Implicit Sampling with Application to Filtering and Data Assimilation

New Methods for Complex Mechanics, Advanced Materials, Interfaces, and Stochastics

Alexandre Chorin
Lawrence Berkeley National Laboratory
Berkeley, CA, 94720.

The effective sampling of multidimensional probability densities is one of the key problems in computational science. As is well know, the difficulty lies in the fact that even in simple cases, there are too many states to be listed, and the fraction of states that has a significant probability is too small for standard sampling; the significant samples have to be searched for, for example by Markov-chain Monte Carlo. This can be quite slow, in particular when the free energy landscapes are not convex; furthemore, the samples produced have persistent correlations.

We have developed a methodology which finds independent high-probability samples by solving an a random equation, each sample being independent of all the others. Suppose you want to sample a probability density function (pdf) $P = P(x)$. Write $F = -logP$, and find ϕ, the minimum of F. Pick a reference random variable ξ, for example a Gaussian with independent components (this choice is not unique and the Gaussian choice here does not imply a Gaussian approximation at any stage). Find samples x of P by solving the equation $F(x) - \phi = \xi^T\xi/2$ for each sample ξ. If F is convex one can make the mapping $\xi \rightarrow x$ one-to-one and onto, and then the most likely samples of x are near the minimum of F and therefore have high probability. The algebraic equation is a single equation in many variables, and one can exploit this degeneracy to construct efficient algoriths for finding solutions, and also to construct $\xi \rightarrow x$ mappings that solve the equation even when F is not convex. Note also that the normalization of P need not be known in advance (because its logarithm cancels out with ϕ), which is important in applications to statistical mechanics.

I will present in detail the application of these ideas to filtering and data assimilation, as in meteorology, oceanography, and geophysics, where the task is to determine the state of a system (and possibly some of its parameters) from a model written as stochastic differential equation (SDE) complemented by a stream of noisy data. The SDE and the data jointly define a time dependent pdf, and one can estimate this pdf recursively by a collection of samples ("particles"), and then estimate the state as a statistic of the particles. The difficulty lies in sampling the pdf efficiently. At the $(n+1)$-th step of the recursion, the SDE by itself defines a pdf $P(x|x^n)$, where x^n is the outcome of the previous step; in standard sampling methods, this last pdf (or some related "prior") is sampled and subsequently weighted by the Bayesian factor $P(b^{n+1}|x)$ which accounts for a new datum b^{n+1}. The problem with such weighting schemes is that in most cases the prior becomes nearly singular with respect to the pdf one is looking for, so that most particles have low probability and the algorithm becomes very expensive. With implicit sampling, one can simply set F equal to minus the log of the desired conditional probability, which is known a priori up to a constant, and then solve the algebraic equation above for each particle to find high probability samples. I will present numerical examples that demonstrate the efficiency and accuracy of the resulting algorithm.

(This work is joint with E. Atkins, M. Morzfeld, and X. Tu).