Singular Likelihoods to Prevent Particle Filter Collapse

Gregor Robinson, Ian Grooms, William Kleiber
Applied Mathematics, CU Boulder

2018 April 16


Ian, what's that dark place over there?
Importance sampling.

Particle filtering is an importance sampling approach to the Bayesian estimation problem for dynamical systems.

Main points:

- Particle filtering basics
- Why particle filters are cool in theory
- Mode of failure in high-dimensional practice
- Our approach to reduce the curse of dimensionality
- Example on a toy model

IMPORTANCE SAMPLING

To approximate 'target' distribution μ:

1. Draw ensemble of N_e 'particles' from some 'proposal' distribution (that's easy to simulate)
2. assign each particle a positive weight $0 < w^{(i)} < 1$ such that $\sum_i w^{(i)} = 1$ and specially rigged so that

$$\mu \approx \mu_{N_e} := \sum_{i=1}^{N_e} w^{(i)} \delta(\boldsymbol{x} - \boldsymbol{x}^{(i)})$$

Simple computations can give us the right weights so that for any $\varphi \in C_b$, $\langle \mu_{N_e}, \varphi \rangle \to \langle \mu, \varphi \rangle$ as $N_e \to \infty$.
(Weak convergence.)

SEQUENTIAL IMPORTANCE SAMPLING (SIS)

1. Start with posterior (analysis) importance sample (ensemble) at assimilation timestep $k-1$
2. Propagate by proposal kernel $\mathbb{P}\left(\mathbf{x}_k^{(i)}|\mathbf{x}_{0:k-1}^{(i)}, \mathbf{y}_k\right)$
3. Reweight so ensemble $\left\{\left(\mathbf{x}_k^{(i)}, w_k^{(i)}\right)\right\}$ becomes valid importance approximation of posterior at assimilation timestep k:

$$w_k^{(i)} \propto w_{k-1}^{(i)} \frac{\mathbb{P}\left(\mathbf{y}_k|\mathbf{x}_k^{(i)}\right) \mathbb{P}\left(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}\right)}{\mathbb{P}\left(\mathbf{x}_k^{(i)}|\mathbf{x}_{0:k-1}^{(i)}, \mathbf{y}_k\right)}.$$

with likelihood $\mathbb{P}\left(\mathbf{y}_k|\mathbf{x}_k\right)$ and transition prior $\mathbb{P}\left(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}\right)$

HERDING CATS (PARTICLES)

Particles can wander from observations.
Closest particle then has much higher likelihood, picking up almost all the weight.
Specifically:

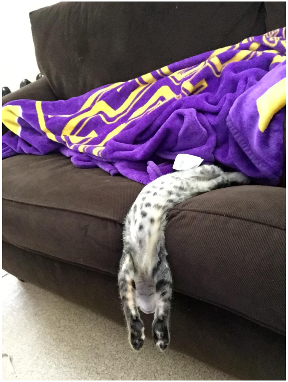
$$w_k^{(i)} \propto \mathbb{P}\left(\mathbf{y}_k | \boldsymbol{x}_k^{(i)}\right) w_{k-1}^{(i)}$$

If one particle has $w^{(i)} \approx 1$ then the UQ is bad. Define

$$\text{Effective Sample Size} = \text{ESS} = \left(\sum_i (w^{(i)})^2\right)^{-1}$$

$$1 \leq \text{ESS} \leq N_e.$$

COLLAPSE



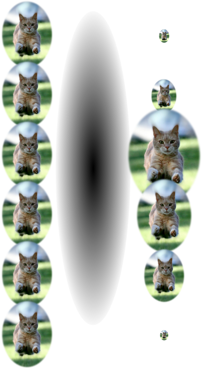
The PF has 'collapsed' when ESS $\ll N$.

To fix, resample: eliminate particles with small weights, replicate ones with large weights, resetting all weights to $1/N$ (gross simplification).

Prior

Prior

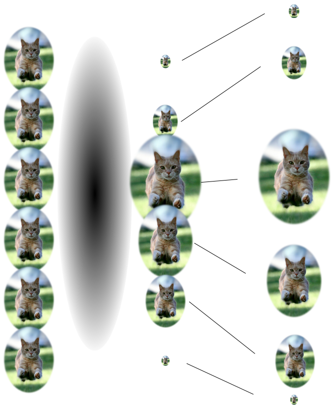
Observation



Prior

Observation

Dynamic

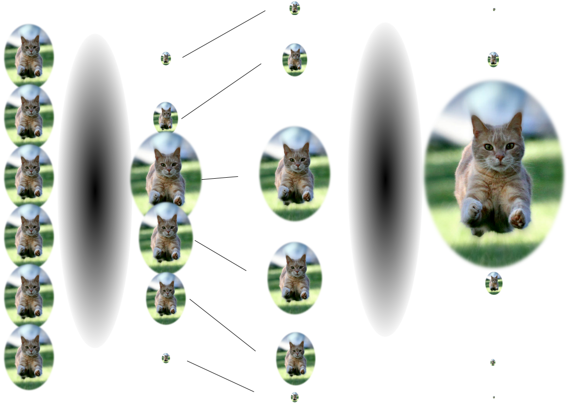


Prior

Observation

Dynamic

COLLAPSE!

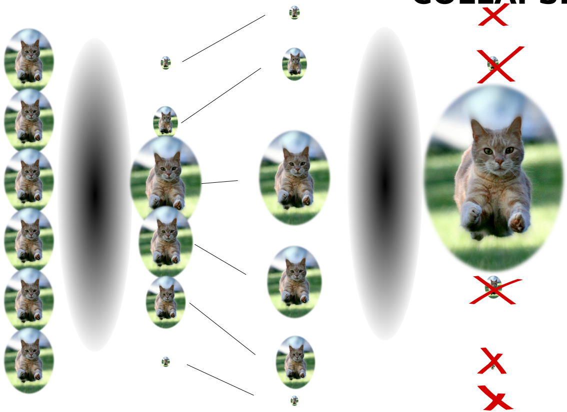


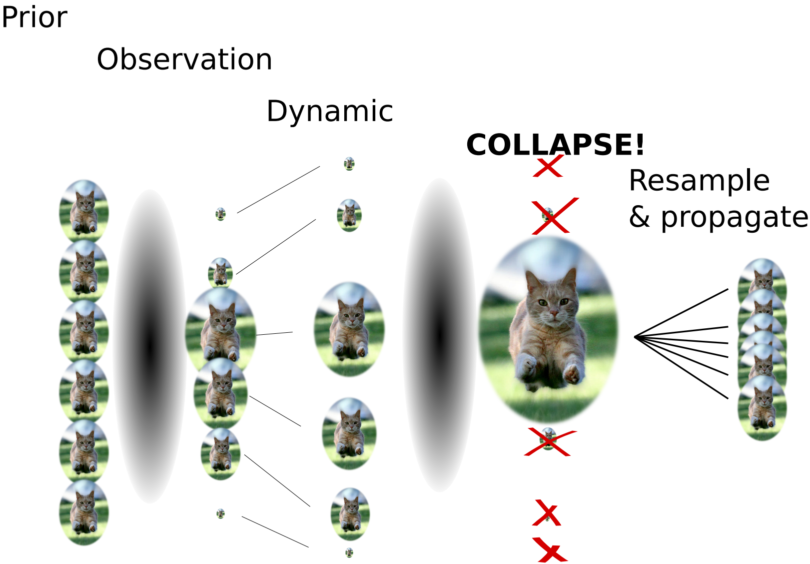
Prior

Observation

Dynamic

COLLAPSE!




Prior
Observation
Dynamic
COLLAPSE!
Resample
& propagate

SEQUENTIAL IMPORTANCE SAMPLING WITH RESAMPLING (SIR)

SIR weakly converges[1] to the Bayesian posterior as $N_e \to \infty$ with extremely permissive constraints on prior, transition kernel, and likelihood.

That's useful for UQ of non-Gaussian and nonlinear problems!

The SIR PF does not work well for high-dimensional problems.

[1] D. Crisan

Snyder et al.[2] consider SIR with linear Gaussian obs:

$$\boldsymbol{y} = \mathbf{H}\boldsymbol{x} + \boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}(0, \mathbf{R})$$

To avoid collapse, they show you need $N_e \sim \exp\{\tau^2/2\}$, where

$$\tau^2 = \sum_k \lambda_k^2 \left(\frac{3}{2}\lambda_k^2 + 1 \right)$$

Where, for the 'standard proposal',

$$\mathbf{P} = \mathrm{Cov}[\mathbf{R}^{-1/2}\mathbf{H}\boldsymbol{x}]$$

and λ_k^2 are the eigenvalues of $\mathbf{P}$.

[2] Bengtsson, Bickel, & Li 2008; Snyder, Bengtsson, Bickel, & Anderson 2008

$\exists$ optimal proposal and related ways to improve required number of particles by incorporating observations into proposal:

- Snyder, Bengtsson, & Morzfeld (2015) showed that the number of required particles is exponential in the 'effective dimension' even in best case scenario of optimal proposal[3]
- Chorin, Morzfeld, & Tu (2009–present) develop an 'implicit' PF that approximates the optimal proposal
- Ades & van Leeuwen (2009–present) develop an 'equivalent weights' PF that is related to (but not equivalent to) the optimal proposal

[3] See also Agapiou, Papaspiliopoulos, Sanz-Alonso, and Stuart (2017) for precise non-asymptotic results

FIDDLING WITH THE LIKELIHOOD

In practice (e.g. for satellite obs) the likelihood is rarely known precicesly. Common to assume spatially-uncorrelated observation errors for serial processing in EnKF.

If we can choose $\mathbf{R}$ to increase EnKF computational efficiency, why not choose it to avoid PF collapse?

FIDDLING WITH THE LIKELIHOOD

Revisit Snyder estimate: need $N_e \sim \exp\{\tau^2/2\}$, where

$$\tau^2 = \sum_k \lambda_k^2 \left(\frac{3}{2}\lambda_k^2 + 1\right)$$

$$\mathbf{P} = \mathrm{Cov}[\mathbf{R}^{-1/2}\mathbf{H}\boldsymbol{x}]$$

Idea: increase variance at scales that don't matter that much.

For geophysical forecast, viscous damping means that small scales don't matter much.

Consider the case where $\mathbf{H} = \mathbf{I}$, and $\mathbf{R}$ and $\mathrm{Cov}[\boldsymbol{x}]$ are simultaneously diagonalizable. Then

$$\lambda_k^2 = \frac{\sigma_k^2}{\gamma_k^2}$$

where σ_k^2 are eigvals of $\mathrm{Cov}[\boldsymbol{x}]$, and γ_k^2 are eigvals of $\mathbf{R}$.

$\boldsymbol{x}$ is an ordinary random field – realizations are, e.g., continuously differentiable – so the spectrum must decay.

$$\lim_{k \to \infty} \sigma_k^2 = 0$$

$\sigma_k \to 0$ is good news: even if the grid is refined ad infinitum, τ^2 might converge to a finite value

$$\tau^2 = \sum_k \lambda_k^2 \left(\frac{3}{2} \lambda_k^2 + 1 \right), \quad \lambda_k^2 = \frac{\sigma_k^2}{\gamma_k^2}$$

But presumably the observation error is also a continuous field, so $\gamma_k \to 0$ also. This is BAD for τ^2.

But it's widely accepted to use a spatially-uncorrelated obs error model, which has constant γ_k. This is good for τ^2.

Why not use $\gamma_k^2 \to \infty$, which is even better for τ^2?

$\sigma_k \to 0$ is good news: even if the grid is refined ad infinitum, τ^2 might converge to a finite value

$$\tau^2 = \sum_k \lambda_k^2 \left(\frac{3}{2} \lambda_k^2 + 1 \right), \quad \lambda_k^2 = \frac{\sigma_k^2}{\gamma_k^2}$$

But presumably the observation error is also a continuous field, so $\gamma_k \to 0$ also. This is BAD for τ^2.

But it's widely accepted to use a spatially-uncorrelated obs error model, which has constant γ_k. This is good for τ^2.

Why not use $\gamma_k^2 \to \infty$, which is even better for τ^2?

Changing the likelihood changes the posterior.

Changing the likelihood changes the posterior. Yes, but...

- a white likelihood already does that (sometimes)
- and you needn't change the posterior *much*:

Consider again a fully-Gaussian case with $\mathbf{H} = \mathbf{I}$ and simultaneously-diagonalizable $\mathbf{R}$ and $\mathrm{Cov}[\boldsymbol{x}]$.

The spectrum of the posterior is

$$\frac{\sigma_k^2 \gamma_k^2}{\sigma_k^2 + \gamma_k^2}$$

At small scales (large k), the posterior variance is small, regardless of how you choose γ_k (because σ_k is small).

As long as γ_k is correct at large scales, the posterior will be correct at large scales.

CONNECTION TO SMOOTHING

Our approach is like truncation/projection onto a large-scale subspace, but with a gradual cutoff.

Assuming jagged observation errors is equivalent to smoothing the observations by applying $\mathbf{R}^{-1/2}$, and then assuming uncorrelated errors

$$\hat{\boldsymbol{y}} = \mathbf{R}^{-1/2}\boldsymbol{y} = \mathbf{R}^{-1/2}\mathbf{H}\boldsymbol{x} + \hat{\boldsymbol{\xi}}, \quad \hat{\boldsymbol{\xi}} \sim \mathcal{N}(0, \mathbf{I})$$

Note that this is *not* equivalent to assuming uncorrelated obs error and then smoothing.

Example

As a the linear SPDE

$$\frac{du}{dt} = \left(-b - c\frac{d}{dx} + \nu\frac{d^2}{dx^2}\right)u + F_t, \tag{1}$$

in a 2π-periodic domain where the forcing is Gaussian, white in time, with spatial spectrum $(1+|k|)^{-1}$.
2048 Fourier modes, $b = 1$, $c = 2\pi$, and $\nu = 1/9$.

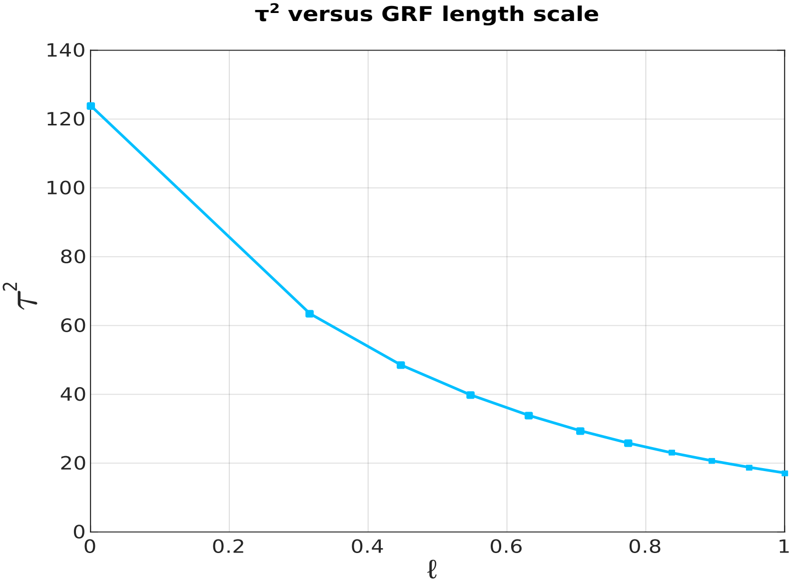
Observations are taken on a regular spatial grid, at discrete times; the true obs errors are smooth (Gaussian, zero-mean).

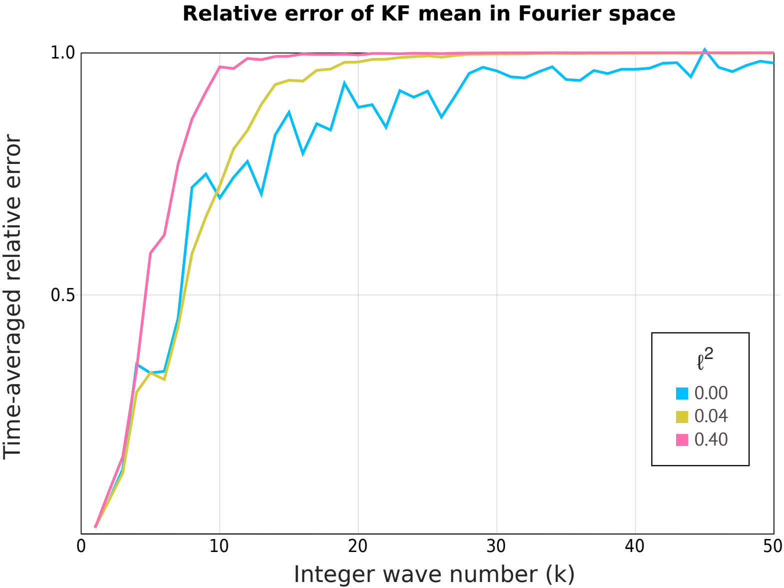
We compare posteriors using true $\mathbf{R}$, $\mathbf{R} = \gamma^2\mathbf{I}$, or a second-order finite-difference discretization of $\gamma^2(1 - \ell^2\partial_x^2)$.

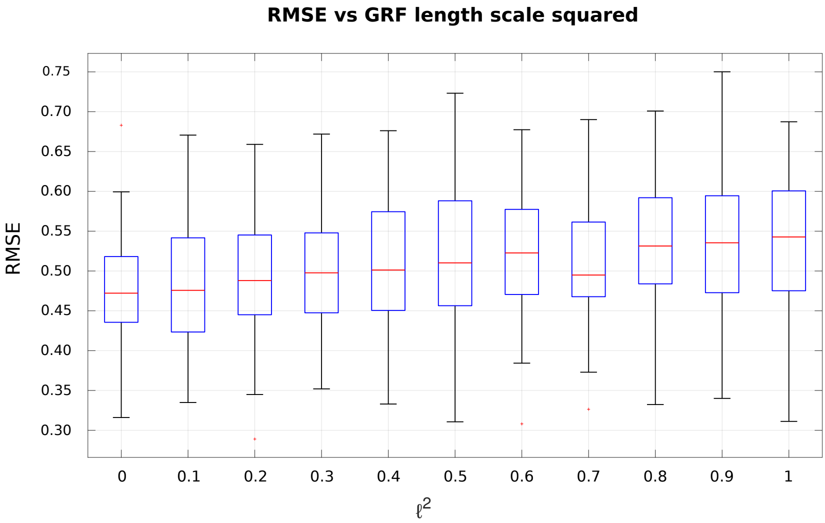
First we compute the exact solution of the filtering problem using the Kalman filter. The filter covariance converges exponentially to a steady state; we use this information to compute τ^2:

64 obs:

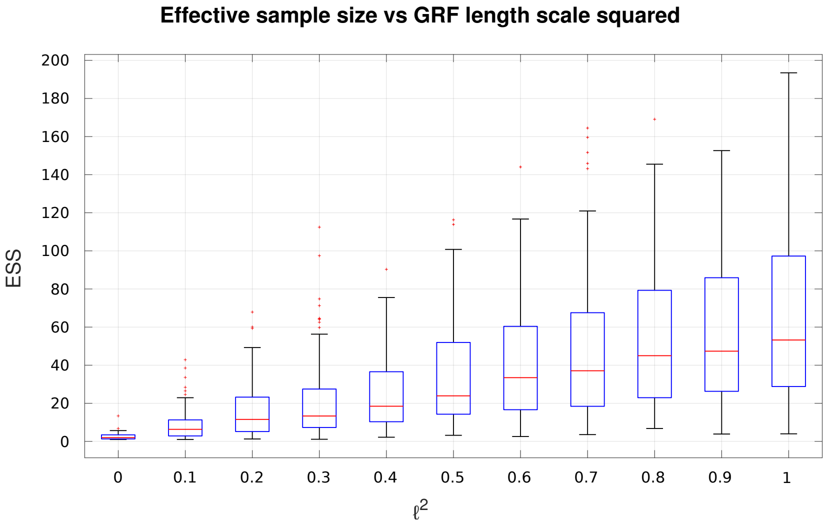
- True $\mathbf{R}$: $\tau^2 = 87$, $N \approx 10^{19}$
- $\mathbf{R} = \gamma^2\mathbf{I}$: $\tau^2 = 162$, $N \approx 10^{35}$
- Generalized, $\ell^2 = 2$: $\tau^2 = 16$, $N \approx 3000$


τ² versus GRF length scale
140
120
100
80
60
40
20
0
0
0.2
0.4
0.6
0.8
1
τ^2
ℓ


Relative error of KF mean in Fourier space
Time-averaged relative error
1.0
0.5
Integer wave number (k)
0
10
20
30
40
50
ℓ^2
0.00
0.04
0.40



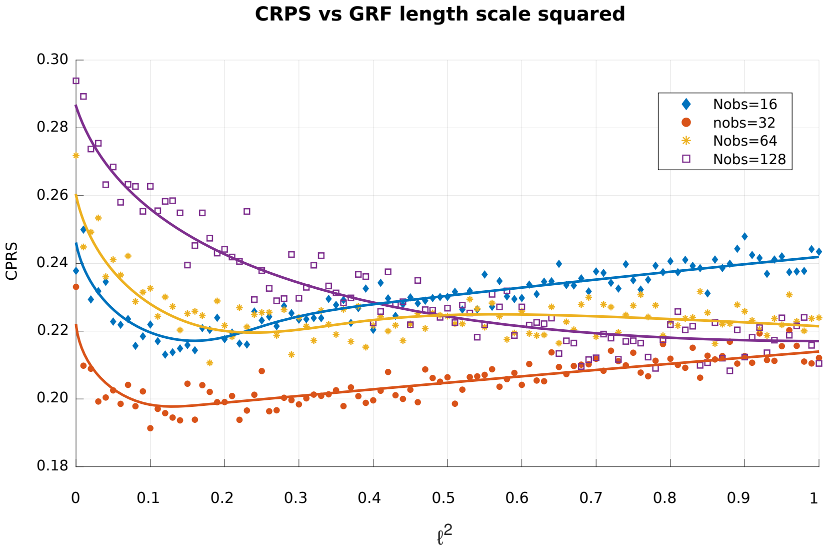
RMSE minimally suffers.



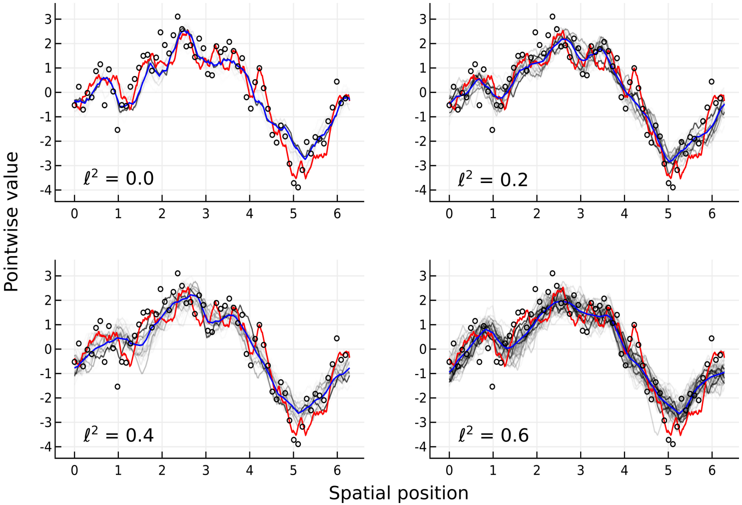
Ensemble size = 400

Quantifying uncertainty quantification

We use the Continuous Ranked Probability Score (CRPS) to quantify the skill of our full PF estimate.


CRPS vs GRF length scale squared
CPRS
ℓ^2
0.30
0.28
0.26
0.24
0.22
0.20
0.18
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Nobs=16
nobs=32
Nobs=64
Nobs=128

Weighted samples for different ℓ^2



WHY USE $\gamma^2(1 - \ell^2\partial_x^2)$?

There's a link between PDEs and Gaussian random fields: the discretization of an elliptic, self-adjoint PDE approximates the precision matrix of a random field. [4]

We reverse this: a discretization of a self-adjoint elliptic PDE approximates the covariance matrix of a jagged random field.

Admits covariance structure that is exploitable for computational efficiency (including multiresolution) even on nonuniform grids.

[4] E.g. Lindgren, Rue, & Lindström, J R Stat Soc 2011

Conclusions & Future Directions

- Using a generalized random field for the obs error model can reduce incidence of PF collapse.
- The price to pay is that the posterior is only accurate on large scales; in practice that might be OK.
- We plan to continue development of approaches to discretizing **R**, esp. for scattered obs, and to apply to real meteorological data.
- Our approach probably won't be a silver bullet, but can be combined with implicit sampling/optimal-proposal and with localization.

Thanks to Jeff Anderson, Greg Beylkin & Chris Snyder for helpful discussions.

THE END!

Questions?

gregor.robinson@colorado.edu

@precompact