5.2 RAT

---

1. Let's say we have the following event space and the empirical data:

VB	VBD	VBG	VBN	VBP	VBZ
5	10	4	8	6	7

   What will be the probability distribution that maximize entropy with the following feature?

$$f_{past} = \{\mathrm{VBD}, \mathrm{VBN}\}, E[f_{past}] = ½$$

2. Suppose we have a 1 feature maxent model built over observed data as shown. This time our one feature is picking out *ends-with*(*vowel*). Work out what the expectation of that feature is and choose the constructed model's probability distribution over the four possible outcomes

	ends-with(*vowel*)	*ends-with*(*consonent*)
starts-with(capital)	**1**	**1**
starts-with(lower)	**2**	**2**

$$f = \{\textit{ends-with}(\textit{vowel})\}$$

3. Which of the following is ***not*** true of joint models $P(c, d)$ with the marginal constraint?
   A) Computing the expectation of each feature is more time-consuming with the marginal constraint
   B) P(c,d) is zero if d does not occur in our empirical data
   C) Maximizing P(c,d) is equivalent to maximizing P(c|d)
   D) The model is useful when the space C×D is too huge to enumerate

4. Suppose a certain feature f_i matches 5 times over the training data (C,D). That is, it's empirical expectation is 5. Suppose further that we train a smoothed maxent model with $\sigma^2 = 1$ and that the feature gets a weight of $\lambda_i = 2$ in the resulting model. What will the empirical expectation of the feature on the training data D be?
   Recall:

$$\delta \log P(C, \lambda | D) / \delta \lambda_i = \text{actual}(f_i, C) - \text{predicted}(f_i, \lambda) - \lambda_i / \sigma^2$$