Kernel Sequential Monte Carlo

Ingmar Schuster* (Paris Dauphine)
Heiko Strathmann* (University College London)
Brooks Paige (Oxford)
Dino Sejdinovic (Oxford)

* equal contribution

April 25, 2016

Section 1

Outline

Section 2

Introduction

Importance Sampling estimators

- Importance Sampling identity

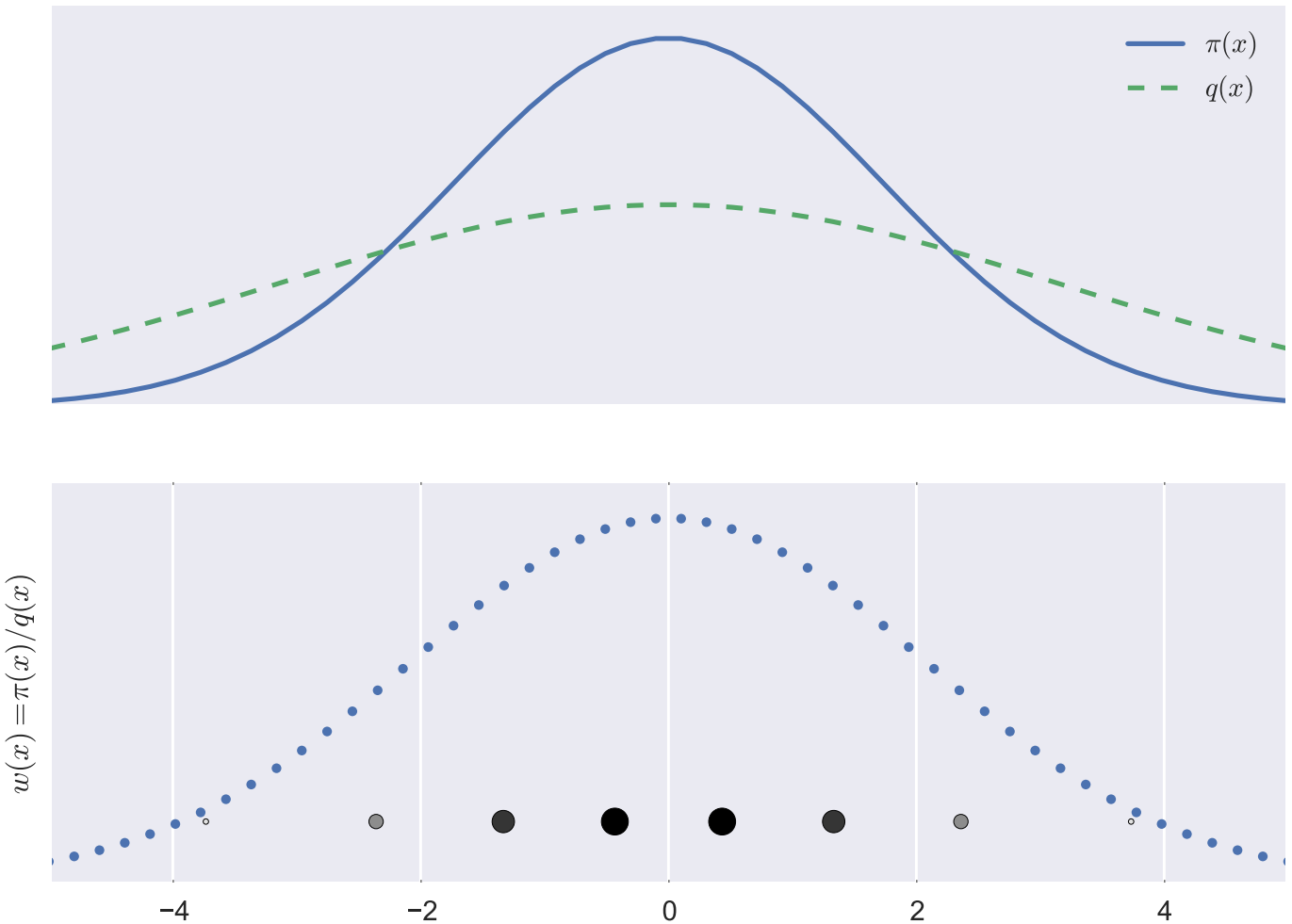
$$H = \int \pi(x)h(x)\mathrm{d}x = \int \frac{\pi(x)}{q(x)}h(x)q(x)\mathrm{d}x$$
$$\approx \frac{1}{w_\Sigma}\sum_{i=1}^{N} w(X_i)h(X_i)$$

  where $X_i \sim q$ iid, $w(X) = \pi(X)/q(X)$ called unnormalized importance weight, $w_\Sigma = \sum_{i=1}^{N} w(X_i)$

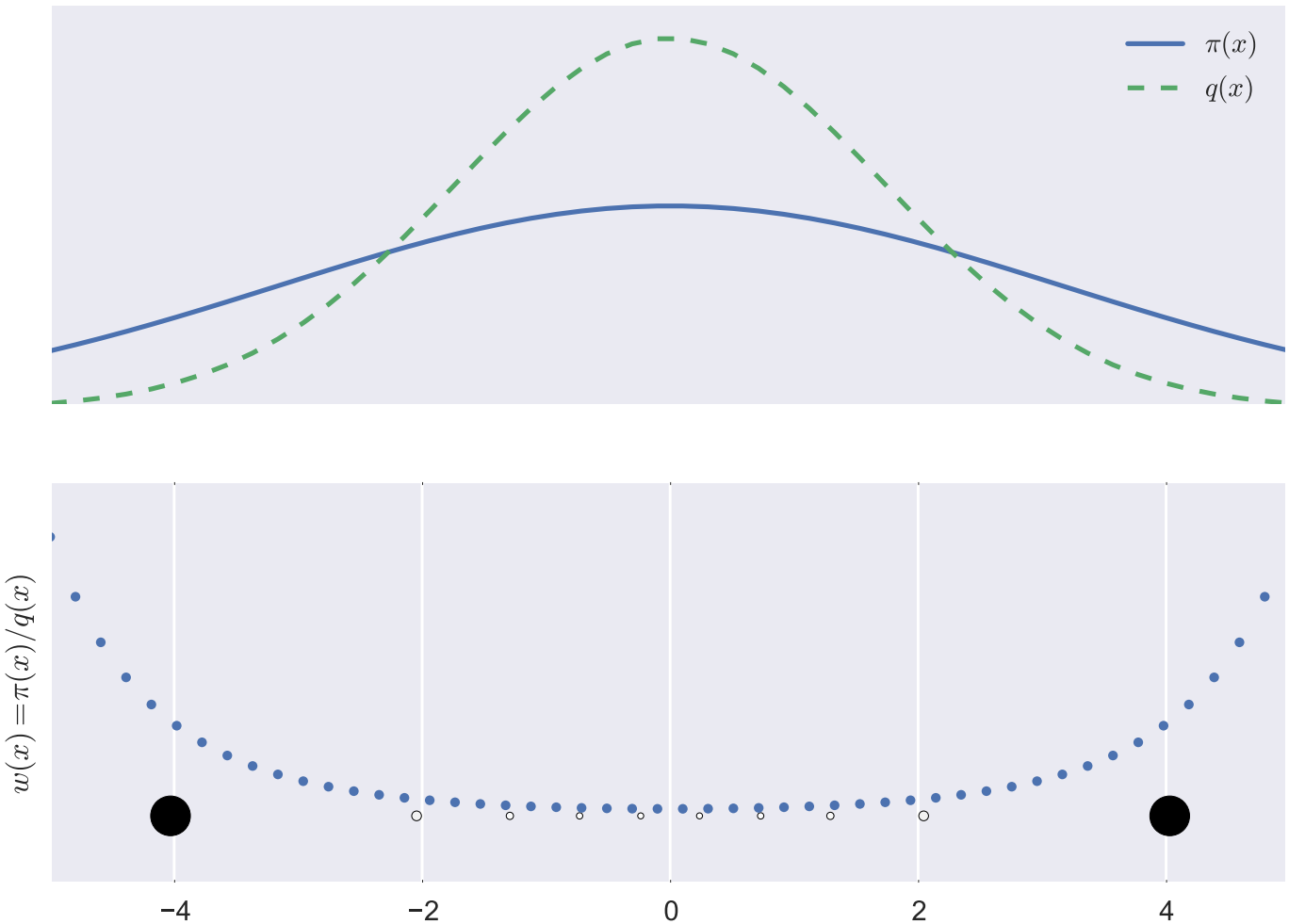
- PMC identity: for any law g over proposals

$$H = \iint \frac{\pi(x)}{q_t(x)}h(x)\ \mathrm{d}q_t(x)\ \mathrm{d}g(q_t) = \frac{1}{w_\Sigma}\sum_{t=1}^{T}\sum_{i=1}^{N} w_t(X_i)h(X_i)$$

Proposal fatter than target



Proposal thinner than target



Population Monte Carlo Cappé et al. (2004)

Input: initial proposal density q_0, unnormalized density π, population size N, sample size m
Output: lists P, W of m samples and weights
Initialize $P = List()$
Initialize $W = List()$
while $len(P) \leq m$ **do**
  construct proposal distribution q_t
  generate set of p samples $\mathbf{X}_t$ from q_t and append it to P
    for all $X \in \mathbf{X}_t$ append weights $\pi(X)/q_t(X)$ to W
end while

Sequential Monte Carlo Samplers

- Approximate integrals with respect to target distribution π_T
- Build upon Importance Sampling: approximate integral of h wrt density π_T using samples following density q (under certain conditions):

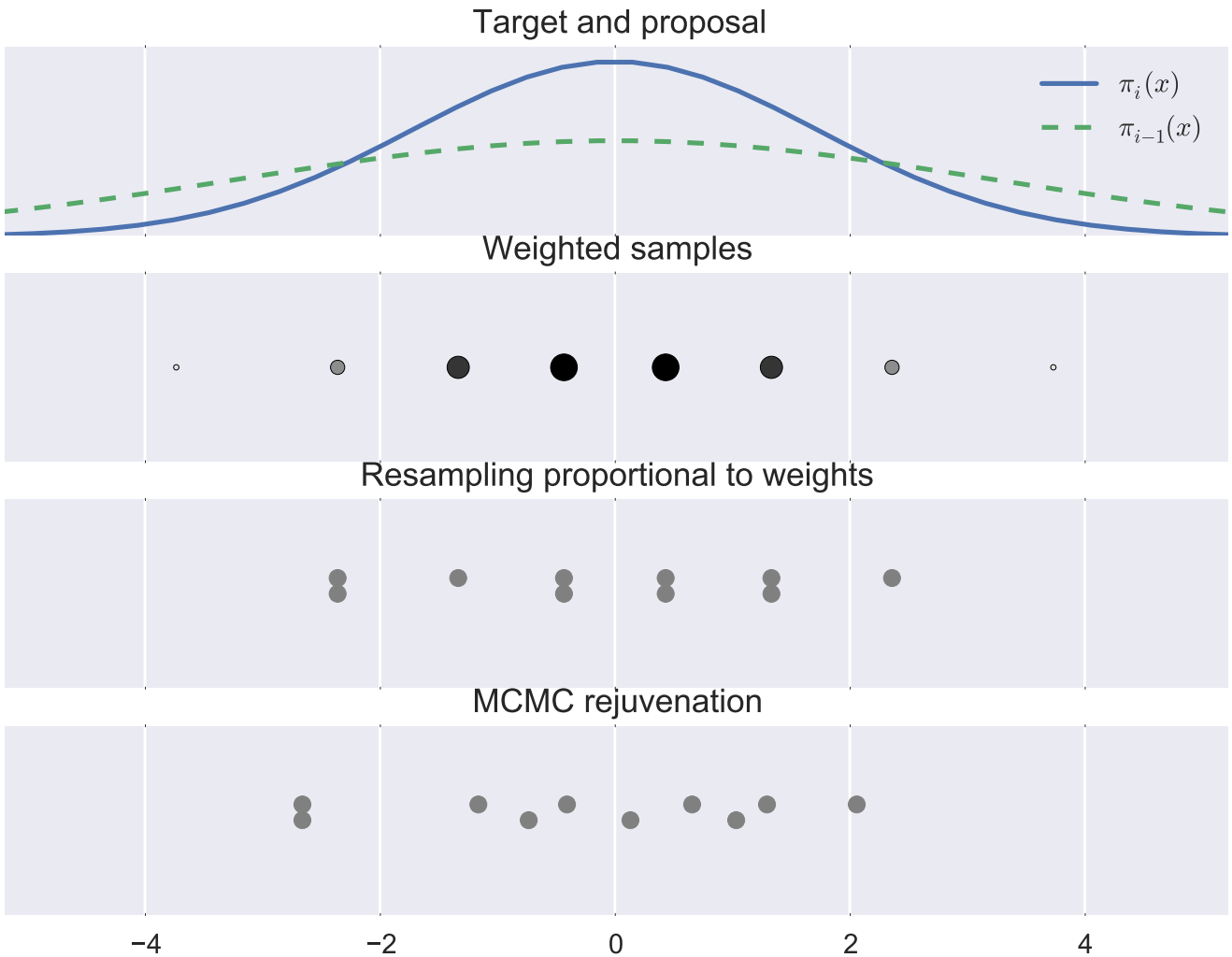
$$\int h(x)\mathrm{d}\pi_T(x) = \int h(x)\frac{\pi_T(x)}{q(x)}\mathrm{d}q(x)$$

- Given prior π_0, build sequence $\pi_0, \ldots, \pi_i, \ldots \pi_T$ such that
  - π_{i+1} is closer to π_T than π_i ($\delta(\pi_{i+1}, \pi_T) < \delta(\pi_i, \pi_T)$ for some divergence δ)
  - sample from π_i can approximate π_{i+1} well using importance weight function $w(\cdot) = \pi_{i+1}(\cdot)/\pi_i(\cdot)$

Sequential Monte Carlo Samplers

- At $i = 0$
  - Using proposal density q_0, generate particles $\{(w_{0,j}, X_{0,j})\}_{j=1}^{N}$ where $w_{0,j} = \pi_0(X_{0,j})/q_0(X_{0,j})$
  - *importance resampling*, resulting in N equally weighted particles $\{(1/N, \bar{X}_{0,j})\}_{j=1}^{N}$
  - *rejuvenation move* for each $\bar{X}_{0,j}$ by Markov Kernel leaving π_0 invariant
- At $i > 0$
  - approximate π_i by $\{(\pi_i(X_{i-1,j})/\pi_{i-1}(X_{i-1,j}), X_{i-1,j})\}_{j=1}^{N}$
  - resampling
  - rejuvenation leaving π_i invariant
  - if $\pi_i \neq \pi_T$, repeat

A visual SMC iteration



Sequential Monte Carlo Samplers

- estimate evidence Z_T of π_T by

$$Z_T \approx Z_0 \prod_{i=1}^{T} \frac{1}{N} \sum_j w_{i,j}$$

  (aka normalizing constant, marginal likelihood)
- Can be adaptive in rejuvenation steps without diminishing adaptation as required in adaptive MCMC
- Will construct rejuvenation using RKHS-embedding of particles

Intractable Likelihoods and Evidence

- intractable likelihoods arise in many models (e.g. nonconjugate latent variable models)
- for unbiased likelihood estimates, SMC/PMC still valid
- simple case: estimate likelihood using IS or SMC, leads to IS^2 (Tran et al., 2013) and SMC^2 (Chopin et al., 2011)
- results in noisy Importance Weights, but approximation of evidence (probability of model given data) is still valid (Tran et al., 2013, Lemma 3)
- *cannot easily use information on geometry of π for efficient inference* (e.g. gradients unavailable)

Kernel emulators

- In the following: adapt RKHS-based emulators to PMC and SMC in intractable likelihood settings for adapting to target geometry
- Using pd kernel $\mathbf{k}(\cdot, \cdot)$ we can
  - adapt to local covariance (Sejdinovic et al., 2014)
  - use gradient information of infinite exponential family approximation to π (Strathmann et al., 2015)
- Emulators used for constructing proposals q_t and use
  - importance correction in PMC
  - Metropolis-Hastings correction within in SMC rejuvenation moves

  for samples $X \sim q_t$

Kernel Emulators

- Local covariance: let $\mathbf{k}'(y, x) = \nabla_y \mathbf{k}(y, x)$ and $\mu(y) = \int \mathbf{k}'(y, x) \mathrm{d}\pi(x)$ then

$$K(y) = \int (\mathbf{k}'(y, x) - \mu(y))^2 \mathrm{d}\pi(x)$$

- Gradient emulation
  - fit infinite exponential family approximation

$$\tilde{q}(y) = \exp(f(y) - A(f))$$

where $f(y) = \langle f, \mathbf{k}(y, \cdot) \rangle_{\mathcal{H}}$ is the inner product between natural parameters f and sufficient statistics $\mathbf{k}(y, \cdot)$ in $\mathcal{H}$ by minimizing

$$\int (\nabla_y \log \pi(y) - \nabla_y f(y))^2 \mathrm{d}\pi(y)$$

  - use gradients information of $\log \tilde{q}$ in proposals

Section 3

Kernel SMC

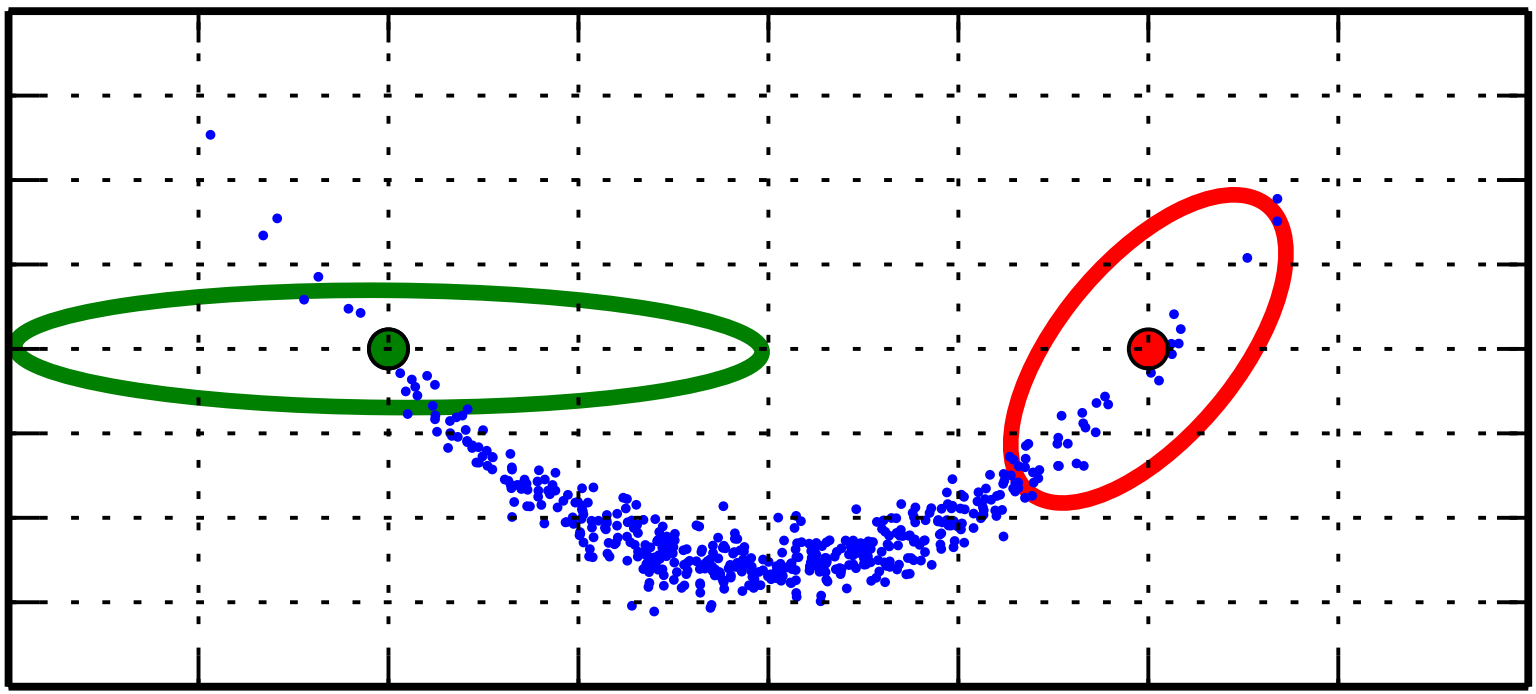
Kernel Gradient Importance sampling

- Use $\mathcal{N}(\cdot|X + \delta_1 \nabla_X \log \tilde{q}(X), \delta_2 C)$ proposals with importance weighting in PMC
- C is a fit to global covariance of target π
- resulting in Kernel Gradient Importance Sampling (KGRIS)
- variant of Gradient Importance Sampling (Schuster, 2015)

Kernel Adaptive SMC Sampler

- Use artificial sequence of distributions leading from prior π_0 to posterior π_T
- rejuvenation with MH moves using $\mathcal{N}(\cdot|X, \delta K(X))$ proposals
- resulting in Kernel Adaptive SMC (KASMC)
- similar to Adaptive SMC sampler, a special case when using a linear kernel (Fearnhead and Taylor, 2013)

KASMC versus ASMC



green: ASMC / KASMC with linear kernel
red: KASMC with Gaussian RBF kernel

Section 4

Implementation Details

Construction of Target Sequence

- For artificial distribution sequence we used geometric bridge

$$\pi_i \propto \pi_0^{1-\rho_i} \pi_T^{\rho_i}$$

where $(\rho_i)_{i=1}^T$ is an increasing sequence satisfying $\rho_T = 1$

- another standard choice in Bayesian Inference is adding datapoints one after another

$$\pi_i(X) = \pi(X|d_1, \ldots, d_{\lfloor \rho_i D \rfloor})$$

resulting in Iterated Batch Importance Sampling (Chopin, 2002, IBIS)

Stochastic approximation, variance reduction

- Free scaling parameters can be tuned for optimal scaling of MCMC using stochastic approximation framework of Andrieu and Thoms (2008)
  - asymptotically optimal acceptance rate for Random Walk MH is $\alpha_{opt} = 0.234$ (Rosenthal, 2011)
  - tune single parameter δ_i by

$$\delta_{i+1} = \delta_i + \lambda_i(\hat{\alpha}_i - \alpha_{opt})$$

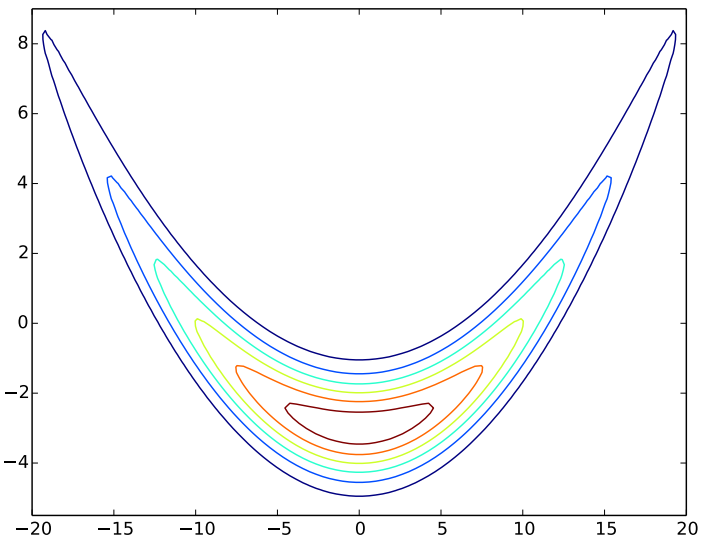
    for non-increasing $\lambda_1, \ldots, \lambda_T$
- used Random Fourier Features (Rahimi and Recht, 2007) for efficient on-line updates of emulators
- used weighted updates and Rao-Blackwellization for variance reduction in estimated emulators

Section 5

Evaluation

Synthetic nonlinear target (Banana)

- Synthetic target: Banana distribution in 8 dimensions, i.e. Gaussian with twisted second dimension



Synthetic nonlinear target (Banana)

- Compare performance of Random-Walk rejuvenation with asymptotically optimal scaling ($\nu = 2.38/\sqrt{d}$), ASMC and KASMC with Gaussian RBF kernel
- Fixed learning rate of $\lambda = 0.1$ to adapt scale parameter using stochastic approximation
- Geometric bridge of length 20
- 30 Monte Carlo runs
- Report Maximum Mean Discrepancy (MMD) using polynomial kernel of order 3: distance of moments up to order 3 between ground truth samples and samples produced by each method

Synthetic nonlinear target (Banana)

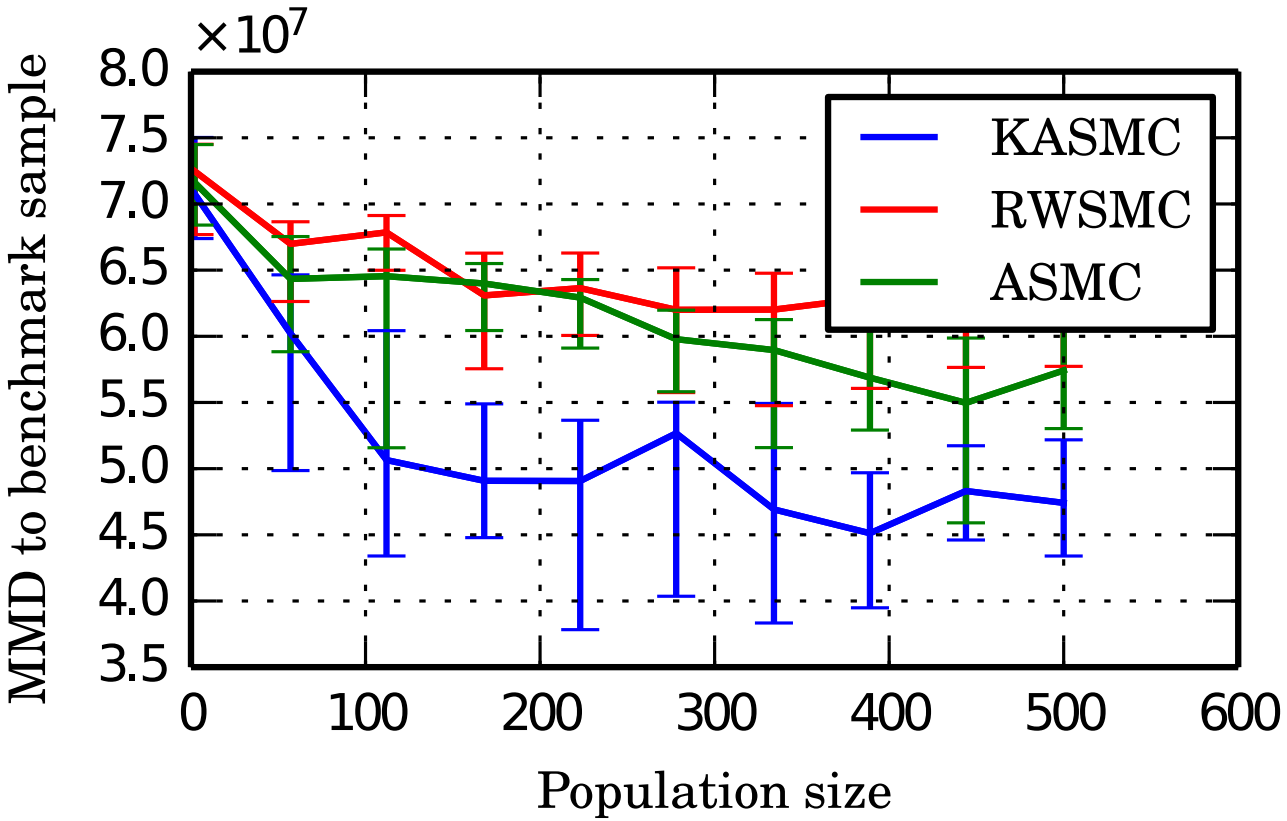


Figure: Improved convergence of all mixed moments up to order 3 of KASMC compared to ASMC and RW-SMC.

Evidence approximation for intractable likelihoods

- in classification using Gaussian Processes (GP), logistic transformation renders likelihood intractable
- likelihood can be unbiasedly estimated using Importance Sampling from EP approximation
- estimate model evidence when using ARD kernel in the GP
- particularly hard because noisy likelihoods means noisy importance weights
- ground truth by averaging evidence estimate over 20 long running SMC algorithms

Evidence approximation for intractable likelihoods

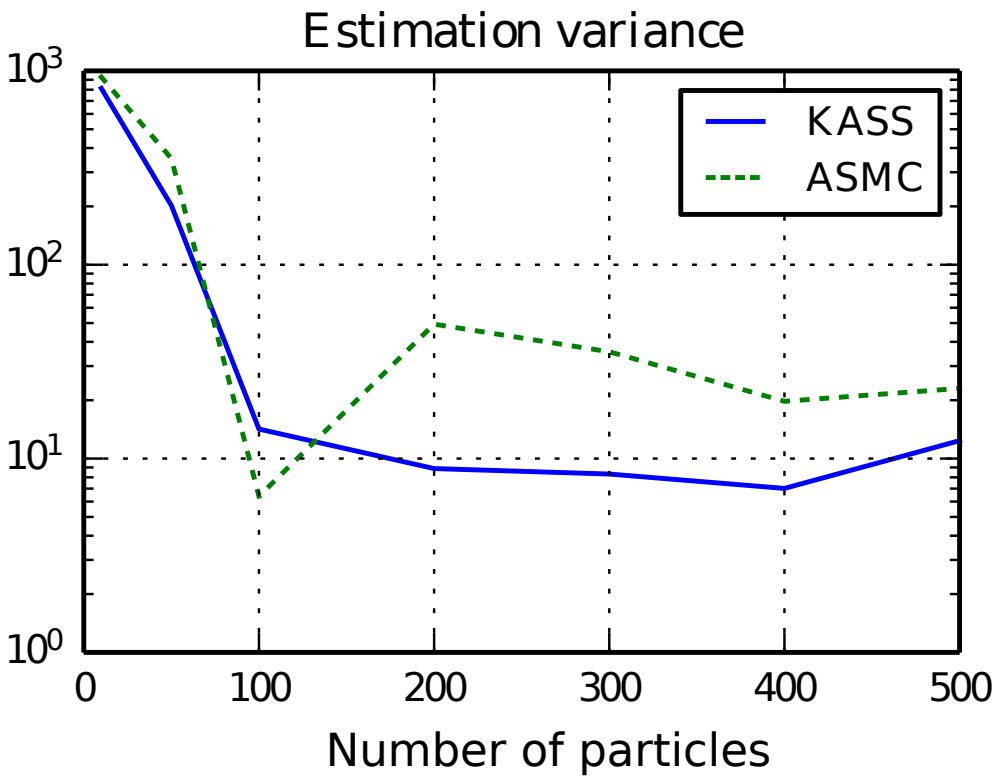
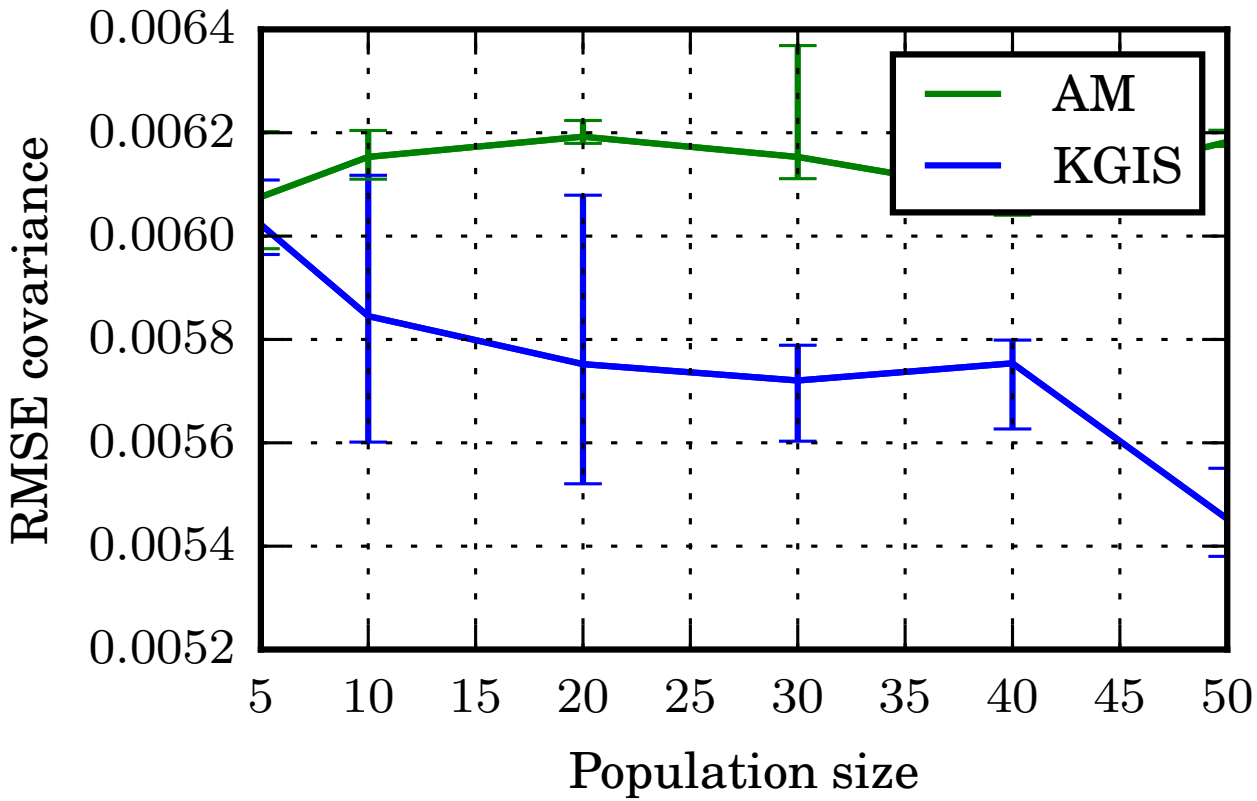


Figure: Monte Carlo Variance, KASMC in blue, ASMC in green.

Stochastic volatility model with intractable likelihood

- Stochastic volatility particularly challenging class of bayesian inverse problems
- time series as a high-dimensional nuisance variable
- models have to capture the non-linearities in the data (Barndorff-Nielsen and Shephard, 2001)
- concentrate on the prediction of daily volatility of asset prices, reusing the model and dataset studied by Chopin et al. (2011) (nuisance of dimension $d = 753$)
- report RMSE of target covariance estimate

KGRIS with Stochastic volatility



Section 6

Conclusion

Conclusion (1)

- Developed Kernel SMC framework
- KSMC exploits kernel emulators of target structure
- combines these with general SMC/PMC advantages for multimodal targets and evidence estimation
- especially attractive when likelihoods are intractable

Conclusion (2)

- evaluated on several challenging models where it was clearly improving statistical efficiency
  - KASMC exhibits better MMD for Banana
  - less MC variance than ASMC in evidence estimation for GP classification
  - KGRIS clearly improves covariance estimates in Stochastic Volatility model

Thanks!

Literature I

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing*, 18(November):343–373.

Barndorff-Nielsen, O. E. and Shephard, N. (2001). Non-Gaussian Ornstern-Uhlenbeck-Based Models and Some of Their Uses in Financial Economics. *Journal of the Royal Statistical Society. Series B*, 63(2):167–241.

Cappé, O., Guillin, a., Marin, J. M., and Robert, C. P. (2004). Population Monte Carlo. *Journal of Computational and Graphical Statistics*, 13(4):907–929.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, 89(3):539–552.

Literature II

Chopin, N., Jacob, P. E., and Papaspiliopoulos, O. (2011). SMC^2: an efficient algorithm for sequential analysis of state-space models. 0(1):1–27.

Fearnhead, P. and Taylor, B. M. (2013). An Adaptive Sequential Monte Carlo Sampler. *Bayesian Analysis*, (2):411–438.

Rahimi, A. and Recht, B. (2007). Random Features for Large Scale Kernel Machines. In *Neural Information Processing Systems*, number 1, pages 1–8.

Rosenthal, J. S. (2011). Optimal Proposal Distributions and Adaptive MCMC. In *Handbook of Markov Chain Monte Carlo*, chapter 4, pages 93–112. Chapman & Hall.

Literature III

Schuster, I. (2015). Consistency of Importance Sampling estimates based on dependent sample sets and an application to models with factorizing likelihoods. *arXiv preprint*, pages 1–14.

Sejdinovic, D., Strathmann, H., Garcia, M. L., Andrieu, C., and Gretton, A. (2014). Kernel Adaptive Metropolis-Hastings. *arXiv*, 32.

Strathmann, H., Sejdinovic, D., Livingstone, S., Szabo, Z., and Gretton, A. (2015). Gradient-free Hamiltonian Monte Carlo with efficient Kernel Exponential Families. In *Neural Information Processing Systems*.

Tran, M.-N., Scharth, M., Pitt, M. K., and Kohn, R. (2013). Importance sampling squared for Bayesian inference in latent variable models.