Discriminative Learning in Speech Recognition

Yueng-Tien, Lo
g96470198@csie.ntnu.edu.tw
Speech Lab, CSIE
National Taiwan Normal University

Reference
Xiaodong He and Li Deng. "Discriminative Learning in Speech Recognition,“
outline

• introduction
• Discriminative Learning Criteria of MMI, MCE and MPE/MWE
• The common rational-function form for objective functions of MMI, MCE, and MPE/MWE
• Optimizing Rational Functions By Growth Transformation
• Discriminative Learning for Discrete HMMs Based on the GT Framework
Introduction (1/3)

• Discriminative learning has become a major theme in recent statistical signal processing and pattern recognition research including practically all areas of speech and language processing.

• A key to understanding the speech process is the dynamic characterization of its sequential or variable-length pattern.

• Two central issues in the development of discriminative learning methods for sequential pattern recognition are:
  1. construction of the objective function for optimization
  2. actual optimization techniques
Introduction(2/3)

• There is a pressing need for a unified account of the numerous discriminative learning techniques in the literature.

• To fulfill this need while providing insights into the discriminative learning framework for sequential pattern classification and recognition.

• It is our hope that the unifying review and insights provided in the article will foster more principled and successful applications of discriminative learning in a wide range of signal processing disciplines, speech processing or otherwise.
• In addition to providing a general overview on the classes of techniques (MMI, MCE, and MPE/MWE), this article has a special focus on three key areas in discriminative learning.

• First, it provides a unifying view of the three major discriminative learning objective functions, MMI, MCE, and MPE/MWE, for classifier parameter optimization, from which insights to the relationships among them are derived.

• Second, we describe an efficient approach of parameter estimation in classifier design that unifies the optimization techniques for discriminative learning.

• The third area is the algorithmic properties of the MCE and MPE/MWE based learning methods under the parameter estimation framework of growth transformation for sequential pattern recognition using HMMs.
**Discriminative Learning Criteria of MMI, MCE and MPE/MWE (1/2)**

- MMI (maximum mutual information), MCE (minimum classification error), and MPE/MWE (minimum phone error/minimum word error) are the three most popular discriminative learning criteria in speech and language processing, which are the main subject of this paper.

- To set up the stage, we denote by $\Lambda$ the set of classifier parameters that needs to be estimated during the classifier design. For instance in speech and language processing, a (generative) joint distribution of observing a data sequence $X$ given the corresponding labeled word sequence $S$ can be written as follows:

$$p(X,S|\Lambda) = p(X|S,\Lambda)p(S)$$
Discriminative Learning Criteria of MMI, MCE and MPE/MWE (2/2)

- it is assumed that the parameters in the “language model” $P(S)$ are not subject to optimization.

- Given a set of training data, we denote by $R$ the total number of training tokens.

- In this paper, we focus on supervised learning, where each training token consists of an observation data sequence: $X_r = x_{r,1}, \ldots, x_{r,T_r}$, and its correctly labeled (e.g., word) pattern sequence: $S_r = W_{r,1}, \ldots, W_{r,N_r}$, with $W_{r,i}$ being the $i$-th word in word sequence $S_r$.

- We use a lower case variable $S_r$ to denote all possible pattern sequences that can be used to label the $r$-th token, including the correctly labeled sequence $S_r$ and other sequences.
Maximum Mutual Information (MMI) (1/3)

- In the MMI-based classifier design, the goal of classifier parameter estimation is to maximize the mutual information $I(X,S)$ between data $X$ and their corresponding labels/symbols $S$.

- From the information theory perspective, mutual information provides a measure of the amount of information gained, or the amount of uncertainty reduced, regarding $S$ after seeing $X$.

- Mutual information $I(X,S)$ is defined as

$$I(X,S) = \sum_{x,s} p(X,S) \log \frac{p(X,S)}{p(X)p(S)} = \sum_{x,s} p(X,S) \log \frac{p(S|X)}{p(S)} = H(S) - H(S|X) \quad (2)$$

where $H(S) = -\sum_s p(S) \log p(S)$ is the entry of $S$, and $H(S|X)$ is the conditional entropy given data $X$:

$$H(S|X) = -\sum_{x,s} p(X,S) \log p(S|X)$$

When $p(S|X)$ is based on model $\mathcal{A}$, we have

$$H(S|X) = -\sum_{x,s} p(X,S) \log p(S|X,A) \quad (3)$$
Maximum Mutual Information (MMI) (2/3)

• Assume that the parameters in $P(S)$ (“language model”) and hence $H(S)$ is not subject to optimization. Consequently, maximizing mutual information of (2) becomes equivalent to minimizing $H(S|X)$ of (3) on the training data. When the tokens in the training data are drawn from an i.i.d. distribution, $H(S|X)$ is given by

$$H(S|X) = -\frac{1}{R} \sum_{r=1}^{R} \log p(S_r|X_r, \Lambda) = -\frac{1}{R} \sum_{r=1}^{R} \log \frac{p(X_r, S_r | \Lambda)}{p(X_r)}.$$  

• Therefore, parameter optimization of MMI based discriminative learning is to maximize the following objective function:

$$O_{\text{MMI}}(\Lambda) = \sum_{r=1}^{R} \log \frac{p(X_r, S_r | \Lambda)}{p(X_r)} = \sum_{r=1}^{R} \log \frac{p(X_r, S_r | \Lambda)}{\sum_{S_r} p(X_r, S_r | \Lambda)} \quad (4)$$

• The objective function $O_{\text{MMI}}$ of (4) is a sum of logarithms. For comparisons with other discriminative training criteria in following sections, we construct the monotonically increasing function of exponentiation for (4). This gives

$$\tilde{O}_{\text{MMI}}(\Lambda) = \exp[O_{\text{MMI}}(\Lambda)] = \prod_{r=1}^{R} \frac{p(X_r, S_r | \Lambda)}{\sum_{S_r} p(X_r, S_r | \Lambda)} \quad (5)$$
Maximum Mutual Information (MMI) (3/3)

- It should be noted that $\tilde{O}_{\text{MMI}}$ and $O_{\text{MMI}}$ have the same set of maximum points, because maximum points are invariant to monotonically increasing transforms. For comparisons with other discriminative training criteria, we rewrite each factor in (5) as

$$
\frac{p(X_r, S_r | \Lambda)}{\sum_{s_r} p(X_r, s_r | \Lambda)} = 1 - \sum_{s_r \neq S_r} P(s_r | X_r, \Lambda) = 1 - \sum_{s_r} \left( 1 - \delta(s_r, S_r) \right) P(s_r | X_r, \Lambda).
$$

- We define (6) as the model-based expected utility for token $X_r$, which equals one minus the model-based expected loss for that token.
Minimum “Phone” or “Word” Errors (MPE/MWE)(1/2)

- In contrast to MMI and MCE described earlier that are typically aimed at large segments of pattern sequences (e.g., at string or even super-string level obtained by concatenating multiple pattern strings in sequence), MPE aims at the performance optimization at the sub-string pattern level.
- The MPE objective function that needs to be maximized is defined as

\[
O_{\text{MPE}}(\Lambda) = \sum_{r=1}^{R} \frac{\sum_{s_r} p(X_r, s_r | \Lambda) A(s_r, S_r)}{\sum_{s_r} p(X_r, s_r | \Lambda)}
\]

- where \(A(s_r, S_r)\) is the raw phone (sub-string) accuracy count in the sentence string \(S_r\).

The raw phone accuracy count \(A(s_r, S_r)\) is defined as the total phone (sub-string) count in the reference string \(S_r\) minus the sum of insertion, deletion and substitution errors of \(s_r\) computed based on \(S_r\).
Minimum “Phone” or “Word” Errors (MPE/MWE)(2/2)

• The MPE criterion (18) equals the model-based expectation of the raw phone accuracy count over the entire training set. This relation can be seen more clearly by rewriting (18) as

\[
O_{\text{MPE}}(\Lambda) = \sum_{r=1}^{R} \sum_{s_r} P(s_r \mid X_r, \Lambda) A(s_r, S_r)
\]

where \( p(s_r \mid X_r, \Lambda) = \frac{p(X_r, s_r \mid \Lambda)}{p(X_r \mid \Lambda)} = \frac{p(X_r, s_r \mid \Lambda)}{\sum_{s_r} p(X_r, s_r \mid \Lambda)} \) is the model-based posterior probability

• Based on raw word accuracy count \( A_i(s_r, S_r) \), we have the equivalent definition of the MWE criterion:

\[
O_{\text{MWE}}(\Lambda) = \sum_{r=1}^{R} \frac{\sum_{s_r} p(X_r, s_r \mid \Lambda) A_i(s_r, S_r)}{\sum_{s_r} p(X_r, s_r \mid \Lambda)}
\]  

(19)
Discussions (single-token level)

- At the single-token level, the MMI criterion uses a model-based expected utility of (6) while the MCE criterion uses a classifier-dependent smoothed empirical utility defined by (9), (13), and (15). Likewise, the MPE/MWE criterion also uses a model-based expected utility, but the utility is computed at the sub-string level; e.g., at the phone or word level. We note that for mathematical tractability reasons, in this paper, a specific misclassification measure (12) is used for MCE. As a consequence, the smoothed empirical utility (15) takes the same form as (6) (though they are derived from different motivations). This can be directly seen by substituting (14) to (15).
\[
\frac{p(X_r, S_r | \Lambda)}{\sum_{s_r} p(X_r, s_r | \Lambda)} = 1 - \sum_{s_r \neq S_r} P(s_r | X_r, \Lambda) = 1 - \sum_{s_r} \left(1 - \delta(s_r, S_r)\right) P(s_r | X_r, \Lambda). \tag{6}
\]

\[
d_r(X_r, \Lambda) = -g_{S_r}(X_r, \Lambda) + G_{S_r}(X_r, \Lambda) \tag{9}
\]

\[
l_r(d_r(X_r, \Lambda)) = \frac{1}{1 + e^{-\alpha d_r(X_r, \Lambda)}} \tag{13}
\]

\[
\begin{aligned}
g_{S_r}(X_r, \Lambda) &= \log p^n(X_r, S_r | \Lambda) \\
G_{S_r}(X_r, \Lambda) &= \log \sum_{i=1}^{N} p^n(X_r, s_{r,i} | \Lambda) \tag{12}
\end{aligned}
\]

\[
u_r(d_r(X_r, \Lambda)) = 1 - l_r(d_r(X_r, \Lambda)). \tag{15}
\]
Discussions (multiple-token level)

- At the multiple-token level, by comparing (5), (17), (18), and (19), it is clear that MMI training maximizes a product of model-based expected utilities of training tokens, while MCE training maximizes a summation of smoothed empirical utilities over all training tokens and MPE/MWE training maximizes a summation of model-based expected utilities (computed on sub-string units). The difference between the product and the summation forms of the utilities differentiates MMI from MCE/MPE/MWE. This difference causes difficulties in extending the original GT/EBW formulas proposed for MMI to other criteria.
\[ O_{MCE}(\Lambda) = R(1 - L_{MCE}(\Lambda)) = \sum_{r=1}^{R} u_r(d_r(X_r, \Lambda)) = \sum_{r=1}^{R} \frac{p(X_r, S_r | \Lambda)}{\sum_{s_r} p(X_r, s_r | \Lambda)} \] (17)

\[ O_{MPE}(\Lambda) = \frac{\sum_{r=1}^{R} \sum_{s_r} p(X_r, s_r | \Lambda) A(s_r, S_r)}{\sum_{s_r} p(X_r, s_r | \Lambda)} \] (18)

\[ O_{MVE}(\Lambda) = \sum_{r=1}^{R} \sum_{s_r} p(X_r, s_r | \Lambda) A_2(s_r, S_r) \frac{p(X_r, s_r | \Lambda)}{\sum_{s_r} p(X_r, s_r | \Lambda)} \] (19)

\[ \tilde{O}_{MFI}(\Lambda) = \exp\left[ O_{MFI}(\Lambda) \right] = \prod_{r=1}^{R} \frac{p(X_r, S_r | \Lambda)}{\sum_{s_r} p(X_r, s_r | \Lambda)} \] (5)
The Common Rational-Function form for Objective functions of MMI, MCE, and MPE/MWE

- we show that the objective functions in discriminative learning based on the MMI, MCE and MPE/MWE criteria can be mapped to a canonical rational-function form where the denominator function is constrained to be positive valued.

- This canonical rational-function form has the benefit of offering insights into the relationships among MMI, MCE, and MPE/MWE based classifiers and it facilitates the development of a unified classifier parameter optimization framework for applying MMI, MCE, and MPE/MWE objective functions in sequential pattern recognition tasks.
Rational-Function Form for the Objective Function of MMI

- Based on (5), the canonical rational-function form for MMI objective function can be constructed as:

\[
\tilde{O}_{\text{MMI}}(\Lambda) = \frac{p(X_1 \ldots X_R, S_1 \ldots S_R | \Lambda)}{\sum_{s_1 \ldots s_R} p(X_1 \ldots X_R, s_1 \ldots s_R | \Lambda)} = \frac{\sum_{s_1 \ldots s_R} p(X_1 \ldots X_R, s_1 \ldots s_R | \Lambda) C_{\text{MMI}}(s_1 \ldots s_R)}{\sum_{s_1 \ldots s_R} p(X_1 \ldots X_R, s_1 \ldots s_R | \Lambda)}
\]

where

\[
C_{\text{MMI}}(s_1 \ldots s_R) = \prod_{r=1}^{R} \delta(s_r, S_r)
\]

- is a quantity that depends only on the sentence sequence \( s_1, \ldots, s_R \), and \( \delta(s_r, S_r) \) is the Kronecker delta function, i.e.,

\[
\delta(s_r, S_r) = \begin{cases} 
1 & \text{if } s_r = S_r \\
0 & \text{otherwise}
\end{cases}
\]

In (20), the first step uses the common assumption that different training tokens are independent of each other.
Rational-Function Form for the Objective Function of MCE(1/3)

• Unlike the MMI case where the rational-function form can be obtained through a simple exponential transformation, the objective function of MCE as given in (17) is a sum of rational functions rather than a rational function in itself (i.e., a ratio of two polynomials)

• The gradient descent based sequential learning using GPD has two main drawbacks:
  1. it is a sample-by-sample learning algorithm. Algorithmically, it is difficult for GPD to parallelize the parameter learning process, which is critical for large scale tasks.
  2. it is not a monotone learning algorithm and it does not have a monotone learning function to determine the stopping point of the discriminative learning.

• The derivation of the rational-function form for the objective function of MCE is as follows:
\[ O_{MCE}(\Lambda) = R(1 - L_{MCE}(\Lambda)) = \sum_{r=1}^{R} u_r(d_r(X_r, \Lambda)) = \sum_{r=1}^{R} \frac{p(X_r, S_r | \Lambda)}{\sum_{s_r} p(X_r, S_r | \Lambda)} \]  

(17)
Rational-Function Form for the Objective Function of MCE(2/3)

\[ O_{MCE}(\Lambda) = \frac{\sum_{r=1}^{R} \sum_{s_1} p(X_r, s_r | \Lambda) \delta(s_r, S_r)}{\sum_{s_1} p(X_1, s_1 | \Lambda)} + \frac{\sum_{r=2}^{R} p(X_r, s_r | \Lambda) \delta(s_r, S_r)}{\sum_{s_2} p(X_2, s_2 | \Lambda)} + \ldots + \frac{\sum_{r=R}^{R} p(X_R, s_R | \Lambda) \delta(s_R, S_R)}{\sum_{s_R} p(X_R, s_R | \Lambda)} \]

\[ = \frac{\sum_{s_1} p(X_1, s_1 | \Lambda)p(X_2, s_2 | \Lambda)[\delta(s_1, S_1) + \delta(s_2, S_2)]}{\sum_{s_1} p(X_1, s_1 | \Lambda)p(X_2, s_2 | \Lambda)} + O_3 + \ldots + O_R \]

\[ = \frac{\sum_{s_1} p(X_1, s_1 | \Lambda)[C_{MCE}(s_1s_2)]}{\sum_{s_2} p(X_1, s_1 | \Lambda)} + O_3 + \ldots + O_R \]

\[ = \frac{\sum_{s_1} p(X_1, s_1, s_2, s_3 | \Lambda)[C_{MCE}(s_1s_2s_3)]}{\sum_{s_1} p(X_1, s_1, s_2, s_3 | \Lambda)} + O_4 + \ldots + O_R \]

\[ = \frac{\sum_{s_1} p(X_1, \ldots, X_R, s_1 \ldots s_R | \Lambda) C_{MCE}(s_1 \ldots s_R)}{\sum_{s_1} p(X_1, \ldots, X_R, s_1 \ldots s_R | \Lambda)} \]

(23)
Rational-Function Form for the Objective Function of MCE(3/3)

- Where

\[
C_{MCE}(s_1...s_R) = \sum_{r=1}^{R} \delta(s_r, S_r) . C_{MCE}(s_1, ..., s_R)
\]

\(C_{MCE}(s_1...s_R)\) can be interpreted as the string accuracy count for \(s_1, ..., s_R\), which takes an integer value between zero and \(R\) as the number of correct strings in \(s_1, ..., s_R\).

- As it will be further elaborated, the rational-function form (23) for the MCE objective function will play a pivotal role in our study of MCE-based discriminative learning.
Rational-Function Form for the Objective Function of MPE/MWE(1/2)

- Similar to MCE, the MPE/MWE objective function is also a sum of multiple (instead of a single) rational functions, and hence it is difficult to derive GT formulas

- An important finding is that the same method used to derive the rational-function form (23) for the MCE objective function can be applied directly to derive the rational-function form for MPE/MWE objective functions as defined in (18) and (19)
Rational-Function Form for the Objective Function of MPE/MWE(2/2)

\[ O_{MWE}(\Lambda) = \sum_{s_1 \ldots s_R} \frac{p(X_1 \ldots X_R, s_1 \ldots s_R | \Lambda) C_{MWE}(s_1 \ldots s_R)}{\sum_{s_1 \ldots s_R} p(X_1 \ldots X_R, s_1 \ldots s_R | \Lambda)} \]

where \( C_{MWE}(s_1 \ldots s_R) = \sum_{r=1}^{R} A_j(s_r, S_r) \).

\[ O_{MPE}(\Lambda) = \sum_{s_1 \ldots s_R} \frac{p(X_1 \ldots X_R, s_1 \ldots s_R | \Lambda) C_{MPE}(s_1 \ldots s_R)}{\sum_{s_1 \ldots s_R} p(X_1 \ldots X_R, s_1 \ldots s_R | \Lambda)} \]

where \( C_{MPE}(s_1 \ldots s_R) = \sum_{r=1}^{R} A(s_r, S_r) \), and
• The main result in this section is that all three discriminative learning objective functions, MMI, MCE, and MPE/MWE, can be formulated in a unified canonical rational-function form as follows:

\[
O(\Lambda) = \frac{\sum_{s_1 \ldots s_R} p(X_1 \ldots X_R, s_1 \ldots s_R \mid \Lambda) \cdot C_{DT}(s_1 \ldots s_R)}{\sum_{s_1 \ldots s_R} p(X_1 \ldots X_R, s_1 \ldots s_R \mid \Lambda)}
\]  

(26)

where the summation over \( s=s1 \ldots sR \) in (26) denotes all possible labeled sequences (both correct and incorrect ones) for all \( R \) training tokens.
## Comments and Discussions

<table>
<thead>
<tr>
<th>Objective Functions</th>
<th>$C_{DT}(s_r)$</th>
<th>$C_{DT}(s_1 \ldots s_R)$</th>
<th>Label Sequence Set Used in DT</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCE (N-best)</td>
<td>$\delta(s_r, S_r)$</td>
<td>$\sum_{r=1}^{R} C_{DT}(s_r)$</td>
<td>${S_r, s_{r,1}, \ldots, s_{r,N}}$</td>
</tr>
<tr>
<td>MCE (one-best)</td>
<td>$\delta(s_r, S_r)$</td>
<td>$\sum_{r=1}^{R} C_{DT}(s_r)$</td>
<td>${S_r, s_{r,1}}$</td>
</tr>
<tr>
<td>MPE</td>
<td>$A(s_r, S_r)$</td>
<td>$\sum_{r=1}^{R} C_{DT}(s_r)$</td>
<td>all possible label sequences</td>
</tr>
<tr>
<td>MWE</td>
<td>$A_l(s_r, S_r)$</td>
<td>$\sum_{r=1}^{R} C_{DT}(s_r)$</td>
<td>all possible label sequences</td>
</tr>
<tr>
<td>MMI</td>
<td>$\delta(s_r, S_r)$</td>
<td>$\prod_{r=1}^{R} C_{DT}(s_r)$</td>
<td>all possible label sequences</td>
</tr>
</tbody>
</table>

Table 1: $C_{DT}(s_1 \ldots s_R)$ in the unified rational-function form for MMI, MCE, and MPE/MWE objective functions. The set of “competing token candidates” distinguishes N-best and one-best versions of the MCE. Note that the overall $C_{DT}(s_1 \ldots s_R)$ is constructed from its constituents $C_{DT}(s_r)$’s in individual string tokens by either summation (for MCE, MPE/MWE) or product (for MMI).
Optimizing Rational Functions By Growth Transformation(1/2)

- GT-based parameter optimization refers to a family of batch-mode, iterative optimization schemes that “grow” the value of the objective function upon each iteration.

- the new set of model parameter $\Lambda$ is estimated from the current model parameter set $\Lambda'$ through a transformation $\Lambda = T(\Lambda')$ with the property that the target objective function “grows” in its value $O(\Lambda) > O(\Lambda')$ unless $\Lambda = \Lambda'$.
The goal of GT based parameter optimization is to find an optimal $\Lambda$ that maximizes the objective function $O(\Lambda)$ which is a rational function of the following form:

$$O(\Lambda) = \frac{G(\Lambda)}{H(\Lambda)}$$

For example, $O(\Lambda)$ can be one of the rational functions of (20), (23), (24) and (25) for the MMI,MCE, and MPE/MWE objective functions, respectively, or the general rational-function (26). In the general case of (26), we have

$$G(\Lambda) = \sum_{s} p(X, s | \Lambda) C(s), \text{ and } H(\Lambda) = \sum_{s} p(X, s | \Lambda)$$  \hspace{1cm} (28)

where we use short-hand notation $s=s1 \ldots sR$ to denote the labeled sequences of all $R$ training tokens/sentences, and $X=X1 \ldots XR$, to denote the observation data sequences for all $R$ training tokens.
$$\tilde{O}_{MM} (\Lambda) = \frac{p(X_1 \ldots X_R, S_1 \ldots S_R | \Lambda)}{\sum_{s_1 \ldots s_R} p(X_1 \ldots X_R, s_1 \ldots s_R | \Lambda)} = \frac{\sum_{s_1 \ldots s_R} p(X_1 \ldots X_R, s_1 \ldots s_R | \Lambda) \ C_{MPE} (s_1 \ldots s_R)}{\sum_{s_1 \ldots s_R} p(X_1 \ldots X_R, s_1 \ldots s_R | \Lambda)}$$

(20)

$$= \frac{\sum_{s_1 \ldots s_R} p(X_1 \ldots X_R, s_1 \ldots s_R | \Lambda) \ C_{MCE} (s_1 \ldots s_R)}{\sum_{s_1 \ldots s_R} p(X_1 \ldots X_R, s_1 \ldots s_R | \Lambda)}$$

(23)

$$O_{MPE} (\Lambda) = \frac{\sum_{s_1 \ldots s_R} p(X_1 \ldots X_R, s_1 \ldots s_R | \Lambda) \ C_{MPE} (s_1 \ldots s_R)}{\sum_{s_1 \ldots s_R} p(X_1 \ldots X_R, s_1 \ldots s_R | \Lambda)}$$

(24)

where \( C_{MPE} (s_1 \ldots s_R) = \sum_{r=1}^{R} A(s_r, S_r) \), and

$$O_{MWE} (\Lambda) = \frac{\sum_{s_1 \ldots s_R} p(X_1 \ldots X_R, s_1 \ldots s_R | \Lambda) \ C_{MWE} (s_1 \ldots s_R)}{\sum_{s_1 \ldots s_R} p(X_1 \ldots X_R, s_1 \ldots s_R | \Lambda)}$$

(25)

where \( C_{MWE} (s_1 \ldots s_R) = \sum_{r=1}^{R} A_l(s_r, S_r) \).
Primary Auxiliary Function

- The GT-based optimization algorithm will constructs an auxiliary function of the following form:

\[ F(\Lambda; \Lambda') = G(\Lambda) - O(\Lambda')H(\Lambda) + D \]

where \( D \) is a quantity independent of the parameter set

\( \Lambda \) is the model parameter set to be estimated

by applying GT to another model parameter set \( \Lambda' \)

Substituting \( \Lambda = \Lambda' \) into , we have

\[ F(\Lambda'; \Lambda') = G(\Lambda') - O(\Lambda')H(\Lambda') + D = D \]

Hence,

\[ F(\Lambda; \Lambda') - F(\Lambda'; \Lambda') = F(\Lambda; \Lambda') - D = G(\Lambda) - O(\Lambda')H(\Lambda) \]

\[ = H(\Lambda)\left(\frac{G(\Lambda)}{H(\Lambda)} - O(\Lambda')\right) = H(\Lambda)(O(\Lambda) - O(\Lambda')) \]
Second Auxiliary Function

May still be too difficult to optimize directly, and a second auxiliary function can be constructed

\[ V(\Lambda; \Lambda') = \sum_{s} \sum_{q} \sum_{\chi} f(\chi, q, s, \Lambda') \log f(\chi, q, s, \Lambda) \]

\[ F(\Lambda; \Lambda') = \sum_{s} \sum_{q} \sum_{\chi} f(\chi, q, s, \Lambda) \]