A Statistical Interpretation of the Maximum Subarray Problem

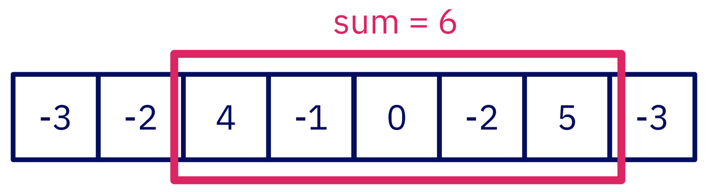
Dennis Wei *IBM Research*
Dmitry Malioutov *Millenium Management*

ICASSP 2023
4 - 10 JUNE, RHODES ISLAND, GREECE

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**

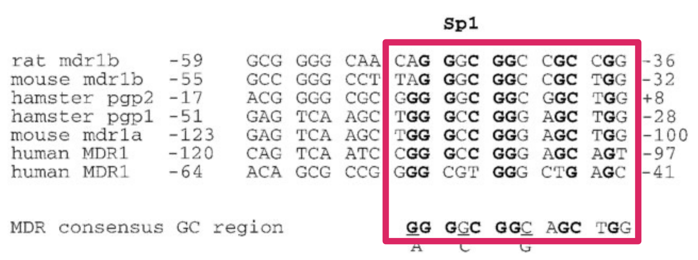


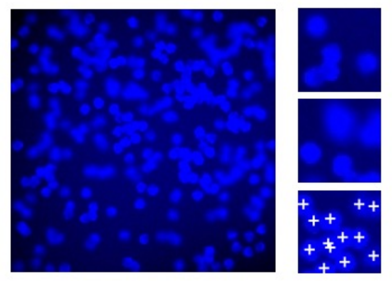
Efficient $O(N)$ algorithm by Kadane [1]

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

Applications:

- Biomolecular sequence analysis [2,3]
- Image processing, computer vision (2-D) [4]





Penalized Maximum Subarray from Exponential Families

Assume $w_1, \dots, w_N$ i.i.d. ~ exponential family

$$f(w_t) = h(w_t)\exp(\eta w_t + \eta'^T T(w_t) - A(\eta, \eta'))$$

- other sufficient statistics
- log-partition function
- natural parameter — interval: $\eta = \eta_1$; background: $\eta = \eta_0$
- w_t itself is one of the sufficient statistics

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

$$\delta = \frac{A(\eta_1, \eta') - A(\eta_0, \eta')}{\eta_1 - \eta_0}$$

Proposition: Penalty falls between interval mean and background mean

$$\mu_0 \le \delta \le \mu_1$$

Example: **Gaussian**

$$\delta = \frac{\mu_0 + \mu_1}{2}$$

Example: **Poisson** with rates λ_0, λ_1

$$\delta = \frac{\lambda_1 - \lambda_0}{\log \lambda_1 - \log \lambda_0}$$

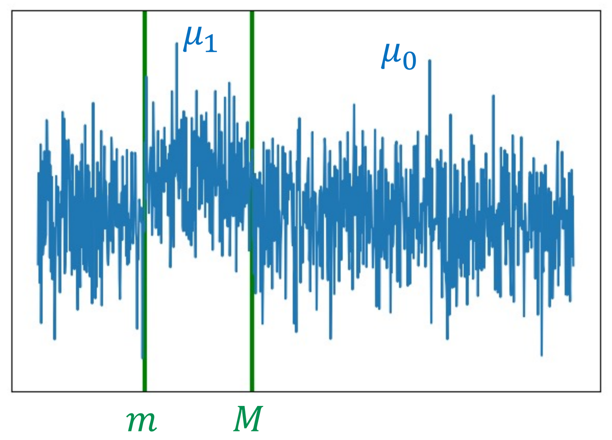
In practice, can set δ based on prior knowledge of $\mu_1 - \mu_0$

A Statistical Localization Problem

Sequence of random variables $w_1, \dots, w_N$

Interval $w_m, \dots, w_M$ has mean μ_1 different from background mean μ_0

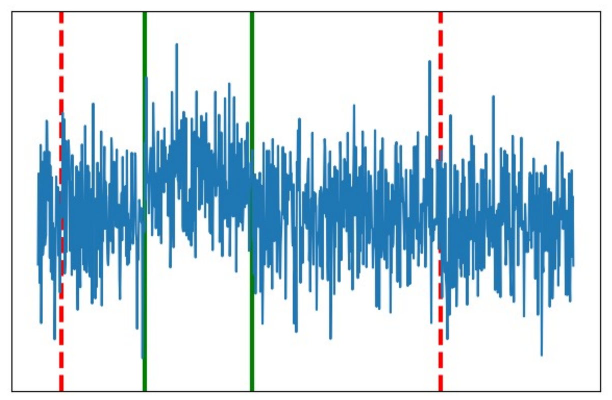
Estimate m, M from one observation of $w_1, \dots, w_N$



Naïve Maximum Subarray Fails Completely

Naïve maximum subarray:

$$\hat{m}, \hat{M} = \arg\max_{m,M} \sum_{m}^{M} w_t$$



Penalized and Constrained Versions Succeed

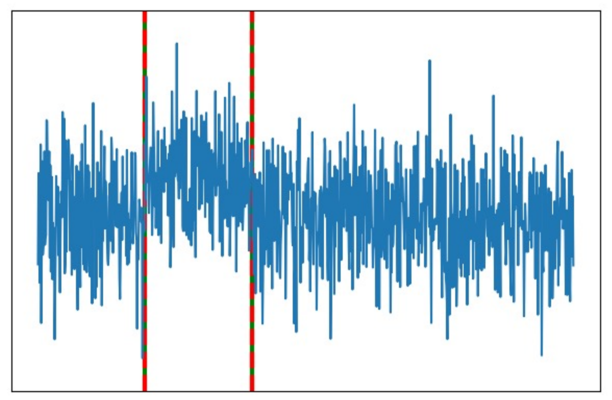
1) Penalized:

$$\hat{m}, \hat{M} = \arg\max_{m,M} \sum_{t=m}^{M} (w_t - \delta)$$

2) Constrained:

$$\hat{m}, \hat{M} = \arg\max_{m,M} \sum_{t=m}^{M} w_t \quad \text{s.t.} \quad M - m + 1 \le K$$

1) is the Lagrangean of 2)

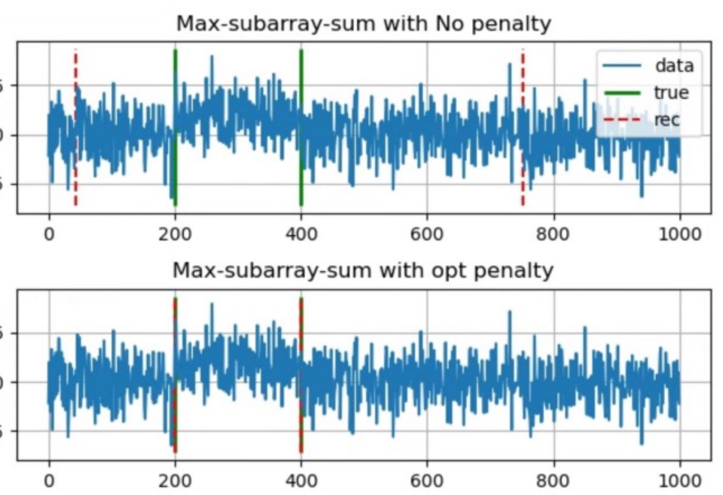


Localization Error Analysis

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

$$\mathbb{E}[\hat{M} - M \mid \hat{M} \ge M] = \frac{N - M}{2}$$

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

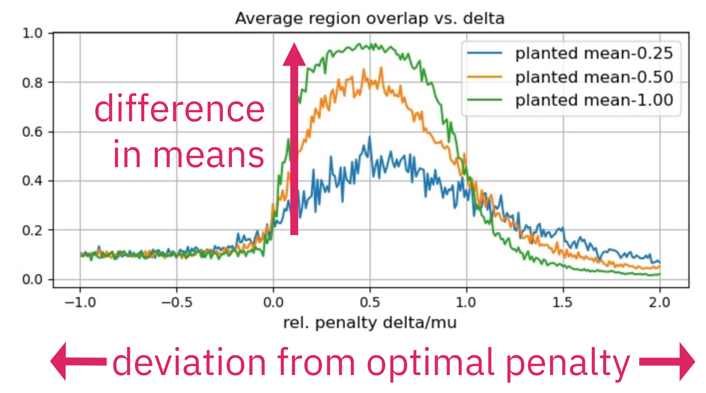


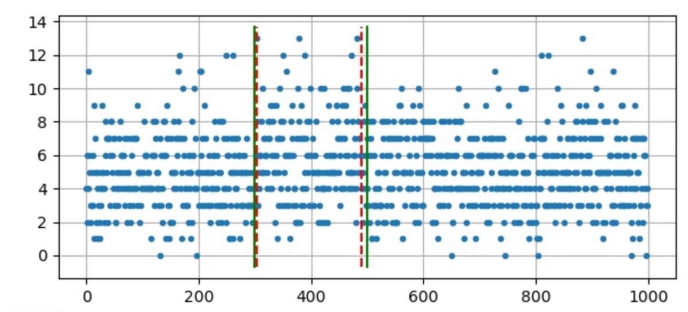
Numerical Simulations

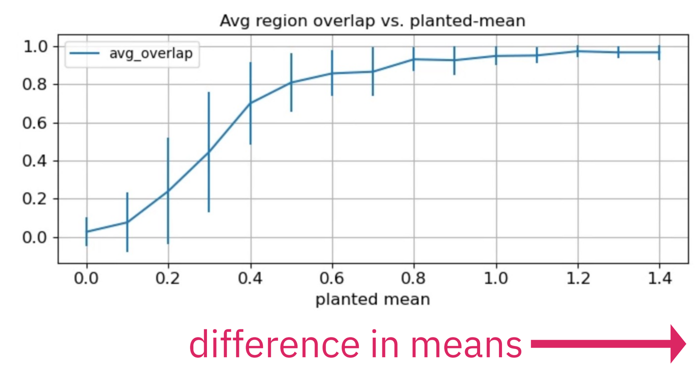
Recovery of Planted Intervals

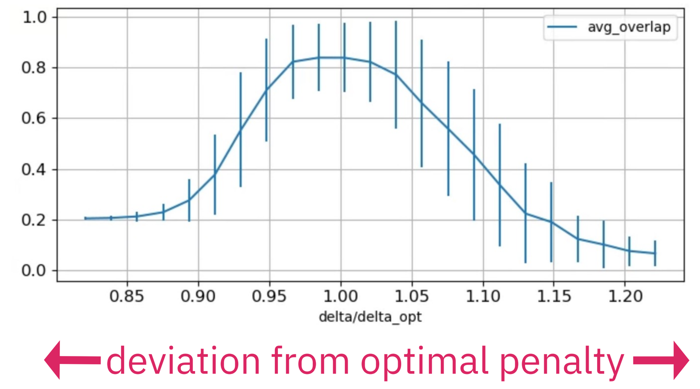
Gaussian ($\mu_0 = 0, \mu_1 = \mu$)

Poisson ($\lambda_0 = 5, \lambda_1 = 6$)

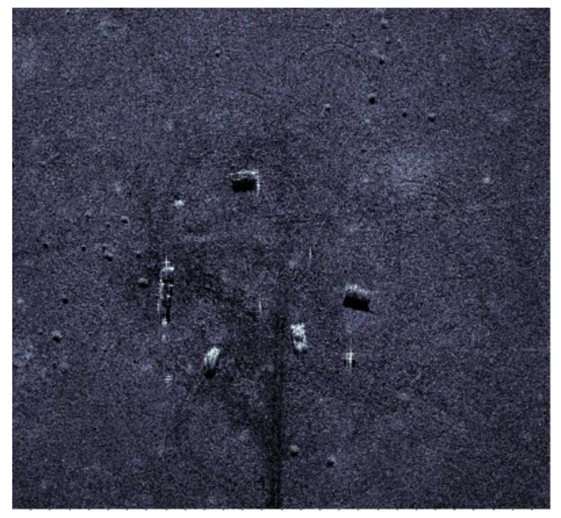


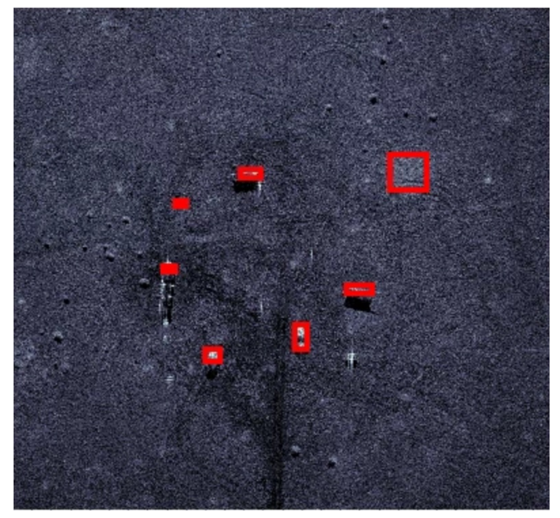




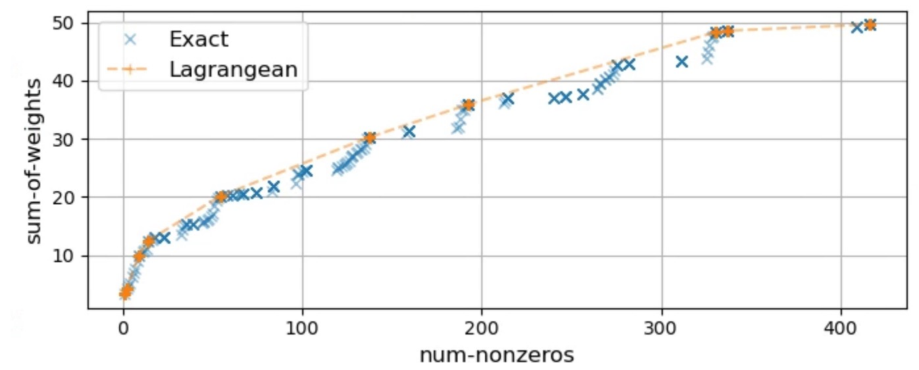


2-D Example: SAR Vehicle Localization





Penalized vs. Constrained Formulations



Penalized solutions appear to lie on convex hull of constrained solutions

References

[1] J. Bentley (1984), "Programming pearls: Algorithm design techniques." Communications of the ACM, 27(9):865-873.

[2] Y.-L. Lin et al. (2002), "Efficient algorithms for locating the length-constrained heaviest segments with applications to biomolecular sequence analysis." Journal of Computer and System Sciences, 65(3):570-586.

[3] J. V. Thottassery et al. (1999), "Sp1 and Egr-1 have opposing effects on the regulation of the rat *Pgp2/mdr1b* gene." J. Biol. Chem., 274(5):3199-3206.

[4] V. Lempitsky and A. Zisserman (2010), "Learning to count objects in images." Advances in Neural Information Processing Systems (NeurIPS).