PROBLEM 1.

(a) By the chain rule and 'conditioning reduces entropy' we have $H(U^nV^n) = H(U^n) + H(V^n|U^n) = H(U^n) + \sum_{i=1}^n H(V_i|U^nV^{i-1}) \leq H(U^n) + \sum_{i=1}^n H(V_i|U_i)$.

(b) By Fano's inequality, for every i, $H(V_i|U_i) \leq h_2(p_i) + p_i \log(|\mathcal{U}| - 1)$. Therefore, $\frac{1}{n}\sum_{i=1}^n H(V_i|U_i) \leq \frac{1}{n}\sum_{i=1}^n h_2(p_i) + \left(\frac{1}{n}\sum_{i=1}^n p_i\right) \log(|\mathcal{U}| - 1)$. Since $h_2(\cdot)$ is a concave function, $\frac{1}{n}\sum_{i=1}^n h_2(p_i) \leq h_2(\frac{1}{n}\sum_{i=1}^n p_i)$.

(c) Using (a) and (b) we find $\frac{1}{n}H(V^n) \leq \frac{1}{n}H(U^nV^n) \leq \frac{1}{n}H(U^n) + h_2(\frac{1}{n}\sum_{i=1}^n p_i) + \left(\frac{1}{n}\sum_{i=1}^n p_i\right) \log(|\mathcal{U}| - 1)$.

Therefore, $\frac{1}{n}H(V^n) - \frac{1}{n}H(U^n) \leq h_2(\frac{1}{n}\sum_{i=1}^n p_i) + \left(\frac{1}{n}\sum_{i=1}^n p_i\right) \log(|\mathcal{U}| - 1)$. Taking the limits, since entropy is a continuous function, we get, $H_V - H_U \leq h_2(p) + p \log(|\mathcal{U}| - 1)$. The result follows by noting that the same argument hold if we swap U and V.

The problem shows that if two processes are close in the sense that the fraction of indices in which they disagree is small, then they must have a small difference in entropy rate — one can interpret this as a continuity property.

Problem 2.

(a) We partition the set $\mathcal{U}$ with subsets of size 2^j where $j \in J$. Then we map the elements of the subset of size 2^j to $\{0,1\}^j$. For every j, both the domain and the codomain have the same cardinality therefore we can find an injective map. Since we use a different codomain for every j the whole map c from $\mathcal{U}$ to $\{0,1\}^*$ is also injective.

(b) Suppose that $L = i$. There are 2^i many elements which get mapped to all possible length i binary sequences which all have the same probability. Therefore, conditioned on $L = i$ all elements of $\{0,1\}^i$ have equal probability 2^{-i}. But $p_{X^i}(x^i) = 2^{-i}$ for every $x^i \in \{0,1\}^i$ is the probability distribution of i i.i.d. $\{0,1\}$-valued random variables $X_1, \ldots, X_i$ with $\Pr(X_1 = 0) = \Pr(X_1 = 1) = \frac{1}{2}$.

(c) From (b) we see that conditioned on $L = i$, W is uniformly distributed on $\{0,1\}^i$, and thus $H(W|L = i) = i$. Consequently, $H(W|L) = \sum_{i \in J} \Pr(L = i) H(W|L = i) = \sum_{i \in J} \Pr(L = i) i = \mathbb{E}[L]$.

(d) Note that for any $j \in J$ $2^j \leq |\mathcal{U}|$. Therefore, all elements of J are less than or equal to $\log|\mathcal{U}|$. Therefore, L can take at most $1 + \log|\mathcal{U}|$ — the additional 1 because it can also take the value 0. Since the maximum entropy that a random variable over a set A can have is $\log|A|$, $H(L) \leq \log(1 + \log|\mathcal{U}|)$.

(e) Note that W is uniformly distributed on a set of $|\mathcal{U}|$ values and L is a function of W. Therefore, $\log|\mathcal{U}| = H(W) = H(W, L) = H(L) + H(W|L) = H(L) + \mathbb{E}[L] \leq \mathbb{E}[L] + \log(1 + \log|\mathcal{U}|)$, which gives us the required lower bound on $\mathbb{E}[L]$.

Thus, by the method outlined in (a), from a uniformly distributed U with entropy $\log|\mathcal{U}|$ we can generate, in expectation at least $\log|\mathcal{U}| - o(\log|\mathcal{U}|)$ i.i.d. and fair random bits. It is easy to see that no deterministic method can generate more than $\log|\mathcal{U}|$ such bits (and further thought reveals that equality in general is not possible).

Problem 3.

(a) $\mathbb{E}[\log f(U)] - D(p_U||p_V) = \mathbb{E}\left[\log f(U) - \log\left(\frac{p_U(U)}{p_V(U)}\right)\right] = \mathbb{E}\left[\log \frac{p_V(U)}{p_U(U)} f(U)\right] \leq \log \mathbb{E}\left[\frac{p_V(U)}{p_U(U)} f(U)\right] = \log \mathbb{E}[f(V)]$.

Where the last equality is because $\mathbb{E}[f(V)] = \sum_{u \in \mathcal{U}} p_V(u) f(u) = \sum_{u \in \mathcal{U}} p_U(u) \frac{p_V(u)}{p_U(u)} f(u)$.

(b) The only inequality that we used in part (a) is Jensen's inequality, which is tight if for every u, $\frac{p_V(u)}{p_U(u)} f(u)$ is a constant. Therefore, choosing $f(u) = \frac{p_U(u)}{p_V(u)}$ we have equality. Since the inequality holds for every nonnegative f and it is an equality for a specific nonnegative f the result follows.

(c) Note that $D(p_{\tilde{U}}||p_{\tilde{V}}) = \mathbb{E}[\log \tilde{f}(\tilde{U})] - \log \mathbb{E}[\tilde{f}(\tilde{V})]$ for some function $\tilde{f} : \tilde{\mathcal{U}} \to \mathbb{R}^+$. Let $f(\cdot) = \tilde{f}(g(\cdot))$.
Therefore, $\mathbb{E}[\log \tilde{f}(\tilde{U})] - \log \mathbb{E}[\tilde{f}(\tilde{V})] = \mathbb{E}[\log f(U)] - \log \mathbb{E}[f(V)] \leq D(p_U||p_V)$. Where the last inequality holds from part (a).

(d) Let $g(x, y) = \mathbb{1}_{[x=y]}$ be the indicator function which takes value 1 when $x = y$. This is the function $g(\cdot)$ defined on the domain of X, Y and X', Y' and it takes values in $\{0, 1\}$. Let $p_U = p_{XY}$ and $p_V = p_X p_Y$. Then using the notation of part (c), $\tilde{U}$ is a Bernoulli random variable with parameter p_e and $\tilde{V}$ is a Bernoulli random variable with parameter q_e. Consequently, using part (c) we get, $I(X;Y) = D(p_{XY}||p_X p_Y) = D(p_U||p_V) \geq D(p_{\tilde{U}}||p_{\tilde{V}}) = D_2(p_e||q_e) = p_e \log \frac{p_e}{q_e} + (1 - p_e) \log \frac{1-p_e}{1-q_e}$.

(e) Now suppose that X is uniformly distributed. Then X' is also uniformly distributed. Since Y' is independent of X', $\Pr(X' \neq Y') = \frac{|\mathcal{U}|-1}{|\mathcal{U}|}$. Therefore, by part (d)
$\log|\mathcal{U}| - H(X|Y) = H(X) - H(X|Y) = I(X;Y) \geq p_e \log p_e + (1 - p_e) \log(1 - p_e) + p_e \log \frac{|\mathcal{U}|}{|\mathcal{U}|-1} + (1 - p_e) \log|\mathcal{U}| = \log|\mathcal{U}| - h_2(p_e) - p_e \log(|\mathcal{U}| - 1)$. Subtracting $\log|\mathcal{U}|$ from all sides gives us,
$H(X|Y) \leq h_2(p_e) + p_e \log(|\mathcal{U}| - 1)$, which is exactly the Fano's inequality.

The result of parts (a) and (b) is known as the Donsker–Varadhan characterization of divergence. Part (c) is known as the data processing equality of divergence. Part (d) is a (slight) generalization of Fano's inequality.

PROBLEM 4.

(a) Note that the Lempel-Ziv algorithm parses the sequence u^∞ as $a, b, aa, ab, ba, bb, aaa, aab, \ldots$ which are the sequences we concatenated to construct u^∞ at the first place. By the time it reaches a segment of length n for the first time it will have seen all the length $n-1$ segments. For example, by the time it reaches 000 it will have already encoded $00, 01, 10, 11$. Therefore there will be at least 8 elements in its dictionary. Therefore, it will use at least n bits to describe any segment of length n. Therefore the bits/letter that the LZ algorithm produces must be greater than or equal to 1. That is, $\rho_{LZ}(u^\infty) \geq 1$.

We know that $\rho_{LZ}(u^\infty) \leq \rho_{FSM-IL}(u^\infty)$. Note that for the simple finite state machine which outputs 0 when it sees a and 1 when it sees b (which is a 1-state information lossless machine) the compressibility is 1. Therefore, $\rho_{FSM-IL}(u^\infty) \leq 1$. So we conclude that $\rho_{LZ}(u^\infty) = 1$.

(b) If the process is stationary yes, as we have seen in the class. However, If U^∞ takes the value u^∞ deterministically, it will have entropy rate $\lim_{n\to\infty} \frac{1}{n} H(X^n) = \lim_{n\to\infty} 0 = 0$, yet the LZ compressibility of the sequence is 1. Therefore, the answer to the question is no.

(c) The LZ algorithm will parse v^∞ as $x_1, x_2, x_3, \ldots$. By the time it encodes x_m it will have output at most $m \log m$ many bits. The input length that it encodes with m iterations is $\sum_{i=1}^{m} i = \Theta(m^2/2)$. Therefore the compressibility of the sequence v^∞ is $\lim\limits_{m\to\infty} \dfrac{2m \log m}{m^2} = 0$.

(d) We fix an $s-$state information lossless finite-state machine. Note that whichever state the machine is at when it begins to encode the segment x_n, it has to produce at least $m \log \dfrac{m}{8s^2}$ many bits when the encoding of the segment x_n is finished. If we take the parsing as the parsing that LZ produces for the sequence x_n. We have that $\lim\limits_{n\to\infty} \dfrac{1}{n} \text{length}(y_n) \geq \lim\limits_{n\to\infty} \dfrac{m_{LZ}(u^n)}{n} \log \dfrac{m_{LZ}(u^n)}{8s^2} = \lim\limits_{n\to\infty} \dfrac{m_{LZ}(u^n)}{n} \log m_{LZ}(u^n) = \lim\limits_{n\to\infty} \dfrac{1}{n} \text{length}(LZ(u^n)) = \rho_{LZ}(u^\infty) = 1$. The result follows from the definition of the limit.

(e) Note that for any fixed information lossless finite state machine, large enough n, say, for all $n \geq n_0(\epsilon)$, $\text{length}(y_n) \geq (1-\epsilon)n = (1-\epsilon)\,\text{length}(x_n)$. Then, $\text{length}(y_1 y_2 \ldots y_n) = \sum_{i=1}^{n} \text{length}(y_i) = \sum_{i=1}^{n_0-1} \text{length}(y_i) + \sum_{i=n_0}^{n} \text{length}(y_i) \geq (1-\epsilon) \sum_{i=n_0}^{n} \text{length}(x_i)$. Therefore, $\dfrac{\text{length}(y_1 y_2 \ldots y_n)}{\text{length}(x_1 x_2 \ldots x_n)} \geq (1-\epsilon) \dfrac{\text{length}(x_{n_0} x_{n_0+1} \ldots x_n)}{\text{length}(x_1 x_2 \ldots x_n)}$. Upon observing that $\lim_{n\to\infty} \dfrac{\text{length}(x_{n_0} \ldots x_n)}{\text{length}(x_1 \ldots x_n)} = 1$ we conclude $\lim\limits_{n\to\infty} \dfrac{\text{length}(y_1 y_2 \ldots y_n)}{\text{length}(x_1 x_2 \ldots x_n)} \geq 1 - \epsilon$. This is true for any ϵ. Therefore, $\lim\limits_{n\to\infty} \dfrac{\text{length}(y_1 y_2 \ldots y_n)}{\text{length}(x_1 x_2 \ldots x_n)} \geq 1$. The limit we wrote is also the compressibility $\rho_{c_M}(v^\infty)$. Therefore, $\rho_{c_M}(v^\infty) \geq 1$.

The IL-FSM we have chosen was arbitrary therefore the finite state compressibility of v^∞ (taking infimum over all IL FSM's) is also greater than or equal to 1. It cannot be larger than 1 because of the simple FSM described in part (a). Therefore, $\rho_{FSM-IL}(v^\infty) = 1$.

In (c–e) we see that the inequality in $\rho_{LZ} \leq \rho_{FSM-IL}$ can be strict. In (a–b) we see that LZ need not always compress to the entropy rate for non-stationary processes.