Sampling from posterior distribution is crucial in various statistical techniques

Posterior can be hard to sample (hurdle a lot of applications)

SGD can be viewed as a Markov chain

property as a Markov chain seems to be largely overlooked

stationary distribution? mixing time? mixing time relationship to convergence rate?

ubiquitous in

SGD is widely used.

SGD has been viewed as Markov chain.

No previous work has analyzed SGD as sampling technique.

(sampling meth formulation)

The posterior sampling method is used to great advantage by a board range of statistical methodologies, such as EM algorithm (xxxx). From the Bayesian point of view, E-step of the EM algorithm can be seen as calculating the posterior distribution of the parameters of interest (Wing xxxx). Another crucial arena of posterior sampling is Bayesian modeling. For example, (German bayes image xxx ) studied image restoration as a MAP problem on lattice graph. They proposed a Markov Random Field - Gibbs class of method for solving the posterior inference.

Typically, to implement such algorithms, one must be able to sample from the posterior distribution $p(\theta|y,z)$. In general, however, the posterior does not have closed form and sampling could be difficult and sometimes intractable. Various alternatives have been proposed such as Monte-Carlo method to approximate the posterior in EM algorithm (xxxx). Tanner and Wong (xxx) considered data augmentation approach for calculating posterior. Analysis of posterior is studied in Rubin (xxx) as a Bayesian Bootstrap approach.

To formalize the goal, the observed data $y$ follows a probability distribution indexed by $\theta$ : $p(y | \theta)$. We want to generate i.i.d. samples from $p(\theta | y)$.

Our approach to the problem (xxx) is related to classical stochastic gradient descent algorithms [RM51, PJ92], where one assumes access to samples $y\_1, …, y\_n$ from the distribution $p$ and performs gradient updates using $$\nabla P(\theta ;y)$$. This reduces to the randomized incremental subgradient method of Nedic and Bertsekas [xxxxxx] when $P$ is concentrated on a set of $n$ points, giving an objective of the form

$$p(x) = \frac{1}{n}\sum\limits\_{i = 1}^n {{p\_i}(x)} $$.

More generally, our problem belongs to the family of xxxxx with xxxxxx, where the goal is to xxxxxx. Classical analysis of SGD focus on xxxxx in nature and generally do not provide xxxxxx result. Our method borrows from yyyyyyy methodology [zzzzzzz], but we generalize them in that we do not assume yyyyyy.

The main result of this paper is that performing stochastic gradient steps with specifically chosen step size, xxxxx results in a provably accurate and efficient sampling procedure. The convergence to the stationary distribution is governed by problem-dependent terms (namely the xxxx of the functions F, and xxxx) familiar from previous results on Markov chain [yyyyyy] as well as terms dependent on the rate at which xxxxxxx. Our xxx main theorems characterize the convergence of SGD in terms of xxxx parameter, which is xxxxx (in a sense we make precise later). In particular, we show that the stationary distribution is xxxxx with high probability.

Posterior sampling 有优势

(一次状态转移有优势than metropolis Hastings)

求gradient efficient

The remainder of the paper is organized as follows. The next section contains a description of the algorithm we analyze and our main technical results. Following that, we show that the classic stochastic gradient descent can be viewed as a special case of our algorithm in Section 3. We expand on these results and provide an upper bound of the mixing time throughout Section 4, and give numerical examples to compare our result with other sampling methods like xxxxxx. We conclude with discussion in Section 5.

Though a wide variety of stochastic optimization methods for solving the problem (2) have been explored in an extensive literature [RM51, PJ92, NB01, NJLS09], most approaches have imposed the restrictive assumption that it is possible to obtain independent and identically distributed samples ξ from the distribution Π. We relax this assumption and instead assume that we receive samples ξ from a stochastic process P indexed by time t, where the stochastic process P converges to the stationary distribution Π. This is a natural relaxation, because in many circumstances

the distribution Π is unknown—for example in statistical applications—and we cannot receive independent samples. In other scenarios, it may be hard to even draw samples from Π efficiently, such as when Ξ is a high-dimensional space or is a combinatorial space, but it is possible to design Markov chains that converge to the distribution Π [JS96]. Further, in computational applications, it is often unrealistic to assume that one actually has access to a source of independent randomness, so studying the effect of correlation is natural and important [IZ89].

