Optimistic Search Strategy: Change-point Detection for Large-scale Data via Adaptive Logarithmic Queries

Solt Kovács

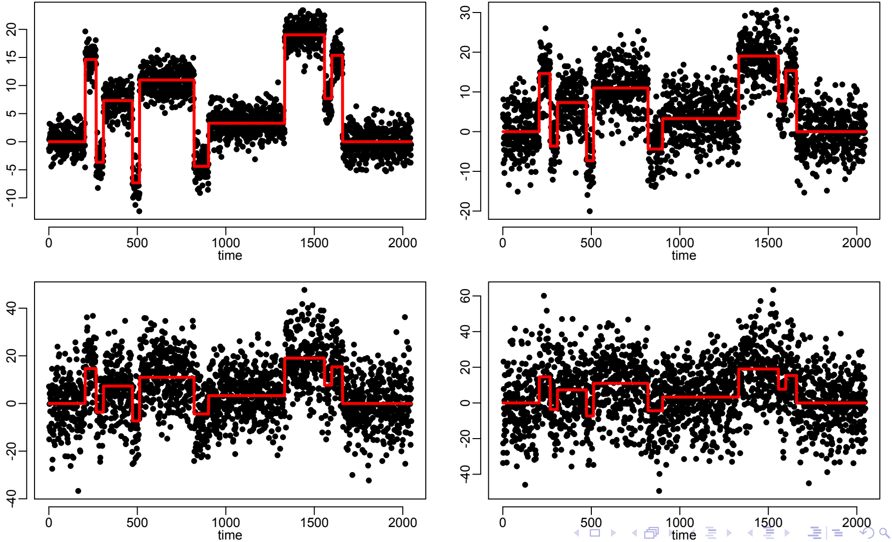
Seminar for Statistics
ETH Zürich

Brighton, 2022

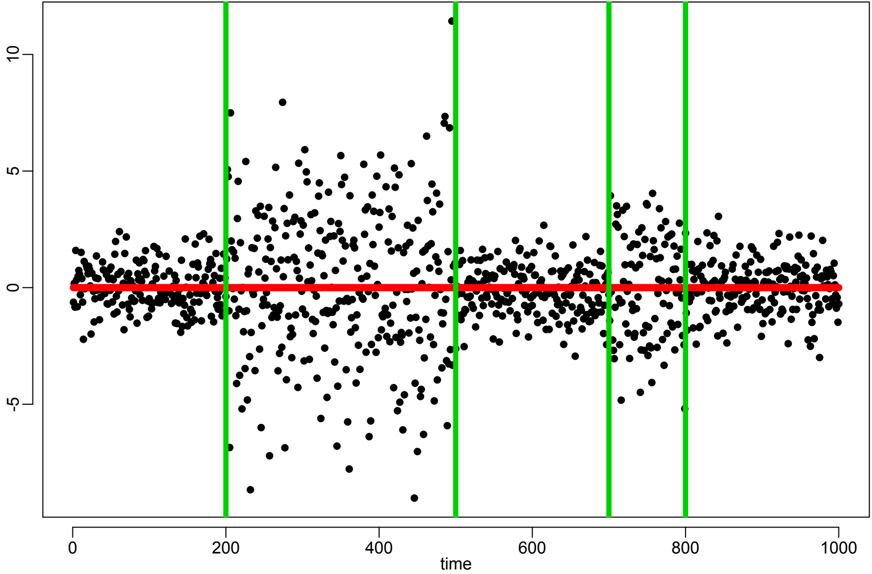
joint work with

- Housen Li (U. Göttingen)
- Lorenz Haubner (ETH Zürich)
- Axel Munk (U. Göttingen)
- Peter Bühlmann (ETH Zürich)

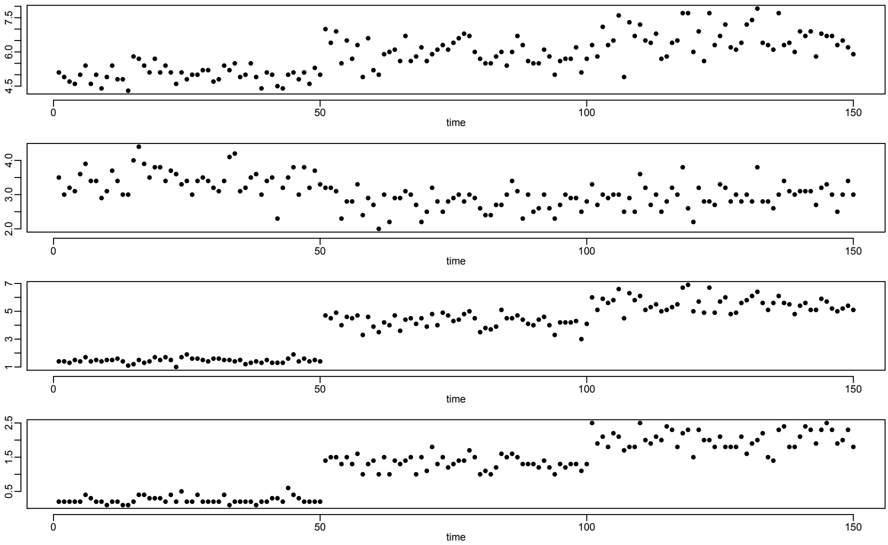
(offline) change point detection



Another example

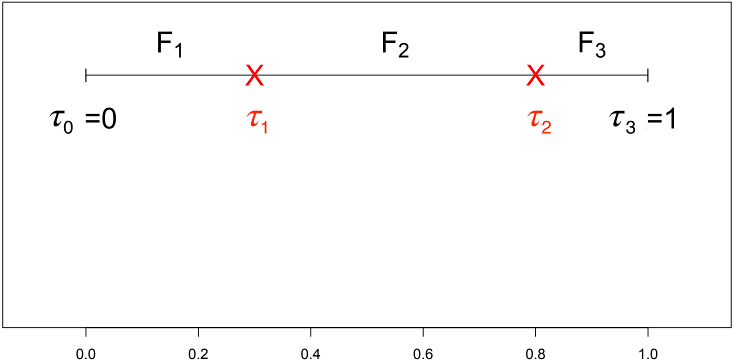


Last example



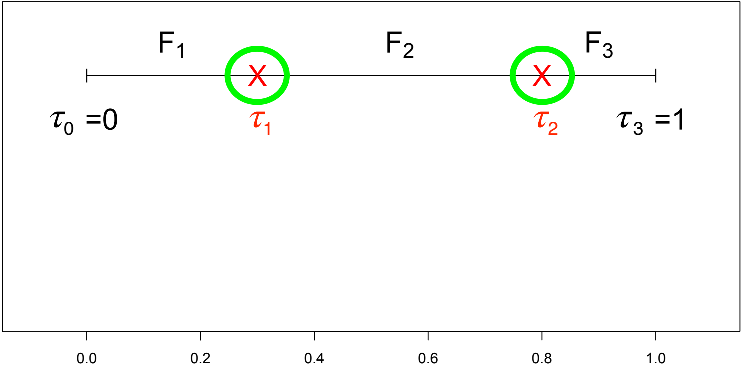
Change point detection in general

- observe ordered $X_i \in \mathbb{R}^p, i = 1, \ldots, T$
- notation: we map the T observations to $[0, 1]$



Goals of change point detection

- estimate the number of change points κ
- estimate the location of change points $\tau_1, \ldots, \tau_\kappa$



Change points in big data - challenges

- algorithms/optimization
- (non)parametric assumptions
- missing values
- dependence
- ...

Change points in big data - challenges

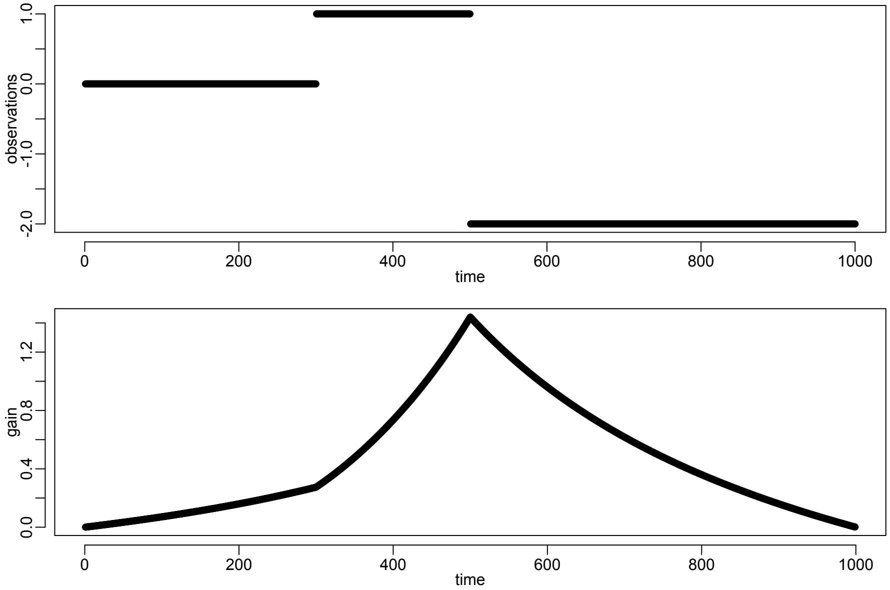
- algorithms/optimization
- (non)parametric assumptions
- missing values
- dependence
- ...

Finding change points

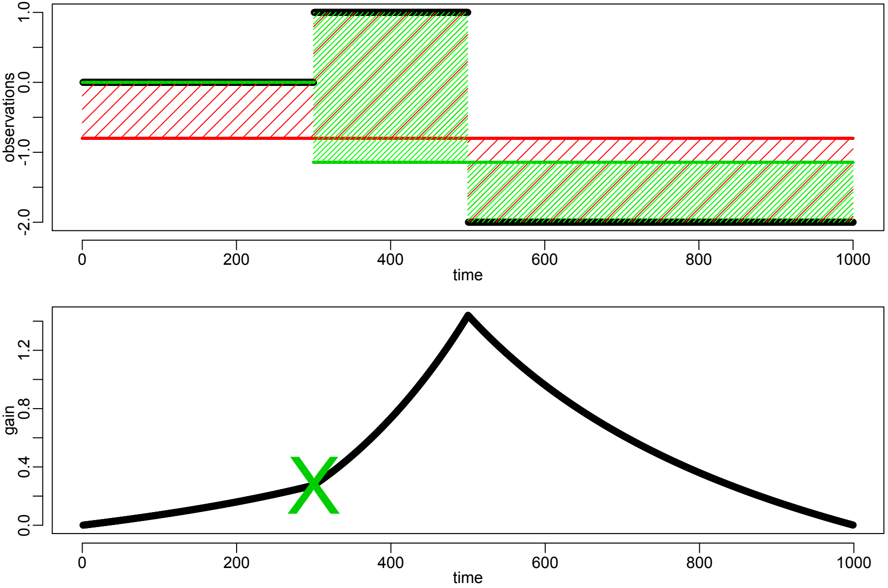
How do we search usually?

– Full grid of all possible split points $2, \ldots, T$

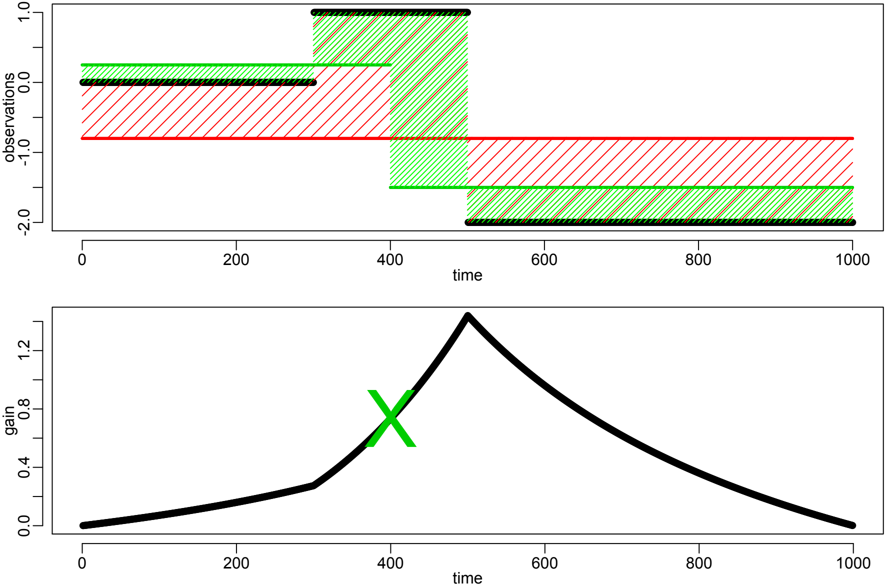
Finding change points



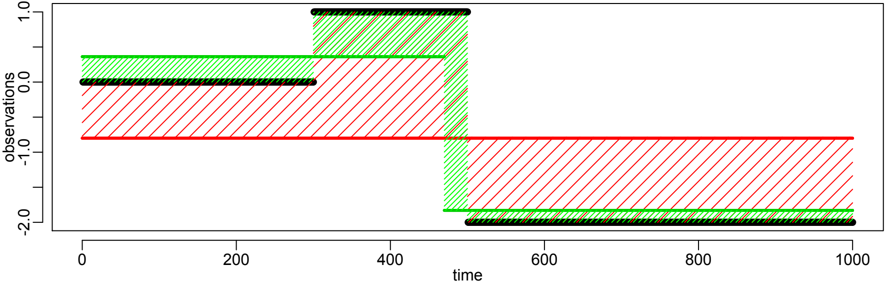
Finding change points

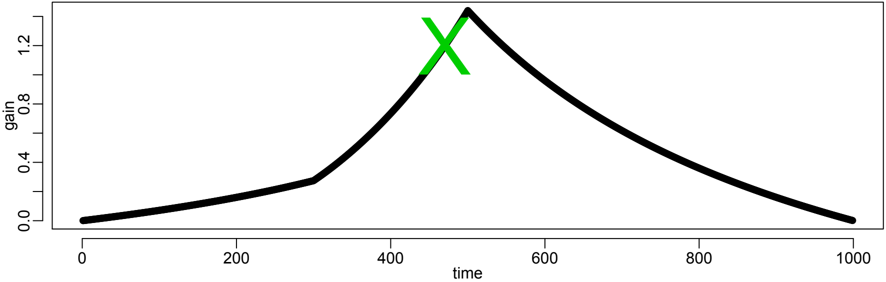


Finding change points

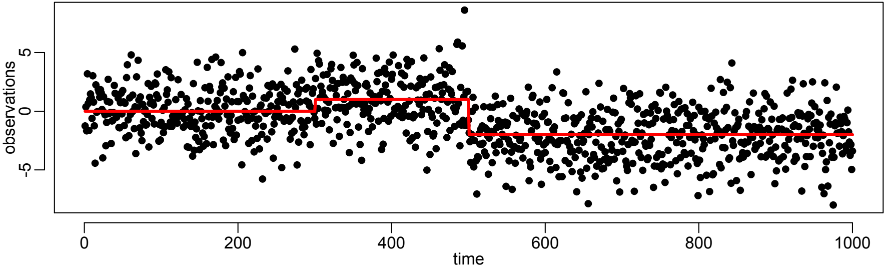


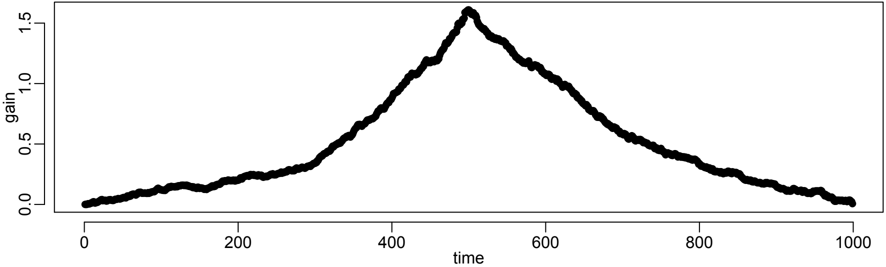
Finding change points



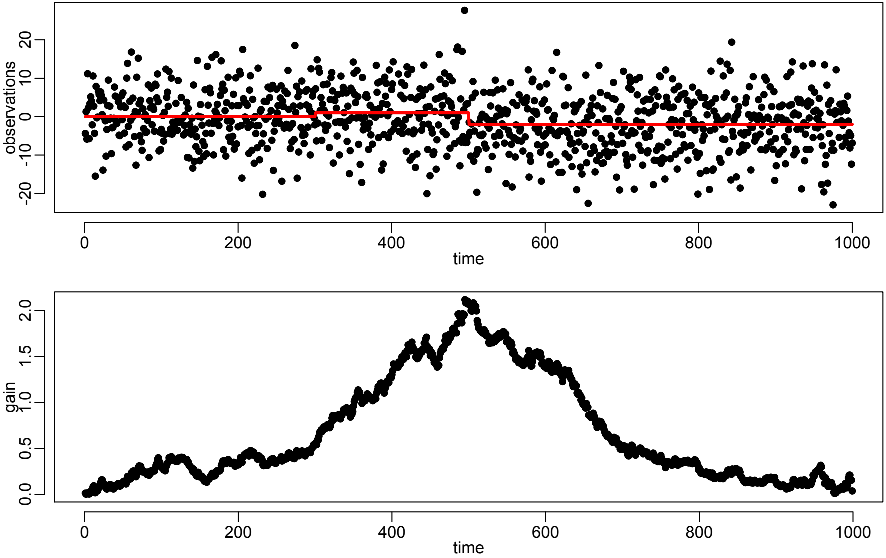


Finding change points





Finding change points



Searching for the best split - Motivation

Assume for now a **single change point**!

How do we search usually?
– Full grid of all possible split points $2, \ldots, T$

What if model fits are expensive?
– e.g. graphical Lasso, Lasso, neural network, Random Forest, ...
– Infeasible computationally (for large T)

Searching for the best split - Motivation

Assume for now a **single change point**!

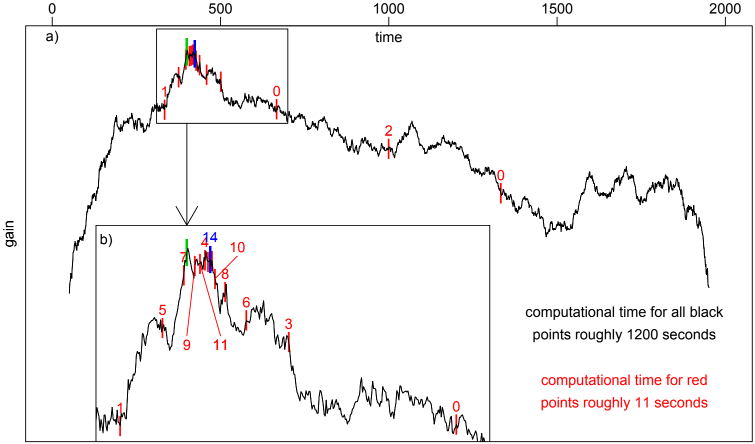
How do we search usually?
– Full grid of all possible split points $2, \ldots, T$

What if model fits are expensive?
– e.g. graphical Lasso, Lasso, neural network, Random Forest, ...
– Infeasible computationally (for large T)

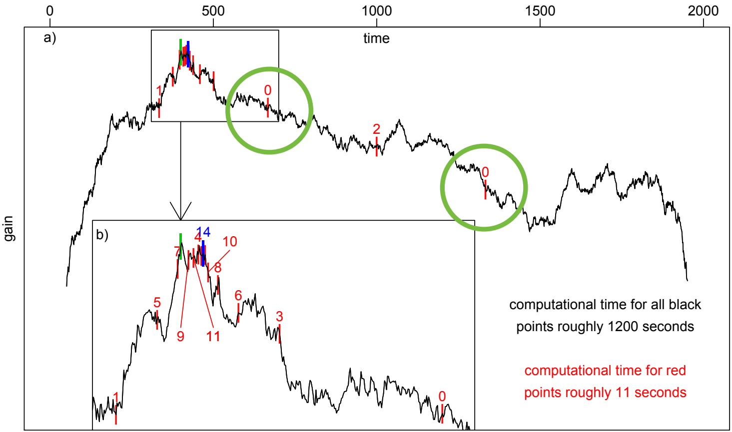
Is it really necessary to consider the full grid?
– No: Optimistic Search (OS) strategies with only $O(\log T)$ evaluations!

Searching for the best split - An example

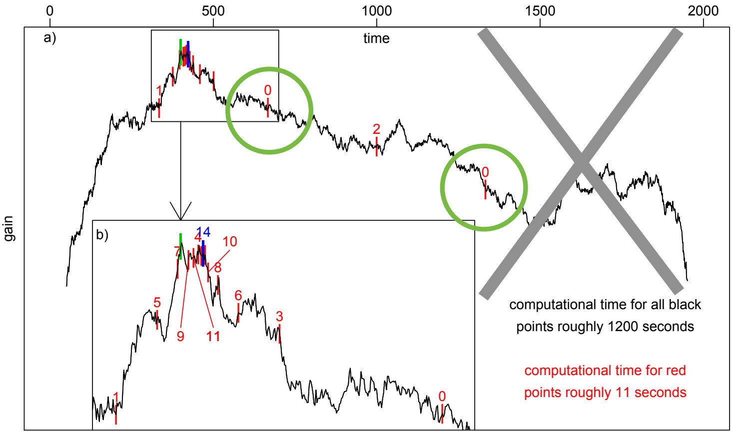
full grid search vs. (naive) Optimistic Search



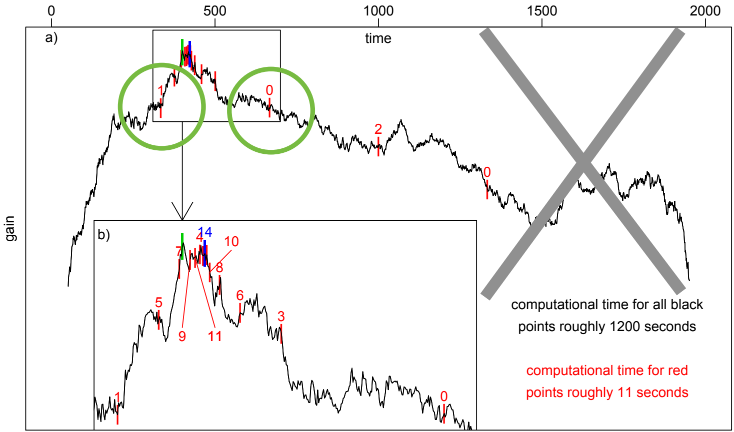
What happens?



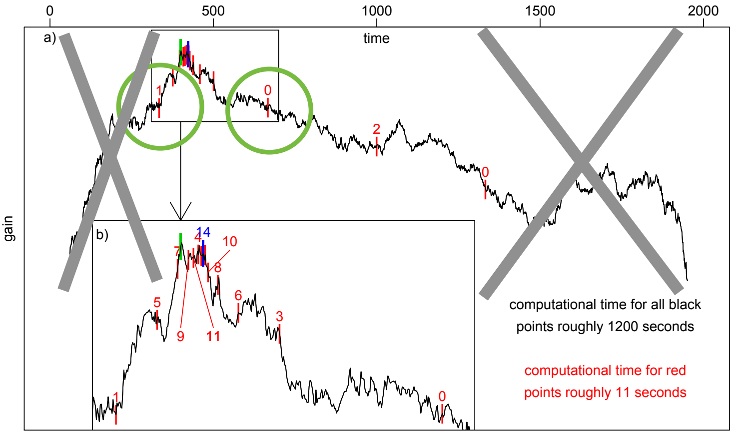
What happens?



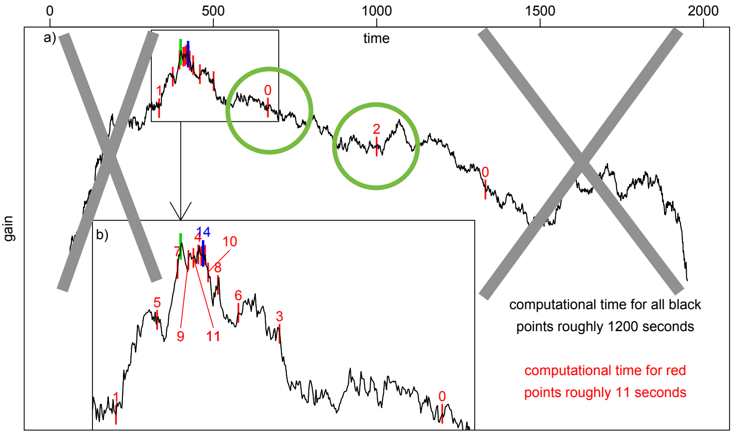
What happens?



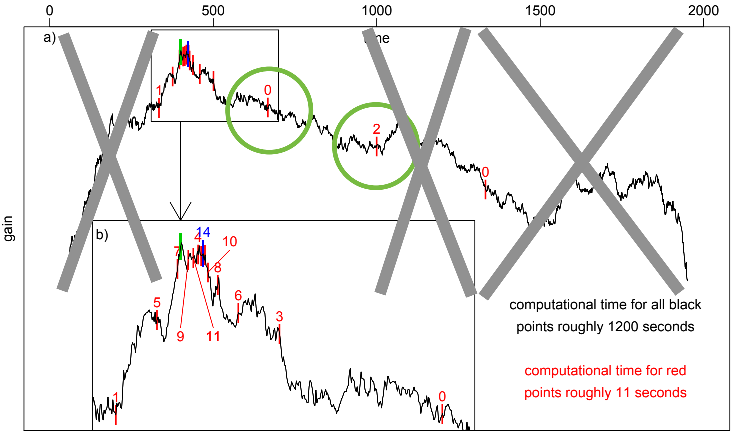
What happens?



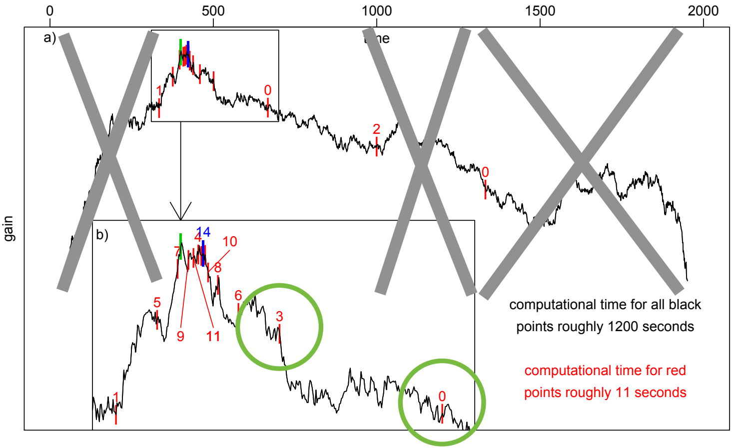
What happens?



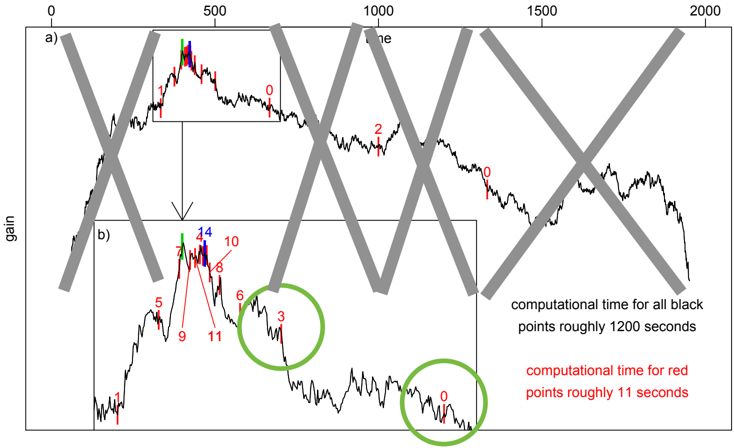
What happens?



What happens?



What happens?



What happens? - Summary

At each step:

- evaluate a point in the middle of the remaining longer segment
- compare the gains
- keep the more promising side (with higher gain)

What happens? - Summary

At each step:

- evaluate a point in the middle of the remaining longer segment
- compare the gains
- keep the more promising side (with higher gain)

Why only $O(\log T)$ evaluations?

What happens? - Summary

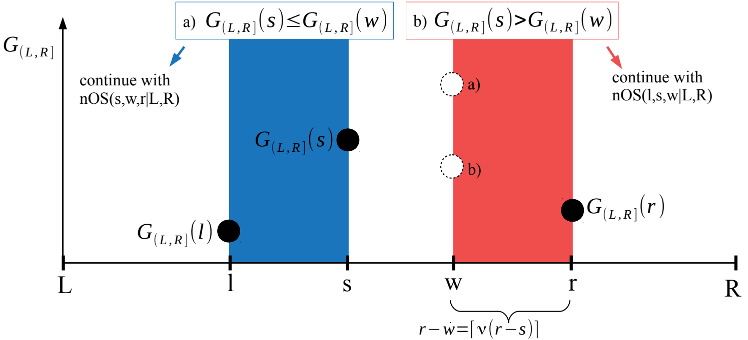
At each step:

- evaluate a point in the middle of the remaining longer segment
- compare the gains
- keep the more promising side (with higher gain)

Why only $O(\log T)$ evaluations?

- Guaranteed to discard at least $1/4$ of the current search interval in each step

What happens? - Visually



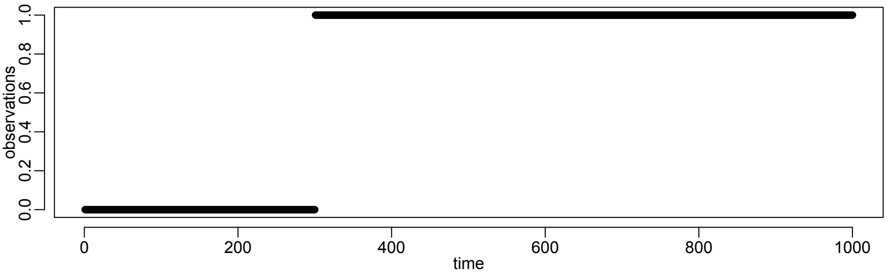
Procedure similar to Golden section search [Kiefer, 1953, Avriel and Wilde, 1966, Avriel and Wilde, 1968]

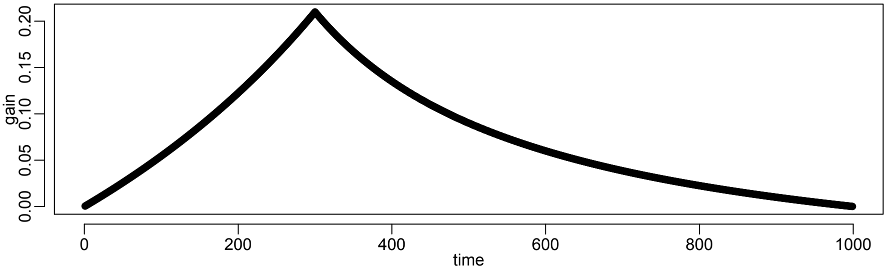
Any guarantees?

Population/noiseless cases:

- For a unimodal function (e.g. single change point case), naive OS returns global maximum in $O(\log T)$ steps.

Population/noiseless case





Any guarantees?

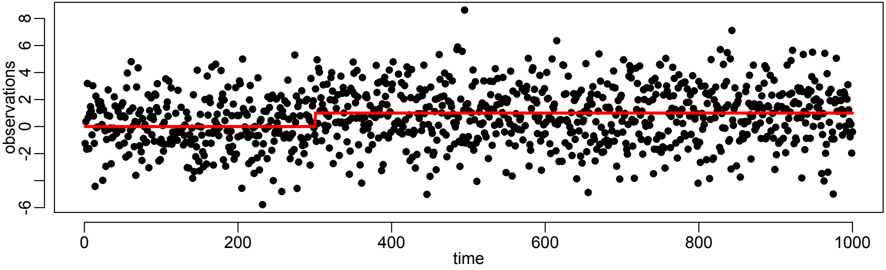
Population/noiseless cases:

- For a unimodal function (e.g. single change point case), naive OS returns global maximum in $O(\log T)$ steps.

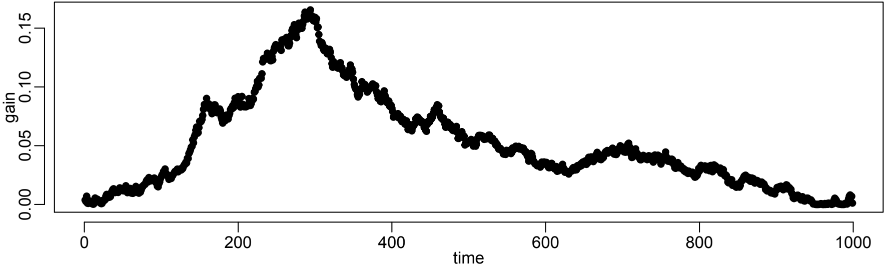
What about noisy cases?

- which model?

Noisy cases



observations
8
6
4
2
0
-2
-6
0
200
400
600
800
1000
time



gain
0.15
0.10
0.05
0.00
0
200
400
600
800
1000
time

univariate Gaussian changes in mean

Assume independent observations $X_1, \ldots, X_T$ and that

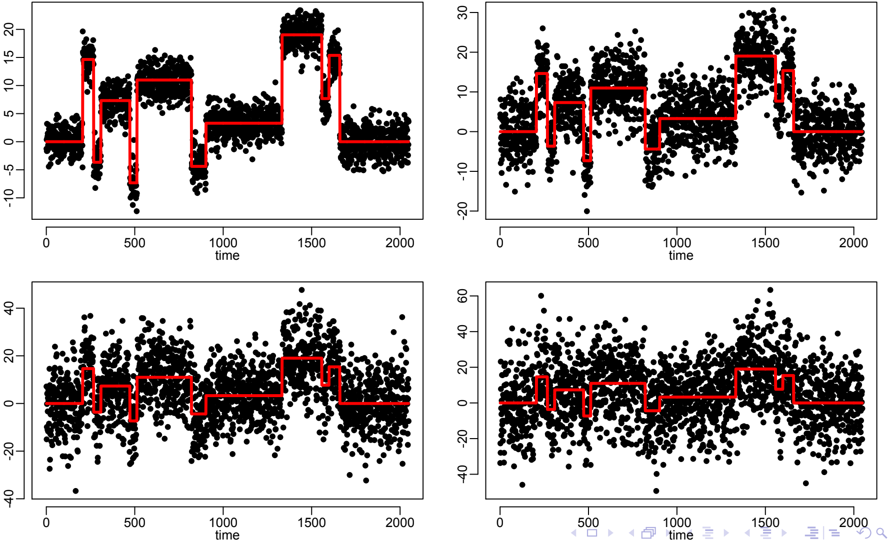
$$X_{\tau_0 T+1}(= X_1), \ldots, X_{\tau_1 T} \sim \mathcal{N}(\mu_0, \sigma^2) \qquad (= F_0)$$
$$\vdots$$
$$X_{\tau_\kappa T+1}, \ldots, X_{\tau_{\kappa+1} T}(= X_T) \sim \mathcal{N}(\mu_\kappa, \sigma^2) \qquad (= F_\kappa)\,,$$

where $\{\tau_i \,:\, i = 1, \ldots, \kappa\}$ gives the location of change points satisfying

$$0 = \tau_0 < \tau_1 < \cdots < \tau_{\kappa+1} = 1 \quad \text{and} \quad \tau_i T \in \mathbb{N}\,,$$

means $\mu_i \neq \mu_{i-1}$ for $i = 1, \ldots, \kappa$ give the levels on segments, and the common standard deviation $\sigma > 0$ is known. Assume w.l.o.g. $\sigma = 1$.

univariate Gaussian changes in mean



univariate Gaussian changes in mean

Define the minimal segment length λ as

$$\lambda \equiv \lambda_T = \min_{i=0,\ldots,\kappa} (\tau_{i+1} - \tau_i) \, ,$$

and the minimal jump size δ as

$$\delta \equiv \delta_T = \min_{i=1,\ldots,\kappa} \delta_i \qquad \text{with} \quad \delta_i = |\mu_i - \mu_{i-1}| \; .$$

univariate Gaussian changes in mean

We use the CUSUM statistics as "gain":

$$\mathrm{CS}_{(l,r]}(s) = \sqrt{\frac{r-s}{n(s-l)}}\sum_{t=l+1}^{s} X_t - \sqrt{\frac{s-l}{n(r-s)}}\sum_{t=s+1}^{r} X_t \,,$$

with integers $0 \le l < s < r \le T$ and $n = r - l$.

Theory for naive OS

- For the univariate Gaussian change in mean model above with a **single** change point (i.e. $\kappa = 1$), using CUSUM statistics as “gain”:

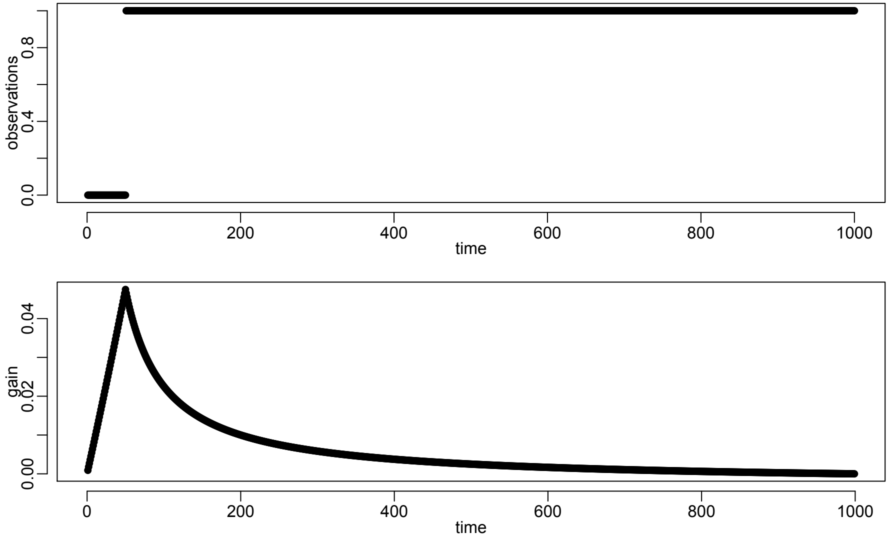
  Naive OS is consistent with **optimal localization error** if the ratio of shorter versus longer segment is not “too unbalanced”.

Theory for naive OS

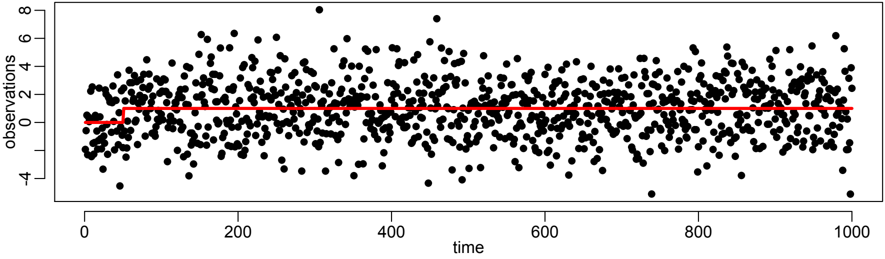
- For the univariate Gaussian change in mean model above with a **single** change point (i.e. $\kappa = 1$), using CUSUM statistics as "gain":

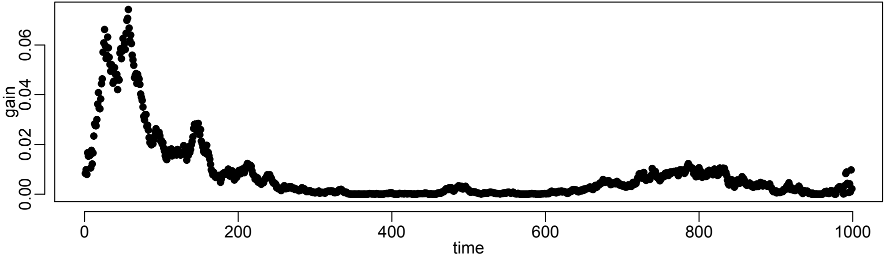
  Naive OS is consistent with **optimal localization error** if the ratio of shorter versus longer segment is not "too unbalanced". Formally, require $\delta\lambda\sqrt{T} \geq C_0\sqrt{\log T}$.

Failure of naive OS

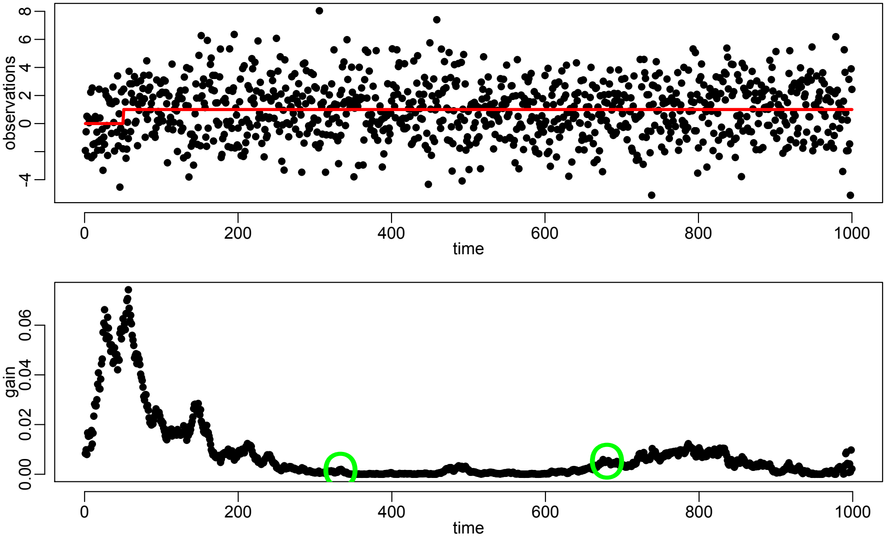


Failure of naive OS

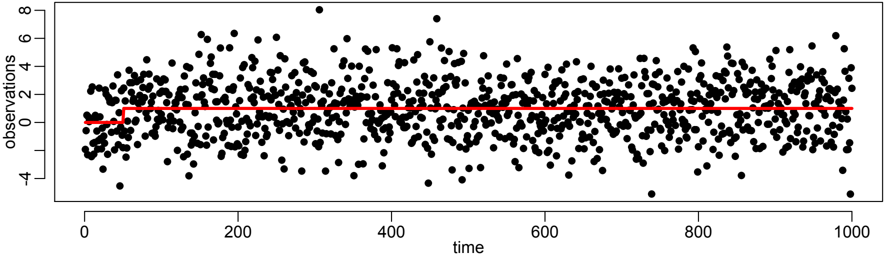


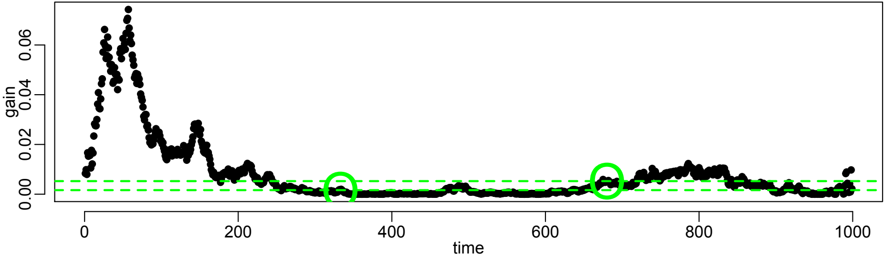


Failure of naive OS



Failure of naive OS





Theory for naive OS

- For the univariate Gaussian change in mean model above with a **single** change point (i.e. $\kappa = 1$), using CUSUM statistics as "gain":

  Naive OS is consistent with **optimal localization error** if the ratio of shorter versus longer segment is not "too unbalanced". Formally, require $\delta\lambda\sqrt{T} \geq C_0\sqrt{\log T}$.

  Suboptimal compared to weakest possible condition $\delta\sqrt{\lambda}\sqrt{T} \geq \sqrt{\log \log T}$ (see [Liu et al., 2019]).

Theory for naive OS

- For the univariate Gaussian change in mean model above with a **single** change point (i.e. $\kappa = 1$), using CUSUM statistics as "gain":

  Naive OS is consistent with **optimal localization error** if the ratio of shorter versus longer segment is not "too unbalanced". Formally, require $\delta\lambda\sqrt{T} \geq C_0\sqrt{\log T}$.

  Suboptimal compared to weakest possible condition $\delta\sqrt{\lambda}\sqrt{T} \geq \sqrt{\log\log T}$ (see [Liu et al., 2019]).

- Can we improve?

Theory for naive OS

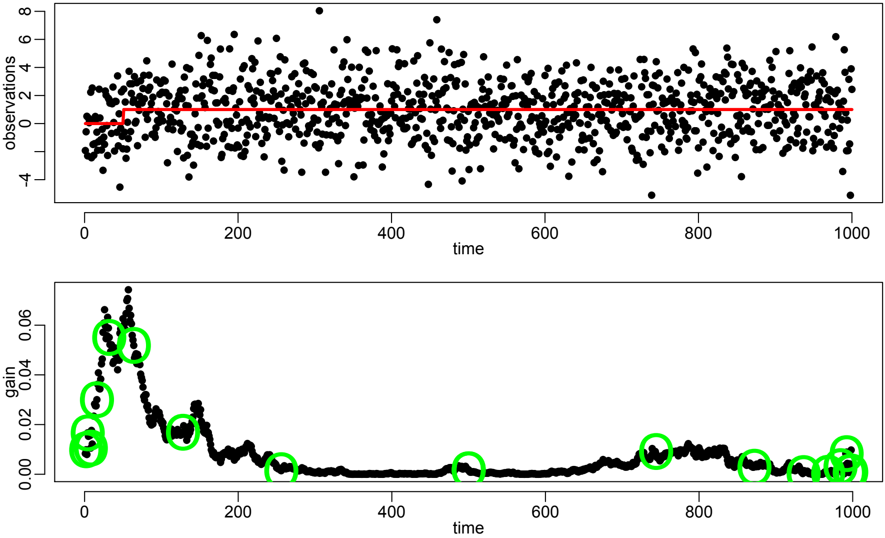
- For the univariate Gaussian change in mean model above with a **single** change point (i.e. $\kappa = 1$), using CUSUM statistics as "gain":

  Naive OS is consistent with **optimal localization error** if the ratio of shorter versus longer segment is not "too unbalanced". Formally, require $\delta\lambda\sqrt{T} \geq C_0\sqrt{\log T}$.

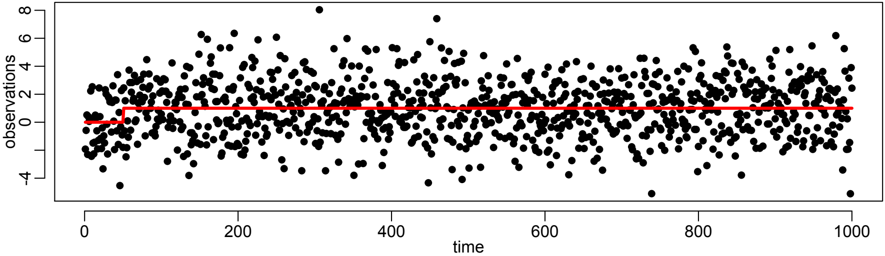
  Suboptimal compared to weakest possible condition $\delta\sqrt{\lambda}\sqrt{T} \geq \sqrt{\log \log T}$ (see [Liu et al., 2019]).

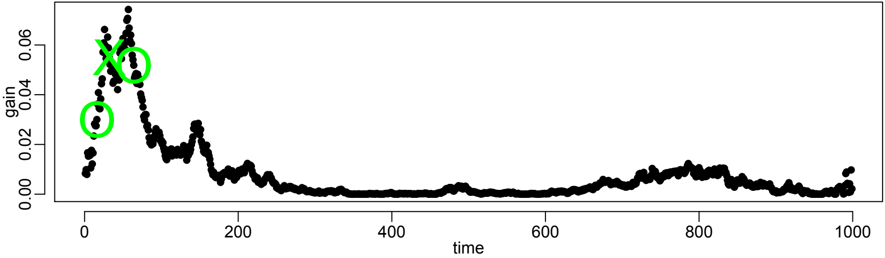
- Can we improve? Yes, **advanced Optimistic Search**!

Advanced Optimistic Search



Advanced Optimistic Search





advanced Optimistic Search

Idea:

- motivated by [Liu et al., 2019] and [Kovács et al., 2020], check dyadic points $\{2, 4, 8, 16, \ldots, T-16, T-8, T-4, T-2\}$
- around maximum select a suitable starting region
- do naive OS in the selected region

advanced Optimistic Search

Idea:

- motivated by [Liu et al., 2019] and [Kovács et al., 2020], check dyadic points $\{2, 4, 8, 16, \dots, T-16, T-8, T-4, T-2\}$
- around maximum select a suitable starting region
- do naive OS in the selected region

Theory:

- For the univariate Gaussian change in mean model above with a **single** change point (i.e. $\kappa = 1$) the **advanced OS** is (nearly) **minimax optimal**.

advanced Optimistic Search

Idea:

- motivated by [Liu et al., 2019] and [Kovács et al., 2020], check dyadic points $\{2, 4, 8, 16, \ldots, T-16, T-8, T-4, T-2\}$
- around maximum select a suitable starting region
- do naive OS in the selected region

Theory:

- For the univariate Gaussian change in mean model above with a **single** change point (i.e. $\kappa = 1$) the **advanced OS** is (nearly) **minimax optimal**. Formally, we require $\delta\sqrt{\lambda T} \geq C_0\sqrt{\log T}$.

Multiple change points?

Combinations possible with

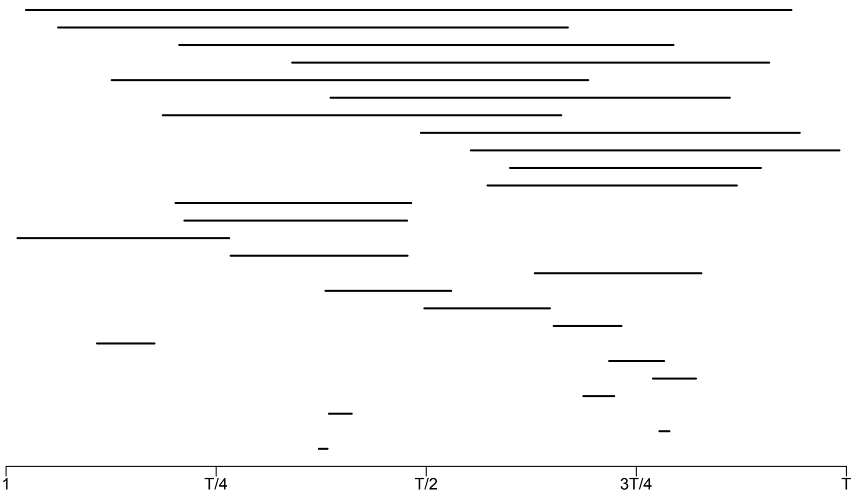
- Binary Segmentation [Vostrikova, 1981]
- Seeded Binary Segmentation [Kovács et al., 2020]
- Wild Binary Segmentation [Fryzlewicz, 2014]
- Circular Binary Segmentation [Olshen et al., 2004]
- ...

Multiple change points?

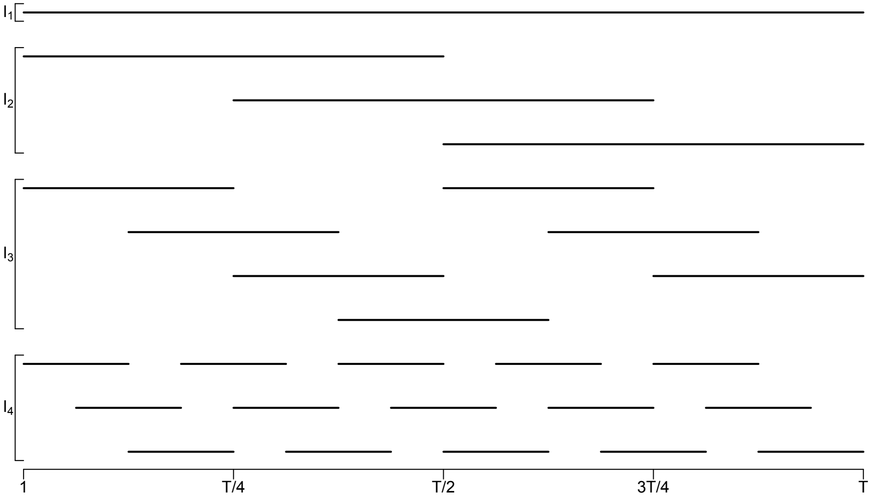
Combinations possible with

- Binary Segmentation [Vostrikova, 1981]
- Seeded Binary Segmentation [Kovács et al., 2020]
- Wild Binary Segmentation [Fryzlewicz, 2014]
- Circular Binary Segmentation [Olshen et al., 2004]
- ...

Random intervals [Fryzlewicz, 2014]



Seeded intervals [Kovács et al., 2020]



Multiple change points?

Optimistic Seeded Binary Segmentation (OSeedBS, with narrowest over threshold selection):

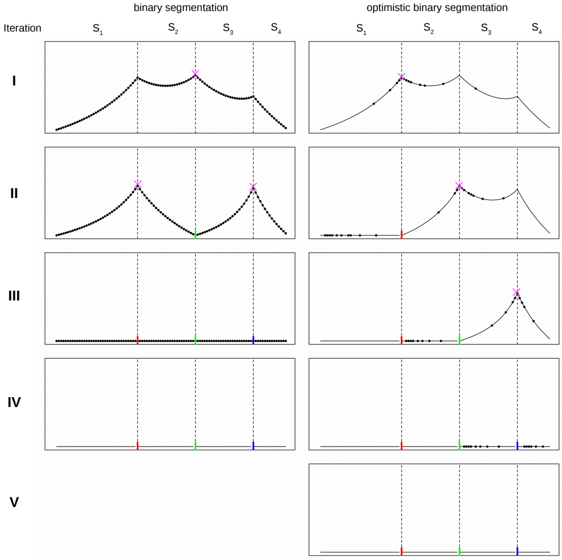
- **Minimax optimality** and **worst case** $O(T)$ computational cost for the above univariate Gaussian change in mean model with multiple change points, i.e. $\kappa > 1$
- **Sublinear** computational **cost possible** under additional assumptions

Optimistic Seeded BS (OSeedBS)

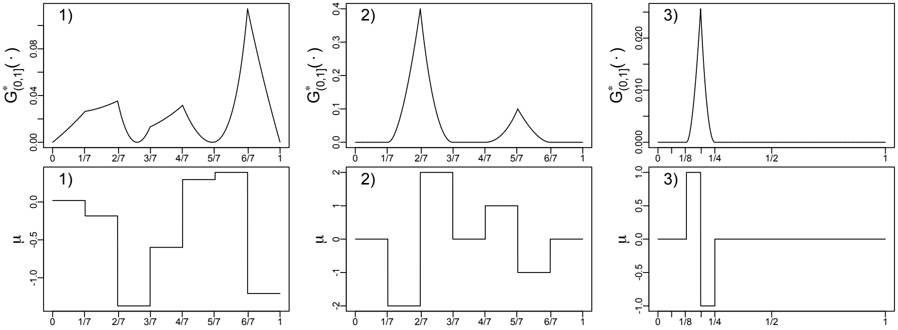
scenario		SeedBS	OSeedBS
"easy"	number of intervals	$O(\log T)$	$O(\log T)$
	number of evaluations	$O(T \log T)$	$O(\log T \cdot \log T)$
"difficult"	number of intervals	$O(T)$	$O(T)$
	number of evaluations	$O(T \log T)$	$O(T)$

Hence, OSeedBS has a **sublinear number of evaluations** in "intermediate" cases when no need to generate all $O(T)$ intervals.

Optimistic Binary Segmentation



Special cases for multiple change points



Some ideas on how to tackle these special cases (maybe for discussion?)

Computational costs for OSeedBS

So far, we considered the number of evaluations. Why?

- recall first example from introduction with high-dimensional Gaussian graphical models

Computational costs for OSeedBS

So far, we considered the number of evaluations. Why?

- recall first example from introduction with high-dimensional Gaussian graphical models
- no cheap updates of fits for neighbouring split points in complex models (e.g. lasso, graphical lasso, random forest, time series fits, etc.)

Computational costs for OSeedBS

So far, we considered the number of evaluations. Why?

- recall first example from introduction with high-dimensional Gaussian graphical models
- no cheap updates of fits for neighbouring split points in complex models (e.g. lasso, graphical lasso, random forest, time series fits, etc.)
- hence, in more complex scenarios with such fits, driving cost is the number of evaluations

Computational costs for OSeedBS

How about the univariate Gaussian case?

If **cumulative sums** have been **pre-computed**:

- then cost of an evaluation $O(1)$
- the (possibly **sublinear**) number of evaluations equals the actual computational costs

Computational costs for OSeedBS

How about the univariate Gaussian case?

If **cumulative sums** have been **pre-computed**:

- then cost of an evaluation $O(1)$
- the (possibly **sublinear**) number of evaluations equals the actual computational costs

If **cumulative sums** are **not available**:

- calculating cumulative sums is $O(T)$

Computational costs for OSeedBS

How about the univariate Gaussian case?

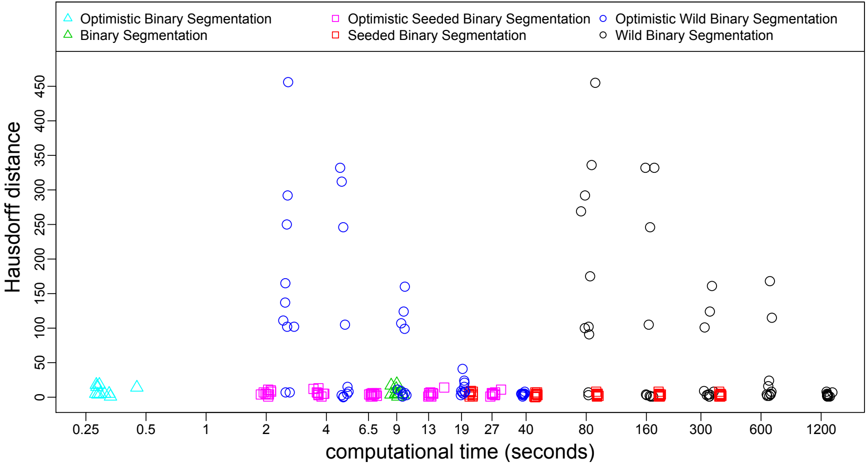
If **cumulative sums** have been **pre-computed**:

- then cost of an evaluation $O(1)$
- the (possibly **sublinear**) number of evaluations equals the actual computational costs

If **cumulative sums** are **not available**:

- calculating cumulative sums is $O(T)$
- **worst case cost** $O(T)$ with NOT selection independently of number of change points (and statistically optimal)

Multiple change points? - An example



Multiple change points?

Overall:

- Multiple change points more challenging
- Many combinations possible with existing techniques with differing computational or theoretical advantages

Main messages

Optimistic Search strategies:

- very strong (asymptotic) guarantees in single change point cases (in univariate Gaussian change in mean model)

Main messages

Optimistic Search strategies:

- very strong (asymptotic) guarantees in single change point cases (in univariate Gaussian change in mean model)
- extendable to multiple change point cases

Main messages

Optimistic Search strategies:

- very strong (asymptotic) guarantees in single change point cases (in univariate Gaussian change in mean model)
- extendable to multiple change point cases
- extendable to multivariate/high-dim setups change point cases

Main messages

Optimistic Search strategies:

- very strong (asymptotic) guarantees in single change point cases (in univariate Gaussian change in mean model)
- extendable to multiple change point cases
- extendable to multivariate/high-dim setups change point cases
- extendable to other distributions

Main messages

Optimistic Search strategies:

- very strong (asymptotic) guarantees in single change point cases (in univariate Gaussian change in mean model)
- extendable to multiple change point cases
- extendable to multivariate/high-dim setups change point cases
- extendable to other distributions
- very few evaluations of the gain function, hence, super fast

Main messages

Optimistic Search strategies:

- very strong (asymptotic) guarantees in single change point cases (in univariate Gaussian change in mean model)
- extendable to multiple change point cases
- extendable to multivariate/high-dim setups change point cases
- extendable to other distributions
- very few evaluations of the gain function, hence, super fast
- applicable and most useful in more complex/costly multivariate, high-dimensional, etc. scenarios

Main messages

Optimistic Search strategies:

- very strong (asymptotic) guarantees in single change point cases (in univariate Gaussian change in mean model)
- extendable to multiple change point cases
- extendable to multivariate/high-dim setups change point cases
- extendable to other distributions
- very few evaluations of the gain function, hence, super fast
- applicable and most useful in more complex/costly multivariate, high-dimensional, etc. scenarios
- simulation results: speedup (of orders of magnitude)

Outlook

- could be applied to speed up many other multiple change point techniques (IDetect, CBS, ...)
- could be used in sequential/online setups
- ...

Overall conclusions

optimistic search: `arxiv.org/abs/2010.10194`

Overall conclusions

optimistic search: `arxiv.org/abs/2010.10194`

seeded BS: `arxiv.org/abs/2002.06633`

Overall conclusions

optimistic search: `arxiv.org/abs/2010.10194`

seeded BS: `arxiv.org/abs/2002.06633`

Change-Point Detection for **Graphical Models** in the Presence of **Missing Values**: in JCGS

Overall conclusions

optimistic search: `arxiv.org/abs/2010.10194`

seeded BS: `arxiv.org/abs/2002.06633`

Change-Point Detection for **Graphical Models** in the Presence of **Missing Values**: in JCGS

Thank you for your attention!

Literature I

Avriel, M. and Wilde, D. J. (1966).
Optimality proof for the symmetric Fibonacci search technique.
Fibonacci Quarterly, 4:265–269.

Avriel, M. and Wilde, D. J. (1968).
Golden block search for the maximum of unimodal functions.
Management Science, 14(5):307–319.

Fryzlewicz, P. (2014).
Wild binary segmentation for multiple change-point detection.
The Annals of Statistics, 42(6):2243–2281.

Kiefer, J. (1953).
Sequential minimax search for a maximum.
Proceedings of the American Mathematical Society, 4:502–506.

Kovács, S., Li, H., Bühlmann, P., and Munk, A. (2020).
Seeded binary segmentation: a general methodology for fast and optimal change point detection.
arXiv:2002.06633.

Liu, H., Gao, C., and Samworth, R. J. (2019).
Minimax rates in sparse, high-dimensional changepoint detection.
arXiv:1907.10012.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004).
Circular binary segmentation for the analysis of arraybased DNA copy number data.
Biostatistics, 5(4):557–572.

Literature II

Vostrikova, L. Y. (1981).
Detecting 'disorder' in multidimensional random processes.
Soviet Mathematics Doklady, 24:270–274.