

Step size selection in Frank Wolf

Overview

Problem statement

Minimize $f(\mathbf{x})$

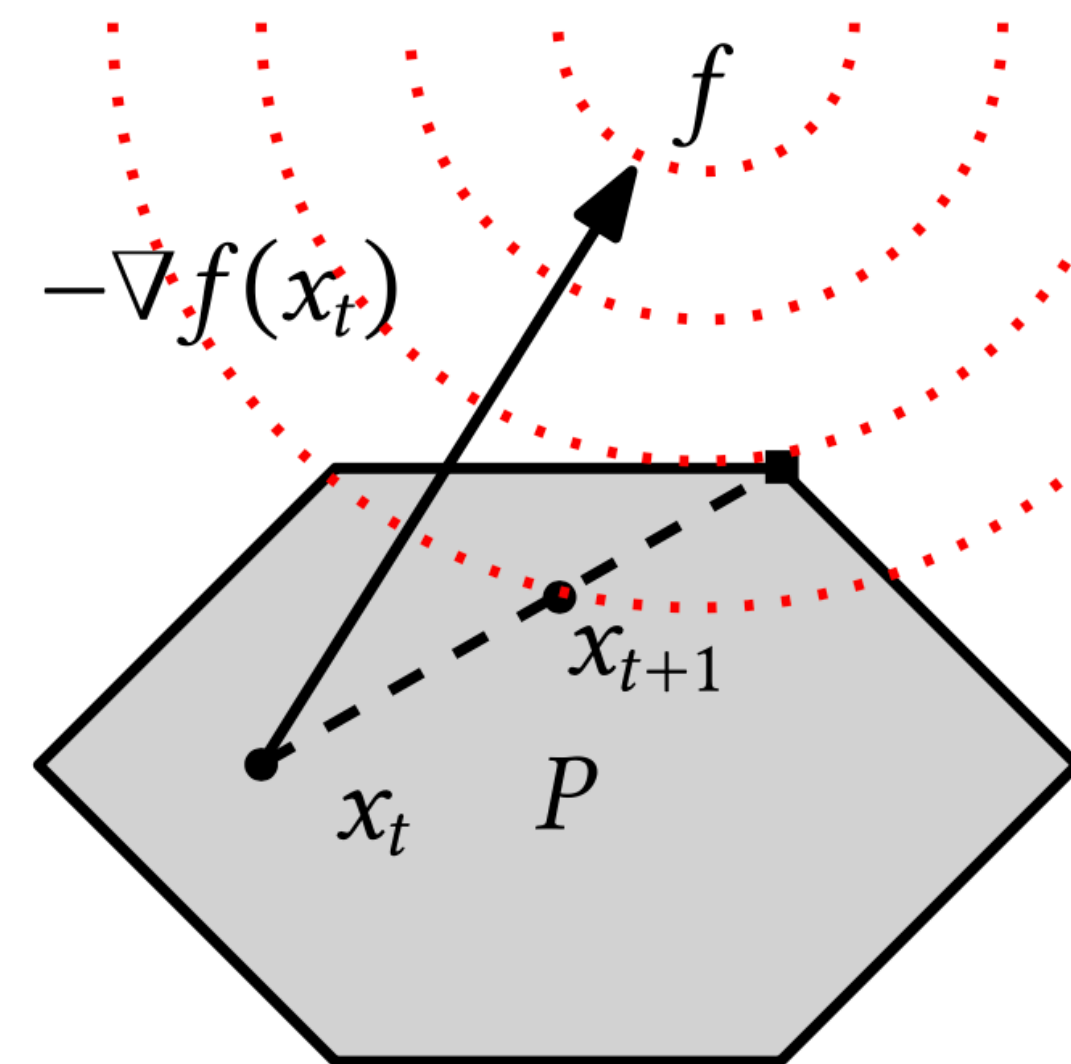
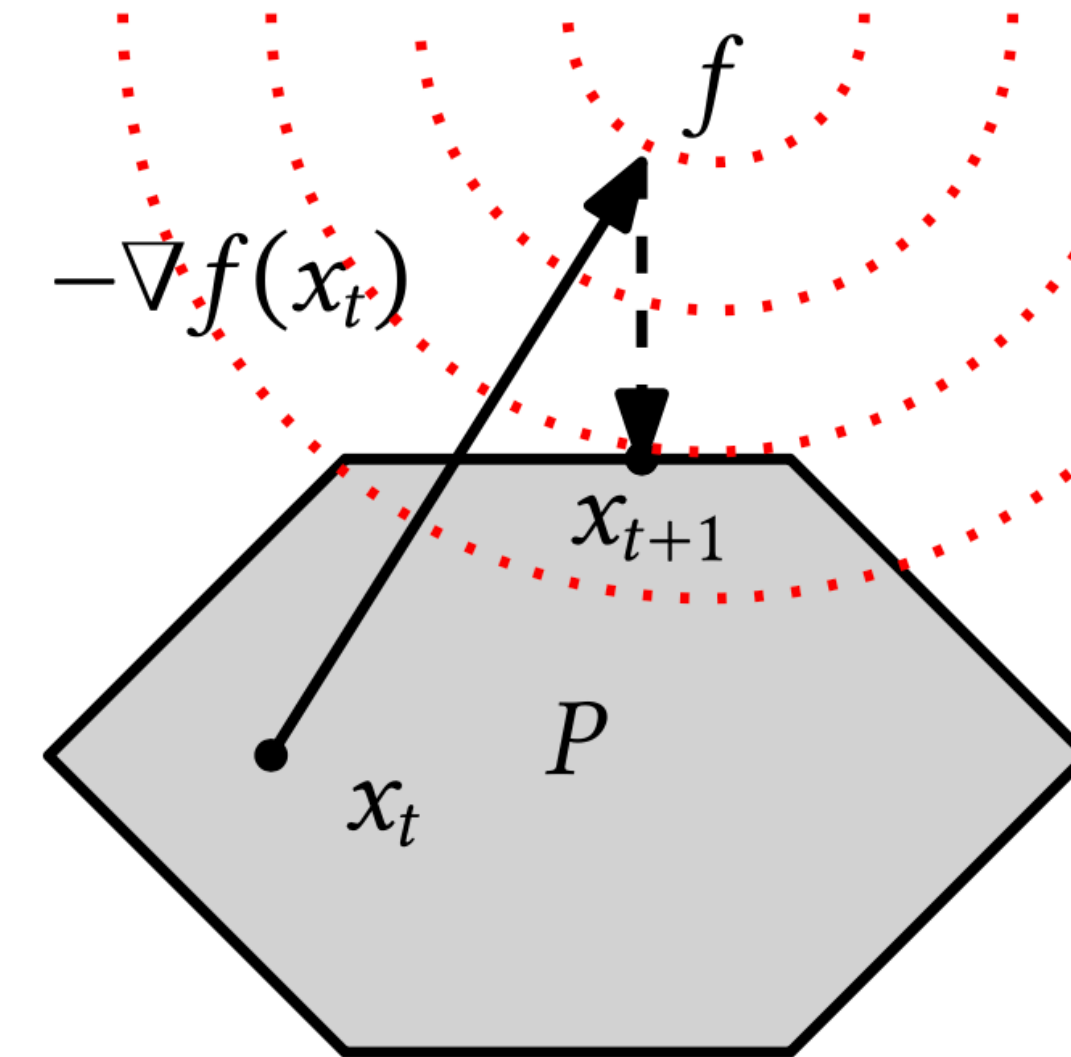
Subject to $\mathbf{x} \in \mathcal{D}$

- \mathcal{D} - **compact, convex** set in a vector space
- $f: \mathcal{D} \rightarrow \mathbf{R}$ is a **convex, L -smooth**, function

Linear Minimization Oracle

$$s = \arg \min_{s \in \mathcal{D}} s^T \nabla f(x_k)$$

- Finds a vector s in feasible set \mathcal{D} , which aligns most with $\nabla f(x_k)$.
- Vector s has the largest projection on $-\nabla f(x_k)$. Usually a vertex of the domain.



Algorithm

1. $s_k = \arg \min_{s \in \mathcal{D}} s^T \nabla f(x_k)$
 2. $x_{k+1} = (1 - \gamma_k)x_k + \gamma_k s_k$
- , where $\gamma_k \in [0,1]$ is a step-size.
- Both $x_k, s_k \in \mathcal{D}$. Convex combination of them is going to remain in the set. $x_{k+1} \in \mathcal{D}$

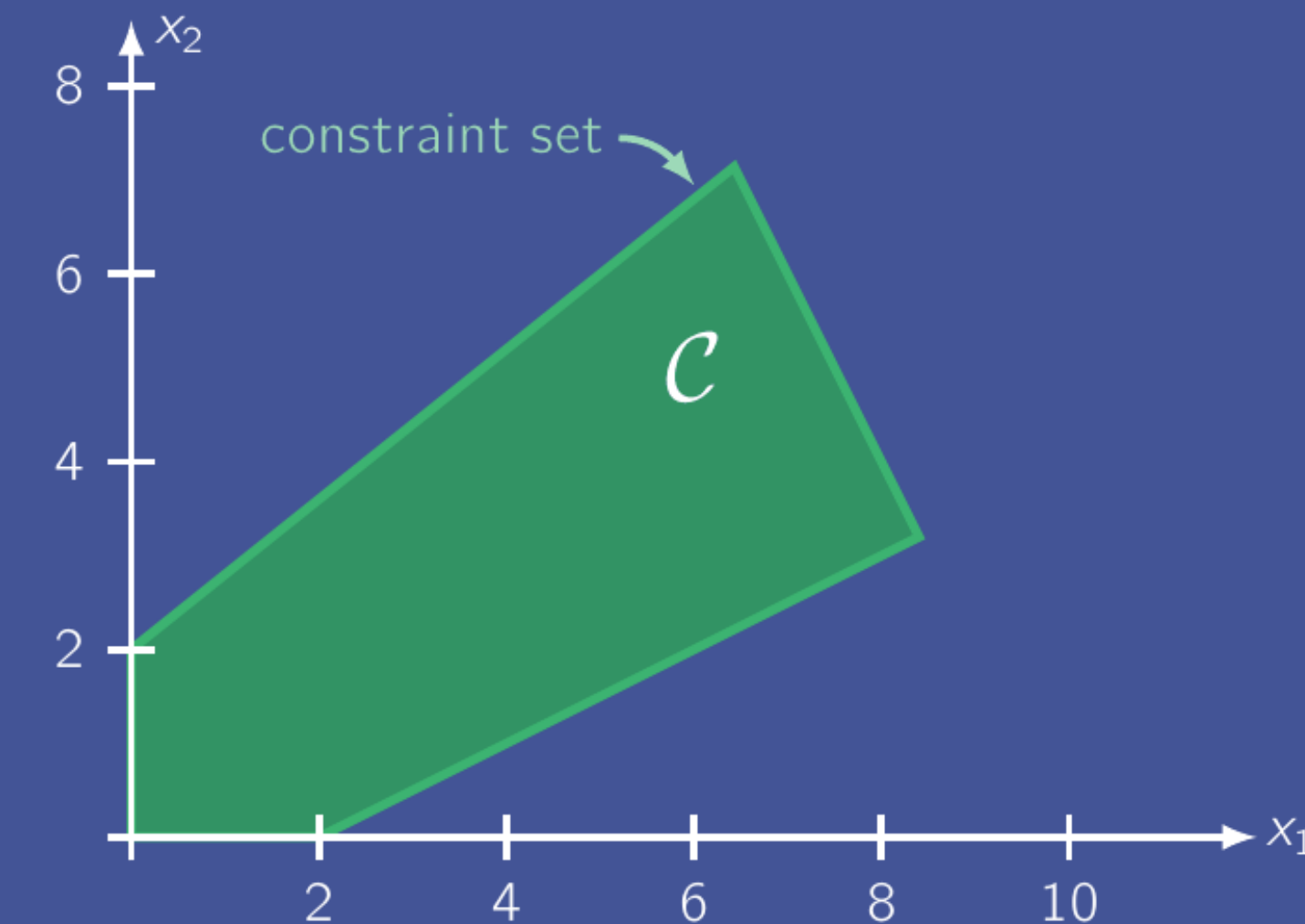
Conditional Gradient

To minimize f over \mathcal{C} , create $\{x^k\}$ via updates

$$s^{k+1} = \operatorname{argmin}_{s \in \mathcal{C}} s^T \nabla f(x^k)$$

$$x^{k+1} = x^k + \alpha_k (s^{k+1} - x^k)$$

Step size is often set to be $\alpha_k = 2/(k+2)$



Properties

- Convergence rate $O(1/k)$
- No need to do projection step (linear optimization vs. quadratic)
- Solve high dimensional problems
- Sparse solutions
- Designed for smooth, convex f
- Poor performance near optimum
- Complex, non-linear boundaries increase computation costs.

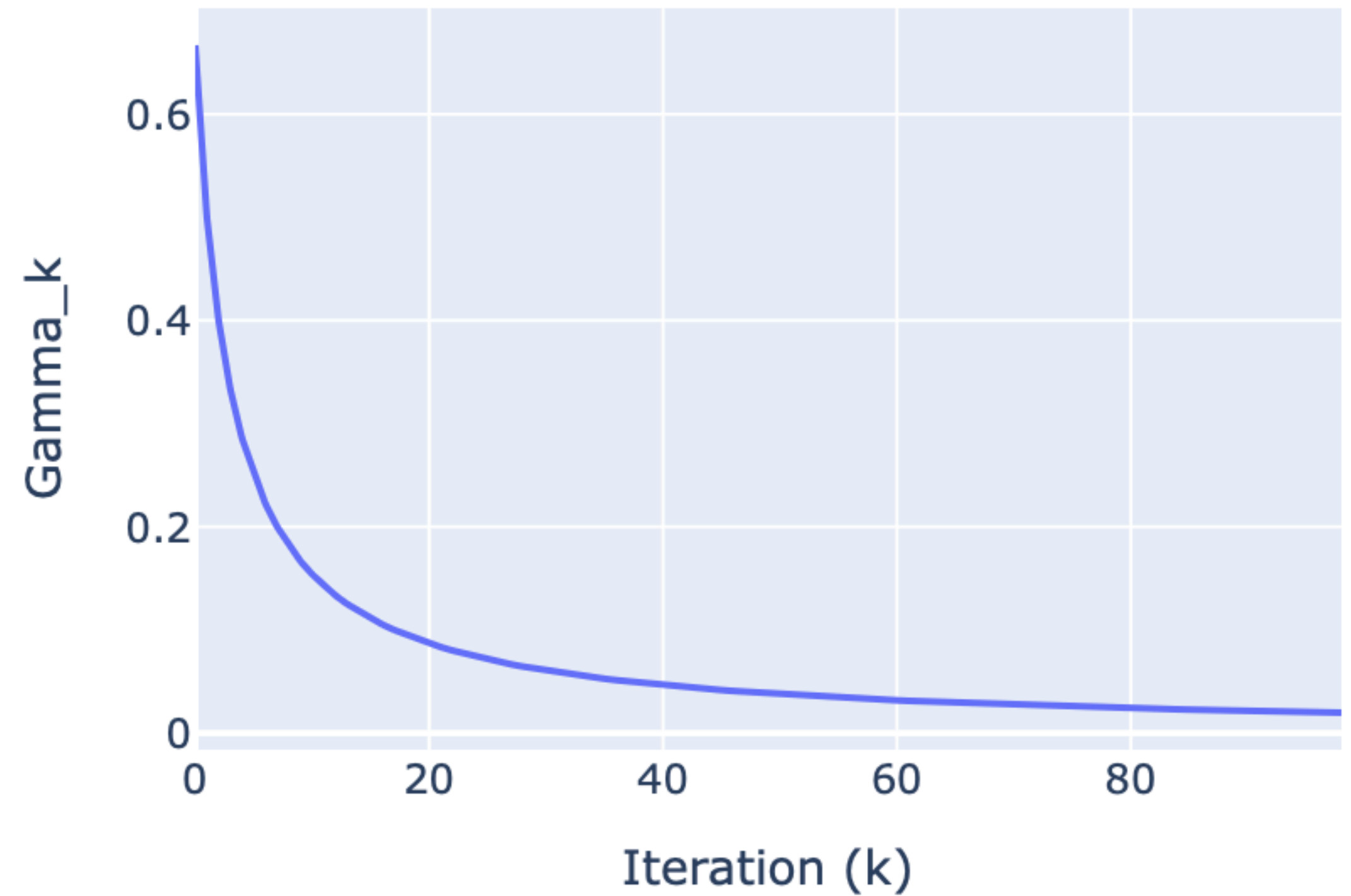
Step size

Line step size

$$\gamma_k = \frac{2}{k+2}$$

- Straightforward and cheap to compute.
- Slow convergence near optimum
- No function adaptation

Line step size

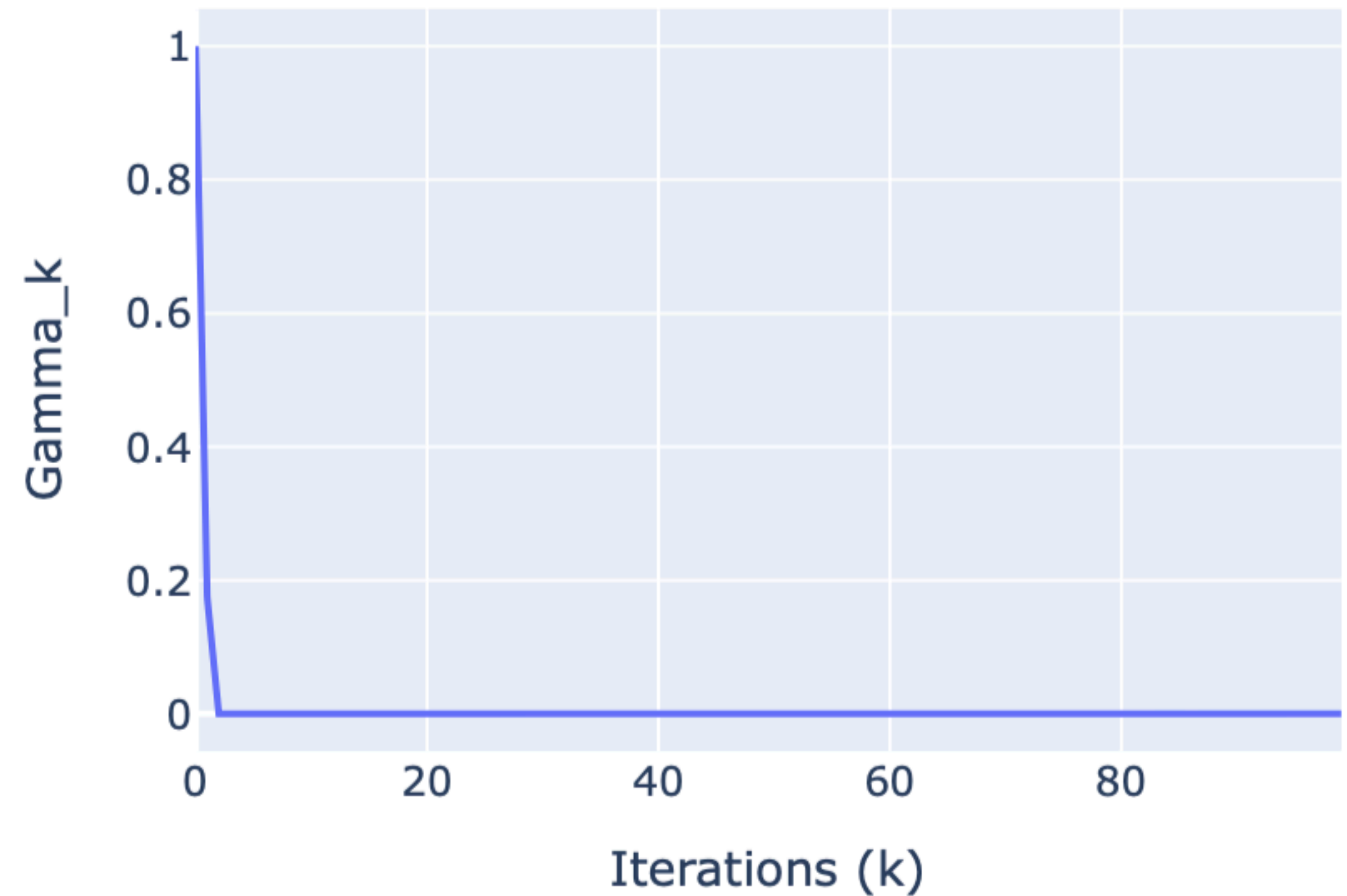


Exact line-search

$$\gamma_k = \arg \min f(x_k + \gamma_k(s_k - x_k))$$

- Ensures highest decrease per iteration
- Costly optimization problem

Exact line search

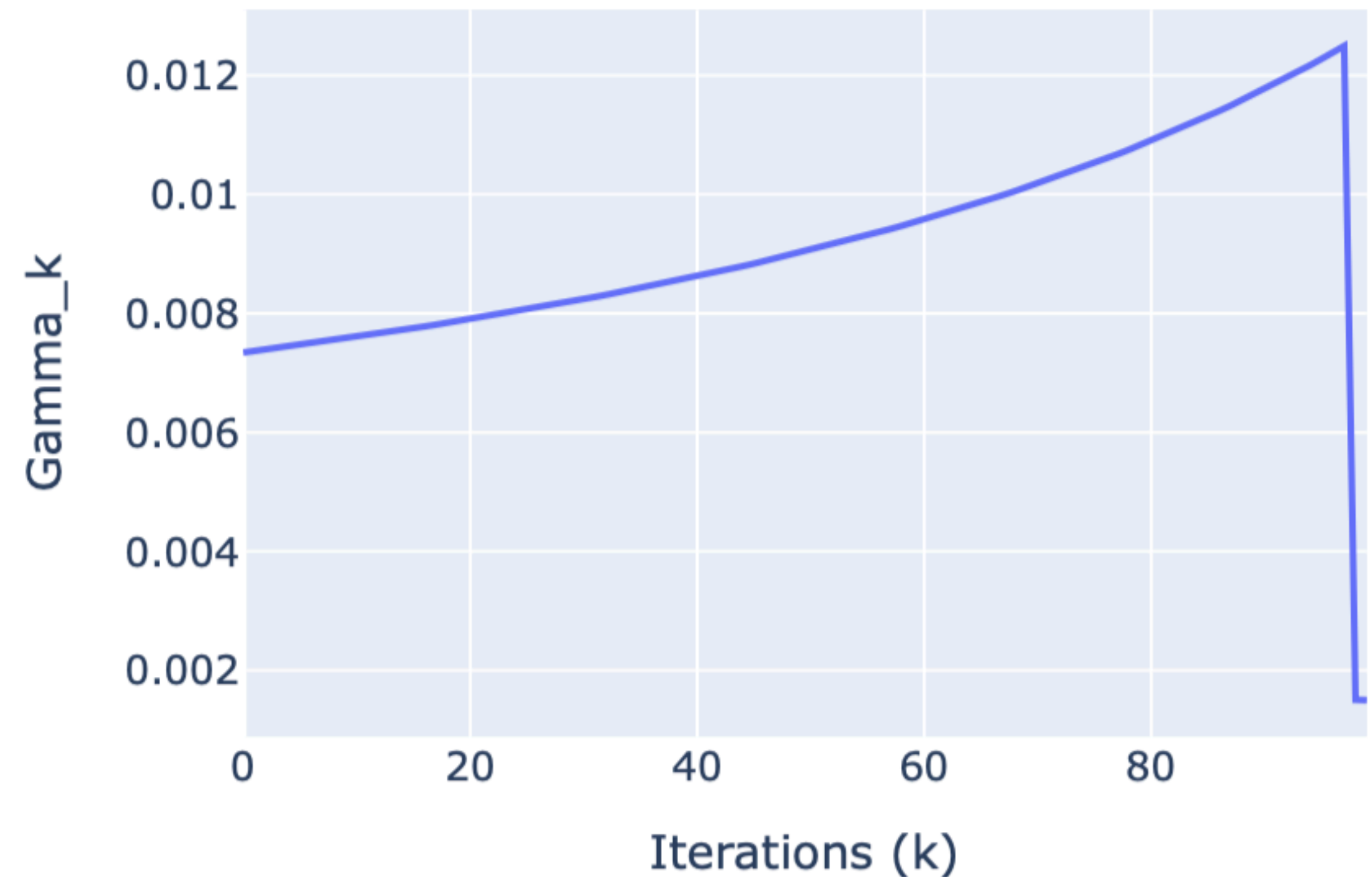


Demyanov-Rubinov

$$\gamma_k = \min\left\{\frac{-\nabla f(x)^T(s_k - x_k)}{L\|s_k - x_k\|^2}, 1\right\}$$

- Goes to zero as we approach the optimum
- Responsive for geometry of f
- Require access to L
- Unstable for small denominator (near optimum)

Demyanov Rubinov



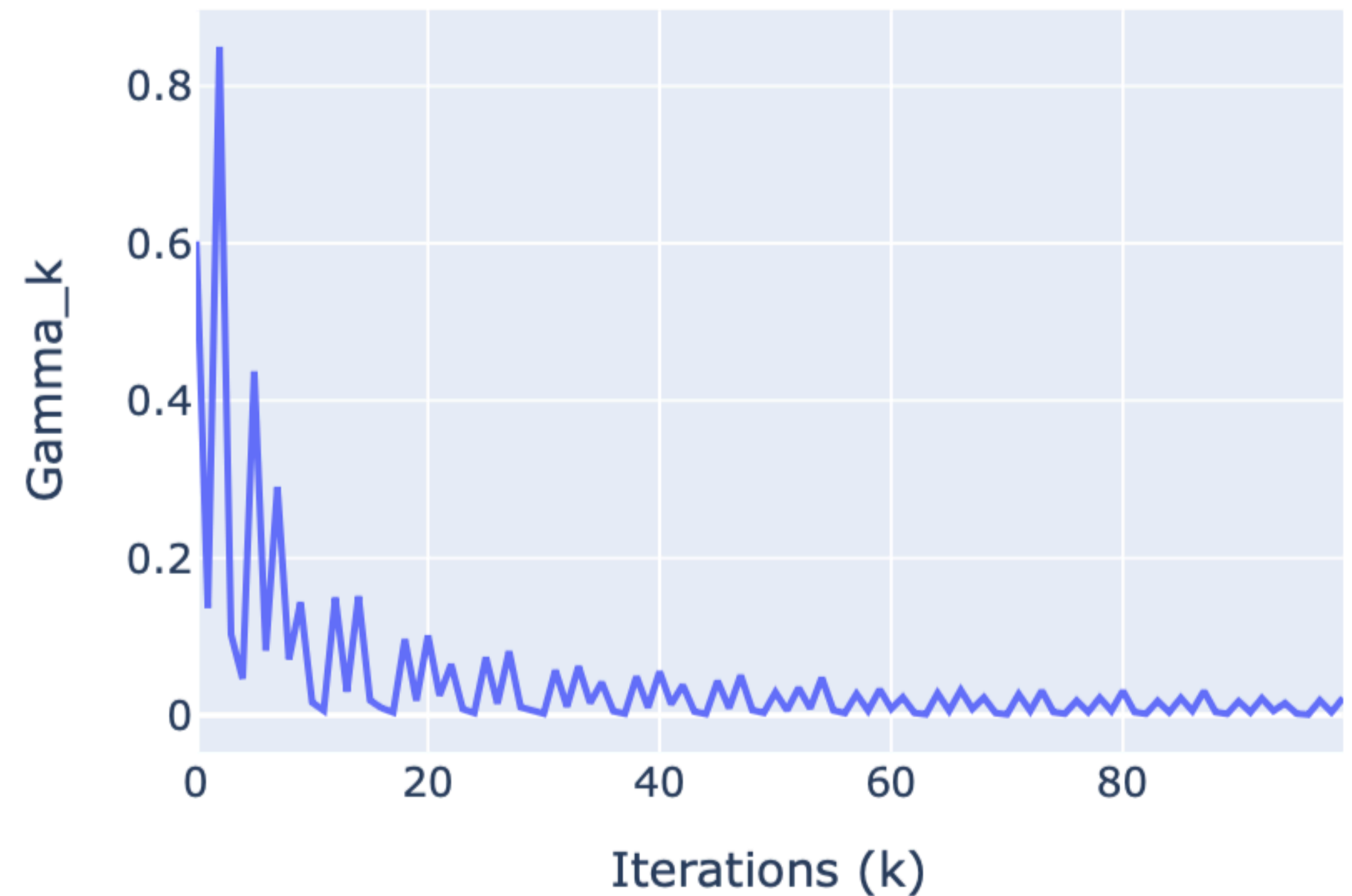
Backtracking line search

$$\gamma_k = \min\left\{\frac{-\nabla f(x)^T(s_k - x_k)}{M\|s_k - x_k\|^2}, 1\right\}$$

M_t is approximation of L

- Doesn't require L
- Adaptive and stable progress
- Require multiple evaluation of f

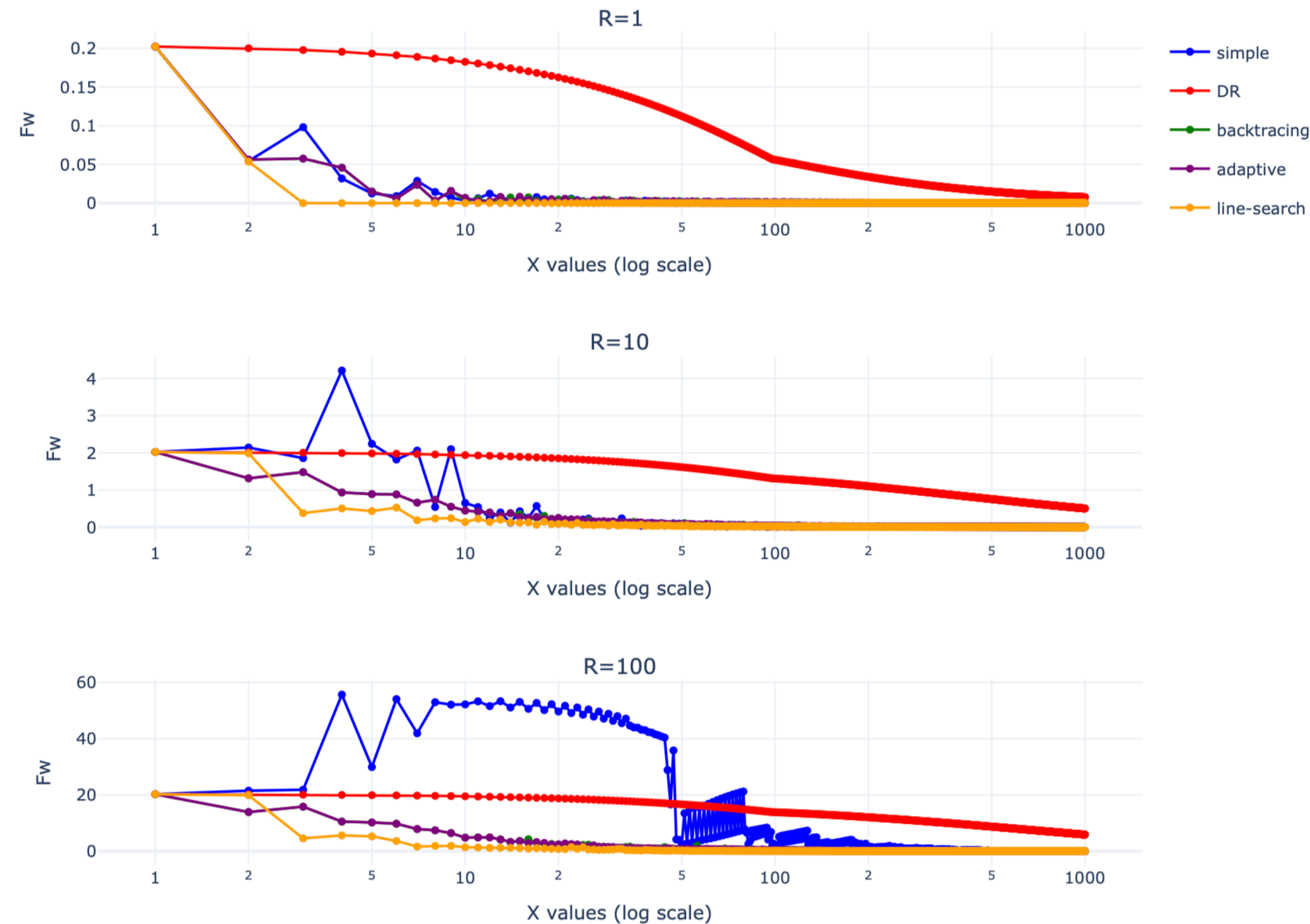
Backtracking line search



Experiments

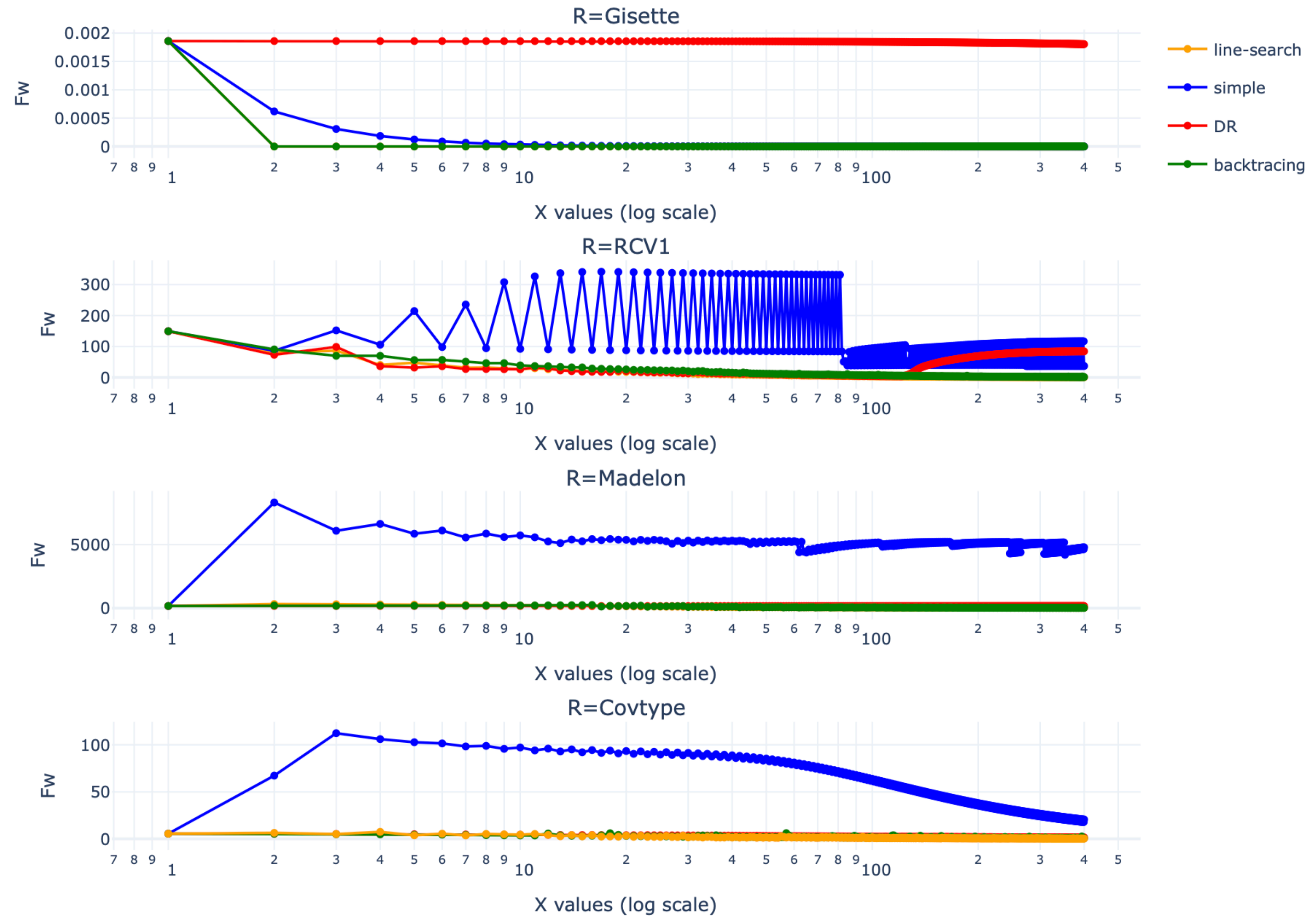
HW (Mushrooms)

Line Plots for Metric: fw (Grouped Legend)



Benchmark

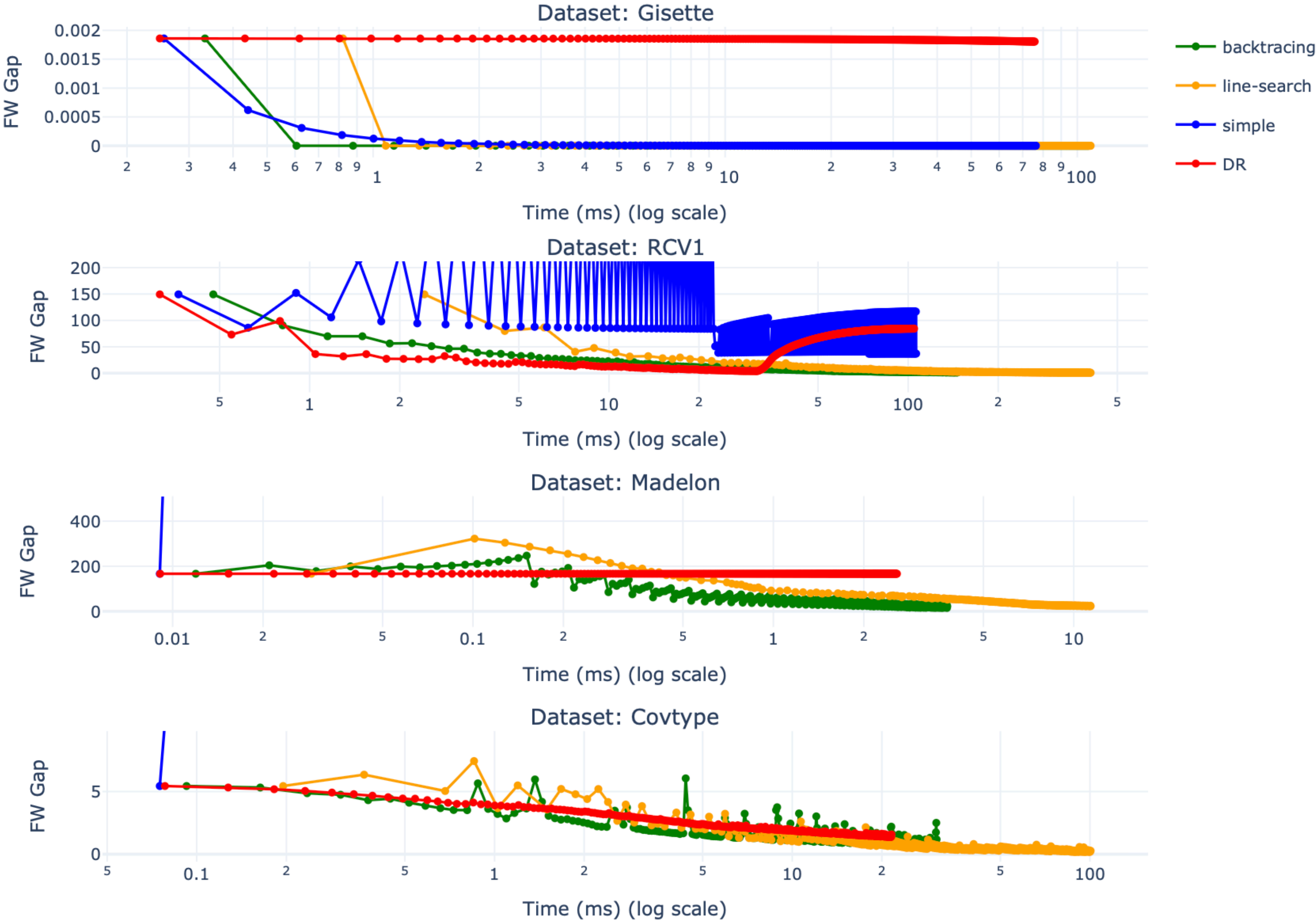
- Gisette: Binary classification on 5000 features $R = 6e^{-3}$
- RCV1: Binary classification on 47236 features $R = 2e^4$
- Madelon: Binary classification on 500 features $R = 20$
- Covtype: Binary classification on 54 features $R = 200$



Performance

	Simple	DR	Backtracking	Line-Search
Gisette	1:16 5.24 it/s	1:15 5.27 it/s	1:49 3.66 it/s	1:49 3.67 it/s
RCV1	1:43 3.87 it/s	1:41 3.94 it/s	2:25 2.75 it/s	6:41 1 it/s
Madelon	0:02 157.22 it/s	0:02 159 it/s	0:03 107 it/s	0:11 35.84 it/s
Covtype	0:21 18.84 it/s	0:21 18.67 it/s	0:30 13.02 it/s	1:39 4.02 it/s

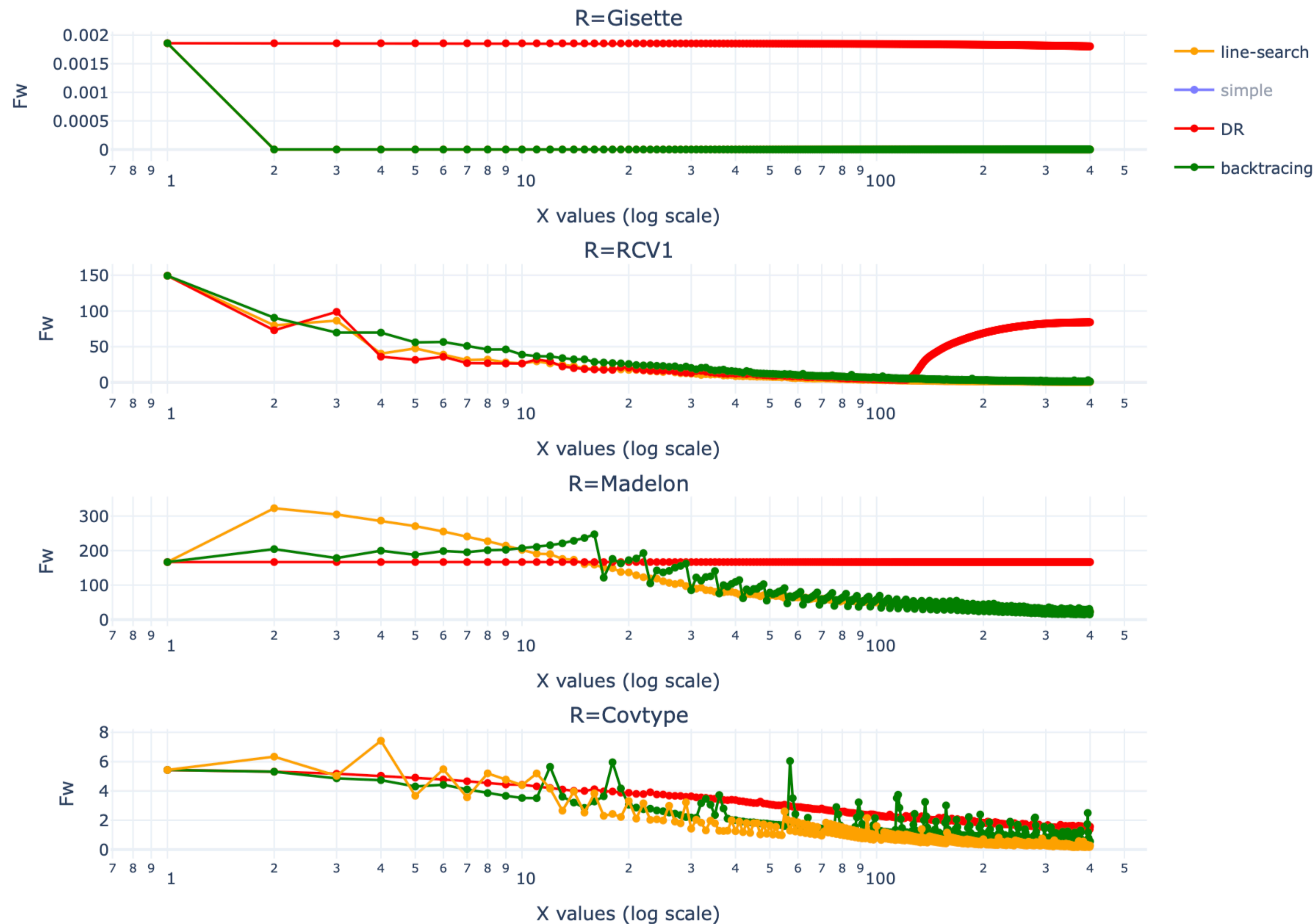
Benchmark. Performance experiment



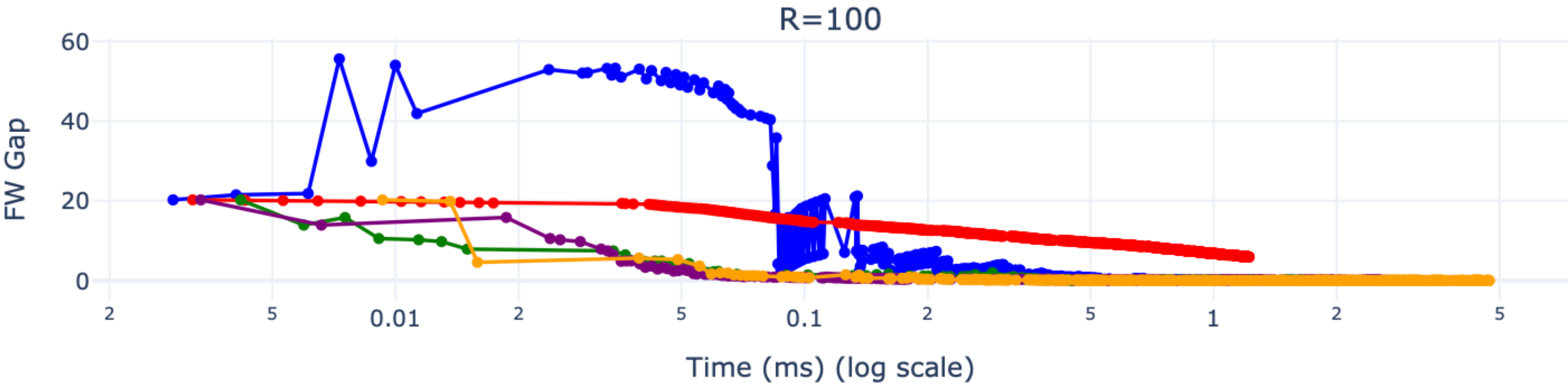
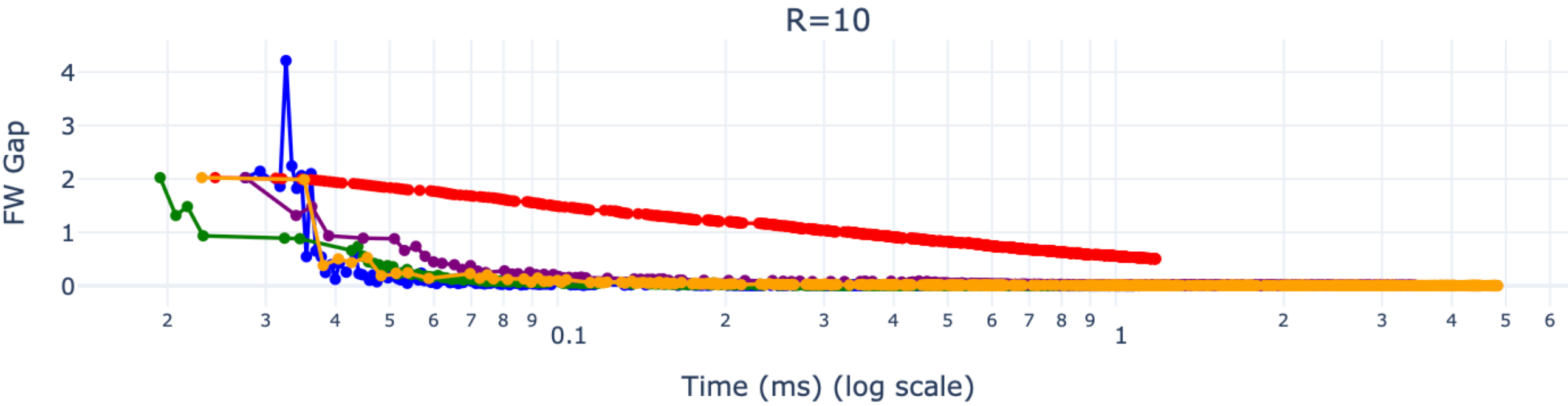
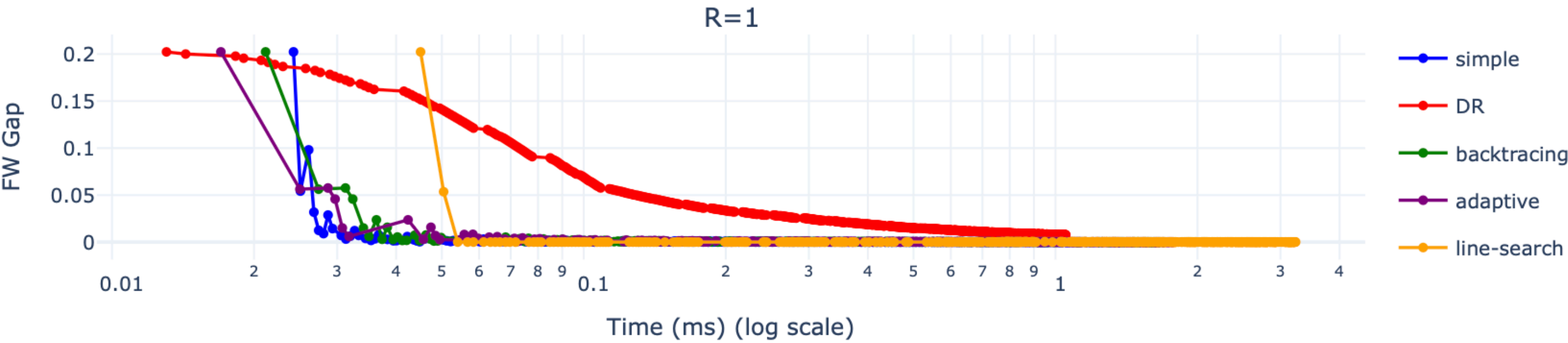
Appendix

The figure consists of four subplots, each representing a different dataset: R=Gisette, R=RCV1, R=Madelon, and R=Covtype. Each subplot shows the Fw (Y-axis) versus X values (log scale) (X-axis). The X-axis is labeled with values 1, 10, 100, 2, 3, 4, 5, indicating a logarithmic scale. The Y-axis is labeled Fw. The legend indicates four methods: line-search (orange), simple (blue), DR (red), and backtracing (green).

- R=Gisette:** The Y-axis ranges from 0 to 0.002. The DR method (red) maintains a high Fw value around 0.0018 across all X values. The backtracing method (green) starts at approximately 0.0018 for X=1 and drops sharply to 0 for X=2, remaining at 0 for all subsequent X values.
- R=RCV1:** The Y-axis ranges from 0 to 150. The DR method (red) starts at 150, drops to around 70 at X=2, fluctuates, and then rises sharply to approximately 80 for X values greater than 100. The backtracing method (green) starts at 150 and generally decreases, reaching near 0 for X values greater than 100.
- R=Madelon:** The Y-axis ranges from 0 to 300. The line-search method (orange) starts at 150, peaks at over 300 at X=2, and then gradually decreases. The DR method (red) remains relatively constant around 150. The backtracing method (green) starts at 150, peaks at around 200 at X=2, and then decreases, reaching near 0 for X values greater than 100.
- R=Covtype:** The Y-axis ranges from 0 to 8. The line-search method (orange) starts at approximately 5.5, peaks at around 6.5 at X=4, and then decreases. The DR method (red) starts at approximately 5.5 and decreases gradually. The backtracing method (green) starts at approximately 5.5 and decreases, reaching near 0 for X values greater than 100.



Mushrooms dataset. Performance experiment



Cancer

- Breast Cancer:

Cancer dataset

