Balanced truncation for Bayesian inference

Elizabeth Qian

September 19, 2022 | Symposium on Inverse Problems
University of Potsdam Campus Griebnitzsee

Inverse problems play a central role in diverse disciplines:

- Geophysics
- Oceanography
- Atmospheric science
- Medical imaging

with **critical applications for humanity's future**,
e.g., in climate and earth resource modeling.

Bayesian inference views the inverse problem through a probabilistic lens:

- Uncertain model parameters $\mathbf{p} \in \mathbb{R}^d$ are endowed with a **prior** probability distribution

- Measured data, $\mathbf{m} \in \mathbb{R}^{d_{\text{obs}}}$, are obtained by applying a forward map, $\mathbf{G}: \mathbb{R}^d \to \mathbb{R}^{d_{\text{obs}}}$, polluted by additive measurement noise $\boldsymbol{\epsilon}$

$$\mathbf{m} = \mathbf{G}(\mathbf{p}) + \epsilon$$

- This measurement model defines a **likelihood** distribution for $\mathbf{m}|\mathbf{p}$

Bayesian inference views the inverse problem through a probabilistic lens:

- After data $\mathbf{m}$ are obtained, Bayes' theorem is used to compute a **posterior** distribution for $\mathbf{p}|\mathbf{m}$

$$\mathbb{P}(\mathbf{p}|\mathbf{m}) \propto \mathbb{P}(\mathbf{m}|\mathbf{p})\,\mathbb{P}(\mathbf{p})$$

- The posterior reflects our updated view of the probability of the parameters conditioned on the measurements

In application, **computational challenges** to the use of Bayesian inference are posed by

- high dimensional parameter $\mathbf{p}$
- expensive forward model $\mathbf{G}$

For example:

- $\mathbf{p}$ is the initial condition of a spatially discretized time-dependent PDE,
- measurements are obtained at times $t_i > 0$, so that
- evaluating $\mathbf{G}$ requires simulating the PDE

Inference problem formulation: linear dynamical system

Unknown parameter is initial state:

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t), \quad \mathbf{x}(0) = \mathbf{p}$$

Measurements at n times $t_1, \ldots, t_n > 0$:

$$\mathbf{m}_i = \mathbf{C}\mathbf{x}(t_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \Gamma_{\text{out}})$$

where $\mathbf{A} \in \mathbb{R}^{d \times d}$ and $\mathbf{C} \in \mathbb{R}^{d_{\text{out}} \times d}$

Inference problem formulation: prior, forward model, likelihood

We assume a Gaussian prior: $\mathbf{p} \sim \mathcal{N}(0, \Gamma_{\mathrm{pr}})$

For n measurements at times $t_1, \dots, t_n$, we have:

$$\mathbf{G} = \begin{bmatrix} \mathbf{C}e^{\mathbf{A}t_1} \\ \vdots \\ \mathbf{C}e^{\mathbf{A}t_n} \end{bmatrix}, \qquad \mathbf{\Gamma}_{\mathrm{obs}} = \begin{bmatrix} \Gamma_{\mathrm{out}} & & \\ & \ddots & \\ & & \Gamma_{\mathrm{out}} \end{bmatrix}$$

The likelihood is Gaussian: $\mathbf{m}|\mathbf{p} \sim \mathcal{N}(\mathbf{Gp}, \Gamma_{\mathrm{obs}})$

Inference problem solution: the posterior

Gaussian prior: $\mathbf{p} \sim \mathcal{N}\left(0, \Gamma_{\text{pr}}\right)$

Gaussian likelihood: $\mathbf{m}|\mathbf{p} \sim \mathcal{N}\left(\mathbf{G}\mathbf{p}, \Gamma_{\text{obs}}\right)$

Gaussian posterior: $\mathbf{p}|\mathbf{m} \sim \mathcal{N}\left(\mu_{\text{pos}}, \Gamma_{\text{pos}}\right)$

where $\mu_{\text{pos}} = \Gamma_{\text{pos}}\mathbf{G}^{\top}\Gamma_{\text{obs}}^{-1}\mathbf{m}$, $\Gamma_{\text{pos}} = \left(\mathbf{H} + \Gamma_{\text{pr}}^{-1}\right)^{-1}$, and $\mathbf{H} = \mathbf{G}^{\top}\Gamma_{\text{obs}}^{-1}\mathbf{G}$.

Gaussian posterior: $\mathbf{p}|\mathbf{m} \sim \mathcal{N}\left(\mu_{\text{pos}}, \Gamma_{\text{pos}}\right)$

where $\mu_{\text{pos}} = \Gamma_{\text{pos}} \mathbf{G}^\top \Gamma_{\text{obs}}^{-1} \mathbf{m}$, $\Gamma_{\text{pos}} = \left(\mathbf{H} + \Gamma_{\text{pr}}^{-1}\right)^{-1}$, and $\mathbf{H} = \mathbf{G}^\top \Gamma_{\text{obs}}^{-1} \mathbf{G}$.

Challenge: computing the posterior is expensive when

- $\boldsymbol{p}$ is high-dimensional (d is large) and
- $\mathbf{G}$ is only implicitly available through evolving the high-dimensional dynamical system.

Solution: reduce dimension of $\boldsymbol{p}$ and $\mathbf{G}$ via

Balanced truncation for Bayesian inference

Preview

Balanced truncation can be **naturally adapted** to Bayesian inference for linear dynamics.

We also **make connections** between

- established **system-theoretic model reduction** analysis and
- theoretical **linear inference results** from [Spantini et al. SISC 2015]

to show that, in certain settings, the resulting reduced model

- is **balanced**, **stable**, has a computable **error bound**,
- and recovers an **optimal** posterior covariance approximation.

Background

1. Balanced truncation for linear time-invariant systems
2. Optimal posterior approximation for linear Gaussian inference

Linear time-invariant (LTI) systems

The system with input $\boldsymbol{u}(t) \in \mathbb{R}^{d_{\text{in}}}$,

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}(t)$$
$$\mathbf{y}(t) = \mathbf{F}\mathbf{x}(t)$$

has infinite reachability and observability Gramians:

$$\mathbf{P} = \int_0^\infty e^{\mathbf{A}t}\mathbf{B}\mathbf{B}^\top e^{\mathbf{A}^\top t}\,\mathrm{d}t, \quad \mathbf{Q} = \int_0^\infty e^{\mathbf{A}^\top t}\mathbf{F}^\top \mathbf{F} e^{\mathbf{A}t}\,\mathrm{d}t$$

Reachability and observability energies

$\mathbf{P}, \mathbf{Q}$ define reachability and observability energies:

$$\|\mathbf{x}\|_{\mathbf{P}^{-1}}^2 = \mathbf{x}^\top \mathbf{P}^{-1} \mathbf{x}, \qquad \|\mathbf{x}\|_{\mathbf{Q}}^2 = \mathbf{x}^\top \mathbf{Q} \mathbf{x}$$

- **low reachability energy = easy to reach** from the origin (requiring only small controls)
- **high observability energy = easy to observe** (large contribution to the output)

Goal: retain rank-r subspace of directions that are both easy to observe and easy to reach.

The balanced truncation subspace

The state directions retained by balanced truncation maximize the Rayleigh quotient,

$$\frac{\mathbf{x}^\top \mathbf{Q} \mathbf{x}}{\mathbf{x}^\top \mathbf{P}^{-1} \mathbf{x}} = \frac{\|\mathbf{x}\|_{\mathbf{Q}}^2}{\|\mathbf{x}\|_{\mathbf{P}^{-1}}^2}$$

These are the generalized eigenvectors of the pencil $(\mathbf{Q}, \mathbf{P}^{-1})$:

$$\mathbf{Q}\mathbf{v} = \delta^2 \mathbf{P}^{-1} \mathbf{v}$$

The δ are the *Hankel singular values* of the LTI system.

Balanced truncation: the reduced model

Balanced truncation obtains a reduced model of size r,

$$\dot{\mathbf{x}}_r(t) = \mathbf{A}_r \mathbf{x}_r + \mathbf{B}_r \mathbf{u}(t)$$
$$\mathbf{y}_r(t) = \mathbf{F}_r \mathbf{x}(t)$$

by **transforming** to the balanced (generalized eigenvector) basis and **truncating** to the leading r states, so that

$$\mathbf{A}_r \in \mathbb{R}^{r \times r}, \quad \mathbf{B}_r \in \mathbb{R}^{r \times d_{\text{in}}}, \quad \mathbf{F}_r \in \mathbb{R}^{d_{\text{out}} \times r}$$

Properties of balanced truncation models

If the original LTI system is linearly stable and minimal, then:

1. The reduced model is balanced: its infinite reachability and observability gramians are diagonal and equal.
2. The reduced model is linearly stable.
3. The reduced output error is bounded by:

$$\|\mathbf{y}(t) - \mathbf{y}_r(t)\|_{L^2(\mathbb{R})} \leq 2 \sum_{j=r+1}^{d} \delta_j \|\mathbf{u}(t)\|_{L^2(\mathbb{R})}$$

Balanced truncation **exploits low-dimensional structure in the input-output map** of an LTI system to reduce its state dimension.

A different low-dimensional structure arises in many Bayesian inference problems:

- because measured **data are only informative in a low-rank subspace** of the parameter space.

Background

1. Balanced truncation for linear time-invariant systems
2. Optimal posterior approximation for linear Gaussian inference

Exploiting low-rank informativeness

Prior: $\mathbf{p} \sim \mathcal{N}(0, \Gamma_{\mathrm{pr}})$

Posterior: $\mathbf{p}|\mathbf{m} \sim \mathcal{N}(\mu_{\mathrm{pos}}, \Gamma_{\mathrm{pos}})$

Note: $\Gamma_{\mathrm{pos}} = \left(\mathbf{H} + \Gamma_{\mathrm{pr}}^{-1}\right)^{-1}$

Thus: $\Gamma_{\mathrm{pos}} \preccurlyeq \Gamma_{\mathrm{pr}}$

Γ_{pos} shrinks relative to Γ_{pr} only in directions where data is informative

Motivates approximating Γ_{pos} by

$$\hat{\Gamma}_{\mathrm{pos}} = \Gamma_{\mathrm{pr}} - KK^{\top}$$

where K is low-rank

Posterior covariance approximation

From Spantini et al. SISC 2015:

Seek $\hat{\Gamma}_{\text{pos}}$ in class of rank-r negative semidefinite updates to Γ_{pr}:

$$\mathcal{M}_r = \{\Gamma_{\text{pr}} - KK^\top : \text{rank}(K) \leq r\}$$

Measure approximation quality using Förstner distance for symmetric positive definite matrices:

$$d_{\mathcal{F}}(A, B) = \sum_{i=1}^{d} \ln^2(\sigma_i)$$

where σ_i are the generalized eigenvalues of (A, B) satisfying

$$Av_i = \sigma_i B v_i$$

Posterior covariance approximation

From Spantini et al. SISC 2015:

Seek $\hat{\Gamma}_{\text{pos}}$ in class of rank-r negative semidefinite updates to Γ_{pr}:

$$\mathcal{M}_r = \{\Gamma_{\text{pr}} - KK^\top : \text{rank}(K) \leq r\}$$

Optimal approximation

$$\min_{M \in \mathcal{M}_r} d_{\mathcal{F}}(\Gamma_{\text{pos}}, M) \equiv \hat{\Gamma}_{\text{pos}} = \Gamma_{\text{pr}} - K_* K_*^\top$$

determined by generalized eigenvalue problem of $\left(H, \Gamma_{\text{pr}}^{-1}\right)$:

$$H v_i = \tau_i^2 \Gamma_{\text{pr}}^{-1} v_i$$

Posterior covariance approximation

From Spantini et al. SISC 2015:

Optimal approximation

$$\min_{M \in \mathcal{M}_r} d_{\mathcal{F}}(\Gamma_{\text{pos}}, M) \equiv \hat{\Gamma}_{\text{pos}} = \Gamma_{\text{pr}} - K_* K_*^\top$$

determined by generalized eigenvalue problem of $\left(H, \Gamma_{\text{pr}}^{-1}\right)$:

$$H v_i = \tau_i^2 \Gamma_{\text{pr}}^{-1} v_i$$

Main result (Spantini): optimal covariance update directions of K_* are the dominant eigendirections of the above pencil.

Key connections

Balanced truncation for LTI systems:

- Generalized eigenvalue problem for $(\mathbf{Q}, \mathbf{P}^{-1})$

Optimal posterior covariance approximation:[Spantini 2015]

- Generalized eigenvalue problem for $\left(\mathbf{H}, \Gamma_{\mathrm{pr}}^{-1}\right)$

We identify natural analogies between:

- Reachability Gramian $\mathbf{P}$ and prior covariance Γ_{pr}
- Observability Gramian $\mathbf{Q}$ and Fisher information matrix $\mathbf{H}$

… to propose a balanced truncation approach for Bayesian inverse problems for LTI systems.

Balanced truncation for Bayesian inference

Uniting system-theoretic model reduction with linear inference results

Reachability and the prior covariance

Recall our inference setting:

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t), \quad \mathbf{x}(0) = \mathbf{p}$$

Suppose we 'spin up' the system from $t = -\infty$ with white noise:

$$\mathrm{d}\mathbf{x} = \begin{cases} \mathbf{A}\mathbf{x}\,\mathrm{d}t + \mathbf{B}\,\mathrm{d}\mathbf{W}(t), & t < 0, \\ \mathbf{A}\mathbf{x}\,\mathrm{d}t, & t \geq 0 \end{cases}$$

Then, a natural prior is the stationary distribution at $t = 0$:

$$\mathbb{E}[\mathbf{x}(0)] = 0,$$

$$\mathbb{E}[\mathbf{x}(0)\mathbf{x}^\top(0)] = \int_0^\infty e^{\mathbf{A}\tau}\mathbf{B}\mathbf{B}^\top e^{\mathbf{A}^\top\tau}\,\mathrm{d}\tau$$

leading to

$$\mu_{\mathrm{pr}} = 0, \quad \Gamma_{\mathrm{pr}} = \mathbf{P}$$

Prior covariance compatibility

Spin-up process can be done for any arbitrary $\mathbf{B}$ that we choose.

Can any prior covariance be interpreted this way? **No.**

Definition: A prior covariance is *compatible* with the linear system dynamics if $\mathbf{A}\Gamma_{\mathrm{pr}} + \Gamma_{\mathrm{pr}}\mathbf{A}^{\top} \preceq 0.$

Compatibility allows the prior covariance to be interpreted as a reachability Gramian without explicitly defining $\mathbf{B}$.

Observability and the Fisher information

Recall our measurement model: $\mathbf{m} = \mathbf{G}\mathbf{p} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Gamma_{\text{obs}})$

$$\mathbf{G} = \begin{bmatrix} \mathbf{C}e^{\mathbf{A}t_1} \\ \vdots \\ \mathbf{C}e^{\mathbf{A}t_n} \end{bmatrix}, \quad \boldsymbol{\Gamma}_{\text{obs}} = \begin{bmatrix} \Gamma_{\text{out}} & & \\ & \ddots & \\ & & \Gamma_{\text{out}} \end{bmatrix}$$

The Fisher information matrix is

$$\mathbf{H} = \mathbf{G}^\top \Gamma_{\text{obs}}^{-1} \mathbf{G} = \sum_{i=1}^{n} e^{\mathbf{A}^\top t_i} \mathbf{C}^\top \Gamma_{\text{out}}^{-1} \mathbf{C} e^{\mathbf{A} t_i}$$

Compare to the LTI observability Gramian: $\mathbf{Q} = \int_0^\infty e^{\mathbf{A}^\top t} \mathbf{F}^\top \mathbf{F} e^{\mathbf{A}t} \, \mathrm{d}t$

The limit of continuous observations

Fisher information:

$$\mathbf{H} = \sum_{i=1}^{n} e^{\mathbf{A}^\top t_i} \mathbf{C}^\top \Gamma_{\text{out}}^{-1} \mathbf{C} e^{\mathbf{A} t_i}$$

Observability Gramian:

$$\mathbf{Q} = \int_0^\infty e^{\mathbf{A}^\top t} \mathbf{F}^\top \mathbf{F} e^{\mathbf{A} t} \, \mathrm{d}t$$

Proposition [Q. et al. Journal of Scientific Computing 2022]:

Summary: Suppose $\mathbf{F} = \Gamma_{\text{out}}^{-1/2} \mathbf{C}$ and the measurement times t_i are Δt apart. Then, as $n \to \infty$ and $\Delta t \to 0$, an appropriate rescaling of $\mathbf{H}$ converges to $\mathbf{Q}$.

Significance: Directions of higher observability energy correspond to directions most informed by data in an idealized measurement model.

Main idea and result

We propose the use of a balanced truncation reduced model based on the pencil $\left(\mathbf{Q}, \Gamma_{\mathrm{pr}}^{-1}\right)$, where

- $\mathbf{F} = \Gamma_{\mathrm{out}}^{-1/2}\mathbf{C}$ is used to define the infinite observability Gramian
- Γ_{pr} is a compatible prior covariance

Theorem [Q. et al. Journal of Scientific Computing 2022]:

- This reduced model is **stable, balanced,** and has an **error bound** in terms of the tail sum of the Hankel singular values.
- Further, in the limit of infinite observations, the reduced model **leads to the Spantini optimal posterior covariance approximation.**

Numerical experiments

Tests for two model reduction benchmarks

Both examples downloadable from slicot.org

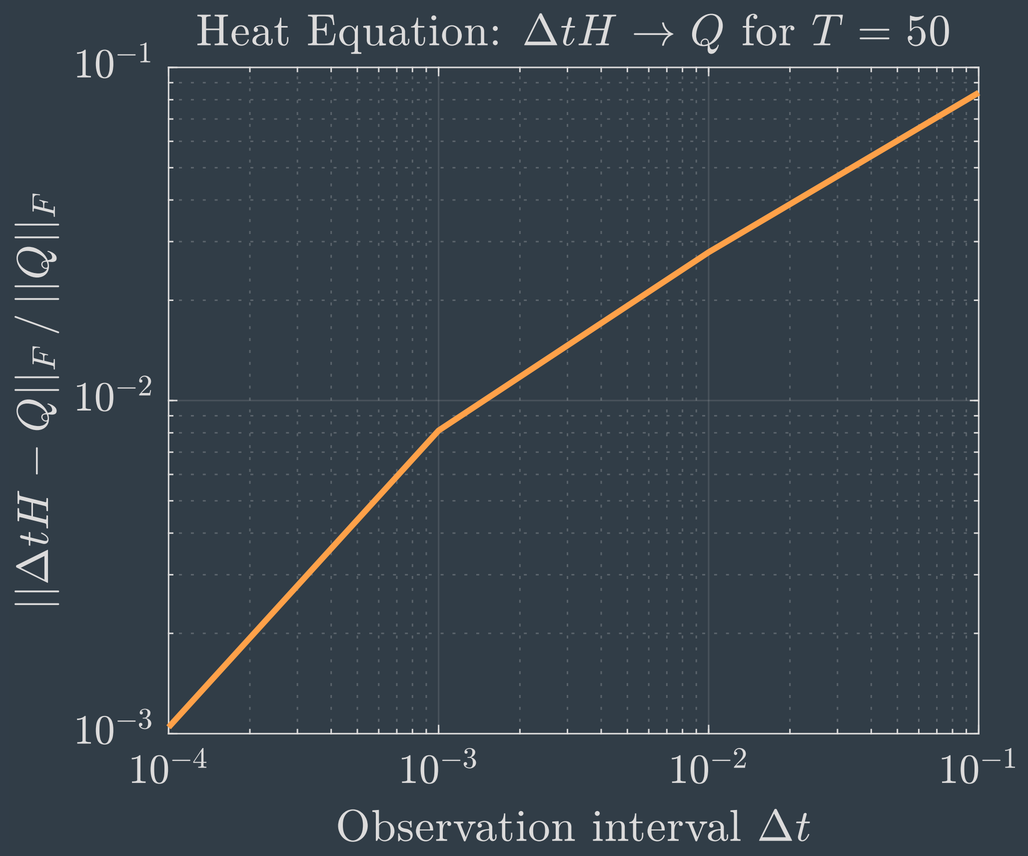
1. Heat equation in 1D rod
2. ISS1R structural model – flex modes of Zvezna service module

For both problems, we compare the posterior mean and covariance approximations obtained via:

1. The **Spantini** optimal low-rank update approach based on $\left(H, \Gamma_{\text{pr}}^{-1}\right)$
2. **BT-Q:** Our proposed balanced truncation approach based on $\left(Q, \Gamma_{\text{pr}}^{-1}\right)$
3. **BT-H:** A variant of our proposed BT approach based on $\left(H, \Gamma_{\text{pr}}^{-1}\right)$

Heat equation problem

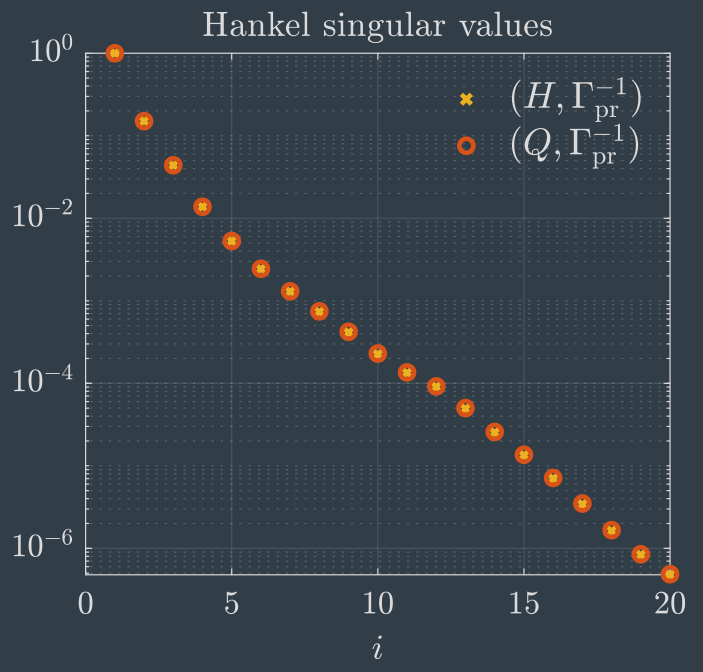
- $B = I$ used to define compatible prior
- True initial condition drawn from prior
- Output is temperature at 2/3 rod length
- Measurements made at
  $$\{\Delta t, 2\Delta t, \ldots, n\Delta t \equiv T\}$$
- 10% measurement noise added to output

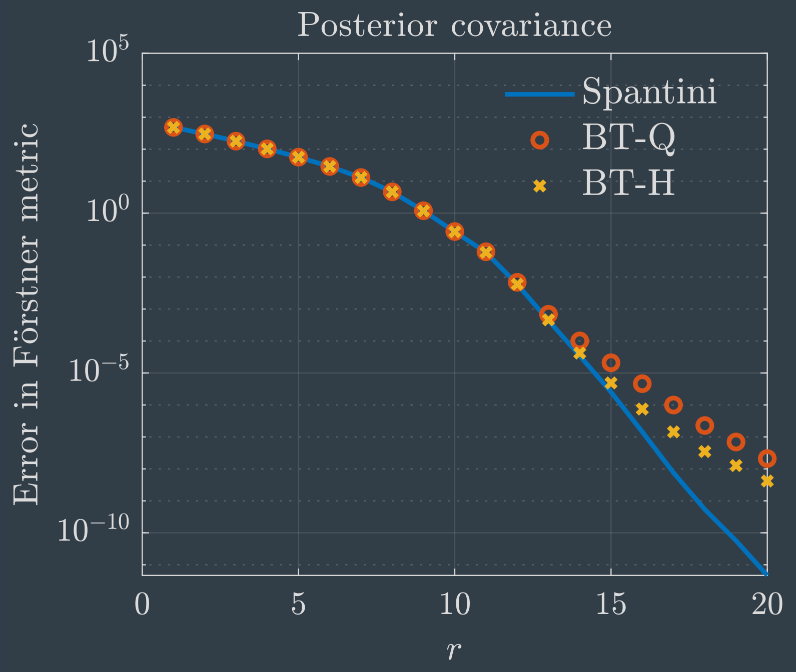


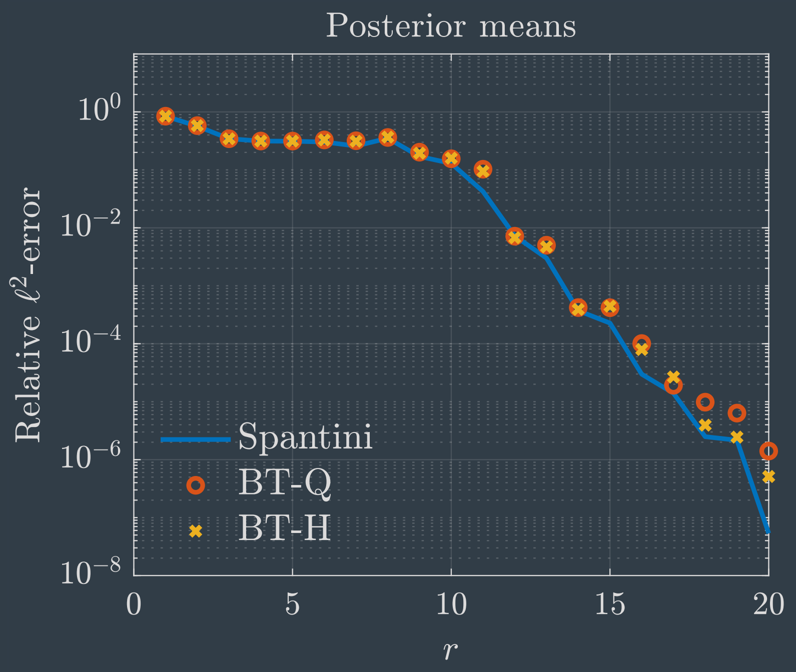
Heat equation: idealized measurements

"near continuous and forever" measurements:

- $\Delta t = 10^{-4}$ measurement spacing, $T = 50$
- Leads to $\Delta t H \approx Q$ with 0.1% relative Frobenius norm error
- BT reduction from $d = 200$ to $r = 20$ yields near-optimal posterior approximation



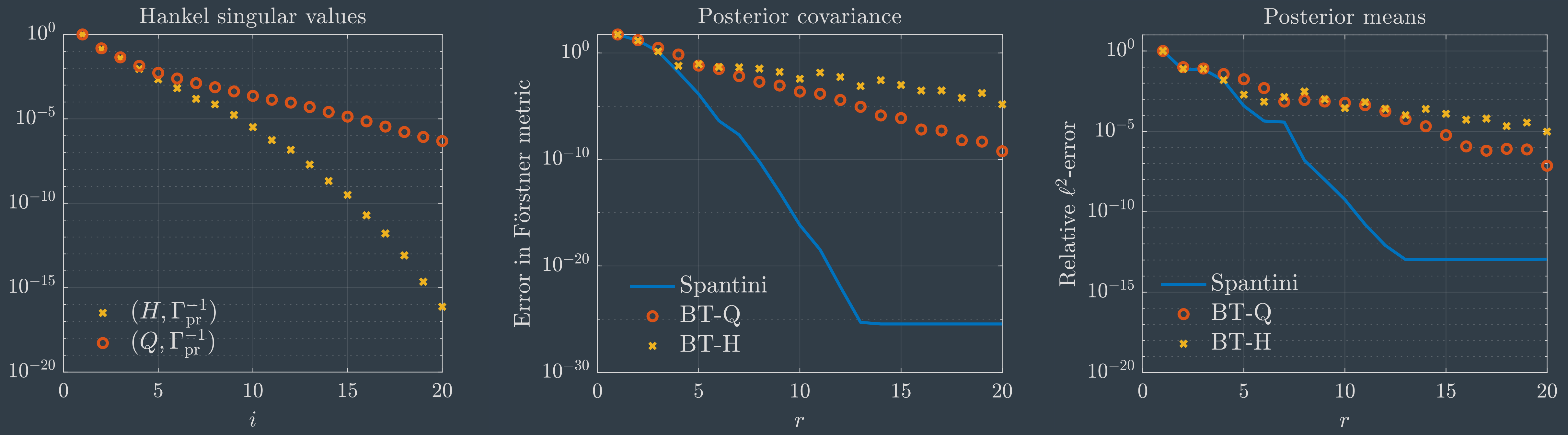




Heat equation: limited measurements

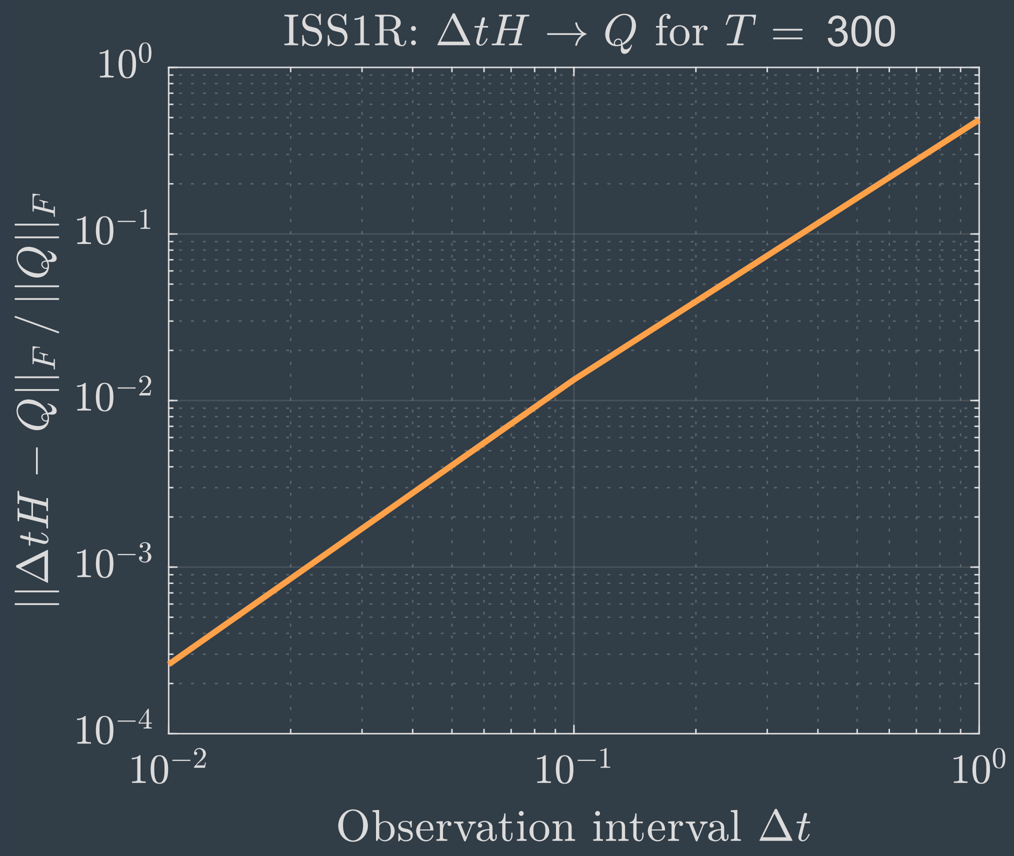
Limited coarse measurements:

- $\Delta t = 10^{-1}$ measurement spacing, $T = 10$
- Leads to $\Delta t H$ with 15% Frobenius norm error relative to Q
- BT reduction from $d = 200$ to $r = 20$ yields sub-optimal posterior approximation



ISS1R problem

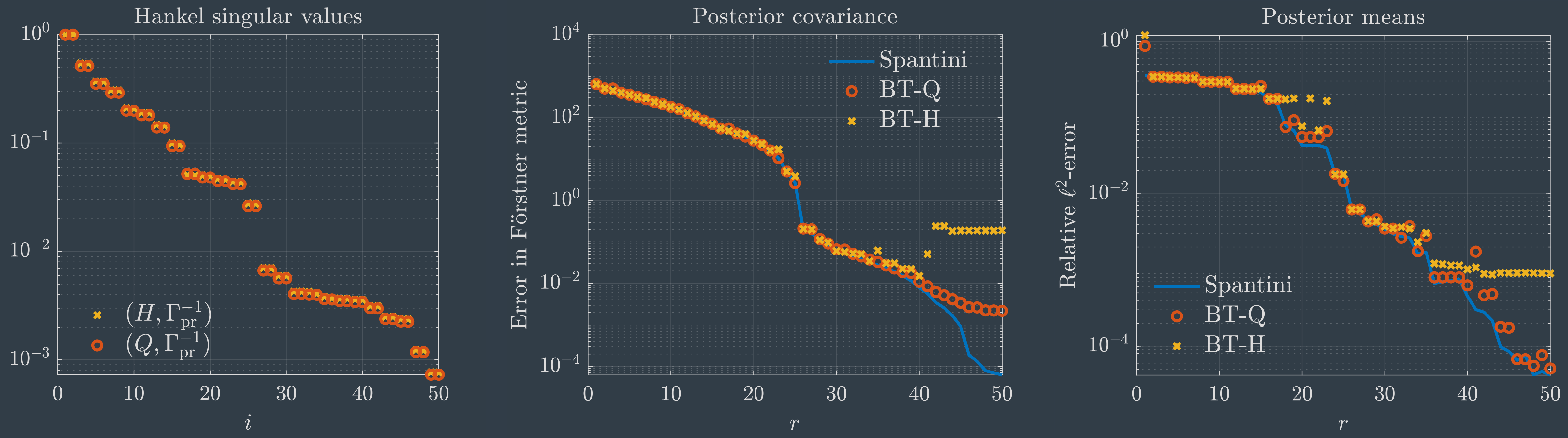
- Provided input port matrix B corresponds to roll/pitch/yaw jets; this is used to define compatible prior
- True initial condition drawn from prior
- Outputs are roll/pitch/yaw gyro readings
- Measurements made at
$$\{\Delta t, 2\Delta t, \dots, n\Delta t \equiv T\}$$
- 10% measurement noise added to output



ISS1R: idealized measurements

"near continuous and forever" measurements:

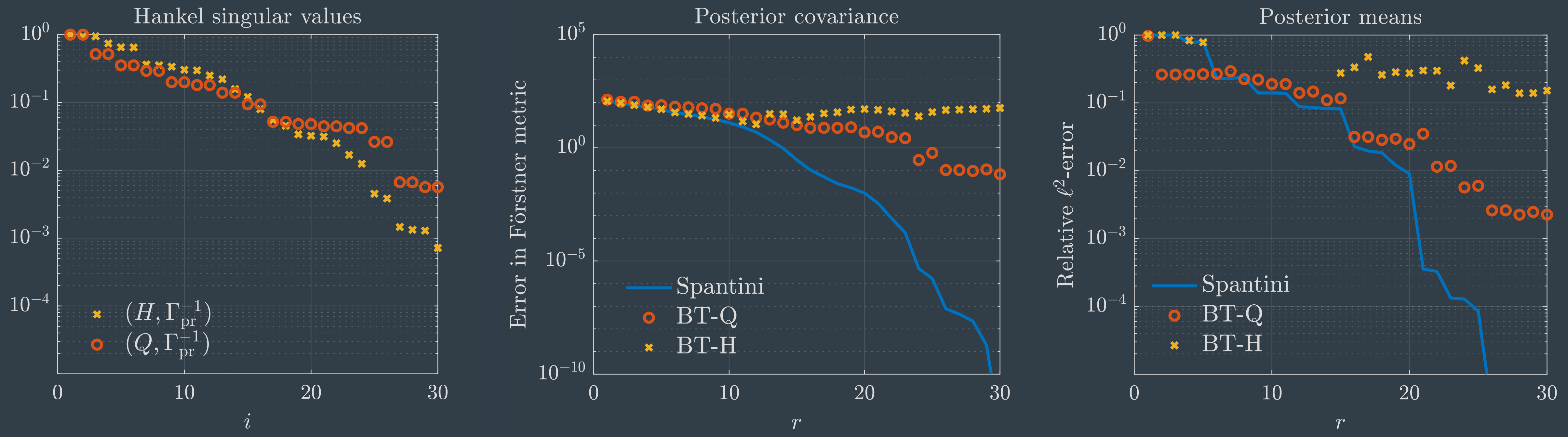
- $\Delta t = 10^{-1}$ measurement spacing, $T = 300$
- Leads to $\Delta t H \approx Q$ with 1% relative Frobenius norm error
- BT-Q reduction from $d = 270$ to $r = 50$ yields near-optimal posterior approximation



ISS1R: limited measurements

Limited coarse measurements:

- $\Delta t = 1$ measurement spacing, $T = 10$
- Leads to $\Delta t H \approx Q$ with 53% relative Frobenius norm error
- BT-Q reduction from $d = 270$ to $r = 30$ yields sub-optimal posterior approximation



Summary

Balanced truncation for Bayesian inference:

- $(\mathbf{Q}, \Gamma_{\text{pr}}^{-1})$ generalized eigenvalue problem defines reduced model
- stable, balanced, and subject to a computable error bound
- recovers the optimal posterior covariance approximation in certain limits
- **cheaply computable** and gives **accurate** posterior approximations in practical settings

LTI system theory

- $(\mathbf{Q}, \mathbf{P}^{-1})$ generalized eigenvalue problem defines balanced truncation model
- Reduced model is stable, balanced, subject to computable error bound

Linear Gaussian inference:

- $(\mathbf{H}, \Gamma_{\text{pr}}^{-1})$ generalized eigenvalue problem defines optimal posterior approximation

Future directions

Workhorse algorithms for Bayesian inference typically require 1000s of simulations

model reduction is a key enabler.

Many potential directions result from cross-pollination between existing system theory and work in Bayesian inference, including:

- Time-limited balanced truncation: see Josie König's poster later today!
- Nonlinear methods: see [Zahm et al. 2018] on the Bayesian side and [Benner & Goyal] for quadratic BT

Thank you!

Email: eqian@caltech.edu ➡️ eqian@gatech.edu

Web: https://www.elizabethqian.com

Qian, Tabeart, Beattie, Gugercin, Jiang, Kramer, and Narayan, “Model reduction of linear dynamical systems via balancing for Bayesian inference”, Journal of Scientific Computing 91(29) 2022.

Acknowledgements:

ICERM

This material is based upon work supported by the National Science Foundation under Grant No. DMS-1439786 and by the Simons Foundation Grant No. 50736 while the authors were in residence at the Institute for Computational and Experimental Research in Mathematics in Providence, RI, during the “Model and dimension reduction in uncertain and dynamic systems” program.

EQ was supported in part by the Fannie and John Hertz Foundation. JMT was partially supported by EPSRC grant EP/S027785/1. AN was partially supported by NSF DMS-1848508.