Statistica Sinica **34** (2024), 1-18
doi:https://doi.org/10.5705/ss.202022.0340

ESTIMATING BOLTZMANN AVERAGES FOR PROTEIN STRUCTURAL QUANTITIES USING SEQUENTIAL MONTE CARLO

Zhaoran Hou and Samuel W.K. Wong*

University of Waterloo

Abstract: Sequential Monte Carlo (SMC) methods are widely used to draw samples from intractable target distributions. Weight degeneracy can hinder the use of SMC when the target distribution is highly constrained. As a motivating application, we consider the problem of sampling protein structures from the Boltzmann distribution. This paper proposes a general SMC method that propagates multiple descendants for each particle, followed by resampling to maintain the desired number of particles. A simulation study demonstrates the efficacy of the method for tackling the protein sampling problem, compared to existing SMC methods. As a real data example, we estimate the number of atomic contacts for a key segment of the SARS-CoV-2 viral spike protein.

Key words and phrases: Monte Carlo methods, particle filter, protein structure analysis, SARS-CoV-2.

1. Introduction

Sequential Monte Carlo (SMC) methods, also known as particle filters, are simulation-based Monte Carlo algorithms for sampling from a target distribution. SMC originated from on-line inference problems in dynamic systems, where observations arrive sequentially and interest lies in the posterior distribution of hidden state variables (Liu and Chen (1998)). Subsequent developments include the Rao-Blackwellised particle filter and its extensions (Casella and Robert (1996); Chen and Liu (2000); Andrieu and Doucet (2002); Chen et al. (2010); Johansen, Whiteley and Doucet (2012)) and the class of particle Markov Chain Monte Carlo algorithms (Andrieu, Doucet and Holenstein (2010); Kantas et al. (2015); Chopin and Singh (2015)). These methods specialize in handling dynamic systems and their structure of hidden states. SMC has also been adapted as a useful approach for sampling from general high-dimensional probability distributions (Liu (2001); Del Moral, Doucet and Jasra (2006); Wang, Wang and Bouchard-Côté (2020)); this is the setting considered in this paper.

We begin with a review of the relevant SMC concepts for general sampling problems following Doucet, de Freitas and Gordon (2001). Assume we have a vector of random variables $(\mathbf{x}_0, \ldots, \mathbf{x}_T)$, denoted by $\mathbf{x}_{0:T}$, with continuous support

*Corresponding author.