Finite-Time Analysis of Projected Langevin Monte Carlo

Authored by:

Sebastien Bubeck
Ronen Eldan
Joseph Lehec

Abstract

We analyze the projected Langevin Monte Carlo (LMC) algorithm, a close cousin of projected Stochastic Gradient Descent (SGD). We show that LMC allows to sample in polynomial time from a posterior distribution restricted to a convex body and with concave log-likelihood. This gives the first Markov chain to sample from a log-concave distribution with a first-order oracle, as the existing chains with provable guarantees (lattice walk, ball walk and hit-and-run) require a zeroth-order oracle. Our proof uses elementary concepts from stochastic calculus which could be useful more generally to understand SGD and its variants.

1 Paper Body

A fundamental primitive in Bayesian learning is the ability to sample from the posterior distribution. Similarly to the situation in optimization, convexity is a key property to obtain algorithms with provable guarantees for this task. Indeed several Markov Chain Monte Carlo methods have been analyzed for the case where the posterior distribution is supported on a convex set, and the negative log-likelihood is convex. This is usually referred to as the problem of sampling from a log-concave distribution. In this paper we propose and analyze a new Markov chain for this problem which could have several advantages over existing chains for machine learning applications. We describe formally our contribution in Section 1.1. Then in Section 1.2 we explain how this contribution relates to various line of work in different fields such as theoretical computer science, statistics, stochastic approximation, and machine learning. 1.1

Main result

Let K ? Rn be a convex set such that 0 ? K, K contains a Euclidean ball of radius r ¿ 0 and is contained in a Euclidean ball of radius R. Denote PK the Euclidean projection on K (i.e., PK (x) = argminy?K —x ? y— where — ? — denotes the Euclidean norm in Rn), and k ? kK the gauge of K defined

by kxkK = inf{t ? 0; x ? tK}, x ? Rn . Let f : K ? R be a L-Lipschitz and ?-smooth convex function, that is f is differentiable and satisfies ?x, y ? K, —?f (x) ? ?f (y)— ? ?—x ? y—, and —?f (x)— ? L. We are interested in the problem of sampling from the probability measure ? on Rn whose density with respect to the Lebesgue measure is given by: Z d? 1 = exp(?f (x))1{x ? K}, where Z = exp(?f (y))dy. dx Z y?K 1

We denote m = E? —X—, and M = E [k?kK], where ? is uniform on the sphere Sn?1 = {x ? Rn : —x— = 1}. In this paper we study the following Markov chain, which depends on a parameter ? ¿ 0, and where ?1 , ?2 , . . . is an i.i.d. sequence of standard Gaussian random variables in Rn , and X 0 = 0,

? ? X k+1 = PK X k ? ?f (X k) + ??k . (1) 2 We call the chain (1) projected Langevin Monte Carlo (LMC). Recall that the total variation distance between two measures ?, ? is defined as TV(?, ?) = supA —?(A) ? ?(A)— where the supremum is over all measurable sets A. With a slight abuse of notation we sometimes write TV(X, ?) where X is a random variable distributed according to e n) (respectively ?) e means that there exists c ? R, C ¿ 0 such that ?. The notation vn = O(u vn ? Cun logc (un) (respectively ?). Our main result shows that for an appropriately chosen step-size and number of iterations, one has convergence in total variation distance of the iterates (X k) to the target distribution ?. 2 and Theorem 1 Let ? ¿ 0. One has TV(X N , ?) ? ? provided that ? = N1 m ? !!

6

? 8 1 n + RL 1 2 2 6 e (n + RL) (M + L/r) nm max N =? , 22 ?m(L + R) . ?16 r ? Note that by viewing ?, L, r as numerical constants, using M ? 1/r, and assuming R ? n and m ? n3/4 , the bound reads 9 6 e n m . N =? ?22 ? Observe also that if f is constant, that is ? is the uniform measure on K, then L = 0, m ? n, and ? e one can show that M = O(1/ n), which yields the bound:

n 11 e N =? . ?2 1.2

Context and related works

There is a long line of works in theoretical computer science proving results similar to Theorem 1, starting with the breakthough result of Dyer et al. [1991] who showed that the lattice walk mixes in e 23) steps. The current record for the mixing time is obtained by Lov?asz and Vempala [2007], O(n e 4) for the hit-and-run walk. These chains (as well as other popular chains who show a bound of O(n such as the ball walk or the Dikin walk, see e.g. Kannan and Narayanan [2012] and references therein) all require a zeroth-order oracle for the potential f , that is given x one can calculate the value f (x). On the other hand our proposed chain (1) works with a first-order oracle, that is given x one can calculate the value of ?f (x). The difference between zeroth-order oracle and firstorder oracle has been extensively studied in the optimization literature (e.g., Nemirovski and Yudin [1983]), but it has been largely ignored in the literature on polynomial-time sampling algorithms. We also note that hit-and-run and LMC are the only chains which are rapidly mixing from any starting point (see Lov?asz and Vempala [2006]), though they have this property for seemingly very different reasons. When initialized in a corner of the convex body, hit-and-run might take a long time to take a step, but once it moves it escapes very

far (while a chain such as the ball walk would only do a small step). On the other hand LMC keeps moving at every step, even when initialized in a corner, thanks for the projection part of (1). Our main motivation to study the chain (1) stems from its connection with the ubiquitous stochastic gradient descent (SGD) algorithm. In general this algorithm takes the form xk+1 = PK (xk ? ??f (xk) + ?k) where ?1 , ?2 , . . . is a centered i.i.d. sequence. Standard results in approximation theory, such as Robbins and Monro [1951], show that if the variance of the noise Var(?1) is of smaller order than the step-size ? then the iterates (xk) converge to the minimum of f on K (for a step-size decreasing sufficiently fast as a function of the number of iterations). For the specific noise 2

sequence that we study in (1), the variance is exactly equal to the step-size, which is why the chain deviates from its standard and well-understood behavior. We also note that other regimes where SGD does not converge to the minimum of f have been studied in the optimization literature, such as the constant step-size case investigated in Pflug [1986], Bach and Moulines [2013]. The chain (1) is also closely related to a line of works in Bayesian statistics on Langevin Monte Carlo algorithms, starting essentially with Tweedie and Roberts [1996]. The focus there is on the unconstrained case, that is K = Rn . In this simpler situation, a variant of Theorem 1 was proven in the recent paper Dalalyan [2014]. The latter result is the starting point of our work. A straightforward way to extend the analysis of Dalalyan to the constrained case is to run the unconstrained chain with an additional potential that diverges quickly as the distance from x to K increases. However it seems much more natural to study directly the chain (1). Unfortunately the techniques used in Dalalyan [2014] cannot deal with the singularities in the diffusion process which are introduced by the projection. As we explain in Section 1.3 our main contribution is to develop the appropriate machinery to study (1). In the machine learning literature it was recently observed that Langevin Monte Carlo algorithms are particularly well-suited for large-scale applications because of the close connection to SGD. For instance Welling and Teh [2011] suggest to use mini-batch to compute approximate gradients instead of exact gradients in (1), and they call the resulting algorithm SGLD (Stochastic Gradient Langevin Dynamics). It is conceivable that the techniques developed in this paper could be used to analyze SGLD and its refinements introduced in Ahn et al. [2012]. We leave this as an open problem for future work. Another interesting direction for future work is to improve the polynomial dependency on the dimension and the inverse accuracy in Theorem 1 (our main goal here was to provide the simplest polynomial-time analysis). 1.3

Contribution and paper organization

As we pointed out above, Dalalyan [2014] proves the equivalent of Theorem 1 in the unconstrained case. His elegant approach is based on viewing LMC as a discretization of the diffusion process dXt = dWt ? 21 ?f (Xt), where (Wt) is a Brownian motion. The analysis then proceeds in two steps, by deriving first the mixing time of the diffusion process, and then showing that the discretized process is ?close? to its continuous version. In Dalalyan [2014] the first step

is particularly transparent as he assumes ?-strong convexity for the potential f , which in turns directly gives a mixing time of order 1/?. The second step is also simple once one realizes that LMC (without projection) can be viewed as the diffusion process dX t = dWt ? 12 ?f (X?b ?t c). Using Pinsker?s inequality and Girsanov?s formula it is then a short calculation to show that the total variation distance between X t and Xt is small. The constrained case presents several challenges, arising from the reflection of the diffusion process on the boundary of K, and from the lack of curvature in the potential (indeed the constant potential case is particularly important for us as it corresponds to ? being the uniform distribution on K). Rather than a simple Brownian motion with drift, LMC with projection can be viewed as the discretization of reflected Brownian motion with drift, which is a process of the form dXt = dWt ? 21 ?f (Xt)dt ? ?t L(dt), where Xt ? K, ?t ? 0, L is a measure supported on {t ? 0 : Xt ? ?K}, and ?t is an outer normal unit vector of K at Xt . The term ?t L(dt) is referred to as the Tanaka drift. Following Dalalyan [2014] the analysis is again decomposed in two steps. We study the mixing time of the continuous process via a simple coupling argument, which crucially uses the convexity of K and of the potential f . The main difficulty is in showing that the discretized process (X t) is close to the continuous version (Xt), as the Tanaka drift prevents us from a straightforward application of Girsanov?s formula. Our approach around this issue is to first use a geometric argument to prove that the two processes are close in Wasserstein distance, and then to show that in fact for a reflected Brownian motion with drift one can deduce a total variation bound from a Wasserstein bound. In this extended abstract we focus on the special case where f is a constant function, that is ? is uniform on the convex body K. The generalization to an arbitrary smooth potential can be found in the supplementary material. The rest of the paper is organized as follows. Section 2 contains the main tehcnical arguments. We first remind the reader of Tanaka?s construction (Tanaka [1979]) of reflected Brownian motion in Section 2.1. We present our geometric argument to bound the 3

Wasserstein distance between (Xt) and (X t) in Section 2.2, and we use our coupling argument to bound the mixing time of (Xt) in Section 2.3. The derivation of a total variation bound from the Wasserstein bound is discussed in Section 2.4. Finally we conclude the paper in Section 3 with some preliminary experimental comparison between LMC and hit-and-run.

2

The constant potential case

In this section we derive the main arguments to prove Theorem 1 when f is a constant function, that is ?f = 0. For a point x ? ?K we say that ? is an outer unit normal vector at x if —?— = 1 and hx ? x0 , ?i ? 0,

?x0 ? K.

For x ? / ?K we say that 0 is an outer unit normal at x. We define the support function hK of K by hK (y) = sup {hx, yi; x ? K} ,

y ? Rn .

Note that hK is also the gauge function of the polar body of K. 2.1

The Skorokhod problem

Let T ? R+ ? {+?} and w : [0, T) ? Rn be a piecewise continuous path with w(0) ? K. We say that x : [0, T) ? Rn and ? : [0, T) ? Rn solve the Skorokhod problem for w if one has x(t) ? K, ?t ? [0, T), x(t) = w(t) + ?(t), ?t ? [0, T), and furthermore ? is of the form Z t ?(t) = ? ?s L(ds), ?t ? [0, T), 0

where ?s is an outer unit normal at x(s), and L is a measure on [0, T] supported on the set {t ? [0, T) : x(t) ? ?K}. The path x is called the reflection of w at the boundary of K, and the measure L is called the local time of x at the boundary of K. Skorokhod showed the existence of such a a pair (x, ?) in dimension 1 in Skorokhod [1961], and Tanaka extended this result to convex sets in higher dimensions in Tanaka [1979]. Furthermore Tanaka also showed that the solution is unique, and if w is continuous then so is x and ?. In particular the reflected Brownian motion in K, denoted (Xt), is defined as the reflection of the standard Brownian motion (Wt) at the boundary of K (existence follows by continuity of Wt). Observe that by It?o?s formula, for any smooth function g on Rn , Z t Z Z t 1 t g(Xt) ? g(X0) = h?g(Xs), dWs i + ?g(Xs) ds ? h?g(Xs), ?s i L(ds). (2) 2 0 0 0 To get a sense of what a solution typically looks like, let us work out the case where w is piecewise constant (this will also be useful to realize that LMC can be viewed as the solution to a Skorokhod problem). For a sequence g1 . . . gN ? Rn , and for ? ¿ 0, we consider the path: w(t) =

N X

gk 1{t ? k?},

t ? [0, (N + 1)?).

k=1

Define (xk)k=0,...,N inductively by x0 = 0 and xk+1 = PK (xk + gk). It is easy to verify that the solution to the Skorokhod problem for w is given by x(t) = x?b ?t c and Rt ?(t) = ? 0 ?s L(ds), where the measure L is defined by (denoting ?s for a dirac at s) L=

N X

—xk + gk ? PK (xk + gk)—?k? ,

k=1

and for s = k?, ?s =

xk + gk ? PK (xk + gk) . —xk + gk ? PK (xk + gk)— 4

2.2

Discretization of reflected Brownian motion

Given the discussion above, it is clear that when f is a constant function, the chain (1) can be viewed as the reflection (X t) of a discretized Brownian motion W t := W?b ?t c at the boundary of K (more precisely the value of X k? coincides with the value of X k as defined by (1)). It is rather clear that the discretized Brownian motion (W t) is ?close? to the path (Wt), and we would like to carry this to the reflected paths (X t) and (Xt). The following lemma extracted from Tanaka [1979] allows to do exactly that. Lemma 1 Let w and w be piecewise continuous path and assume that (x, ?) and (x, ?) solve the Skorokhod problems for w and w, respectively. Then for all time t we have

—x(t) ? x(t)—2 ? —w(t) ? w(t)—2 Z t +2 hw(t) ? w(t) ? w(s) + w(s), ?(ds) ? ?(ds)i. 0

Applying the above lemma to the processes (Wt) and (W t) at time T = N ? yields (note that WT = W T) Z T Z T 2 —XT ? X T — ? ?2 hWt ? W t , ?t iL(dt) + 2 hWt ? W t , ? t iL(dt) 0

0

We claim that the second integral is equal to 0. Indeed, since the discretized process is constant on the intervals [k?, (k + 1)?) the local time L is a positive combination of Dirac point masses at ?, 2?, . . . , N ?. On the other hand Wk? = W k? for all integer k, hence the claim. Therefore Z T hWt ? W t , ?t i L(dt) —XT ? X T —2 ? ?2 0

Using the inequality hx, yi ? kxkK hK (y) we get Z

2

—XT ? X T — ? 2 sup kWt ? W T kK [0,T]

T

hK (?t) L(dt). 0

Taking the square root, expectation and using Cauchy?Schwarz we get " # "Z # T 2

E —XT ? X T — ? 2 E sup kWt ? W T kK E hK (?t) L(dt) . [0,T]

(3)

0

The next two lemmas deal with each term in the right hand side of the above equation, and they will show that there exists a universal constant C such that

1/4 E —XT ? X T — ? C (? log(T /?)) n3/4 T 1/2 M 1/2 . (4) We discuss why the above bound implies a total variation bound in Section 2.4. Lemma 2 We have, for all t ¿ 0 Z E 0

t

nt hK (?s) L(ds) ? . 2

Proof By It?o?s formula d—Xt —2 = 2hXt , dWt i + n dt ? 2hXt , ?t i L(dt). Now observe that by definition of the reflection, if t is in the support of L then hXt , ?t i ? hx, ?t i, ?x ? K. In other words hXt , ?t i ? hK (?t). Therefore Z t Z t 2 hK (?s) L(ds) ? 2 hXs , dWs i + nt + —X0 —2 ? —Xt —2 . 0

0

The first term of the right?hand side is a martingale, so using that X0 = 0 and taking expectation we get the result.

5

Lemma 3 There exists a universal constant C such that " # E sup kWt ? W t kK ? C M

p

n? log(T /?).

[0,T]

Proof Note that

"

#

E sup kWt ? W t kK = E [0,T]

max
0?i?N ?1
Yi
where Yi =
kWt ? Wi? kK .
sup t?[i?,(i+1)?)

Observe that the variables (Yi) are identically distributed, let p ? 1 and write ? !1/p ?

N ?1 X ? ? N 1/p kY0 kp . E max Yi ? E ? —Yi —p i?N ?1
i=0

We claim that

? (5) kY0 kp ? C p n ? M for some constant C, and for all p ? 2. Taking this for granted and choosing p = log(N) in the previous inequality yields the result (recall that N = T /?). So it is enough to prove (5). Observe that since (Wt) is a martingale, the process Mt = kWt kK is a sub?martingale. By Doob?s maximal inequality kY0 kp = k sup Mt kp ? 2kM? kp , [0,?]

for every p ? 2. Letting ?n be the standard Gaussian measure on Rn and using Khintchin?s inequality we get Z 1/p Z ? ? p kM? kp = ? kxkK ?n (dx) ? C p? kxkK ?n (dx). Rn
Rn

Lastly, integrating in polar coordinate, it is easily seen that Z ? kxkK ?n (dx) ? C n M. Rn

2.3

A mixing time estimate for the reflected Brownian motion

Given a probability measure ? supported on K, we let ?Pt be the law of Xt when X0 has law ?. The following lemma is the key result to estimate the mixing time of the process (Xt). Lemma 4 Let x, x0 ? K

—x ? x0 — TV(?x Pt , ?x0 Pt) ? ? . 2?t

The above result clearly implies that for a probability measure ? on K, TV(?0 Pt , ?Pt) ? R —x— ?(dx) K ? . Since ? (the uniform measure on K) is stationary for reflected Brownian motion, we 2?t obtain m TV(?0 Pt , ?) ? ? . (6) 2?t In other words, starting from 0, the mixing time of (Xt) is of order m2 . We now turn to the proof of the above lemma. Proof The proof is based on a coupling argument. Let (Wt) be a Brownian motion starting from 0 and let (Xt) be a reflected Brownian motion starting from x:

X0 = x dXt = dWt ? ?t L(dt) 6

where (?t) and L satisfy the appropriate conditions. We construct a reflected Brownian motion (Xt0) starting from x0 as follows. Let ? = inf{t ? 0; Xt = Xt0 }, and for t ¡ ? let St be the orthogonal reflection with respect to the hyperplane (Xt ? Xt0)? . Then up to time ? , the process (Xt0) is defined by (0 X0 = x0 dXt0 = dWt0 ? ?t0 L0 (dt) dWt0 = St (dWt) where L0 is a measure supported on {t ? ? ; Xt0 ? ?K}, and ?t0 is an outer unit normal at Xt0 for all such t. After time ? we just set Xt0 = Xt . Since St is an orthogonal map (Wt0) is a Brownian motion and thus (Xt0) is a reflected Brownian motion starting from x0 . Therefore TV(?x Pt , ?x0 Pt) ? P(Xt 6= Xt0) = P(? ¿ t).

Observe that on [0, ?) dWt ? dWt0 = (I ? St)(dWt) = 2hVt , dWt iVt , where Vt =

Xt ?Xt0 —Xt ?Xt0 — .

So

d(Xt ? Xt0) = 2hVt , dWt iVt ? ?t L(dt) + ?t0 L0 (dt) = 2(dBt) Vt ? ?t L(dt) + ?t0 L0 (dt), where Z

t

hVs , dWs i,

Bt =

on [0, ?).

0

Observe that (Bt) is a one?dimensional Brownian motion. It?o?s formula then gives dg(Xt ? Xt0) = 2h?g(Xt ? Xt0), Vt i dBt ? h?g(Xt ? Xt0), ?t i L(dt) + h?g(Xt ? Xt0), ? 0 ti L0 (dt) + 2?2 g(Xt ? Xt0)(Vt , Vt) dt, for every smooth function g on Rn . Now if g(x) = —x— then ?g(Xt ? Xt0) = Vt so h?g(Xt ? Xt0), Vt i = 1, h?g(Xt ? Xt0), ?t i ? 0 on the support of L, and h?g(Xt ? Xt0), ?t0 i ? 0 1 on the support of L0 . Moreover ?2 g(Xt ? Xt0) = —Xt ?Y P(Xt ?Yt)? where Px? denotes the t— ? 2 orthogonal projection on x . In particular ? g(Xt ? Yt)(Vt) = 0. We obtain —Xt ? Xt0 — ? —x ? x0 — + 2Bt , on [0, ?). Therefore P(? ¿ t) ? P(? 0 ¿ t) where ? 0 is the first time the Brownian motion (Bt) hits the value ?—x ? x0 —/2. Now by the reflection principle —x ? x0 — P(? 0 ¿ t) = 2 P (0 ? 2 Bt ¡ —x ? x0 —) ? ? . 2?t

2.4

From Wasserstein distance to total variation

To conclude it remains to derive a total variation bound between XT and X T using (4). The details of this step are deferred to the supplementary material where we consider the case of a general logconcave distribution. The intuition goes as follows: the processes (XT +s)s?0 and (X T +s)s?0 both evolve according to a Brownian motion until the first time s that one process undergoes a reflection. But if T is large enough and ? is small enough then one can easily get from (4) (and the fact that the uniform measure does not put too much mass close to the boundary) that XT and X T are much closer to each other than they are to the boundary of K. This implies that one can couple them (just as in Section 2.3) so that they meet before one of them hits the boundary.

3

Experiments

Comparing different Markov Chain Monte Carlo algorithms is a challenging problem in and of itself. Here we choose the following simple comparison procedure based on the volume algorithm 7

developed in Cousins and Vempala [2014]. This algorithm, whose objective is to compute the volume of a given convex set K, procedes in phases. In each phase ‘ it estimates the mean of a certain function under a multivariate Gaussian restricted to K with (unrestricted) covariance ¿ In . Cousins and Vempala provide a Matlab implementation of the entire algorithm, where in each phase

the target mean is estimated by sampling from the truncated Gaussian using the hit-and-run (H&R) chain. We implemented the same procedure with LMC instead of H&R, and we choose the step-size ? = 1/(?n2), where ? is the smoothness parameter of the underlying log-concave distribution (in particular here ? = 1/¿2). The intuition for the choice of the step-size is as follows: the scaling in inverse smoothness comes from the optimization literature, while the scaling in inverse dimension squared comes from the analysis in the unconstrained case in Dalalyan [2014].

Time

Estimated normalized volume 2000

4.5 Box H&R Box LMC Box and Ball H&R Box and Ball LMC

4 3.5

Box H&R Box LMC Box and Ball H&R Box and Ball LMC

1800 1600 1400

3

1200 2.5 1000 2 800 1.5 600 1

400

0.5 0

200 1

2

3

4

5

6

7

8

9

0

10

1

2

3

4

5

6

7

8

9

10

We ran the volume algorithm with both H&R and LMC on the ? following set of convex bodies: K = [?1, 1]n (referred to as the ?Box?) and K = [?1, 1]n ? 2n Bn (referred to as the ?Box and Ball?), where n = 10 ? k, k = 1, . . . , 10. The computed volume (normalized by 2n for the ?Box? and by 0.2 ?2n for the ?Box and Ball?) as well as the clock time (in seconds) to terminate are reported in the figure above. From these experiments it seems that LMC and H&R roughly compute similar values for the volume (with H&R

being slightly more accurate), and LMC is almost always a bit faster. These results are encouraging, but much more extensive experiments are needed to decide if LMC is indeed a competitor to H&R in practice.

8

2 References

S. Ahn, A. Korattikara, and M. Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. In ICML 2012, 2012. F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n). In Advances in Neural Information Processing Systems 26 (NIPS), pages 773?781. 2013. B. Cousins and S. Vempala. Bypassing kls: Gaussian cooling and an o? (n3) volume algorithm. Arxiv preprint arXiv:1409.6011, 2014. A. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. Arxiv preprint arXiv:1412.7392, 2014. M. Dyer, A. Frieze, and R. Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. Journal of the ACM (JACM), 38(1):1?17, 1991. R. Kannan and H. Narayanan. Random walks on polytopes and an affine interior point method for linear programming. Mathematics of Operations Research, 37:1?20, 2012. L. Lov?asz and S. Vempala. Hit-and-run from a corner. SIAM J. Comput., 35(4):985?1005, 2006. L. Lov?asz and S. Vempala. The geometry of logconcave functions and sampling algorithms. Random Structures & Algorithms, 30(3):307?358, 2007. A. Nemirovski and D. Yudin. Problem Complexity and Method Efficiency in Optimization. Wiley Interscience, 1983. G. Pflug. Stochastic minimization with constant step-size: asymptotic laws. SIAM J. Control and Optimization, 24(4):655?666, 1986. H. Robbins and S. Monro. A stochastic approximation method. Annals of Mathematical Statistics, 22:400?407, 1951. A. Skorokhod. Stochastic equations for diffusion processes in a bounded region. Theory of Probability & Its Applications, 6(3):264?274, 1961. H. Tanaka. Stochastic differential equations with reflecting boundary condition in convex regions. Hiroshima Mathematical Journal, 9(1):163?177, 1979. L. Tweedie and G. Roberts. Exponential convergence of langevin distributions and their discrete approximations. Bernoulli, 2(4):341?363, 1996. M. Welling and Y.W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In ICML 2011, 2011.

9