Measuring Sample Quality with Stein's Method

Authored by:

Jackson Gorham
Lester Mackey

Abstract

To improve the efficiency of Monte Carlo estimation, practitioners are turning to biased Markov chain Monte Carlo procedures that trade off asymptotic exactness for computational speed. The reasoning is sound: a reduction in variance due to more rapid sampling can outweigh the bias introduced. However, the inexactness creates new challenges for sampler and parameter selection, since standard measures of sample quality like effective sample size do not account for asymptotic bias. To address these challenges, we introduce a new computable quality measure based on Stein's method that bounds the discrepancy between sample and target expectations over a large class of test functions. We use our tool to compare exact, biased, and deterministic sample sequences and illustrate applications to hyperparameter selection, convergence rate assessment, and quantifying bias-variance tradeoffs in posterior inference.

1 Paper Body

When faced with a complex target distribution, one often turns to RMarkov chain Monte Carlo (MCMC) [1] to approximate intractable expectations EP [h(Z)] = X p(x)h(x)dx with asympPn totically exact sample estimates EQ [h(X)] = i=1 q(xi)h(xi). These complex targets commonly arise as posterior distributions in Bayesian inference and as candidate distributions in maximum likelihood estimation [2]. In recent years, researchers [e.g., 3, 4, 5] have introduced asymptotic bias into MCMC procedures to trade off asymptotic correctness for improved sampling speed. The rationale is that more rapid sampling can reduce the variance of a Monte Carlo estimate and hence outweigh the bias introduced. However, the added flexibility introduces new challenges for sampler and parameter selection, since standard sample quality measures, like effective sample size, asymptotic variance, trace and mean plots, and pooled and within-chain variance diagnostics, presume eventual convergence to the target [1] and hence do not account for asymptotic bias. To address this shortcoming, we develop a new measure of sample quality suitable for comparing asymptotically exact, asymptotically biased, and even deterministic sample sequences. The quality

measure is based on Stein?s method and is attainable by solving a linear program. After outlining our design criteria in Section 2, we relate the convergence of the quality measure to that of standard probability metrics in Section 3, develop a streamlined implementation based on geometric spanners in Section 4, and illustrate applications to hyperparameter selection, convergence rate assessment, and the quantification of bias-variance tradeoffs in posterior inference in Section 5. We discuss related work in Section 6 and defer all proofs to the appendix. Notation We denote the ‘2 , ‘1 , and ‘1 norms on Rd by k?k2 , k?k1 , and k?k1 respectively. We will ? often refer to a generic norm k?k on Rd with associated dual norms kwk , supv2Rd :kvk=1 hw, vi ? ? ? for vectors w 2 Rd , kM k , supv2Rd :kvk=1 kM vk for matrices M 2 Rd?d , and kT k , ? supv2Rd :kvk=1 kT [v]k for tensors T 2 Rd?d?d . We denote the j-th standard basis vector by ej , the @ partial derivative @xk by rk , and the gradient of any Rd -valued function g by rg with components (rg(x))jk , rk gj (x). 1

2

Quality Measures for Samples

Consider a target distribution P with open convex support X ? Rd and continuously differentiable density p. We assume that p is known up to its normalizing constant and that exact integration under P is intractable for most functions of interest. We will approximate expectations under P with the aid of a weighted sample, a collection of distinct sample points x1 , . . . , xn 2 X with weights q(xi) encoded in a probability mass function q. ThePprobability mass function q induces a discrete distrin bution Q and an approximation EQ [h(X)] = i=1 q(xi)h(xi) for any target expectation EP [h(Z)]. We make no assumption about the provenance of the sample points; they may arise as random draws from a Markov chain or even be deterministically selected. Our goal is to compare the fidelity of different samples approximating a common target distribution. That is, we seek to quantify the discrepancy between EQ and EP in a manner that (i) detects when a sequence of samples is converging to the target, (ii) detects when a sequence of samples is not converging to the target, and (iii) is computationally feasible. A natural starting point is to consider the maximum deviation between sample and target expectations over a class of real-valued test functions H, dH (Q, P) = sup —EQ [h(X)] h2H

(1)

EP [h(Z)]—.

When the class of test functions is sufficiently large, the convergence of dH (Qm , P) to zero implies that the sequence of sample measures (Qm)m 1 converges weakly to P . In this case, the expression (1) is termed an integral probability metric (IPM) [6]. By varying the class of test functions H, we can recover many well-known probability metrics as IPMs, including the total variation distance, generated by H = {h : X ! R — supx2X —h(x)— ? 1}, and the Wasserstein distance (also known as the Kantorovich-Rubenstein or earth mover?s distance), dWk?k , generated by H = Wk?k , {h : X ! R — supx6=y2X

—h(x) h(y)— kx yk

? 1}.

The primary impediment to adopting an IPM as a sample quality measure is

that exact computation is typically infeasible when generic integration under P is intractable. However, we could skirt this intractability by focusing on classes of test functions with known expectation under P . For example, if we consider only test functions h for which EP [h(Z)] = 0, then the IPM value dH (Q, P) is the solution of an optimization problem depending on Q alone. This, at a high level, is our strategy, but many questions remain. How do we select the class of test functions h? How do we know that the resulting IPM will track convergence and non-convergence of a sample sequence (Desiderata (i) and (ii))? How do we solve the resulting optimization problem in practice (Desideratum (iii))? To address the first two of these questions, we draw upon tools from Charles Stein?s method of characterizing distributional convergence. We return to the third question in Section 4.

3

Stein?s Method

Stein?s method [7] for characterizing convergence in distribution classically proceeds in three steps: 1. Identify a real-valued operator T acting on a set G of Rd -valued1 functions of X for which EP [(T g)(Z)] = 0 for all g 2 G.

(2)

Together, T and G define the Stein discrepancy,

S(Q, T , G) , sup —EQ [(T g)(X)]— = sup —EQ [(T g)(X)] g2G

g2G

EP [(T g)(Z)]— = dT G (Q, P),

an IPM-type quality measure with no explicit integration under P . 2. Lower bound the Stein discrepancy by a familiar convergence-determining IPM dH . This step can be performed once, in advance, for large classes of target distributions and ensures that, for any sequence of probability measures (?m)m 1 , S(?m , T , G) converges to zero only if dH (?m , P) does (Desideratum (ii)). 1 One commonly considers real-valued functions g when applying Stein?s method, but we will find it more convenient to work with vector-valued g.

2

3. Upper bound the Stein discrepancy by any means necessary to demonstrate convergence to zero under suitable conditions (Desideratum (i)). In our case, the universal bound established in Section 3.3 will suffice. While Stein?s method is typically employed as an analytical tool, we view the Stein discrepancy as a promising candidate for a practical sample quality measure. Indeed, in Section 4, we will adopt an optimization perspective and develop efficient procedures to compute the Stein discrepancy for any sample measure Q and appropriate choices of T and G. First, we assess the convergence properties of an equivalent Stein discrepancy in the subsections to follow. 3.1

Identifying a Stein Operator

The generator method of Barbour [8] provides a convenient and general means of constructing operators T which produce mean-zero functions under P (2) . Let (Zt)t 0 represent a Markov process with unique stationary distribution P . Then the infinitesimal generator A of (Zt)t 0 , defined by (Au)(x) = lim (E[u(Zt) — Z0 = x] t!0

u(x))/t

for

u : Rd ! R,

satisfies EP [(Au)(Z)] = 0 under mild conditions on A and u. Hence, a candidate operator T can be constructed from any infinitesimal generator. For example, the overdamped Langevin diffusion, defined by the stochastic differential equation dZt = 12 r log p(Zt)dt + dWt for (Wt)t 0 a Wiener process, gives rise to the generator 1 1 (AP u)(x) = hru(x), r log p(x)i + hr, ru(x)i. (3) 2 2 After substituting g for 12 ru, we obtain the associated Stein operator

(TP g)(x) , hg(x), r log p(x)i + hr, g(x)i. (4) The Stein operator TP is particularly well-suited to our setting as it depends on P only through the derivative of its log density and hence is computable even when the normalizing constant of p is not. If we let @X denote the boundary of X (an empty set when X = Rd) and n(x) represent the outward unit normal vector to the boundary at x, then we may define the classical Stein set ? ? ?? rg(y)k ? ? krg(x) Gk?k , g : X ! Rd sup max kg(x)k , krg(x)k , ? 1 and kx yk x6=y2X hg(x), n(x)i = 0, 8x 2 @X with n(x) defined of sufficiently smooth functions satisfying a Neumann-type boundary condition. The following proposition ? a consequence of integration by parts ? shows that Gk?k is a suitable domain for TP . Proposition 1. If EP [kr log p(Z)k] ¡ 1, then EP [(TP g)(Z)] = 0 for all g 2 Gk?k . Together, TP and Gk?k form the classical Stein discrepancy S(Q, TP , Gk?k), our chief object of study. 3.2

Lower Bounding the Classical Stein Discrepancy

In the univariate setting (d = 1), it is known for a wide variety of targets P that the classical Stein discrepancy S(?m , TP , Gk?k) converges to zero only if the Wasserstein distance dWk?k (?m , P) does [9, 10]. In the multivariate setting, analogous statements are available for multivariate Gaussian targets [11, 12, 13], but few other target distributions have been analyzed. To extend the reach of the multivariate literature, we show in Theorem 2 that the classical Stein discrepancy also determines Wasserstein convergence for a large class of strongly log-concave densities, including the Bayesian logistic regression posterior under Gaussian priors. Theorem 2 (Stein Discrepancy Lower Bound for Strongly Log-concave Densities). If X = Rd , and log p is strongly concave with third and fourth derivatives bounded and continuous, then, for any probability measures (?m)m 1 , S(?m , TP , Gk?k) ! 0 only if dWk?k (?m , P) ! 0. We emphasize that the sufficient conditions in Theorem 2 are certainly not necessary for lower bounding the classical Stein discrepancy. We hope that the theorem and its proof will provide a template for lower bounding S(Q, TP , Gk?k) for other large classes of multivariate target distributions. 3

3.3

Upper Bounding the Classical Stein Discrepancy

We next establish sufficient conditions for the convergence of the classical Stein discrepancy to zero. Proposition 3 (Stein Discrepancy Upper Bound). If X ? Q and Z ? P with r log p(Z) integrable, ? ? S(Q, TP , Gk?k) ? E[kX Zk] + E[kr log p(X) r log p(Z)k] + E r log p(Z)(X Z)¿ r h i i h 2 2 ? E[kX Zk] + E[kr log p(X) r log p(Z)k] + E kr log p(Z)k E kX Zk . One implication of Proposition 3 is that S(Qm , TP , Gk?k) converges to zero whenever Xm ?

Qm converges in mean-square to Z ? P and r log p(Xm) converges in mean to r log p(Z). 3.4

Extension to Non-uniform Stein Sets

The analyses and algorithms in this paper readily accommodate non-uniform Stein sets of the form ? ? ? kg(x)k? krg(x)k? krg(x) rg(y)k? , , ? 1 and c1:3 d supx6=y2X max c c c kx yk 1 2 3 Gk?k , g : X ! R (5) hg(x), n(x)i = 0, 8x 2 @X with n(x) defined

for constants c1 , c2 , c3 ¿ 0 known as Stein factors in the literature. We will exploit this additional flexibility in Section 5.2 to establish tight lower-bounding relations between the Stein discrepancy and Wasserstein distance for well-studied target distributions. For general use, however, we advocate the parameter-free classical Stein set and graph Stein sets to be introduced in the sequel. Indeed, any non-uniform Stein discrepancy is equivalent to the classical Stein discrepancy in a strong sense: Proposition 4 (Equivalence of Non-uniform Stein Discrepancies). For any c1 , c2 , c3 ¿ 0, c1:3 min(c1 , c2 , c3)S(Q, TP , Gk?k) ? S(Q, TP , Gk?k) ? max(c1 , c2 , c3)S(Q, TP , Gk?k).

4

Computing Stein Discrepancies

In this section, we introduce an efficiently computable Stein discrepancy with convergence properties equivalent to those of the classical discrepancy. We restrict attention to the unconstrained domain X = Rd in Sections 4.1-4.3 and present extensions for constrained domains in Section 4.4. 4.1

Graph Stein Discrepancies

Evaluating a Stein discrepancy S(Q, TP , G) for a fixed (Q, P) pair reduces to solving an optimization program over functions g 2 G. For example, the classical Stein discrepancy is the optimum Pn S(Q, TP , Gk?k) = sup i=1 q(xi)(hg(xi), r log p(xi)i + hr, g(xi)i) (6) g

?

?

s.t. kg(x)k ? 1, krg(x)k ? 1, krg(x)

?

rg(y)k ? kx

yk, 8x, y 2 X .

Note that the objective associated with any Stein discrepancy S(Q, TP , G) is linear in g and, since Q is discrete, only depends on g and rg through their values at each of the n sample points xi . The primary difficulty in solving the classical Stein program (6) stems from the infinitude of constraints imposed by the classical Stein set Gk?k . One way to avoid this difficulty is to impose the classical smoothness constraints at only a finite collection of points. To this end, for each finite graph G = (V, E) with vertices V ? X and edges E ? V 2 , we define the graph Stein set, ? ? ? Gk?k,Q,G , g : X ! Rd — 8 x 2 V, max kg(x)k , krg(x)k ? 1 and, 8 (x, y) 2 E, max

?

kg(x) g(y)k? krg(x) rg(y)k? kg(x) g(y) rg(x)(x y)k? kg(x) g(y) rg(y)(x y)k? , , , 2 2 1 1 kx yk kx yk 2 kx yk 2 kx yk

?

?1 ,

the family of functions which satisfy the classical constraints and certain implied Taylor compatibility constraints at pairs of points in E. Remarkably, if the graph G1 consists of edges between all distinct sample points xi , then the associated complete graph Stein discrepancy S(Q, TP , Gk?k,Q,G1) is equivalent to the classical Stein discrepancy in the following strong sense. 4

Proposition 5 (Equivalence of Classical and Complete Graph Stein Discrepancies). If X = Rd , and G1 = (supp(Q), E1) with E1 = {(xi , xl) 2 supp(Q)2 : xi 6= xl }, then S(Q, TP , Gk?k) ? S(Q, TP , Gk?k,Q,G1) ? ?d S(Q, TP , Gk?k), where ?d is a constant, independent of (Q, P), depending only on the dimension d and norm k?k.

Proposition 5 follows from the Whitney-Glaeser extension theorem for smooth functions [14, 15] and implies that the complete graph Stein discrepancy inherits all of the desirable convergence properties of the classical discrepancy. However, the complete graph also introduces order n2 constraints, rendering computation infeasible for large samples. To achieve the same form of equivalence while enforcing only O(n) constraints, we will make use of sparse geometric spanner subgraphs. 4.2

Geometric Spanners

For a given dilation factor t 1, a t-spanner [16, 17] is a graph G = (V, E) with weight kx yk on each edge (x, y) 2 E and a path between each pair x0 6= y 0 2 V with total weight no larger than tkx0 y 0 k. The next proposition shows that spanner Stein discrepancies enjoy the same convergence properties as the complete graph Stein discrepancy. Proposition 6 (Equivalence of Spanner and Complete Graph Stein Discrepancies). If X = Rd , Gt = (supp(Q), E) is a t-spanner, and G1 = (supp(Q), {(xi , xl) 2 supp(Q)2 : xi 6= xl }), then S(Q, TP , Gk?k,Q,G1) ? S(Q, TP , Gk?k,Q,Gt) ? 2t2 S(Q, TP , Gk?k,Q,G1).

Moreover, for any ‘p norm, a 2-spanner with O(?d n) edges can be computed in O(?d n log(n)) expected time for ?d a constant depending only on d and k?k [18]. As a result, we will adopt a 2-spanner Stein discrepancy, S(Q, TP , Gk?k,Q,G2), as our standard quality measure. 4.3

Decoupled Linear Programs

The final unspecified component of our Stein discrepancy is the choice of norm k?k. We recommend the ‘1 norm, as the resulting optimization problem decouples into d independent finite-dimensional linear programs (LPs) that can be solved in parallel. More precisely, S(Q, TP , Gk?k1 ,Q,(V,E)) equals Pd P—V — sup (7) j=1 i=1 q(vi)(ji rj log p(vi) + jji) j 2R

—V — ,

j 2R

d?—V —

s.t. k j k1 ? 1, k j k1 ? 1, and 8 i 6= l : (vi , vl) 2 E, ? — — k j (ei el)k1 — ji jl h j ei ,vi vl i— — max kvjii vljlk , kv , , 1 kv v k2 i vl k 1

1

2

i

ji

l 1

We have arbitrarily numbered the elements vi of the vertex set V so that value gj (vi), and jki represents the gradient value rk gj (vi). 4.4

jl h j el ,vi 2 1 2 kvi vl k1

ji

vl i—

?

? 1.

represents the function

Constrained Domains

A small modification to the unconstrained formulation (7) extends our tractable Stein discrepancy computation to any domain defined by coordinate boundary constraints, that is, to X = (?1 , 1) ? ? ? ? ? (?d , d) with 1 ? ?j ¡ j ? 1 for all j. Specifically, for each dimension j, we augment the j-th coordinate linear program of (7) with the boundary compatibility constraints ? ? — — — — — (v b)— max —vij jibj — , —vijjkibj — , ji 1 (vjji bij)2 j ? 1, for each i, bj 2 {?j , j } R, and k 6= j. (8) 2

ij

j

These additional constraints ensure that our candidate function and gradient values can be extended to a smooth function satisfying the boundary conditions hg(z), n(z)i = 0 on @X . Proposition 15 in the appendix shows that the spanner Stein discrepancy so computed is strongly equivalent to the classical Stein discrepancy on X . Algorithm 1 summarizes the complete solution for computing our recommended, parameter-free spanner Stein discrepancy in the multivariate setting. Notably, the spanner step is unnecessary in the univariate setting, as the complete graph Stein discrepancy S(Q, TP , Gk?k1 ,Q,G1) can be computed directly by sorting the sample and boundary points and only enforcing constraints between consecutive points in this ordering. Thus, the complete graph Stein discrepancy is our recommended quality measure when d = 1, and a recipe for its computation is given in Algorithm 2. 5

Algorithm 1 Multivariate Spanner Stein Discrepancy input: Q, coordinate bounds (?1 , 1), . . . , (?d , d) with 1 ? ?j ¡ j ? 1 for all j G2 Compute sparse 2-spanner of supp(Q) for j = 1 to d do (in parallel) rj Solve j-th coordinate linear program (7) with graph G2 and boundary constraints (8) Pd return j=1 rj

Algorithm 2 Univariate Complete Graph Stein Discrepancy

input: Q, bounds (?,) with 1 ? ? ¡ ? 1 (x(1) , . . . , x(n0)) S ORT({x1 , . . . , xn , ?, } R) P n0 d return sup 2Rn0 , 2Rn0 i=1 q(x(i))(i dx log p(x(i)) + i) ? ? s.t. k k1 ? 1, 8i ? n0 , — i — ? I ? ¡ x(i) ¡ , and, 8i ¡ n0 , ? — i — i i+1 i (x(i) x(i+1))— — i i+1 i i+1 — i+1 — max x—(i+1) , 1 1 x(i) , x(i+1) x(i) , (x x)2 (x (i+1)

2

5

(i)

2

i+1 (x(i) (i+1)
x(i)
x(i+1))—)2
?
?1
Experiments

We now turn to an empirical evaluation of our proposed quality measures. We compute all spanners using the efficient C++ greedy spanner implementation of Bouts et al. [19] and solve all optimization programs using Julia for Mathematical Programming [20] with the default Gurobi 6.0.4 solver [21]. All reported timings are obtained using a single core of an Intel Xeon CPU E5-2650 v2 @ 2.60GHz. 5.1

A Simple Example

We begin with a simple example to illuminate a few properties of the Stein diagnostic. For the target P = N (0, 1), we generate a sequence of sample points i.i.d. from the target and a second sequence i.i.d. from a scaled Student?s t distribution with matching variance and 10 degrees of freedom. The left panel of Figure 1 shows that the complete graph Stein discrepancy applied to the first n Gaussian sample points decays to zero at an n 0.52 rate, while the discrepancy applied to the scaled Student?s t sample remains bounded away from zero. The middle panel displays optimal Stein functions g recovered by the Stein program for different sample sizes. Each g yields a test function h , TP g, featured in the right panel, that best discriminates the sample Q from the target P . Notably, the Student?s t test functions exhibit relatively large magnitude values in the tails of the support. 5.2

Comparing Discrepancies

We show in Theorem 14 in the appendix that, when d = 1, the classical Stein discrepancy is the optimum of a convex quadratically constrained quadratic program with a linear objective, O(n) variables, and O(n) constraints. This offers the opportunity to directly compare the behavior of the graph and classical Stein discrepancies. We will also compare to the Wasserstein distance dWk?k ,

? ? ??? ?
?
g
??
?? ? ?
?
100
1000
?
10000
Number of sample points, n
? ?

? ?

? ?? ?? ? ? ?? ??? ? ? ? ? ? ? ? ?

? ? ???? ? ? ? ? ? ? ? ? ? ? ? ?? ?? ? ? ? ? ?? ? ? ? ? ? ? ? ? ? ?? ? ? ? ? ?? ?

?6 ?3 0

3

6 ?6 ?3 0

x

3

6

4 2 0 ?2 5.0 2.5 0.0 ?2.5

?

?

? ?? ?? ?? ?

? ?? ?? ? ? ? ?? ?? ? ?

?6 ?3 0

3

Sample ?

Gaussian Scaled Student's t

n = 30000

?

? ?

n = 30000

0.01

?

?

? ?

n = 3000

?

0.03

?

2 1 0 ?1 ?2

h = TP g
?
n = 3000
Stein discrepancy
?
Scaled Student's t
Gaussian
? ?
n = 300
?
1.0 0.5 0.0 ?0.5 ?1.0 1.0 0.5 0.0 ?0.5 ?1.0 1.0 0.5 0.0 ?0.5 ?1.0
n = 300
?
0.10
Scaled Student's t
Gaussian
?
6 ?6 ?3 0
3 6
x

Figure 1: Left: Complete graph Stein discrepancy for a N (0, 1) target. Middle / right: Optimal Stein functions g and discriminating test functions h = TP g recovered by the Stein program. 6

seed = 8
seed = 9 ?
? ?
?
? ?
? ?
?? ? ?
?
?
? ?
? ?
?
?
? ?
0.03 0.01
?
? ? ?
? ? ??? ?
? ???
?
?

? ? ?
?? ? ?
?
? ? ?
? ???
?
?
? ?
???
? ?
?
Gaussian
0.030 0.010
? ?
? ?
?
?
?
??
?
? ?
? ? ? ??
0.003 0.001
?
? ? ? ?
?
? ???
?
?
?
?
? ? ? ??? ?
?
?
?
Discrepancy ?
Classical Stein Wasserstein
?
?
? ?
?
?
?
? ?
? ??? ?
Uniform

Discrepancy value
seed = 7 0.30 0.10
??
???
Complete graph Stein
?
?
?
100
1000
10000
100
1000
10000
100
1000
10000
Number of sample points, n

Figure 2: Comparison of discrepancy measures for sample sequences drawn i.i.d. from their targets. which is computable for simple univariate target distributions [22] and provably lower bounds the non-uniform Stein discrepancies (5) with c1:3 = (0.5, 0.5, 1) for P = Unif(0, 1) and c1:3 = (1, 4, 2) for P = N (0, 1) [9, 23]. For N (0, 1) and Unif(0, 1) targets and several random number generator seeds, we generate a sequence of sample points i.i.d. from the target distribution and plot the nonuniform classical and complete graph Stein discrepancies and the Wasserstein distance as functions of the first n sample points in Figure 2. Two apparent trends are that the graph Stein discrepancy very closely approximates the classical and that both Stein discrepancies track the fluctuations in Wasserstein distance even when a magnitude separation exists. In the Unif(0, 1) case, the Wasserstein distance in fact equals the classical Stein discrepancy because TP g = g 0 is a Lipschitz function. 5.3

Selecting Sampler Hyperparameters

Stochastic Gradient Langevin Dynamics (SGLD) [3] with constant step size ? is a biased MCMC procedure designed for scalable inference. It approximates the overdamped Langevin diffusion, but, because no Metropolis-Hastings (MH) correction is used, the stationary distribution of SGLD deviates increasingly from its target as ? grows. If ? is too small, however, SGLD explores the sample space too slowly. Hence, an appropriate choice of ? is critical for accurate posterior inference. To illustrate the value of the Stein diagnostic for this task, we adopt the bimodal Gaussian mixture model (GMM) posterior of [3] as our target. For a range of step sizes ?, we use SGLD with minibatch size 5 to draw 50 independent sequences of length n = 1000, and we select the value of ? with the highest median quality ? either the maximum effective sample size (ESS, a standard diagnostic based on autocorrelation [1]) or the minimum spanner Stein discrepancy ? across these sequences. The average discrepancy computation consumes 0.4s for spanner construction and 1.4s per coordinate

linear program. As seen in Figure 3a, ESS, which does not detect distributional bias, selects the largest step size presented to it, while the Stein discrepancy prefers an intermediate value. The rightmost plot of Figure 3b shows that a representative SGLD sample of size n using the ? selected by ESS is greatly overdispersed; the leftmost is greatly underdispersed due to slow mixing. The middle sample, with ? selected by the Stein diagnostic, most closely resembles the true posterior. 5.4

Quantifying a Bias-Variance Trade-off

The approximate random walk MH (ARWMH) sampler [5] is a second biased MCMC procedure designed for scalable posterior inference. Its tolerance parameter ? controls the number of datapoint likelihood evaluations used to approximate the standard MH correction step. Qualitatively, a larger ? implies fewer likelihood computations, more rapid sampling, and a more rapid reduction of variance. A smaller ? yields a closer approximation to the MH correction and less bias in the sampler stationary distribution. We will use the Stein discrepancy to explicitly quantify this bias-variance trade-off. We analyze a dataset of 53 prostate cancer patients with six binary predictors and a binary outcome indicating whether cancer has spread to surrounding lymph nodes [24]. Our target is the Bayesian logistic regression posterior [1] under a N (0, I) prior on the parameters. We run RWMH (? = 0) and ARWMH (? = 0.1 and batch size = 2) for 105 likelihood evaluations, discard the points from the first 103 evaluations, and thin the remaining points to sequences of length 1000. The discrepancy computation time for 1000 points averages 1.3s for the spanner and 12s for a coordinate LP. Figure 4 displays the spanner Stein discrepancy applied to the first n points in each sequence as a function of the likelihood evaluation count. We see that the approximate sample is of higher Stein quality for smaller computational budgets but is eventually overtaken by the asymptotically exact sequence. 7

diagnostic = ESS

Step size, ? = 5e?05

Step size, ? = 5e?03

Step size, ? = 5e?02

? ? ?

2.0

?

4

?

3

1.0

2 1

?

?

?

diagnostic = Spanner Stein

? ?

?
?2 ?3 ?4 1e?04
1e?03
1e?02
?2 ?1
Step size, ?
(a) Step size selection criteria
0
Spanner Stein discrepancy
0.3
?
0.2 ?
16
? ?
?? ? ? ? ? ? ? ? ? ? ? ? ? ? ???? ? ? ???? ? ???????? ? ??? ?? ? ?? ? ? ?? ?? ?? ??
0.1
3e+03 1e+04 3e+04 1e+05
3
?2 ?1
0
x1
1
2
3
?2 ?1
0
1
2
; Stein discrepancy minimized at ? = 5 ? 10
?
3
1.0
?
2.0 ?
1.5
? ? ?
??? ? ? ?? ? ?? ?? ?? ??? ???? ? ?? ? ? ? ? ? ? ? ? ? ? ? ? ? ??? ???? ?? ? ? ??? ?????
0.5
3e+03 1e+04 3e+04 1e+05
??
? ?? ??? ? ?? ? ? ? ????? ?? ? ? ? ? ? ? ? ? ?? ? ? ? ? ? ? ??? ? ?? ?? ?? ?? ???? ???
.
? ?

?
3
Second moment error 2.5
?
? ?
2
Mean error
?
0.4
20
2
Normalized prob. error
?
24
1
?
(b) 1000 SGLD sample points with equidensity contours of p overlaid
Figure 3: (a) ESS maximized at ? = 5 ? 10 Discrepancy
? ? ??? ?? ? ? ? ? ?? ? ? ?? ? ?? ? ? ? ? ? ? ? ? ? ? ? ? ? ?? ?? ? ? ? ?
? ?? ? ?? ? ? ? ? ? ? ?? ? ? ? ? ? ? ? ?? ? ? ? ? ? ??? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ?? ? ? ? ? ? ?? ? ? ? ? ? ? ?? ? ? ? ? ? ? ? ? ? ? ? ?? ? ? ? ? ?
? ? ? ? ? ? ? ? ?? ? ? ? ? ? ??? ? ? ?? ?? ? ?? ? ? ? ? ? ? ? ? ?? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ? ? ? ?? ? ? ? ? ? ? ? ? ? ? ? ? ? ?? ? ??? ? ?? ? ? ? ?
?? ?? ? ? ? ? ? ?? ? ? ? ? ? ?? ? ??? ? ? ? ? ? ?? ? ? ? ? ? ? ? ? ?? ?? ??
? ? ? ? ? ? ?? ? ? ? ? ?? ? ? ? ? ?? ? ? ? ?? ? ? ?? ? ? ? ? ?? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ?? ? ? ? ?? ? ? ? ? ? ? ? ? ? ? ? ?? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?? ? ? ? ? ? ? ? ? ? ? ? ?? ? ? ? ? ?? ? ?? ? ?
? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?? ? ? ?? ? ?? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?? ? ? ? ?
? ? ? ? ?? ?? ? ? ? ? ? ? ? ? ? ?? ? ? ?? ? ?? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?? ? ?? ? ? ?? ? ? ? ? ? ? ?? ? ? ? ? ? ? ? ?
? ? ? ? ? ?? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?? ? ?? ? ? ? ? ? ? ? ? ? ? ? ? ??
? ? ? ? ? ? ?? ? ? ? ? ?? ? ? ? ? ? ? ?? ? ? ? ? ?? ? ? ? ? ? ?? ? ? ?? ?
? ?? ? ? ? ? ? ? ? ? ? ? ?? ? ? ? ? ? ? ? ? ? ? ?? ? ?? ? ?? ? ? ?? ?
0 ?1
3.0 2.5 2.0 1.5 1.0
?
? ? ? ? ? ? ?? ? ?? ? ? ?? ? ? ? ? ?? ? ? ??? ? ? ? ?? ?? ? ???? ? ?? ?
? ? ? ? ? ? ? ? ? ? ? ?? ?? ? ?? ? ? ?? ?? ?? ??? ? ?? ? ? ? ?? ?? ? ????
? ? ?? ? ?? ?? ? ?? ? ??? ?? ?? ? ? ? ? ? ? ? ?? ? ?? ? ? ????? ? ?? ? ????
?? ? ?? ? ? ? ??? ?? ? ?? ?? ? ?????? ? ?? ? ??? ?? ?? ?? ??? ???? ? ? ? ?
? ? ?? ?? ? ? ? ? ? ? ? ? ? ? ?? ? ? ? ?? ? ?? ??? ??? ? ? ???? ???? ? ? ?
? ? ? ?? ? ? ? ? ? ??? ? ?? ? ? ? ?? ?? ??? ? ?? ? ? ? ?? ? ?? ? ?? ? ? ? ?
?? ?? ? ? ?? ?? ?? ?? ?? ?? ?? ? ?
??? ? ? ? ? ? ? ? ?? ?? ?? ?? ? ? ? ?? ? ?? ??? ? ? ? ? ?? ?? ?? ?? ? ?? ?
? ? ?? ? ?? ?? ?? ? ??? ? ??? ??? ? ? ??? ???? ? ??? ?? ? ?? ?? ? ?? ? ? ?

?? ? ??? ? ? ?? ? ? ? ??? ? ?? ? ? ? ?? ? ?? ? ??? ?? ? ? ? ? ? ? ?? ?? ?? ? ? ? ??? ????? ?? ???? ? ?? ? ? ? ? ? ?? ??? ?? ???? ?? ?? ? ? ? ? ??? ? ? ?? ? ? ? ? ? ? ? ??? ?? ? ? ? ? ? ? ? ? ? ?? ? ?? ? ? ???? ? ? ? ? ? ??? ? ? ??? ? ? ?? ? ? ? ? ? ?? ? ?? ? ? ? ? ? ? ? ? ? ? ?? ? ?? ? ?? ? ? ? ? ?? ?? ? ? ? ? ? ?? ? ?? ? ? ? ?? ? ??? ? ? ? ? ? ?? ??? ? ? ? ? ?? ? ? ?? ? ? ? ? ? ? ? ? ? ? ? ?? ? ?? ? ? ? ?? ? ? ? ? ? ?? ? ?? ? ? ?? ?? ? ? ? ? ??? ?? ?? ? ? ? ?? ? ? ? ? ?? ? ????? ? ? ? ? ?? ?????? ?? ?? ?? ?? ?? ? ? ? ? ? ??? ? ? ??? ?? ? ?? ? ? ??? ? ? ? ?? ? ?? ? ? ??? ? ? ? ? ? ? ?? ? ? ? ?? ??? ???? ? ? ? ?? ? ?? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?? ? ? ?? ? ?? ? ? ? ?? ? ? ?? ? ? ? ? ?? ?? ? ??? ? ? ?? ?? ?? ? ? ? ? ? ? ????? ? ? ?? ? ? ???? ??? ? ? ? ? ??? ?? ? ??? ? ? ? ?? ? ? ??? ? ? ? ? ? ? ? ? ? ? ? ? ?? ? ? ? ?? ? ? ?? ? ?? ? ? ? ? ? ? ? ? ?? ? ?? ?? ?

?

1.5

x2

Log median diagnostic

2.5

1.0 0.5

3e+03 1e+04 3e+04 1e+05

Hyperparameter

? ? ?

??

? ? ?? ?? ? ? ?? ? ?? ???? ?? ? ? ? ? ? ? ? ? ? ? ?? ? ? ? ? ????? ?? ? ?? ?????? ?

?

?=0 ? = 0.1

3e+03 1e+04 3e+04 1e+05

Number of likelihood evaluations

Figure 4: Bias-variance trade-off curves for Bayesian logistic regression with approximate RWMH. To corroborate our result, we use a Metropolis-adjusted Langevin chain [25] of length 107 as a surrogate Q? for the target and compute several error measures for each sample Q: normalized probability max —E[X Z]— error maxl —E[(hX, wl i) (hZ, wl i)]—/kwl k1 , mean error maxjj —EQ?j [Zjj]— , and second moment max

—E[X X

Z Z]—

j k j k j,k error max for X ? Q, Z ? Q? , (t) , 1+e1 t , and wl the l-th datapoint covariate j,k —EQ? [Zj Zk]— vector. The measures, also found in Figure 4, accord with the Stein discrepancy quantification.

5.5

Assessing Convergence Rates

The Stein discrepancy can also be used to assess the quality of deterministic sample sequences. In Figure 5 in the appendix, for P = Unif(0, 1), we plot the complete graph Stein discrepancies of the first n points of an i.i.d. Unif(0, 1) sample, a deterministic Sobol sequence [26], and a deterministic R1 kernel

herding sequence [27] defined by the norm khkH = 0 (h0 (x))2 dx. We use the median value over 50 sequences in the i.i.d. case and estimate the convergence rate for each sampler using the slope of the best least squares affine fit to each log-log plot. The discrepancy computation time averages 0.08s for n = 200 points, andpthe recovered rates of n 0.49 and n 1 for the i.i.d. and Sobol sequences accord with expected O(1/ n) and O(log(n)/n) bounds from the literature [28, 26]. As 0.96 witnessed also in other outpaces its best known bound of p metrics [29], the herding rate of n dH (Qn , P) = O(1/ n), suggesting an opportunity for sharper analysis.

6

Discussion of Related Work

We have developed a quality measure suitable for comparing biased, exact, and deterministic sample sequences by exploiting an infinite class of known target functionals. The diagnostics of [30, 31] also account for asymptotic bias but lose discriminating power by considering only a finite collection of functionals. For example, for a N (0, 1) target, the score statistic of [31] cannot distinguish two samples with equal first and second moments. Maximum mean discrepancy (MMD) on a characteristic Hilbert space [32] takes full distributional bias into account but is only viable when the expected kernel evaluations are easily computed under the target. One can approximate MMD, but this requires access to a separate trustworthy ground-truth sample from the target. Acknowledgments The authors thank Madeleine Udell, Andreas Eberle, and Jessica Hwang for their pointers and feedback and Quirijn Bouts, Kevin Buchin, and Francis Bach for sharing their code and counsel. 8

2 References

[1] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. Handbook of Markov chain monte carlo. CRC press, 2011. [2] C. J. Geyer. Markov chain monte carlo maximum likelihood. Computer Science and Statistics: Proc. 23rd Symp. Interface, pages 156?163, 1991. [3] M. Welling and Y.-W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In Proceedings of the 28th International Conference on Machine Learning, pages 681?688, 2011. [4] S. Ahn, A. Korattikara, and M. Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. In Proceeding of 29th International Conference on Machine Learning (ICML?12), 2012. [5] A. Korattikara, Y. Chen, and M. Welling. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In Proceeding of 31th International Conference on Machine Learning (ICML?14), 2014. [6] A. M?uller. Integral probability metrics and their generating classes of functions. Advances in Applied Probability, 29(2):pp. 429?443, 1997. [7] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory, pages 583?602, Berkeley, CA, 1972. University of California Press. [8] A. D. Barbour. Stein?s method and Poisson process convergence. J. Appl. Probab.,

(Special Vol. 25A): 175?184, 1988. A celebration of applied probability. [9] L. HY. Chen, L. Goldstein, and Q.-M. Shao. Normal approximation by Steins method. Springer Science & Business Media, 2010. [10] S. Chatterjee and Q.-M. Shao. Nonnormal approximation by Steins method of exchangeable pairs with application to the Curie?Weiss model. Annals of Applied Probability, 21(2):464?483, 2011. [11] G. Reinert and A. R?ollin. Multivariate normal approximation with Steins method of exchangeable pairs under a general linearity condition. Annals of Probability, 37(6):2150?2173, 2009. [12] S. Chatterjee and E. Meckes. Multivariate normal approximation using exchange-able pairs. Alea, 4: 257?283, 2008. [13] E. Meckes. On Steins method for multivariate normal approximation. In High dimensional probability V: The Luminy volume, pages 153?178. Institute of Mathematical Statistics, 2009. ? [14] G. Glaeser. Etude de quelques alg'ebres tayloriennes. J. Analyse Math., 6:1?124; erratum, insert to 6 (1958), no. 2, 1958. [15] P. Shvartsman. The Whitney extension problem and Lipschitz selections of set-valued mappings in jetspaces. Transactions of the American Mathematical Society, 360(10):5529?5550, 2008. [16] P. Chew. There is a planar graph almost as good as the complete graph. In Proceedings of the Second Annual Symposium on Computational Geometry, SCG ?86, pages 169?177, New York, NY, 1986. ACM. [17] D. Peleg and A. A. Sch?affer. Graph spanners. Journal of Graph Theory, 13(1):99?116, 1989. [18] S. Har-Peled and M. Mendel. Fast construction of nets in low-dimensional metrics and their applications. SIAM Journal on Computing, 35(5):1148?1184, 2006. [19] Q. W. Bouts, A. P. ten Brink, and K. Buchin. A framework for computing the greedy spanner. In Proceedings of the Thirtieth Annual Symposium on Computational Geometry, SOCG?14, pages 11:11? 11:19, New York, NY, 2014. ACM. [20] M. Lubin and I. Dunning. Computing in operations research using Julia. INFORMS Journal on Computing, 27(2):238?248, 2015. [21] Gurobi Optimization. Gurobi optimizer reference manual, 2015. URL http://www.gurobi.com. [22] S. S. Vallender. Calculation of the Wasserstein distance between probability distributions on the line. Theory of Probability & Its Applications, 18(4):784?786, 1974. [23] C. D?obler. Stein?s method of exchangeable pairs for the Beta distribution and generalizations. arXiv:1411.4477, 2014. [24] A. Canty and B. D. Ripley. boot: Bootstrap R (S-Plus) Functions, 2015. R package version 1.3-15. [25] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. Bernoulli, pages 341?363, 1996. [26] R. E. Caflisch. Monte carlo and quasi-monte carlo methods. Acta numerica, 7:1?49, 1998. [27] Y. Chen, M. Welling, and A. Smola. Super-samples from kernel herding. In Proceeding of 26th Uncertainty in Artificial Intelligence (UAI?10), 2010. [28] E. del Barrio, E. Gin, and C. Matrn. Central limit theorems for the Wasserstein distance between the empirical and the true distributions. Ann. Probab., 27(2):1009?1071, 04 1999. [29] F. Bach, S. Lacoste-Julien, and G. Obozinski. On the equivalence between herding and conditional gradient algorithms. In Proceeding of 29th International Conference on Machine Learning (ICML?12), 2012. [30] A. Zellner and C.-K. Min. Gibbs sampler convergence criteria. Journal of the American Statistical Association, 90(431):921?927, 1995. [31] Y. Fan, S. P. Brooks, and A. Gelman.

Output assessment for monte carlo simulations via the score statistic. Journal of Computational and Graphical Statistics, 15(1), 2006. [32] A. Gretton, K. M Borgwardt, M. Rasch, B. Sch?olkopf, and A. J. Smola. A kernel method for the twosample-problem. In Advances in Neural Information Processing Systems, pages 513?520, 2006.

9