Uncertainty Quantification and Performance guarantees for stochastic processes

Luc Rey-Bellet

University of Massachusetts Amherst

Ki-Net at Duke, April 2019

Work supported by **NSF** and **AFOSR**

Collaborators on this and related projects

- Paul Dupuis (Brown),
- Sung-Ha Hwang (KAIST)
- Markos Katsoulakis (UMass Amherst)
- Yannis Pantazis (FORTH Crete)
- Jeremiah Birrell (UMass Amherst)
- Konstantinos Gourgoulias (UMass Amherst)
- Jinchao Feng (UMass Amherst)
- Jie Wang (UMass Amherst)
- Sosung Baek (KAIST)

- More coming.....

- Performance guarantees for hypocoercive MCMC samplers, by Jeremiah Birrell, Luc Rey-Bellet. (In preparation)

- Uncertainty Quantification for Markov Processes via Variational Principles and Functional Inequalities, by Jeremiah Birrell, Luc Rey-Bellet (submitted)

- , Sensitivity Analysis for Rare Events based on Rnyi Divergence, by Paul Dupuis, Markos A. Katsoulakis, Yannis Pantazis, Luc Rey-Bellet (to appear in Annals of Applied Probability)

- How biased is your model? Concentration Inequalities, Information and Model Bias, by Konstantinos Gourgoulias, Markos A. Katsoulakis, Luc Rey-Bellet, Jie Wang (To appear in IEEE, Transactions on Information Theory)

- Scalable Information Inequalities for Uncertainty Quantification, by Markos A. Katsoulakis, Luc Rey-Bellet, Jie Wang (J. Comp. Phys.)

Basic question: Uncertainty quantification

$\to$ **Baseline model P (= probability measure on $\mathcal{X}$)**. Think of it as a (tractable) model you use to compute or calculate.

NOT TO BE TRUSTED!!

$\to$ **Quantities of interests (QoI)** such as

- $E_P[f]$ (Expectation)
- $\mathrm{Var}_P(f)$ (Variance) or $\frac{\mathrm{Cov}_P(f,g)}{\sqrt{\mathrm{Var}_P(f)\mathrm{Var}_P(g)}}$ (correlation), ...
- $\Lambda_{P,f}(c) = \log E_P[e^{cf}]$ (risk sensitive functional)
- $\log P(A)$ (probability of some rare event)
- and so on

$\rightarrow$ Family of **alternative models** Q. Think of it as describing the true but unknowable model. Set

$$\mathcal{Q}_\eta \quad = \quad \{Q \text{ is } \eta \text{ "close" to } P\}$$

Think of something like

$$\mathcal{Q}_\eta = \{Q \,:\, R(Q||P) \leq \eta\} \qquad R(Q||P) = E_Q\left[\log \frac{dQ}{dP}\right] \text{ relative entropy}$$

It measures the allowed information loss.

Given an observable quantity f can one find **uncertainty bounds** or performance guarantees

$$\inf_{Q \in \mathcal{Q}_\eta} \mathbf{E}_Q[f] \leq \mathbf{E}_P[f] \leq \sup_{Q \in \mathcal{Q}_\eta} \mathbf{E}_Q[f]\,.$$

$\rightarrow$ **Robustness** , Book by Hansen (Nobel 2011) and Sargent (Nobel 2013)
$\rightarrow$ Operation research, Finance, etc.... $\rightarrow$

The bounds should be **tight** and **computable** (numerically or analytically).

Challenge: Scalable bounds for probabilities on high-dimensional spaces

Long-time regime ($T \to \infty$) : Typical example: two ergodic Markov processes X_t and Y_t with path space measures $P_{0:T}$ and $Q_{0:T}$ and stationary measures μ_P and μ_Q

In this case we assume there is rate of information loss

$$\frac{1}{T} R(Q_{0:T}||P_{0:T}) \to r(Q||P)$$

We want steady states UQ bounds, control e.g. on

$$E_{\mu_P}[f] - E_{\mu_Q}[f]$$

especially if μ_P and/or μ_Q is not know explicitly

Seemingly unrelated: performance guarantees for sampling

Think of a MCMC where $\mu = \mu_P$ is your target distribution sampled using X_t and we are trying to evaluate

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T f(X_s) = \int f d\mu \text{ with } X_0 \sim \mu_0$$

How do we evaluate the performance of the Markov process X_t starting for the initial measure μ_0 as a MCMC algorithm?

- Practical: use the sample variance

$$T \mathrm{Var}_{P_{\mu_0}} \left[\frac{1}{T} \int_0^T f(X_s) \right]$$

to build asymptotic confidence intervals using central limit theorem.

Drawback: how do you choose T to be in the CLT regime...

• Mixing times: Use spectral gaps estimates to compute mixing times (need explicit constants). Geometric ergodicity, L^2 estimates, etc....

Explicit bounds on $\mathrm{dist}(\mu_T, \mu)$ where $X_T \sim \mu_T$

Drawback: in practice we often do not sample μ_T but use ergodic averages (empirical measure)

• Concentration inequalities (My favorite for today). Construct explicit rigorous finite T confidence intervals using concentration inequalities such as Bernstein type inequalities

$$P_{\mu_0}\left(\frac{1}{T}\int_0^T f(X_s) - \int f d\mu > r\right) \le \left\|\frac{d\mu_0}{d\mu}\right\|_{L^2(\mu)} \exp\left(-t\frac{r^2}{2(\sigma^2 + Mr)}\right)$$

with explicit constants σ^2 and M.

YES : Obtain explicit performance guarantees if we use finite time samples. But it may be too pessimistic.

What's wrong with CKP? Scalability

Czsizar-Kullback-Pinsker

$$|E_Q[f] - E_P[f]| \le \sqrt{2R(Q||P)}\ \ \|f - E_P[f]\|_\infty$$

Take Markov measures $P = P^{0:T}$ and $Q = Q^{0:T}$ on the time window $[0, T]$ and

$$F_T = \frac{1}{T}\int_0^T f(X_s)\, ds\,.$$

Then we have

$$\|F_T\|_\infty = \|f\|_\infty = O(1) \text{ and } R(Q^{0:T}||P^{0:T}) = O(T)$$

CKP scales terribly poorly with T, the LHS is $O(1)$ but the RHS diverges like $\sqrt{T}$.

But

$$\text{Var}_{P^{0:T}}[F_T] = O\left(\frac{1}{T}\right)$$

so one would need the variance instead of the sup norm.

Gibbs Variational principle

- Relative entropy (a.k.a Kullback-Leibler divergence).

$$R(Q \,||\, P) = \begin{cases} E_Q\left[\log \frac{dQ}{dP}\right] & \text{if } Q \ll P \\ +\infty & \text{otherwise} \end{cases}$$

$R(Q \,||\, P)$ is a divergence, that is $R(Q \,||\, P) \geq 0$ and $R(Q \,||\, P) = 0$ if and only if $Q = P$.

- Gibbs variational principle for the relative entropy: (convex duality).

$$\log E_P\left[e^f\right] = \sup_Q \{E_Q[f] - R(Q||P)\}$$

with the supremum attained if and only if

$$dQ = dQ^f = \frac{e^f dP}{E_P[e^f]}$$

Gibbs information inequality

From the Gibbs variational principle, for any Q and $c \geq 0$

$$\mathbf{E}_Q[\pm cf] \leq \log \mathbf{E}_P\left[e^{\pm cf}\right] + R(Q||P)\,.$$

Optimize over c:

Theorem (Gibbs Information inequality)

$$-\underbrace{\inf_{c>0}\left\{\frac{\Lambda(-c)+R(Q||P)}{c}\right\}}_{=\Xi_{\mathbf{P},-\mathbf{f}}(\mathbf{R}(\mathbf{Q}||\mathbf{P}))} \leq \mathbf{E}_{\mathbf{Q}}[\mathbf{f}] - \mathbf{E}_{\mathbf{P}}[\mathbf{f}] \leq \underbrace{\inf_{c>0}\left\{\frac{\Lambda(c)+R(Q||P)}{c}\right\}}_{=\Xi_{\mathbf{P},\mathbf{f}}(\mathbf{R}(\mathbf{Q}||\mathbf{P}))}$$

$$\Xi_{P,f}(\eta) \equiv \inf_{c>0}\left\{\frac{\Lambda(c)+\eta}{c}\right\} \qquad \Lambda(c) = \log \mathbf{E}_P\left[e^{c(f-\mathbf{E}_P[f])}\right]$$

How good is it? (Long history... Dupuis; Bobkov; Boucheron, Lugosi. Massart; Breuer,Czizsar, etc...)

Some convex analysis: UQ vs LDP

• Given $f : \mathcal{X} \to \mathbb{R}$ in $L^1(P)$ consider the centered cumulant generating function

$$\Lambda(c) = \log \mathbf{E}_P \left[e^{c(f - \mathbf{E}_P[f])} \right]$$

This is a convex function which we **assume** to be finite in a nbd of 0.

• Legendre-Fenchel transform

$$\Lambda^*(x) = \sup_c \{xc - \Lambda(c)\}$$

This is the rate function in Cramer's theorem and $\Lambda^*(x) \geq 0$ and $= 0$ iff $x = 0$.

• Inverse function (two branches) (Fenchel-Young)

$$(\Lambda^*)^{-1}_{\pm}(\eta) = \inf_{c \geq 0} \left\{ \frac{\Lambda(\pm c) + \eta}{c} \right\}$$

Key role in UQ!

Properties of the Gibbs information inequality

- $\Xi_{P,f}(R(Q||P)$ is a **divergence**, i.e.

$\Xi_{P,f}(R(Q||P)) \geq 0$ and $\Xi_{P,f}(R(Q||P)) = 0$ if and only if $Q = P$ or $f = const$

- **Tightness I**: Family of alternative models

$$\mathcal{Q}_\eta = \{Q\,;\, R(Q||P) \leq \eta\}$$

There exists a maximizing measure $Q_\eta \in \mathcal{Q}_\eta$ such that

$$\sup_{Q \in \mathcal{Q}_\eta} E_Q[f] - E_P[f] = E_{Q_\eta}[f] - E_P[f] = \Xi_{P,f}(\eta)$$

Moreover Q_η has the form (Cramer's tilting)

$$\frac{dQ_\eta}{dP} = \frac{e^{c(\eta)f}}{E_P[e^{c(\eta)f}]} \text{ with } c \quad \text{such that} \quad R(Q_\eta||Q) = \eta$$

• (**Tightness II**) Given P and Q assumed to be mutually absolutely continuous then for

$$f = \log \frac{dQ}{dP}$$

we have

$$E_Q[f] - E_P[f] = R(Q||P) + R(P||Q) = \Xi_{P,f}(R(Q||P))$$

(symmetrized relative entropy)

• **Linearization** For small η

$$\Xi_{P,f}(\eta) = \sqrt{2\mathrm{Var}_P[f]\eta} + \frac{1}{3}\sqrt{\mathrm{Var}_P[f]}S(f)\eta + O(\eta^{3/2})$$

where $S(f) = \frac{E[|f-E_P[f]|^3]}{\mathrm{Var}_P[f]^{3/2}}$ is the skewness.

Making it computable with concentration inequalities

Some examples: (Much more in Gourgoulias, Katsoulakis, R.-B., Wang).

- If $a \le f \le b$ we have Hoeffding's inequality

$$\Lambda(c) \le \frac{c^2(b-a)^2}{8} \le \frac{c^2\|f - \mathbf{E}_P[f]\|_\infty}{2}$$

and then

$$\Xi_{P,f}(\eta) \le \sqrt{2\eta}\|f - \mathbf{E}_P[f]\|_\infty \quad \text{(Cziszar-Kullback Pinsker).}$$

- If f is bounded and $\mathrm{Var}_P[f] = \sigma^2$ then we have Bernstein inequality

$$\Lambda(c) \le \frac{c^2\sigma^2}{2(1 - c\|f - \mathbf{E}_P[f]\|_\infty)}$$

and then

$$\Xi_{P,f}(\eta) \le \sqrt{2\mathrm{Var}_P[f]\eta} + \|f - \mathbf{E}_P[f]\|_\infty \eta$$

This beats Pinsker if η is not too big (especially if σ^2 is small) and captures the exact small η asymptotics.

- Many more.....

Scalability for ergodic Markov processes

Baseline process:

- Ergodic continuous time Markov process X_t on state space $\mathcal{X}$
- path-space measure $P_{\mu_0}^{0:T}$ and with stationary distribution μ.
- Infinitesimal generator $\mathcal{L}$ (acting on $L^2(\mu)$).

Alternative process:

- Ergodic continuous time stochastic process Y_t on state space $\mathcal{X}$ (not necessarily Markovian!).
- path-space measure $Q_{\nu_0}^{0:T}$ with $Q_{\nu_0}^{0:T} \ll P_{\mu_0}^{0:T}$ and assume that

  $$r(Q||P) = \lim_{T\to\infty} \frac{1}{T} R(Q_{\nu_0}^{0:T}||P_{\mu_0}^{0:T})$$ relative entropy rate exists

Steady state UQ bounds for ergodic Markov processes

Consider ergodic averages $\frac{1}{T}\int_0^T f(X_s)\,ds$ then using the Gibbs UQ bound one the steady state bias bound

$$\xi_{P,-f}(r(Q||P)) \leq \underbrace{\lim_{T\to\infty}\frac{1}{T}\int_0^T f(Y_s)ds}_{\text{true process}} - \underbrace{E_\mu[f]}_{\text{baseline}} \leq \xi_{P,f}(r(Q||P))$$

where

$$\xi_{P,f}(\eta) = \inf_{c>0}\left\{\frac{\lambda(c)+\eta}{c}\right\}$$

$$\lambda(c) = \lim_{T\to\infty}\frac{1}{T}\log E_{P^{0:T}_{\mu_0}}\left[\exp\left(c\int_0^T (f(X_s)-E_\mu[f])ds\right)\right]$$

- (Linearization:) Under suitable assumptions one can linearize

$$\xi_{P,f}(r(Q||P)) = \sqrt{2\sigma^2(f) r(Q||P)} + O(r(Q||P))$$

where $\sigma^2(f)$ is the asymptotic variance (CLT)

$$\sigma^2(f) = 2\int_0^\infty \langle (f - E_\mu[f]) \, , \, e^{\mathcal{L}t}(f - E_\mu[f]) \rangle_{L^2(\mu)} \, .$$

• Main idea is to consider the Feynmann-Kac semi group

$$e^{T(\mathcal{L}+V)}h(x) = E_{P^{0:T}_{\delta_x}}\left[e^{\int_0^T V(X_s)ds}h(X_t)\right]$$

and to use the (finite T!) bound using Lumer-Philips Theorem Liming Wu valid also for non-symmetric generators

$$\frac{1}{T}\log\|e^{T(\mathcal{L}+V)}\|_{L^2(\mu)} \leq \sup\left\{\langle g\,,\,\mathcal{L}g\rangle_{L^2(\mu)} + \int V|g|^2 d\mu\,,\,\|g\|^2 = 1\right\}.$$

to derive we use concentration inequalities for Markov process .

We relie then on results from Wu, and Cattiaux , Guillin, and Guillin, Leonard, Wu, Yao, and Gao, Guillin, Wu, going back to Villani and many others.

Poincaré inequalities and bounded f

Assume a Poincaré inequality (spectral gap)

$$\mathrm{Var}_\mu[f] \le -\alpha \langle f\,,\,\mathcal{L}f\rangle_{L^2(\mu)}\,, \quad f \in D(\mathcal{L})$$

• Theorem: For bounded f and general $\mathcal{L}$ a functional analytic lemma gives ($\widetilde{f} = f - E_\mu[f]$)

$$\lambda(c) \le \frac{c^2 \alpha \mathrm{Var}_\mu[\mathrm{f}]}{1 - \alpha c \|\widetilde{f}\|_\infty}$$

$$\xi_{P,f}(\eta) \le 2\sqrt{\alpha \mathrm{Var}_\mu[f]\eta} + \alpha \|\widetilde{f}\|_\infty \eta$$

• Theorem: For bounded f and symmetric $\mathcal{L}$ we can use the asymptotic variance

$$\lambda(c) \leq \frac{c^2\sigma^2(f)}{2(1-\alpha c\|\widetilde{f}\|_\infty)}$$

and thus

$$\xi_{P,f}(\eta) \leq \sqrt{2\sigma^2(f)\eta} + \alpha\|\widetilde{f}\|_\infty \eta$$

(This is sharp for small η).

Log-Sobolev inequalities and unbounded f

Assume a stronger Log-Sobolev inequality

$$\mathbf{E}_\mu[f^2 \log(f^2)] - \mathbf{E}_\mu[f^2] \log \mathbf{E}_\mu[f^2] \leq -\beta\langle f\,, \mathcal{L} f\rangle \quad f \in D(\mathcal{L})$$

Then using the Gibbs variational principle get the bound

$$\begin{aligned} \xi_{P,f}(\eta) &= \inf_{c>0} \left\{ \frac{\log E_\mu \left[e^{c(f - E_\mu[f])} \right]}{c} + \frac{\beta\eta}{c} \right\} \\ &= \sqrt{2\beta \mathrm{Var}_\mu[f]\eta} + O(\eta) \qquad (1) \end{aligned}$$

and we can work another round of concentration inequalities to obtain explicit constants depending on the tails of μ and f. It is all reduced to the steady state, no more dynamics!.

Example

Langevin equation

$$dX = -\nabla V + J\nabla V + \sqrt{2}dW_t$$

for any any antisymmetric J has invariant measure $d\mu = e^{-V}dx$ and we have

$$\mathcal{L} = \underbrace{\Delta - \nabla V \nabla}_{\text{symmetric}} + \underbrace{J\nabla V \nabla}_{\text{antisymmetric}}$$

Assume $V(x) \sim |x|^\beta$

- Spectral gap for $\beta > 1$
- Log Sobolev for $\beta > 2$ so UQ bounds for $V(X)$ itself.

For $1 < b \leq 2$ we can use F- Sobolev inequalities to consider unbounded f.

Hypocoercive samplers

Goal: To sample from $\nu(dq) \propto e^{-\beta V(q)}dq$ extending the phase space and sample from the measure

$$\mu(dp, dq) = \nu(dq)\pi(dp) \propto e^{-\beta(V(q)+p^2/2m)}dpdq$$

You can use other distribution of p too.

Why?: Add extra dimensions to escape your bad karma.... Make the dynamics irreversible to get faster (This idea has been around for quite a while but is quite popular.)

- Ex1: Langevin equation

$$dq_t = \frac{p_t}{m}dt, \quad dp_t = \left(-\nabla V(q_t) - \gamma\frac{p_t}{m}\right)dt + \sqrt{\frac{2\gamma}{\beta}}dW_t$$

(2) $$\mathcal{L} = \underbrace{\left(\frac{p^T}{m}\right)\nabla_q - \nabla V^T \nabla_p}_{T=-T^*} + \underbrace{\frac{1}{\beta}(\Delta_p - \gamma\left(\frac{p}{M}\right)^T\nabla_p)}_{S=S^*}$$

• Ex2: Randomized Hamiltonian Monte-Carlo.

The particle follow Hamiltonian equation of motions

$$dq_t = \frac{p_t}{m}dt, \quad dp_t = -\nabla V(q_t)$$

without noise or dissipation for a random amount of time at which we resample the momentum according to the stationary measure.

With the projection $\Pi f = \int f(p,q)d\pi(p)$ the generator is

(3) $$\mathcal{L} = \underbrace{\left(\frac{p^T}{m}\right)\nabla_q - \nabla V^T \nabla_p}_{T=-T^*} + \underbrace{\lambda(\Pi - I)}_{S=S^*}$$

- EX 3: Bouncy particle sampler.

The particle follow straight lines for a random time. At updating time one either resample the momentum according to the stationary measure *or the particle "bounces"*, i.e., it undergoes a Newtonian elastic collision on the hyperplane tangential to the gradient of the energy and the momentum is updated according to the rule

$$(4) \qquad r(q)p = p - \frac{p^T \nabla V(q)}{\|\nabla V\|^2} \nabla V \quad Rf(p,q) = f(q, r(q)p)$$

$$(5) \qquad \mathcal{L} = \underbrace{\left(\frac{p}{m}\right)^T \nabla_q}_{\text{free motion}} + \underbrace{\left[\left(\frac{p}{m}\right)^T \nabla V(q)\right]^+ (R - I)}_{\text{bouncing}} + \underbrace{\lambda(\Pi - I)}_{\text{noise}}$$

- Zig-zag sampler..... etc...

Hypocoercvity

Dolbeaut-Mouhot-Schmeiser (Langevin)
Andrieu-Durmus-Nüsken-Roussel
after many other works (Villani, Hereau-Nier, Hairer-Eckmann).

Idea: The dynamics is not coercive (no Poincaré inequality in $L^2(\mu)$ for $\mathcal{L}$), but there exists a scalar product equivalent to $L^2(\mu)$ where a Poincar'e inequality holds!

$$\langle f, g\rangle_\epsilon = \langle f, f\rangle + \epsilon\langle f, (B + B^*)g\rangle.$$

$$B = (1 + (T\Pi)^*(T\Pi))^{-1}(-T\Pi)^*$$

and T is the antisymmetric part of the generator

Modified Poincaré inequality:

$$\langle -\mathcal{L}g, g\rangle_\epsilon \geq \Lambda(\epsilon)\mathrm{Var}_\mu(f) \tag{6}$$

and $\Lambda(\epsilon)$ is explicitly expressed in terms of the Poincaré inequality for $\nu(dq)$ the spectral gap of the noise operator and the potential V....

Performance guarantees for hypocoercive samplers

New results (Jermiah Birell and L. R.-B.)

Theorem (Bernstein type inequalities for hypocoercive sampler)
For bounded f we have

$$P_{\mu_0}\left(\left|\frac{1}{T}\int_0^T f(X_t)dt - \int f d\mu\right| \geq r\right)$$

$$\leq a(\epsilon)\left\|\frac{d\mu_0}{d\mu}\right\|_{L^2(\mu)} \exp\left(-T\frac{b(\epsilon)\Lambda(\epsilon)r^2}{4\mathrm{Var}_\mu[f] + 2c(\epsilon)\|f - E_\mu[f]\|r}\right)$$

where $a(\epsilon), b(\epsilon), c(\epsilon)$ only depends on ϵ.

You can use this to derive non asymptotic confidence intervals for $\int f d\mu$, i.e. as well as UQ bounds for alternative process

$$\xi_{P,f}(\eta) \leq \sqrt{2a(\epsilon)\Lambda(\epsilon)\mathrm{Var}_\mu[f]\eta} + b(\epsilon)\Lambda(\epsilon)\|f - E_\mu[f]\|_\infty \eta$$

where η is the relative entropy rate.