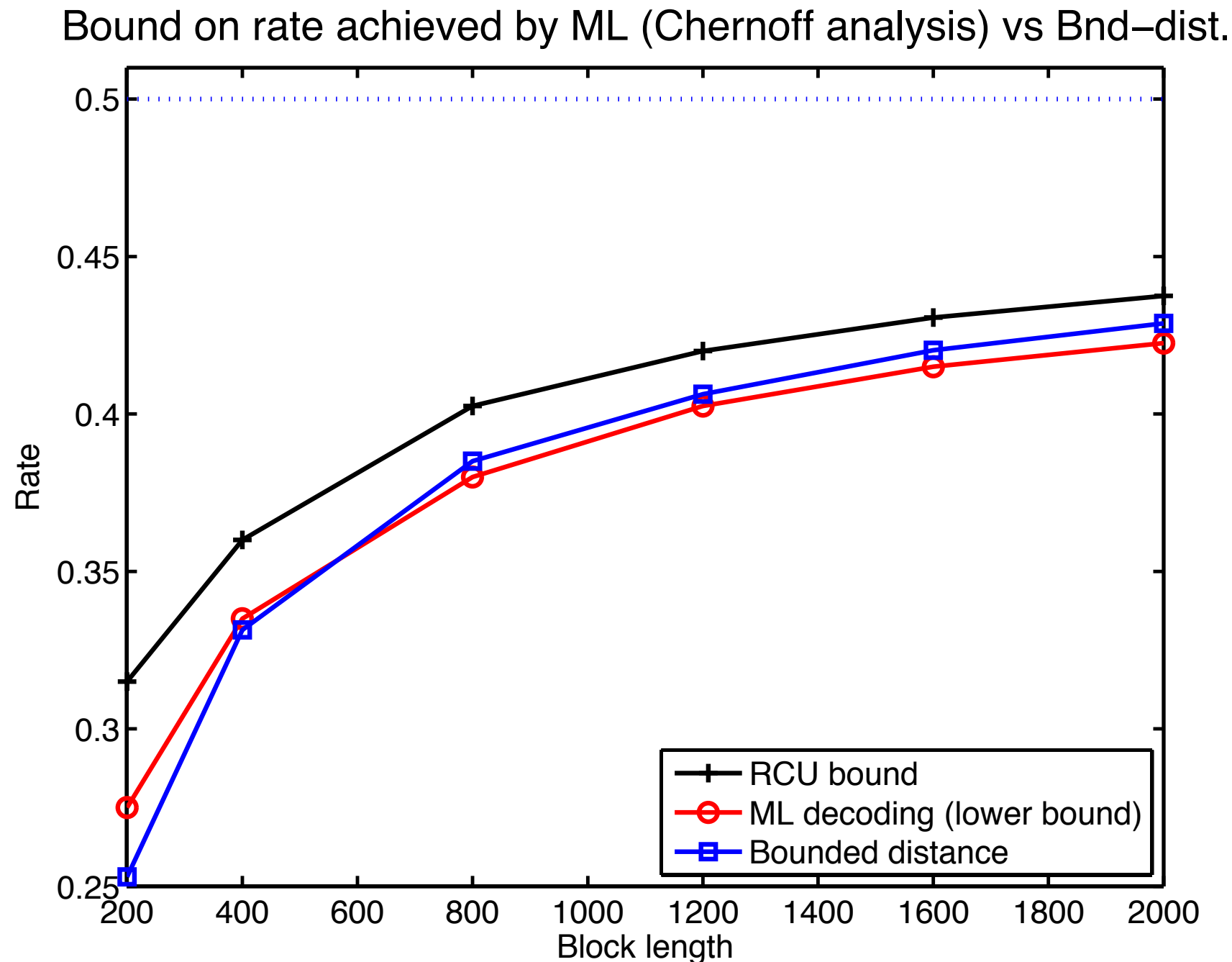
$$\Pr[\text{error}] \leq \sum_{\Delta=0}^n \binom{n}{\Delta} p^{\Delta} (1-p)^{n-\Delta} \min \left\{ 1, (M-1) \sum_{s=0}^{\Delta} \binom{n}{s} 2^{-n} \right\}$$

Remarks:

- This bound is called the “random coding union bound” (RCU) in Polyanskiy-Poor-Verdú (PPV '10)
- Here we condition on the distance between the observation and Tx c.w. to get the coupling, taking the expectation of the conditional probability that some other codeword happens to be closer to the observation.
- Holds for all block lengths
- Have *not* applied a Chernoff bound
- Can plot this bound for non-trivial block lengths
- Aside: to plot  $\binom{n}{k}$  for large  $n$  &  $k$ , it is better numerically to compute

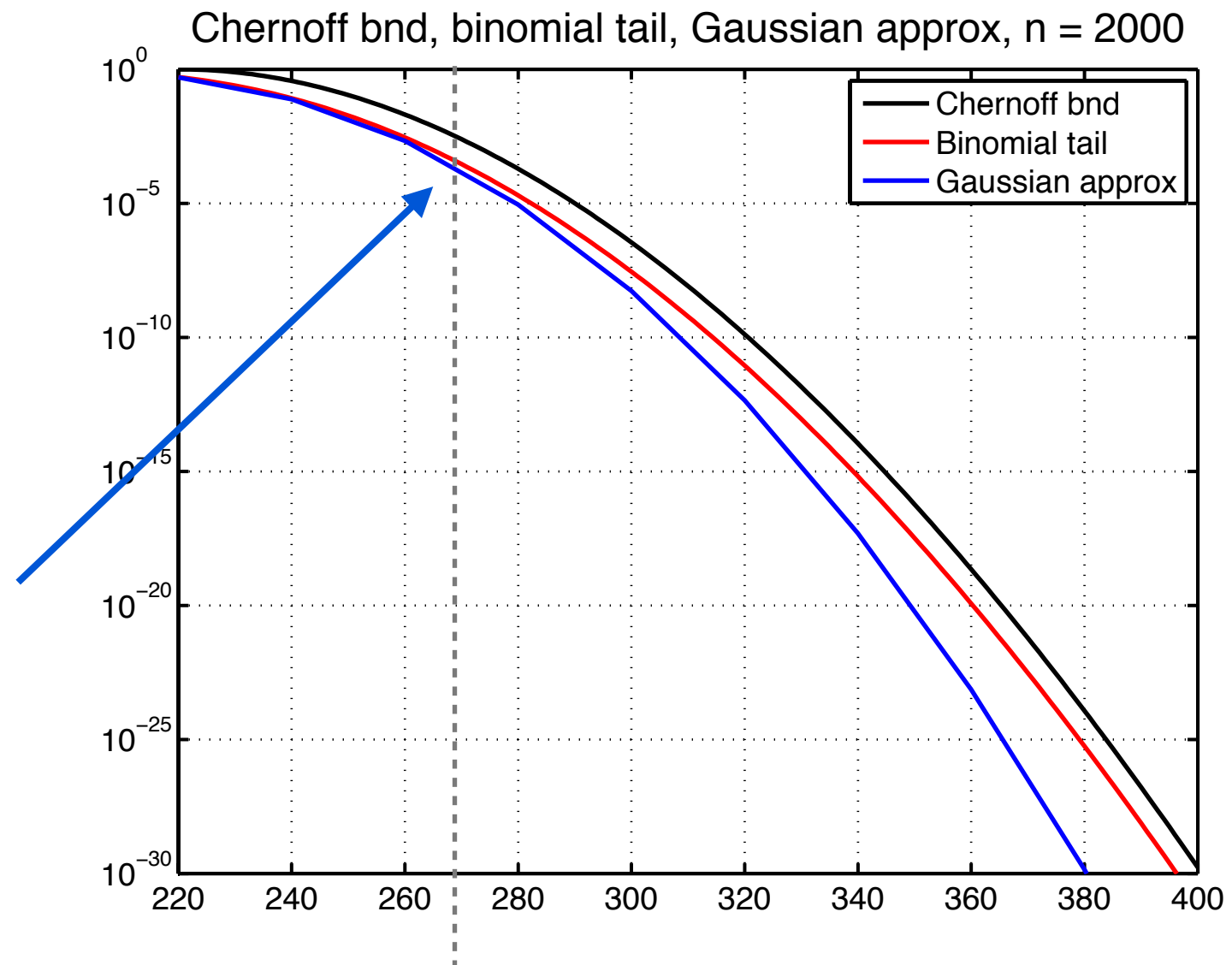
$$\log \binom{n}{k} = \sum_{\ell=\max\{k, n-k\}}^n \log \ell - \sum_{\ell=1}^{\min\{k, n-k\}} \log \ell$$

# Compare: RCU, Bounded-dist, ML via Chernoff



- Now we see the ML decoder (RCU bound) outperforms bounded distance.
- Can we characterize how we approach capacity as block length gets large?
- Can we cleanly break into “outage” and “confusion” events as before?

# Recall CDF of Gaussian approximated better near mean



Will need to bound events more similar to those in the bounded distance dec.

In this regime a Gaussian approximation looks good, we need to correct the approximation to turn it into a bound

threshold 268  
for  $\Pr[\text{err}] = 0.001$

# Turn Gaussian approx into a bound via Berry-Esseen

Theorem (Berry-Esseen for i.i.d. Bernoulli r.v.)

For  $X_1, \dots, X_n \sim \text{i.i.d. Bern}(p)$

$$\left| \Pr \left[ \sum_{i=1}^n (X_i - \mu) \geq \lambda \sqrt{n\sigma^2} \right] - Q(\lambda) \right| \leq \frac{B}{\sqrt{n}}$$

*alternately*

$$\left| \Pr \left[ \sum_{i=1}^n (X_i - \mu) \leq \lambda \sqrt{n\sigma^2} \right] - Q(-\lambda) \right| \leq \frac{B}{\sqrt{n}}$$


where  $Q(\lambda) = \frac{1}{\sqrt{2\pi}} \int_{\lambda}^{\infty} e^{-y^2/2} dy$  and  $\mu = p$ ,  $\sigma^2 = p(1 - p)$ .

- Note:
- (i) Bounds the absolute difference between CDF and tail of a Gaussian
  - (ii) Can be generalized to other (and non-i.i.d.) distributions
  - (iii) The Berry-Esseen constant “ $B$ ” in this case is about 2.5

As before split into “outage” and “confusion events”

$$\Pr[\text{error}] \leq \sum_{\Delta=0}^n \binom{n}{\Delta} p^{\Delta} (1-p)^{n-\Delta} \min \left\{ 1, M \sum_{s=0}^{\Delta} \binom{n}{s} 2^{-n} \right\}$$

Now, choose  $M = \frac{1}{\sum_{s=0}^K \binom{n}{s} 2^{-n}}$  for some  $K$  to be specified

$$\Pr[\text{error}] \leq \sum_{\Delta=0}^K \binom{n}{\Delta} p^{\Delta} (1-p)^{n-\Delta} \left[ M \sum_{s=0}^{\Delta} \binom{n}{s} 2^{-n} \right] + \sum_{\Delta=K+1}^n \binom{n}{\Delta} p^{\Delta} (1-p)^{n-\Delta}$$


**Confusion**: some other c.w. too close

**Outage**: true c.w. too far away

Holds because  $M \sum_{s=0}^{\Delta} \binom{n}{s} 2^{-n}$  is strictly increasing in  $\Delta$  and with above choice for  $M$  certainly  $M \sum_{s=0}^{\Delta} \binom{n}{s} 2^{-n}$  exceeds 1 for  $\Delta > K$

Question: How should we pick  $K$ ?

# What's new: scaling we choose for K (and analysis)

$$K = np + \sqrt{np(1-p)} Q^{-1} \left( \epsilon - \frac{B+G}{\sqrt{n}} \right)$$

Diagram illustrating the scaling for K (and analysis). The formula is shown with annotations:

- $np$  is labeled "mean".
- $\sqrt{np(1-p)}$  is labeled "standard deviation".
- $Q^{-1} \left( \epsilon - \frac{B+G}{\sqrt{n}} \right)$  is labeled "# stnd. dev. above mean (B is Berry-Esseen const, G is to be discussed)".
- The entire term  $Q^{-1} \left( \epsilon - \frac{B+G}{\sqrt{n}} \right)$  is circled in red.
- The parameter  $\epsilon$  is labeled "target  $Pr[err]$ , which was  $\epsilon = 0.001$  in our example".

Recall from error exponent story:

- Margin the error exponent threshold (the  $n\delta$ ) was above the mean increased **linearly** with block length
- Here the margin the threshold  $K$  is above the mean increases only as the **square-root** of the block length

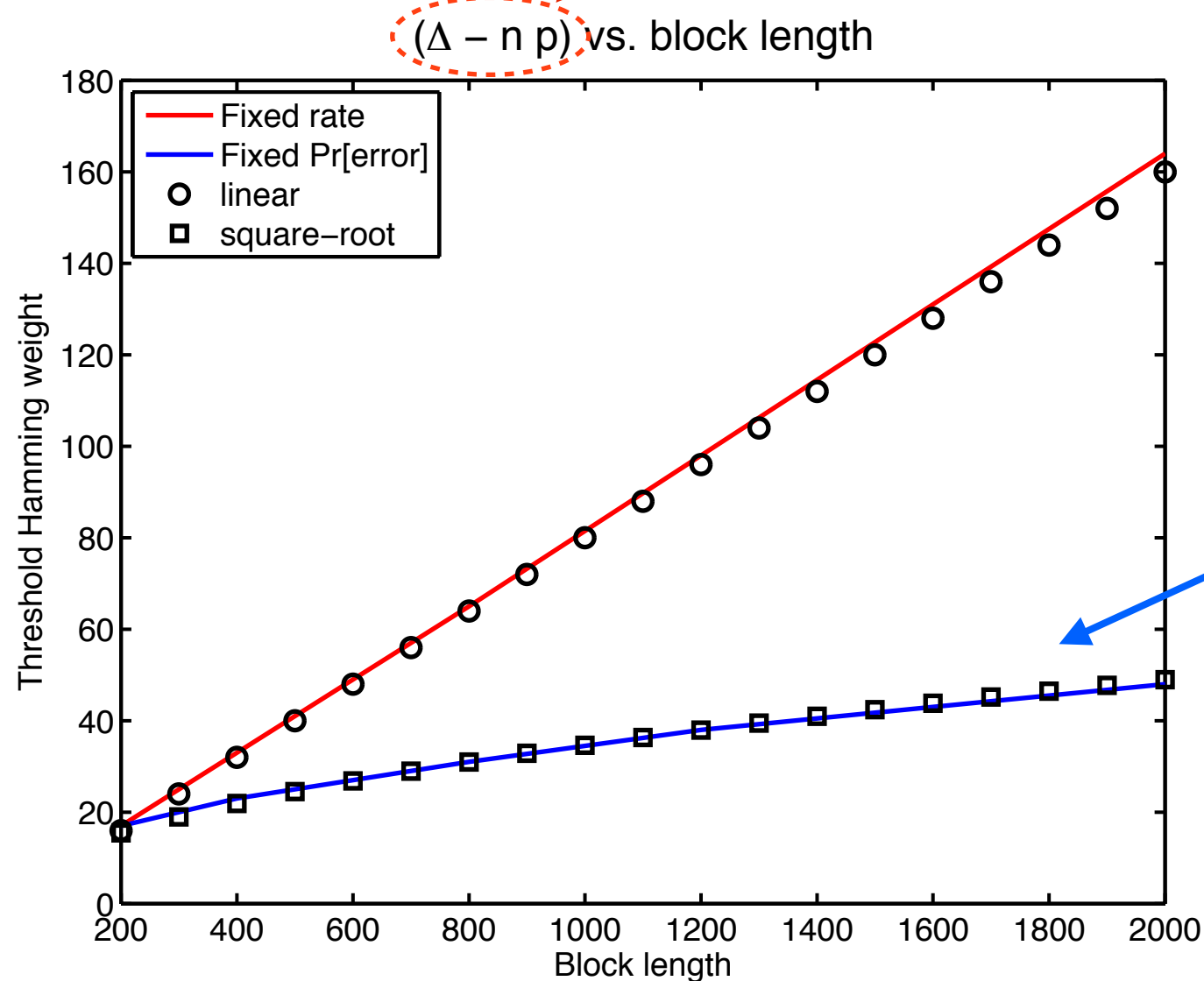
Q1: Where have we seen this scaling before?

Q2: What do we gain by making this choice for  $K$ ?



# What's new: in the fixed-error analysis

we plotted the margin above the mean Hamming weight



Recall this increased as the square-root of the block length

In current context threshold is the  $K$  parameter

$$\Pr[\text{error}] \leq \sum_{\Delta=0}^K \binom{n}{\Delta} p^{\Delta} (1-p)^{n-\Delta} \left[ M \sum_{s=0}^{\Delta} \binom{n}{s} 2^{-n} \right] + \sum_{\Delta=K+1}^n \binom{n}{\Delta} p^{\Delta} (1-p)^{n-\Delta}$$

# Apply Berry-Esseen to outage term

First, bound outage:  $\sum_{\Delta > K} \binom{n}{\Delta} p^{\Delta} (1-p)^{n-\Delta} = \Pr \left[ \sum_{i=1}^n X_i > K \right]$

$$\begin{aligned} \Pr \left[ \sum_{i=1}^n X_i > K \right] &= \Pr \left[ \sum_{i=1}^n (X_i - p) > K - np \right] \\ &= \Pr \left[ \sum_{i=1}^n (X_i - p) > \sqrt{np(1-p)} Q^{-1} \left( \epsilon - \frac{B+G}{\sqrt{n}} \right) \right] \\ &\leq Q \left( Q^{-1} \left( \epsilon - \frac{B+G}{\sqrt{n}} \right) \right) + \frac{B}{\sqrt{n}} \\ &= \epsilon - \frac{B+G}{\sqrt{n}} + \frac{B}{\sqrt{n}} = \epsilon - \frac{G}{\sqrt{n}} \end{aligned}$$

plays role of  $\lambda$

Choice  $\epsilon - \frac{B+G}{\sqrt{n}}$  allows us to cancel out Berry-Esseen const  $B$

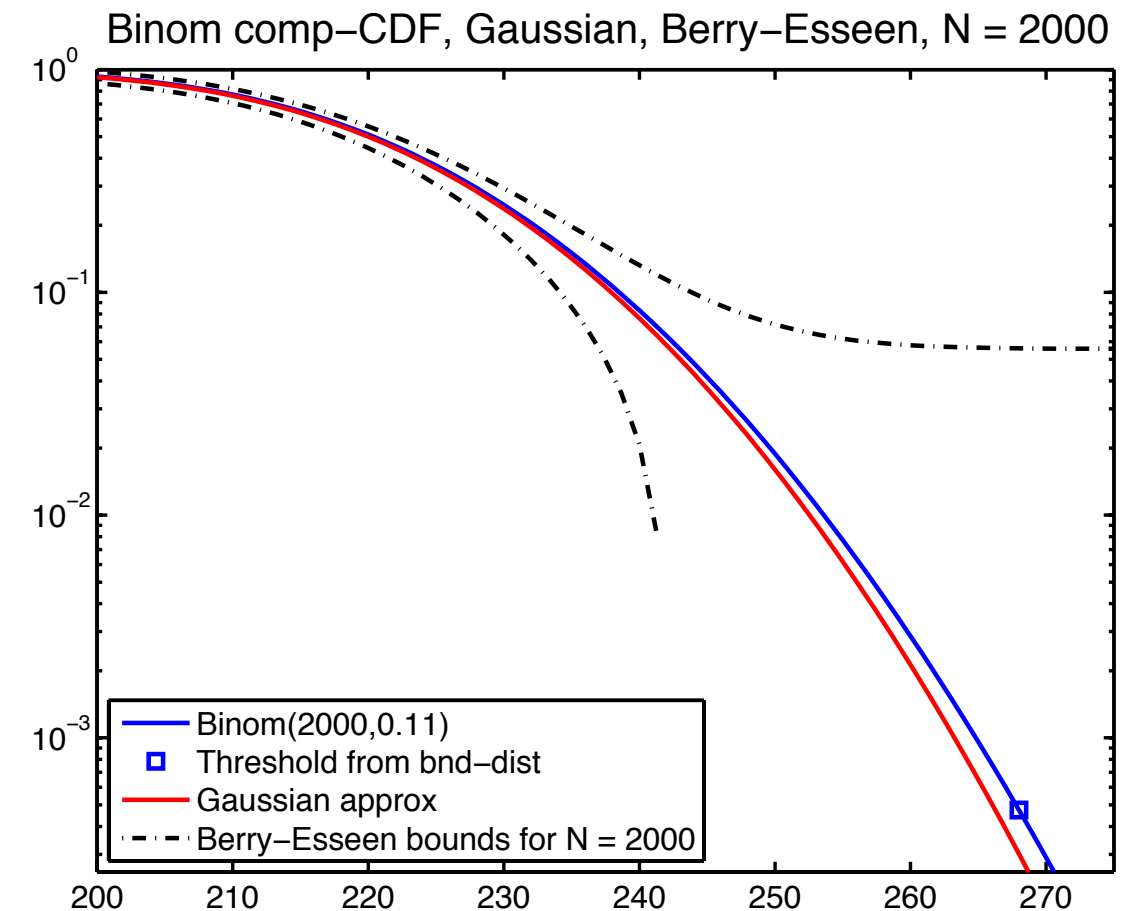
# A caveat

When going through step

$$Q\left(Q^{-1}\left(\epsilon - \frac{B+G}{\sqrt{n}}\right)\right) = \epsilon - \frac{B+G}{\sqrt{n}}$$

Need

$$\epsilon - \frac{B+G}{\sqrt{n}} > 0$$

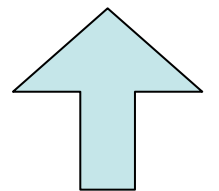


For our running example of  $n = 2000$ ,  $\sqrt{n} \simeq 45$  and with  $B = 2.57$  and  $G = 2.68$  ( $G$  is yet to be defined) we compute

$$\epsilon - \frac{B+G}{\sqrt{n}} = -0.1166$$

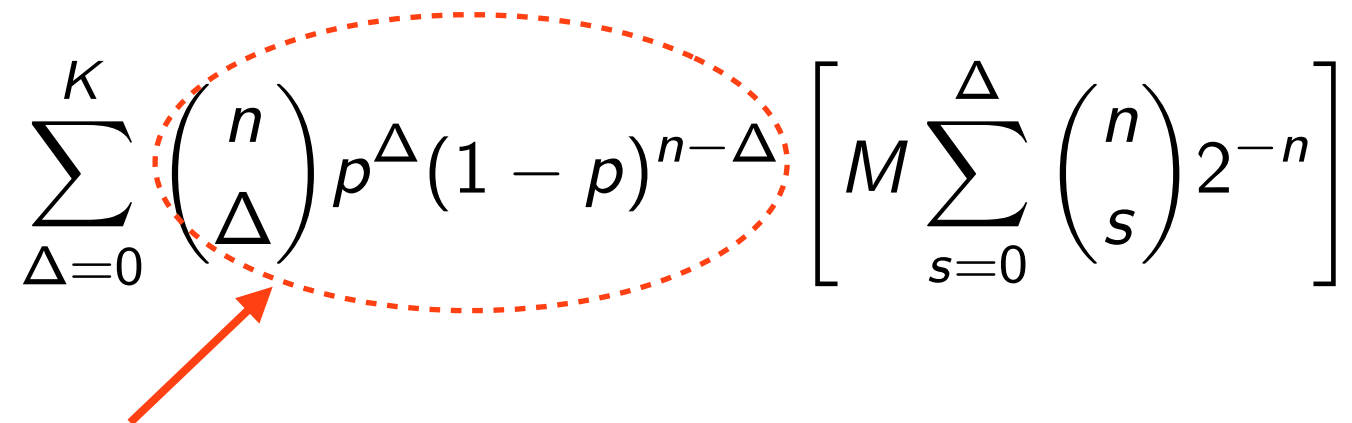
I.e., the approximation we made here is valid only for large  $n$

- This analysis doesn't give bnds for all  $n$  (as did earlier ones)
- Nevertheless, turns out to be quite accurate even for small  $n$
- It *will* tell us how rate approaches cap. as  $n \rightarrow \infty$



Berry-Esseen bounds quite loose relative to an error of 0.001 at block length of  $n = 2000$

## Next bound the “confusion” term

$$\sum_{\Delta=0}^K \left( \binom{n}{\Delta} p^{\Delta} (1-p)^{n-\Delta} \right) \left[ M \sum_{s=0}^{\Delta} \binom{n}{s} 2^{-n} \right]$$


One term in a Binomial distribution, will bound by the max term:

First:

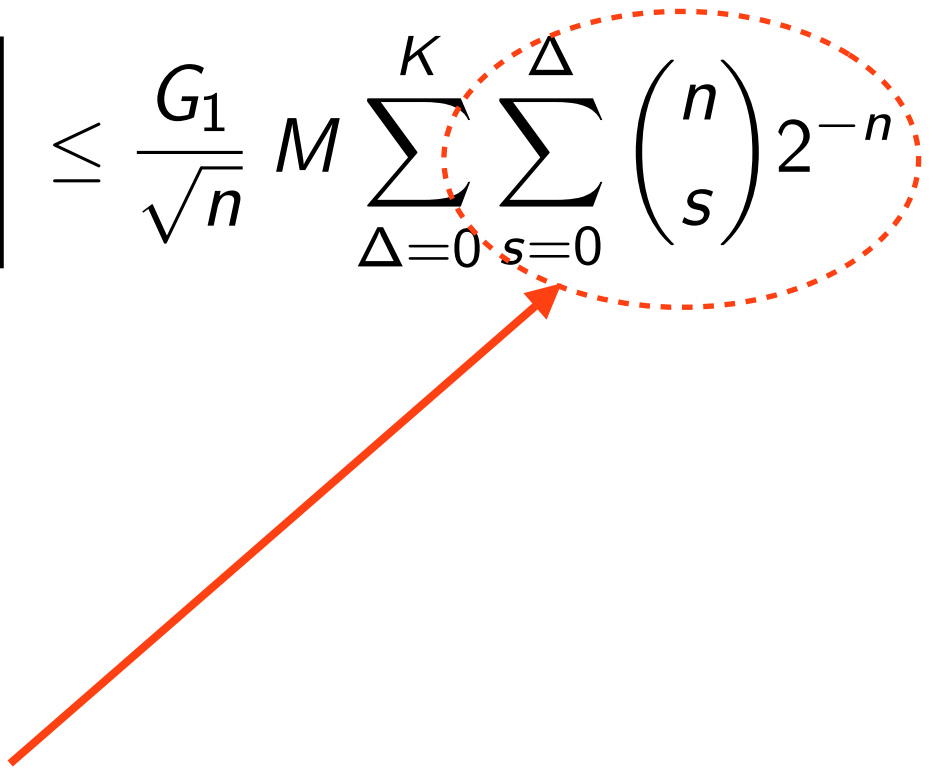
$$\binom{n}{\Delta} p^{\Delta} (1-p)^{n-\Delta} < \frac{G_1}{\sqrt{n}}$$

- If  $np \in \mathbb{Z}$  then mean of Binomial is the mode (see C&T)
- Then apply Stirling's approx to get the  $1/\sqrt{n}$

$$\sqrt{2\pi n} 2^{n \ln n - n} < n! < e^{1/12} \sqrt{2\pi n} e^{n \ln n - n}$$

- For reference:  $G_1 = (e^{1/12}) / \sqrt{2\pi p(1-p)}$

## Continue with the “confusion” term

$$\sum_{\Delta=0}^K \binom{n}{\Delta} p^{\Delta} (1-p)^{n-\Delta} \left[ M \sum_{s=0}^{\Delta} \binom{n}{s} 2^{-n} \right] \leq \frac{G_1}{\sqrt{n}} M \sum_{\Delta=0}^K \sum_{s=0}^{\Delta} \binom{n}{s} 2^{-n}$$


Note that this is the tail of a Binomial distribution (again)

We've seen two types of bounds today:

- Chernoff “large-deviation” bounds
- Berry-Esseen correction of Gaussian approximation

Which do you think is the right to use here?

# Should use a large deviations bound

Large deviation because:  $\Delta \leq K = np + O(\sqrt{n}) \ll \frac{n}{2}$  if  $n$  is large

So, the events are far from the mean, & Berry-Esseen won't be good

After some manipulation (See PPV '10) get

$$\begin{aligned} \sum_{\Delta=0}^K \binom{n}{\Delta} p^{\Delta} (1-p)^{n-\Delta} \left[ M \sum_{s=0}^{\Delta} \binom{n}{s} 2^{-n} \right] &\leq \frac{G_1}{\sqrt{n}} M \sum_{\Delta=0}^K \sum_{s=0}^{\Delta} \binom{n}{s} 2^{-n} \\ &\leq \frac{G_1}{\sqrt{n}} \left( \frac{1-r}{1-2r} \right)^2 = \frac{G}{\sqrt{n}} \end{aligned}$$

Note that the constant  $r$  can be chosen in the range  $p < r < 0.5$ , where  $r$  play the role of the fraction of flips in the large deviation result

# Combine two analyses

$$\Pr[\text{error}] \leq \sum_{\Delta=0}^k \binom{n}{\Delta} p^{\Delta} (1-p)^{n-\Delta} \left[ M \sum_{s=0}^{\Delta} \binom{n}{s} 2^{-n} \right] + \sum_{\Delta=K+1}^n \binom{n}{\Delta} p^{\Delta} (1-p)^{n-\Delta}$$

$$\leq \frac{G_1}{\sqrt{n}}$$

This is the “confusion” event  
 The dominant event is **far** from the mean  
 Bound using a large-deviation technique

$$\leq \epsilon - \frac{G_1}{\sqrt{n}}$$

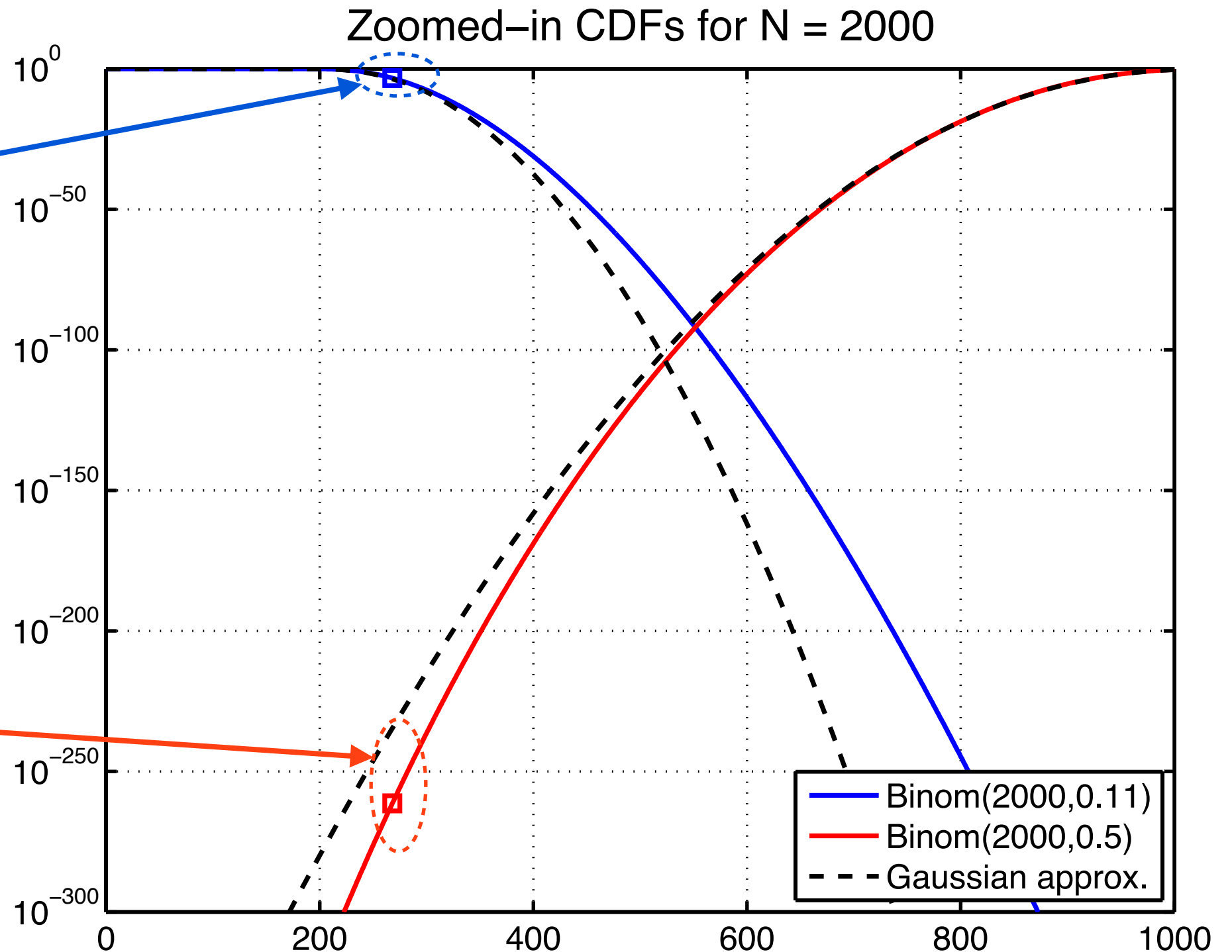
This is the “outage” event  
 The dominant event is **closer** to the mean  
 Bound using a CLT (Berry-Esseen) technique

$$\text{Final bound: } \Pr[\text{error}] \leq \frac{G_1}{\sqrt{n}} + \left( \epsilon - \frac{G_1}{\sqrt{n}} \right) = \epsilon$$

# Visualize the need to combine bounding techniques

Gaussian approximation is accurate in this region, Chernoff is loose

Gaussian approximation terrible here, but saw earlier that Chernoff quite good





# Counting codeword: what rate did we get?

Recall our earlier choice:

$$M = \left[ 2^{-n} \sum_{s=0}^K \binom{n}{s} \right]^{-1}$$

Get lower-bound on  $M$ :

$$\begin{aligned} 2^{-n} \sum_{s=0}^K \binom{n}{s} &\leq 2^{-n} \sum_{s=0}^K \binom{n}{K} \left( \frac{K}{n-K} \right)^{K-s} \\ &\leq 2^{-n} \binom{n}{K} \sum_{s=0}^{\infty} \left( \frac{K}{n-K} \right)^s \\ &= 2^{-n} \binom{n}{K} \frac{n-K}{n-2K} \end{aligned}$$

where  $\frac{\binom{n}{s}}{\binom{n}{K}} \leq \left( \frac{K}{n-K} \right)^{K-s}$ , useful for bounding terms in a Binomial distribution, follows using  $\frac{n!}{m!} \leq n^{n-m}$

# Counting codeword: calculate log-number of msgs

$$\log M \geq n - \log \binom{n}{K} - \log \frac{n-K}{n-2K}$$

$O(1)$  terms

Applying Stirling's approximation to  $\binom{n}{K}$  we get

$$-\log \binom{n}{K} \geq -0.5 \log \left( \frac{n}{K(n-K)} \right) - nH_B \left( \frac{K}{n} \right) - \log \left( \frac{e^{1/12}}{\sqrt{2\pi}} \right)$$

Recall “Big-O” notation:  $f(x) = O(g(x))$  as  $x \rightarrow \infty$  if and only if  $\exists A > 0$  and an  $x_0$  such that  $|f(x)| \leq A|g(x)|$  for all  $x > x_0$

$$\log M \geq n - 0.5 \log \left( \frac{n}{K(n-K)} \right) - nH_B \left( \frac{K}{n} \right) + O(1)$$

# Final step: linearize via Taylor's formula

$$\log M \geq n - 0.5 \log \left( \frac{n}{K(n-K)} \right) - nH_B \left( \frac{K}{n} \right) + O(1)$$

Apply Taylor's formula and the choice of K to these two terms

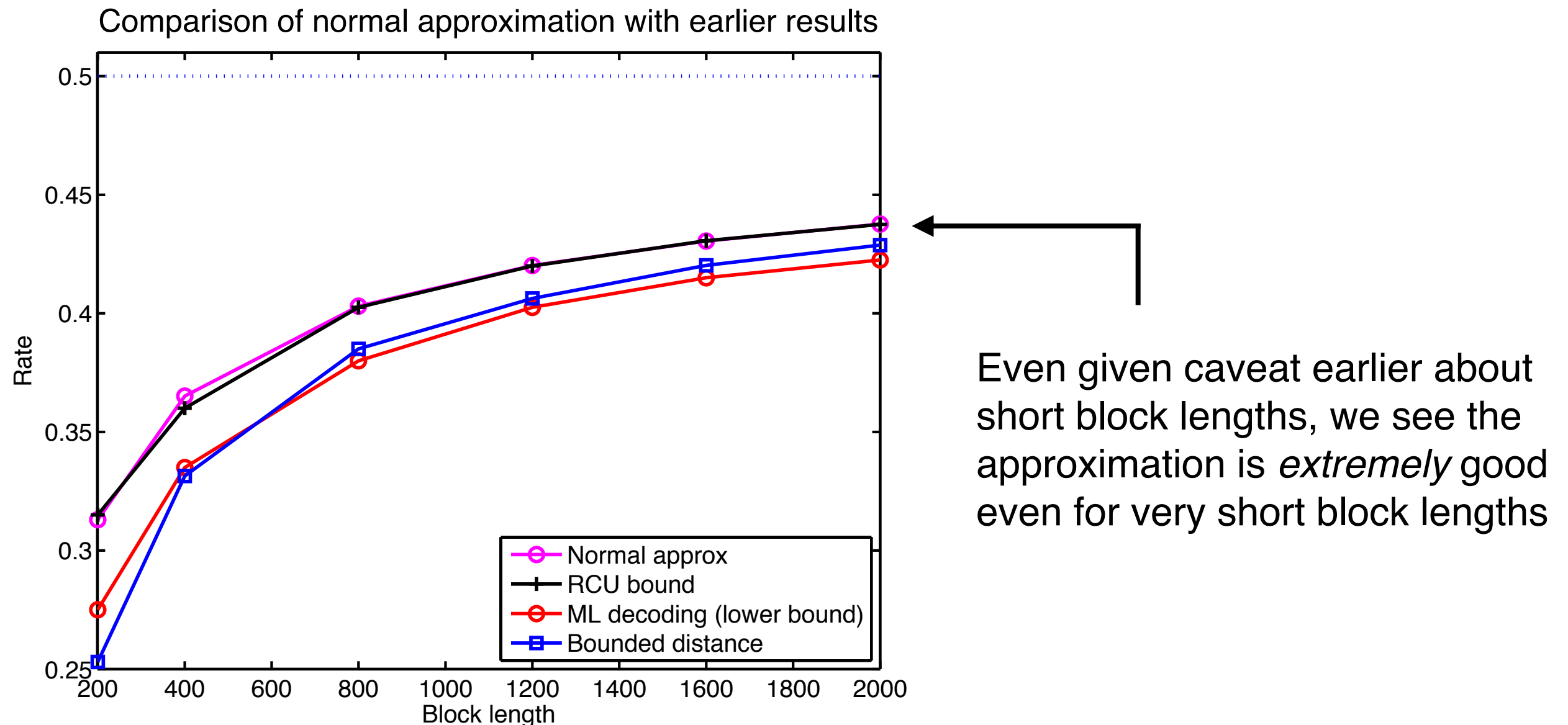
$$K = np + \sqrt{np(1-p)} Q^{-1} \left( \epsilon - \frac{B+G}{\sqrt{n}} \right)$$

$$\log M \geq n - nH_B(p) + 0.5 \log n - \log \frac{1-p}{p} \sqrt{np(1-p)} Q^{-1} \left( \epsilon - \frac{B+G}{\sqrt{n}} \right) + O(1)$$

$$\log \frac{1-p}{p} = \frac{d}{dx} H_B(x) \Big|_{x=p} \text{ comes from Taylor approximation}$$

Incorporate into  $O(1)$  by appealing to Taylor's again

# Compare approximation to earlier bounds



There is a matching converse:

$$\log M \leq n(1 - H_B(p)) + 0.5 \log n - \sqrt{nV} Q^{-1}(\epsilon) + O(1)$$

Capacity term

$V$  is called the channel “dispersion”

# Normal approximation BSC & DMCs

For the BSC we've seen ( $M^*$  is largest possible  $M$ ):

$$\log M^*(n, \epsilon) = nC - \sqrt{nV} Q^{-1}(\epsilon) + 0.5 \log n + O(1)$$

# Normal approximation BSC & DMCs

For the BSC we've seen ( $M^*$  is largest possible  $M$ ):

$$\log M^*(n, \epsilon) = nC - \sqrt{nV} Q^{-1}(\epsilon) + 0.5 \log n + O(1)$$

For DMCs w/ a unique capacity-achieving input dist. can show:

$$\log M^*(n, \epsilon) = nC - \sqrt{nV} Q^{-1}(\epsilon) + O(\log n)$$

where the channel “dispersion” can be calculated as

$$V = \text{var}_{p_X p_{Y|X}} [i(X, Y)]$$

Cap-achieving input dist

channel law

Information density

Compare order terms to see expansion a bit more exact for BSCs

# Recap: three decoders

## Bounded distance

- Easy to analyze & gave us theme of analysis, split into “outage” and “confusion” events
- Identified regimes of interest: fixed rate and fixed  $\Pr[\text{err}]$
- Allowed us to see scaling of threshold

## ML analysis for fixed rate (error exponents)

- Introduced coupling between “outage” & “confusion”
- Applied Chernoff bounds to each

## ML analysis for fixed $\Pr[\text{err}]$ (non-asymptotic)

- Finer analysis of coupled event
- Got to the “normal” approximation: use Berry-Esseen for “outage” and large deviation for “confusion”
- Understood how rate can approach capacity
- Normal approximation quite accurate even for small  $n$

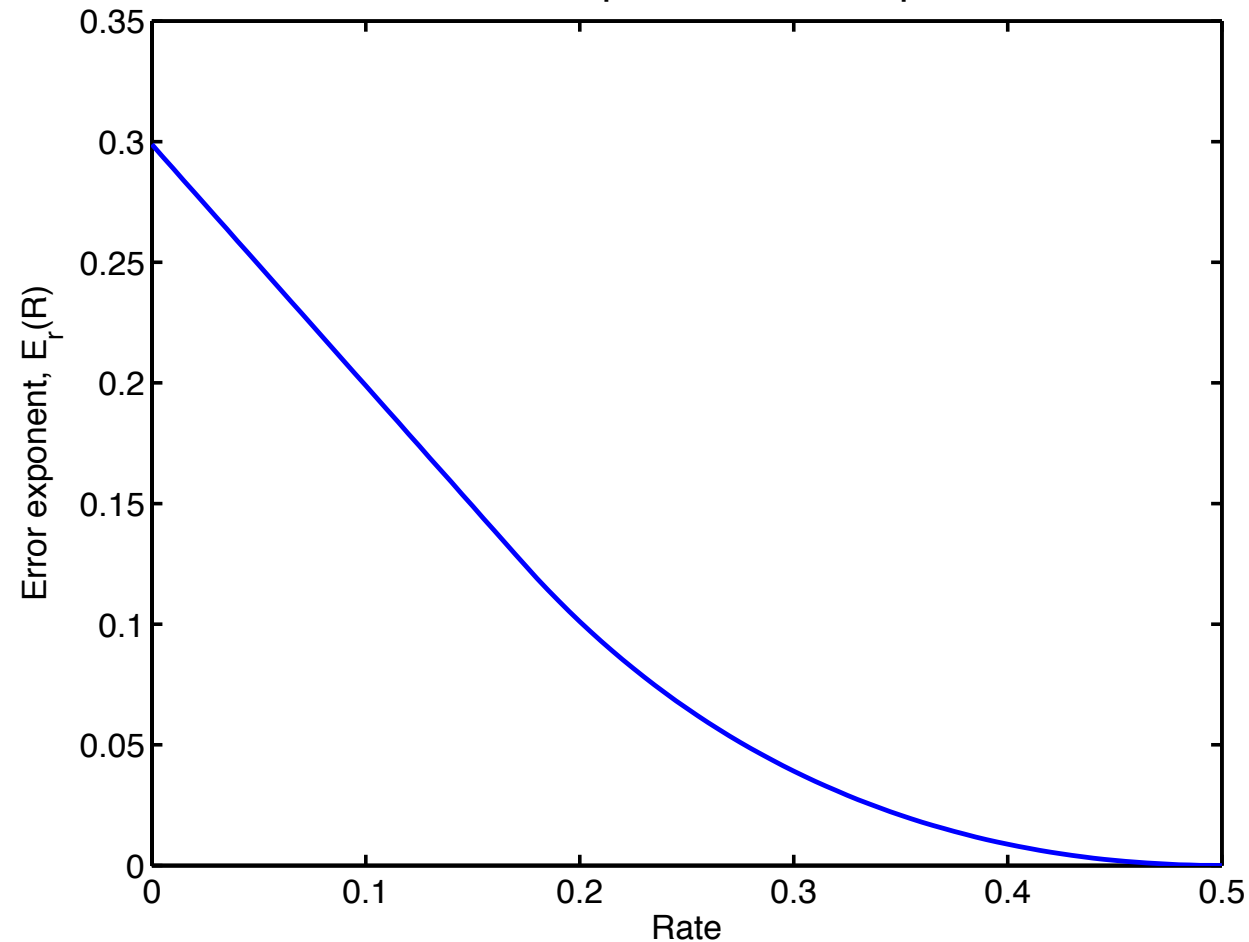
# Talk objectives: recap

- Beyond typicality decoding: Analyzed ML decoding
- As block length increases:
  - How quickly does error drop? Error exponents
  - How quickly do you approach capacity? Normal approx
- Intro to tools used to answer such questions:
  - Large deviations: Chernoff bounds
  - Gaussian approximations: Berry-Esseen
- When to use which type of tool & connections between: Chernoff when tail starts far from mean, Berry-Esseen when tail starts close to mean.
- Will try to illustrate general techniques and results in simplest illustrative context: BSC Showed general forms



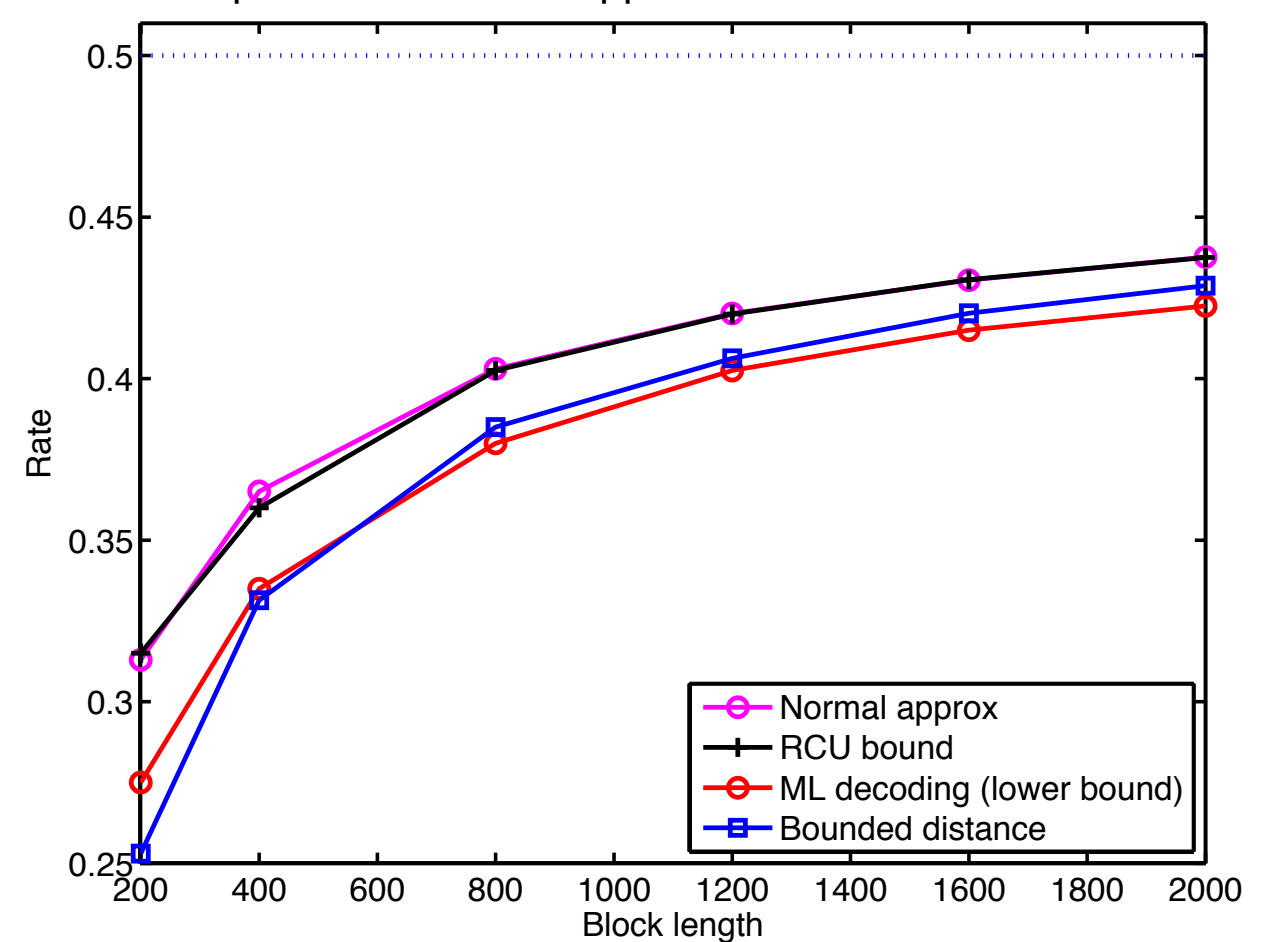
# Talk results: two interesting & connected problems

Plot of error exponent vs rate,  $p = 0.11$



Error exponents

Comparison of normal approximation with earlier results



Finite block length analysis