The Intrinsic Dimension of Importance Sampling

Sergios Agapiou

Department of Statistics, University of Warwick

Joint work with: **O. Papaspiliopoulos, D. Sanz-Alonso and A. M. Stuart**

Reading-Warwick Data Assimilation Day, June 23, 2015

Enabling Quantification of

EQUIP

Uncertainty for Inverse Problems

Three related problems

Bayesian Inverse Problems

$$\left.\begin{array}{ll} Prior: & u \sim {\color{blue}\mu_0} \\ Data: & y = \mathcal{G}(u) + \eta \end{array}\right\} \quad Posterior: \quad u|y \sim {\color{red}\mu^y}$$

Filtering

$$\begin{tikzcd}
& \boxed{v_{j+1}|Y_j} \arrow[dr, "\text{Inverse problem}"] & \\
\boxed{v_j|Y_j} \arrow[ur, "\text{Signal dynamics}"] \arrow[dr, "\text{Inverse problem}"'] \arrow[rr, dashed] & & \boxed{{\color{red} v_{j+1}|Y_{j+1}}} \\
& \boxed{v_j|Y_{j+1}} \arrow[ur, "\text{Signal dynamics}"'] &
\end{tikzcd}$$

General Framework

$${\color{red}\mu}(du) \propto g(u){\color{blue}\pi}(du), \quad unknown\ normalizing\ constant\ {\color{blue}\pi}(g)^{-1}$$

Table of Contents

General Framework: Autonormalized IS

- **Aim**: estimate expectations of functions $f : \mathcal{X} \to \mathbb{R}$ wrt probability measure μ.
- **Challenge**: can only access π, $\mu(du) \propto g(u)\pi(du)$.
- Implicitly assume $0 < \pi(g) < \infty$.

$$\mu(f) = \frac{\pi(fg)}{\pi(g)} \approx \frac{\frac{1}{N}\sum_{i=1}^N f(u^i)g(u^i)}{\frac{1}{N}\sum_{j=1}^N g(u^j)}, \quad u^i \sim \pi,$$
$$= \sum_{i=1}^N w^i f(u^i) =: \mathcal{I}^N(f),$$

where

$$w^i = \frac{g(u^i)}{\sum_j g(u^j)}.$$

- $\mathcal{I}^N(f)$ biased estimator of $\mu(f)$. SLLN suggests consistent.

Non-asymptotic Theorem 1; $\mu(du) \propto g(u)\pi(du)$

$$\rho := \frac{\pi(g^2)}{\pi(g)^2} \in [1, \infty]$$

- Assume $\pi(g^2) < \infty$ st $\rho < \infty$.

Theorem (A., Papaspiliopoulos, Sanz-Alonso, Stuart '15)

Bias: $$\sup_{|f|\le 1} \left|\mathbb{E}[\mathcal{I}^N(f) - \mu(f)]\right| \le 12\frac{\rho}{N}$$

MSE: $$\sup_{|f|\le 1} \mathbb{E}\left[\left(\mathcal{I}^N(f) - \mu(f)\right)^2\right] \le 4\frac{\rho}{N}$$

- Minimal assumptions on g, strong assumptions on f.

Non-asymptotic Theorem 2; $\mu(du) \propto g(u)\pi(du)$

Assume

- $\pi(g^k) < \infty$ for all $k \geq 2$ (often holds).
- $\pi(|f|^{2+\epsilon}) < \infty$, for $\epsilon > 0$ arbitrarily small.

Theorem (A., Papaspiliopoulos, Sanz-Alonso, Stuart '15)

$$\text{Bias:} \qquad \left|\mathbb{E}[\mathcal{I}^N(f) - \mu(f)]\right| \leq \frac{C_{\text{bias}}}{N}$$

$$\text{MSE:} \qquad \mathbb{E}\left[\left(\mathcal{I}^N(f) - \mu(f)\right)^2\right] \leq \frac{C_{\text{MSE}}}{N}$$

- Constants only involve π-moments of f, g and fg.
- Strong assumptions on g, minimal assumptions on f.

Comments

- Theorem 2 generalizes: conjugate assumptions on f and g.
- Recall

$$\mathcal{I}^N(f) = \frac{\frac{1}{N}\sum_{i=1}^N f(u^i)g(u^i)}{\frac{1}{N}\sum_{j=1}^N g(u^j)} =: \frac{\pi^N(fg)}{\pi^N(g)}$$

- Decomposition

$$\mathcal{I}^N(f) - \mu(f) = \frac{\pi^N(fg)}{\pi^N(g)} - \frac{\pi(fg)}{\pi(g)}$$

$$= \frac{\pi^N(fg) - \pi(fg)}{\pi(g)} + \pi^N(fg)\Big(\frac{1}{\pi^N(g)} - \frac{1}{\pi(g)}\Big)$$

- Theorem 2 requires careful handling of 2nd term: follow DL09.
- Marcinkewicz-Zygmund inequality

$$\|\pi^N(h) - \pi(h)\|_t \le C_t \|h(u_1) - \pi(h)\|_t N^{-\frac{1}{2}}, \quad \forall t \ge 2.$$

The ratio $\rho = \pi(g^2)/\pi(g)^2$

- ρ captures the variance of the weights w^i.
- Appears in Theorem 1: smaller ρ better error estimates.
- *Effective Sample Size*

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

- For large N, SLLN gives

$$ESS(N) \approx \frac{N}{\rho}.$$

- For efficient IS need small ρ.

Table of Contents

IS for Linear Bayesian Inverse Problems

- $\mathcal{X}, \mathcal{Y}$ separable Hilbert spaces.
- Interested in recovering $u \in \mathcal{X}$ from noisy indirect data $y \in \mathcal{Y}$.

Bayesian Inverse Problems

$$\left.\begin{aligned} &Prior\ (proposal): \ u \sim \mu_0 = N(0, \sigma\Sigma) \\ &Data: \ y = Ku + \eta \in \mathcal{Y}, \quad \eta \sim N(0, \gamma\Gamma) \end{aligned}\right\} \ Posterior: \ u|y \sim \mu^y$$

- Sensible notion of dimension? When $\mu_0 \ll \mu^y$? Size of ρ?

IS for Linear Bayesian Inverse Problems

- $\mathcal{X}, \mathcal{Y}$ separable Hilbert spaces.
- Interested in recovering $u \in \mathcal{X}$ from noisy indirect data $y \in \mathcal{Y}$.

Bayesian Inverse Problems

$$\left.\begin{array}{l} \textit{Prior (proposal)}: \; u \sim \mu_0 = N(0, \sigma\Sigma) \\ \textit{Data}: \; y = Ku + \eta \in \mathcal{Y}, \quad \eta \sim N(0, \gamma\Gamma) \end{array}\right\} \textit{Posterior}: \; u|y \sim \mu^y$$

- Sensible notion of dimension? When $\mu_0 \ll \mu^y$? Size of ρ?

Key: how informative the data is relative to the prior

- eigenvalues of $A := \Sigma^{1/2} K^* \Gamma^{-1} K \Sigma^{1/2}$
- value of $\lambda := \gamma/\sigma$

Two notions of effective dimension: efd and τ

$$\boxed{\tau := \frac{1}{\lambda}\mathrm{Tr}\,(A)} \quad \boxed{\mathrm{efd} := \mathrm{Tr}\left((\lambda I + A)^{-1}A\right)}$$

Motivation for $\boldsymbol{\tau}$: "collapse" of IS occurs iff $\tau = \infty$, BBL08.

Motivation for efd: Machine learning and SIP, Z02, LM14.

- Different behaviour as $\lambda \to 0$ (small noise compared to prior scaling).
- τ does not capture behaviour of A as $\lambda \to 0$.

Connection between τ, efd, ρ and $\mu^y \ll \mu_0$

Theorem (A., Papaspiliopoulos, Sanz-Alonso, Stuart '15)

Let $\nu(du, dy) = \mathbb{P}(dy|u)\mu_0(du)$ and assume A bdd. The following are equivalent:

i) $\text{efd} < \infty$.

ii) $\tau < \infty$.

iii) $\|\Gamma^{-\frac{1}{2}} K u\| < \infty$, μ_0-almost surely.

iv) For ν-a.a. y, μ^y is absolutely continuous w.r.t. μ_0 and

$$\frac{d\mu^y}{d\mu_0}(u) \propto \exp\left(-\frac{\gamma}{2}\|Ku\|_\Gamma^2 + \gamma\langle y, Ku\rangle_\Gamma\right) =: g(u; y)$$

with

$$0 < \mu_0\big(g(\cdot; y)\big) < \infty.$$

v) It holds $0 < g(u; y) < \infty$ ν-a.s. and for ν-a.a. y

$$\rho := \frac{\mu_0\left(g(\cdot; y)^2\right)}{\mu_0(g(\cdot; y))^2} < \infty.$$

Connection between τ, efd, ρ and $\mu^y \ll \mu_0$

Theorem (A., Papaspiliopoulos, Sanz-Alonso, Stuart '15)

Let $\nu(du, dy) = \mathbb{P}(dy|u)\mu_0(du)$ and assume A bdd. The following are equivalent:

i) $\text{efd} < \infty$.

ii) $\tau < \infty$.

iii) $\|\Gamma^{-\frac{1}{2}}Ku\| < \infty$, μ_0-almost surely.

iv) For ν-a.a. y, μ^y is absolutely continuous w.r.t. μ_0 and

$$\frac{d\mu^y}{d\mu_0}(u) \propto \exp\left(-\frac{\gamma}{2}\|Ku\|_\Gamma^2 + \gamma\langle y, Ku\rangle_\Gamma\right) =: g(u; y)$$

with

$$0 < \mu_0\big(g(\cdot; y)\big) < \infty.$$

v) It holds $0 < g(u; y) < \infty$ ν-a.s. and for ν-a.a. y

$$\rho := \frac{\mu_0\left(g(\cdot; y)^2\right)}{\mu_0(g(\cdot; y))^2} < \infty.$$

Diagonal Inverse Problems

$$y = Ku + \eta, \quad \eta \sim N(0, \gamma\Gamma), \quad u \sim N(0, \sigma\Sigma), \quad \lambda := \gamma/\sigma.$$

Assumption

- K^*K, Γ and Σ commute, hence diagonal in same basis.
- Eigenvalues of $A = \Sigma^{\frac{1}{2}} K^* \Gamma^{-1} K \Sigma^{\frac{1}{2}} : \{j^{-\beta}\}_{j=1}^{\infty}, \quad \beta \geq 0.$
- Sequence of d-dim problems corresponding to A_d with eigenvalues $\{j^{-\beta}\}_{j=1}^{d}$.

$$\tau = \tau(d, \lambda, \beta), \quad \mathrm{efd} = \mathrm{efd}(d, \lambda, \beta), \quad \rho = \rho(d, \lambda, \beta).$$

$$\tau(\infty, \lambda, \beta) = \frac{1}{\lambda} \sum_{j=1}^{\infty} j^{-\beta} < \infty \iff \beta > 1 \iff \mu_{\infty}^{y} \ll \mu_{0,\infty}.$$

Diagonal Inverse Problems

Theorem (A., Papaspiliopoulos, Sanz-Alonso, Stuart '15)

- Let $\beta > 1$ and $\lambda > 0$ fixed. As $d \to \infty$,

$$\rho(d, \lambda, \beta) \nearrow \rho(\infty, \lambda, \beta) < \infty.$$

- Let $\beta > 1$, $d = \infty$. As $\lambda \to 0$, $\mathrm{efd}(\lambda) \approx \lambda^{-1/\beta}$ and

$$\mathbb{P}\Big[\rho(\lambda) \geq c_1 \exp\big(c_2\, \mathrm{efd}(\lambda)\big)\Big] \longrightarrow 1. \qquad \text{(small noise)}$$

- Let $0 \leq \beta \leq 1$ and $\lambda > 0$ fixed. As $d \to \infty$, $\mathrm{efd}(d) \approx d^{1-\beta}$ and

$$\mathbb{P}\Big[\rho(d) \geq c_1 \exp\big(c_2\, \mathrm{efd}(d)\big)\Big] \longrightarrow 1. \qquad \text{(large } d\text{)}$$

efd is the universally important quantity.

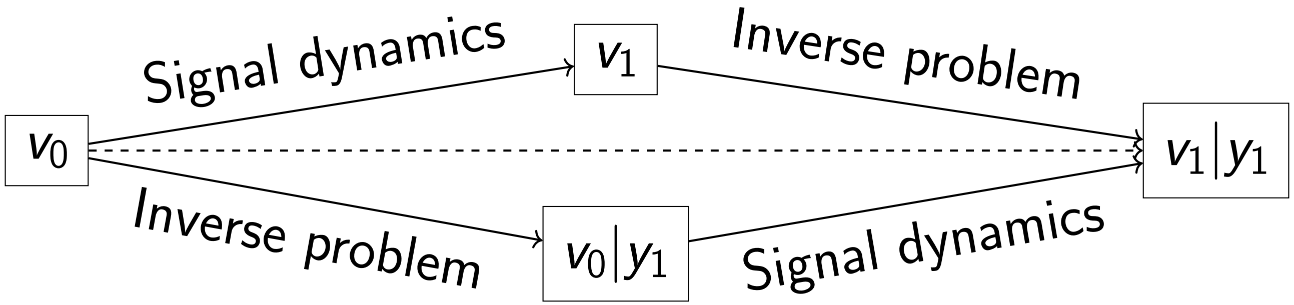
Table of Contents

IS for Filtering

Filtering

Signal: $v_1 = Mv_0 + N(0, Q), \qquad v_0 \sim N(0, P) = \mathbb{P}_0.$

Data: $y_1 = Hv_1 + N(0, R)\ .$ Target: $\mathbb{P}(u|y_1)$, $u = (v_0, v_1)$.



Standard proposal: $\pi_{st}(du) := \mathbb{P}_0(dv_0)\mathbb{P}(dv_1|v_0)$.

Optimal proposal: $\pi_{op}(du) := \mathbb{P}_0(dv_0)\mathbb{P}(dv_1|v_0, y_1)$.

IS collapse props for two proposals relate to collapse props of corresponding inverse problem.

Table of Contents

Highlights

- **General framework**:
  - Balance between assumptions on test function and change of measure.
- **Linear inverse problem**:
  - Introduced adequate notion of dimension.
  - Showed its relevance for importance sampling.
  - Emphasized the importance of absolute continuity.
- **Filtering**: extend analysis.

http://www.sergiosagapiou.com

- S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, A. M. Stuart, *The intrinsic dimension of importance sampling*, in preparation.
- P. Rebeschini, R. van Handel, *Can local particle filters beat the curse of dimensionality?*, Annals of Applied Probability, 2015.
- P. Doukhan, G. Lang, *Evaluation of moments of a ratio with application to regression estimation*, Bernoulli, 2009.
- T. Bengtsson, P. Bickel, B. Li, *Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems*, Probability and Statistics: essays in honor of David. A. Freedman, 2008.
- T. Zhang, *Effective dimension and generalization of kernel learning*, Advances in Neural Information Processing Systems, 2002.
- S. Lu, P. Mathé, *Discrepancy based model selection in statistical inverse problems*, Journal of Complexity, 2014.