Theorem[section] [theorem]Lemma
[theorem]Example [theorem]Definition

# The intrinsic dimension of importance sampling


Omiros Papaspiliopoulos
www.econ.upf.edu/~omiros

ICREA
upf. Universitat Pompeu Fabra Barcelona
Department of Economics and Business
Barcelona GSE Graduate School of Economics

Jointly with:

- Sergios Agapiou (U of Cyprus)
- Daniel Sanz-Alonso (U of Warwick $\rightarrow$ Brown)
- Andrew M. Stuart (U of Warwick $\rightarrow$ Caltech)

# Summary

*"Our purpose in this paper is to overview various ways of measuring the computational complexity of importance sampling, to link them to one another through transparent mathematical reasoning, and to create cohesion in the vast published literature on this subject. In addressing these issues we will study importance sampling in a general abstract setting, and then in the particular cases of Bayesian inversion and filtering."*

# Outline

1. Importance sampling
2. Linear inverse problems & intrinsic dimension
3. Dynamic linear inverse problems: sequential IS
4. Outlook

# Autonormalised IS

$$\mu(\phi) = \frac{\pi(\phi g)}{\pi(g)},$$

$$\mu^N(\phi) := \frac{\frac{1}{N}\sum_{n=1}^{N}\phi(u^n)g(u^n)}{\frac{1}{N}\sum_{m=1}^{N}g(u^m)}, \quad u^n \sim \pi \quad \text{i.i.d.}$$

$$= \sum_{n=1}^{N} w^n \phi(u^n), \qquad w^n := \frac{g(u^n)}{\sum_{m=1}^{N} g(u^m)},$$

$$\mu^N := \sum_{n=1}^{N} w^n \delta_{u^n}$$

# Quality of IS and metrics

Distance between random measures[1]

$$d(\mu, \nu) := \sup_{|\phi|\leq 1} \left[\mathbb{E}\big(\mu(\phi) - \nu(\phi)\big)^2\right]^{\frac{1}{2}},$$

Interested in $d(\mu^N, \mu)$

[1] Rebeschini, P. and van Handel, R. (2013). Can local particle filters beat the curse of dimensionality? *arXiv preprint arXiv:1301.6585*

Divergence metrics between target and proposal:

- $D_{\chi^2}(\mu\|\pi) := \pi\left(\left[\frac{g}{\pi(g)} - 1\right]^2\right) = \rho - 1$;
  $\rho = \pi(g^2)/\pi(g)^2$
- $D_{\mathrm{KL}}(\mu\|\pi) = \pi\left(\frac{g}{\pi(g)} \log \frac{g}{\pi(g)}\right)$

and is known[2] that

$$\rho \geq e^{D_{\mathrm{KL}}(\mu\|\pi)}$$

[2] Th. 4.19 of Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities*. Oxford University Press, Oxford

**Theorem**

*Let*

$$\rho := \frac{\pi(g^2)}{\pi(g)^2} < \infty\,.$$

*Then,*

$$d(\mu^N, \mu)^2 := \sup_{|\phi|\leq 1} \mathbb{E}\left[\left(\mu^N(\phi) - \mu(\phi)\right)^2\right]$$
$$\leq \frac{4}{N}\rho = \frac{4}{N}(1 + D_{\chi^2}(\mu\|\pi))$$

Slutsky's lemmas yield for $\overline{\phi} := \phi - \mu(\phi)$

$$\sqrt{N}(\mu^N(\phi) - \mu(\phi)) \Longrightarrow N\left(0, \frac{\pi(g^2\overline{\phi}^2)}{\pi(g)^2}\right)$$

# ESS

$$ESS(N) := \left(\sum_{n=1}^{N} (w^n)^2\right)^{-1} = N \frac{\pi^N(g)^2}{\pi^N(g^2)}$$

If $\pi(g^2) < \infty$, for large $N$

$$ESS(N) \approx N/\rho\,; \quad d(\mu^N, \mu)^2 \lessapprox \frac{4}{ESS(N)}$$

# Non-square integrable weights

Otherwise, extreme value theory[3] suggests that if density of weights has tails $\gamma^{-a-1}$, for $1 < a < 2$,

$$\mathbb{E}\left[\frac{N}{ESS(N)}\right] \approx C\,N^{-a+2}\,.$$

In any case, whenever $\pi(g) < \infty$, $w^{(N)} \to 0$ as $N \to \infty$[4].

---

[3] e.g. McLeish, D. L. and O'Brien, G. L. (1982). The expected ratio of the sum of squares to the square of the sum.
*Ann. Probab.*, 10(4):1019–1028

[4] e.g. Downey, P. J. and Wright, P. E. (2007). The ratio of the extreme to the sum in a random sequence.
*Extremes*, 10(4):249–266

# Weight collapse: "unbounded degrees of freedom"

$$\pi_d(du) = \prod_{i=1}^{d} \pi_1(du(i)),\ tm_d(du) = \prod_{i=1}^{d} \mu_1(du(i)),$$

where $\mu_\infty$ and $\pi_\infty$. Then

$$\rho_d \approx e^{c^2 d}$$

and a non-trivial calculation[5] shows unless $N$ grows exponentially with $d$, $w^{(N)} \to 1$

[5] Bickel, P., Li, B., Bengtsson, T., et al. (2008). Sharp failure rates for the bootstrap particle filter in high dimensions.
In *Pushing the limits of contemporary statistics: Contributions in honor of Jayanta K. Ghosh*, pages 318–329. Institute of Mathematical Statistics

# Weight collapse: singular limits

Suppose

$$g(u) = \exp\big(-\epsilon^{-1} h(u)\big)$$

where $h$ unique minimun at $u^*$. Laplace approximation yields

$$\rho_\epsilon \approx \sqrt{\frac{h''(u^\star)}{4\pi\epsilon}}.$$

# Literature pointers

- The metric is introduced in [Del Moral, 2004]; neat formulation of [Rebeschini and van Handel, 2013]. Concurrent work for $L^1$ error in [Chatterjee and Diaconis, 2015]. Other concentration inequalities available, e.g. Th 7.4.3 of [Del Moral, 2004] but based on covering numbers. We provide an alternative concentration with more assumptions on $g$ and less on $\phi$ following [Doukhan and Lang, 2009]
- More satisfactory are results on concentrations for interacting particle systems, but those typically assume very strong assumptions on both weights and transition dynamics, see e.g. [Del Moral and Miclo, 2000]
- For algebraic deterioration if importance sampling in Bayesian learning problems see [Chopin, 2004]

# Outline

1. Importance sampling
2. Linear inverse problems & intrinsic dimension
3. Dynamic linear inverse problems: sequential IS
4. Outlook

# Bayesian linear inverse problem in Hilbert spaces

$$y = Ku + \eta, \qquad \text{on} \in (\mathcal{H}, \langle \cdot, \cdot \rangle, \|\cdot\|)$$

$$\eta \sim N(0, \Gamma) \quad u \sim N(0, \Sigma)$$

$\Gamma, \Sigma : \mathcal{H} \to \mathcal{H} \quad \eta \in \mathcal{Y} \supseteq \mathcal{H},\ u \in \mathcal{X} \supseteq \mathcal{H} \quad K : \mathcal{X} \to \mathcal{Y}$

E.g. linear regression, signal deconvolution.

# Bayesian inversion/learning

Typically, $K$ bounded linear operator with ill-conditioned generalised inverse

$u|y \sim \mathbb{P}_{u|y} = N(m, C)$

$$
\begin{aligned}
C^{-1} &= \Sigma^{-1} + K^*\Gamma^{-1}K, \\
C^{-1}m &= K^*\Gamma^{-1}y.
\end{aligned}
$$

(or Schur's complement to get different inversions)

# Connection to importance sampling

This learning problem is entirely tractable and amenable to simulation/approximation. However, we take it as a tractable test case to understand importance sampling:

$$\pi(du) \equiv N(0, \Sigma) \qquad \mu(du) \equiv N(m, C)$$

Absolute continuity not obvious!

# The key operator & an assumption

$$S := \Gamma^{-\frac{1}{2}} K \Sigma^{\frac{1}{2}}, \quad A := S^* S$$

Assume that the spectrum of $A$ consists of a countable number of eigenvalues:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_j \geq \cdots \geq 0$$

$$\tau := \mathrm{Tr}(A)$$ [6]

[6] finiteness of which used as necessary sufficient condition for no collapse by Bickel, P., Li, B., Bengtsson, T., et al. (2008). Sharp failure rates for the bootstrap particle filter in high dimensions. In *Pushing the limits of contemporary statistics: Contributions in honor of Jayanta K. Ghosh*, pages 318–329. Institute of Mathematical Statistics

# dof & effective number of parameters

$$efd := \mathrm{Tr}((I + A)^{-1} A)$$

has been used within the Statistics/Machine Learning community[7] to quantify the effective number of parameters within Bayesian or penalized likelihood frameworks

Here we have obtained an equivalent expression to the one usually encountered in the literature; it is also valid in the Hilbert space framework

[7] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit.
*J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(4):583–639,

Section 3.5.3 of Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer New York

# Relating measures of intrinsic dimension

**Lemma**

$$\frac{1}{\|I + A\|}\tau \leq \mathit{efd} \leq \tau.$$

Hence,

$$\tau < \infty \iff \mathit{efd} < \infty$$

**Theorem**

*Let $\mu = \mathbb{P}_{u|y}$ and $\pi = \mathbb{P}_u$. The following are equivalent:*

i) *efd* $< \infty$;

ii) $\tau < \infty$;

iii) $\Gamma^{-1/2} Ku \in \mathcal{H}$, *$\pi$-almost surely;*

iv) *for $\nu_y$-almost all $y$, the posterior $\mu$ is well defined as a measure in $\mathcal{X}$ and is absolutely continuous with respect to the prior with*

$$\frac{d\mu}{d\pi}(u) \propto \exp\left(-\frac{1}{2}\left\|\Gamma^{-1/2}Ku\right\|^2 + \frac{1}{2}\langle \Gamma^{-1/2}y, \Gamma^{-1/2}Ku\rangle\right)$$
$$=: g(u;y),$$

*where* $0 < \pi\big(g(\cdot;y)\big) < \infty$.

**Remark**

*Notice that polynomial moments of $g$ are equivalent to re-scaling $\Gamma$ hence (among other moments)*

$$\rho = \frac{\pi\left(g(\cdot;y)^2\right)}{\pi\left(g(\cdot;y)\right)^2} < \infty \iff \tau < \infty$$

**Remark**

$$\begin{aligned}\tau &= \mathrm{Tr}\big((C^{-1} - \Sigma^{-1})\Sigma\big) = \mathrm{Tr}\big((\Sigma - C)C^{-1}\big),\\ efd &= \mathrm{Tr}\big((\Sigma - C)\Sigma^{-1}\big) = \mathrm{Tr}\big((C^{-1} - \Sigma^{-1})C\big).\end{aligned}$$

# Spectral jump

Suppose that $A$ has eigenvalues $\{\lambda_i\}_{i=1}^{d_u}$ with $\lambda_i = L \gg 1$ for $1 \leq i \leq k$, and

$$\sum_{i=k+1}^{d_u} \lambda_i \ll 1.$$

Then $\tau(A) \approx Lk$, *efd* $\approx k$ and for large $k, L$:

$$\rho \gtrapprox L^{\frac{efd}{2}}$$

hence $\rho$ grows exponentially with *number* of relevant eigenvalues, but *algebraically* with their size

# Spectral cascade

**Assumption**

*$\Gamma = \gamma I$ and that $A$ has eigenvalues $\left\{\frac{j^{-\beta}}{\gamma}\right\}_{j=1}^{\infty}$ with $\gamma > 0$, and $\beta \geq 0$. We consider a truncated sequence of problems with $A(\beta, \gamma, d)$, with eigenvalues $\left\{\frac{j^{-\beta}}{\gamma}\right\}_{j=1}^{d}$, $d \in \mathbb{N} \cup \{\infty\}$. Finally, we assume that the data is generated from a fixed underlying infinite dimensional truth $u^\dagger$,*

$$y = Ku^\dagger + \eta, \quad Ku^\dagger \in \mathcal{H},$$

*and for the truncated problems the data is given by projecting $y$ onto the first $d$ eigenfunctions of $A$.*

- $\rho$ grows **algebraically** in the small noise limit $(\gamma \to 0)$ if the nominal dimension $d$ is finite.
- $\rho$ grows **exponentially** in $\tau$ or *efd* as the nominal dimension grows $(d \to \infty)$, or as the prior becomes rougher $(\beta \searrow 1)$.
- $\rho$ grows **factorially** in the small noise limit $(\gamma \to 0)$ if $d = \infty$, and in the joint limit $\gamma = d^{-\alpha}, d \to \infty.$ The exponent in the rates relates naturally to *efd*.

# Literature pointers

- Bayesian conjugate inference with linear models and Shur dates back to [Lindley and Smith, 1972], with infinite dimensional extension in [Mandelbaum, 1984] and with precisions in [Agapiou et al., 2013]
- Bayesian formulations of inverse problems is now standard and has been popularised by [Stuart, 2010] (see however [Papaspiliopoulos et al., 2012] for early foundations in the context of SDEs)

- Absolute continuity between Gaussian measures in infinite-dimensional Hilbert spaces is not at all guaranteed; see the notion of Cameron-Martin space and the so-called Feldman-Hajek theorem [Da Prato and Zabczyk, 1992]. It is common in the literature to assume conditions under which prior and posterior are equivalent, hence there exists a likelihood. Our theorem shows that they are equivalent to assuming finite intrinsic dimension!

# Outline

1. Importance sampling
2. Linear inverse problems & intrinsic dimension
3. Dynamic linear inverse problems: sequential IS
4. Outlook

# Setting (first step towards data assimilation)

$$
\begin{aligned}
v_1 &= M v_0 + \xi, \quad v_0 \sim N(0, P), \quad \xi \sim N(0, Q), \\
y_1 &= H v_1 + \zeta, \quad \zeta \sim N(0, R).
\end{aligned}
$$

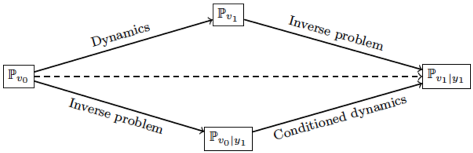


FIG 1. *Filtering step decomposed in two different ways. The upper path first pushes forward the measure $\mathbb{P}_{v_0}$ using the signal dynamics, and then incorporates the observation $y_1$. The lower path assimilates the observation $y_1$ first, and then propagates the conditioned measure using the signal dynamics. The standard proposal corresponds to the upper decomposition and the optimal one to the lower decomposition.*

Behaviour of filtering model determined by inverse problem

$$y_1 = Ku + \eta, \quad u \sim \mathbb{P}_u, \quad \eta \sim N(0, \Gamma),$$

with

- $K = (0,\ H)$, $\Gamma = R$ for standard proposal
- $K = (HM,\ 0)$, $\Gamma = R + HQH^*$ for locally optimal proposal

**Theorem**

$$\tau_{st} \geq \tau_{op}$$

For example, if $H = Q = R = M = I$ but $\mathrm{Tr}(P) < \infty$, then $\tau_{op} < \infty$ and $\tau_{st} = \infty$:

- Inverse problems perspective: prior is regularising but if propagated not so, hence a bad inverse problem
- State-space model perspective: very informative data! Predictive distribution is singular with respect to filter

# Literature pointers

- One-step filtering is only analysed for simplicity. It is however a necessary step for PF. This is done in various recent works, e.g. [Bengtsson et al., 2008]. [Chorin and Morzfeld, 2013] consider filters initialised at stationary covariances; they also define a notion of intrinsic dimension of a data assimilation problem as the Frobenius norm of this covariance, which is at odds with both $\tau$ and *efd*, and does not seem to be appropriate for characterising stability of PFs
- Optimal proposal is only locally optimal in multi-step problems, although it has some interesting characterisations, see [Chopin and Papaspiliopoulos, 2016].

# Outline

1. Importance sampling
2. Linear inverse problems & intrinsic dimension
3. Dynamic linear inverse problems: sequential IS
4. Outlook

# Outlook

Degrees of freedom have been defined for non-linear Bayesian hierarchical model - see DIC of [Spiegelhalter et al., 2002]. It is thus natural to try and extend this work for nonlinear inverse problems, and this might be a real advantage of *efd* vs $\tau$

The formulation of MCMC algorithms on Hilbert spaces provided a whole new set of tools for designing and analysing theoretically algorithms, see e.g. the recent [Cotter et al., 2013]. We see this work as the importance sampling analogue. The conversion of some of the understanding to new algorithms is a priority

Very similar ideas are being developed for deterministic and quasi Monte Carlo integration, see e.g. [Kuo and Sloan, 2005]

- Agapiou, S., Larsson, S., and Stuart, A. M. (2013).
Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems.
*Stochastic Processes and their Applications*, 123(10):3828–3860.
- Bengtsson, T., Bickel, P., Li, B., et al. (2008).
Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems.
In *Probability and statistics: Essays in honor of David A. Freedman*, pages 316–334. Institute of Mathematical Statistics.
- Bickel, P., Li, B., Bengtsson, T., et al. (2008).
Sharp failure rates for the bootstrap particle filter in high dimensions.
In *Pushing the limits of contemporary statistics: Contributions in honor of Jayanta K. Ghosh*, pages 318–329. Institute of Mathematical Statistics.
- Bishop, C. M. (2006).
*Pattern recognition and machine learning*.
Springer New York.
- Boucheron, S., Lugosi, G., and Massart, P. (2013).
*Concentration inequalities*.
Oxford University Press, Oxford.
- Chatterjee, S. and Diaconis, P. (2015).
The sample size required in importance sampling.
*arXiv preprint arXiv:1511.01437*.
- Chopin, N. (2004).
Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference.
*Annals of statistics*, pages 2385–2411.
- Chopin, N. and Papaspiliopoulos, O. (2016).
*A concise introduction to sequential Monte Carlo*.
- Chorin, A. J. and Morzfeld, M. (2013).
Conditions for successful data assimilation.
*Journal of Geophysical Research: Atmospheres*, 118(20):11–522.
- Cotter, S. L., Roberts, G. O., Stuart, A. M., and White, D. (2013).
MCMC methods for functions: modifying old algorithms to make them faster.
*Statistical Science*, 28(3):424–446.

- Da Prato, G. and Zabczyk, J. (1992).
  *Stochastic equations in infinite dimensions*.
  Cambridge university press.
- Del Moral, P. (2004).
  *Feynman-Kac Formulae*.
  Springer.
- Del Moral, P. and Miclo, L. (2000).
  *Branching and interacting particle systems approximations of Feynman-Kac formulae with applications to non-linear filtering*.
  Springer.
- Doukhan, P. and Lang, G. (2009).
  Evaluation for moments of a ratio with application to regression estimation.
  *Bernoulli*, 15(4):1259–1286.
- Downey, P. J. and Wright, P. E. (2007).
  The ratio of the extreme to the sum in a random sequence.
  *Extremes*, 10(4):249–266.
- Kuo, F. Y. and Sloan, I. H. (2005).
  Lifting the curse of dimensionality.
  *Notices of the AMS*, 52(11):1320–1328.
- Lindley, D. V. and Smith, A. F. M. (1972).
  Bayes estimates for the linear model.
  *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–41.
- Mandelbaum, A. (1984).
  Linear estimators and measurable linear transformations on a Hilbert space.
  *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 65(3):385–397.
- McLeish, D. L. and O'Brien, G. L. (1982).
  The expected ratio of the sum of squares to the square of the sum.
  *Ann. Probab.*, 10(4):1019–1028.
- Papaspiliopoulos, O., Pokern, Y., Roberts, G., and Stuart, A. (2012).
  Nonparametric estimation of diffusions: A differential equations approach.
  *Biometrika*, 99(3):511–531.

- Rebeschini, P. and van Handel, R. (2013).
  Can local particle filters beat the curse of dimensionality?
  *arXiv preprint arXiv:1301.6585*.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002).
  Bayesian measures of model complexity and fit.
  *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(4):583–639.
- Stuart, A. M. (2010).
  Inverse problems: a Bayesian perspective.
  *Acta Numerica*, 19:451–559.