[redacted] [redacted]

1.(a) No #

Consider the following example:

X, Y_1, and Y_2 are 3 RVs on $\{0,1\}$ with joint pmf

$$\begin{cases} P((X,Y_1,Y_2) = (1,0,0)) = \frac{1}{4} \\ P((X,Y_1,Y_2) = (0,1,0)) = \frac{1}{4} \\ P((X,Y_1,Y_2) = (0,0,1)) = \frac{1}{4} \\ P((X,Y_1,Y_2) = (1,1,1)) = \frac{1}{4}. \end{cases}$$

Or equivalently, $(X, Y_1, Y_2) \sim \mathrm{Unif}(\{(1,0,0), (0,1,0), (0,0,1), (1,1,1)\})$.

It immediately follows that $H(X, Y_1, Y_2) = \log_2 4 = 2$.

Note that X, Y_1, and Y_2 have a marginal distribution $\mathrm{Ber}(\frac{1}{2})$, and thus $H(X) = H(Y_1) = H(Y_2) = 1$.

In addition, note that (X, Y_1), (X, Y_2), (Y_1, Y_2) have the same distribution $\mathrm{Unif}(\{(0,0), (0,1), (1,0), (1,1)\})$.

Thus $H(X, Y_1) = H(X, Y_2) = H(Y_1, Y_2) = \log_2 4 = 2$.

Now we have

$$I(X;Y_1) = H(X) + H(Y_1) - H(X,Y_1) = 1+1-2=0,$$

$$I(X;Y_2) = H(X) + H(Y_2) - H(X,Y_2) = 1+1-2=0, \text{ but}$$

$$I(X;Y_1,Y_2) = H(X) + H(Y_1,Y_2) - H(X,Y_1,Y_2)$$

$$= 1+2-2=1 \neq 0,$$

2. No #

We can take arbitrary iid X and Y_1, say $X, Y_1 \overset{iid}{\sim} Ber(\frac{1}{2})$, and then set $Y_2 = Y_1$.

Then $I(X;Y_1) = I(X;Y_2) = 0$ since X and Y_1 are independent and $Y_2 = Y_1$.

But $I(Y_1, Y_2) = I(Y_1, Y_1) = H(Y_1) - H(Y_1|Y_1)$

$$= H(Y_1) = 1 \neq 0.$$

2-(a) We calculate $H(X,Z|Y)$ in two different ways:

① By chain rule of conditional entropy, we have

$$H(X,Z|Y) = H(Z|Y) + H(X|Y,Z) \quad \cdots \quad (2\text{-}1)$$

② Since $X \to Y \to Z$, we have X and Z are conditionally independent given Y. Thus,

$$\begin{aligned} H(X,Z|Y) &= \sum_y P_Y(y)\, H(X,Z|Y=y) \\ &= \sum_y P_Y(y) \big(H(X|Y=y) + H(Z|Y=y)\big) \\ &= H(X|Y) + H(Z|Y), \quad \cdots \quad (2\text{-}2) \end{aligned}$$

Comparing (2-1) and (2-2), we have $H(X|Y) = H(X|Y,Z)$.

Q.E.D.

(b) From (a) and the fact that conditioning reduces entropy, we have

$$H(X|Y) = H(X|Y,Z) \le H(X|Z).$$

Q.E.D.

(c) From (b), we have

$$I(X;Y) - I(X;Z)$$
$$= (H(X) - H(X|Y)) - (H(X) - H(X|Z))$$
$$= H(X|Z) - H(X|Y) \geq 0.$$

Thus $I(X;Y) \geq I(X,Z)$. QED.

(d) From (a), we have

$$I(X;Z|Y) = H(X|Y) - H(X|Y,Z) = 0.$$ QED.

3. Following the terminology in class, we seek to prove that for any given collection of RVs $X_1, \cdots, X_n$, the following set function g is submodular:

$$g: 2^N \to \mathbb{R}_{\geq 0}$$

$$g(\{i_1, i_2, \cdots, i_\ell\}) = H(X_{i_1}, X_{i_2}, \cdots, X_{i_\ell}).$$

More precisely, we seek to show for any S_1 and S_2 such that

$$S_1 = \{i_1, \cdots, i_\ell\} \subseteq S_2 = \{i_1, \cdots, i_\ell, i_{\ell+1}, \cdots i_m\} \subseteq \{1, \cdots, n\}$$

and for any $j \in \{1, \cdots, n\} \setminus S_2$, we have

$$g(S_1 \cup \{j\}) - g(S_1) \geq g(S_2 \cup \{j\}) - g(S_2),$$

or equivalently

$$H(X_{i_1}, \cdots, X_{i_\ell}, X_j) - H(X_{i_1}, \cdots, X_{i_\ell})$$
$$\geq H(X_{i_1}, \cdots, X_{i_m}, X_j) - H(X_{i_1}, \cdots, X_{i_m}).$$

$$\cdots (3\text{-}1)$$

Note that by the chain rule of entropy, we have

$$H(X_{i_1}, \dots, X_{i_l}, X_j) - H(X_{i_1}, \dots, X_{i_l})$$
$$= H(X_j | X_{i_1}, \dots, X_{i_l}). \quad \cdots (3\text{-}2)$$

And similarly,

$$H(X_{i_1}, \dots, X_{i_m}, X_j) - H(X_{i_1}, \dots, X_{i_m})$$
$$= H(X_j | X_{i_1}, \dots, X_{i_m}) \quad \cdots (3\text{-}3).$$

Furthermore, since conditioning reduces entropy, we have

$$H(X_j | X_{i_1}, \dots, X_{i_m})$$
$$= H(X_j | X_{i_1}, \dots, X_{i_l}, X_{i_{l+1}}, \dots, X_{i_m})$$
$$\leq H(X_j | X_{i_1}, \dots, X_{i_l}). \quad \cdots (3\text{-}4)$$

Combining (3-2), (3-3) and (3-4) proves (3-1). QED

4.(a) I write $\mathbb{1}_{\{\cdot\}}$ for the indicator function instead to avoid confusion with mutual information.

Following the hint, define $\hat{P}_n'$ by

$$\hat{P}_n'(x) = \frac{1}{n}\sum_{j=n+1}^{2n} \mathbb{1}_{\{X_j = x\}} \quad \text{for each } x \in \mathcal{X}.$$

Then we have for each $x \in \mathcal{X}$ that

$$\begin{aligned}
\hat{P}_{2n}(x) &= \frac{1}{2n}\sum_{i=1}^{2n} \mathbb{1}_{\{X_i = x\}} \\
&= \frac{1}{2n}\left(\sum_{i=1}^{n} \mathbb{1}_{\{X_i = x\}} + \sum_{j=n+1}^{2n} \mathbb{1}_{\{X_j = x\}}\right) \\
&= \frac{1}{2}\hat{P}_n + \frac{1}{2}\hat{P}_n'.
\end{aligned}$$

Thus, by the convexity of KL divergence, we have

$$\begin{aligned}
D(\hat{P}_{2n} \| P) &= D\left(\tfrac{1}{2}\hat{P}_n + \tfrac{1}{2}\hat{P}_n' \,\|\, P\right) \\
&\le \frac{1}{2}D(\hat{P}_n \| P) + \frac{1}{2}D(\hat{P}_n' \| P).
\end{aligned}$$

$$\cdots (4\text{-}1)$$

Taking expectation on both sides of (4-1), we have

$$\mathbb{E}[D(\hat{P}_{2n} \| P)]$$

$$\leq \frac{1}{2}\mathbb{E}[D(\hat{P}_n \| P)] + \frac{1}{2}\mathbb{E}[D(\hat{P}_n' \| P)]. \quad \cdots (4\text{-}2)$$

Finally, note that since $(X_i)_{i \geq 1}$ are iid, $\hat{P}_n'$ has the same distribution as $\hat{P}_n$, and thus (4-2) becomes

$$\mathbb{E}[D(\hat{P}_{2n} \| P)] \leq \mathbb{E}[D(\hat{P}_n \| P)]. \qquad \text{QED.}$$

(b) Following the hint, define for each $k \in \{1, \cdots, n\}$ and for each $x \in \mathcal{X}$ that

$$\hat{P}_n^{(k)}(x) = \frac{1}{n-1} \sum_{\substack{i=1 \\ i \neq k}}^{n} \mathbb{1}_{\{X_i = x\}}.$$

Then, we have for each $x \in \mathcal{X}$ that

$$\sum_{k=1}^{n} \hat{P}_n^{(k)}(x) = \frac{1}{n-1} \sum_{k=1}^{n} \sum_{\substack{i=1 \\ i \neq k}}^{n} \mathbb{1}_{\{X_i = x\}}$$

(switch summation)

$$= \frac{1}{n-1} \sum_{i=1}^{n} \sum_{\substack{k=1 \\ k \neq i}}^{n} \mathbb{1}_{\{X_i = x\}}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} (n-1)\, \mathbb{1}_{\{X_i = x\}}$$

$$= n \hat{P}_n(x).$$

That is, we have

$$\hat{P}_n = \sum_{k=1}^{n} \frac{1}{n} \hat{P}_n^{(k)}.$$

Then, by convexity of $D(\cdot\|\cdot)$ again, we have

$$D(\hat{P}_n \| P) = D\left(\sum_{k=1}^{n} \frac{1}{n} \hat{P}_n^{(k)} \| P\right)$$

$$\leq \sum_{k=1}^{n} \frac{1}{n} D\left(\hat{P}_n^{(k)} \| P\right)$$

$$= \frac{1}{n} \sum_{k=1}^{n} D\left(\hat{P}_n^{(k)} \| P\right). \quad \cdots (4\text{-}3).$$

Finally, by the iid property of $(X_i)_{i \geq 1}$ again, all the $\hat{P}_n^{(k)}$ have the same distribution as $\hat{P}_n^{(n)} = \hat{P}_{n-1}$, and thus taking expectation on both sides of (4-3) yields

$$\mathbb{E}[D(\hat{P}_n \| P)] \leq \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}[D(\hat{P}_{n-1} \| P)]$$

$$= \mathbb{E}[D(\hat{P}_{n-1} \| P)].$$

QED.

5. (a) $H(X) = -0.6 \log 0.6 - 0.4 \log 0.4 \simeq \underline{0.9710}$ #

(b) A sequence x^n falls in $A_\varepsilon^{(n)}$ if and only if

$$H(X) - \varepsilon \leq -\frac{1}{n} \log p(x^n) \leq H(X) + \varepsilon.$$

Putting $H(X) = 0.9710$ and $\varepsilon = 0.1$ gives

$$0.8710 \leq -\frac{1}{n} \log p(x^n) \leq 1.0710. \quad \cdots \quad (5\text{-}1)$$

According to the table, the sequences with $k = 11, \cdots, 19$ one's satisfy (5-1).

Thus, $A_{0.1}^{(25)}$ contains <u>the sequences with $k=11, \cdots, 19$ ones.</u> #

The probability of this set can be found by

$$\sum_{k=11}^{19} \binom{n}{k} p^k (1-p)^{1-k}, \text{ where } p = 0.6 \text{ and } n = 25.$$

By table, this value is approximately $\underline{0.9362}$ #.

The cardinality of this set can be found by

$$\sum_{k=11}^{19} \binom{n}{k}, \text{ which is } \underline{26366510}\ \#.$$

(c) Note that $P(X=1)=0.6 > P(X=0)=0.4$.

Therefore, this smallest set with prob. 0.9, denoted as B, should contain as many 1's as possible. More explicitly, B should contain all sequences with 25 1's, and then 24 1's, and so on, until the probability of B is no less than 0.9.

That is, one seek to find k_0 s.t.

$$\sum_{k=k_0}^{25} \binom{25}{k} p^k (1-p)^{n-k} < 0.9 \text{ but}$$

$$\sum_{k=k_0-1}^{25} \binom{25}{k} p^k (1-p)^{n-k} \geq 0.9.$$

A numerical computation shows $k_0 = 13$, and thus B contains all the sequences with $13, 14, \cdots, 25$ 1's.

The probability of these sequences is

$$\sum_{k=13}^{25} \binom{25}{k} p^k (1-p)^{n-k} \approx 0.8462.$$

The rest probability $P_0 = 0.9 - 0.8462 = 0.0538$ can be fulfilled by collecting $n_0 = \left\lceil \frac{0.0538}{p^{12}(1-p)^{13}} \right\rceil \approx 3680673$

sequences with 12 1's into B.

Therefore, the number of elements in B is

$$\sum_{k=13}^{25} \binom{25}{k} + n_0 \approx \underline{20457889} \#.$$

(d) The intersection contains all the sequences of 13, ..., 19 1's and n_0 sequences of 12 1's.

Thus there are

$$\sum_{k=13}^{19} \binom{25}{k} + n_0 \approx \underline{20389483} \#$$ elements in the intersection.

The probability is

$$\sum_{k=13}^{19} \binom{25}{k} p^k (1-p)^{n-k} + p_0 \approx \underline{0.8706} \#.$$

6. Let $t_n := \frac{x_n}{y_n}$.

By assumption, $\sum_{n=1}^{\infty} t_n$ converges, and denote this limit to be $A \in \mathbb{R}$.

Now we seek to prove

$$\frac{1}{y_n}\sum_{i=1}^{n} x_i = \frac{1}{y_n}\sum_{i=1}^{n} y_i t_i \to 0 \text{ as } n \to \infty. \quad \cdots (6\text{-}1).$$

Define $S_n = \begin{cases} \sum_{i=1}^{n} t_i, & n \geq 1 \\ 0 & , n=0 \end{cases}$, and then

$t_n = S_n - S_{n-1}$ for $n \geq 1$.

Now, by summation by part, we have

$$\frac{1}{y_n}\sum_{i=1}^{n} y_i t_i = \frac{1}{y_n}\sum_{i=1}^{n} y_i (S_i - S_{i-1})$$

$$= \frac{1}{y_n}\left(\sum_{i=1}^{n} y_i S_i - \sum_{i=1}^{n} y_i S_{i-1}\right)$$

$$= \frac{1}{y_n}\left(y_n S_n + \sum_{i=1}^{n-1} y_i S_i - \sum_{i=2}^{n} y_i S_{i-1}\right)$$

$$= \frac{1}{y_n}\left(y_n s_n + \sum_{i=1}^{n-1} y_i s_i - \sum_{i=1}^{n-1} y_{i+1} s_i\right)$$

$$= s_n + \frac{1}{y_n}\sum_{i=1}^{n-1} s_i (y_i - y_{i+1})$$

$$= s_n - \frac{1}{y_n}\sum_{i=1}^{n-1} s_i (y_{i+1} - y_i) \quad \cdots (6\text{-}2)$$

Taking absolute value on (6-2), by triangle inequality we have

$$\left|\frac{1}{y_n}\sum_{i=1}^{n} t_i y_i\right| = \left|s_n - \frac{1}{y_n}\sum_{i=1}^{n-1} s_i (y_{i+1} - y_i)\right|$$

$$\leq |s_n - A| + \left|\frac{1}{y_n}\sum_{i=1}^{n-1} s_i (y_{i+1} - y_i) - A\right|$$

$$= |s_n - A| + \left|\frac{1}{y_n}\sum_{i=1}^{n-1} s_i (y_{i+1} - y_i) - \frac{1}{y_n}\sum_{i=1}^{n-1} A (y_{i+1} - y_i)\right|$$

$$= |s_n - A| + \left|\frac{1}{y_n}\sum_{i=1}^{n-1} (s_i - A)(y_{i+1} - y_i)\right|$$

$$\leq |s_n - A| + \frac{1}{|y_n|}\sum_{i=1}^{n-1} |s_i - A|\,|y_{i+1} - y_i|$$

$$= |S_n - A| + \frac{1}{|y_n|} \sum_{i=1}^{n-1} |S_i - A| (y_{i+1} - y_i) \cdots (6\text{-}3)$$

where the last equality holds since $(y_i)_{i\geq 1}$ is increasing.

Now let $\varepsilon > 0$.

Since $S_n = \sum_{i=1}^{n} t_i \xrightarrow{n\to\infty} A$ by assumption, $\exists N_1 \in \mathbb{N}$ s.t.

$\forall n \geq N_1$, we have $|S_n - A| < \frac{\varepsilon}{3}$

In addition, since $y_n \xrightarrow{n\to\infty} \infty$ by assumption, $\exists N_2 \in \mathbb{N}$

s.t. $\forall n \geq N_2$, we have $y_n > 0$ and

$$\frac{1}{y_n} \sum_{i=1}^{N_1 - 1} |S_i - A| (y_{i+1} - y_i) < \frac{\varepsilon}{3}$$

Now for $n \geq N_3 := \max(N_1, N_2)$, we have from (6-3) that

$$\left| \frac{1}{y_n} \sum_{i=1}^{n} t_i y_i \right| \leq \frac{\varepsilon}{3} + \frac{1}{y_n} \sum_{i=1}^{N_1 - 1} |S_i - A| (y_{i+1} - y_i)$$

$$+ \frac{1}{y_n} \sum_{i=N_1}^{n-1} |S_i - A| (y_{i+1} - y_i)$$

$$\leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{1}{y_n} \sum_{i=N_1}^{n-1} \frac{\varepsilon}{3} (y_{i+1} - y_i)$$

$$= \frac{2}{3}\varepsilon + \frac{\varepsilon}{3}\left(1 - \frac{y_{N_1}}{y_n}\right)$$

$$\leq \frac{2}{3}\varepsilon + \frac{\varepsilon}{3} = \varepsilon. \quad \cdots (6-4)$$

Since ε is arbitrary, from (6-4) we deduce that

$$\left|\frac{1}{y_n}\sum_{i=1}^{n} x_i\right| = \left|\frac{1}{y_n}\sum_{i=1}^{n} t_i y_i\right| \xrightarrow{n\to\infty} 0, \text{ and}$$

thus $\frac{1}{y_n}\sum_{i=1}^{n} x_i \xrightarrow{n\to\infty} 0.$

QED

7.(a) Define

$$(P_1, \cdots, P_7) = (0.05, 0.08, 0.13, 0.09, 0.30, 0.20, 0.15).$$

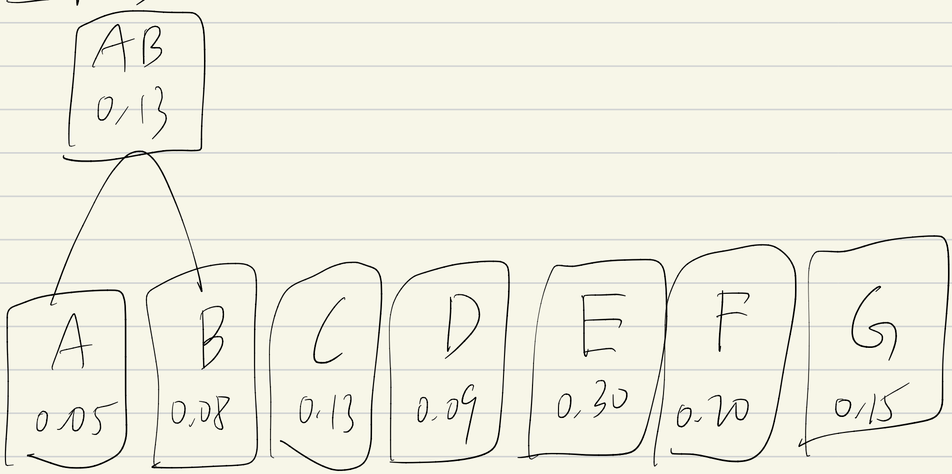
Then $H(X) = -\sum_{i=1}^{7} P_i \log P_i \approx \underline{2.5989}$ #

(b) If there is no probability known, encoding 7 symbols needs a minimum of $\lceil \log_2 7 \rceil = 3$ bits.

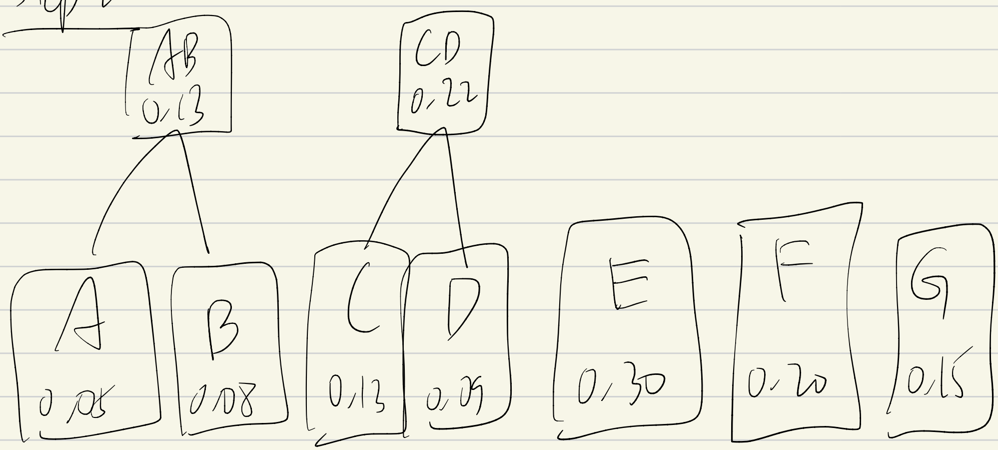
A fixed-length code can be constructed as follows:

$A \to 000$
$B \to 001$
$C \to 010$
$D \to 011$
$E \to 100$
$F \to 101$
$G \to 110$ #

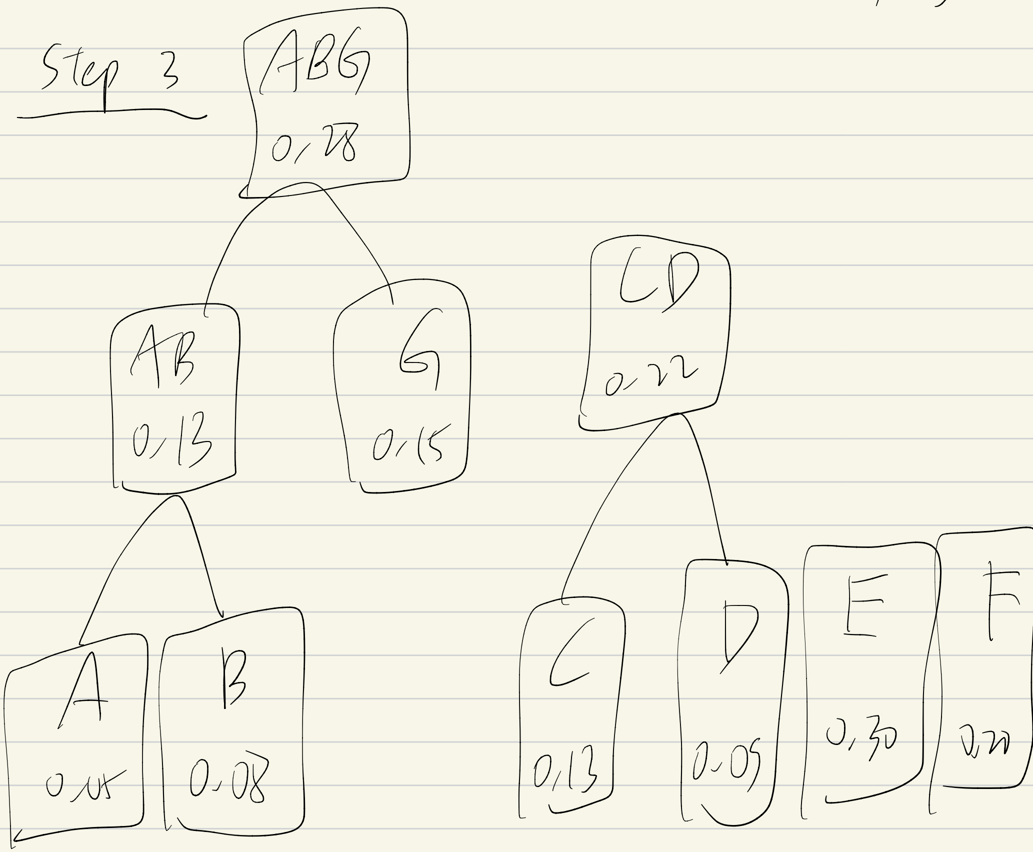
(c) Step 1



Step 2

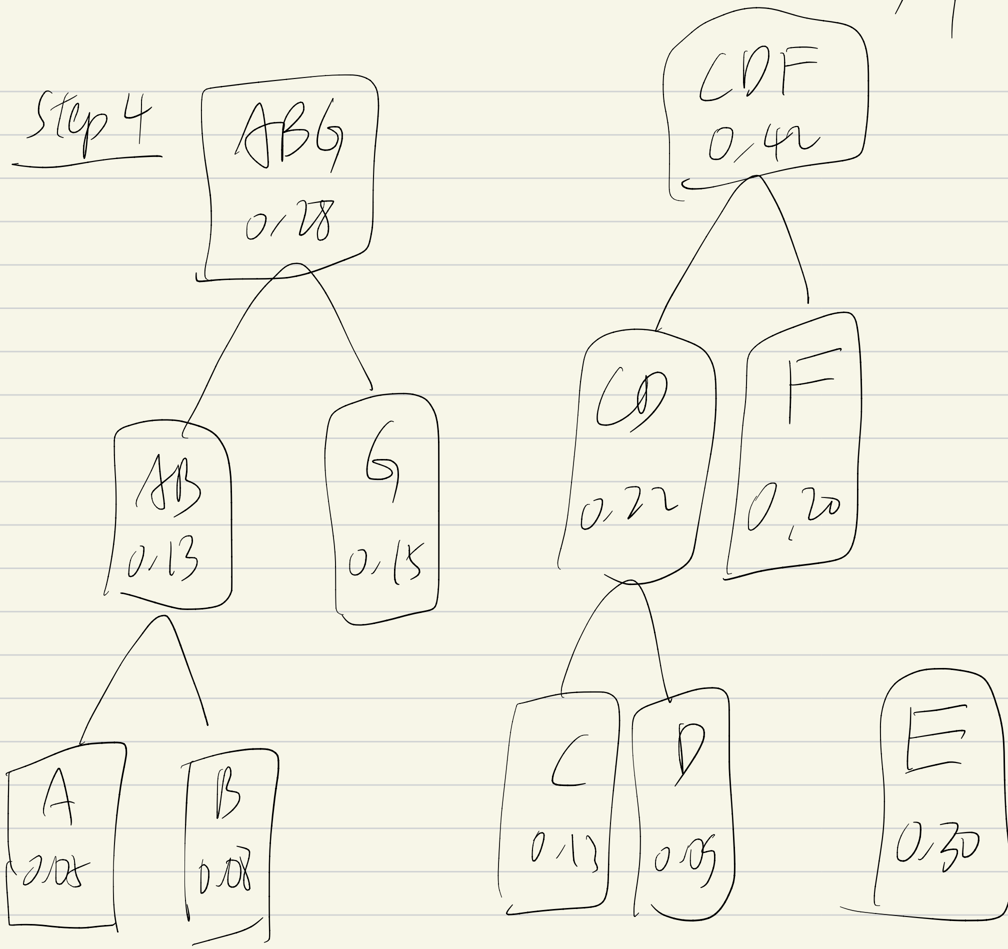


Step 3



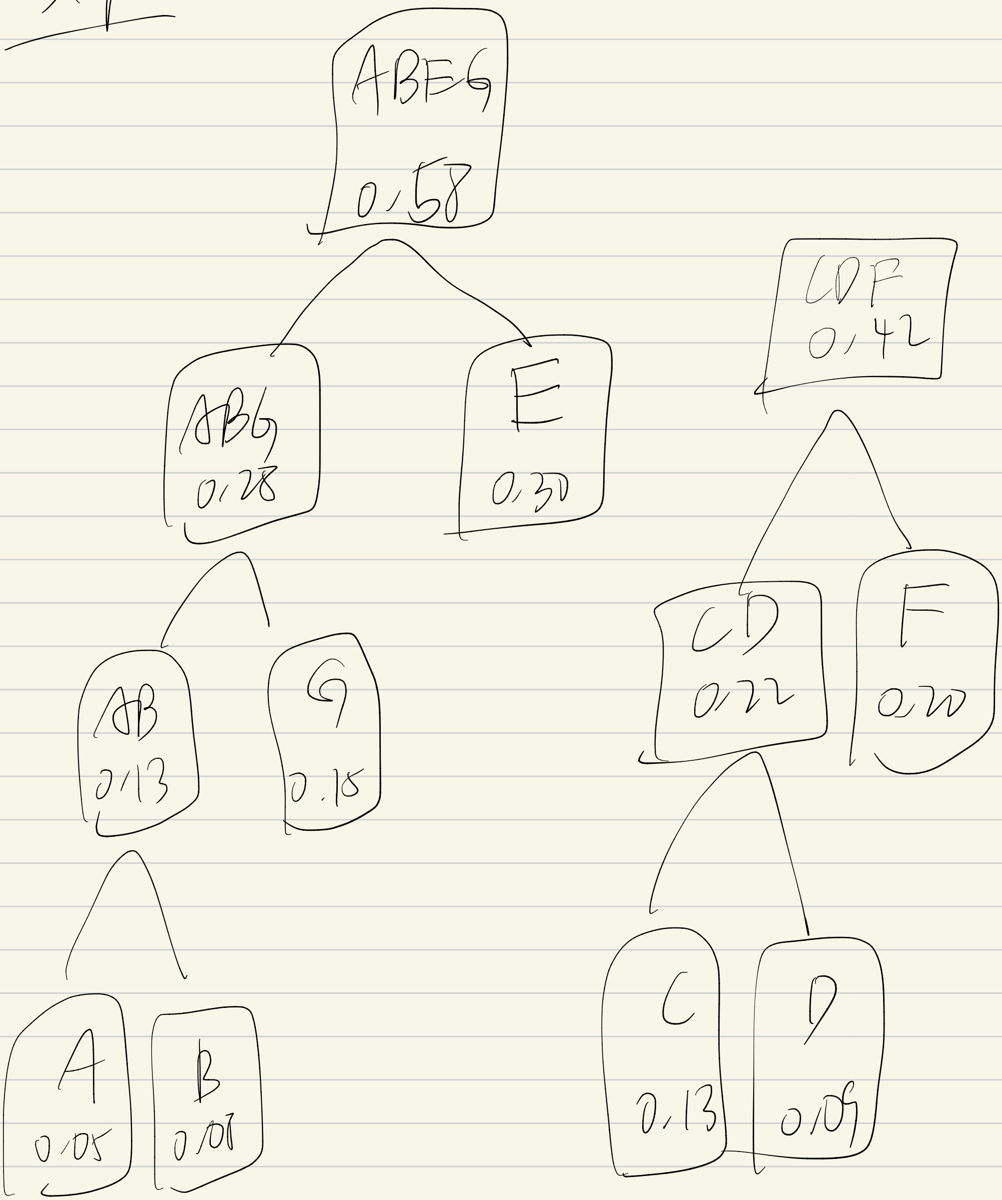
7-4

Step 4



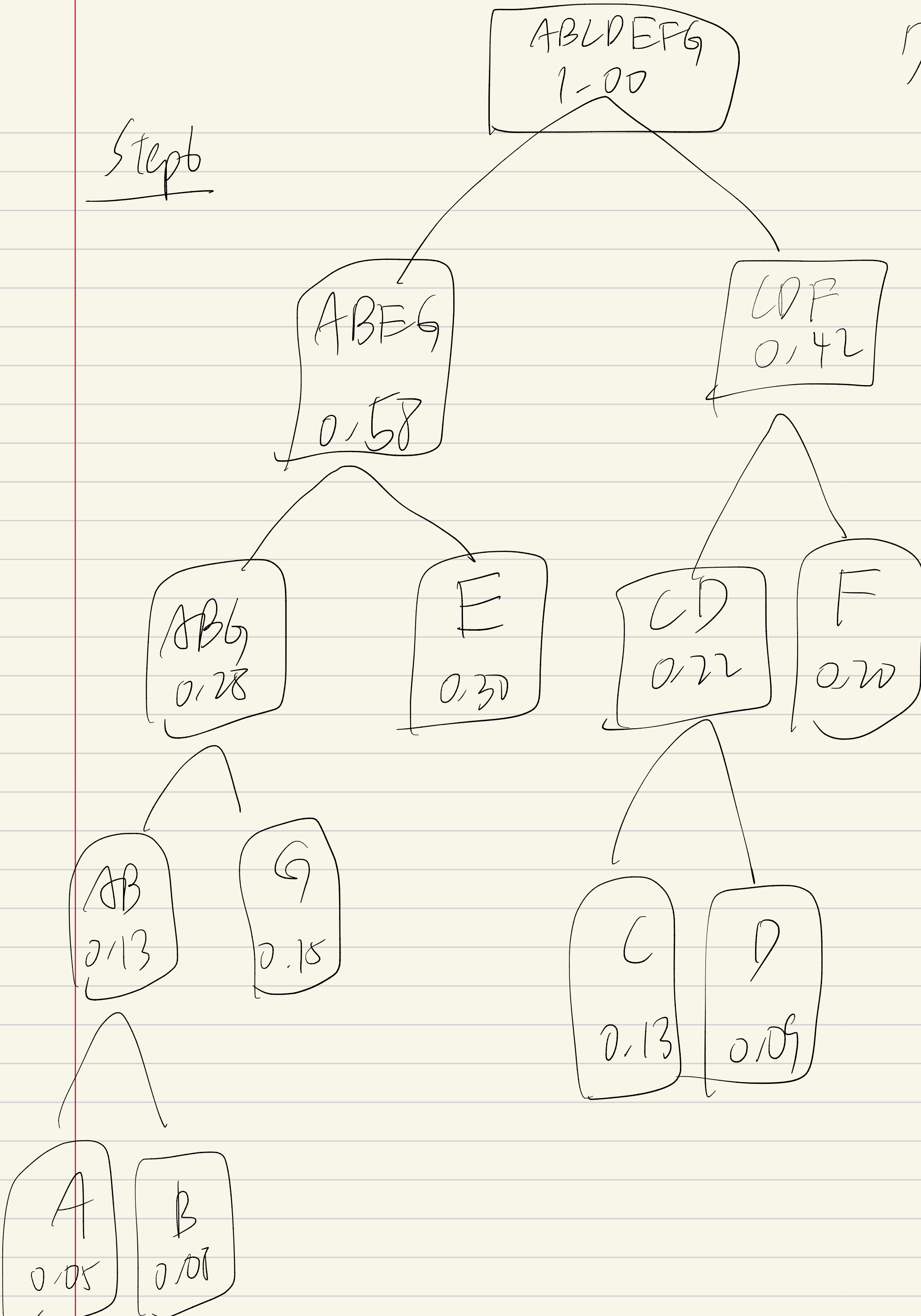
ABG
0.28
CDF
0.42
AB
0.13
G
0.15
CD
0.22
F
0.20
A
0.05
B
0.08
C
0.13
D
0.05
E
0.30

Step5


ABEG
0.58
ABG
0.28
E
0.30
AB
0.13
G
0.15
A
0.05
B
0.08
CDF
0.42
CD
0.22
F
0.20
C
0.13
D
0.09

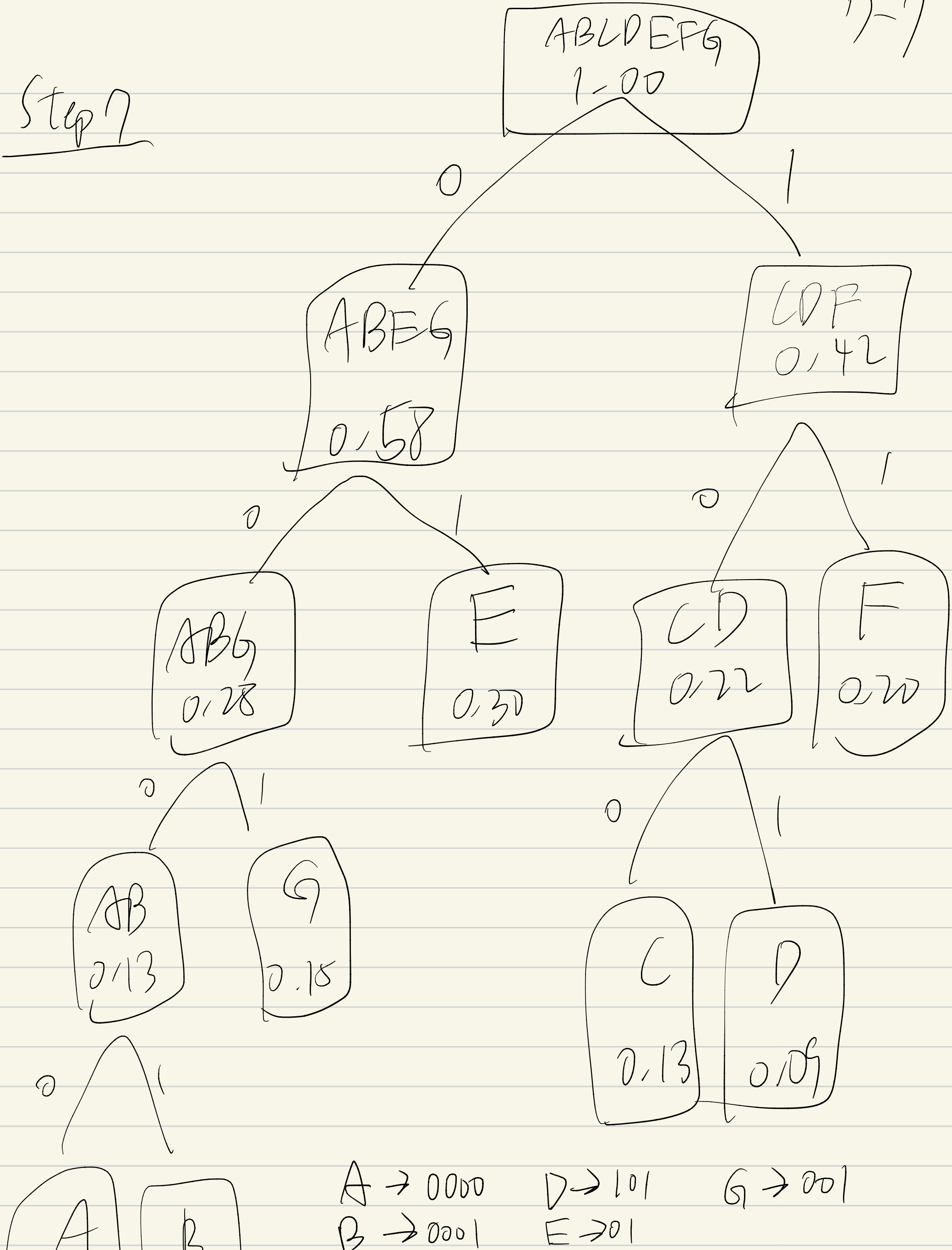
7-6

Step6


ABCDEFG
1.00
ABEG
0.58
CDF
0.42
ABG
0.28
E
0.30
CD
0.22
F
0.20
AB
0.13
G
0.15
C
0.13
D
0.09
A
0.05
B
0.08

7-7

Step 7



A → 0000 D → 101 G → 001
B → 0001 E → 01
C → 100 F → 11

#

We have $(l_1, l_2, l_3, l_4, l_5, l_6, l_7) = (4, 4, 3, 3, 2, 2, 3)$.

Thus, the average code length is

$$\sum_{i=1}^{7} p_i l_i = \underline{2.63} \ \#$$

(d) We compute

$$\sum_{i=1}^{7} 2^{-l_i} = 1 \leq 1,$$ and thus Kraft's inequality

<u>is satisfied</u> #

(e) The minimum code length satisfies

$H(X_1, \dots, X_{10}) \leq 10L < H(X_1, \dots, X_{10}) + 1$, where

$X_i \overset{iid}{\sim} X$.

Thus $10H(X) \leq 10L < 10H(X) + 1$

$$\Rightarrow \underline{H(X) \leq L < H(X) + \frac{1}{10}} \ \#$$

ILLINOIS ECE

7. (A)(1)

sample average part: $\frac{1}{n}\sum_{i=1}^{n} \log_2 P(x_i)$ ↗ using source statistics

$= \frac{1}{8}(3\log_2 0.3 + 2\log_2 0.2 + \log_2 0.05 + \log_2 0.08 + \log_2 0.13)$

$= -2.5955$

(this gives us a minus $\overbrace{\text{sample average}}^{\bar{X}_n}$ of 2.5955 bits)

2) 2.5984 bits

3) using problem 3.1 in T&C (and Chebyshev)

$p\{|\bar{X}_n - M| < \epsilon\} \geq \frac{\sigma^2}{n\epsilon}$

$\epsilon = 1$

$n = 8$ $\sigma^2 = 2$

$M = H(x)$

$= 2.5984$

$\cancel{p\{|\bar{X}_n - 2.5984| \leq 1\} \geq 0.75}$

$p\{|\bar{X}_n - 2.5984|\} \geq \boxed{0.75}$

ILLINOIS ECE

4) $P\{|\bar{X}_n - \mu| < \varepsilon\} \geq 1 - \frac{\sigma^2}{n\varepsilon}$

$$1 - \frac{\sigma^2}{n\varepsilon} \geq 0.99$$

$$n = \left\lceil \frac{\sigma^2}{0.01\varepsilon} \right\rceil = \boxed{200}$$

5) $n = \left\lceil \frac{100 \cdot 0.25}{1^2} \right\rceil = 25$

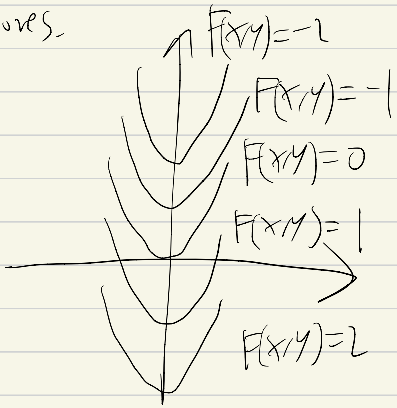
possible # tuples $= 7^{25}$

6) as n gets large

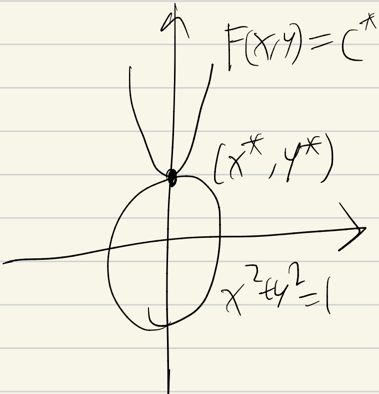
$$q_0 = \frac{2.5989^n}{7^n} \approx \frac{6.054^n}{7^n}$$

7.) $\left\lceil \log_2 2^{2.5989n} \right\rceil = \lceil 2.5999n \rceil$ per tuple

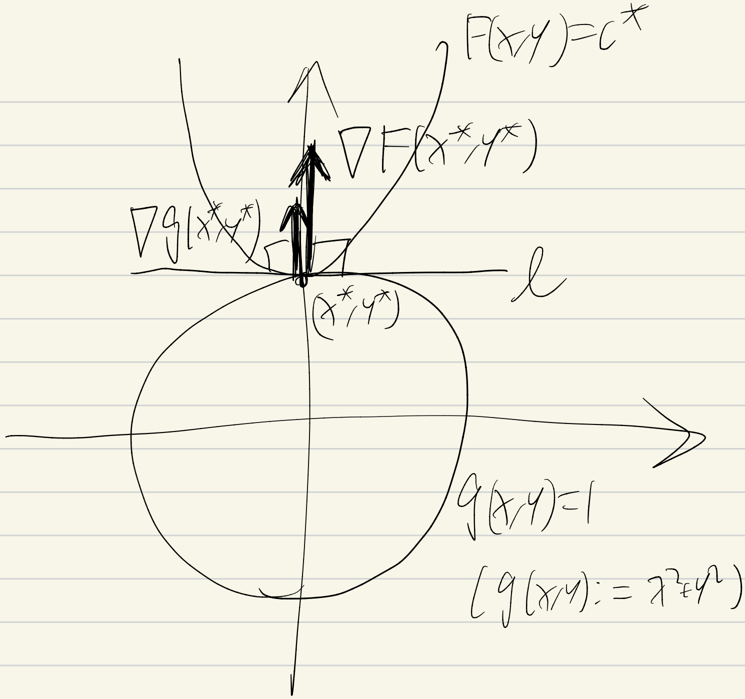
8. For each $c \in \mathbb{R}$, the equation $F(x,y) = 8x^2 - 2y = c$ defines a curve on $\mathbb{R}^2$. As c slightly changes, the curve slightly moves.



In this example, as c decreases, the curve moves upward. To minimize $F(x,y)$ under the constraint $x^2+y^2=1$, one seek to move the curve upward as much as possible while keeping at least an intersection of the unit circle.



For the minimal c^*, it is the case that moving the curve upward any further will make the curve have no intersection of $x^2+y^2=1$. Graphically, we know that in this case, The curve $F(x,y)=c^*$ is tangent to $x^2+y^2=1$ at the intersecting point (x^*,y^*).



Graphically, the normal vector of $F(x,y)=c^*$ at (x^*,y^*), i.e. $\nabla F(x^*,y^*)$, is perpendicular to the tangent line ℓ.

At the same time, the normal vector of $g(x,y)=1$ at (x^*,y^*), i.e. $\nabla g(x^*,y^*)$, is also perpendicular to ℓ.

Since $\nabla F(x^*,y^*)$ and $\nabla g(x^*,y^*)$ are perpendicular to the same line, they are parallel #.

That is, $\exists -\lambda \in \mathbb{R}$ s.t. $\nabla F(x^*,y^*) = -\lambda \nabla g(x^*,y^*)$, here the role of $-\lambda$ is the scaling between two parallel vectors $\nabla F(x^*,y^*)$ and $\nabla g(x^*,y^*)$. #

Now come back to the Lagrange multiplier equations:

$\mathcal{L}(x,y,\lambda) = F(x,y) + \lambda\,(g(x,y)-1).$

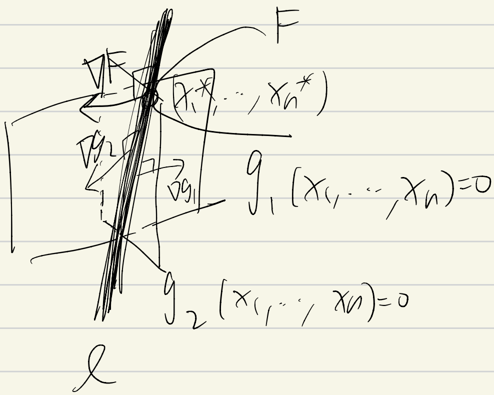
$$\Rightarrow \begin{cases} \dfrac{\partial \mathcal{L}}{\partial x} = \dfrac{\partial F}{\partial x} + \lambda \dfrac{\partial g}{\partial x} = 0 & \cdots\ (8\text{-}1) \\[2ex] \dfrac{\partial \mathcal{L}}{\partial y} = \dfrac{\partial F}{\partial y} + \lambda \dfrac{\partial g}{\partial y} = 0 & \cdots\ (8\text{-}2) \\[2ex] \dfrac{\partial \mathcal{L}}{\partial \lambda} = g(x,y) - 1 = 0 & \cdots\ (8\text{-}3) \end{cases}$$

(8-1) & (8-2) mean exactly that $\nabla F(x^*,y^*) = -\lambda \nabla g(x^*,y^*)$, and (8-3) is simply the constraint.

As a remark, having parallel gradient is just a necessary condition for being a minimal solution. That's why one have to calculate all the solutions satisfying (8-1), (8-2), and (8-3) and find the minimum among these solutions.

In the case where there are multiple (say m) constraints, graphically, $\nabla F(x_1, \dots, x_n)$ is tangent to the intersection of $g_1(x_1, \dots, x_n) = 0$, $g_2(x_1, \dots, x_n) = 0$, $\dots$, and $g_m(x_1, \dots, x_n) = 0$, and thus

$$\nabla F(x_1, \dots, x_n) \in \operatorname{Span}(\nabla g_1(x_1, \dots, x_n), \dots, \nabla g_m(x_1, \dots, x_n)).$$



$$\begin{cases} \ell \perp \nabla g_1(x_1^*, \dots, x_n^*) \\ \ell \perp \nabla g_2(x_1^*, \dots, x_n^*) \\ \ell \perp \nabla F(x_1^*, \dots, x_n^*) \end{cases}$$

That is, $\exists -\lambda_1, \dots, -\lambda_m \in \mathbb{R}$ s.t.

$\nabla F(x_1^*, \dots, x_n^*) = \sum_{i=1}^{m} -\lambda_i \nabla g_i(x_1^*, \dots, x_n^*)$, which explains the Lagrange multiplier with multiple constraints.