1. Niech η i λ będą dwoma rozkładami prawdopodobieństwa na zbiorze skończonym A. Niech X_i, $i = 0, 1, \ldots$ będą niezależnymi zmiennymi losowymi o rozkładzie η. Niech $P_\lambda(X_1, \ldots, X_n)$ będzie prawdopodobeństem otrzymania ciągu $X_1, \ldots, X_n)$ według rozkładu λ. Uzasadnij że

$$\lim_{n\to\infty} \frac{1}{n} \log(P_\lambda(X_1, \ldots, X_n)) = \sum_{a\in A} \eta(a) \log(\lambda(a))$$

z prawdopodobieństwem 1. W szczególności, dla $\eta = \lambda$ z prawej strony dostajemy entropię rozkładu η. Uzasadnij że wyrażenie

$$\sum_{a\in A} \eta(a) \log(\lambda(a))$$

jako funkcja λ przy ustalonym η osiąga maksimum dla $\lambda = \eta$ i dla pozostałych λ wartości są mniejsze. Jaki stąd można wyciąnąć wniosek jeśli mamy n niezależnie wylosowych próbek z rozkładem który nie znamy, ale wiemy że jest jednym ze skończenie wielu rozkładów $\lambda_1, \ldots, \lambda_k$.

Wskazówka: Istnienie granicy wyżej wynika z mocnego prawa wielkich liczb, aby wyznaczyć granicę trzeba policzyć wartość oczekiwaną.

2. Mamy tekst o którym wiemy że jest w jednym z kilku języków, ale nie wiemy dokładnie w którym (np. tekst jest albo po polsku albo po angielsku). Zaproponuj metodę rozpoznawania w jakim tekst jest języku na podstawie prawdopodobieństwa przypisywanego tekstowi przez model(e) Markowa na znakach. Wyprobuj to praktycznie, estymując parametry modelu z długiego tekstu.

3. Przy estymacji Baysa mamy zadany rozkład prawdopodobieństwa μ na parametrze $\lambda \in \mathbb{R}^k$ i dla każdego λ mamy rozkład prawdopodobieństwa P_λ na zbiorze A. Rozkład λ może być ciągły, ale dla uproszczenia zakładamy że A jest skończony i dla dowolnego λ i $a \in A$ mamy $P_\lambda(a) > 0$. Estymacja polega na tym że mając dany wynik $a_1, \ldots, a_n$ z n niezależnych losowań szacujemy λ przez

$$\frac{\int \lambda P_\lambda(a_1, \ldots, a_n) d\mu(\lambda)}{\int P_\lambda(a_1, \ldots, a_n) d\mu(\lambda)}$$

Niech $A = \{0, \ldots, k-1\}$, $P_\lambda(m) = \lambda_m$, przy tym λ ma rozkład jednostajny na zbiorze wektorów o nieujemnych składowych sumujących się do 1. Uzasadnij że jeśli w wylosowanym ciągu wartość m pojawia się l_m razy gdzie $\sum l_m = n$ to Baysowskie oszacowanie dla λ_m to $\frac{l_m+1}{n+k}$.

Wskazówka: Najpierw oblicz całki dla $k = 2$. Następnie zauważ że dla wiekszych k po scałkowaniu po zmiennych innych niż λ_m dostajemy całkę jak dla $k = 2$, tyle że w liczniku i mianowiku pojawi się dodatkowy czynnik $(1 - \lambda_m)^{k-2}$.

4. Zakładamy że prawdziwy rozkład na słowach jest jak w zadaniu 1 z listy 1, tzn. pierwsze 1000 pozycji słownika ma równe prawdopodobieństwa pojawienia się w tekście i odpowadają za 90% tekstu. Tzn. każda z nich pojawia się z z prawdopodobieństwem $9 * 10^{-4}$. Pozostałe 100000 pozycji słownika daje resztę tekstu i też ma równe prawdopodobieństwa (równe 10^{-6}). Przyjmujemy że słowa są losowane ze słownia niezależnie. Losujemy 101000 wyrazów i stosujemy estymację Baysa do oszacowania rozkładu, tzn. zastępujemy prawdopodobieństwo slowa i przez $\frac{c_i+1}{101000+101000}$ gdzie c_i to zaobserwowana ilość wystąpień słowa i. Oszacuj jak to się ma do prawdziwego rozkładu.

5. Przy rozkładzie na słowach jak w zadaniu 4 losujemy 25000 słów. Dzielimy do ma dwie części: pierwsze 20000 tworzy cześć A, pozostałe 5000 część B. Niech c_i będzie zaobserwowaną ilością wystąpień słowa i w części A. Dla liczby k niech T_k będzie zbiorem słów dla których $c_i = k$ i niech t_k będzie łączną ilościa wystąpień słów z T_k w części B. Tym razem przyjmujemy że empiryczne prawdopodobieństwo słowa i to $\frac{t_k}{|T_k|N}$ gdzie $k = c_i$, $N = 20000$ zaś $|T_k|$ oznacza ilość elementów zbioru T_k. Dla pozostałych słów przyjmujemy równe prawdopodobieństwa empiryczne, tak by całość sumowała się do 1. Oszacuj jak to się ma do prawdziwego rozkładu.