Bayesian inference as optimal transportation

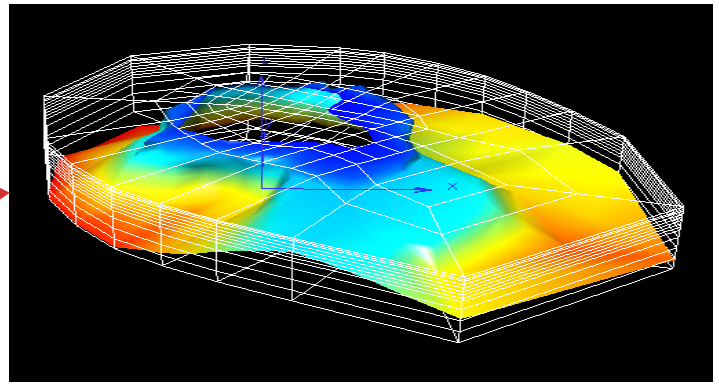
Youssef Marzouk and Tarek Moselhy

[1]Massachusetts Institute of Technology
Department of Aeronautics & Astronautics
Center for Computational Engineering

MIT AEROASTRO

UQ for inverse problems

- ***Example:*** *subsurface flow and transport*



$$d = G(x) + \epsilon$$

forward model (PDE)

observation/model errors

- Data are limited in number, noisy, and *indirect*
- Parameter x may be infinite-dimensional
- We wish to **quantify uncertainty** in x

Statistical inverse problems

- Take a Bayesian approach:

$$\pi^d(x) \equiv p(x \mid d) \propto p(d \mid x)\, p(x)$$

- Key strategies for making this computationally tractable:
  - ***Approximations*** of the forward model: spectral expansions, reduced order models, interpolation/regression
  - Efficient and structure-exploiting **sampling schemes** to explore the posterior distribution

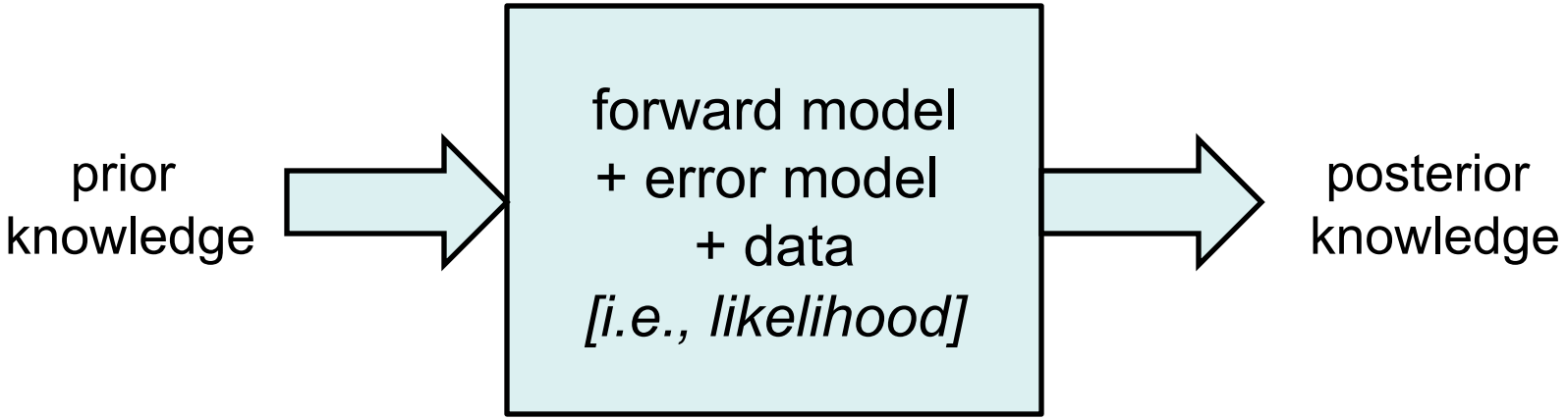
Alternatives to MCMC

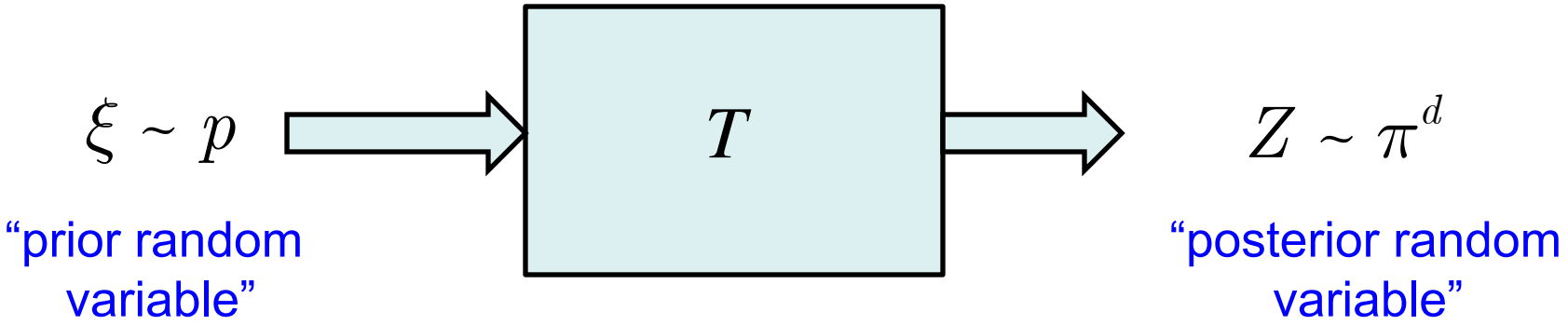
- Yesterday: exploring the posterior distribution with **Markov chain Monte Carlo (MCMC)**

- Is this the only way?

- A few drawbacks of MCMC:
  - Generates a stream of *correlated* samples
  - Proposal design is difficult; potential for poor mixing
  - No clear convergence criteria!
  - Requires a ***large*** number of forward model evaluations
  - Intrinsically serial
  - Posterior normalizing constants require additional effort

A different viewpoint

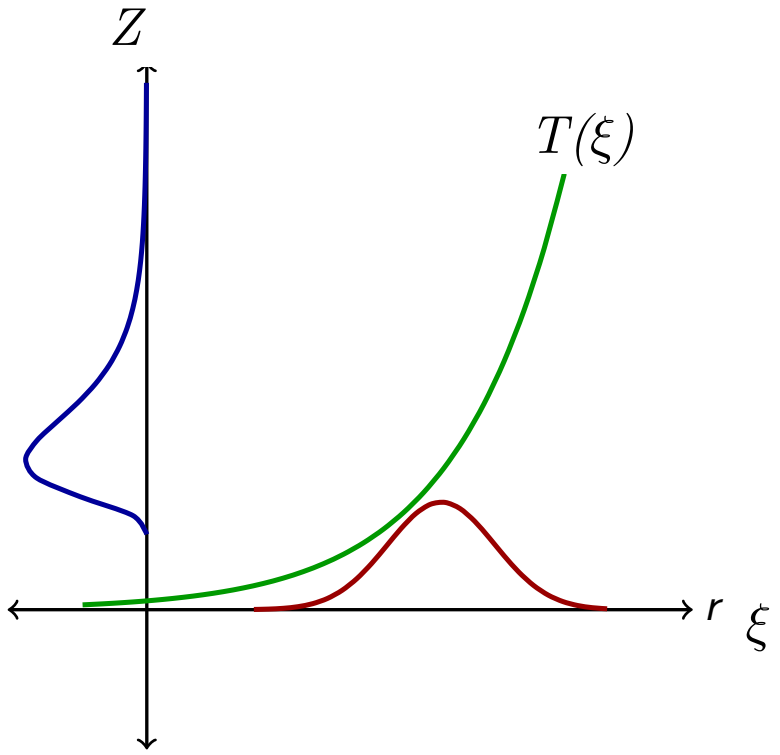




$$Z = T(\xi)$$

Optimal transport

- Deterministic coupling of two random variables, $\xi \sim \mu$, $Z \sim \nu$



- Monge problem: $\min_T \int c\Big(\xi, T(\xi)\Big)\, d\mu(\xi)$, where $T^\sharp \mu = \nu$
- Exists a unique and *monotone* solution for quadratic (and other) **transport costs** $c(x,y)$ [Brenier 1991, McCann 1995]

Optimal transport

- Let π be the target density and p be the density of a reference random variable ξ
  - Given the ability to (i) ***sample*** ξ and (ii) ***evaluate*** π ***up to a normalizing constant****,* one can compute numerical approximations to the optimal map
  - Found through solution of an optimization problem…
  - The reference distribution could be the prior, or it could be something else easy to sample from

Finding the transport map

- Start with the posterior *(target)* density as

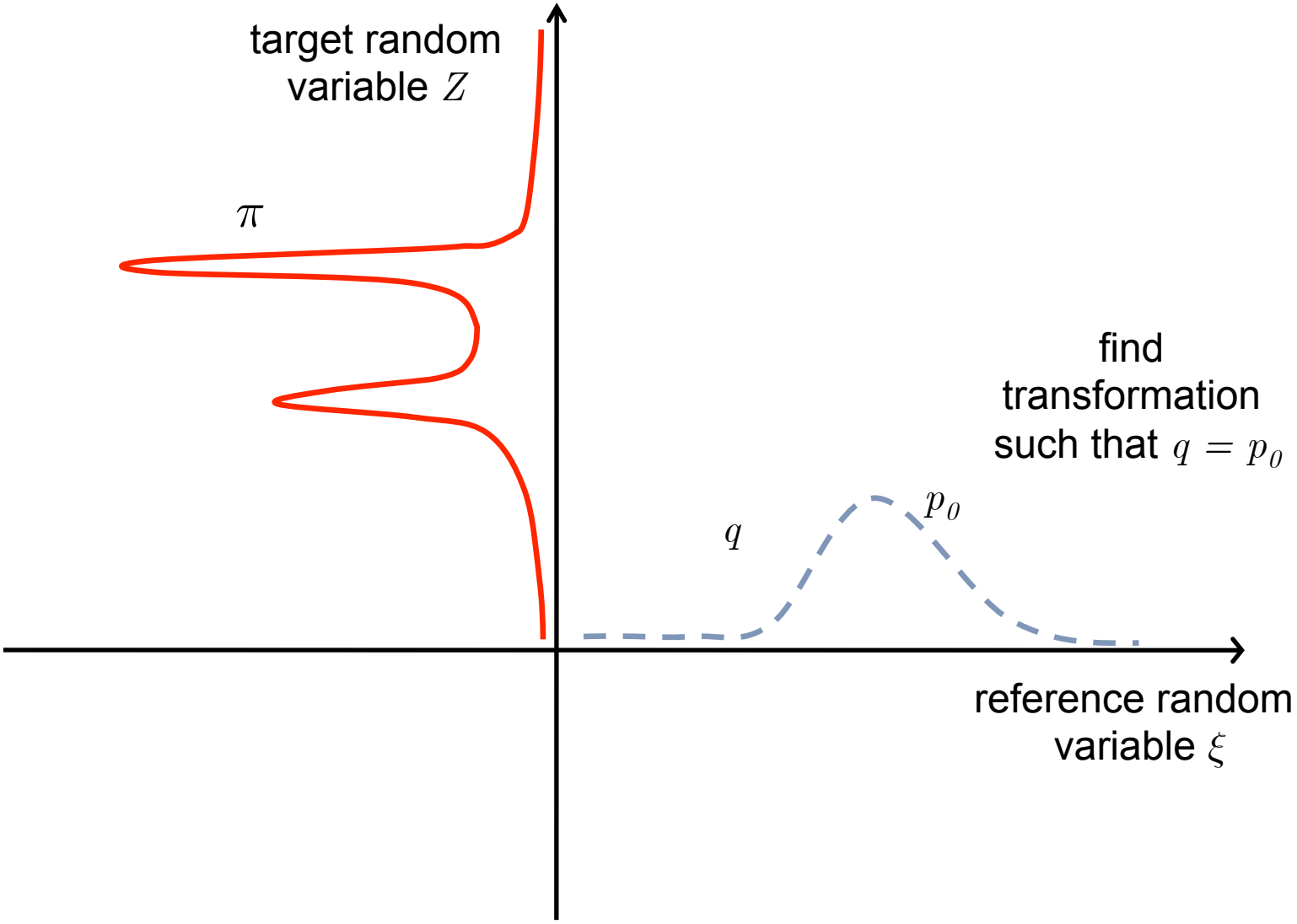
$$\pi^d(z) = \frac{L\left(z;d\right)p\left(z\right)}{\beta}$$

- What if we *knew* a monotone T such that $Z = T(\xi)$?
- Perform a transformation from the target (posterior) to the reference (prior) to get a probability density for ξ

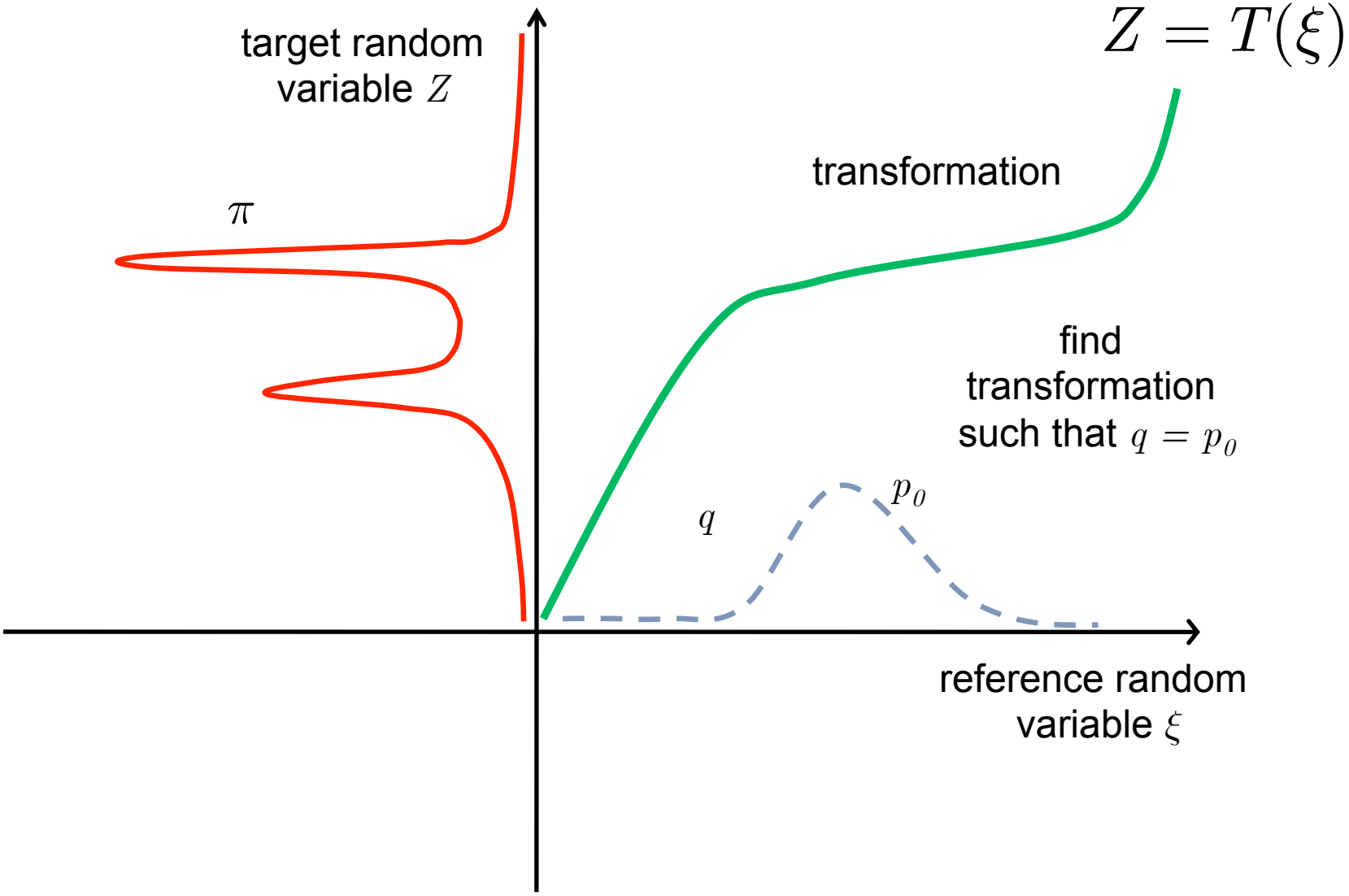
$$\pi^d(z) \quad \xrightarrow{Z = T\left(\xi\right)} \quad q(\xi) = \pi^d\left(T(\xi)\right)\left|\det\frac{\partial T}{\partial \xi}\right|$$

- Compare q to p : suggests a variational approach…

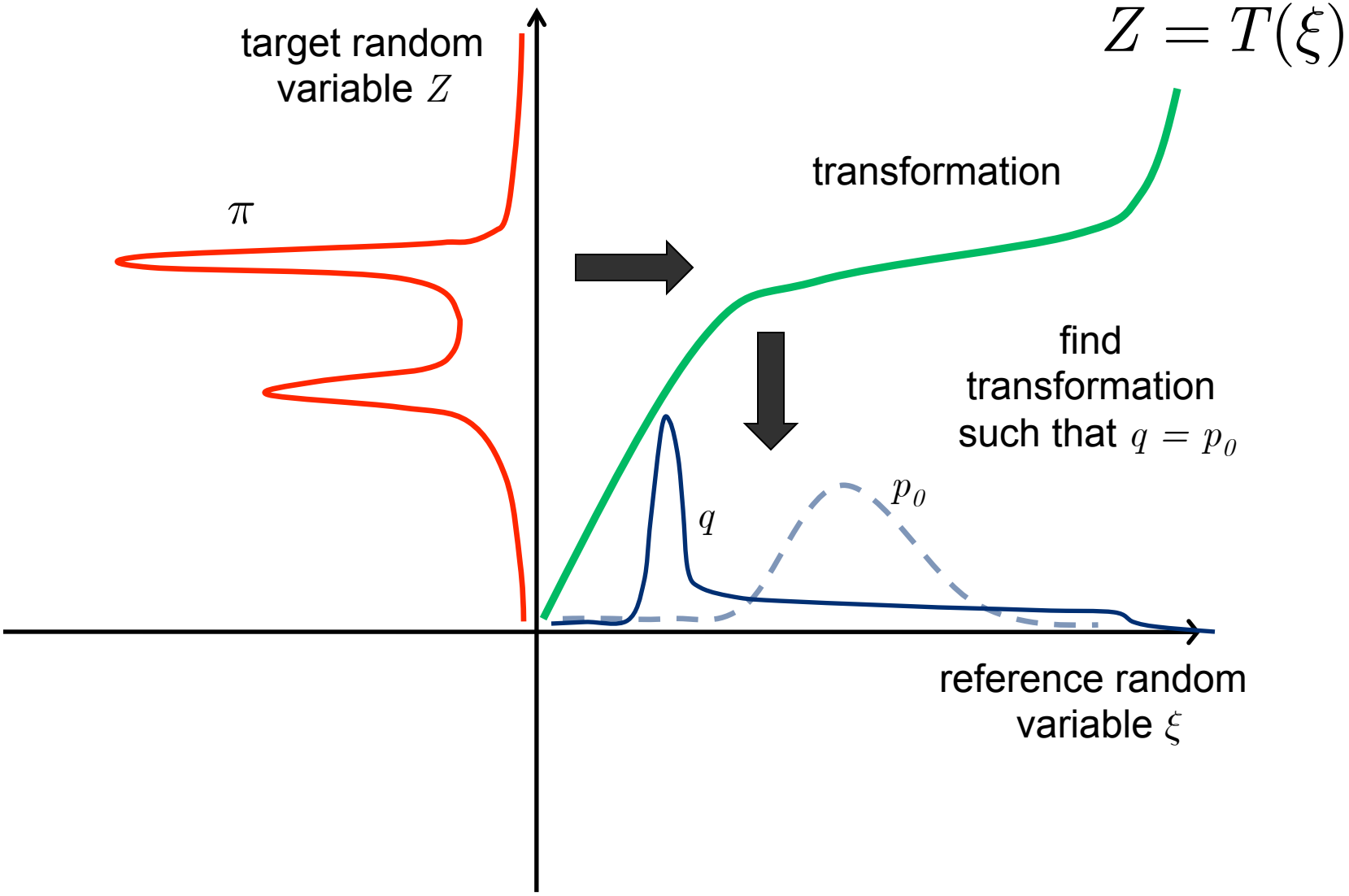
Optimal map schematic



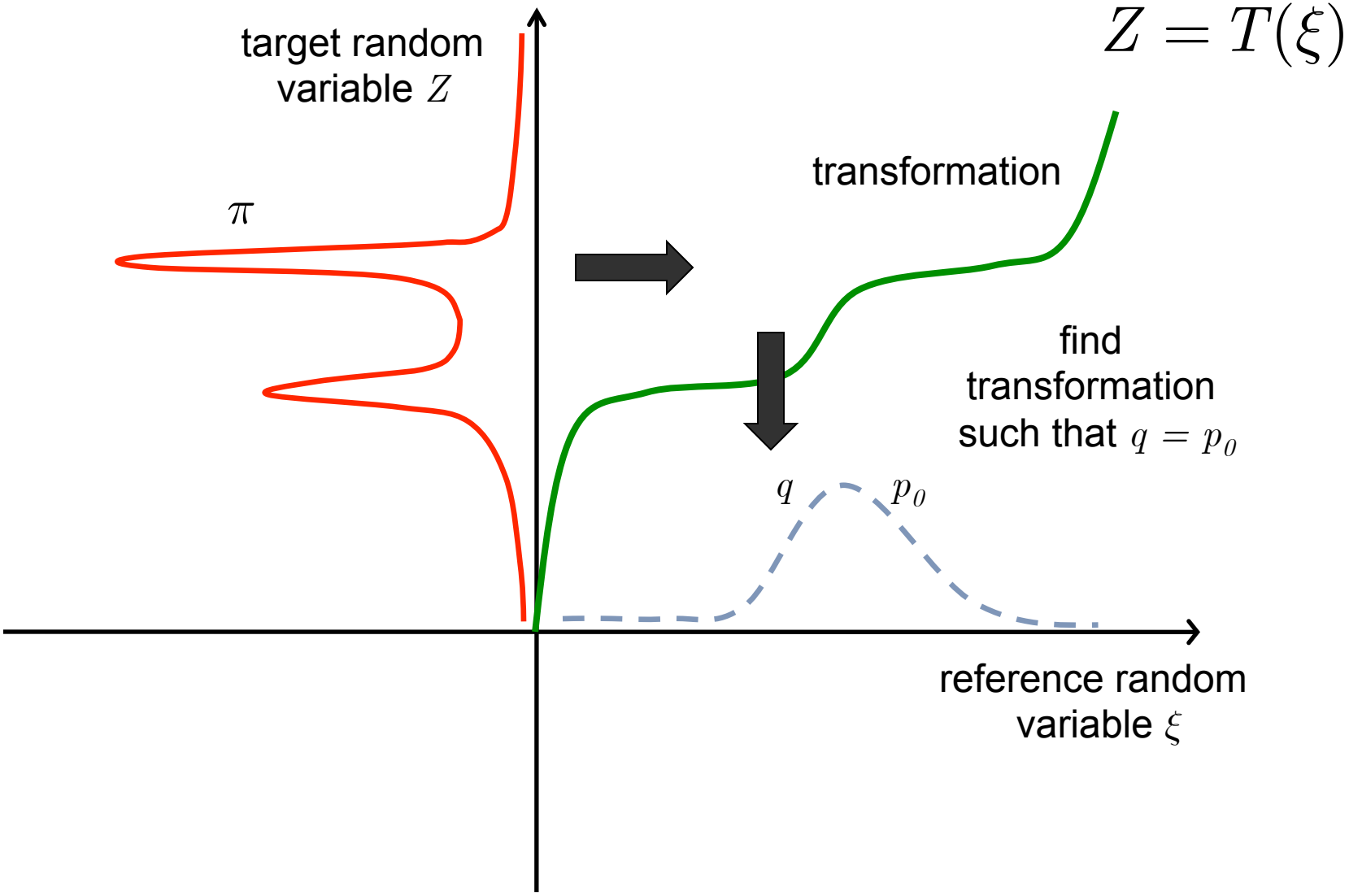
Optimal map schematic



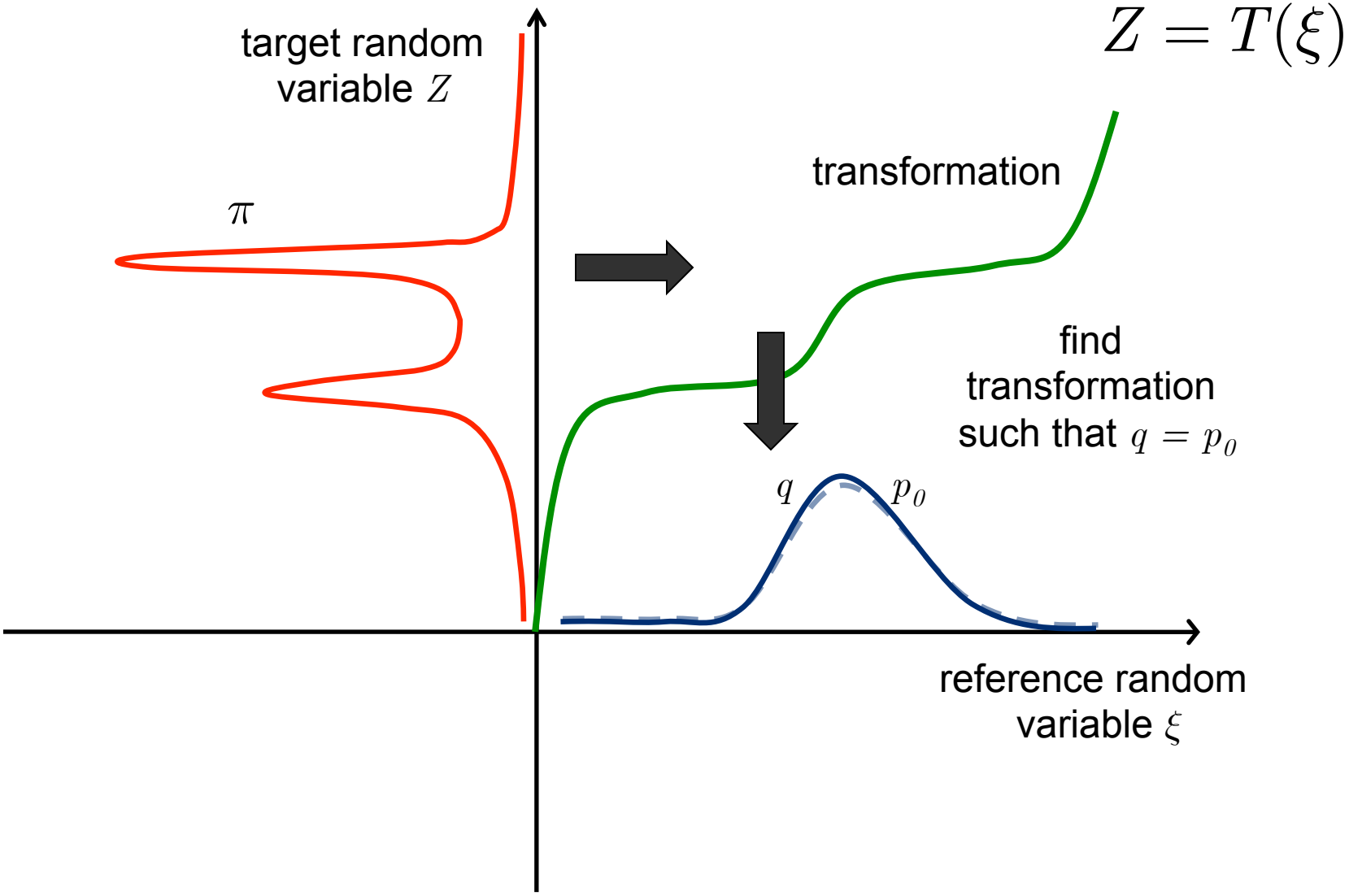
Optimal map schematic



Optimal map schematic



Optimal map schematic



Formulation details

- For simplicity, just consider pointwise equality of densities:

$$p(\xi) = q(\xi) = \pi^d\Big(T(\xi)\Big)\left|\det\frac{\partial T}{\partial \xi}\right| = \frac{L\Big(T(\xi);d\Big)\,p\Big(T(\xi)\Big)}{\beta}\left|\det\frac{\partial T}{\partial \xi}\right|$$

  - Take log of both sides and rearrange:

$$\Phi\Big(x;T\Big) = \log\Big(L\circ T\Big) + \log\Big(p\circ T\Big) + \log\left|\det\frac{\partial T}{\partial x}\right| - \log p = \log\beta$$

  - Hence, find T such that Φ *is* ***constant*** *in* ξ
  - As a byproduct of inference, we obtain the posterior *normalizing constant* β

- Same result can be derived by minimizing Kullback-Leibler divergence or Hellinger distance from p to q

Formulation details

- Alternatively, note that

$$D_{KL}\left(p(\xi)\middle\|q(\xi;T)\right) = \log\beta - \mathbb{E}_{\xi}\left[\Phi(\xi;T)\right]$$

  - Thus we can maximize $\mathbb{E}[\Phi]$ …

Formulation details

- Simpler optimization objective:

$$\max_{T} \mathbb{E}_{\xi}\left[\log\left(\pi \circ T\right)\left|\det \frac{\partial T}{\partial \xi}\right|\right]$$

  with additional structure or penalties (transport costs) to ensure monotonicity of T

- Attributes of the optimization problem:
  - We typically seek a “triangular” structure (Knothe-Rosenblatt map), such that $T_i(\boldsymbol{\xi}) = T_i(\xi_1, \ldots, \xi_i)$
  - Jacobian determinant then becomes easy to compute
  - This is a *stochastic optimization problem;* sample from reference random variable and use sample average approximation (SAA)
  - Gradient-based optimization (e.g., quasi-Newton) for SAA problem
  - Can remove absolute value above; replace with positivity constraints

Optimal maps in context

- Some connections with variational Bayesian methods
- Interesting relationship with ***implicit sampling*** [Chorin & Tu 2009; Chorin, Morzfeld, & Tu 2010; Atkins, Morzfeld, & Chorin 2012]:
  - Implicit sampling maps $\xi^{(i)}$ to $z^{(i)}$ via

$$\log \pi(z) - \text{const} = \log p_0(\xi)$$

  - Contrast with the present Φ equation:

$$\log \pi(z) + \log \left| \det \frac{\partial z}{\partial \xi} \right| - \log \beta = \log p_0(\xi), \quad \text{where } z = T(\xi)$$

  - Implicit sampling omits Jacobian; yields *weighted* samples
  - Optimal maps seek a *global transformation,* while implicit sampling proceeds sample-by-sample

Map representation

- We represent *T* using an orthogonal polynomial expansion (e.g., Hermite chaos)
  - Orthogonality is with respect to the *reference distribution*
  - Why? Analytical expressions for posterior moments in terms of *F*

$$T\left(\xi\right) = F^T \Psi\left(\xi\right)$$

← vector of orthogonal polynomials

↗ matrix of unknown coefficients

- Typical approach:
  - First solve for best linear map
  - *Enrich polynomial space* by iterating over degree; monitor $\mathrm{Var}[\Phi]$ (proportional to $\mathrm{D}_{\mathrm{KL}}(p||q)$) to assess convergence

Properties of the map

- **Potential advantages** (relative to MCMC):
  - Generate *arbitrary* numbers of *independent* posterior samples, without additional forward solves
  - Analytical expressions for posterior moments
  - Clear convergence criterion
  - Key steps are embarrassingly parallel
  - Can propagate posterior distribution through subsequent models (polynomial chaos expansion)
  - Compute posterior normalizing constant (*marginal likelihood*) for use in Bayesian model selection

- **Key questions:** For which target distributions will the map have simple/computable structure? How many degrees of freedom are needed?

Simple linear example

- 100 dimensional problem
- *A* is randomly generated

$$y(x) = Ax, \quad d = y + \epsilon$$

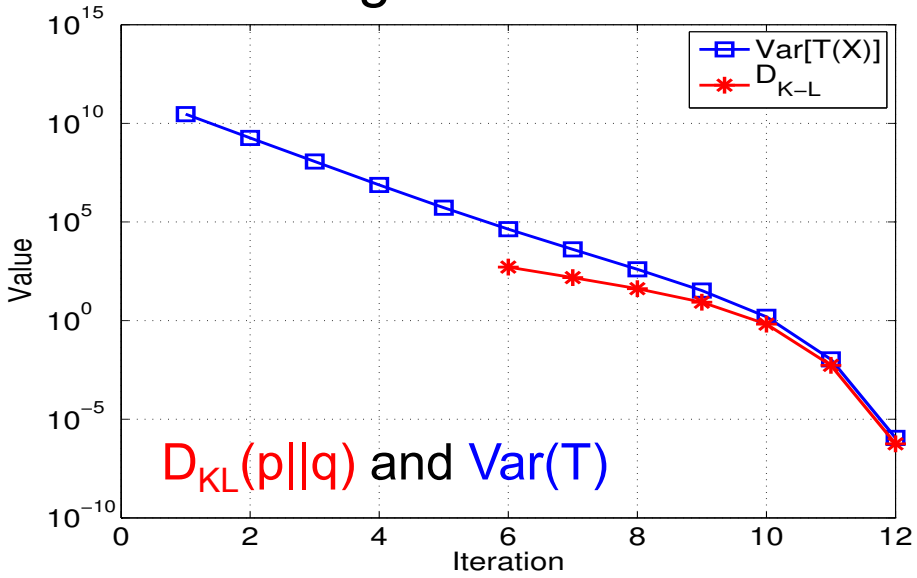
$$X \sim N(0, I), \quad \epsilon \sim N(0, \Sigma_n)$$

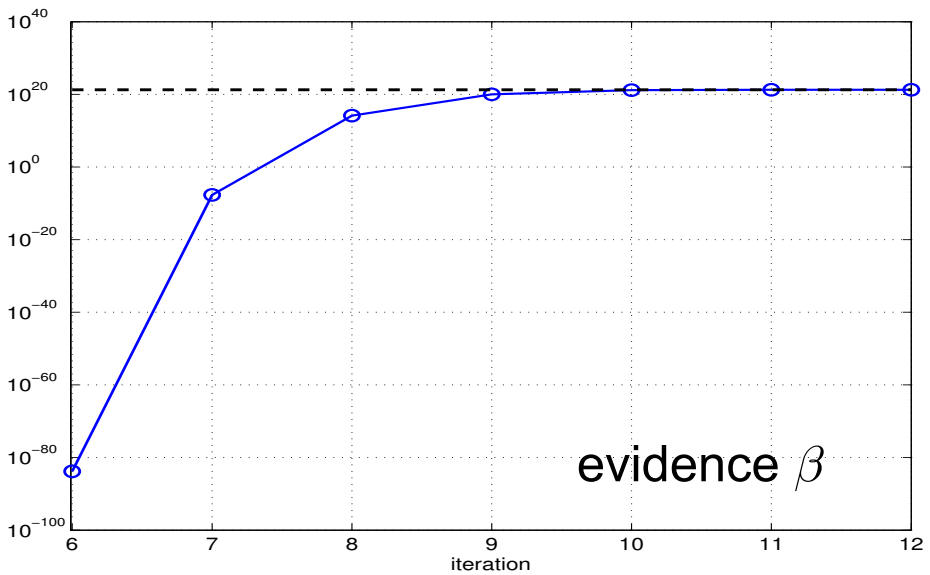
- Gaussian posterior:

$$Z \sim N\left(\mu_z, \Sigma_z\right)$$

$$A \in \mathbb{R}^{8\times 100} \quad y, d \in \mathbb{R}^8$$

- Start iterations from identity map $T(\xi) = \xi$
- Convergence to exact solution in 12 iterations



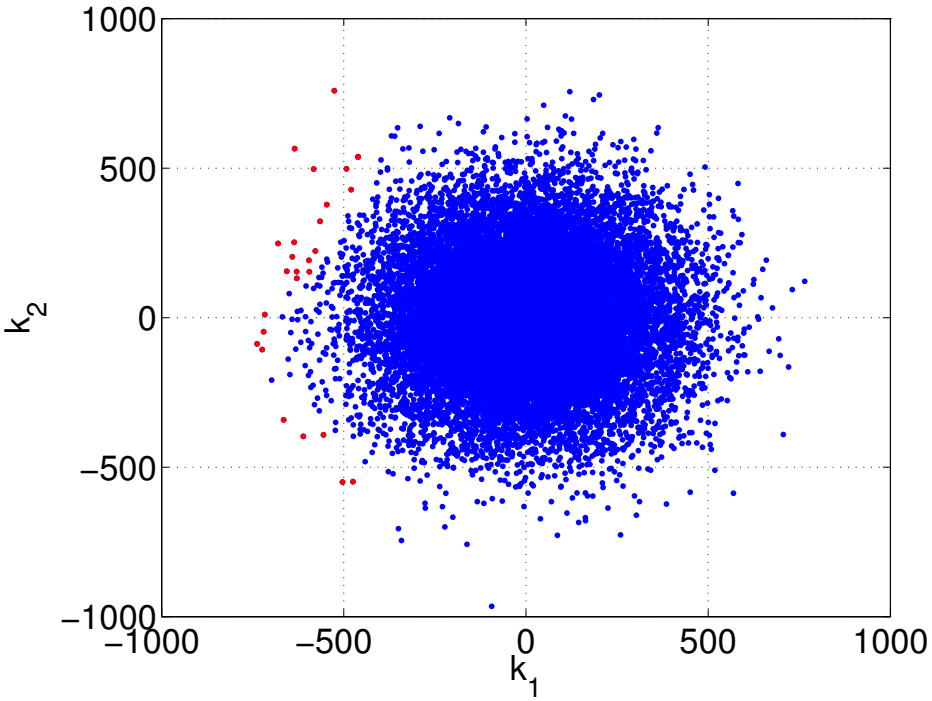


Reaction kinetics

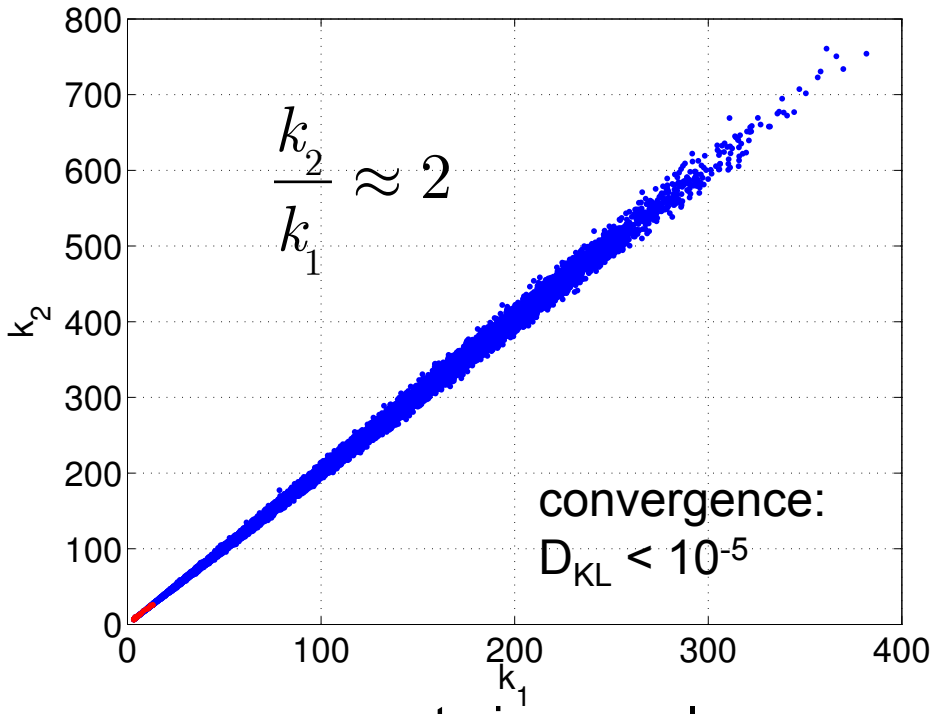
- Five late-time observations of *A*; truth is $k_1 = 1$, $k_2 = 2$
- Gaussian prior
- Infer k_1 and k_2

$$\frac{dA}{dt} = -k_1 A + k_2 B$$

$$\frac{dB}{dt} = k_1 A - k_2 B$$



prior samples

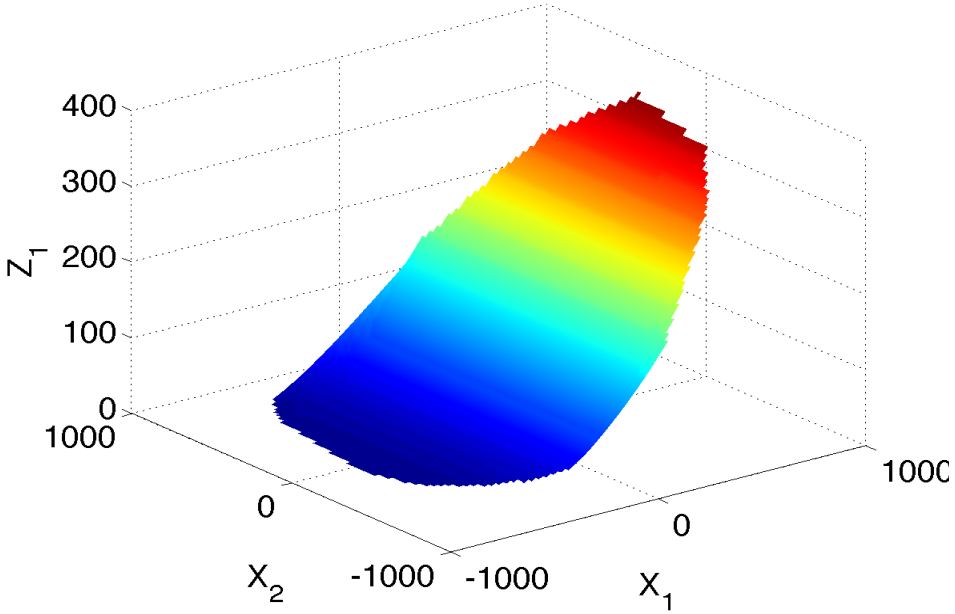


posterior samples

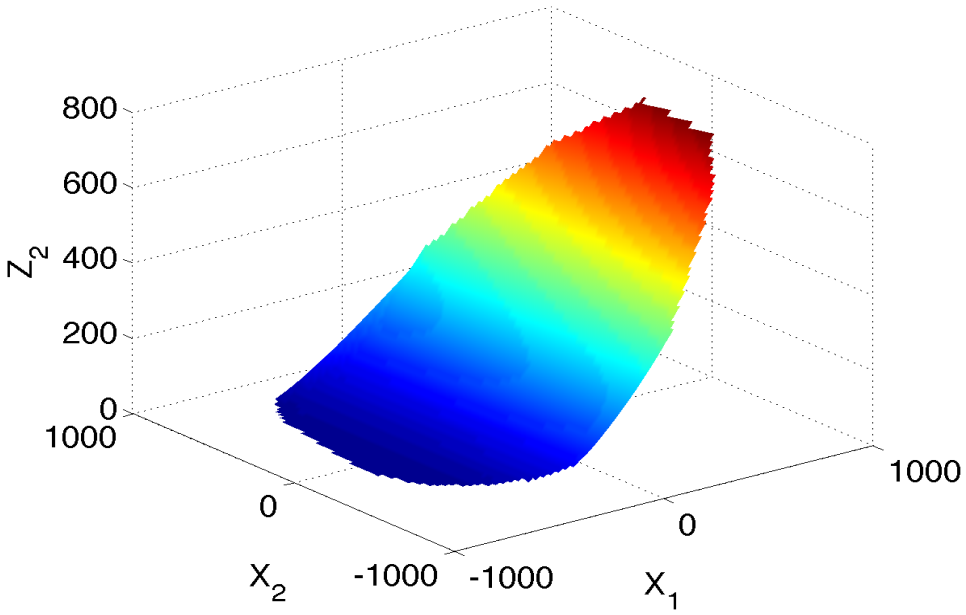
Reaction kinetics: map

- 7^{th} order polynomial map

First component of posterior



Second component of posterior

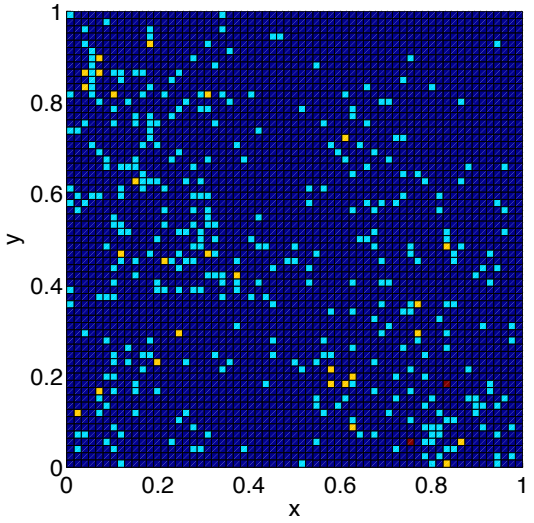


Log-Gaussian Cox point process

- From Christensen 2005, Girolami 2011
- Would like to infer latent intensity field z

$$z \sim N\left(\mu, C(r, r')\right)$$

$$C\left(r, r'\right) = \sigma^2 \exp\left(\left|r - r'\right| / 64\ell\right)$$



64x64 grid
4096 dimensions

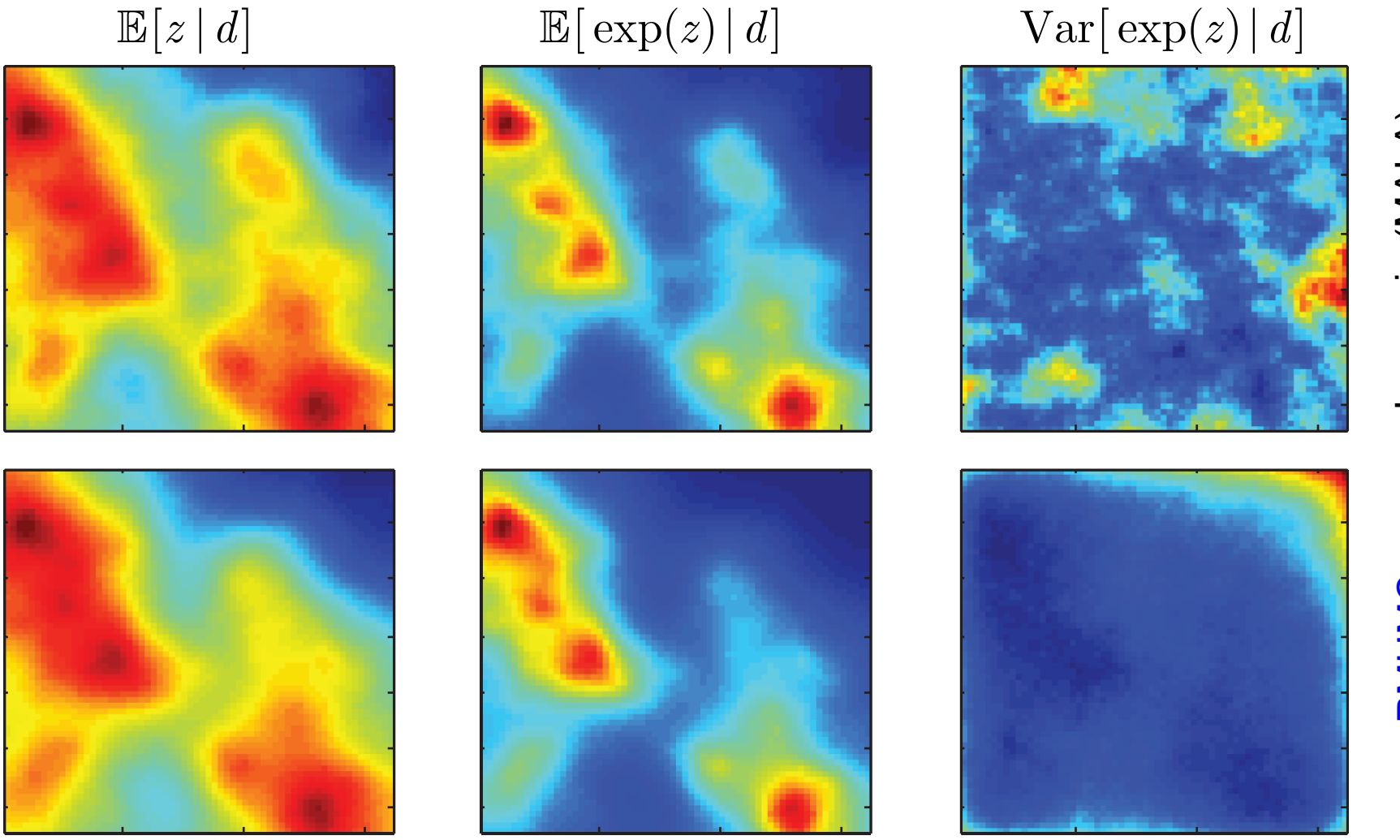
- Observe number of counts d *per* grid cell
  - Poisson distributed with mean $m \exp\left(z\right)$

- Posterior density:

$$p\left(z\middle|d\right) \propto \exp\left(d^T z - m\mathbf{1}^T \exp(z)\right) \exp\left(-\frac{1}{2}\left(z-\mu\right)^T \Sigma^{-1}\left(z-\mu\right)\right)$$

Log-Gaussian Cox point process

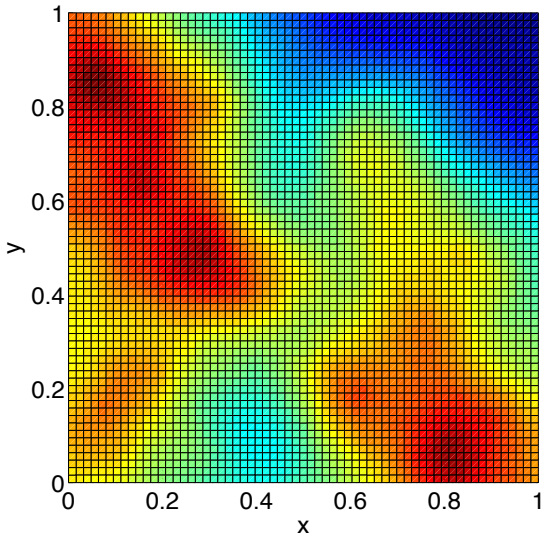
- This is a challenging problem for most MCMC samplers!



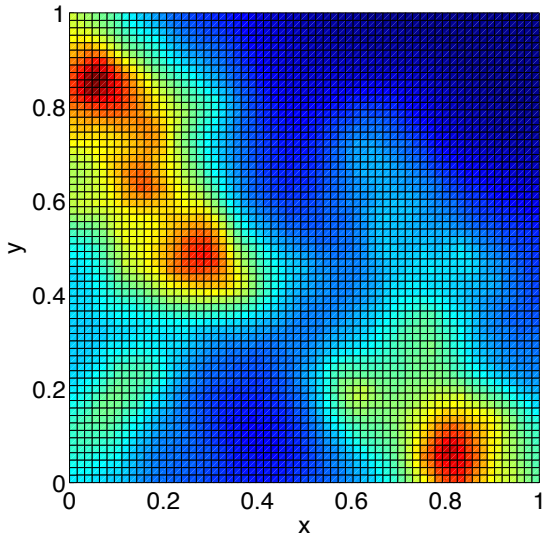
[Girolami & Calderhead, JRSSB 2011]

Log-Gaussian Cox process

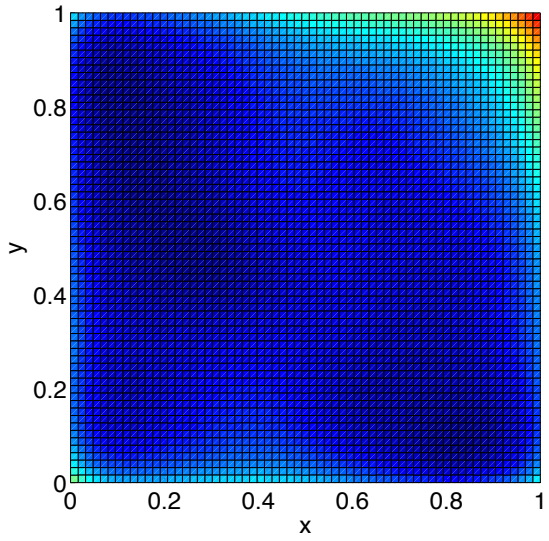
$\mathbb{E}[z \mid d]$



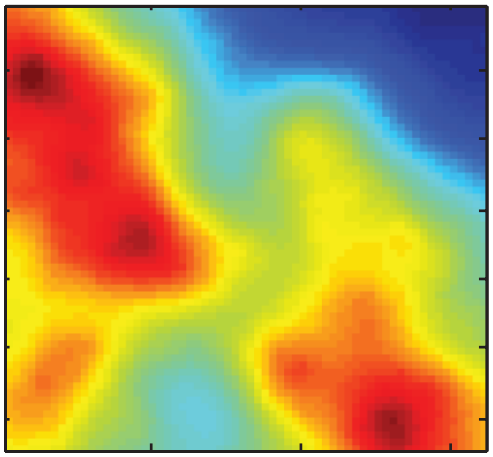
$\mathbb{E}[\exp(z) \mid d]$

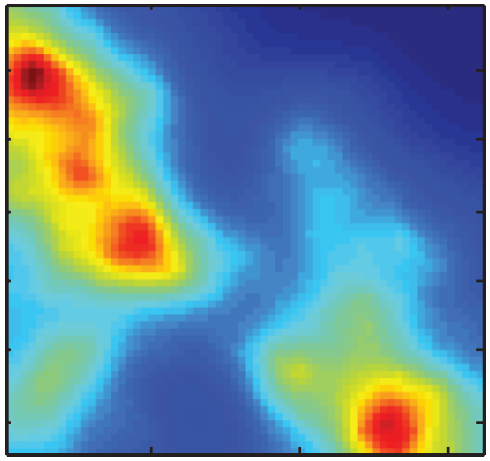


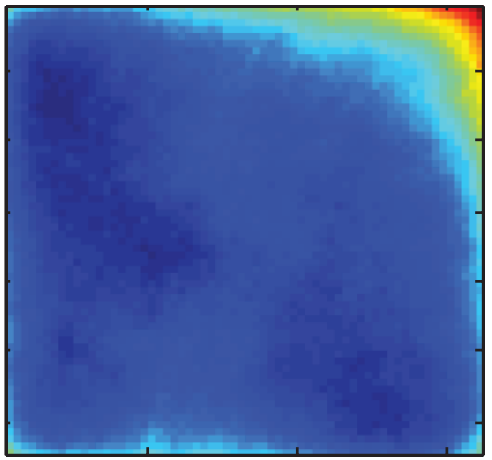
$\mathrm{Var}[\exp(z) \mid d]$



map







RMHMC

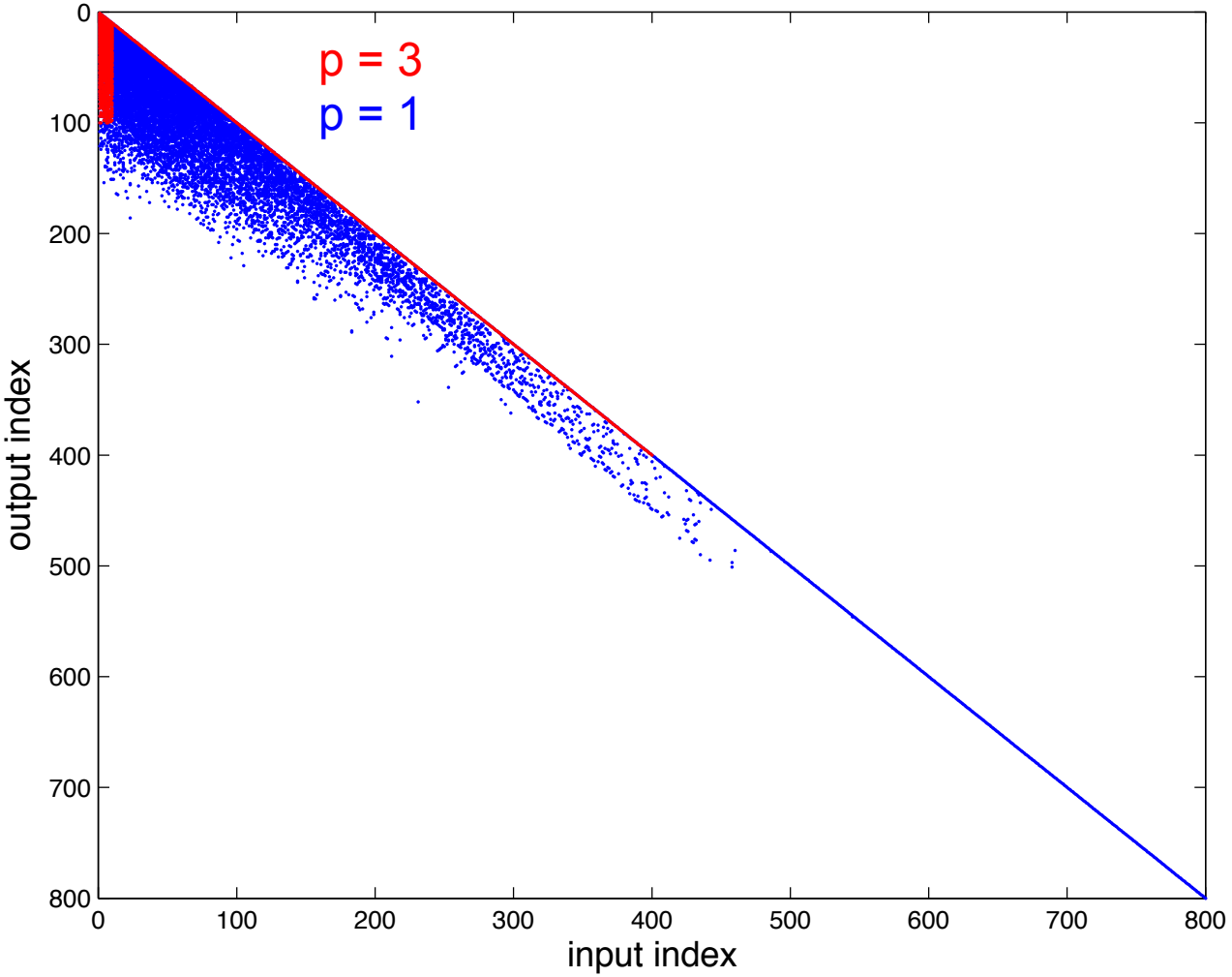
Log-Gaussian Cox process

- Comparison with Riemannian manifold MCMC
  - Triangular map; polynomial transformation in Karhunen-Loève coordinates

method	time	ESS (min, med, max)	s/min ESS	relative speed
MALA [GC2011]	31577	(3, 8, 50)	10605	1
MMALA [GC2011]	634	(26, 84, 174)	24.1	440
RMHMC [GC2011]	2936	(1951, 4545, 5000)	1.5	7070
Map (4096 dims)	1000	*5000*	*0.2*	*53000*

High-dimensional maps

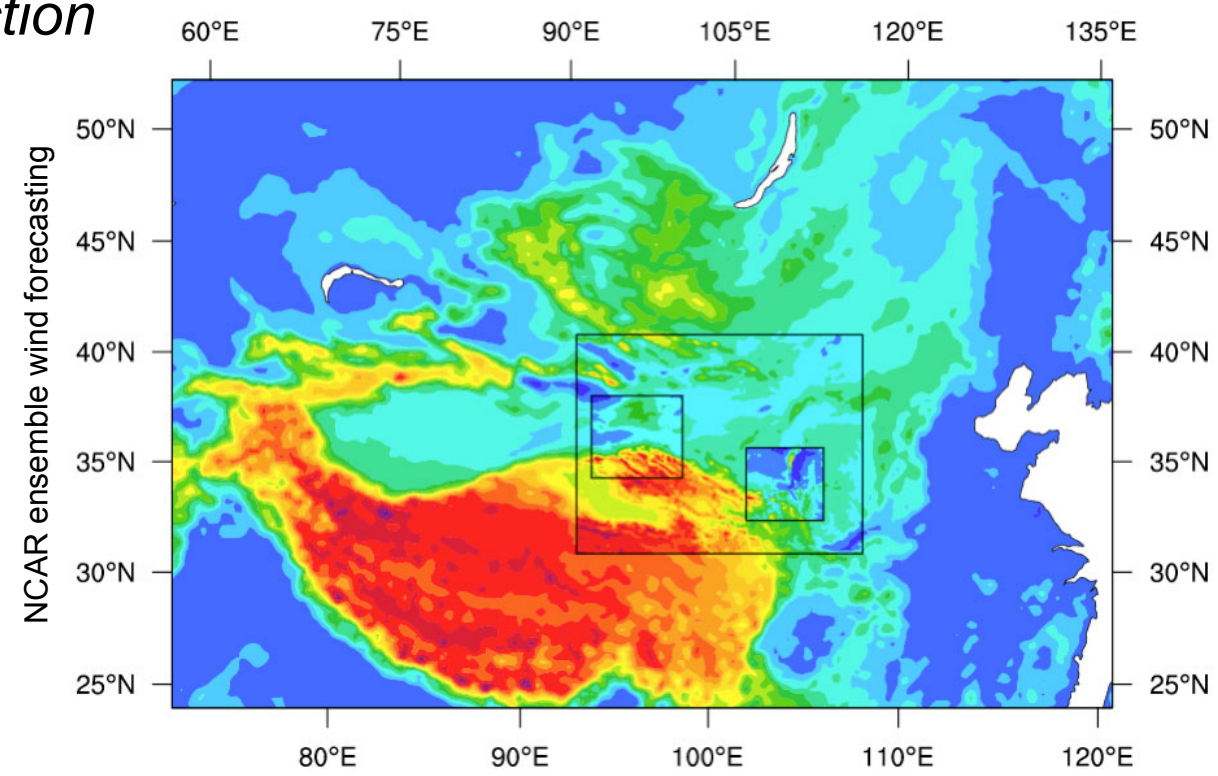
- Why are maps computable (here, even in 4096 dimensions)?



- **Polynomial terms** in final map, log-Gaussian Cox process
  - Consider coefficients above some threshold, here $\varepsilon = 10^{-2}$
  - What degree are the corresponding terms?

From static to *dynamic* problems

- ***Example:*** *atmospheric data assimilation, numerical weather prediction*



⇒ Learn the **state** given noisy observations; quantify **uncertainty** in the state estimate

⇒ An essential step in **predictive** simulation

Bayesian approach

- **Dynamical model** $f(x_t \mid x_{t-1})$
  - Discretization of an SDE; deterministic model plus noise
  - Dynamics typically nonlinear and stochastic
- **Observation operator** $g(y_{t^*} \mid x_{t^*})$
  - Observations y sparse in space and time
- **Posterior distribution (filtering)**

$$p\left(x_t \middle| y_{1:t}\right) \propto g(y_t \mid x_t)\, p(x_t \mid y_{1:t-1})$$

$$= g(y_t \mid x_t) \int f(x_t \mid x_{t-1})\, p(x_{t-1} \mid y_{1:t-1})\, dx_{t-1}$$

  - We wish to *estimate* or *reconstruct* the current state of the system
  - We wish to *quantify uncertainty* in this estimate

Filtering algorithms

- How to characterize the posterior distribution $p(x_t \mid y_{1:t})$ in a recursive fashion?

- **Weighted particle approaches** (e.g., particle filtering, SMC)
  - Represent the posterior as a weighted sum of Dirac measures

$$p(x_t \mid y_{1:t}) \approx \sum_i^N w_i \delta_{X_t^i}(x_t)$$

  - Good proposal distributions are crucial; particles must fall in the regions of high posterior probability!
  - Extensive work in this regard: Doucet 2001, Snyder 2008, VanLeeuwen 2010, Chorin 2009–12
  - Converge to the Bayesian solution as $N \to \infty$

Filtering algorithms

- How to characterize the posterior distribution $p(x_t \mid y_{1:t})$ in a recursive fashion?

- **Ensemble Kalman filtering (EnKF)**
  - Approximate the prior (i.e., forecast distribution) as Gaussian
  - Bayesian update proceeds from the Gaussian assumption
  - **Biased;** does not converge to the true Bayesian solution
  - Can have good tracking performance and good computational efficiency; regularly applied to high-dimensional problems
  - Yet performs poorly in quantifying uncertainty (e.g., reproducing covariance of the Bayesian posterior) [Law & Stuart 2012]

Filtering algorithms

- Two rather contrasting schemes:
  - Continuous versus discrete: sum of a few point masses (PF) versus Gaussian approximation (EnKF)
  - Inference via “local” versus “global” updates

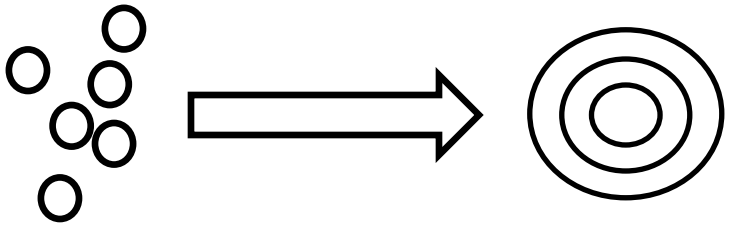
- **We will propose filtering algorithms that use optimal transport maps**
  - Use maps to tackle intrinsic challenges: *approximating the prior*, *moving from prior to posterior,* even *smoothing*

- **Key attributes:**
  - Continuous representation avoids weighted particles (and particle degeneracy)
  - Yet converge to the true Bayesian solution

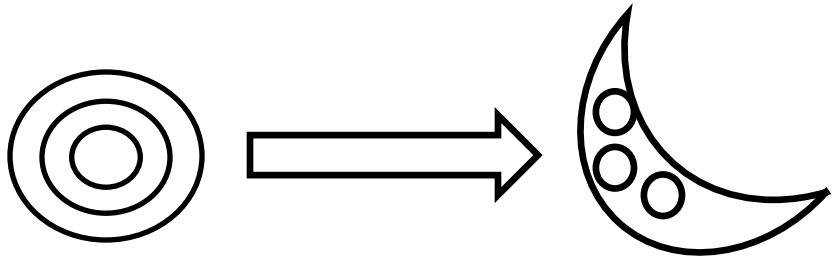
Optimal maps in filtering

- Use transport maps in two different ways

1. **From samples to a map**
   - Given an arbitrary collection of samples $\xi^{(i)}$
   - Prescribe a simple target measure (Gaussian) for Z
   - Find $Z = T(\xi)$, then invert it!
   - This is regularized *density estimation* with a map



2. **From a probability density to samples**
   - Given a complex target measure π
   - Choose a simple reference measure (e.g., Gaussian) for ξ
   - Construct a map from ξ to target
   - Use the map to generate independent samples



Filtering via maps

- Two algorithms: one that separates forecast/analysis, one that tackles them jointly

- **"Double map" algorithm**

$$p\left(x_t \middle| y_{1:t}\right) \propto g(y_t \mid x_t)\, p(x_t \mid y_{1:t-1})$$

$$= g(y_t \mid x_t) \int f(x_t \mid x_{t-1})\, p(x_{t-1} \mid y_{1:t-1})\, dx_{t-1}$$

Step 1: construct a "forecast map" $X_t \mid y_{1:t-1} = T_f(\xi)$

  - Generate samples of $X_t \mid y_{1:t-1}$ and map them to a target Gaussian
  - The *inverse* of this map is T_f
  - Now the inference problem can be transformed to ξ-space…

Filtering via maps

- *Step 2:* construct an "inference map"
  - Write the posterior density in terms of ξ

$$\pi_{\xi}^{t} \equiv p\left(\xi \middle| y_{1:t}\right) \propto p\left(y_t \middle| T_f(\xi)\right)\, p_{\xi}(\xi)$$

  - Complex prior and simple likelihood now transformed to a simple prior and complex likelihood: the **target density**
  - Construct a map T_a from another (Gaussian) reference to this posterior

$$\begin{array}{ccc} \tilde{\xi}_t \sim \pi_{\xi}^{t} & \xleftarrow{\quad T_a \quad} & \xi \sim p_{\xi} \\ \Big\downarrow T_f & & \Big\downarrow T_f \\ X_t \mid y_{1:t} & & X_t \mid y_{1:t-1} \end{array}$$

  - **Final map** is $T = T_f \circ T_a$

Double-map algorithm

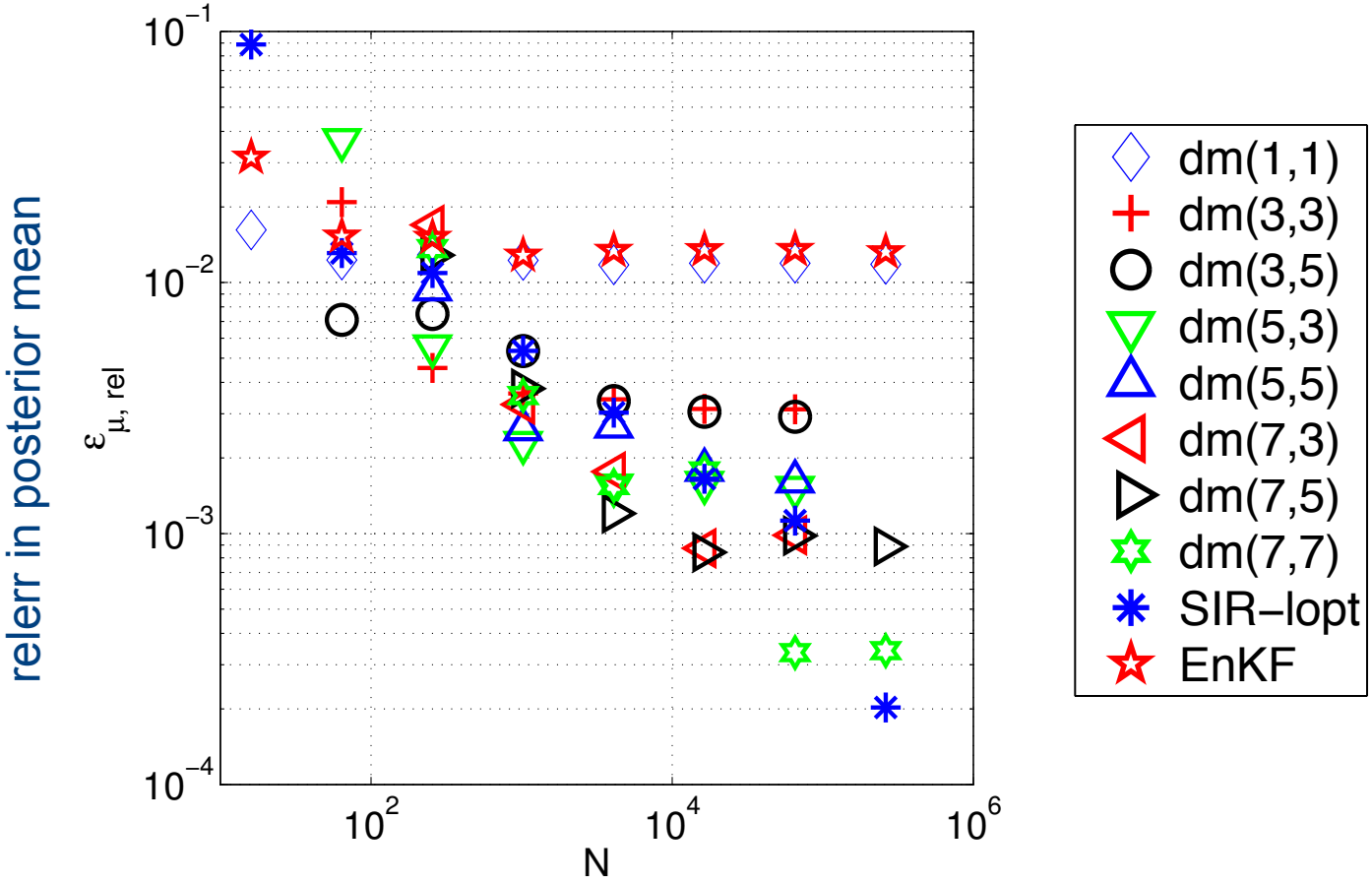
- Comments on the algorithm:
  - **When both maps are linear, algorithm reduces to the EnKF!**
  - Increasing the polynomial order generalizes the EnKF to non-Gaussian distributions
  - Use this freedom to balance computational cost with accuracy
  - Infrequent observations: intermediate-time states are *marginalized* away
  - Forecast distribution is estimated from samples, but the estimate (via map T_f) is continuous; approximate information is "filled in" over the entire space
  - Posterior may concentrate where there are no forecast samples

Numerical examples

- **Example 1: Lorenz-63**
  - Observe every 20 steps (x and z), timestep Δt = 0.01; Gaussian initial condition, stochastic forcing
  - Compare sampling-importance-resampling (SIR) with locally optimal proposal, EnKF, and double map; versus a “gold-standard” SIR posterior
    - Note: we are comparing to the true Bayesian solution, not evaluating tracking error

Numerical examples

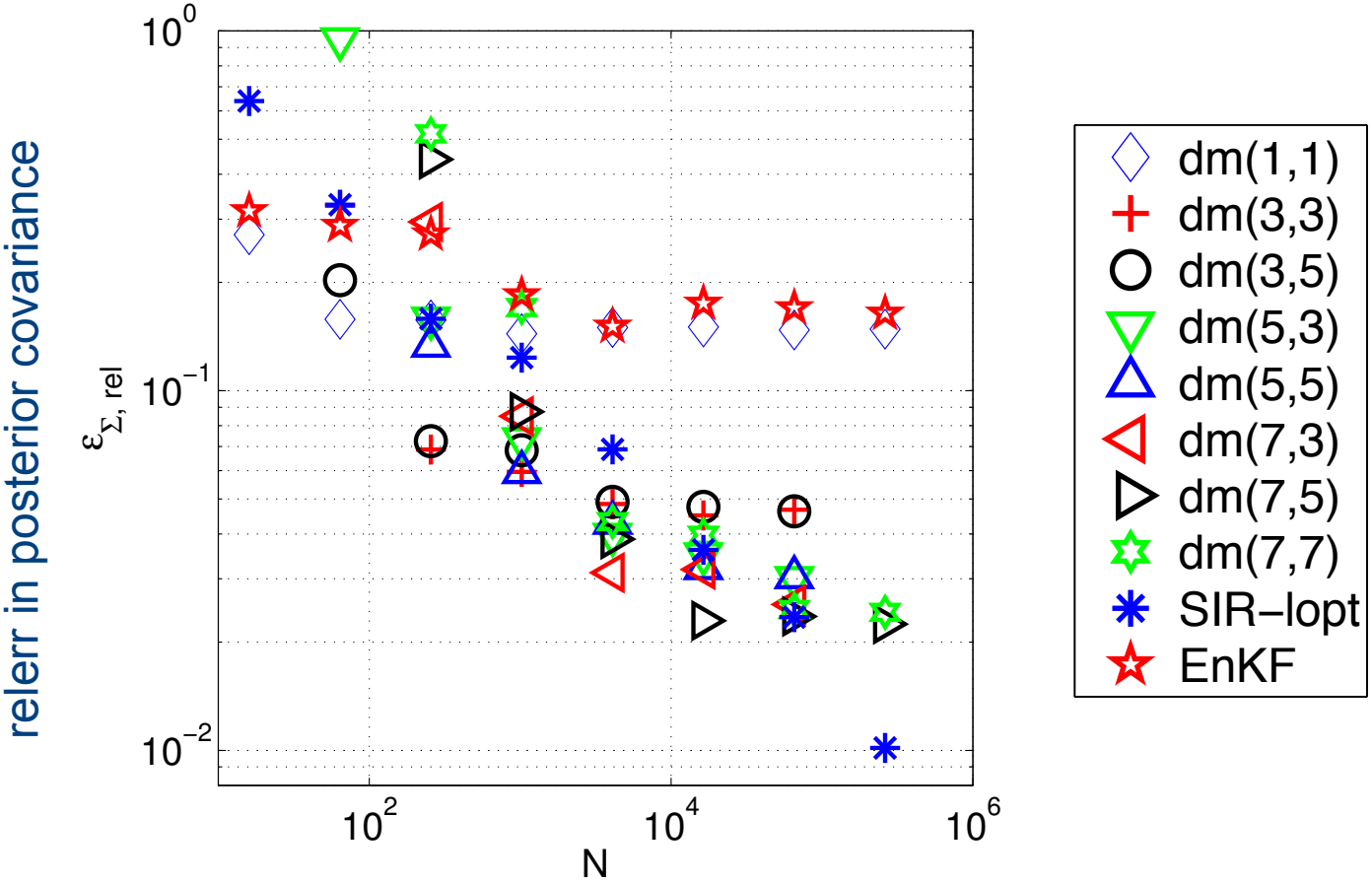
- **Example 1: Lorenz-63** (errors over 1200 timesteps)



- EnKF does not converge as $N \to \infty$
- Maps approach the Bayesian solution as polynomial degree is refined

Numerical examples

- **Example 1: Lorenz-63** (errors over 1200 timesteps)



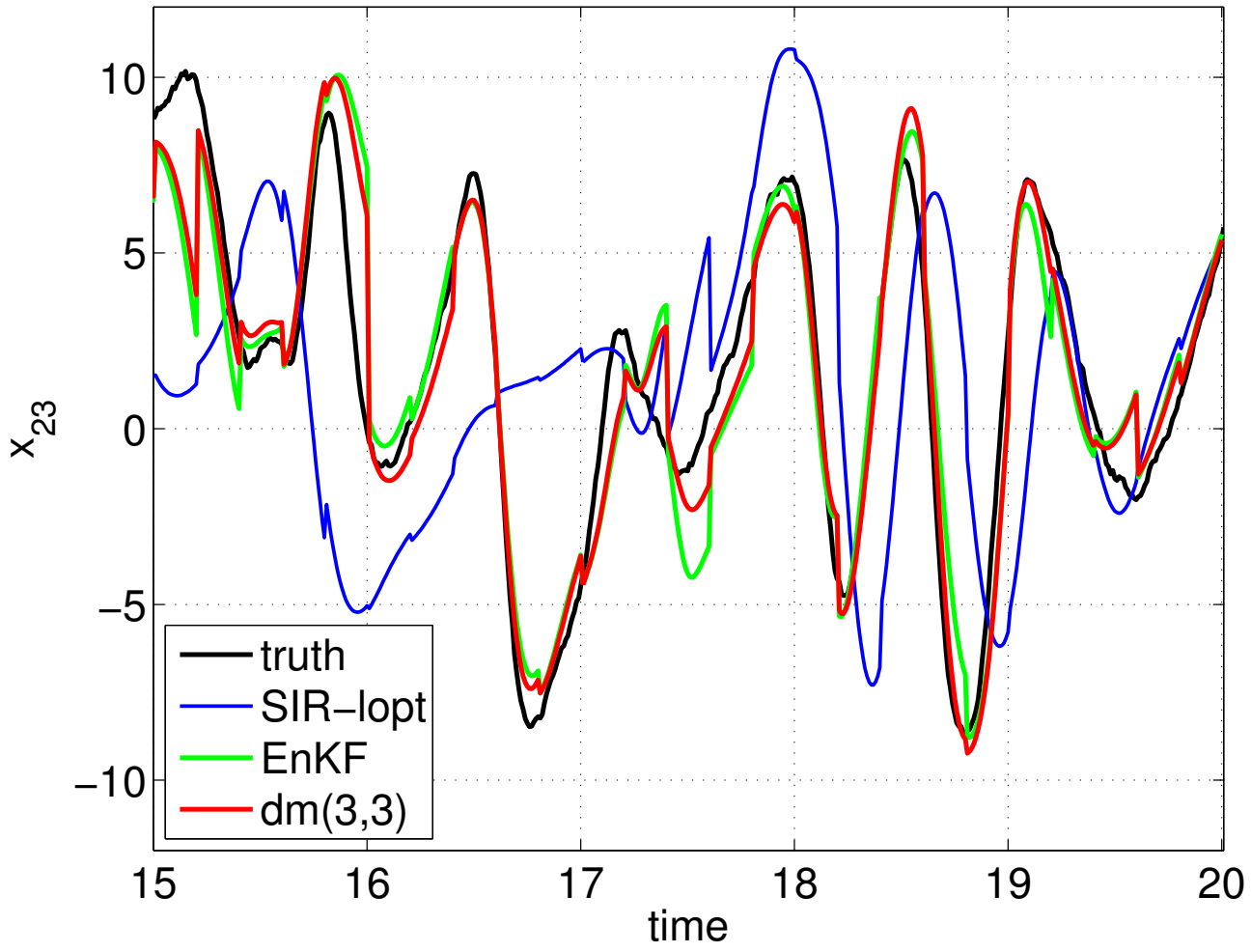
- EnKF does not converge as $N \to \infty$
- Maps approach the Bayesian solution as polynomial degree is refined

Numerical examples

- **Example 2: Lorenz 95**
  - 40-dimensional problem, $F=8$ (strongly chaotic), stochastic forcing
  - Observe every 20 timesteps, odd states only; timestep Δt = 0.01
  - Considered a challenging test configuration for filtering (Bengtsson 2003, van Leeuwen 2010, Lei & Bickel 2011)
  - Compare EnKF, double map, and SIR-lopt

Numerical examples

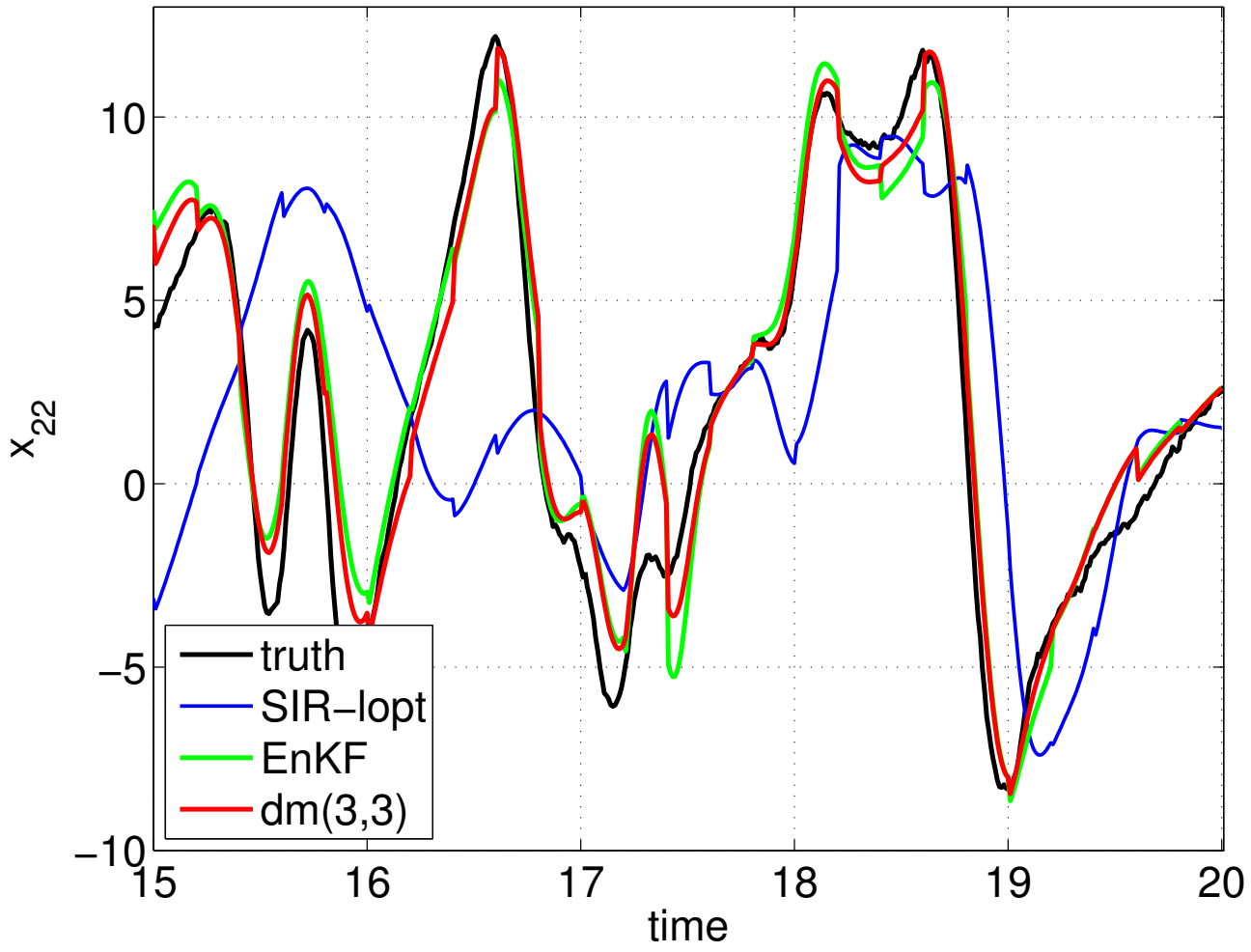
- **Example 2: Lorenz 95** (state tracking performance, 256 particles)



observed state x_{23}

Numerical examples

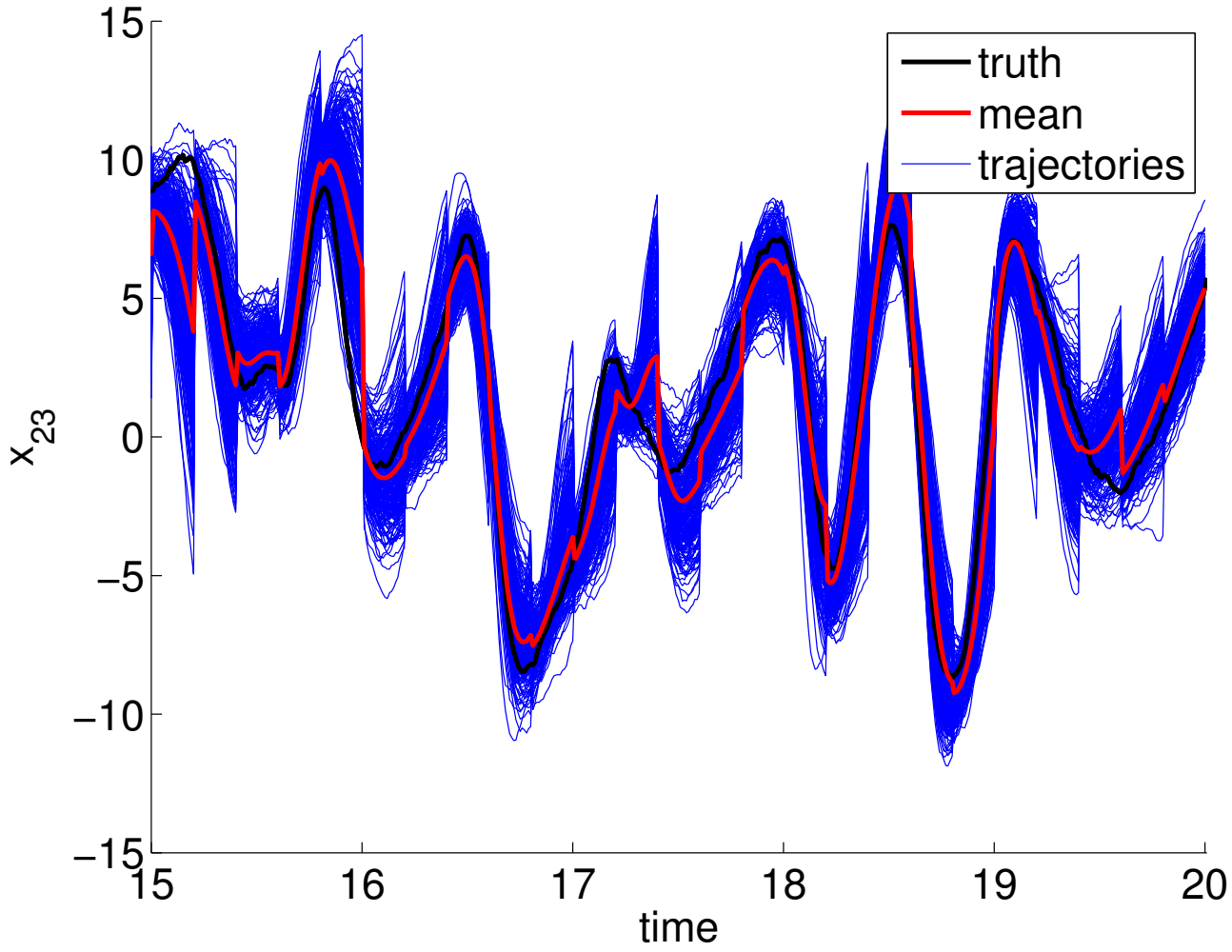
- **Example 2: Lorenz 95** (state tracking performance, 256 particles)



unobserved state x_{22}

Numerical examples

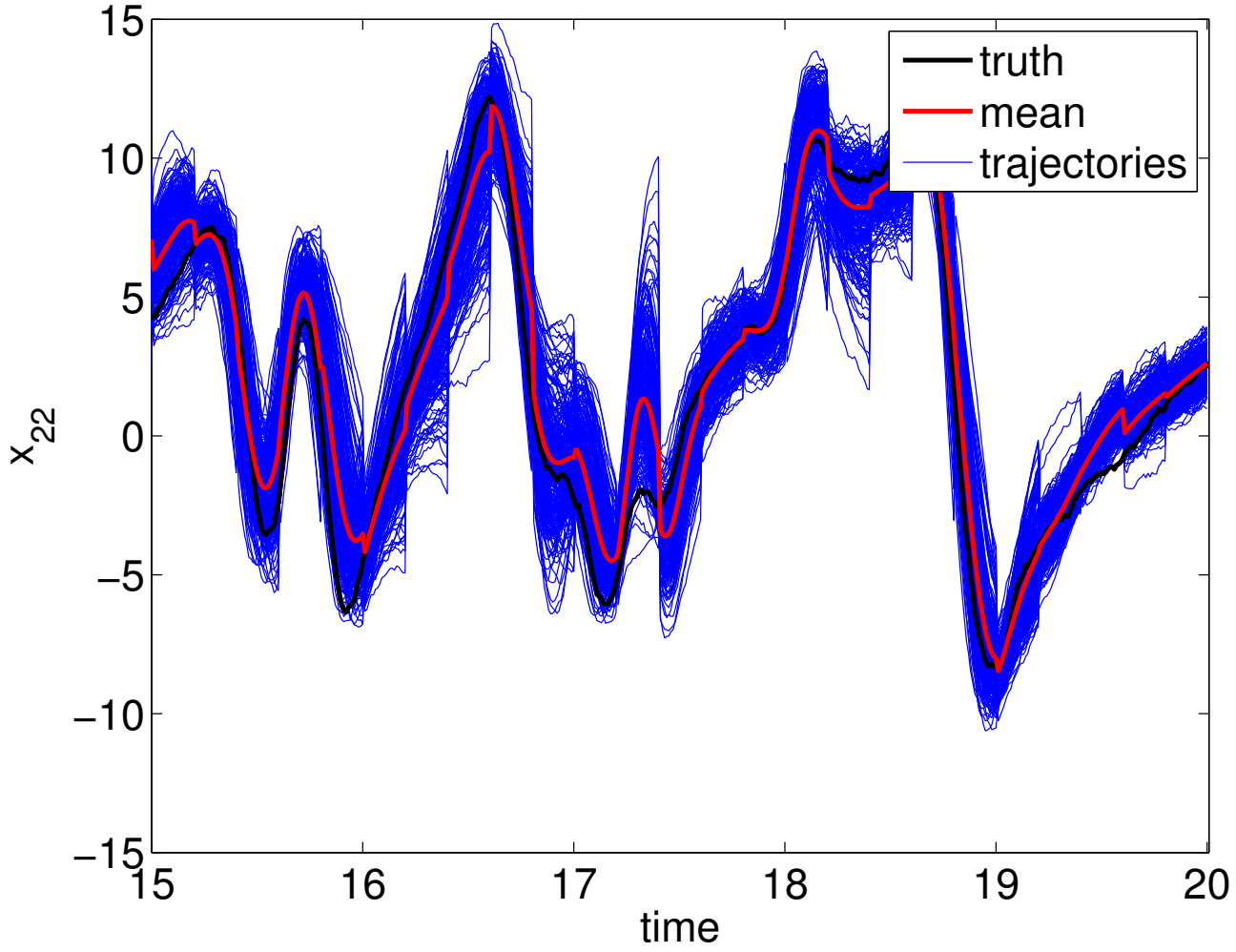
- **Example 2: Lorenz 95** (state tracking performance, 256 particles)



observed state x_{23}

Numerical examples

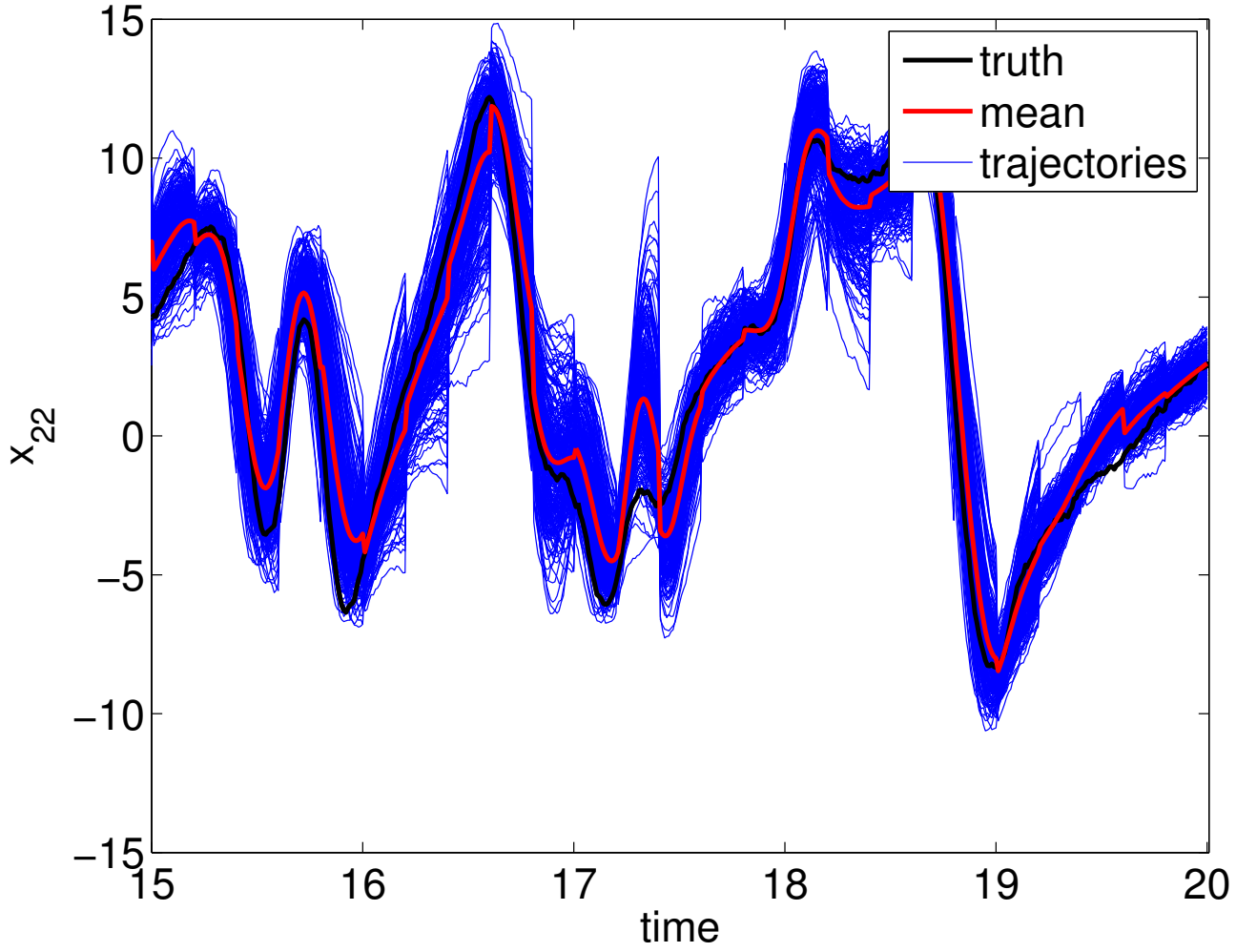
- **Example 2: Lorenz 95** (state tracking performance, 256 particles)



unobserved state x_{22}

Numerical examples

- **Example 2: Lorenz 95** (state tracking performance, 256 particles)



Relative tracking errors (RMSE, 100 twin experiments)

- EnKF: *26.5%*
- dm(3,3): *25.0%*
- SIR-lopt: *112.5%*

Numerical examples

- **Example 2: Lorenz 95**
  - What about convergence to the true posterior? Are we obtaining meaningful *uncertainty* information in the state?
  - Must restrict attention to a single assimilation step; reference Bayesian solution obtained via exhaustive importance sampling
  - Relative errors (percent) in mean and covariance (Frobenius norm), 4096 particles

	SIR-lopt	EnKF	dm (3,3)	dm (5,5)	dm (7,7)
mean	12.8	**2.8**	2.7	2.1	**1.7**
covar	7215	**31**	22	20	**18**

Numerical examples

- **Example 2: Lorenz 95**
  - Computational cost (wallclock time for 10 assimilation steps) and map degrees of freedom (DOF); 512 particles
  - Here, model integration is cheap; cost is dominated by optimization!
  - Yet optimization iterations do not require further evaluation of the dynamical model

	EnKF	dm(1,1)	dm(3,3)	dm(5,5)	dm(7,7)
time [s]	1	7	215	1950	4945
map DOFs	–	194	1329	3635	5197

From double map to joint map

- The Lorenz 95 problem is challenging, but the the observations are in a sense *typical.* Simple approximations of the forecast distribution can succeed. Marginalization supports computational efficiency.

- What about *rare events?* Now one needs to capture the tail of the prior distribution. (The prior and likelihood are *mutually* singular!)
- Directly approximating the tail of the prior is completely impractical; instead we must abandon marginalization and turn to an **all-at-once** approach

Joint map

- **Algorithm 2: Joint map**
  - *Idea:* map to the joint distribution of state trajectories between observations; no forecast/analysis subdivision
  - More robust (no prior approximation), but more expensive!
  - *Key steps:*
    - Rewrite joint distribution

      $$p(x_t, x_{t-1} \mid y_{1:t}) \propto p(y_t \mid x_t) p(x_t \mid x_{t-1}) p(x_{t-1} \mid y_{1:t-1})$$

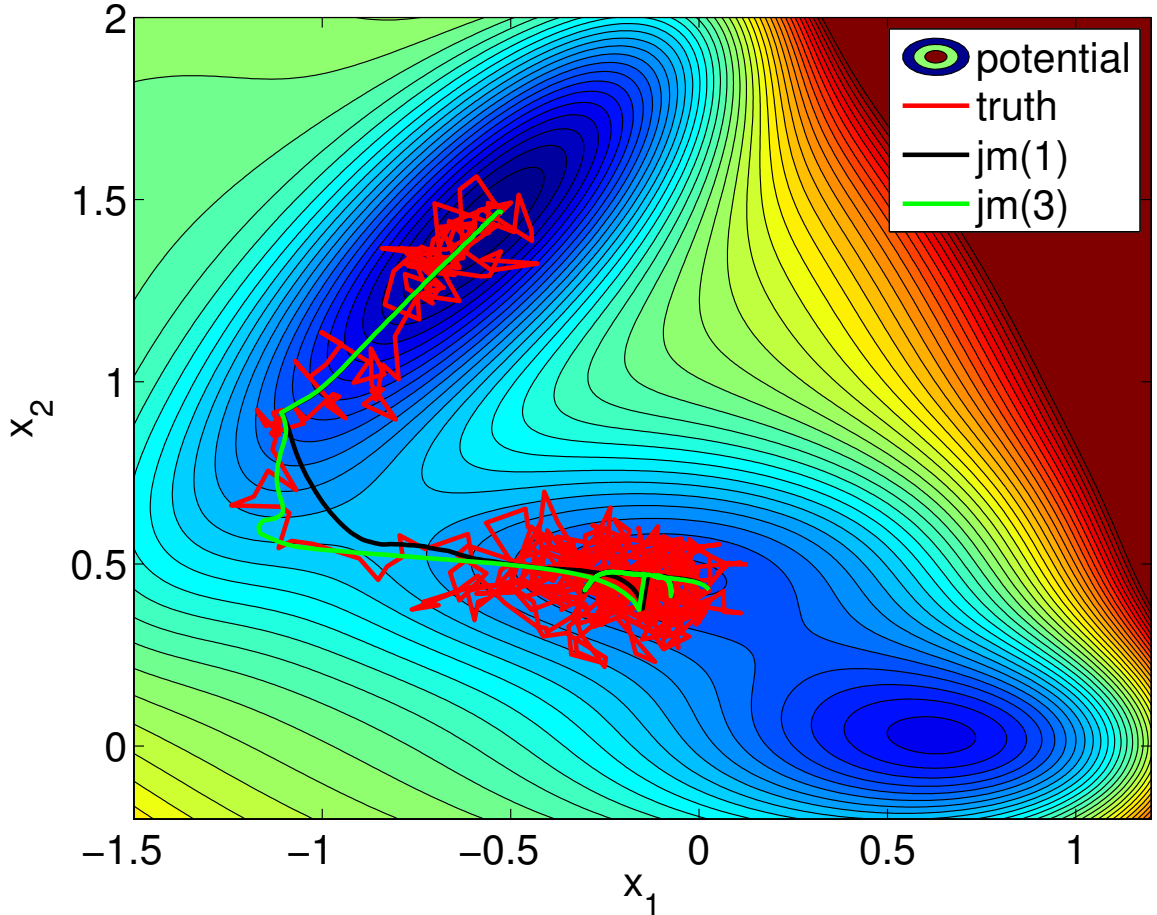
      using map from previous observation step:

      $$p(x_t, \xi \mid y_{1:t}) \propto p(y_t \mid x_t) p(x_t \mid T_{t-1}(\xi)) p_\xi(\xi)$$

    - Find a map from reference (η_1, η_2) to (x_t, ξ)
    - Triangular map structure recovers marginal on x_t
  - Generalize directly to occasional observations; now we are locally *smoothing* the trajectory between $t\text{-}1$ and t
  - Computational cost *grows* with the number of intermediate states…

Numerical examples

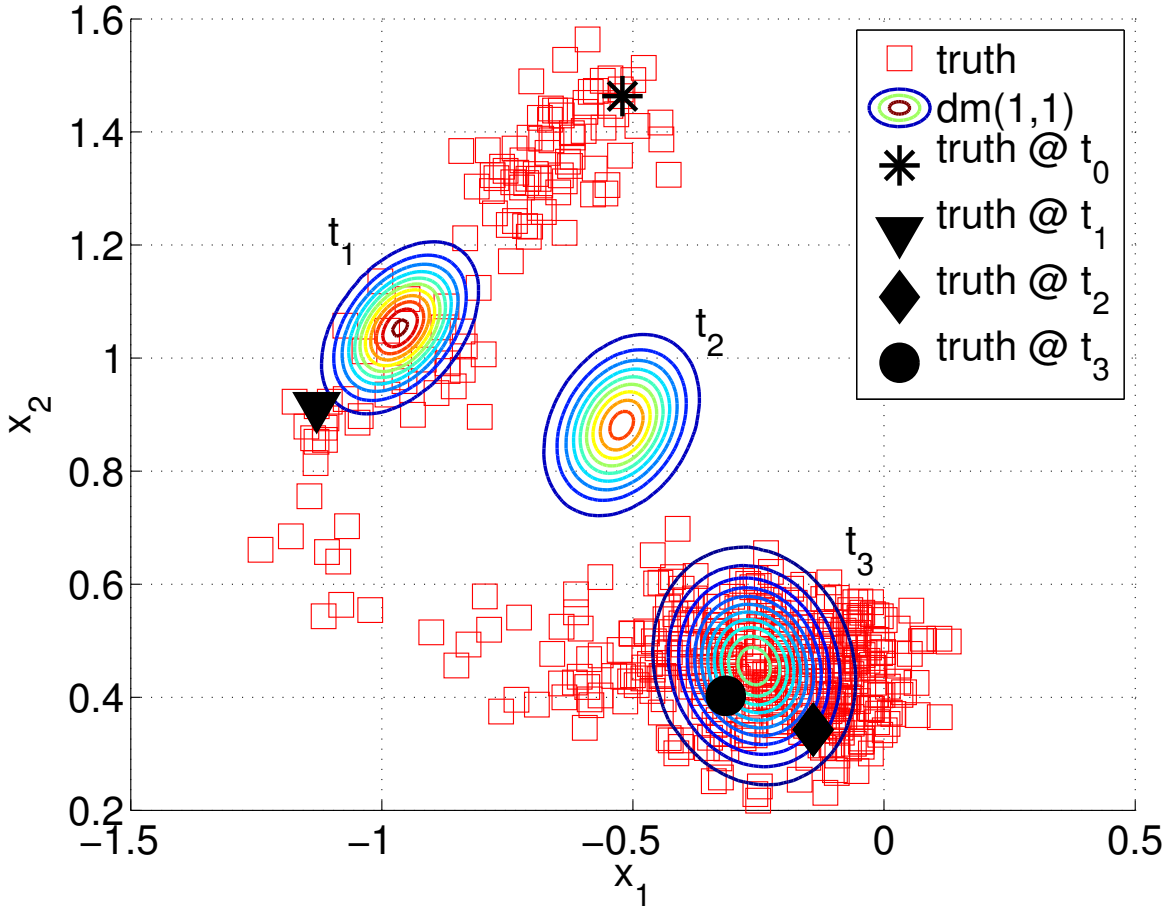
- **Example 3:** particle in a **Mueller-Brown potential**
  - Small noise, rare transition (1 in 10^6); observe x_2 every 100 steps
  - **Joint map:** posterior mean versus true trajectory; local smoothing



SAA with
256 samples

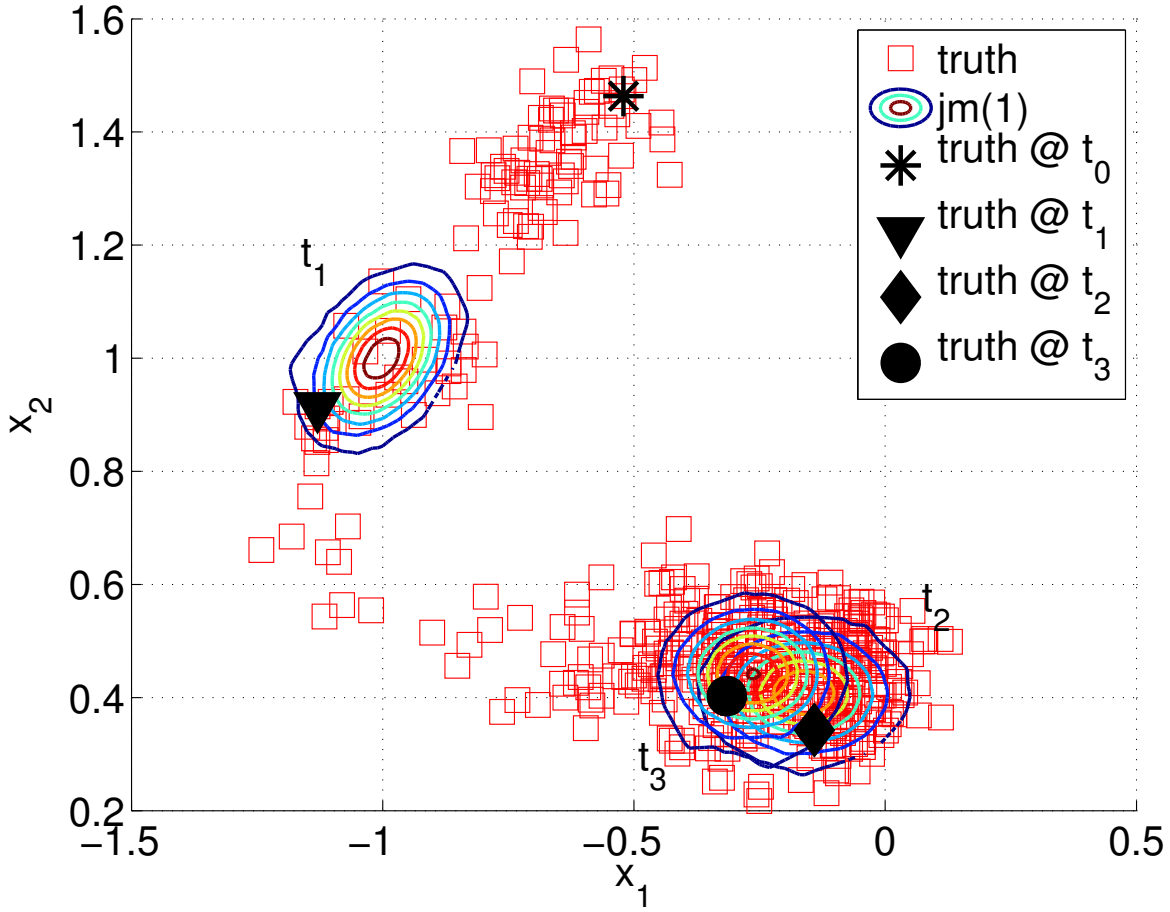
Numerical examples

- **Example 3:** particle in a **Mueller-Brown potential**
  - Small noise, rare transition (1 in 10^6); observe x_2 every 100 steps
  - Compare double map with joint map; filtering distributions



Numerical examples

- **Example 3:** particle in a **Mueller-Brown potential**
  - Small noise, rare transition (1 in 10^6); observe x_2 every 100 steps
  - Compare double map with joint map; filtering distributions



Conclusions

- Bayesian inference with **optimal transport maps**
  - Characterize and simulate from the posterior by solving an **optimization** problem
  - Clear convergence criterion; evidence computed as a byproduct
  - Favorable performance comparisons with MCMC

- A **map-based** approach to Bayesian filtering
  - Based on the construction of transport maps from a reference measure
  - Two algorithms: double map and joint map
  - Both involve the solution of optimization problems
  - Double map generalizes EnKF
  - Continuous representations; convergence to the true Bayesian solution

Conclusions

- **Many open issues:**
  - More efficient optimization approaches
  - Which DOFs are needed? Better adaptive parameterization and enrichment schemes
  - Dimension reduction (ideas from yesterday): using Hessian information to identify directions that change from prior to posterior
  - Filtering: understanding the role of **regularization** in representing the prior; constructing improved regularization schemes in the context of maps

Acknowledgments

- **Support** from *US Department of Energy, Office of Advanced Scientific Computing Research*

- References:
  - T. Moselhy, Y. Marzouk, “Bayesian inference with optimal maps.” *J. Comp. Phys.*, **231**: 7815–7850 (2012). Also at http://arxiv.org/abs/1109.1516.