Non-smooth dynamics perspectives for designing large scale optimization algorithms in stochastic settings

Rachel Kuske, Georgia Tech

In Collaboration with:

Emmanouil Daskalakis (UBC/VCC)
Felix Herrmann (Georgia Tech)
Andre Wibisono (Georgia Tech)

Non-smooth dynamics in optimization algorithms:

Example: look for solution of Ax=b,

Linearized Bregman (LB):

$$\begin{aligned} z_{k+1} &= z_k - t_k A^\top (A x_k - b) \\ x_{k+1} &= S_\lambda(z_{k+1}), \end{aligned} \qquad S_\lambda(z_k) = \max(|z_k| - \lambda, 0)\,\mathrm{sign}(z_k)$$

Yin, et al, 2008

S_λ is a shrinkage or thresholding operator - removes elements below threshold λ

Found in algorithms seeking sparse solutions, e.g. compressed sensing, underdetermined problems

Cai et al, 2009

t_k is a time step

Non-zero entries in solution $x^i = z^i \pm \lambda$

Context + Disclaimer:

Many different options for iterative methods in optimization:

First order methods (e.g. GD), Accelerated (higher order), stochastic, hybrids, non-smooth (projections, thresholds, etc)

Convex, non-convex:

Assumptions for any one method: sparsity, noise, matrix

Recently, more work from dynamics (and control) perspectives

Methods motivated by sparsity

Appended constraint for data match

Basis Pursuit	$\min_x \|x\|_1 \quad \text{subject to} \quad Ax = b\,.$
BPDN	$\min_x \|x\|_1 \quad \text{subject to} \quad Ax = b\,.$

$\lambda \to \infty$

Families of methods, e.g. close cousin of LB

ISTA - Iterative shrinkage (soft) thresholding

$$
\begin{aligned}
z_{k+1} &= \overset{x_k}{\cancel{z_k}} - t_k A^\top (Ax_k - b) \\
x_{k+1} &= S_\lambda(z_{k+1}), \qquad S_\lambda(z_k) = \max(|z_k| - \lambda, 0)\,\mathrm{sign}(z_k)
\end{aligned}
$$

ℓ_1- norm often used for sparse solutions, e.g. compressed sensing, underdetermined problems

$$\min_x \lambda\|x\|_1 + \frac{1}{2}\|x\|_2^2 \quad \text{subject to} \quad Ax = b$$

LB + dynamic time step

Lorentz et al 2014

Motivating applications

Large scale problems, with sparse representation:

Witte, et al, 2015

$$\min_x \|x\|_1 \quad \text{subject to} \quad \sum_{i=1}^{n_s} \|J_i[m_0, q_i]C^* x - b_i\|_2 \leq \sigma.$$

e.g. Recent results in compressed sensing in seismic imaging

Solution: curvelet transform coefficients x

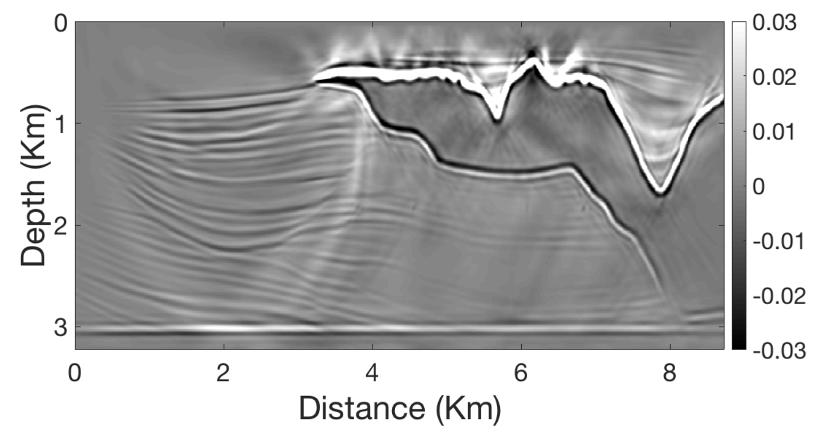
Large number of source experiments

Linearized - gives error/inconsistencies:

$Ax = b + \varepsilon \qquad \mathrm{Var}[\varepsilon] = \sigma^2$

Large ill-conditioned system

Background model parameters



(a) Iteration 21

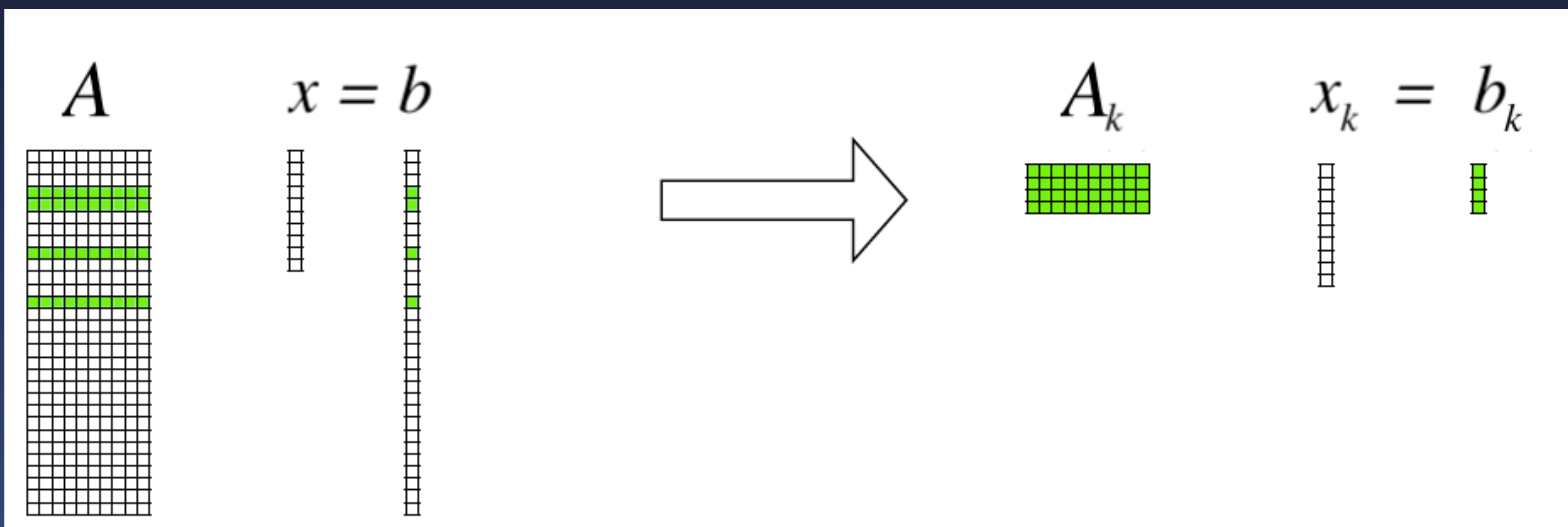
Motivating applications

Large scale problems, with sparse representation:

Focus on LB:

Straightforward implementation
Capitalize on sparsity - rapid progress to sparse solution
Combine with subsampling for large problems

Over-determined | Under-determined



$$
\begin{aligned}
z_{k+1} &= z_k - t_k \boxed{A_k^\top}(\boxed{A_k} x_k - \boxed{b_k}) \\
x_{k+1} &= S_\lambda(z_{k+1}),
\end{aligned}
$$

subsampling on each iteration

Under-determined systems (sub-samples):

Usual gradient descent: may not find sparse solution

Benefit from the presence of "noise", fluctuations, thresholds

Drawback: does not converge unless noise vanishes

$$\mathbb{E}\|\boldsymbol{x}_j - \boldsymbol{x}_\star\|_2^2 \leq \left[1 - \frac{\sigma_{\min}^2(\boldsymbol{A})}{n}\right]^j \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2 + \frac{n\|\boldsymbol{e}\|_\infty^2}{\sigma_{\min}^2(\boldsymbol{A})}$$

Simple Kaczmarz

A is nxd $e = \varepsilon$

$$\mathbb{E}\|\boldsymbol{x}_j - \boldsymbol{x}_\star\|_2^2 \leq \left[1 - \frac{\sigma_{\min}^2(\boldsymbol{A})}{\beta m}\right]^j \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2 + \frac{\beta}{\alpha} \cdot \frac{\|\boldsymbol{e}\|_2^2}{\sigma_{\min}^2(\boldsymbol{A})}$$

Randomized block Kaczmarz
- subsampling size m, with bounds on condition numbers

Needell and Tropp, 2012
(overdetermined, inconsistent, least squares minimization)

Stochastic gradient descent: escape local minima (ML)

Connection with non-smooth dynamics:

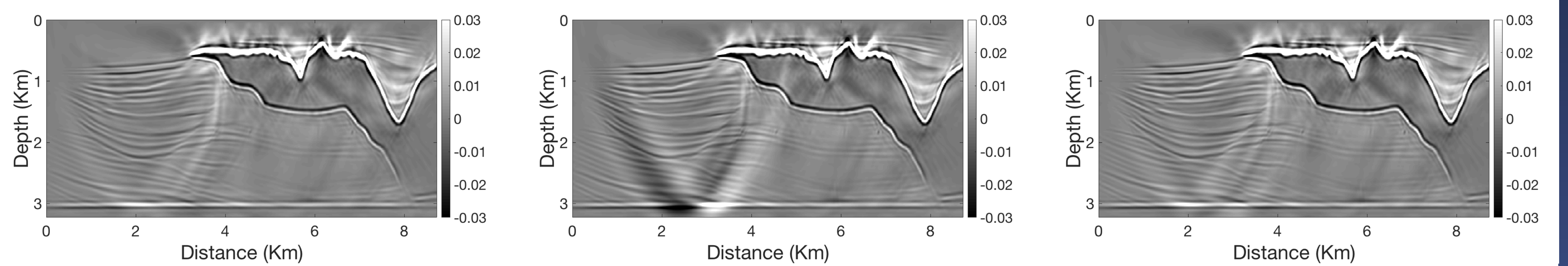
Sources of noise/error/variation/fluctuation:

Subsampling:

Inconsistencies: error due to linearization (data mis-match)

Threshold: search for sparse solution

Evidence of sustained chatter:



(a) Iteration 21 (b) Iteration 22 (c) Iteration 23

Dynamics for real systems:

Can not reach exact solution: noise + large/ expensive system with finite number of iterations

Computing stops during a transient in the algorithm

Want an algorithm that makes fast progress towards the solution

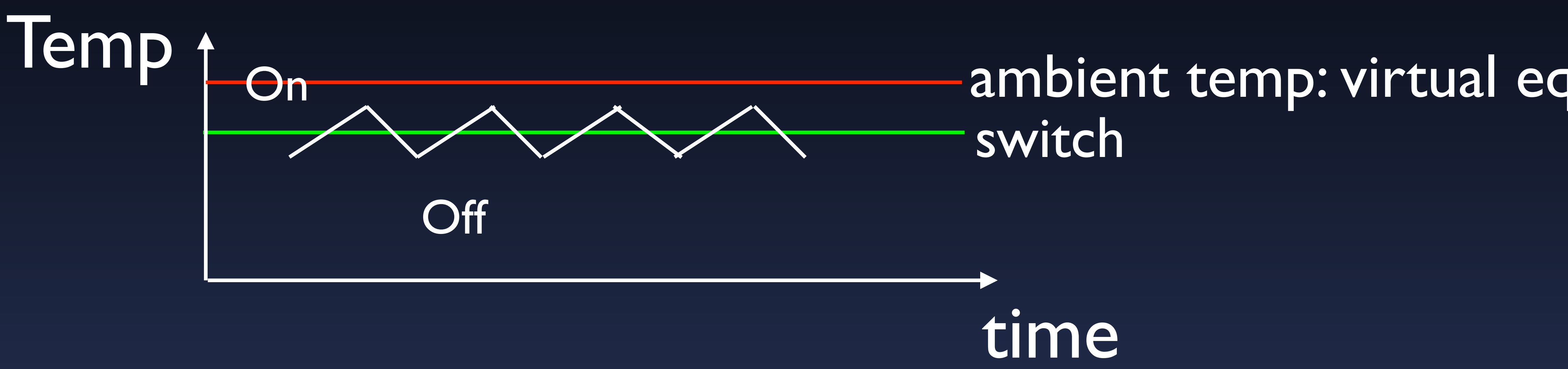
Not necessarily sparse : compressed sensing-like (violate certain assumptions for convergence)

In practice, the features that may aid rapid progress may also impede convergence in later interactions

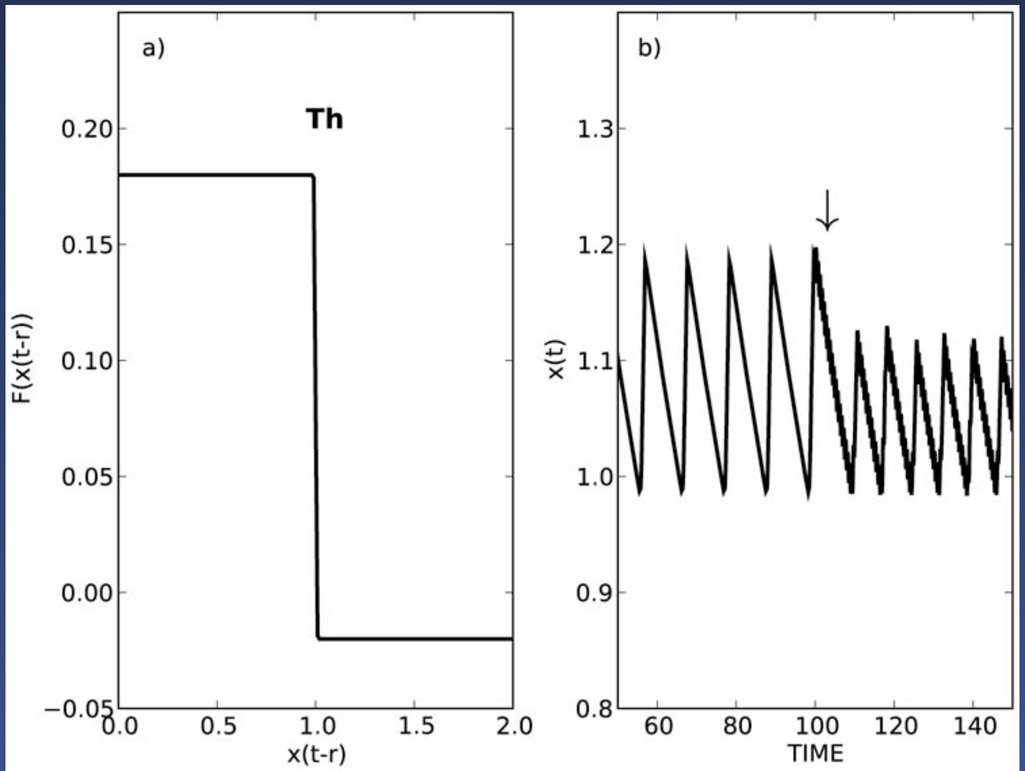
Chatter:

Search for a virtual equilibrium:

Ex: Thermostat set below ambient temperature



Some delay in feedback, otherwise have sliding on switch



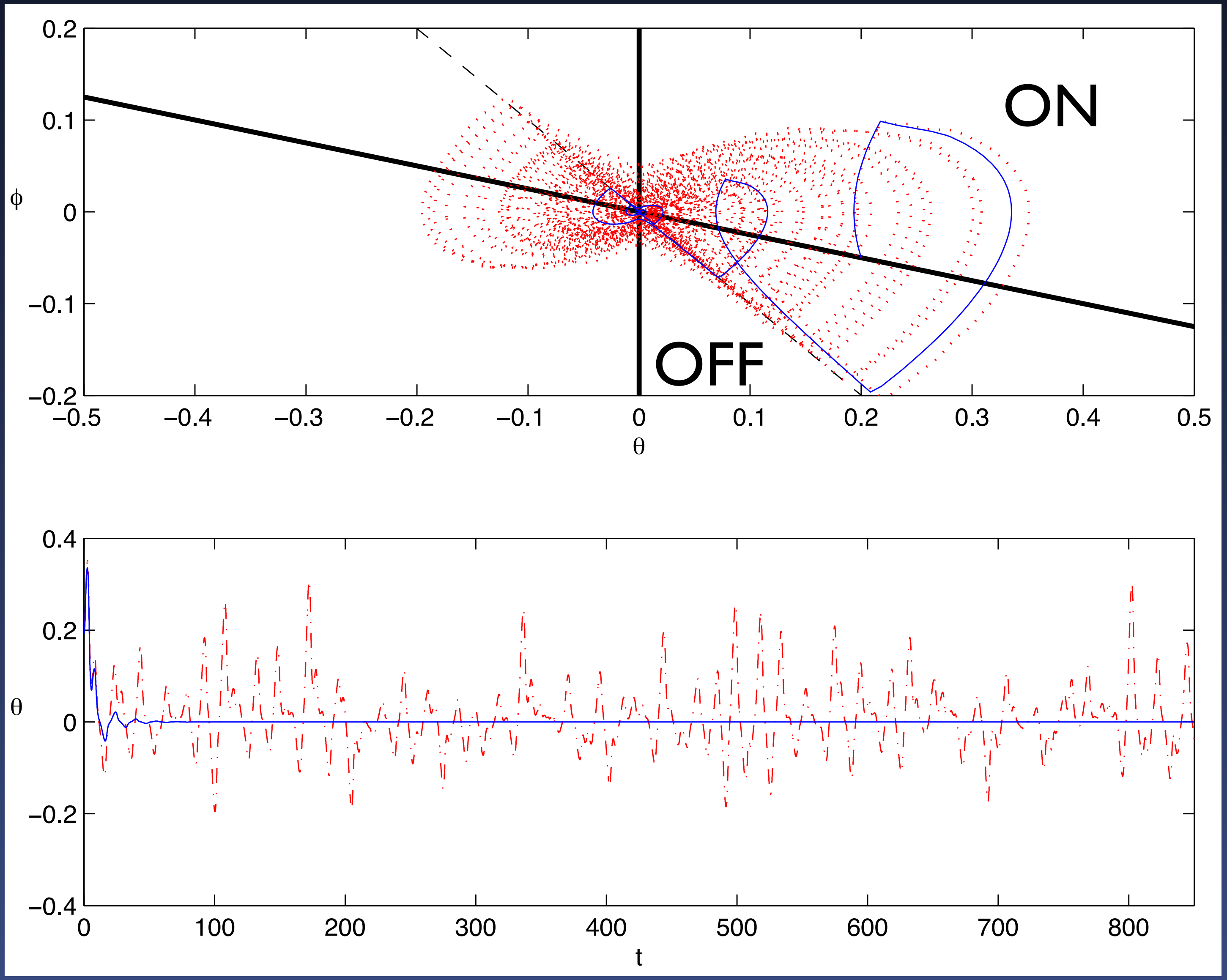
Toy model

$$\frac{dx}{dt} = F(x(t-\tau))x(t)$$

Milton, et al, 2009

Coherence resonance-type route to chatter

On-off control of balance: Inverted pendulum with delayed feedback control



Transient oscillations sustained as spiral via noise

Coherence resonance-type phenomenon - sustained transient oscillations with characteristic frequency

In optimization context: discrete time steps

In general, want to take as large a time step as possible for faster convergence

$$
\begin{aligned}
z_{k+1} &= z_k - t_k A^\top (A x_k - b) \\
x_{k+1} &= S_\lambda(z_{k+1}),
\end{aligned}
$$

$$S_\lambda(z_k) = \max(|z_k| - \lambda, 0)\,\mathrm{sign}(z_k)$$

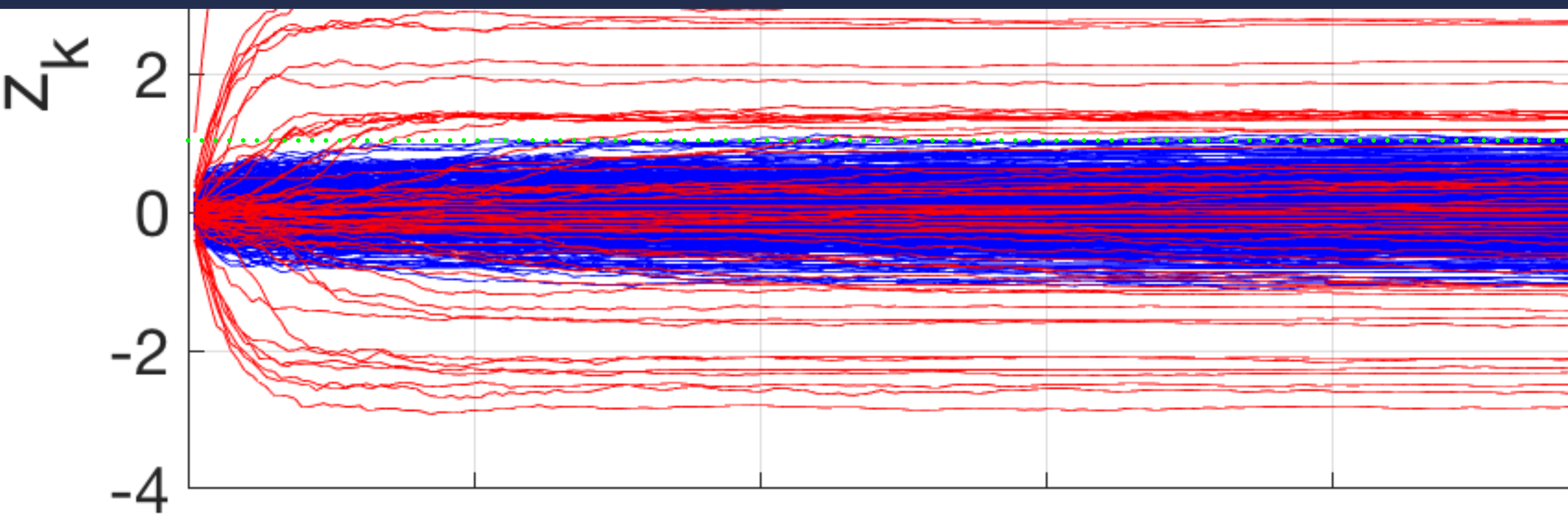
Constant:

$$t_k = \frac{1}{\|A_k\|_2^2}$$

Dynamic:

$$t_k = \frac{\|A_k x_k - b_k\|_2^2}{\|A_k^\top (A_k x_k - b_k)\|_2^2}$$

Entries enter and exit the support, crossing threshold at λ



Analogy to chatter

Taking finite steps at each stage:

In the inconsistent case: previous step approximation to over-determined case - no exact solution, only approximate solution -

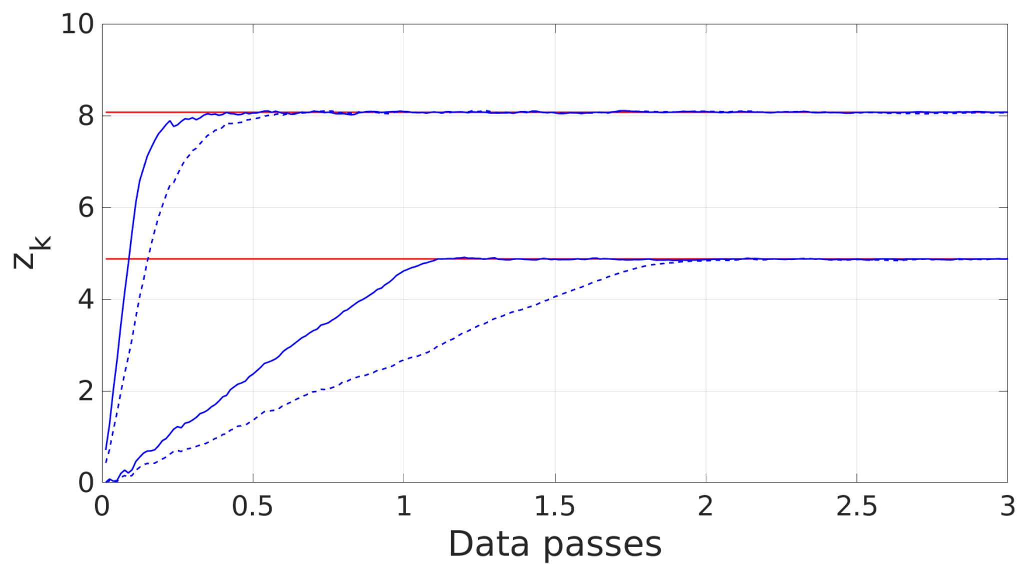
Sparse case: entry $x^i = 0$ in exact solution

Will exceed the threshold at some point, but will (likely) reduce below threshold on next iteration

Test (sparse) problem: track dynamics of entries

Threshold alone does not cause chatter:

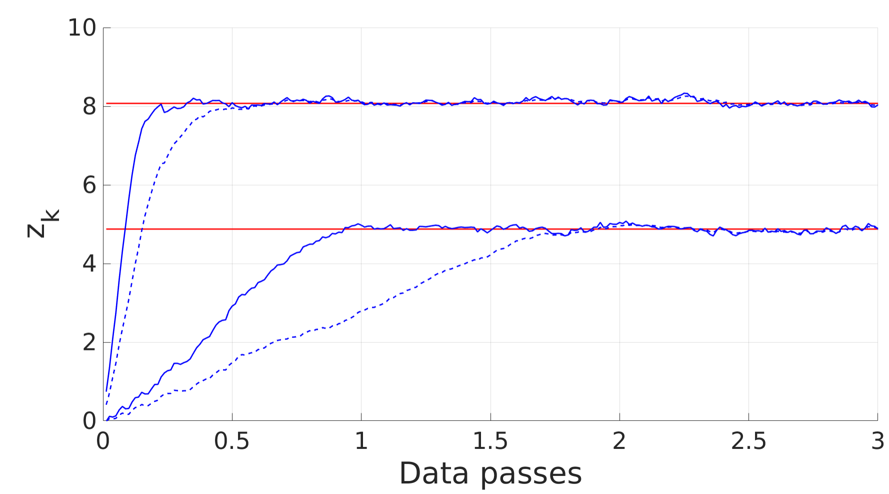
In the consistent case, there is an optimal solution to Ax=b ($\sigma = 0$)



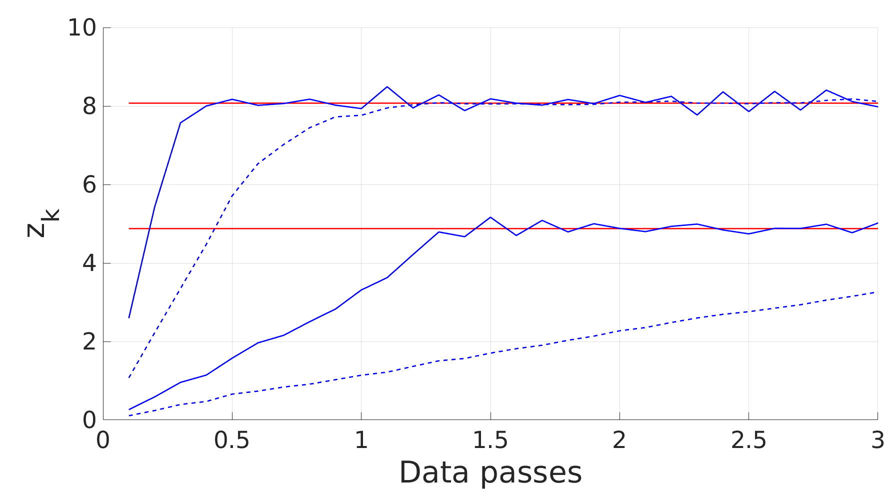
(a) $A_k \in \mathbb{R}^{250\times1000}$, $\sigma = 0$

Sparse example, threshold 4

Under-determined system



(b) $A_k \in \mathbb{R}^{250\times1000}$, $\sigma > 0$

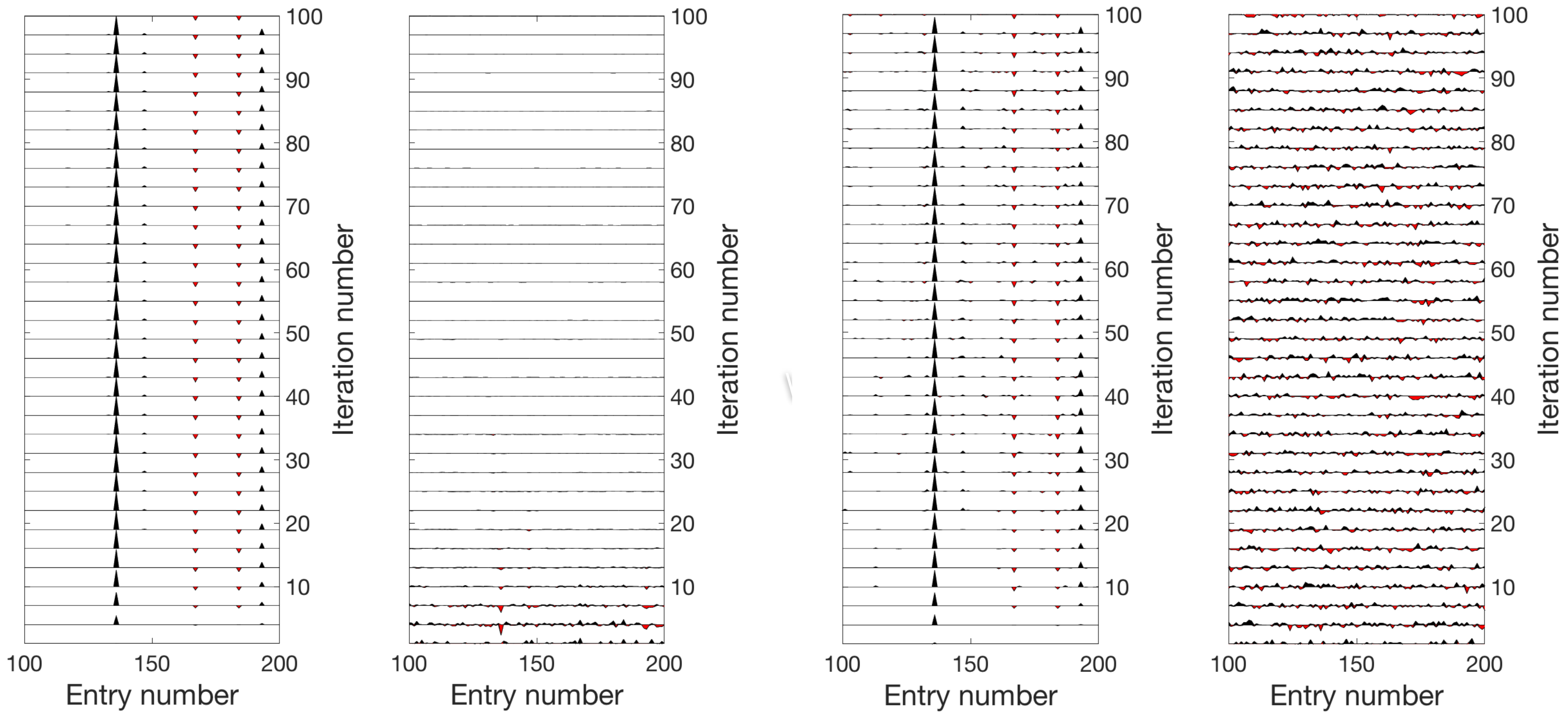


(c) $A_k \in \mathbb{R}^{2000\times1000}$, $\sigma > 0$

Over-determined (inconsistent) system $\sigma \neq 0$

"Wiggle" plots :
classically used for signal traces (in seismic)

Model vs. gradient



(a) Model iterates for $A_k \in \mathbb{R}^{250\times 1000}$, $\sigma = 0$

(b) Gradients for $A_k \in \mathbb{R}^{250\times 1000}$, $\sigma = 0$

(c) Model iterates for $A_k \in \mathbb{R}^{250\times 1000}$, $\sigma > 0$

(d) Gradients for $A_k \in \mathbb{R}^{250\times 1000}$, $\sigma > 0$

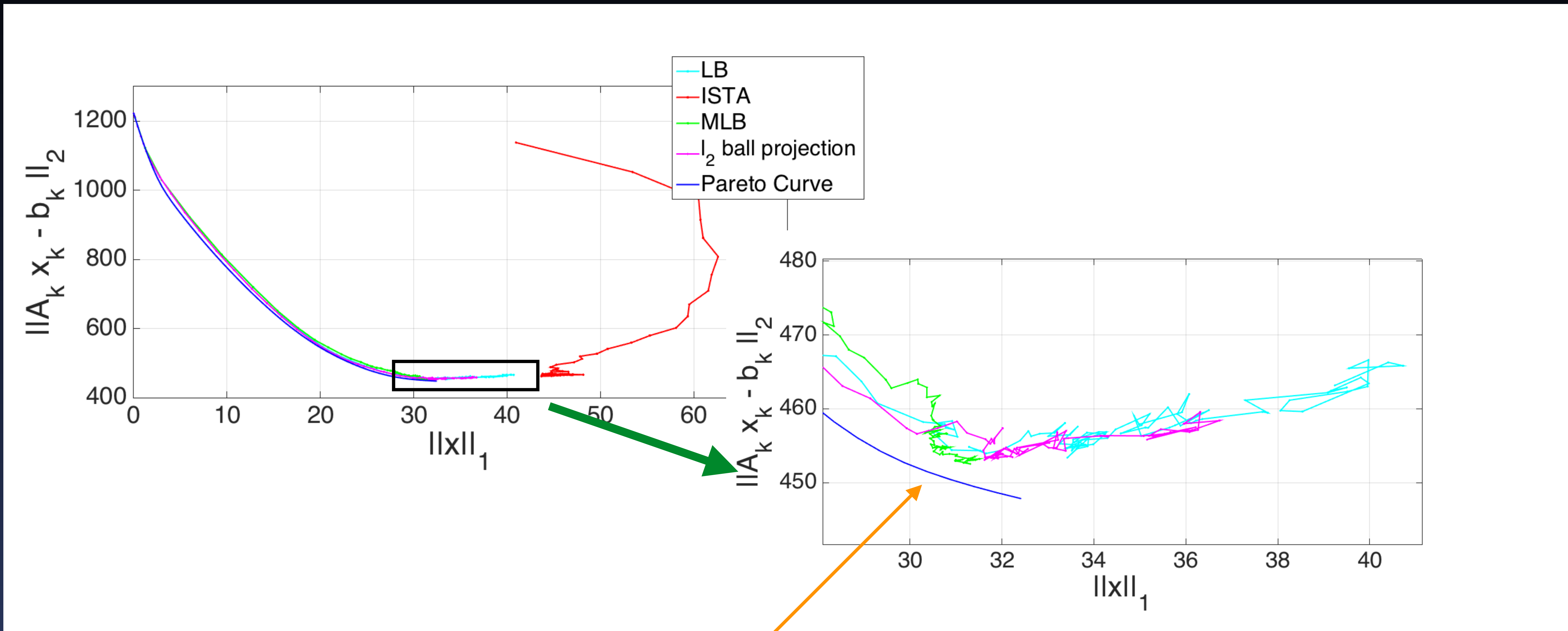
Consistent vs. Inconsistent

Approaches to address cycling: fluctuations about a solution

Projection at each step, based on noise level:
Advantage: Eliminates largest of fluctuations,
Disadvantage: Reduces the solution space - some solutions not allowed, have to approximate the noise level. (Lorenz, et al, 2014)

Reduce step-size: When, and how? Choose specific directions of search
Reduce overshoot (used e.g. in SGD)
Disadvantage: Could slow convergence, could be computationally expense to determine

Evolution in the error vs. sparsity trade-off plane



Compare to the Pareto curve: separates feasible and infeasible solutions

Hennenfent, et al, 2008

Different types of transient behavior - ideally tracking the Pareto curve (LB uses threshold only in gradient term - samples transients)

Modified LB (MLB) algorithm

Specific features of regular crossing of threshold: Frequent change of gradient

In contrast to changes in gradient due to subsampling or change of gradient due to noise

$$
\begin{aligned}
z_{k+1} &= z_k - \tau_k \odot A_k^\top (A_k x_k - b_k) \\
x_{k+1} &= S_\lambda(z_{k+1}),
\end{aligned}
$$

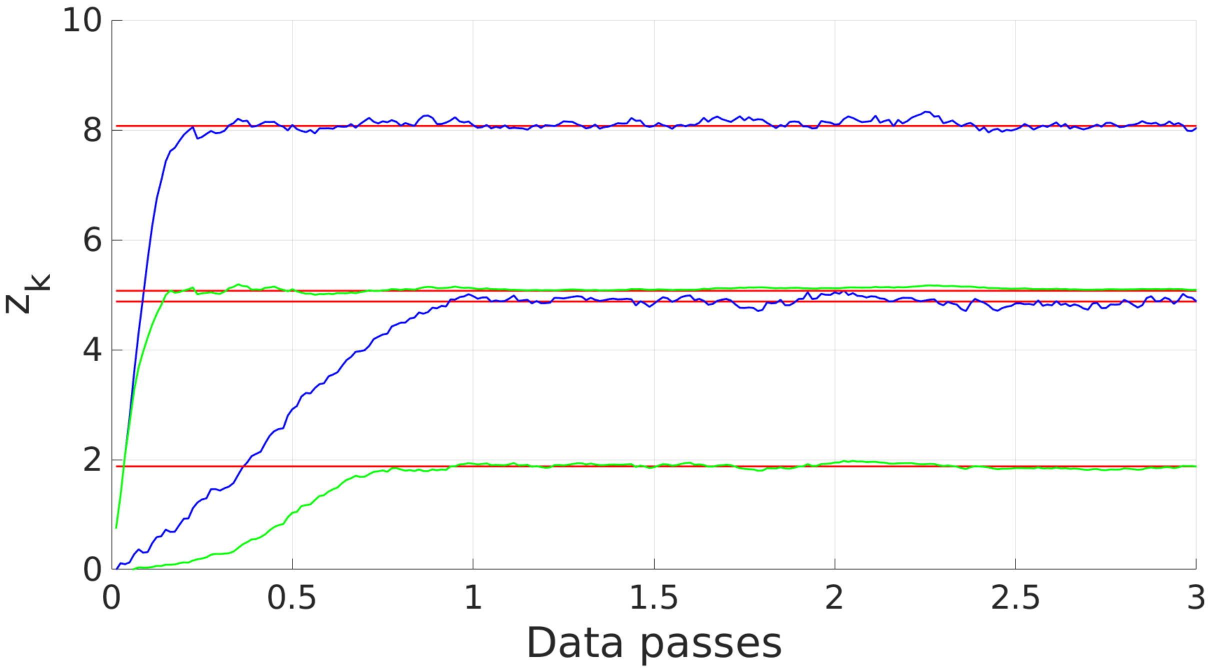
$$
\tau_k[i] = t_k \frac{\left| \sum_{j=1}^{k} \operatorname{sign}([A_j^\top (A_j x_j - b_j)]_i) \right|}{k}
$$

Factor in definition of the time step: element by element adjustment of time step

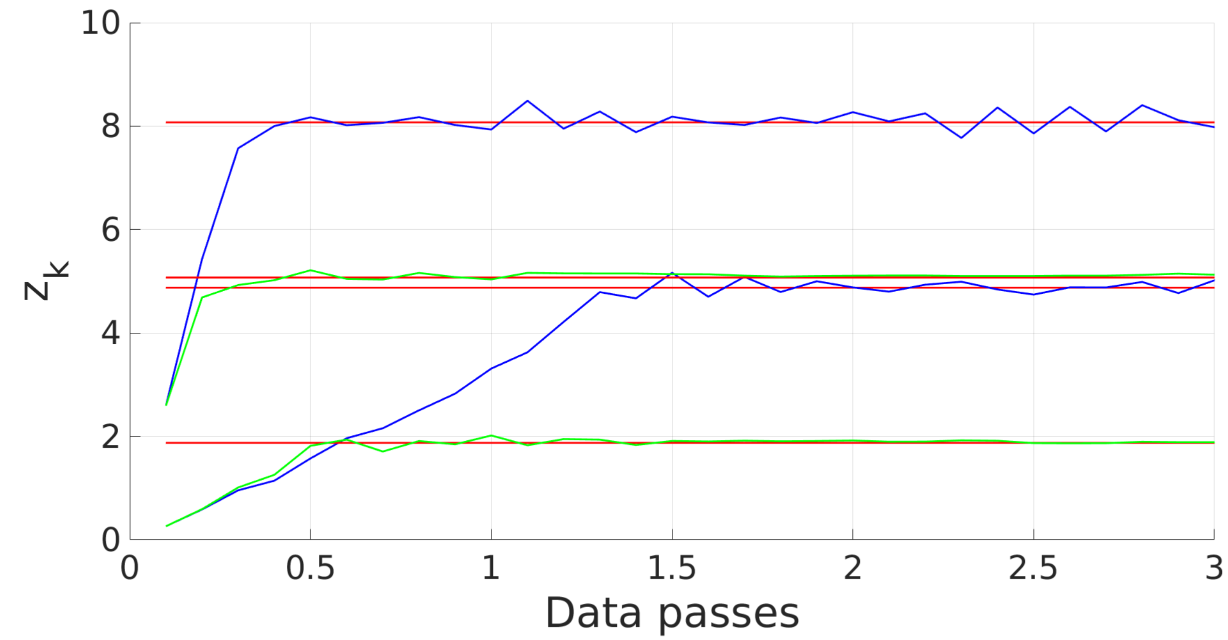
No chatter - no change in time step, progress towards the correct value continues

Chatter sets in - time step decreases

MLB vs LB

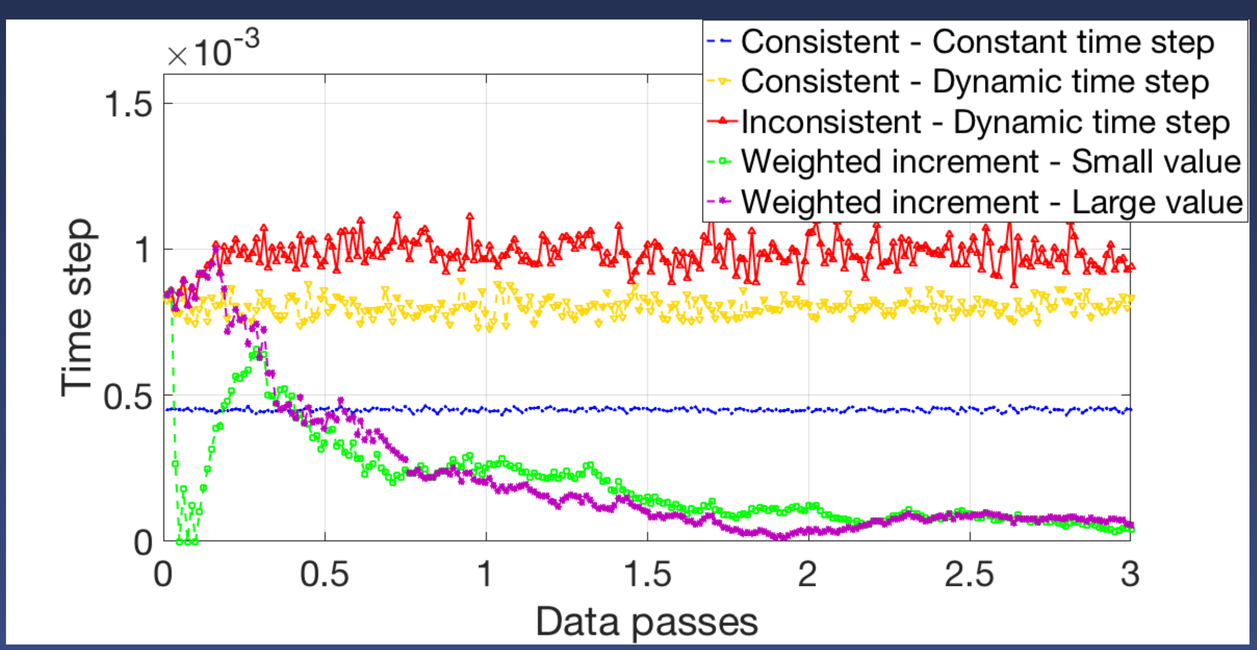


(a) Dynamic time-step, $A_k \in \mathbb{R}^{250\times 1000}$



(b) Dynamic time-step, $A_k \in \mathbb{R}^{2000\times 1000}$

MLB solution settles in at target value, as time step is reduced

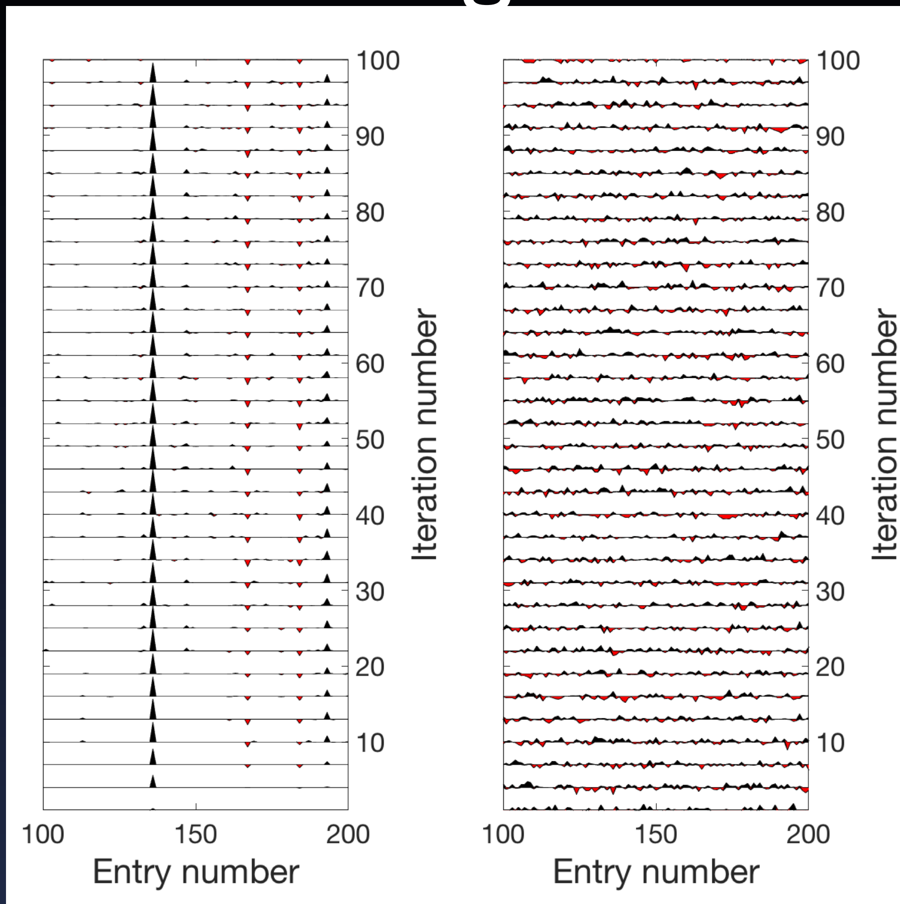


Note: different choice of λ for different schemes -

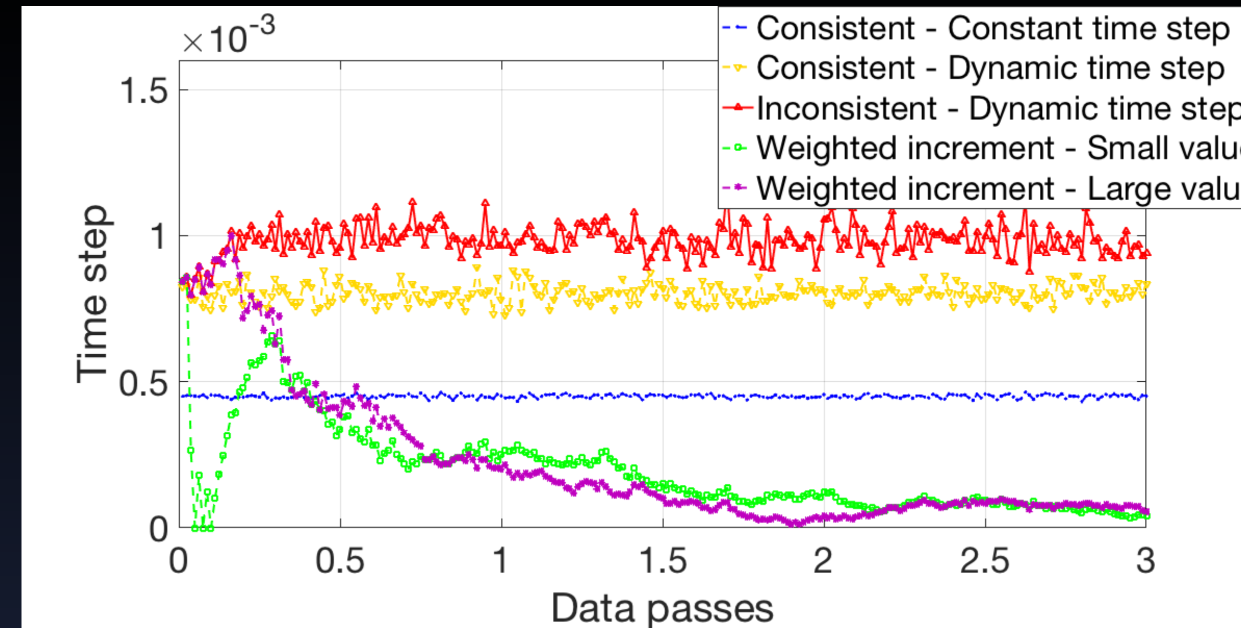
$$x^i = z^i \pm \lambda$$

Daskalakis,K, Herrmann 2017

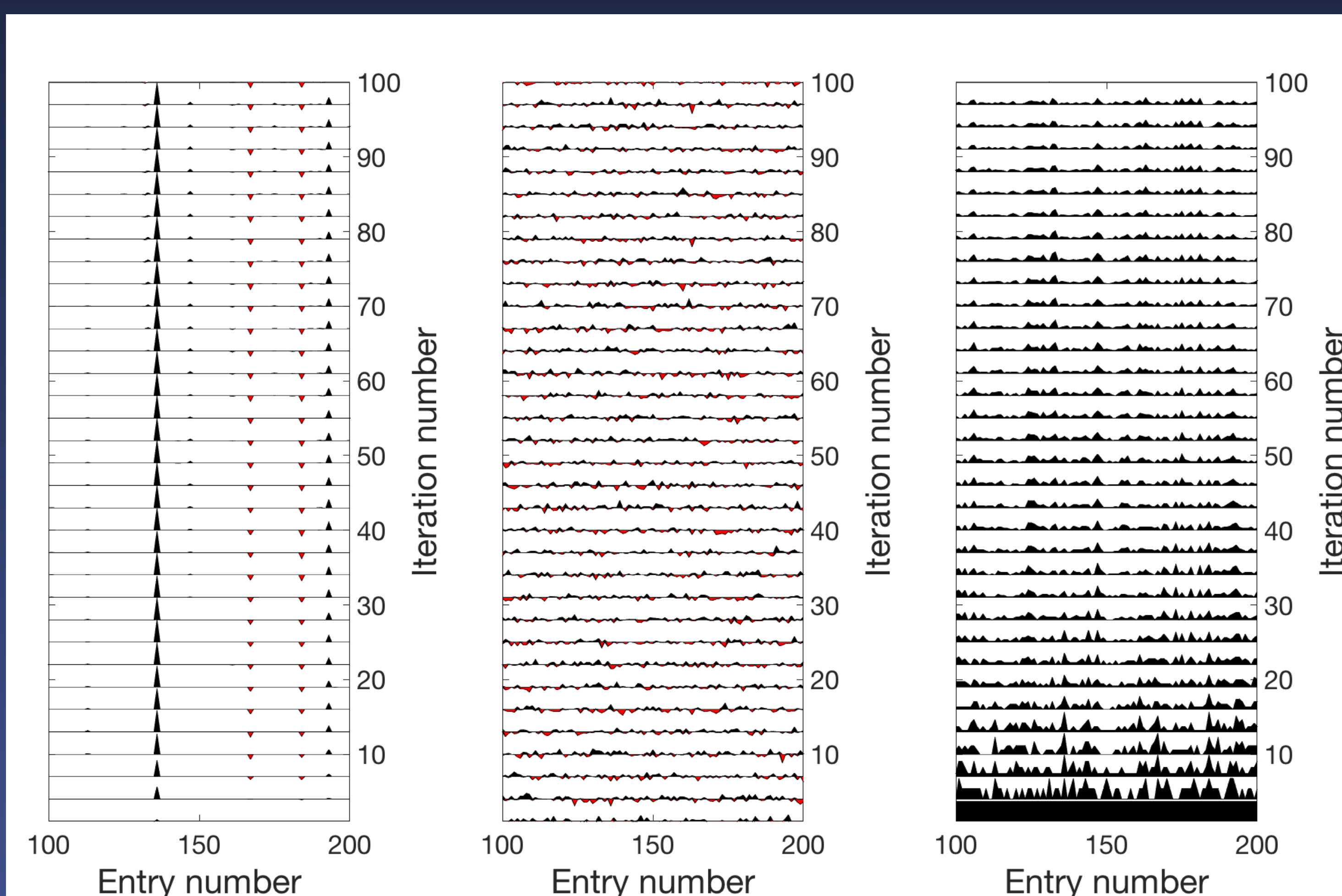
Model vs. gradient



LB

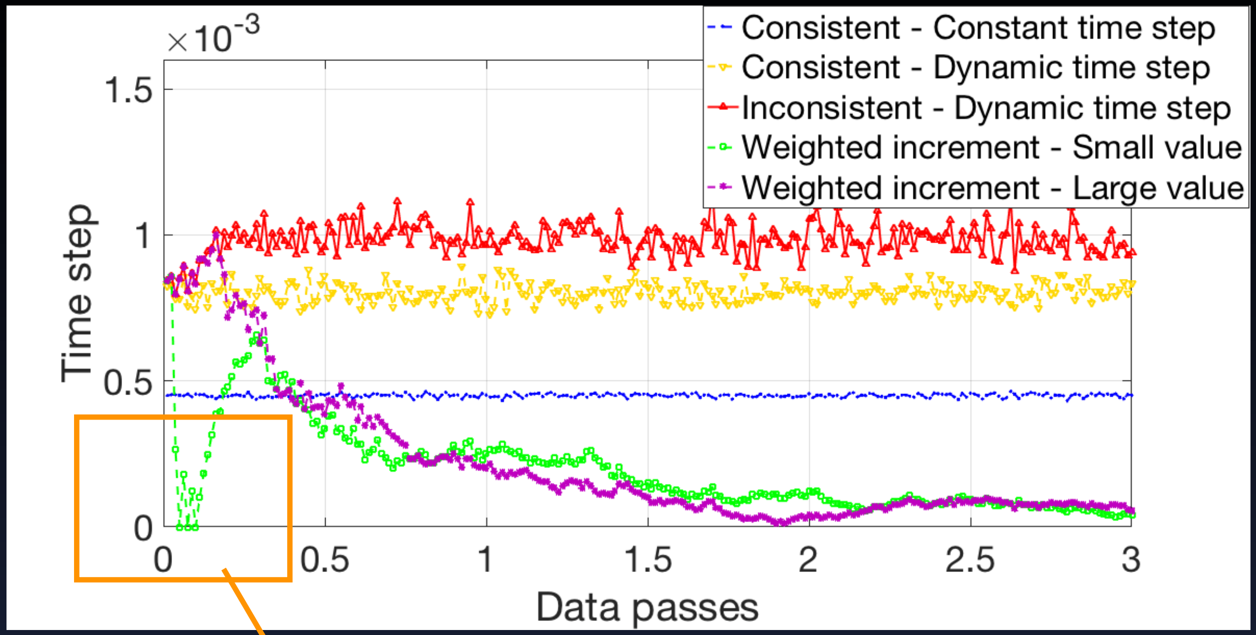


Time-step

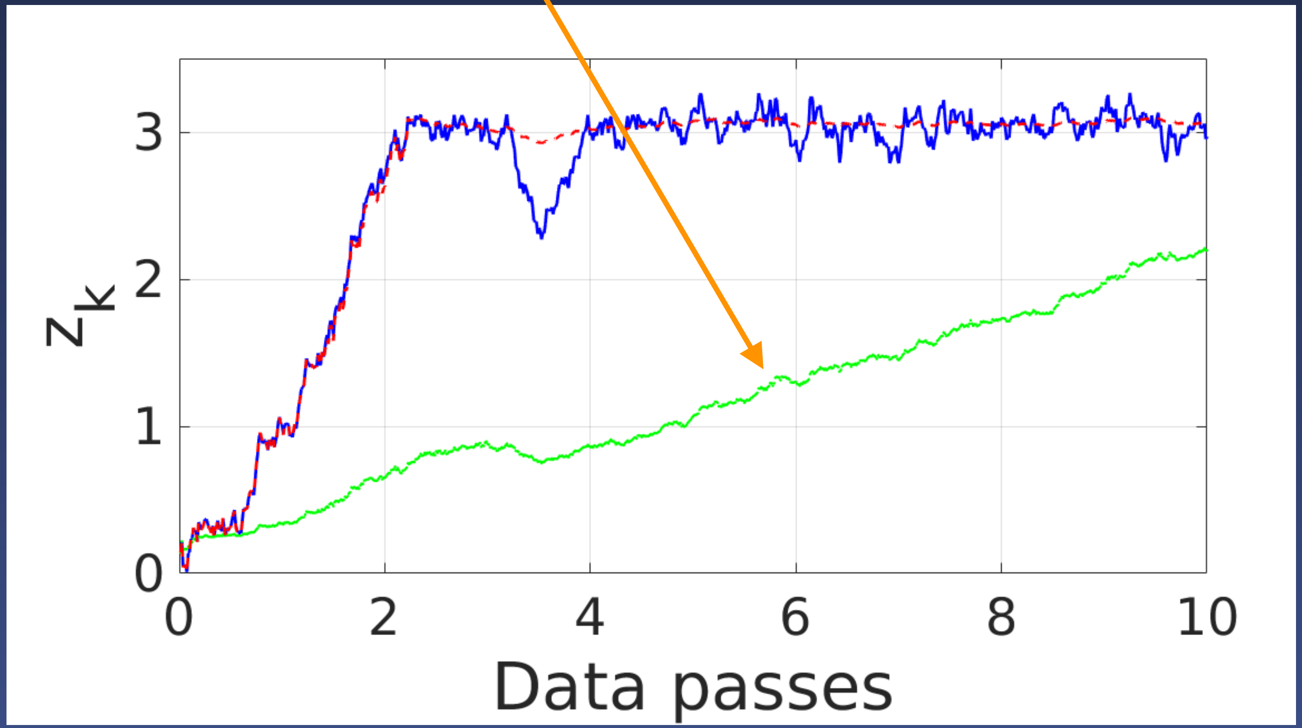


MLB

Recall variable time step: entry dependent



Small time step for smaller entries: slower convergence = long transients for certain entries



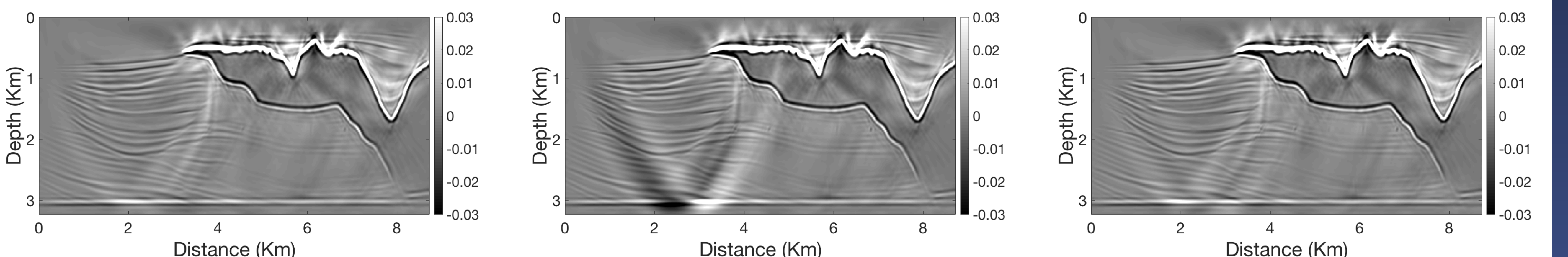
Same threshold of LB and MLB

Slow convergence of small entries

Is this a problem for sparse solutions?

In real life, solutions are compressible: entries in solution decrease in magnitude with some exponent

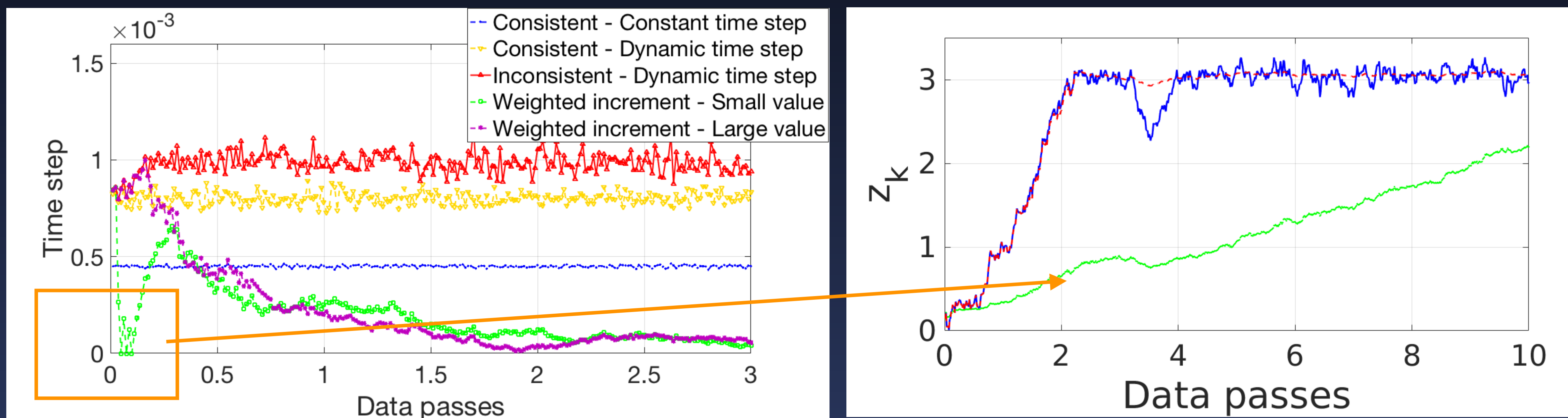
Implications:
separating solution from noise is tricky when resolving small entries



(a) Iteration 21 (b) Iteration 22 (c) Iteration 23

Slow convergence of small entries

Typically small entries below threshold - move slowly to threshold for MLB, due to chatter removal variable time step



MLB+T: Include threshold detection: use entry-specific time step from MLB only after entry crosses threshold

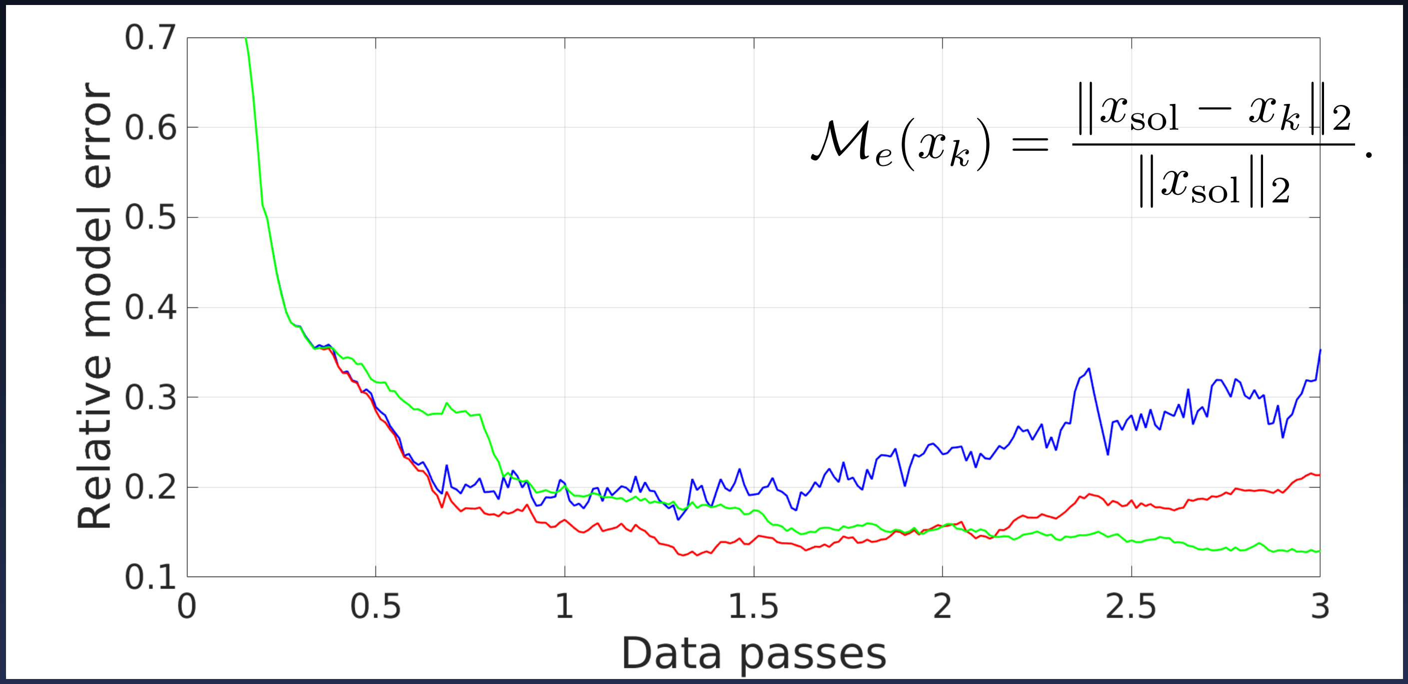
MLB+T = LB for entries not yet crossing the threshold

Slow convergence of small entries

Implications: separating small entries from noise

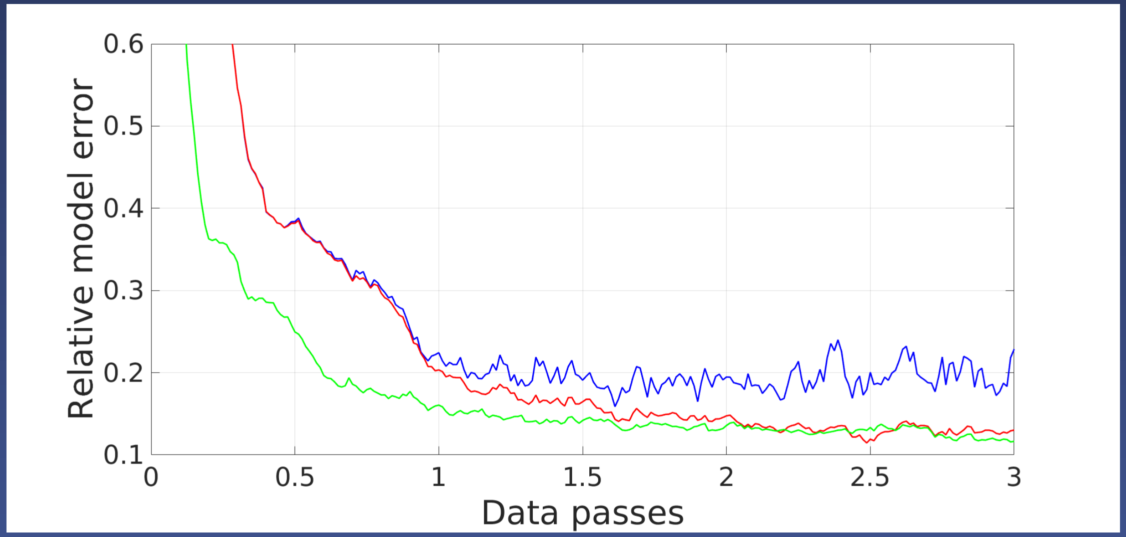
Danger of over-fitting the noise:

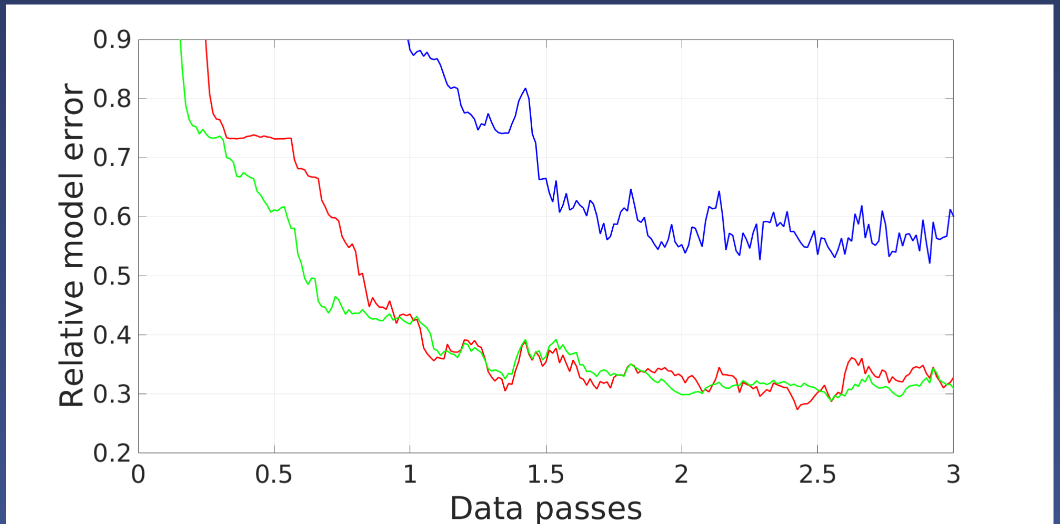
Small entries are not rejected after crossing the threshold



Several transient phases in LB method:

1. Honing in on large entries
2. Iterate to include small entries
3. When to stop to avoid over-fitting of noise?





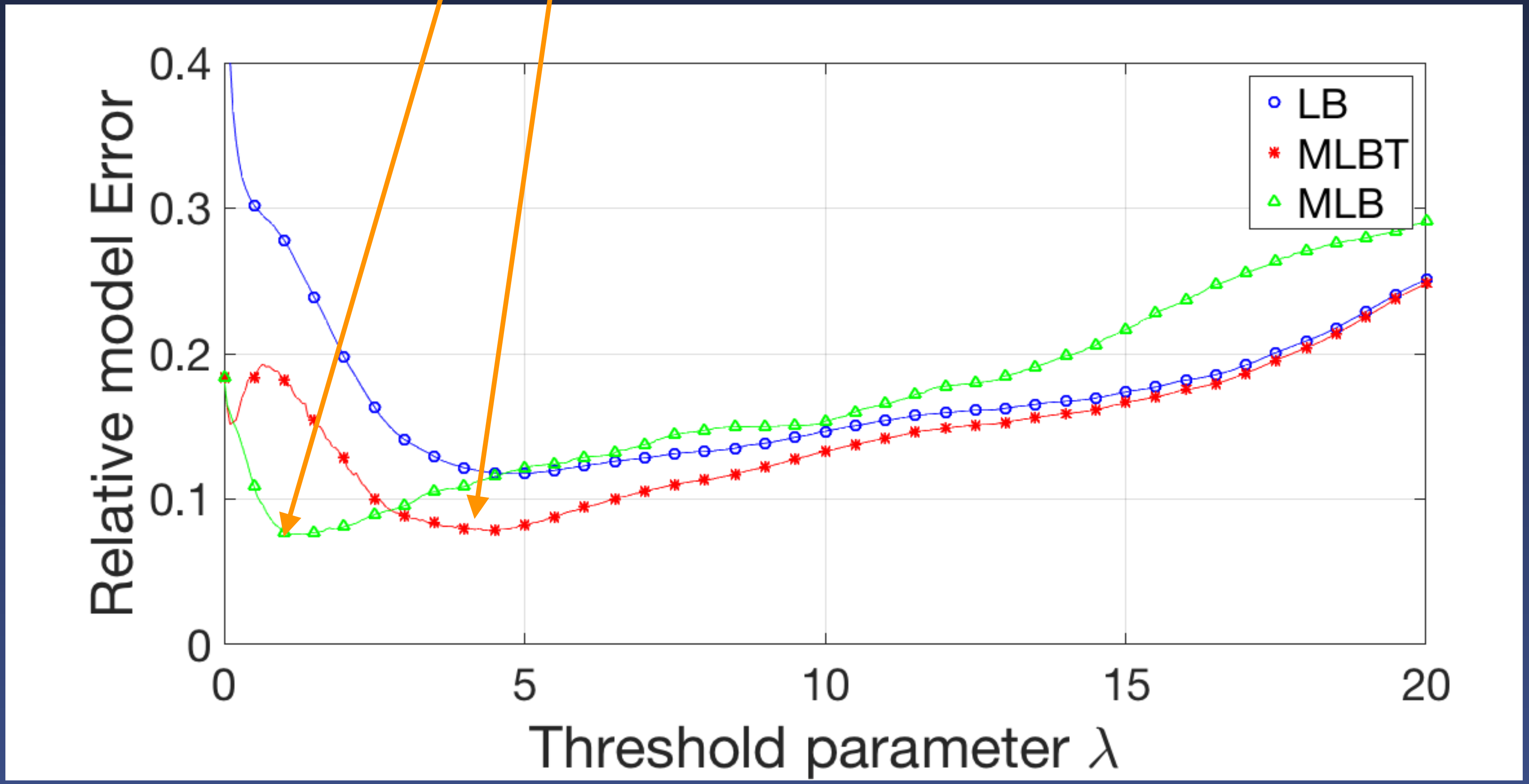
Estimate for λ, using dynamics of LB

1. Honing in on large entries
2. Iterate to include small entries - approaching size of noise
3. When to stop to avoid over-fitting of noise?

$$\lambda = N_d \max_{i \in \mathcal{S}} (|t_k [A_k^T \varepsilon]^i|)$$

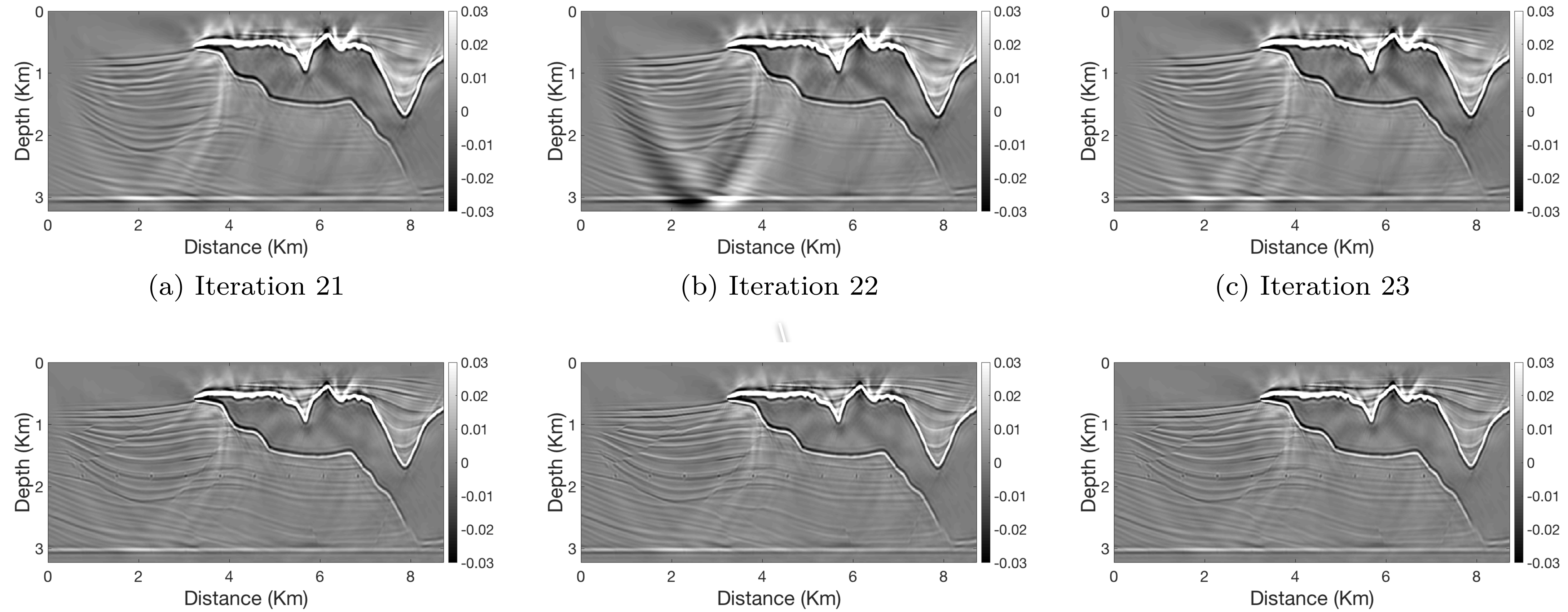
N_d = number of data passes

Computing transients



Daskalakis,K, Herrmann 2019

Large scale problem:



(a) Iteration 21 (b) Iteration 22 (c) Iteration 23

(d) Iteration 21, with the proposed modification (e) Iteration 22, with the proposed modification (f) Iteration 23, with the proposed modification

MLBT - similar results for reduction in data passes: 15-20%

Limitations for approximating λ :

Typically level of noise not known - estimates used in LB + projection

Model error not known - instead residual used

Sources of non-smooth and stochastic dynamics

Thresholds - connection to sparsity

$$\min_x \|x\|_1 \quad \text{subject to} \quad Ax = b\,.$$

Projections: use of error bounds to reduce search space

Representations: e.g. ReLU commonly used in ML

Online/Streaming Applications:

Network perspectives: ML

Non-convexity - use of methods such as stochastic gradient descent

(recursive) Layers, CNN's

DS perspectives:

Landscape perspectives: interacting particle systems. (e.g. Rotskoff, et al 2018; Mei, 2019)

Lagrangian formulation for accelerated methods

Wibisono, et al 2016

Direction dependent time step

Yezzi, et al 2018

Modified equations: cts approximations of discrete algorithm + correction - connections to multiple scale dynamics

Potential for noise sustained oscillations: accelerated methods, without thresholds

DS perspectives:

Potential for noise sustained oscillations (without thresholds)

$$x_{k+1} = x_k + \beta_1(x_k - x_{k-1}) - t_k \nabla f(x_n + \beta_2(x_k - x_{k-1}))$$

Accelerated (higher order) methods: e.g. Nesterov, Heavy ball, etc

Inconsistent:

coherence resonance-type result

Larger β_j

Reduced noise