A Statistical Interpretation of the Maximum Subarray Problem

—

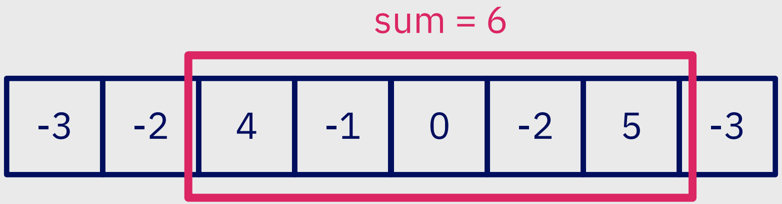
ICASSP 2023

Dennis Wei
IBM Research

Dmitry Malioutov
Millenium Management

Maximum Subarray Problem

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



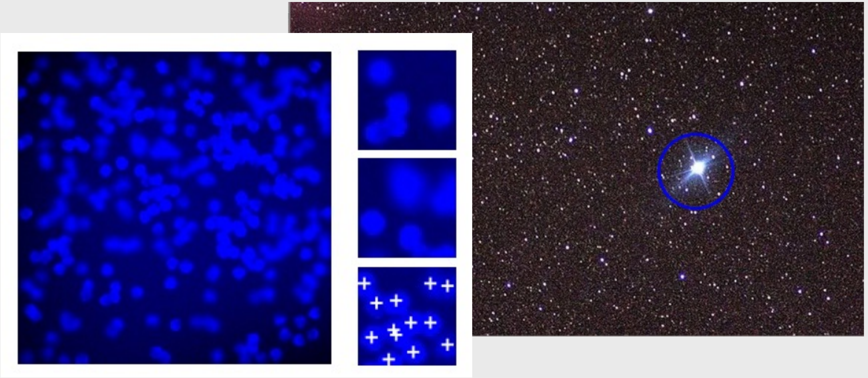
Applications:

- Biomolecular sequence analysis

					Sp1				
rat mdr1b	-59	GCG	GGG	CAA	CAG	GGC	GGC	CGC	CGG	-36
mouse mdr1b	-55	GCC	GGG	CCT	TAG	GGC	GGC	CGC	TGG	-32
hamster pgp2	-17	ACG	GGG	CGC	GGG	GGC	GGC	GGC	TGG	+8
hamster pgp1	-51	GAG	TCA	AGC	TGG	GCC	GGG	AGC	TGG	-28
mouse mdr1a	-123	GAG	TCA	AGC	TGG	GCC	GGG	AGC	TGG	-100
human MDR1	-120	CAG	TCA	ATC	CGG	GCC	GGG	AGC	AGT	-97
human MDR1	-64	ACA	GCG	CCG	GGG	CGT	GGG	CTG	AGC	-41
MDR consensus GC region					GG	GGC	GGC	AGC	TGG	
					A	C	G			

Thottassery et al. (1999), *J. Biol. Chem.* 274(5):3199-206

- Image processing, computer vision (2-D)



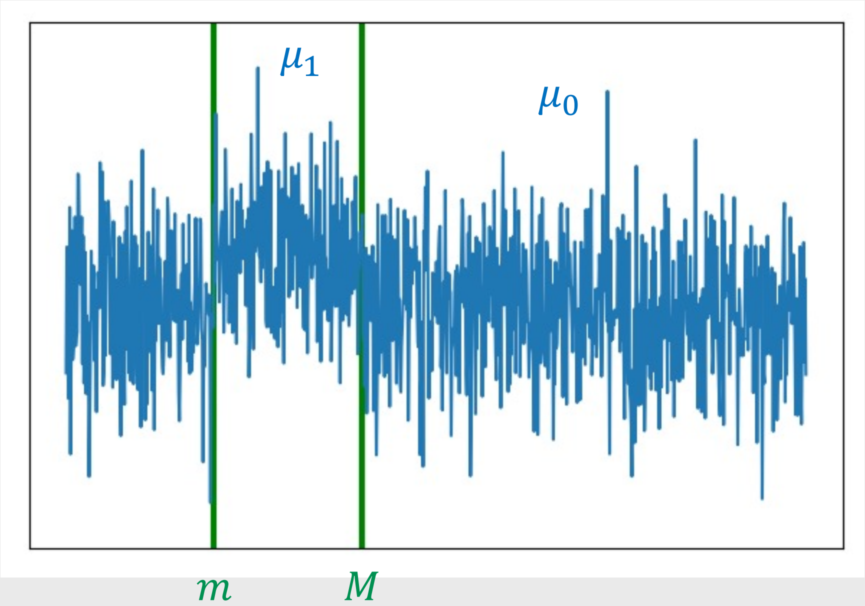
Lempitsky & Zisserman (2010), *NeurIPS*

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

Localize the interval (estimate m, M) from observation of $w_1, \dots, w_N$



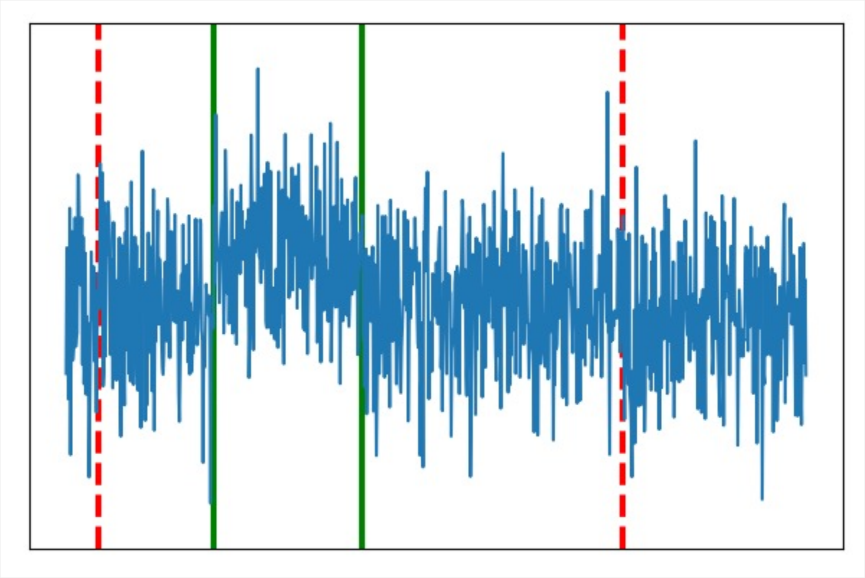
Motivating Experiment

Apply maximum subarray:

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

Efficient $O(N)$ algorithm by Kadane

Fails at localization!



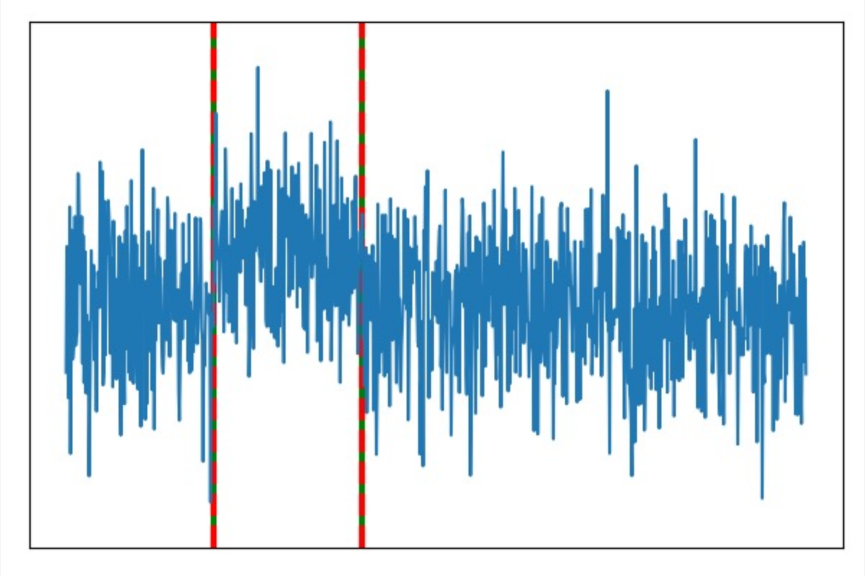
Fixing the Failure

1) Penalized maximum subarray

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

2) Constrained maximum subarray

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



Focus on 1) in this talk

See paper for Lagrangean + convex hull relationship between 1) and 2)

Penalized Max Subarray from Exponential Families

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

other sufficient statistics

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

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 boundaries m, M reduces to penalized max subarray

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

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

Penalty Value for Exponential Families

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

Proposition: Penalty falls between interval mean and background mean

$$\mu_0 \leq \delta \leq \mu_1$$

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

Example: Gaussian

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

Example: Poisson with rates λ_0, λ_1

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

Localization Analysis

Penalized maximum subarray

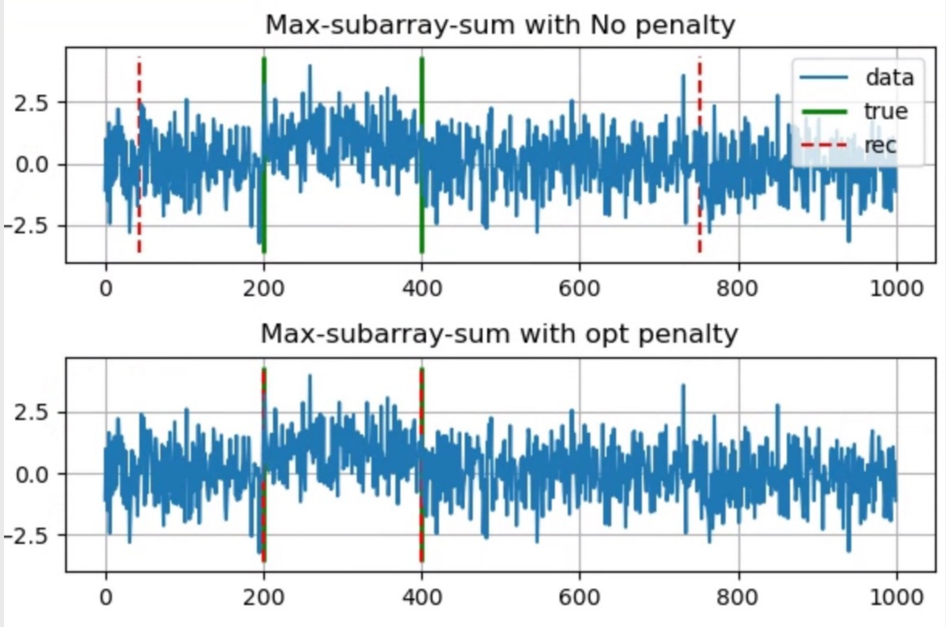
$$\widehat{m}, \widehat{M} = \arg\max_{m,M} \sum_{m}^{M} (w_t - \delta)$$

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

$$\mathbb{E}\left[\widehat{M} - M \mid \widehat{M} \geq M\right] = \frac{N - M}{2}$$

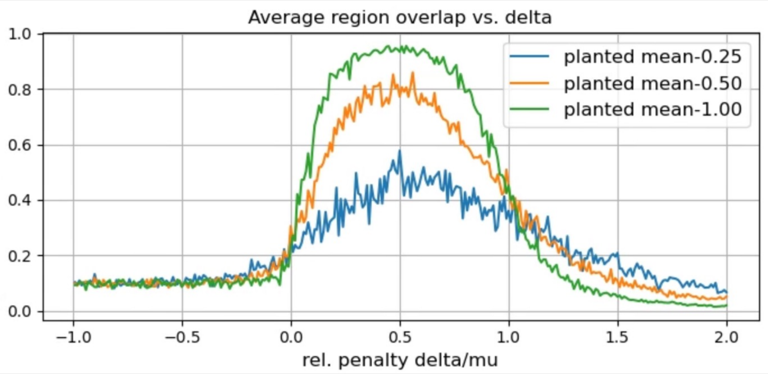
length of array

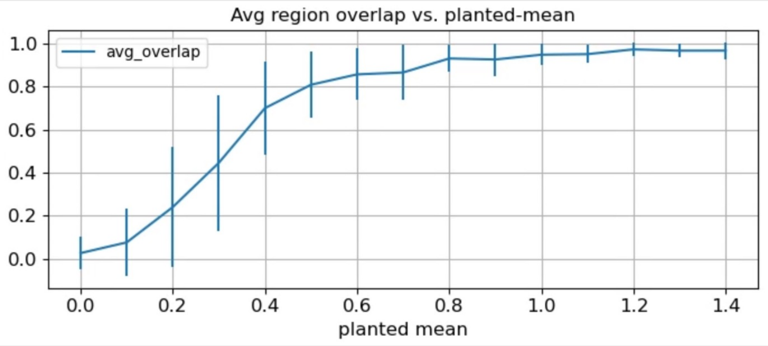
Lemma: For $\delta > 0$, error independent of N



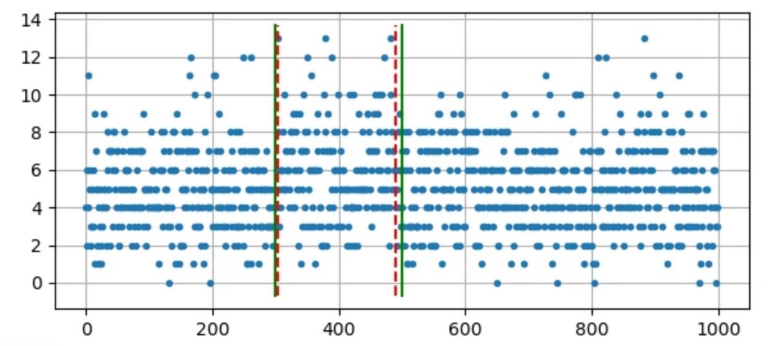
Numerical Simulations

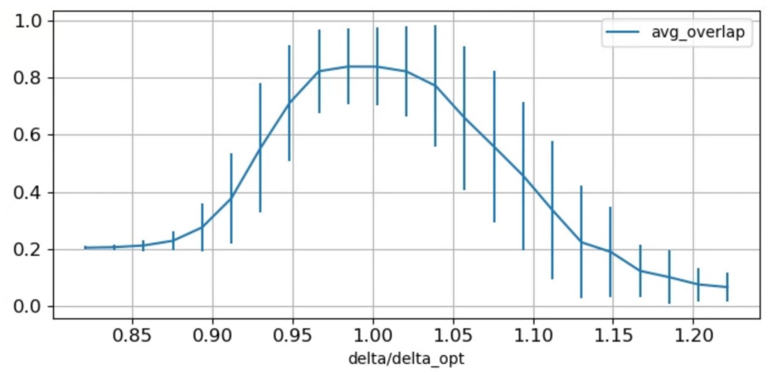
Gaussian $(\mu_0 = 0, \mu_1 = \mu)$



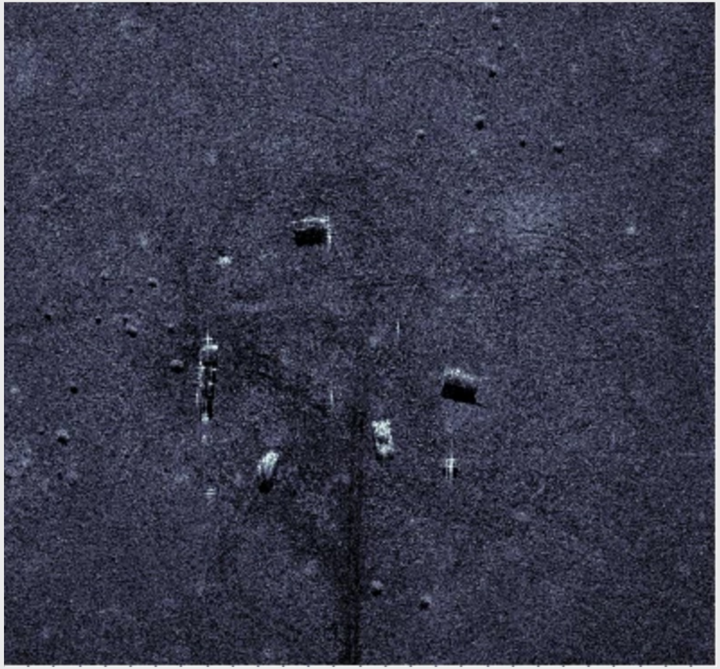


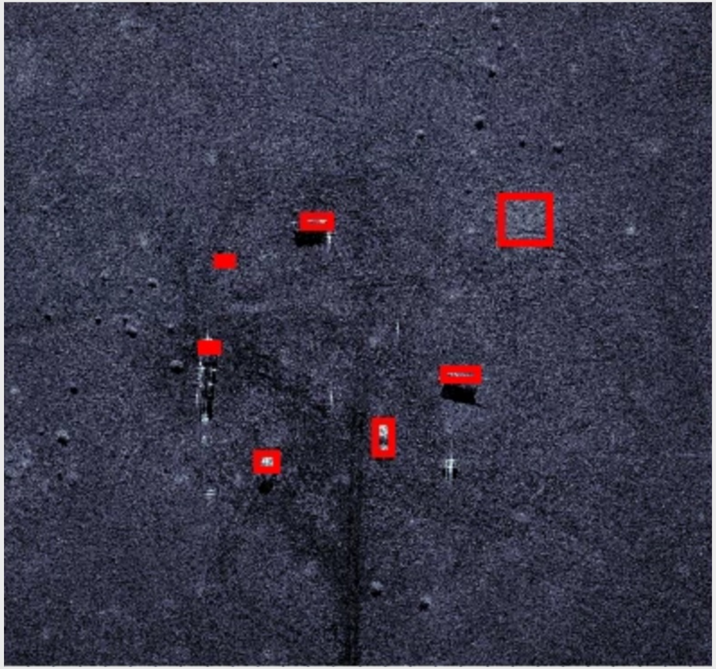
Poisson $(\lambda_0 = 5, \lambda_1 = 6)$





Works in 2-D Also





Summary

Statistical localization problem inspired by maximum subarray

Naïve max subarray fails to localize while penalized and constrained versions succeed

Penalized version results from exponential families

Paper: https://arxiv.org/abs/2304.13307