Discussion on "Exploration vs. Exploitation in Adaptive MCMC" (Scott Schmidler)

Feng Liang

University of Illinois at Urbana-Champaign

Jan 4, 2010

Adaptive MCMC

- aMCMC: pick an optimal proposal distribution $Q_{\theta_n}(x, dy)$ from a parametric family, where $\theta_n = \theta_n(X_1, \ldots, X_n)$ depends on the previous draws.
- Is it still a valid algorithm?
  - Many results (convergence, π-ergodicity, LLN) have been established to answer this question.
  - "Although more theoretical work on adaptive sampling can be expected, the existing body of results provides sufficient justification and guidelines to build adaptive MH samplers for challenging problems." (Giordani and Kohn, 2009)

The key question from Scott's talk: Is it worth doing it?

- A friendly reminder from Scott: don't forget about the original goal of designing adaptive algorithms; they are useful only if they do better than their non-adaptive counterparts.
- Existing work: optimizing $=$ achieving the best acceptance ratio α^*, where α^* is related to the optimality of the asymptotic variance $\mathrm{Var}_\theta(f)$ (Andrieu and Thoms, 2008).

Scott's Framework

I: Use mixing time to characterize finite sample performance

- Of practical interest is the finite sample performance

$$\mathrm{MSE}(f) = \mathrm{Bias}(f)^2 + \mathrm{Var}(f).$$

- A proper way to characterize the finite sample performance of an algorithm is through the convergence rate of its mixing time.
- Interesting and surprising results: some adaptive algorithms do not improve qualitatively on the mixing times of their non-adaptive Markov chain counterparts.

II: Exploitation (reduce variance) vs Exploration (reduce bias)

- Classify algorithms as "exploratory" versus "exploitative", based on their ability to improve mixing time versus autocorrelation for multimodal target distributions.
- Propose to combine adaptations of both types, which is likely to be superior to using any single existing adaptive algorithm.

- I want to congratulate Scott on this inspiring talk!
- Next: Comments and questions on general aMCMC from a non-expert user.

A Decision Theoretic Study on the Optimality of Adaptive Algorithms

- A family of transition kernels $P_\theta(x, dy)$ with $\theta \in \Theta$.
- A utility function $L(\pi, \theta)$ measures the performance of the transition kernel P_θ wrt a target distribution π.
- The optimal transition kernel for a given target distribution is P_{θ^*} where
$$\theta^* = \arg\min_\theta L(\pi, \theta).$$
- An adaptive algorithm defines a sequence of mappings θ_n from $(X_1, \ldots, X_n)$ to Θ.

- Possible Results
  - Asymptotic optimality $\theta_n \longrightarrow \theta^*$
  - Oracle property: establish bounds on $L(\pi, \theta_n) - L(\pi, \theta^*)$.
- Potential connection with other areas.
  - In function estimation with regularization, a typical strategy is to restrict the parameter space $\Theta_1 \subset \Theta_2 ... \subset \Theta_n$ to achieve the optimal bias and variance trade-off.
  - I've seen similar techniques used in other talks.
  - Any deeper connection?

f-dependent Criteria

- MSF(f): a natural choice for the performance measure.
- Someone's treasure is someone else's trash: optimality for a function f might not result in optimality for another function f'.
- My question to the experts: can you design adaptive algorithms for certain families of f, such as linear or polynomial functions?

- Motivation I: in aMCMC, the key is to find a proposal distribution that approximates the target π well. In most talks, this discrepancy is measured by the KL-divergence.
  However, consider

$$p_1(x) \propto \exp(x)\mathbf{1}_{x\geq 0}, \quad p_2(x) \propto \exp(x-\epsilon)\mathbf{1}_{x\geq \epsilon},$$

  we have $\mathrm{KL}(p_1, p_2) = \infty$, but $\mathbb{E}_1 f(X) \approx \mathbb{E}_2 f(X)$ for small ϵ.
- Motivation II: Sampling a high-dimensional Normal distribution is difficult since it's hard to learn the covariance structure.
  However, if $f(X) = a^t X$, then we only need to sample a 1-dim Normal.

What Else Can We Utilize from the Past

- aMCMC: tune the proposal distribution "on the fly" by utilizing the history $(X_1, \ldots, X_n)$.
- A complete history $(X_1, Y_1, X_2, Y_2, \ldots, X_n, Y_n)$
- Can we utilize the rejected samples? Not for sampling but for learning.
- The key step in aMCMC: approximate the target distribution π, where π is unknown but can be evaluated at any point x (up to a constant).
- At each step, we have already compute the value of π at Y_i, then we should use that information even if we do not accept Y_i.

Other Random Thoughts

- Connection to active learning
- Connection to manifolds learning

Conclusion

- Some complaints from the potential users
  - Leaning within learning
  - Tuning parameters for tuning parameters
  - Adaptation within adaptation
  - Is it worth all the efforts? Can't we just run a two-stage approach?

Conclusion

- Some complaints from the potential users
  - Leaning within learning
  - Tuning parameters for tuning parameters
  - Adaptation within adaptation
  - Is it worth all the efforts? Can't we just run a two-stage approach?
- There is no question on the need and importance to develop theory and methodology for aMCMC.
- Positive side: More work need to be done for aMCMC! We can start planning to come to AdapskIV, AdapskV...