1. The exam has 8 questions with a total of 120 points. You have 3 hours to take the exam. Questions have different numbers of points so please allocate your time to each question accordingly. The exam has 20 pages, with the last two provided as space for longer answers.

2. All logarithms are base 2, unless otherwise stated.

3. Please write all answers in the designated area underneath the question. If you need more room for your answer, please indicate as such, and continue your response on pages 19 or 20. No additional pages will be allowed.

4. Scratch paper will be provided and collected at the end of the exam, but will not be graded.

5. All answers should be justified, unless otherwise stated.

6. The exam is closed book but you are allowed two double-sided sheets of notes. No other materials are allowed.

7. You can utilize results from lectures and homeworks without proof.

Good luck!

Name:

SUID:
1. (16 points) True-false questions (no explanations required). A correct answer gets 2 points. An incorrect answer gets 0 points. Leaving the answer blank gets 1 point. (Knowing you don’t know has value.)

Please shade your answer in completely to receive full credit.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>F</td>
</tr>
</tbody>
</table>

(a) The capacity of a discrete memoryless channel is bounded between 0 and 1.  
(b) For any continuous random variable $X$ and any function $f$, $h(f(X)) \leq h(X)$.  
(c) $h(X)$ indicates the expected number of bits we need to transmit to perfectly reconstruct $X$.  
(d) It is possible for 3 random variables $X, Y, Z$ to satisfy $H(Z|X) > 0, H(Z|Y) > 0$, and $H(Z|X, Y) = 0$.  
(e) Arithmetic coding has faster encoding than Huffman coding, but attains an asymptotically worse compression rate.  
(f) LZ-77 can achieve compression arbitrarily close to the entropy rate with a bounded window size.  
(g) Network coding is needed to achieve the min-cut bound for communicating from a single source to a single destination over a network with noiseless links.  
(h) If the pair $(x^n, y^n)$ is jointly typical with respect to an i.i.d. source then $(\pi(x^n), \pi(y^n))$ is also jointly typical for any permutation $\pi$.

Solution:

(a) False: consider a perfect channel that noiselessly transmits $\mathcal{X} = \mathcal{Y} = \{1, 2, 3, 4\}$, which has a capacity of 2 bits / channel use. This statement holds for binary input channels (note that $C \leq \log(\min(|\mathcal{X}|, |\mathcal{Y}|))$).

(b) False: consider $X \sim \text{Unif}([0,1])$, and $f(x) = 2x$. Then, $f(X) \sim \text{Unif}([0,2])$, and $h(X) = 0$ while $h\left(f(X)\right) = 1$.

(c) False: we need to transmit an infinite number of bits to perfectly reconstruct $X$.

(d) True: $X, Y$ i.i.d. Bern(1/2), $Z = X \oplus Y$.

(e) False: Arithmetic coding does yield faster encoding than Huffman, and attains the same asymptotic compression rate.
(f) False: LZ-77 requires an increasing window size to get arbitrarily close to capacity.

(g) False: for a single source with a single destination, a max-flow type algorithm is sufficient (no network coding required).

(h) True: since the source is i.i.d. \( p(x^n) = p(\pi(x^n)) \), \( p(y^n) = p(\pi(y^n)) \), and \( p(x^n, y^n) = \prod p(x_i, y_i) = p(\pi(x^n), \pi(y^n)) \).
2. (14 points) Recall the setting in class, where Alice wishes to communicate a message $W$ to Bob in the presence of an eavesdropping adversary Eve. Alice and Bob share some common randomness in the form of a random key $K$ (independent of $W$), which Eve does not know. To accomplish this task, Alice encodes the message $W$ to a ciphertext $C$ with the use of $K$.

(a) (2 points) Draw a system diagram summarizing the situation.

(b) (6 points) As we showed, if Alice and Bob wish to communicate without Eve learning anything about $M$, we must have $H(K) \geq H(W)$. Justify each of the below steps in a proof of this statement

\[
H(W) \overset{(a)}{=} H(W|C) \\
\overset{(b)}{=} H(K,W|C) - H(K|C,W) \\
\overset{(c)}{\leq} H(K|C) + H(W|C,K) \\
\overset{(d)}{\leq} H(K).
\]

(c) Now consider the one-time pad, where $W$ is uniform on $\{0,1\}^n$, $K$ is uniform on $\{0,1\}^n$ (independent of $W$), and $C = K \oplus W$ (entrywise xor).

i. (2 points) Show that the ciphertext $C$ is independent of the message $W$.

ii. (4 points) Explain why all the inequalities in part (b) are tight under the one-time pad.

**Solution:**

(a) As discussed in class, our model is below:
(b) a) comes from the fact that $I(W; C) = 0$, b) comes from chain rule of entropy, c) comes from nonnegativity of entropy and chain rule, d) from the fact that $W$ is a deterministic function of $K, C$ (decoding), and that conditioning reduces entropy.

One common mistake in this question was claiming that since $K$ and $W$ are independent, $H(K, W|C) = H(K|C) + H(W|C)$. However, this is not a valid application of the chain rule, as we are now conditioning on $C$. This is because, conditioned on $C$, $K$ and $W$ are no longer independent, and so we are left with $H(K, W|C) = H(K|C) + H(W|K, C)$.

(c) i. We see that for all values of $c, w$, we have that

$$P(C = c| W = w) = P(C = w \oplus k| W = w)$$

$$= P(K = c \oplus w| W = w)$$

$$= 2^{-n},$$

as $K$ is independent of $W$, and $K$ is uniform over $\{0, 1\}^n$.

ii. Inequality c is tight as $K = W \oplus C$, i.e. $K$ is a deterministic function of $(W, C)$. Inequality d is tight as $K$ and $C$ are independent by a symmetric argument to the previous part (b.i).
3. (12 points) Let $W(y|x)$ be a symmetric discrete memoryless channel with binary input $X \in \{0, 1\}$. The capacity-achieving distribution is uniform on $\{0, 1\}$ and the resulting capacity is $C(W)$ bits per channel use. Consider a polar code of block length $N = 4$ to be used over this channel. The construction of such a code is based on a linear transformation between four i.i.d. binary symbols $U_1, U_2, U_3, U_4$ and four channel input symbols $X_1, X_2, X_3, X_4$. The corresponding channel output symbols are $Y_1, Y_2, Y_3, Y_4$.

(a) (4 points) The polar code creates 4 effective channels $W^{++}, W^{+-}, W^{-+}, W^{--}$. Write down the capacities $C(W^{++}), C(W^{+-}), C(W^{-+}), C(W^{--})$ of these 4 channels in terms of information quantities involving the above random variables.

(b) (4 points) Among the capacities in part (a), state which ones are definitely greater than or equal to $C(W)$, which ones are definitely less than or equal to $C(W)$, and which ones can be greater than, equal to or less than $C(W)$. Justify your answers.

(c) (4 points) Suppose we would like to send at rate $3/4$ bits per channel use. Which of the $U_i$’s should be frozen, and which should be set as information bits? What are the dimensions of the resulting generator matrix mapping the information bits to the channel input symbols? Explain.

Solution:

(a) $$
C(W^{--}) = I(U_1; Y_1, Y_2, Y_3, Y_4) \\
C(W^{-+}) = I(U_2; Y_1, Y_2, Y_3, Y_4|U_1) \\
C(W^{+--}) = I(U_3; Y_1, Y_2, Y_3, Y_4|U_1, U_2) \\
C(W^{+++}) = I(U_4; Y_1, Y_2, Y_3, Y_4|U_1, U_2, U_3)
$$

(b) We know that the polarization operation increases the capacity for the plus channel and decreases the capacity for the minus channel, i.e. $C(W^+) \geq C(W) \geq C(W^-)$. Thus, we can similarly conclude that $C(W^{++}) \geq C(W^+) \geq C(W^-) \geq C(W^{--})$. However, for the other two synthetic channels, this depends on the actual channel at hand.

Hence, $C(W^{--}) \leq C(W)$, $C(W^{++}) \geq C(W)$, and the other 2 are channel dependent.

(c) We should freeze $U_1$ to be 0, and transmit on $U_2, U_3, U_4$, as $C_1$ is the smallest. The dimensions of the generator matrix are $4 \times 3$. 
4. (10 points) Consider the binary symmetric channel with crossover probability 0, i.e. a noiseless binary channel. Clearly, the channel capacity is one bit per channel use and communication at exactly this rate with zero probability of error can be achieved without coding at all. Let us see how a random code performs on this very simple channel.

A rate \( R \) random code is defined as follows: for each message \( W = m \in \{1, 2, \ldots, 2^{nR}\} \) we select a sequence \( X^n(m) \) uniformly at random in \( \{0, 1\}^n \), independently of all other \( X^n(m') \) for \( m' \neq m \). The received sequence through the noiseless channel is \( Y^n = X^n(m) \). The decoder knows the random code, and returns \( \hat{m} \) if there is a unique message \( \hat{m} \) that is coded to \( Y^n \), and declares an error otherwise.

(a) (4 points) Suppose \( 0 < R < 1 \). Compute the limit of the average error probability as \( n \to \infty \). Is the code reliable?

(b) (6 points) Now suppose \( R = 1 \), i.e. we want to send at exactly the capacity using a random code. As a function of \( n \), give an expression for the exact probability that the decoder declares an error. Compute the limit of this error probability as \( n \to \infty \). Is this code reliable? Compare this to the performance under no coding.

(Hint: you may find the below fact useful:

\[
 e^x = \lim_{m \to \infty} (1 + x/m)^m,
\]

where \( e \) is Euler’s constant, \( e \approx 2.718 \).

Solution:

(a) We bound this error probability as in class. For a fixed message \( m \), we have for a random code that

\[
\mathbb{P}(\text{error}) = \mathbb{E}_W [\mathbb{P}(\text{fail to decode } m)]
\]

\[
= \mathbb{E}_W \left[ \mathbb{P} \left( \bigcup_{m' \neq m} \{ X^n(m) = X^n(m') \} \right) \right]
\]

\[
\leq \left( 2^{nR} - 1 \right) \mathbb{E}_W [\mathbb{P}(X^n(m) = X^n(m'))]
\]

\[
\leq 2^{nR} \times 2^{-n} = 2^{n(R-1)}
\]

\[
\to 0.
\]

Where in the second line we picked some \( m' \neq m \) (since each \( m' \neq m \) has an identically distributed codeword). Since the error probability goes to 0, the code is reliable.

(b) We analyze this error probability, noting that the probability of an error is equal to the probability of a collision given that message \( m \) was transmitted, by symmetry.
We analyze the complement of this; the probability that there is no error. Since this is equivalent for all $m$, we analyze it for a fixed $m$.

\[
\mathbb{P}(\text{no error}) = \mathbb{P}(\cap_{m' \neq m} \{X^n(m) \neq X^n(m')\}) \\
= (\mathbb{P}(X^n(m) \neq X^n(m')))^{2^n-1} \\
= (1 - 2^{-n})^{2^n-1} \\
\to e^{-1}.
\]

Thus, the error probability of the random code is asymptotically $1 - 1/e$, which cannot be made arbitrarily small (is not decreasing with $n$). Hence, the code is not reliable.

We see that random codes perform worse than no coding for transmitting at rate $R = 1$, but for any $R < 1$ they can still achieve vanishing error probability.
5. (14 points) For $X, X_1, X_2, \ldots, X_n$ drawn i.i.d from a distribution $p(x)$, the typical set is defined as
\[ A^{(n)}_\epsilon = \left\{ x^n : \left| -\frac{1}{n} \log p(x^n) - H(X) \right| \leq \epsilon \right\} . \]
In this question we consider the case where $X \sim \text{Bern}(p)$.

(a) (3 points) Show that the condition that a sequence $x^n$ is in the typical set is solely a function of $k$, the number of ones in $x^n$, and not on any other aspects of $x^n$.

(b) (3 points) Express this condition on $k$ explicitly in terms of $n, \epsilon, p$.

(c) (2 points) State the Asymptotic Equipartition Property (AEP) for $X_1, \ldots, X_n$.

(d) (4 points) Using part (b) or otherwise, what does the AEP say about the fraction of 1’s in $X_1, \ldots, X_n$ for $n$ large when $p \neq 0.5$?

(e) (2 points) What can the AEP say about the fraction of 1’s in the case when $p = 0.5$?

Solution:

(a) Since the $X_i$ are i.i.d., we have that $p(x^n) = \prod_{i=1}^n p(x_i) = p^k(1-p)^{n-k}$, which is just a function of $k$, the number of heads in $x^n$, and not any other feature.

(b) Simplifying the condition for being in the typical set, we have that
\[
-\frac{1}{n} \log p(x^n) - H(X) = -\frac{k}{n} \log p - (1 - \frac{k}{n}) \log(1 - p) - (-p \log p - (1 - p) \log(1 - p))
= \left( \frac{k}{n} - p \right) \log \left( \frac{1-p}{p} \right) .
\]

Thus, we can see that in order for $x^n$ to be in the typical set, it must satisfy (assuming $p \neq 1/2$)
\[
\left| \left( \frac{k}{n} - p \right) \log \left( \frac{1-p}{p} \right) \right| \leq \epsilon \iff |k - np| \leq \frac{n \epsilon}{\log \left( \frac{1-p}{p} \right)} .
\]

(c) For $X^n$ i.i.d. $p(x)$, this states that
\[
\mathbb{P}(A^{(n)}_\epsilon) \to 1, \quad \text{as } n \to \infty ,
\]
or equivalently that
\[
\mathbb{P} \left( \left| -\frac{1}{n} \log p(X^n) - H(X) \right| \leq \epsilon \right) \to 1 .
\]
Note that $-\frac{1}{n} \log p(X^n)$ is a random variable, and so we need to specify in what sense it converges to $H(X)$ (in probability).
(d) For all \( p \neq 1/2 \) the AEP implies the WLLN, i.e. that

\[
P \left( \left| \frac{1}{n} \sum_{i=1}^{n} X_i - p \right| \leq \epsilon \right) \to 1.
\]

This stems from the fact that for the WLLN we wish to show that for any fixed \( p \neq 1/2 \) and \( \epsilon > 0 \), that for any \( \delta > 0 \) we have that there exists an \( n_0 \) such that \( |\sum_{i=1}^{n} X_i - np| \leq n\epsilon \) with probability at least \( 1 - \delta \) for all \( n \geq n_0 \). This is obtained from the AEP by selecting \( \epsilon' = \epsilon \log \left( \frac{1-p}{p} \right) \), and since we know that the AEP holds for all \( \epsilon > 0 \), we have that there exists an \( n'_0 \) such that for all \( n \geq n'_0 \)

\[
1 - \delta \leq P(A_n^{(n)}) = P \left( |k - np| \leq \frac{n\epsilon'}{\log \left( \frac{1-p}{p} \right)} \right) = P (|k - np| \leq n\epsilon
\]

(e) The AEP cannot say anything about the fraction of ones when \( p = 1/2 \). This is because we cannot pick any such \( \epsilon' > 0 \) for the statement to hold.

Intuitively, this is because for \( p \neq 1/2 \) we have that the probability of a sequence is a direct function of the number of heads \( k \), and so a sequence having a typical (expected) probability means that it has a typical number of ones. A conversion rate needs to be payed in the convergence guarantee \( (\log \frac{1-p}{p}) \). However, for \( p = 1/2 \), the probability of the sequence is independent of the number of heads, and so there is no equivalence between the two.
6. (16 points) One common drawback of compression schemes like Huffman coding is that they rely heavily upon knowing the distribution of the input. If this is not the case, we can end up with very bad compression schemes.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>1/2</td>
<td>1/4</td>
<td>1/8</td>
<td>1/8</td>
</tr>
<tr>
<td>$q(x)$</td>
<td>1/3</td>
<td>1/3</td>
<td>1/6</td>
<td>1/6</td>
</tr>
</tbody>
</table>

(a) (3 points) Given the distribution $p(x)$, construct an optimal prefix code for $X \sim p(x)$. How does the expected length of this code compare to $H(p)$?

(b) (2 points) If instead $X \sim q(x)$, what is the expected length of compressing $X$ according to the code from part (a) which assumes (incorrectly) that $X \sim p(x)$? How does this compare to the entropy $H(q)$?

(c) (4 points) If the assumed distribution $p(x)$ were perturbed by at most 1/8 (i.e. move at most 1/8 probability mass between symbols), what is the most the compression rate for $q(x)$ could increase by?

(d) (3 points) Now suppose $p(x)$ is any distribution on an arbitrary but finite alphabet such that $p(x)$ is a negative integer power of 2 for any $x$. How does the expected length of the optimal prefix code compare to $H(p)$?

(e) (4 points) If instead $X \sim q(x)$, where $q(x)$ is any arbitrary distribution on the same alphabet as $p$, what is the expected length of compressing $X$ according to the code from part (d) assuming $X \sim p(x)$? Express your answer in terms of information measures relating $p$ and $q$, and compare to the entropy $H(q)$.

Solution:

(a) $a \mapsto 0$, $b \mapsto 10$, $c \mapsto 110$, $d \mapsto 111$. The expected length of this code is thus $2/4 + 1/2 + 3 \times 3/16 + 3 \times 1/16 = 7/4$, which is equal to $H(p)$.

(b) The expected length is now $1/3 + 2/3 + 1 = 2$, which is larger than the entropy $H(q) = \frac{2}{3} \log 3 + \frac{1}{3} \log 6 = \log 3 + \frac{1}{3} \approx 1.93 < 2$.

(c) If our density for $q(x)$ is perturbed by moving all the mass away from symbol d (e.g. moving it to c), we would have the new distribution $q' = [1/2, 1/4, 1/4, 0]$. However, since $q(x)$ places no mass on the symbol d, an optimal code for $q(x)$ would assign it an infinitely long string, yielding an infinitely bad compression rate ($D(p||q) = +\infty$ due to support misalignment).

Another way of thinking about this is that there is an infinitely large alphabet of possible symbols, containing all other letters in the English alphabet, all letters in the Greek alphabet, etc. Since all of these have 0 probability of occurring under $p$, however, we can safely assign them an infinitely long bitstring. In this
question, we see that once the distribution $p$ places no mass on $d$, it now treats this symbol like all the other 0 probability symbols (e.g. $w, z, \alpha$), and does not assign a codeword to it (i.e., it has an infinitely long bitstring). This means that this new codebook will achieve an infinitely worse compression rate for $q(x)$.

(d) Since $p(x_i) = 2^{-\ell_i}$ for all $x_i$ for some $\ell_i$, we have that assigning codeword lengths of $\ell_i = -\log p(x_i)$ for symbol $x_i$ is possible (due to Kraft’s inequality) and optimal. An alternative justification for this is that the probability distribution is dyadic. This yields

$$\mathbb{E}_p \left[ \log \frac{1}{p(X)} \right] = H(p)$$

Thus, the expected length of the optimal prefix code is equal to $H(p)$.

(e) Since the codewords are of length $\ell_i = \log \frac{1}{p(x_i)}$, we have that

$$\mathbb{E}_q \left[ \log \frac{1}{p(X)} \right] = \mathbb{E}_q \left[ \log \frac{q(X)}{p(X)} - \log q(X) \right] = D(q||p) + H(q).$$

This is larger than $H(q)$ as the relative entropy $D(q||p)$ is nonnegative. In general, we only have that our expected length $L$ for $X \sim q(x)$ satisfies (when using codewords of length $\ell_i = \lceil -\log p(x_i) \rceil$)

$$H(q) + D(q||p) \leq L \leq H(q) + D(q||p) + 1.$$

In order to obtain equality, we need that $p(x)$ is dyadic.
7. (18 points) Consider the Gaussian channel:

\[ Y = X + Z, \]

where \( Z \sim N(0, \sigma^2) \) independent of \( X \), i.e. the pdf of \( Z \) is

\[ f(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{z^2}{2\sigma^2}}. \]

There is an average input power constraint \( P \). In the parts below, it is ok to leave your answers in terms of integrals that can be computed numerically.

(a) (1 point) Give a formula for the capacity \( C \) of this channel.

(b) In the Gaussian channel model, the input \( X \) can take on a continuum of values. In an actual digital communication system, signaling is done over a discrete set of input values rather than over a continuum. The simplest communication scheme uses binary signalling: \( X = a \) and \( X = -a \).

i. (1 point) What value should \( a \) take to meet the power constraint?

ii. (1 point) Write down the conditional distribution of the output \( Y \) given the discrete input.

iii. (3 points) Derive a formula for the maximum rate of reliable communication \( R_1 \) when the input is constrained by binary signalling.

iv. (2 points) How does \( R_1 \) compare to the capacity \( C \) of the Gaussian channel? Justify your answer.

(c) In the Gaussian channel model, the received signal \( Y \) is real-valued. In an actual digital communication system, the received signal is quantized into a discrete set of values. The simplest quantizer is a two-level quantizer \( Q \): \( Q(y) = -1 \) if \( y < 0 \) and \( Q(y) = +1 \) if \( y > 0 \).

i. (3 points) Derive an expression for the maximum rate of reliable communication \( R_2 \) under binary signalling and two-level quantization of the output.

ii. (2 points) How does \( R_2 \) compare to \( R_1 \) and to the channel capacity \( C \)? Justify.

iii. (2 points) What happens to the rate \( R_2 \) if the quantizer’s output is not plus or minus 1 but plus or minus 2?

(d) (3 points) Let \( \Delta > 0 \) and suppose the digital communication system uses signalling at levels \( X = i\Delta \), where \( i \) is any integer, and a quantizer that quantizes the output \( Y \) to the largest integer multiple of \( \Delta \) less than or equal to \( Y \). Let \( R(\Delta) \) be the maximum rate of reliable communication under these constraints. What is the limit of \( R(\Delta) \) as \( \Delta \to 0 \).

Solution:

(a) \( C = \frac{1}{2} \log(1 + P/\sigma^2). \)

(b) i. \( a = \sqrt{P}. \)
ii. \[
\begin{align*}
f_{Y|X=a}(y) &= \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(y-a)^2}{2\sigma^2}} \\
f_{Y|X=-a}(y) &= \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(y+a)^2}{2\sigma^2}}
\end{align*}
\]

iii. Channel capacity is
\[
R_1 = \max I(X;Y) = \max h(Y) - h(Y|X) = h(Y) - \frac{1}{2} \log(2\pi e\sigma^2)
\]
\[
= -\int_y \frac{1}{2\sqrt{2\pi}\sigma^2} \left( e^{-\frac{(y-a)^2}{2\sigma^2}} + e^{-\frac{(y+a)^2}{2\sigma^2}} \right) \log \left( \frac{1}{2\sqrt{2\pi}\sigma^2} \left( e^{-\frac{(y-a)^2}{2\sigma^2}} + e^{-\frac{(y+a)^2}{2\sigma^2}} \right) \right) dy \\
- \frac{1}{2} \log(2\pi e\sigma^2),
\]

as
\[
f_Y(y) = \frac{1}{2\sqrt{2\pi}\sigma^2} \left( e^{-\frac{(y-a)^2}{2\sigma^2}} + e^{-\frac{(y+a)^2}{2\sigma^2}} \right).
\]

iv. We have that \(R_1 < C\), as this \(Y\) stemming from a quantized \(X\) still has a second moment constraint; \(X\) and \(Z\) are independent, so \(E[Y^2] = E[X^2] + E[Z^2] = P + \sigma^2\). However, this \(Y\) is not Gaussian, as we would have had transmitting \(X \sim \mathcal{N}(0, P)\), and so while \(h(Y|X)\) has stayed the same \(h(Y)\) has decreased (as Gaussian maximizes entropy for a given second moment constraint).

(c) i. We have that
\[
R_2 = I(X;Q(Y)) = h(Q(Y)) - h(Q(Y)|X) = 1 - H(\Phi(-\sqrt{P/\sigma^2})),
\]
where \(\Phi\) is the CDF of a standard Gaussian.

This is because using symmetry we see that \(h(Q(Y)|X) = h(Q(Y)|X = a)\), and defining \(Z \sim \mathcal{N}(0, 1)\) we have that
\[
H(Q(Y)|X) = H(Q(Y)|X = a) = H_b(\mathbb{P}(Y > 0 | X = a)) = H_b(\mathbb{P}(Z > \sqrt{P/\sigma^2})) = H_b(\mathbb{P}(Z < -\sqrt{P/\sigma^2})).
\]

ii. \(R_2 \leq R_1 < C\) by data processing inequality.

iii. No change, mutual information is label invariant.
(d) We have that similarly to the homework,

\[
R(\Delta) = I(X^\Delta; Y^\Delta) \\
\rightarrow I(X; Y) \\
= \frac{1}{2} \log(1 + P/\sigma^2).
\]
8. (20 points) An i.i.d. Bern(1/2) source \( \{X_i\} \) is to be sent to the decoder. More concretely, a block of symbols \( X^n = (X_1, X_2, \ldots, X_n) \) is to be sent. Some side information \( Y^n \) is given to the decoder, where

\[
Y_i = \begin{cases} 
X_i & i \in S \\
e & i \not\in S
\end{cases}
\]

Here \( S \) is a size \( n/2 \) random subset of \( \{1, 2, \ldots n\} \), uniformly chosen from all such subsets. \( S \) is independent of \( X^n \). (We assume \( n \) is even.) “e” indicates an erasure.

(a) (2 points) Suppose \( n = 4 \) and \( X^4 = x^4 \), where \( x_1 = 0 \), \( x_2 = 1 \), \( x_3 = 1 \), \( x_4 = 0 \). Write down \( x^4 \) and all possible values of \( Y^4 \) given \( X^4 = x^4 \).

(b) (2 points) Compute \( H(X^4|Y^4 = y^4) \) for each of the values of \( Y^4 \) from part (a). Compute \( H(X^4|Y^4) \).

(c) (1 point) Without the side information \( Y^n \) at the decoder, how many bits need to be sent so that the decoder can reconstruct \( X^n \) exactly?

(d) (2 points) If the side information \( Y^n \) is available at the encoder as well as the decoder, how many bits are needed so that the decoder can reconstruct \( X^n \) exactly?

(e) We now seek an efficient scheme that can communicate at a rate arbitrarily close to the rate in part (d), but with side information known only at the decoder. Efficiency here means that the encoding and decoding complexity scales only polynomially with \( n \) rather than exponentially in \( n \).

i. (3 points) Recall that a rate \( k/m \) random linear code has codewords of the form \( Gu^k + v^m \), where \( u^k \) is a vector of \( k \) information bits, \( G \) is a random \( m \) by \( k \) matrix of i.i.d. Bern(1/2) entries, and \( v^m \) is a random \( m \)-dimensional vector of i.i.d. Bern(1/2) entries. \( G \) and \( v^m \) are independent of each other, and are further independent of the information bits \( u^k \). In a homework, we have shown that such a code can achieve rates arbitrarily close to the capacity of the BSC(\( p \)) channel. Explain why encoding can be done efficiently. In the special case of \( p = 0 \), explain why decoding can also be done efficiently at all rates strictly less than the capacity. (You can assume all linear algebraic operations over a finite field, such as matrix addition, matrix-vector multiplication, matrix-matrix multiplication, matrix inversion, solving a system of linear equations, etc, can be performed in time polynomial in the dimensions of the matrices and vectors.)

ii. (10 points) Using the previous part or otherwise, construct an efficient compression scheme (efficient encoding and decoding) that can achieve rates arbitrarily close to that in part (d) but with side information only available at the decoder. The error probability of reconstructing \( X^n \) should vanish as \( n \to \infty \).

Solution:
(a) \( x^4 = 0110, Y^4 \in \{ee10, e1e0, e11e, 0ee0, 0ee1, 01ee\} \).

(b) \( H(X^4|Y^4 = y^4) = 2 \) for all \( y^4 \), and so \( H(X^4|Y^4) = 2 \).

(c) \( H(X^n) = n \)

(d) \( I(X^n; Y^n) = H(X^n) - H(X^n|Y^n) = n - n/2 = n/2 \). No uncertainty about \( X_i \) on the coordinates where \( Y_i \) not erased, 1 bit of uncertainty when \( Y_i \) is erased, \( n/2 \) erasures.

(e) A. Encoding is performed via a matrix vector product and an addition, and so can be done in polynomial time. Decoding for \( p = 0 \) entails solving a linear system of equations, which can be performed in polynomial time (assuming the system is solvable, which happens whp if we are transmitting at rate less than capacity).

Note that random linear codes are always memory efficient; this question was focusing on computational efficiency.

B. We use the previous part, and construct a random linear code with \( m = n(1/2 + \epsilon), k = n \). The receiver receives these \( m \) linear combinations, as well as the vector \( Y^n \); this allows them to exactly solve for and remove the effects of these \( n/2 \) symbols from the linear code. This leaves us with an \( m \times (n/2) \) generator matrix \( G \) (with i.i.d. Bern(1/2) entries), which we observe noiselessly. Thus, using the previous part we can decode in linear time (encoding is again simply a matrix vector product and an addition).

Note that naively invoking Slepian-Wolf and utilizing a random hash function mapping \( X^n \to \{1, 2, \ldots, 2^{n/2}\} \) would not yield an efficient scheme. The decoder in this case would need to iterate over all \( x^n \) and determine whether they were both jointly typical with the observed \( y^n \), and had the same hash as was transmitted. The complexity of this unstructured method is exponential in \( n \) (and as such only received 2 points).