A Statistical Interpretation of the Maximum Subarray Problem

Dennis Wei  IBM Research
Dmitry Malioutov  Millennium Management

We study a noisy localization problem inspired by the classical maximum subarray problem. While the naïve solution fails completely, penalized and constrained versions can succeed and are theoretically justified.

Maximum Subarray Problem
Given an array of numbers, find contiguous subarray with largest sum

Efficient O(N) algorithm by Kadane [1]

Some generalizations also have O(N) algorithms [2]

Applications:
• Biomolecular sequence analysis [2,3] • Image processing, computer vision

Penalized Maximum Subarray from Exponential Families
Assume \( w_1, \ldots, w_N \) i.i.d. – exponential family

Then maximum likelihood estimate of \( m, M \) reduces to penalized max subarray with optimal penalty

\[
\delta = \frac{A(m, \eta) - A(b, \eta)}{\eta_1 - \eta_0}
\]

Proposition: Penalty falls between interval mean and background mean

Example: Gaussian

Example: Poisson with rates \( \lambda_0, \lambda_1 \)

In practice, can set \( \delta \) based on prior knowledge of \( \mu_1 - \mu_0 \)

A Statistical Localization Problem
Sequence of random variables \( w_1, \ldots, w_N \)

Interval \( w_m, \ldots, w_M \) has mean \( \mu_1 \) different from background mean \( \mu_0 \)

Estimate \( m, M \) from one observation of \( w_m, \ldots, w_N \)

Penalized and Constrained Versions Succeed

1) Penalized:

\[
\delta = \arg \max_{w_0, \ldots, w_K} \sum_{j=0}^{K} (w_j - \delta)
\]

2) Constrained:

\[
\delta = \arg \max_{w_0, \ldots, w_K} \sum_{j=0}^{K} w_j \text{ s.t. } M - m + 1 \leq K
\]

1) is the Lagrangean of 2)

Localization Error Analysis

Lemma: For naïve max subarray \( \delta = 0 \), expected localization error

\[
\mathbb{E}[M - m | M \geq m] = \frac{N}{2}
\]

Lemma: For penalized version \( \delta > 0 \), error independent of \( N \)

Penalized vs. Constrained Formulations

Reference

[2] Y.-L. Lou et al. (2003), "Efficient algorithms for locating the length-constrained heaviest segments with applications to biological sequence analysis."
