METROPOLIS-HASTINGS VIA CLASSIFICATION

By Tetsuya Kaji and Veronika Ročková

The University of Chicago

Booth School of Business
5807 South Woodlawn Avenue
Chicago, IL 60637
tkaji@ChicagoBooth.edu
Veronika.Rockova@ChicagoBooth.edu

This paper develops a Bayesian computational platform at the interface between posterior sampling and optimization in models whose marginal likelihoods are difficult to evaluate. Inspired by adversarial optimization, namely Generative Adversarial Networks (GAN) [18], we reframe the likelihood function estimation problem as a classification problem. Pitting a Generator, who simulates fake data, against a Classifier, who tries to distinguish them from the real data, one obtains likelihood (ratio) estimators which can be plugged into the Metropolis-Hastings algorithm. The resulting Markov chains generate, at a steady state, samples from an approximate posterior whose asymptotic properties we characterize. Drawing upon connections with empirical Bayes and Bayesian mis-specification, we quantify the convergence rate in terms of the contraction speed of the actual posterior and the convergence rate of the Classifier. Asymptotic normality results are also provided which justify inferential potential of our approach. We illustrate the usefulness of our approach on simulated data.

1. Introduction. Many contemporary statistical applications require inference for models which are easy to simulate from but whose likelihoods are impossible to evaluate. This includes implicit (simulator-based) models [9], defined through an underlying generating mechanism, or models prescribed through intractable likelihood functions.

Statistical inference for intractable models has traditionally relied on some form of likelihood approximation (see [20] for a recent survey). For example, [9] propose kernel log-likelihood estimates obtained from simulated realizations of an implicit model. Approximate Bayesian Computation (ABC) [5, 33, 38] is another simulation-based approach which obviates the need for likelihood evaluations by (1) generating fake data $\widetilde{X}_\theta$ for parameter values θ sampled from a prior, and (2) weeding out those pairs $(\widetilde{X}_\theta, \theta)$ for which $\widetilde{X}_\theta$ has low fidelity to observed data. The discrepancy between observed and fake data is evaluated by first reducing the two datasets to a vector of summary statistics and then measuring the distance between them. Both the distance function and the summary statistics are critical for inferential success. [11] proposed a semi-automated approach that approximates the posterior mean (a summary statistic that guarantees first-order accuracy) using a linear model regressing parameter samples onto simulated data. [22] elaborated on this strategy using deep neural networks which are expected to yield better approximations to the posterior mean. Beyond the choice of summary

MSC2020 subject classifications: 60J05, 62C10, 62F15.

Keywords and phrases: Approximate Bayesian Computation, Classification, Generative Adversarial Networks, Likelihood-free Inference, Metropolis-Hastings Algorithm, Markov Chain Monte Carlo.

statistics, ABC has to be deployed with caution for Bayesian model choice [36, 27]. Synthetic likelihood (SL) [43, 32] is another approach for carrying out inference in intractable models by constructing a proxy Gaussian likelihood for a vector of summary statistics. Implicit in the success of both ABC and SL is the assumption that the generating process can produce simulated summary statistics that adequately represent the observed ones. If this compatibility is not satisfied (e.g. in misspecified models), both SL [12] and ABC [13] can provide unreliable estimates. Avoiding the need for summary statistics, [21] proposed to use discriminability of the observed and simulated data as a discrepancy measure in ABC. Their accepting/rejecting mechanism separates samples based on a discriminator's ability to tell the real and fake data apart. Similarly as their work, our paper is motivated by the observation that distinguishing two data sets is usually easier if they were simulated with very different parameter values. However, instead of deploying this strategy inside ABC, we embed it directly inside the Metropolis-Hastings algorithm.

The Metropolis-Hastings (MH) method generates ergodic Markov chains through an accept-reject mechanism which depends in part on likelihood ratios comparing proposed candidate moves and current states. For many latent variable models, the marginal likelihood is *not* available in closed form making direct application of MH impossible (see [8] for examples). The pseudo-marginal likelihood method [1] offers a remedy by replacing likelihood evaluations with their importance sampling estimates. Many variants of this approach have been proposed to correct for bias [29], reduce the variance of the likelihood ratio estimator [8] or to make sure that the resulting chain produces samples from the actual (not only approximate) posterior [4]. The idea of using likelihood approximations within MH dates back to at least [29] and has been implemented in a series of works (see e.g [30] and [4] and references therein).

Our approach is fundamentally different from existing approximate MH algorithms. Inspired by adversarial optimization, namely the Generative Adversarial Networks (GAN) [18], we reframe the likelihood (ratio) estimation problem as classification problem. Similarly as with GANs, we pit two agents (a Generator and a Classifier) against one another. Assessing the similitude between the fake data, outputted by the Generator, and observed data, the Classifier provides likelihood estimators which can be deployed inside MH. The resulting algorithm provides samples from an approximate posterior. GANs have been successful in learning distributions over complex objects, such as images, and have been coupled with MH in [40] to sample from the likelihood. Their method is an elaboration of the discriminator rejection sampling [3] which uses an importance sampling post-processing step using information from a trained discriminator. These two strategies are very different from our proposal here which is concerned with posterior inference about model parameters in a likelihood-free environment.

Our contributions are both methodological and theoretical. We develop a personification of Metropolis-Hastings algorithm for intractable likelihoods based on Classification, further referred to as MHC. We consider two variants: (1) a fixed generator design which may yield biased samples, and (2) a random generator design which may yield unbiased samples with increased variance. We then describe how and when the two can be combined in order to provide posterior samples with an asymptotically correct location and spread. Our theoretical analysis consists of new convergence rate results for a posterior residual (an approximation error) associated with the Classifier. These rates are then shown to affect the rate of convergence of the stationary distribution, in a similar way as the ABC tolerance level affects the convergence rate of ABC posteriors [14]. Theoretical characterizations of the related pseudo-marginal method have been, so far, limited to convergence properties of the Markov chain such as mixing rates [1, 8].

Here, we provide a rigorous asymptotic study of the stationary distribution including convergence rates (drawing upon connections to empirical Bayes and Bayesian misspecification), asymptotic normality results and, in addition, polynomial mixing time characterizations of the Markov chain. We demonstrate the usefulness of our approach on simulated data, including the famous Ricker model [34] analyzed earlier by multiple authors [20, 43, 11].

The paper is structured as follows. Section 2 and 3 introduce the classification-based likelihood ratio estimator and the MHC sampling algorithm. Section 4 then describes the asymptotic properties of the stationary distribution. Section 5 shows demonstrations on simulated data and, finally, Section 6 wraps up with a discussion.

Notation The following notation will be used throughout the manuscript. We employ the operator notation for expectation, e.g., $P_0 f = \int f dP_0$ and $\mathbb{P}_m^\theta f = \frac{1}{m}\sum_{i=1}^m f(X_i^\theta)$. The *$\varepsilon$-bracketing number* $N_{[]}(\varepsilon, \mathcal{F}, d)$ of a set $\mathcal{F}$ with respect to a premetric d is the minimal number of ε-brackets in d needed to cover $\mathcal{F}$.[1] The *δ-bracketing entropy integral* of $\mathcal{F}$ with respect to d is $J_{[]}(\delta, \mathcal{F}, d) := \int_0^\delta \sqrt{1 + \log N_{[]}(\varepsilon, \mathcal{F}, d)} d\varepsilon$. We denote the usual Hellinger semi-metric for independent observations as $d_n^2(\theta, \theta') = \frac{1}{n}\sum_{i=1}^n \int (\sqrt{p_{\theta,i}} - \sqrt{p_{\theta',i}})^2 \mathrm{d}\mu_i$. Next, $K(p_{\theta_0}^{(n)}, p_\theta^{(n)}) = \sum_{i=1}^n K(p_{\theta_0,i}, p_{\theta,i})$ denotes the Kullback-Leibler divergence between product measures and $V_2(f, g) = \int f |\log(f/g)|^2 \mathrm{d}\mu$. Define $\langle a, b\rangle = \sum_{i=1}^d a_i b_i$ for $a, b \in \mathbb{R}^d$.

2. Likelihood Estimation with a Classifier. Our framework consists of a series of i.i.d. observations $\{X_i\}_{i=1}^n \in \mathcal{X}$ realized from a probability measure P_{θ_0} indexed by a parameter $\theta_0 \in \Theta$ which is endowed with a prior $\Pi_n(\cdot)$. We assume that P_θ, for each $\theta \in \Theta$, admits a density p_θ. Our objective is to draw observations from the posterior density given $X^{(n)} = (X_1, \ldots, X_n)'$ defined through

$$\pi_n(\theta \mid X^{(n)}) = \frac{p_\theta^{(n)}(X^{(n)})\pi(\theta)}{\int_\Theta p_\vartheta^{(n)}(X^{(n)})\, d\Pi(\vartheta)}, \tag{2.1}$$

where $p_\theta^{(n)} = \prod_{i=1}^n p_\theta(X_i)$. Our focus is on situations where the likelihood $p_\theta^{(n)}$ is too costly to evaluate but can be readily sampled from.

We develop a Bayesian computational platform at the interface between sampling and optimization inspired by Generative Adversarial Networks (GAN) [18]. The premise of GAN's is to discover rich distributions over complex objects arising in artificial intelligence applications through simulation. The learning procedure consists of two entities pitted against one another. A Generator aims to deceive an Adversary by simulating samples that resemble the observed data while, at the same time, the Adversary learns to tell the fake and real data apart. This process iterates until the generated data are indistinguishable by the Adversary. While GAN's have found their usefulness in simulating from distributions over images, here we forge new connections to Bayesian posterior simulation.

Similarly as with GAN's, we assume a Generator transforming a set of latent variables $\widetilde{X} \in \widetilde{\mathcal{X}}$ to collect samples from P_θ through a deterministic mapping $T_\theta : \widetilde{\mathcal{X}} \to \mathcal{X}$, i.e. $T_\theta(\widetilde{X}) \sim P_\theta$ for $\widetilde{X} \sim \widetilde{P}$ for some distribution $\widetilde{P}$ on $\widetilde{\mathcal{X}}$. This implies that we can draw one set of m observations $\widetilde{X}^{(m)}$ and filter them through T_θ to obtain a sample $\widetilde{X}_\theta^{(m)} = T_\theta(\widetilde{X}^{(m)})$ from P_θ for any $\theta \in \Theta$. A similar latent variable framework has been

[1] A *premetric* on $\mathcal{F}$ is a function $d : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ such that $d(f, f) = 0$ and $d(f, g) = d(g, f) \geq 0$.

considered in the pseudo-marginal literature [8] where $\widetilde{X}^{(m)}$ are latent variables in a hierarchical models with intractable marginal likelihoods. There is no shortage of examples of such hierarchical models in population genetics [38], econometrics [19], or ecology [4] (including the famous Ricker model [34] analyzed later in Section 5.1). Our approach differs considerably from pseudo-marginal methods. It rests on the availability of a cross-entropy classifier (used by the Adversary in the GAN framework) which can be deployed to obtain an estimator of the likelihood.

The classification problem with the cross-entropy loss is defined through

$$\max_{D\in\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\log D(X_i)+\frac{1}{m}\sum_{i=1}^{m}\log(1-D(X_i^\theta))\right], \tag{2.2}$$

where $\mathcal{D}$ is a set of measurable classification functions $D:\mathcal{X}\to[0,1]$ (1 for 'real' and 0 for 'fake' data) and where $X_i^\theta = T_\theta(\widetilde{X}_i)$ for $X_i\sim\widetilde{P}$ for $i=1,\dots,m$ are the 'fake' data outputted by the Generator. If an oracle were to furnish the true model p_0, the solution to (2.2) can be shown to satisfy (see [18, Proposition 1])

$$D_\theta(X) := \frac{p_{\theta_0}(X)}{p_{\theta_0}(X)+p_\theta(X)} \quad \text{for } X\in\mathcal{X}. \tag{2.3}$$

Reorganizing the terms in (2.3), the likelihood can be written in terms of the discriminator function as

$$p_\theta^{(n)}(X^{(n)}) = p_{\theta_0}^{(n)}(X^{(n)})\exp\left(\sum_{i=1}^{n}\log\frac{1-D_\theta(X_i)}{D_\theta(X_i)}\right). \tag{2.4}$$

The oracle discriminator $D_\theta(\cdot)$ depends on p_0 but can be estimated by simulation. Indeed, one can deploy the Generator to simulate the fake data $\widetilde{X}_\theta^{(m)} = T_\theta(\widetilde{X}^{(m)})$ and train a Classifier to distinguish them from $X^{(n)}$. The Classifier outputs an estimate $\hat{D}_{n,m}^\theta$ which can be plugged into (2.4) to obtain the following likelihood estimator

$$\widehat{p}_\theta^{(n)}(X^{(n)}) = p_{\theta_0}^{(n)}(X^{(n)})\exp\left(\sum_{i=1}^{n}\log\frac{1-\hat{D}_{n,m}^\theta(X_i)}{\hat{D}_{n,m}^\theta(X_i)}\right) = p_\theta^{(n)}(X^{(n)})\mathrm{e}^{u_\theta(X^{(n)})}, \tag{2.5}$$

where

$$u_\theta(X^{(n)}) := \sum_{i=1}^{n}\left(\log\frac{1-\hat{D}_{n,m}^\theta}{1-D_\theta}-\log\frac{\hat{D}_{n,m}^\theta}{D_\theta}\right) \tag{2.6}$$

will be further referred to as the log-posterior residual. In other words, (2.5) is a deterministic functional of auxiliary random variables $\widetilde{X}^{(m)}$ and the observed data $X^{(n)}$ and can be computed (up to a norming constant) from $\hat{D}_{n,m}^\theta$. The posterior density $\pi_n(\theta\mid X^{(n)})$ can be then estimated by replacing D_θ with $\hat{D}_{n,m}^\theta$ in the likelihood expression to obtain

$$\widehat{\pi}_{n,m}(\theta\mid X^{(n)}) := \exp\left(\sum_{i=1}^{n}\log\frac{1-\hat{D}_{n,m}^\theta(X_i)}{\hat{D}_{n,m}^\theta(X_i)}\right)\pi(\theta) \propto \pi_n(\theta\mid X^{(n)})\mathrm{e}^{u_\theta(X^{(n)})}. \tag{2.7}$$

Two observations ought to be made. First, the estimator (2.7) targets the posterior density only up to a norming constant. This will not be an issue in Bayesian algorithms involving posterior density ratios (such as the Metropolis-Hastings algorithm considered here). Second, the estimator (2.7) performs exponential tilting of the original posterior, where the quality of the approximation crucially depends on the statistical properties of $u_\theta(X^{(n)})$. Note that $u_\theta(X^{(n)})$ depends also on the latent data $\widetilde{X}_\theta^{(m)}$. We devote the entire Section 4.1 to statistical properties of $u_\theta(X^{(n)})$.

3. Metropolis Hastings via Classification. The Metropolis-Hastings (MH) algorithm is one of the mainstays of Bayesian computation. The deployment of unbiased likelihood estimators within MH has shown great promise in models whose likelihoods are not available [5, 1, 2]. In the previous section, we have shown how classification can be deployed to obtain estimates of likelihood ratios. This suggests a compelling question: *Can we deploy these classification-based estimators within MH?* This section explores this intriguing possibility and introduces a new MH variant that we further refer to as MHC, Metropolis Hastings via Classification.

Our objective is to simulate values from an (approximate) posterior distribution $\Pi_n(\cdot \mid X^{(n)})$ with a density $\pi_n(\theta \mid X^{(n)}) \propto p_\theta^{(n)}(X^{(n)})\pi(\theta)$ over $(\Theta, \mathscr{B})$ using the MH routine. Recall that MH simulates a Markov chain according to the transition kernel

$$K(\theta, \theta') := \rho(\theta, \theta')q(\theta' \mid \theta) + \delta_\theta(\theta') \int_\Theta (1 - \rho(\theta, \tilde{\theta}))q(\tilde{\theta} \mid \theta)\mathrm{d}\tilde{\theta},$$

where

$$\rho(\theta, \theta') := \min\left\{ \frac{p_{\theta'}^{(n)}(X^{(n)})\pi(\theta')}{p_\theta^{(n)}(X^{(n)})\pi(\theta)} \frac{q(\theta \mid \theta')}{q(\theta' \mid \theta)}, 1 \right\}. \tag{3.1}$$

and where $q(\cdot \mid \theta)$ is a proposal density generating candidate values θ' for the next move.

It is often the case in practice that we cannot directly evaluate $p_\theta^{(n)}(X^{(n)})$ but have access to its (unbiased) estimator. The pseudo-marginal approach [1] deploys an importance sampling estimator obtained via data augmentation through the introduction of auxiliary latent variables, say $\widetilde{X}_\theta^{(m)}$. In its simplest form (Monte Carlo Within Metropolis (MCWM) described in [1]), this method requires independently simulating m replicates of the auxiliary data for each likelihood evaluation at each iteration. Other variants have been suggested where latent data are recycled from the previous iterations (Grouped Independence MH (GIMH) described in [5]) or using correlated latent variables are for the numerator and the denominator of the acceptance ratio [8].

In this work, we propose replacing $p_\theta^{(n)}$ in the acceptance ratio (3.1) with the classification-based likelihood estimator (2.5) outlined in Section 2. This estimator, similarly as with the pseudo-marginal method, also relies on the introduction of latent variables $\widetilde{X}_\theta^{(m)}$. As we have seen earlier, it can be rewritten in terms of the estimated discriminator as

$$\widehat{p}_\theta^{(n)}(X^{(n)}) \propto \exp\left(\sum_{i=1}^n \log \frac{1 - \hat{D}_{n,m}^\theta(X_i)}{\hat{D}_{n,m}^\theta(X_i)} \right). \tag{3.2}$$

The evaluation of $\widehat{p}_\theta^{(n)}(X^{(n)})$ can be carried out by merely computing $\hat{D}_{n,m}^\theta(X_i)$ where $\hat{D}_{n,m}^\theta$ is a trained classifier distinguishing $X^{(n)}$ from $\widetilde{X}_\theta^{(m)}$. Putting the pieces together, one can replace the intractable likelihood ratio in the acceptance probability (3.1) with

$$\rho_u(\theta, \theta') := \min\left\{ \frac{\widehat{p}_{\theta'}^{(n)}(X^{(n)})\pi(\theta')}{\widehat{p}_\theta^{(n)}(X^{(n)})\pi(\theta)} \frac{q(\theta \mid \theta')}{q(\theta' \mid \theta)}, 1 \right\}. \tag{3.3}$$

Note that the proportionality constant in the likelihood expression (3.2) cancels out in (3.3), allowing $\rho_u(\theta, \theta')$ to be directly computable. We consider two variants. The first one, called a *fixed generator design*, assumes that the randomness of $\hat{D}_{n,m}^\theta$, for each given θ and $X^{(n)}$, is determined by latent variables $\widetilde{X}^{(m)}$ *shared* by all steps of the algorithm. This corresponds to the case when m auxiliary data points $\widetilde{X}_\theta^{(m)} =$

$\{\widetilde{X}_i^\theta\}_{i=1}^m$ are obtained through a *deterministic* mapping $\widetilde{X}_i^\theta = T_\theta(\widetilde{X}_i)$ for some $\widetilde{X}_i \sim \widetilde{P}$ that are not changed throughout the algorithm. The second version, called a *random generator design*, assumes that the underlying latent variables variables $\widetilde{X}^{(m)} = \{\widetilde{X}_i\}_{i=1}^m$ are refreshed at each step. While the difference between these two versions is somewhat subtle, we will see important bias-variance implications. We provide details for each of these two variants below.

3.1. *Fixed Generator MHC.* The fixed generator algorithm is summarized in Algorithm 1 below.

ALGORITHM 1 (Fixed Generator). Draw $\widetilde{X}^{(m)} = \{\widetilde{X}_i\}_{i=1}^m \sim \widetilde{P}$. Initialize $\theta^{(0)}$, generate $\widetilde{X}_{\theta^{(0)}}^{(m)}$ and repeat the following.

(1) Given $\theta^{(t)}$, generate $\theta' \sim q(\cdot \,|\, \theta^{(t)})$.
(2) Generate $\widetilde{X}_{\theta'}^{(m)} = \{\widetilde{X}_i^{\theta'}\}_{i=1}^m$ according to $\widetilde{X}_i^{\theta'} = T_{\theta'}(\widetilde{X}_i)$.
(3) Compute $\hat{D}_{n,m}^{\theta'}$ from $X^{(n)}$ and $\widetilde{X}_{\theta'}^{(m)}$ and using $\rho_u(\cdot,\cdot)$ defined in (3.3) set

$$\theta^{(t+1)} = \begin{cases} \theta' & \text{with probability } \rho_u(\theta^{(t)}, \theta'), \\ \theta^{(t)} & \text{with probability } 1 - \rho_u(\theta^{(t)}, \theta'), \end{cases}$$

using $\widehat{p}_\theta(X^{(n)})$ defined in (3.2).

It is natural to inquire whether and how the likelihood approximation affects the stationary distribution of the resulting Markov chain. Due to the exponential tilt $\mathrm{e}^{u_\theta(X^{(n)})}$ in the likelihood approximation (2.5), Algorithm 1 does *not* yield the correct posterior $\pi_n(\theta \,|\, X^{(n)})$ at its steady state. Indeed, under standard assumptions (see Section 7.3.1 of [35]), the stationary distribution of the Markov chain, conditional on $\widetilde{X}^{(m)}$, writes as (see e.g. Theorem 7.2 in [35])

$$\pi_n^\star(\theta \,|\, X^{(n)}) = \frac{p_\theta^{(n)}(X^{(n)}) \times \mathrm{e}^{u_\theta(X^{(n)})} \times \pi(\theta)}{\int_\Theta p_\theta^{(n)}(X^{(n)}) \times \mathrm{e}^{u_\theta(X^{(n)})} \times \pi(\theta)\mathrm{d}\theta}. \tag{3.4}$$

We do not view his property as unsurmountable. Other approximate MH algorithms (e.g the MCWM pseudo-marginal method of [1]) *also* do not yield $\pi_n(\theta \,|\, X^{(n)})$ at stationarity. However, the samples generated by Algorithm 1 will be distributed according an approximate posterior (3.4) whose statistical properties we describe in detail in Section 4. In Section 3.4, we further quantify the speed of MHC convergence in large samples under the assumption of asymptotic normality. As will be seen in Section 4, the exponential tilt induces certain bias where the pseudo-posterior (3.4) concentrates around a projection of the true parameter θ_0. Despite the bias, the curvature of the approximate posterior can be shown to match the curvature of the actual posterior (under differentiability assumptions in Section 4.1). The random generator version, introduced in the next section, works the other way around. It can lead to a correct location (no bias) but at the expense of enlarged variance.

3.2. *Random Generator MHC.* The random generator variant proceeds as in Algorithm 1 but refreshes $\widetilde{X}^{(m)} \sim \widetilde{P}$ at each step before computing the acceptance ratio. We summarize the algorithm below, dropping the subscript m in $\widetilde{X}^{(m)}$ for simplicity.

ALGORITHM 2 (Random Generator). Initialize $\theta^{(0)}$ and $\widetilde{X}^{(0)}$ and repeat the following.

(1) Given $\theta^{(t)}$, generate $\theta' \sim q(\cdot \mid \theta^{(t)})$ and $\widetilde{X}' \sim \widetilde{q}(\widetilde{X}' \mid \widetilde{X}^{(t)})$.
(2) Generate $\widetilde{X}_{\theta'} = \{\widetilde{X}_i^{\theta'}\}_{i=1}^m$ according to $\widetilde{X}_i^{\theta'} = T_{\theta'}(\widetilde{X}_i')$.
(3) Compute $\hat{D}_{n,m}^{\theta'}$ from $X^{(n)}$ and $\widetilde{X}_{\theta'}'$ and using

$$\widetilde{\rho}_u(\theta, \widetilde{X}; \theta', \widetilde{X}') = \min\left\{ \frac{\widehat{p}_{\theta'}^{(n)}(X^{(n)})\pi(\theta')}{\widehat{p}_{\theta}^{(n)}(X^{(n)})\pi(\theta)} \frac{q(\theta \mid \theta')}{q(\theta' \mid \theta)} \frac{\widetilde{q}(\widetilde{X} \mid \widetilde{X}')}{\widetilde{q}(\widetilde{X}' \mid \widetilde{X})}, 1 \right\}. \tag{3.5}$$

with $\widehat{p}_{\theta}^{(n)}(X^{(n)})$ obtained from (3.2) set

$$(\theta^{(t+1)}, \widetilde{X}^{(t+1)}) = \begin{cases} (\theta', \widetilde{X}') & \text{with probability } \widetilde{\rho}_u(\theta, \widetilde{X}; \theta', \widetilde{X}'), \\ (\theta^{(t)}, \widetilde{X}^{(t)}) & \text{with probability } 1 - \widetilde{\rho}_u(\theta, \widetilde{X}; \theta', \widetilde{X}'). \end{cases}$$

To glean more insights into this variant, it is helpful to regard $(\theta^{(t)}, \widetilde{X}^{(t)})$ jointly as a Markov chain with an augmented proposal density $q(\theta', \widetilde{X}' \mid \theta, \widetilde{X}) = q(\theta' \mid \theta)\widetilde{q}(\widetilde{X}' \mid \widetilde{X})$ where $\widetilde{q}(\widetilde{X}' \mid \widetilde{X})$ possibly depends on $\widetilde{X}$. In order to make the dependence on $\widetilde{X}$ in $u_\theta(X^{(n)})$ more transparent, we will denote the posterior residual defined in (2.6) with $u_\theta(X^{(n)}, \widetilde{X})$ going forward. It can be seen that the marginal stationary distribution of the augmented Markov chain under Algorithm 2 equals

$$\widetilde{\pi}_n^\star(\theta \mid X^{(n)}) := \int \pi_n^\star(\theta \mid X^{(n)}) \mathrm{d}\widetilde{X}, \tag{3.6}$$

where $\pi_n^\star(\theta \mid X^{(n)})$ was defined earlier in (3.4) and depends on $\widetilde{X}$ through $u_\theta(X^{(n)}, \widetilde{X})$. The following characterization will be useful for establishing statistical properties of $\widetilde{\pi}_n^\star(\theta \mid X^{(n)})$ later in Section 4. From (3.4), we can write

$$\widetilde{\pi}_n^\star(\theta \mid X^{(n)}) \propto p_\theta^{(n)}(X^{(n)}) \times \mathrm{e}^{\widetilde{u}_\theta(X^{(n)})} \times \pi(\theta) \tag{3.7}$$

where

$$\widetilde{u}_\theta(X^{(n)}) = \log \int \mathrm{e}^{u_\theta(X^{(n)}, \widetilde{X})} \mathrm{d}\widetilde{X}. \tag{3.8}$$

Assuming almost-sure positivity of the joint proposal density $q(\theta', \widetilde{X}' \mid \theta, \widetilde{X})$, it can be verified (e.g. from Corollary 4.1 in [39]) that the marginal distribution of $\theta^{(t)}$ after t steps of Algorithm 2 converges in total variation to $\widetilde{\pi}_n^\star(\theta \mid X^{(n)})$. Interestingly, from (3.7) we can see that the stationary distribution (3.6) from Algorithm 2 has *the same functional form* as the stationary distribution (3.4) from Algorithm 1. The only difference is replacing $u_\theta(X^{(n)})$ with an averaged-out version $\widetilde{u}_\theta(X^{(n)})$ in (3.8). Integration may inflate the stationary distribution (3.6) by making it more spread-out compared to the fixed generator sampler. However, the exponential tilting factor $\widetilde{u}_\theta(X^{(n)})$ is averaged out. While $u_\theta(X^{(n)})$ in (2.6) is *fixed* in $\widetilde{X}$ (creating a non-vanishing bias term), $u_\theta(X^{(n)})$ in (3.8) can average out to 0 (depending on $\widetilde{q}(\cdot \mid \cdot)$), erasing the bias and yielding the actual posterior as the stationary distribution.

3.3. *Debiasing.* Algorithm 1 and 2 can be combined to produce a more realistic representation of the true posterior. We mentioned that Algorithm 1, under the differentiability assumptions, has the same curvature as the actual posterior but has a non-vanishing shift. Algorithm 2, on the other hand, has a reduced bias due to the averaging aspect in (3.8). We can thus diminish the bias of the fixed generator design by shifting the location towards the mean of samples obtained with the random generator. This leads to the following hybrid procedure.

ALGORITHM 3 (Bias correction).

1. Generate a sample $\{\theta_1^{(t)}\}_{t=1}^T$ using Algorithm 1.
2. Generate a sample $\{\theta_2^{(t)}\}_{t=1}^T$ using Algorithm 2.
3. Debias $\{\theta_1^{(t)}\}$ using $\{\theta_2^{(t)}\}$, that is, construct a sample $\{\theta^{(t)}\}$ by

$$\theta^{(t)} := \theta_1^{(t)} - \frac{1}{T}\sum_{s=1}^{T}\theta_1^{(s)} + \frac{1}{T}\sum_{s=1}^{T}\theta_2^{(s)}.$$

While Algorithms 1 and 2 can be deployed as a standalone, the de-biasing variant might increase the quality of the samples. Note that if $\widetilde{u}_\theta(X^{(n)}) = 0$, Algorithm 2 will be unbiased yielding the actual posterior as its stationary distribution. In Section 4.1 we theoretically justify Algorithm 3 by providing sufficient conditions under which it yields samples from an object which has the same limit as the actual posterior.

3.4. *Mixing Properties of MHC.* A critical issue for MCMC algorithms is the determination of the number of iterations needed for the result to be approximately a sample from the distribution of interest. This section sheds light on the mixing rate of Algorithm 1. Under standard assumptions on $q(\cdot \,|\, \cdot)$ (such as positivity almost surely, see Corollary 4.1 in [39]), the distribution of the MHC Markov chain after t steps will converge to $\pi_n^\star(\theta \,|\, X^{(n)})$ from any initialization in Θ in total variation as $t \to \infty$. [28] derive necessary and sufficient conditions for the Metropolis algorithms (with independent or symmetric candidate distributions) to converge at a geometric rate to a prescribed continuous distribution. [6] studied the speed of convergence of MH when both $n \to \infty$ *and* $d \to \infty$ where $\theta \in \Theta \subset \mathbb{R}^d$.

We can reformulate their sufficient conditions for showing polynomial mixing times of MHC. Recall that the stationary distribution $\pi_n^\star(\theta \,|\, X^{(n)})$ of the MHC sampler in (3.4) normalized to a compact set $K \subset \Theta$, writes as

$$\Pi_K^\star(B) = \int_B \pi_n^\star(\theta \,|\, X^{(n)}) / \int_K \pi_n^\star(\theta \,|\, X^{(n)}).$$

We are interested in bounding the number of steps needed to draw a random variable from Π_K^* with a given precision. We denote with Π_K^{*t} the distribution obtained after t steps of the MHC algorithm starting from Π_K^{*0}. It is known (see e.g. [26]) that the total variation distance between Q and Q_t can be bounded by

$$\|\Pi_K^* - \Pi_K^{*t}\|_{TV} \leq \sqrt{M}(1 - \phi^2/2)^t,$$

where M is a constant which depends on the initial distribution Π_K^{*0} and ϕ is the *conductance* of the Markov chain defined, e.g., in (3.13) in [6]. To obtain bounds on the conductance, the Markov chain needs to transition somewhat smoothly (see assumption D1 and D2 in [6]). These assumptions pertain to the continuity of the transitioning measure and are satisfied by the Gaussian random walk with a suitable choice of the proposal variance (see Section 3.2.4 in [6]) The following Lemma summarizes Theorem 2 of [6] in the context of Algorithm 1 under asymptotic normality assumptions examined in more detail in Section 4.5.

LEMMA 3.1. *(Mixing Rate) Under Assumptions (4.16)-(4.17) and a Gaussian random walk $q(\cdot \,|\, \cdot)$ satisfying Lemma 4 of [6], the global conductance ϕ of the Markov chain obtained from Algorithm 1 satisfies $1/\phi = \mathcal{O}(d)$ in $P_{\theta_0}^{(n)}$-probability. In addition, the minimal number of MCMC iterations needed to achieve $\|\Pi_K^* - \Pi_K^{*t}\|_{TV} < \epsilon$ is $\mathcal{O}(d^2 \log(M/\epsilon))$ for some suitable constant M depending on the initial distribution Π_K^{*0}.*

MHC thus attains bounds on the mixing rate that are *polynomial* in d (i.e. rapid mixing) under suitable Bernstein-von Mises conditions formalized later in Section 4.5. This section investigates how fast the Markov chain converges to its target $\pi_n^\star(\theta \,|\, X^{(n)})$ as the number of iterations t grows. In Section 4.3 (resp. Section 4.4), we investigate a fundamentally different question. We assess the speed at which the target $\pi_n^\star(\theta \,|\, X^{(n)})$ shrinks around the truth θ_0 (resp. a Kullback-Leibler projection) as n grows.

4. Properties of the Stationary Distribution. We now shift attention from the computational aspects of MHC to its potential as a statistical inference procedure. To understand the qualitative properties of the MHC scheme, we provide an asymptotic study of its stationary distribution, drawing upon its connections to empirical Bayes methods (Section 4.3) and Bayesian misspecification (Section 4.4). Before delving into the stationary distribution, however, we first derive rates of convergence for the posterior residual $u_\theta(X^{(n)})$ in (2.6) which plays a fundamental role.

4.1. *Convergence of the Posterior Residual.* We denote the sample objective function in (2.2) with $\mathbb{M}_{n,m}^\theta(D) := \mathbb{P}_n \log D + \mathbb{P}_m^\theta \log(1-D)$, where we employed the operator notation for expectation, e.g., $\mathbb{P}_n f = \frac{1}{m}\sum_{i=1}^m f(X_i)$ and $\mathbb{P}_m^\theta f = \frac{1}{m}\sum_{i=1}^m f(X_i^\theta)$. Throughout this section, we will use a simplified notation u_θ instead of $u_\theta(X^{(n)})$ and similarly for p_θ and $p_\theta^{(n)}$. We denote by P the probability measure that encompasses all randomness, e.g., as $O_P(1)$.[2] The estimated Classifier is seen to satisfy

$$\hat{D}_{n,m}^\theta := \max_{D \in \mathcal{D}_n} \mathbb{M}_{n,m}^\theta(D)$$

where $\mathcal{D}_n$ constitutes a sieve of classifiers that expands with the sample size and that is not too rich (as measured by the bracketing entropy $N_{[]}(\varepsilon, \mathcal{F}, d)$). In practice, the estimator $\hat{D}_{n,m}^\theta$ can be obtained by deploying a variety of classifiers ranging from logistic regression to deep learning (see Assumption 3 in [23] for a sieve construction using neural network classifiers). The discrepancy between two classifiers will be measured by a Hellinger-type distance

$$d_\theta(D_1, D_2) := \sqrt{h_\theta(D_1, D_2)^2 + h_\theta(1-D, 1-D_\theta)^2},$$

where $h_\theta(D_1, D_2) = \sqrt{(P_{\theta_0} + P_\theta)(\sqrt{D_1} - \sqrt{D_2})^2}$ (see [23] and [31] for more discussion about this discrepancy). The rate of convergence of the Classifier was previously established by [23] under assumptions reviewed below. In the following, we denote with $\mathcal{D}_{n,\delta}^\theta := \{D \in \mathcal{D}_n : d_\theta(D, D_\theta) \leq \delta\}$ the neighborhood of the oracle classifier within the sieve.

Assumption 1. Assume that n/m converges and that an estimator $\hat{D}_{n,m}^\theta$ exists that satisfies $\mathbb{M}_{n,m}^\theta(\hat{D}_{n,m}^\theta) \geq \mathbb{M}_{n,m}^\theta(D_\theta) - O_P(\delta_n^2)$ for a nonnegative sequence δ_n. Moreover, assume that the bracketing entropy integral[3] satisfies $J_{[]}(\delta_n, \mathcal{D}_{n,\delta_n}^\theta, d_\theta) \lesssim \delta_n^2\sqrt{n}$ and that there exists $\alpha < 2$ such that $J_{[]}(\delta, \mathcal{D}_{n,\delta}^\theta, d_\theta)/\delta^\alpha$ is decreasing in δ.

Under Assumption 1, [23] conclude the following convergence rate result for the classifier established.

[2] We may think of this P as the "canonical representation" [42, Problem 1.3.4].

[3] See the notation section

THEOREM 4.1 (23, Theorem 1'). *Let Assumption 1 hold for a given $\theta \in \Theta$, then* $d_\theta(\hat{D}^\theta_{n,m}, D_\theta) = O^*_P(\delta_n)$.

While [23] focused mainly on the convergence of $\hat{D}^\theta_{n,m}$, here we move the theoretical investigation of GAN's further by establishing the rate of convergence of $u_\theta(\cdot)/n$ as well as its limiting shape. To this end, we assume the following support compatibility assumption, a refinement of the bounded likelihood ratio condition in nonparametric maximum likelihood (Theorem 3.4.4 in [42] and Lemma 8.7 in [15]).

ASSUMPTION 2. There exists $M > 0$ such that for every $\theta \in \Theta$, $P_{\theta_0}(p_{\theta_0}/p_\theta)$ and $P_{\theta_0}(p_{\theta_0}/p_\theta)^2$ are bounded by M and

$$\sup_{D \in \mathcal{D}^\theta_{n,\delta_n}} P_{\theta_0}\left(\frac{D_\theta}{D} \,\middle|\, \frac{D_\theta}{D} \geq \frac{25}{16}\right) < M, \quad \sup_{D \in \mathcal{D}^\theta_{n,\delta_n}} P_{\theta_0}\left(\frac{1-D_\theta}{1-D} \,\middle|\, \frac{1-D_\theta}{1-D} \geq \frac{25}{16}\right) < M$$

for δ_n in Assumption 1. The brackets in Assumption 1 can be taken so that $P_{\theta_0}\left(\sqrt{\frac{u}{\ell}} - 1\right)^2 = O(d_\theta(u,\ell)^2)$ and $P_{\theta_0}\left(\sqrt{\frac{1-\ell}{1-u}} - 1\right)^2 = o(d_\theta(u,\ell))$.

The following Theorem will be crucial for understanding theoretical properties of our MHC sampling algorithm, where the rate of convergence of $u_\theta(\cdot)/n$ will be seen to affect the rate of convergence of the stationary distribution of our Markov chains.

THEOREM 4.2 (Posterior Residual Convergence Rate). *Let Assumptions 1 and 2 hold for a given $\theta \in \Theta$, then*

$$u_\theta/n = \mathbb{P}_n\left(\log \frac{1-\hat{D}^\theta_{n,m}}{1-D_\theta} - \log \frac{\hat{D}^\theta_{n,m}}{D_\theta}\right) = O^*_P(\delta_n).$$

PROOF. Section 7.1

One seemingly pessimistic conclusion from Theorem 4.2 is that $u_\theta(\cdot)$ *does not* vanish. [23] show that if the true likelihood ratio has a low-dimensional representation and an appropriate neural network is used for the discriminator, the rate δ_n depends only on the underlying dimension and not on the original dimension of X_i. In spite of the non-vanishing tilting term $u_\theta(X^{(n)})$, it turns out that Algorithm 1 can be refined (de-biased) to produce reasonable samples as long as $\hat{D}^\theta_{n,m}$ estimates the score well (see Section 3.3).

4.2. *Convergence of u_θ under Differentiability.* In this section, we show quadratic approximability for u_θ at a much faster rate than Theorem 4.2 when the model and the classifier are differentiable in some suitable sense.

ASSUMPTION 3 (Differentiability of p_θ). There exists $\theta_0 \in \Theta \subset \mathbb{R}^d$ such that $P_0 = P_{\theta_0}$. The model $\{p_\theta\}$ is *differentiable in quadratic mean at* θ_0, that is, there exists a measurable function $\dot{\ell}_{\theta_0} : \mathcal{X} \to \mathbb{R}^d$ such that[4]

$$\int \left[\sqrt{p_{\theta_0+h}} - \sqrt{p_{\theta_0}} - \frac{1}{2} h' \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}}\right]^2 = o(\|h\|^2).$$

[4]Integration is understood with respect to some dominating measure.

This is a classical assumption (see e.g. Section 5.5 of [41]) which implies local asymptotic normality.

EXAMPLE 1 (Normal Location-scale Model). Let $X_i \sim P_0 = N(0,1)$ and $P_\theta = N(\mu, \sigma^2)$ where $\theta = (\mu, \sigma^2)$ is the unknown parameter and $\theta_0 = (0,1)$ the true parameter. This model satisfies Assumption 3 with the score $\dot{\ell}_{\theta_0}(x) = \begin{bmatrix} x \\ (x^2-1)/2 \end{bmatrix}$ and the Fisher information matrix $I_{\theta_0} = \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}$.

Going back to (2.5), we write $\widehat{p}_\theta(X^{(n)}) = \prod_{i=1}^n \hat{p}_\theta(X_i)$, where

$$\hat{p}_\theta = p_{\theta_0} \frac{1 - \hat{D}^\theta_{n,m}}{\hat{D}^\theta_{n,m}} \tag{4.1}$$

is an estimator of p_θ that is possibly unscaled so that $\int \hat{p}_\theta$ may not be one. The scaling constant will be denoted by $c_\theta := \int \hat{p}_\theta$. In general, $\hat{p}_\theta$ is not observable since p_{θ_0} is not available. From (2.6), we can see that

$$u_\theta = n\mathbb{P}_n \log \frac{1 - \hat{D}^\theta_{n,m}}{\hat{D}^\theta_{n,m}} - n\mathbb{P}_n \log \frac{1 - D_\theta}{D_\theta} = n\mathbb{P}_n \log \frac{\hat{p}_\theta}{p_{\theta_0}} - n\mathbb{P}_n \log \frac{p_\theta}{p_{\theta_0}}.$$

Under Assumption 3, van der Vaart [41, Theorem 7.2] derives convergence of the second term above in the local neighborhood of θ_0. In Theorem 4.3 below, we derive convergence of the first term under the following assumption.

ASSUMPTION 4 (Differentiability of $\hat{p}_\theta$).

(i) The estimator $\hat{p}_\theta$ in (4.1) is *differentiable in quadratic mean in probability at* θ_0 *with a cubic rate*, that is,

$$\int \left[\sqrt{\hat{p}_{\theta_0+h}} - \sqrt{\hat{p}_{\theta_0}} - \frac{1}{2} h' \dot{\ell}_{\theta_0} \sqrt{\hat{p}_{\theta_0}} \right]^2 = O_P(\|h\|^3),$$

where $\dot{\ell}_{\theta_0} : \mathcal{X} \to \mathbb{R}^d$ is the score function in Assumption 3.

(ii) Dependence of $\mathbb{P}_n$ and $\hat{p}_\theta$ is asymptotically ignorable in the sense that for every compact $K \subset \mathbb{R}^d$,

$$\sup_{h \in K} \left| n(\mathbb{P}_n - P_{\theta_0}) \left(\sqrt{\frac{\hat{p}_{\theta_0+h/\sqrt{n}}}{\hat{p}_{\theta_0}}} - 1 - \frac{h' \dot{\ell}_{\theta_0}}{2\sqrt{n}} \right) \right| \longrightarrow 0,$$

$$\sup_{h \in K} \left| n(\mathbb{P}_n - P_{\theta_0}) \left(\sqrt{\frac{\hat{p}_{\theta_0+h/\sqrt{n}}}{\hat{p}_{\theta_0}}} - 1 \right)^2 \right| \longrightarrow 0$$

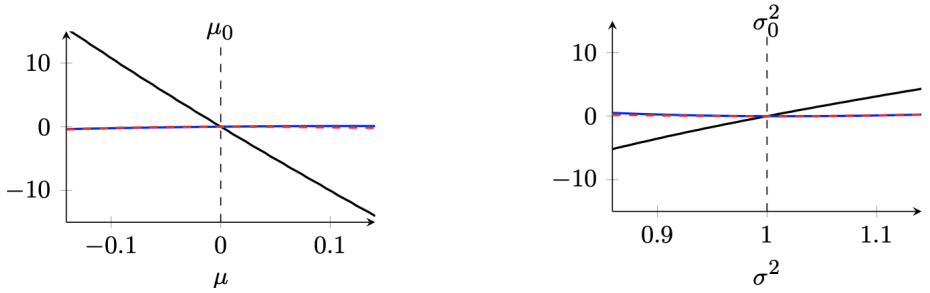
in outer probability.

(iii) The scaling factor is asymptotically linear in the sense that there exists a sequence of $\mathbb{R}^d$-valued random variables $\dot{c}_{n,\theta_0}$ such that for every compact $K \subset \mathbb{R}^d$,

$$\sup_{h \in K} \left| n(c_{\theta_0+h/\sqrt{n}} - c_{\theta_0}) - \sqrt{n} h' \dot{c}_{n,\theta_0} \right| \longrightarrow 0$$

in outer probability.

Assumption 4 (i) requires that $\hat{p}_\theta$ estimates the score well and is smoother than once differentiable. If $\hat{p}_\theta$ is twice differentiable in θ, then it holds with $O_P(\|h\|^4)$. Assumption 4 (ii) requires that the dependence of $\mathbb{P}_n$ and $\hat{p}_\theta$ be ignored asymptotically. If $\mathbb{P}_n$



(a) The black line $n(c_\theta - c_{\theta_0})$; the blue line $n(\mathbb{P}_n - P_{\theta_0})(\sqrt{\hat{p}_\theta/\hat{p}_{\theta_0}} - 1 - (\theta-\theta_0)'\dot{\ell}_{\theta_0}/2)$; the red line $n(\mathbb{P}_n - P_{\theta_0})(\sqrt{\hat{p}_\theta/\hat{p}_{\theta_0}} - 1)^2$. σ^2 is fixed at σ_0^2.

(b) The black line $n(c_\theta - c_{\theta_0})$; the blue line $n(\mathbb{P}_n - P_{\theta_0})(\sqrt{\hat{p}_\theta/\hat{p}_{\theta_0}} - 1 - (\theta-\theta_0)'\dot{\ell}_{\theta_0}/2)$; the red line $n(\mathbb{P}_n - P_{\theta_0})(\sqrt{\hat{p}_\theta/\hat{p}_{\theta_0}} - 1)^2$. μ is fixed at μ_0.

Fig 1: Illustration of Assumption 4 (ii–iii) in Example 1. $n = m = 5000$.

and $\hat{p}_\theta$ were independent, it would follow from Chebyshev's or Markov's inequality. Assumption 4 (iii) requires that the quadratic curvature of the scaling constant vanishes asymptotically.

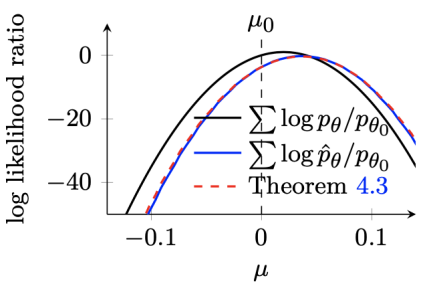
In general, Assumption 4 is not verifiable since the likelihood is not available. To illustrate the assumption, we verify that it holds in the toy example of the normal location-scale model from Example 1.

EXAMPLE 1 (Normal Location-scale Model, continuing from p. 11). The oracle discriminator of P_0 from P_θ is $D_\theta(x) = [1 + \exp(-\frac{1}{2}\log\sigma^2 + \frac{x^2}{2} - \frac{(x-\mu)^2}{2\sigma^2})]^{-1}$. Let us use the logistic regression using regressors $(1, x, x^2)$ to estimate D_θ, i.e., $D_\theta(x) = [1 + \exp(-\beta_0 - \beta_1 x - \beta_2 x^2)]^{-1}$. Thus, the true parameter for the logistic regression is $\beta = (\beta_0, \beta_1, \beta_2) = (\frac{1}{2}\log\sigma^2 + \frac{\mu^2}{2\sigma^2}, -\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2} - \frac{1}{2})$. Let $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ be the estimator of β. Then,

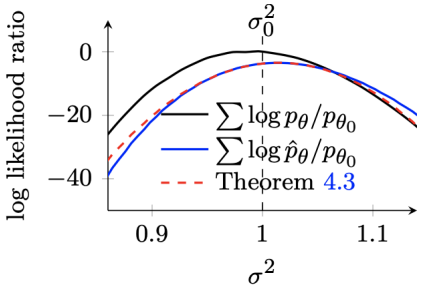
$$\hat{p}_\theta(x) = \frac{\exp(-\frac{x^2}{2} - \hat{\beta}_0 - \hat{\beta}_1 x - \hat{\beta}_2 x^2)}{\sqrt{2\pi}} \qquad \text{and} \quad c_\theta = \frac{\exp(-\hat{\beta}_0 + \frac{1}{2}\frac{\hat{\beta}_1^2}{1+2\hat{\beta}_2})}{\sqrt{1+2\hat{\beta}_2}}.$$

Being a MLE, $\hat{\beta}$ is regular and efficient, so $\sqrt{n}(\hat{\beta} - \beta) = \Delta + o_P(1)$ for a normal vector Δ. Moreover, if we generate X_i^θ through $X_i^\theta = \mu + \sigma\widetilde{X}_i$, $\widetilde{X}_i \sim N(0,1)$, there is one-to-one correspondence between $X_i^{\theta_1}$ and $X_i^{\theta_2}$ for every θ_1 and θ_2, so the dependence of Δ on θ disappears as $n \to \infty$ for otherwise a more efficient estimator exists to contradict efficiency. Therefore, the formula for $\hat{p}_\theta$ implies that Assumption 4 (i) is satisfied with the oracle score function $\dot{\ell}_{\theta_0}$; since $\hat{p}_\theta$ is twice differentiable, it holds with a faster rate of $O_P(\|h\|^4)$. Finally, we check Assumption 4 (ii) and (iii) by simulation. Figure 1 shows the supremands of Assumption 4 (ii) and (iii) as functions of $\theta = (\mu, \sigma^2)$. The black lines plot $n(c_\theta - c_{\theta_0})$ as we change θ; they are linear and its quadratic curvatures are ignorable. The blue lines represent $n(\mathbb{P}_n - P_{\theta_0})(\sqrt{\hat{p}_\theta/\hat{p}_{\theta_0}} - 1 - (\theta-\theta_0)'\dot{\ell}_{\theta_0}/2)$ and the red lines $n(\mathbb{P}_n - P_{\theta_0})(\sqrt{\hat{p}_\theta/\hat{p}_{\theta_0}} - 1)^2$; compared to the values of $n(c_\theta - c_{\theta_0})$, both are uniformly ignorable.

With Assumption 4, the estimated log likelihood asymptotes to a quadratic function that has the oracle curvature but a different center.



(a) True log likelihood, estimated log likelihood, and quadratic approximation by Theorem 4.3. $\sigma^2 = \sigma_0^2$.



(b) True log likelihood, estimated log likelihood, and quadratic approximation by Theorem 4.3. $\mu = \mu_0$.

Fig 2: Illustration of Theorem 4.3 in Example 1. $n = m = 5000$.

THEOREM 4.3. *Let p_θ and $\hat{p}_\theta$ satisfy Assumptions 3 and 4 and $\int(\sqrt{\hat{p}_{\theta_0}} - \sqrt{p_{\theta_0}})^2 = O_P(\delta_n^2)$ for some $\delta_n = o(n^{-1/4})$. Then, for every compact $K \subset \mathbb{R}^d$,*

$$\sup_{h\in K}\left| n\mathbb{P}_n \log \frac{\hat{p}_{\theta_0+h/\sqrt{n}}}{\hat{p}_{\theta_0}} + \frac{1}{2}h'I_{\theta_0}h - \sqrt{n}\mathbb{P}_n h'\dot{\ell}_{\theta_0} + \sqrt{n}\hat{P}_{\theta_0}h'\dot{\ell}_{\theta_0} - \sqrt{n}h'\dot{c}_{n,\theta_0}\right| \longrightarrow 0$$

in outer probability.

PROOF. Section 7.2

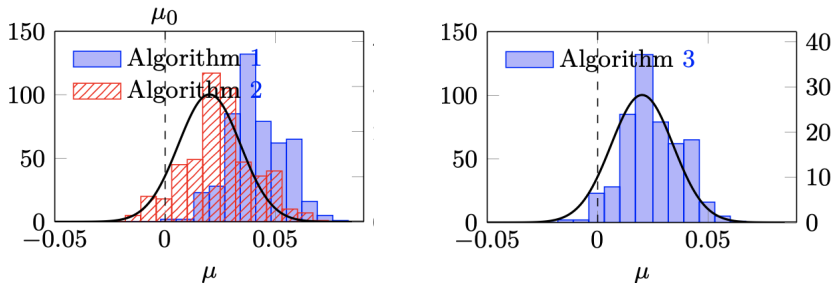
REMARK 1. (Linear u_θ) One important conclusion from this section is linearity of u_θ. Indeed, van der Vaart [41, Theorem 7.2] and Theorem 4.3 imply that

$$u_{\theta_0+h/\sqrt{n}} - u_{\theta_0} = h'\sqrt{n}(\dot{c}_{n,\theta_0} - \hat{P}_{\theta_0}\dot{\ell}_{\theta_0}) + o_P(1).$$

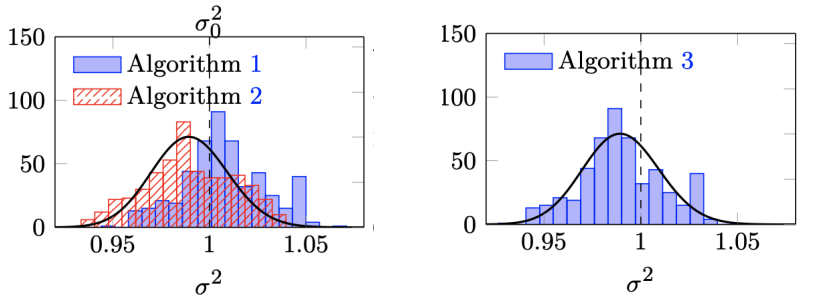
Thus, u_θ converges to a linear function in θ, though the leading linear term $\sqrt{n}(\dot{c}_{n,\theta_0} - \hat{P}_{\theta_0}\dot{\ell}_{\theta_0})$ may diverge if the convergence rate of $\hat{p}_{\theta_0}$ is slow. We revisit linearity of u_θ later in Section 4.4 (Example 2) as one of the sufficient conditions for the Bernstein-von Mises theorem.

The term $-\frac{1}{2}h'I_{\theta_0}h + \sqrt{n}\mathbb{P}_n h'\dot{\ell}_{\theta_0}$ is the quadratic curve to which the true log likelihood ratio converges. The linear term $h'\sqrt{n}(\dot{c}_{n,\theta_0} - \hat{P}_{\theta_0}\dot{\ell}_{\theta_0})$ shifts the center of the quadratic curve but not the curvature. The term $n\mathbb{P}_n \log \frac{\hat{p}_{\theta_0}}{p_{\theta_0}}$ changes the level of the quadratic curve. Theorem 4.3 has a very important consequence regarding the limiting shape of the stationary distribution $\pi_n^\star(\theta \,|\, X^{(n)})$ for Algorithm 1 defined in (3.4). It shows that $\pi_n^\star(\theta \,|\, X^{(n)})$ approaches a *biased* normal distribution with the *same variance* as the true posterior. In addition, we have seen in Section 3.2 that the stationary distribution $\widetilde{\pi}_n^\star(\theta \,|\, X^{(n)})$ of Algorithm 2 defined in (3.7) is averaged over the bias. Therefore, if $\mathbb{E}[\dot{c}_{n,\theta_0} - \hat{P}_{\theta_0}\dot{\ell}_{\theta_0} \,|\, X] = 0$, then the stationary distribution of Algorithm 3 converges to the correct normal posterior, i.e. it has the same limit as the actual posterior $\pi_n(\theta \,|\, X^{(n)})$. Theorem 4.3 thus provides a theoretical justification for de-biasing suggested in Section 3.3.

EXAMPLE 1 (Normal location-scale model, continuing from p. 11). Since this model with the logistic classifier satisfies Assumptions 3 and 4, it is susceptible to Theorem 4.3. This is supported by a diagnostics plot in Figure 2 which portrays true



(a) Algorithm 1 and 2. (b) Algorithm 3.



(c) Algorithm 1 and 2. (d) Algorithm 3.

Fig 3: Histograms of the MHC samples of μ and σ^2 in the normal location-scale model. Algorithm 1 (resp. 2) yield more biased (resp. dispersed) samples compared to the true posterior (black curve). Algorithm 3 (on the right) tracks the black curve more closely.

and estimated likelihood ratios. In Figure 2a, μ is varied with σ^2 fixed at σ_0^2 while, in Figure 2b, σ^2 is varied with μ held at μ_0. The difference between the estimated log likelihood (blue) and the quadratic approximation (dashed red) is negligible, demonstrating that the validity of Theorem 4.3 is justifiable. Compared to the oracle log likelihood (black), the estimated log likelihood is shifted by the random term $\sqrt{n}(\dot{c}_{n,\theta_0} - \hat{P}_{\theta_0}\dot{\ell}_{\theta_0})$. The curvature, however, is the same as oracle since the red line curves by the Fisher information I_{θ_0}. Thus, we expect Algorithm 1 to produce a biased sample and Algorithm 2 a dispersed sample. Note that we can compute $\sqrt{n}\hat{P}_{\theta_0}\dot{\ell}_{\theta_0} = c_{\theta_0}\sqrt{n}[-\frac{\hat{\beta}_1}{1+2\hat{\beta}_2}, -\frac{1}{2} + \frac{1}{2(1+2\hat{\beta}_2)} + \frac{\hat{\beta}_1^2}{2(1+2\hat{\beta}_2)^2}]'$, which is asymptotically linear in Δ by the delta method. It is then reasonable to expect that this term has mean zero when averaged over $\tilde{X}$ since $\hat{\beta}$ is asymptotically unbiased. If $\dot{c}_{n,\theta_0}$ also has mean zero, then Algorithm 2 is unbiased and Algorithm 3 recovers the exact normal posterior.

To see that this is indeed the case, we impose a conjugate normal-inverse-gamma prior, $\theta \sim N\Gamma^{-1}(\mu_0, \nu, \alpha, \beta)$, that is, the marginal prior of σ^2 is the inverse-gamma $\Gamma^{-1}(\alpha, \beta)$ and the conditional prior of μ given σ^2 is $N(\mu_0, \frac{\sigma^2}{\nu})$. The posterior is then analytically calculated as $\theta \mid X \sim N\Gamma^{-1}\big(\frac{\nu\mu_0 + n\bar{X}_n}{\nu+n}, \nu+n, \alpha+\frac{n}{2}, \beta + \frac{1}{2}\sum_i (X_i - \bar{X}_n)^2 + \frac{n\nu}{\nu+n}\frac{(\bar{X}_n - \mu_0)^2}{2}\big)$ for $\bar{X}_n := \frac{1}{n}\sum_i X_i$. Figure 3 shows the histograms of Algorithm 1,2 and 3 after $K = 500$ MCMC steps. Since the estimated log likelihood has a rightward bias (as seen from Figure 2), Algorithm 1 produces a sample that is shifted to the right (Figures 3a and 3c). Algorithm 2, on the other hand, gives a sample that is more dispersed than the posterior but is correctly placed, indicating that the random bias has mean zero. Consequently, Algorithm 3 generates a sample that is placed and shaped correctly (Figures 3b and 3d).

4.3. *Posterior Convergence Rates: Connection to Empirical Bayes.* Having quantified the convergence rate of the posterior residual $u_\theta(X^{(n)})$ in Theorem 4.2, we are now ready to explore the convergence rate of the entire stationary distribution without necessarily imposing differentiability assumptions.

Recall that the MHC sampler *does not* reach $\pi_n(\theta \,|\, X^{(n)})$ in steady state. Recall that the stationary distribution (using the fixed generator) takes the form

$$\Pi_n^\star(B \,|\, X^{(n)}) = \frac{\int_B p_\theta^{(n)}/p_{\theta_0}^{(n)} \times \mathrm{e}^{u_\theta} \times \pi(\theta)\mathrm{d}\theta}{\int_\Theta p_\theta^{(n)}/p_{\theta_0}^{(n)} \times \mathrm{e}^{u_\theta} \times \pi(\theta)\mathrm{d}\theta}. \tag{4.2}$$

In the random design, we simply replace u_θ in (4.2) with $\widetilde{u}_\theta$ defined in (3.8). Interestingly, (4.2) can be viewed as an actual posterior under a *tilted prior* with a density $\pi^*(\theta) \propto \mathrm{e}^{u_\theta}\pi(\theta)$. This shifted prior depends on the data $X^{(n)}$ (through $u_\theta(X^{(n)})$) and thereby (4.2) can be loosely regarded as an empirical Bayes (EB) posterior. While EB uses plug-in estimators of prior hyper-parameters, here the data enters the prior in a less straightforward manner. We further the EB connection later in Remark 3.

We first assess the quality of the posterior approximation (4.2) through its concentration rate around the *true parameter* value θ_0 using the traditional Hellinger semi-metric $d_n(\theta, \theta')$. The rate depends on the interplay between the concentration of the *actual* posterior[5] $\Pi_n(\theta \,|\, X^{(n)})$ and the rate at which the residual $u_\theta(X^{(n)})$ in (2.6) diverges. Recall that the rate of $u_\theta(\cdot)/n$ was established earlier in Theorem 4.2. The following Theorem uses assumptions on prior concentration around θ_0 using the typical Kullback-Leibler neighborhood $B_n(\theta_0, \epsilon) = \left\{\theta \in \Theta : K(p_{\theta_0}^{(n)}, p_\theta^{(n)}) \le n\epsilon^2, \frac{1}{n}\sum_{i=1}^n V_2(p_{\theta_0}(X_i), p_\theta(X_i)) \le \epsilon^2\right\}$.

THEOREM 4.4. *Consider the pseudo-posterior distribution $\Pi_n^\star$ defined through (4.2). Suppose that the prior $\Pi_n(\cdot)$ satisfies conditions (3.2) and (3.4) in [16] for a sequence $\varepsilon_n \to 0$ such that $n\varepsilon_n^2 \to \infty$. In addition, let $\widetilde{C}_n$ be such that*

$$P_{\theta_0}^{(n)}\left(\sup_{\theta\in\Theta} |u_\theta(X^{(n)})/n| > \widetilde{C}_n\varepsilon_n^2\right) = o(1) \tag{4.3}$$

and assume that for sets $\Theta_n \subset \Theta$ the prior satisfies

$$\frac{\Pi_n(\Theta\backslash\Theta_n)}{\Pi_n(B_n(\theta_0, \varepsilon_n))} = o(\mathrm{e}^{-2(1+\widetilde{C}_n)n\varepsilon_n^2}). \tag{4.4}$$

Then we have, for any $M_n \to \infty$ such that $\widetilde{C}_n = o(M_n)$,

$$P_{\theta_0}^{(n)}\left[\Pi_n^\star(\theta : d_n(\theta, \theta_0) > M_n\varepsilon_n \,|\, X^{(n)})\right] = o(1) \quad \text{as } n \to \infty.$$

PROOF. The proof is a minor modification of Theorem 4 in [16] and is postponed until Section 7.3.

REMARK 2. Assumption 4.3 can be slightly weakened by considering the aggregate behavior of $u_\theta(X^{(n)})$ around θ_0 with respect to the prior $\Pi_n(\cdot)$. It suffices to assume $P_{\theta_0}^{(n)}\left(I_n(\Pi_n, X^{(n)}, \varepsilon_n) \le \mathrm{e}^{-\widetilde{C}_n n\varepsilon_n^2}\right) = o(1)$ where $I_n(\Pi_n, X^{(n)}, \epsilon) = \int_{B_n(\theta_0,\epsilon)} \mathrm{e}^{u_\theta(X^{(n)})}\mathrm{d}\Pi_n(\theta)$ and that $P_{\theta_0}^{(n)}\left[\sup_{\Theta_n^c \cup d_n(\theta,\theta_0)>\epsilon} |u_\theta(X^{(n)})| > \widetilde{C}_n n\varepsilon_n^2\right] = o(1)$ for any $\epsilon > \varepsilon_n$. Assumption (4.4) is not needed if one is only interested in the concentration inside Θ_n.

[5]Using the usual notion [16], we say that the posterior $\Pi_n(\cdot \,|\, X^{(n)})$ concentrates around θ_0 at the rate ε_n (satisfying $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$) if $P_{\theta_0}\Pi_n[\theta \in \Theta : d_n(\theta, \theta_0) > M\varepsilon_n \,|\, X^{(n)}] \to 0$ as $n \to \infty$ where M possibly depends on n.

Theorem 4.4 shows that the concentration rate of the pseudo-posterior nearly matches the concentration rate of the original posterior ε_n (this is implied by condition (3.2), (3.4) and a variant of (4.4) according to Theorem 4 of [16]) up to an inflation factor $\widetilde{C}_n$ which depends on the rate of $u_\theta(X^{(n)})/n$. If $\widetilde{C}_n = \mathcal{O}(1)$ in (4.3), the rate of the actual posterior and pseudo-posterior will be the same.

REMARK 3. (Connection to Empirical Bayes) Since $\Pi_n^\star(\cdot \,|\, X^{(n)})$ can be regarded as an EB posterior, we could alternatively apply techniques of [10] and [37] to quantify the convergence rate in Theorem 4.4. In particular, we can replace Assumption (4.3) with the following condition to lower-bound the denominator in (4.2)

$$\sup_{\theta \in B_n(\theta_0, \varepsilon_n)} P_{\theta_0}^{(n)} \left[\ln(p_\theta^{(n)}/p_{\theta_0}^{(n)}) + u_\theta < -n\varepsilon_n^2 \right] = o(n\varepsilon_n^2).$$

Instead of relying on the existence of exponential tests (through Lemma 9 in [16]), we can directly assume that for any $\epsilon > \varepsilon_n$ and for all $\theta \in \Theta_n$ such that $d(\theta, \theta_0) > j\epsilon$ for any $j \in \mathbb{N}$ there exists a test $\phi_n(\theta)$ satisfying

$$P_{\theta_0}^{(n)} \phi_n \lesssim \mathrm{e}^{-n\epsilon^2/2} \quad \text{and} \quad \int_{\mathcal{X}} (1 - \phi_n) p_\theta^{(n)} \mathrm{e}^{u_\theta} \leq \mathrm{e}^{-j^2 n\epsilon^2/2}.$$

REMARK 4. (Random Generator) Recall that the stationary distribution $\widetilde{\pi}_n^\star(\theta \,|\, X^{(n)})$ of the random generator MHC version can be written as (4.2) where u_θ is replaced with $\widetilde{u}_\theta$ from (3.8). Theorem 4.4 holds also for the random generator where $\widetilde{C}_n$ is obtained from (4.3) with $\widetilde{u}_\theta$ instead of u_θ. Due to the averaging aspect, we might expect this $\widetilde{C}_n$ to be smaller in the random generator design.

Theorem 4.4 describes the behavior of the pseudo-posterior around the truth θ_0. We learned that the rate is artificially inflated due a bias inflicted by the likelihood approximation, where $\Pi_n^\star(\cdot \,|\, X^{(n)})$ *may not* shrink around θ_0 when ε_n is faster than the rate δ_n established in Theorem 4.2. This suggest that the truth may not be the most natural centering point for the posterior to concentrate around. A perhaps more transparent approach is to consider a different (data-dependent) centering which will allow for a more honest reflection of the contraction speed devoid of any implicit bias. We look into model misspecification for guidance about reasonable centering points.

4.4. *Posterior Convergence Rates: Connection to Misspecification.* In Section 4.3, we reframed the stationary distribution (3.4) as an empirical Bayes posterior by absorbing the term $\mathrm{e}^{u_\theta(X^{(n)})}$ inside the prior. This section pursues a different approach, absorbing $\mathrm{e}^{u_\theta(X^{(n)})}$ inside the likelihood instead. This leads a mis-specified model $\widetilde{P}_\theta^{(n)}$ prescribed by the following likelihood function

$$\widetilde{p}_\theta^{(n)}(X^{(n)}) = \frac{p_\theta^{(n)}(X^{(n)}) \mathrm{e}^{u_\theta(X^{(n)})}}{C_\theta} \quad \text{where} \quad C_\theta = \int_{\mathcal{X}} p_\theta^{(n)}(X^{(n)}) \mathrm{e}^{u_\theta(X^{(n)})} \mathrm{d}X^{(n)}. \tag{4.5}$$

Defining $\widetilde{\pi}(\theta) \propto \pi(\theta) C_\theta$, we can rewrite (3.4) as a posterior density under a mis-specified likelihood and the modified prior $\widetilde{\pi}(\theta)$ as

$$\pi_n^\star(\theta \,|\, X^{(n)}) = \frac{\widetilde{p}_\theta^{(n)}(X^{(n)}) \widetilde{\pi}(\theta)}{\int_\Theta \widetilde{p}_\theta^{(n)}(X^{(n)}) \widetilde{\pi}(\theta) \mathrm{d}\theta}. \tag{4.6}$$

Since the model $\widetilde{p}_\theta^{(n)}$ is mis-specified (i.e. $P_{\theta_0}^{(n)}$ is *not* of the same form as $\widetilde{\mathcal{P}}^{(n)} = \{\widetilde{P}_\theta^{(n)} : \theta \in \Theta\}$), the posterior will concentrate around the point θ^* defined as

$$\theta^* = \arg\min_{\theta\in\Theta} -P_{\theta_0}^{(n)} \log[\widetilde{p}_\theta^{(n)}/p_{\theta_0}^{(n)}] \tag{4.7}$$

which corresponds to the element $\widetilde{P}_{\theta^*}^{(n)} \in \widetilde{\mathcal{P}}^{(n)}$ that is closest to $P_{\theta_0}^{(n)}$ in the KL sense [24]. Unlike in the iid data case studied, e.g., in [24] and [7], our likelihood (4.5) is not an independent product due to the non-separability of the function $u_\theta(X^{(n)})$. The following Theorem 4.5 quantifies concentration in terms of a KL neighborhoods around $\widetilde{P}_{\theta^*}^{(n)}$ defined as

$$B(\epsilon, \widetilde{P}_{\theta^*}^{(n)}, P_{\theta_0}^{(n)}) = \left\{\widetilde{P}_\theta^{(n)} \in \widetilde{\mathcal{P}}^{(n)} : K(\theta^*, \theta_0) \le n\epsilon^2, V(\theta^*, \theta_0) \le n\epsilon^2\right\}, \tag{4.8}$$

where $K(\theta^*, \theta_0) \equiv P_{\theta_0}^{(n)} \log \frac{\widetilde{p}_{\theta^*}^{(n)}}{\widetilde{p}_\theta^{(n)}}$ and $V(\theta^*, \theta_0) = P_{\theta_0}^{(n)} \left|\log \frac{\widetilde{p}_{\theta^*}^{(n)}}{\widetilde{p}_\theta^{(n)}} - K(\theta^*, \theta_0)\right|^2$.

THEOREM 4.5. *Denote with $Q_\theta^{(n)}$ a measure defined through $\mathrm{d}Q_\theta^{(n)} = \frac{p_{\theta_0}^{(n)}}{\widetilde{p}_{\theta^*}^{(n)}} \mathrm{d}P_\theta^{(n)}$ and let $d(\cdot,\cdot)$ be a semi-metric on $\mathcal{P}^{(n)}$. Suppose that there exists a sequence $\varepsilon_n > 0$ satisfying $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$ such that for every $\epsilon > \varepsilon_n$ there exists a test ϕ_n (depending on ϵ) such that for every $J \in \mathbb{N}_0$*

$$P_{\theta_0}^{(n)}\phi_n \lesssim \mathrm{e}^{-n\epsilon^2/4} \quad \textit{and} \quad \sup_{\widetilde{P}_\theta^{(n)} : d(\widetilde{P}_\theta^{(n)}, \widetilde{P}_{\theta^*}^{(n)}) > J\epsilon} Q_\theta^{(n)}(1-\phi_n) \le \mathrm{e}^{-nJ^2\epsilon^2/4}. \tag{4.9}$$

Let $B(\epsilon, \widetilde{P}_{\theta^*}^{(n)}, P_{\theta_0}^{(n)})$ be as in (4.8) and let $\widetilde{\Pi}_n(\theta)$ be a prior distribution with a density $\widetilde{\pi}(\theta) \propto C_\theta \pi(\theta)$. Assume that there exists a constant $L > 0$ such that, for all n and $j \in \mathbb{N}$,

$$\frac{\widetilde{\Pi}_n\left(\theta \in \Theta : j\varepsilon_n < d(\widetilde{P}_\theta^{(n)}, \widetilde{P}_{\theta^*}^{(n)}) \le (j+1)\varepsilon_n\right)}{\widetilde{\Pi}_n\left(B(\epsilon, \widetilde{P}_{\theta^*}^{(n)}, P_{\theta_0}^{(n)})\right)} \le \mathrm{e}^{n\varepsilon_n^2 j^2/8}. \tag{4.10}$$

Then for every sufficiently large constant M, as $n \to \infty$,

$$P_{\theta_0}^{(n)} \Pi_n^\star\left(\widetilde{P}_\theta^{(n)} : d(\widetilde{P}_\theta^{(n)}, \widetilde{P}_{\theta^*}^{(n)}) \ge M\varepsilon_n \,|\, X^{(n)}\right) \to 0. \tag{4.11}$$

PROOF. Section 7.4

REMARK 5. For iid data, [24] introduce a condition involving entropy numbers under misspecification which implies the existence of exponential tests for a testing problem that involves non-probability measures. Since we have a non-iid situation, we assumed the existence of tests directly.

REMARK 6. (Friendlier Metrics) In parametric models indexed by θ in a metric space (Θ, d), it is more natural to characterize the posterior concentration in terms of $d(\cdot,\cdot)$ rather than the Kullback-Leibler divergence[6]. If there exists some metric $d(\cdot,\cdot)$ on Θ such that for some $\underline{C}, \bar{C} > 0$ and for all $\theta, \theta_1, \theta_2 \in \Theta$

$$-\underline{C}\, d^2(\theta, \theta^*) \le P_{\theta_0}^{(n)} \log \frac{\widetilde{p}_\theta^{(n)}}{\widetilde{p}_{\theta^*}^{(n)}} \le -\bar{C}\, d^2(\theta, \theta^*) \quad \text{and} \quad P_{\theta_0}^{(n)} \left(\log \frac{\widetilde{p}_{\theta_1}^{(n)}}{\widetilde{p}_{\theta_2}^{(n)}}\right)^2 \le d^2(\theta_1, \theta_2) \tag{4.12}$$

[6]Hellinger neighborhoods are less appropriate for misspecified models

then the ball $B(\epsilon, \widetilde{P}^{(n)}_{\theta^*}, P^{(n)}_{\theta_0})$ contains a ball of the form $\{\theta \in \Theta : d(\theta, \theta^*) \leq C_1\epsilon\}$ and one can replace (4.10) with a prior concentration condition $\widetilde{\Pi}_n(\theta : d(\theta, \theta^*) \leq C_1\varepsilon_n) \geq \mathrm{e}^{-Ln\varepsilon_n^2}$ for some suitable $C_1, L > 0$ and assume (4.9) in terms of d on Θ to obtain $P^{(n)}_{\theta_0}\Pi^\star_n\left(\theta : d(\theta, \theta^*) \geq M_n\varepsilon_n \,|\, X^{(n)}\right) \to 0$ for any sequence $M_n \to \infty$.

4.5. *Bernstein-von Mises Theorem.* The Bernstein-von Mises (BvM) theorem asserts that the posterior distribution of a parameter in a suitably regular finite-dimensional model is approximately normally distributed as the number of observations grows to infinity. More precisely, if $\theta \to p_\theta$ is appropriately smooth and identifiable and the prior $\Pi_n(\cdot)$ puts positive mass around the true parameter θ_0, then the posterior distribution of $\sqrt{n}(\theta - \widehat{\theta}_n)$ tends to $N(0, I^{-1}_{\theta_0})$ for most observations $X^{(n)}$, where $\widehat{\theta}_n$ is an efficient estimator and I_θ is the Fisher information matrix of the model at θ. In this section, we want to understand the effect of the tilting factor $\mathrm{e}^{u_\theta(X^{(n)})}$ on the limiting shape of the pseudo-posterior in (3.4) that is proportional to $\pi_n(\theta \,|\, X^{(n)})\mathrm{e}^{u_\theta(X^{(n)})}$. Exponential tilting is particularly intuitive for linear $u_\theta(X^{(n)})$ and for Gaussian posteriors.

Example 2. (Linear u_θ) Suppose that the posterior $\pi_n(\theta \,|\, X^{(n)})$ is Gaussian with some mean μ and covariance Σ. This holds approximately in regular models according to the BvM theorem (Theorem 10.1 in [41]). Assume that there exists an invertible mapping $\tau : \Theta \to \Theta$ such that $\theta = \tau(\bar{\theta})$ where the density for $\bar{\theta}$ satisfies

$$\pi_n(\theta \,|\, X^{(n)})\mathrm{e}^{u_\theta(X^{(n)})}\mathrm{d}\theta \propto \pi^*_n(\bar{\theta} \,|\, X^{(n)})\mathrm{d}\bar{\theta}.$$

Assuming the following linear form (justified in Remark 1)

$$u_\theta(X^{(n)}) = a^*(X^{(n)}) + \theta' u^*(X^{(n)}) \tag{4.13}$$

we obtain $\bar{\theta} \sim \mathcal{N}(\mu + \Sigma\, u^*(X^{(n)}), \Sigma)$. In this case, the mapping τ satisfies $\theta = \tau(\bar{\theta}) = \bar{\theta} - \Sigma u^*(X^{(n)})$, implying a location shift. We had concluded a similar property below Theorem 4.3 at the end of Section 4.1.

Example 2 reveals how the behavior of $u^*(X^{(n)})$ affects the centering of the posterior limit (under linearity and Gaussianity) and how it may prevent BvM from occurring when $\widehat{\theta}_n + \frac{1}{n}I^{-1}_{\theta_0}u^*(X^{(n)})$ is *not* an asymptotically efficient estimator. We now turn to more precise statements by recollecting the BvM phenomenon under misspecification in LAN models [25]. The centering and the asymptotic covariance matrix will be ultimately affected by θ^* in (4.7).

Lemma 4.6. *(Bernstein von-Mises) Assume that the posterior (4.6) concentrates around θ^* at the rate ε^*_n and that for every compact $K \subset \mathbb{R}^d$*

$$\sup_{h\in K}\left|\log\frac{\widetilde{p}^{(n)}_{\theta^*+\varepsilon^*_n h}(X^{(n)})}{\widetilde{p}^{(n)}_{\theta^*}(X^{(n)})} - h'\widetilde{V}_{\theta^*}\widetilde{\Delta}_{n,\theta^*} - \frac{1}{2}h'\widetilde{V}_{\theta^*}h\right| \to 0 \quad \textit{in } P^{(n)}_{\theta_0}\textit{-probability} \tag{4.14}$$

for some random vector $\widetilde{\Delta}_{n,\theta^*}$ and a non-singular matrix $\widetilde{V}_{\theta^*}$. Then the pseudo-posterior converges to a sequence of normal distributions in total variation at the rate ε^*_n, i.e.

$$\sup_B\left|\Pi^*_n\left(\varepsilon^{*-1}_n(\theta - \theta^*) \in B \,|\, X^{(n)}\right) - N_{\widetilde{\Delta}_{n,\theta^*},\widetilde{V}_{\theta^*}}(B)\right| \to 0 \quad \textit{in } P^{(n)}_{\theta_0}\textit{-probability.}$$

PROOF. Follows from Theorem 2.1 of [25].

It remains to examine the assumption (4.14). For `iid` data, [25] derived sufficient conditions (Lemma 2.1) for (4.14) to hold. Due to the non-separability of the term $u_\theta(X^{(n)})$, the mis-specified model cannot be regarded as arriving from an `iid` experiment. Below, we nevertheless provide some intuition for when (4.14) is expected to hold in our setup when $u_\theta(X^{(n)})$ is linear, relaxing slightly Lemma 2.1 in [25]. Recall that in Remark 1 we have concluded that under differentiability assumptions, the posterior residual $u_\theta(X^{(n)})$ does converge to a linear function in θ.

LEMMA 4.7. *Assume that $P_{\theta_0}^{(n)} = P_{\theta_0}^n$ with a density $\prod_{i=1}^n p_{\theta_0}(x_i)$ where the function $\theta \to \log p_\theta(x)$ is differentiable at θ^* with a derivative $\dot{\ell}_\theta$. Assume there exists an open neighborhood U of θ^* such that $\left|\log \frac{p_{\theta_1}(x)}{p_{\theta_2}(x)}\right| \le m_{\theta^*}\|\theta_1 - \theta_2\|$ P_{θ_0} $-a.s.\forall \theta_1, \theta_2 \in U$ where m_θ is a square integrable function. Assume that the log-likelihood has a 2nd order Taylor expansion around θ^* (i.e. (7.5) holds). Assume that u_θ is asymptotically linear around θ^* (i.e. (7.6) holds), then (4.14) holds with $\varepsilon_n^* = 1/\sqrt{n}$ and*

$$\widetilde{V}_\theta = V_\theta \quad \text{and} \quad \widetilde{\Delta}_{n,\theta} = V_\theta^{-1}\left[\frac{\dot{C}_\theta}{\sqrt{n}} + \sqrt{n}\mathbb{P}_n\dot{\ell}_\theta + \frac{u^*(X^{(n)})}{\sqrt{n}}\right] \tag{4.15}$$

PROOF. Section 7.5.

The assumptions in Lemma 4.7 are closely related to the ones in Theorem 4.3. The main difference is that Lemma 4.7 is concerned with the behavior of the (misspecified) likelihood around θ^* as opposed to θ^0. While asymptotic normality still occurs in misspecified models, the implied Bayesian credible sets do not automatically yield valid confidence set, as would be the case in correctly specified case [25].

REMARK 7. (Alternative BvM Conditions) [6] characterized related BvM conditions. We restate these conditions utilizing the localized re-parametrization $h = \sqrt{n}(\theta - \theta_0) - s$, where $s = \sqrt{n}(\hat{\theta} - \theta_0)$ is a *zero-mean* vector where $\hat{\theta}$ is some suitable estimator. We first define a localized criterion function

$$\ell(h) \equiv \frac{\widetilde{p}_{\hat{\theta}+h/\sqrt{n}}(X^{(n)})\widetilde{\pi}(\hat{\theta} + h/\sqrt{n})}{\widetilde{p}_{\hat{\theta}}(X^{(n)})\widetilde{\pi}(\hat{\theta})},$$

which corresponds to the normalized pseudo-posterior $\pi^*(\theta \,|\, X^{(n)})/\pi^*(\hat{\theta} \,|\, X^{(n)})$. [6] impose a centered variant of (4.14) requiring that $\ell(h)$ approaches a quadratic form on a closed ball K (such that[7] $\Lambda \equiv \sqrt{n}(\Theta - \theta_0) - s = K \cup K^c$) in the sense that

$$|\log \ell(h) - (-h'Jh)/2| \le \epsilon_1 + \epsilon_2 \times h'Jh/2 \quad \forall h \in K, \tag{4.16}$$

for some matrix $J > 0$ with eigenvalues bounded away from zero. If

$$\epsilon_1 = o(1) \quad \text{and} \quad \epsilon_2 \times \lambda_{max}^2(J)(\sup_{h\in K}\|h\|)^2 = o(1) \quad \text{in } P_{\theta_0}^{(n)}\text{-probability.} \tag{4.17}$$

Theorem 1 of [6] shows that $\ell(h)/\int_\Lambda \ell(h)\mathrm{d}h$ approaches the standard normal density in $P_{\theta_0}^{(n)}$-probability as $n, d \to \infty$. The condition (4.16) (a) allows for mild deviations from smoothness and log-concavity, (b) involves also the prior (unlike (4.14)) but, (c) requires

[7] $\int_K \ell(h)\mathrm{d}h/\int_\Lambda \ell(h)\mathrm{d}h \ge 1 - o_{P_{\theta_0}}(1)$ and $\int_K \phi(h)\mathrm{d}h$ for $\phi(\cdot)$ standard Gaussian density

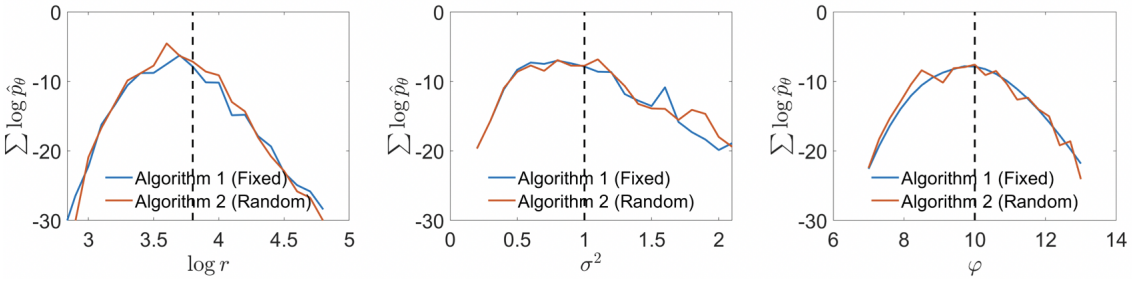


Fig 4: Estimated log likelihood ratio for the Ricker model: (Left) function of $\log r$ fixing $\sigma = \sigma_0$ and $\varphi = \varphi_0$, (Middle) function of σ^2 fixing $r = r_0$ and $\varphi = \varphi_0$, (Right) function of φ fixing $\sigma = \sigma_0$ and $r = r_0$.

the existence of a $\sqrt{n}-$consistent estimator $\hat{\theta}$. Lemma 4.6 is more general, where the rate ε_n^* does not need to be $1/\sqrt{n}$ and where the posterior is allowed to have a non-vanishing bias. The requirement (4.17) imposes certain restrictions on $u_\theta(X^{(n)})$. For example, in the linear case (4.13) one would need $u^\star(X^{(n)}) = o(\sqrt{n})$ in $P_{\theta_0}^{(n)}$-probability from (4.17).

5. Applications. We demonstrate the usefulness of MHC on two examples. The first demonstration is the Ricker model (analyzed previously, e.g., in [20, 43, 11]) where we compare MHC to the pseudomarginal method. The second one is a Bayesian model choice example where we compare MHC to ABC.

5.1. *The Ricker model.* The Ricker model is a classic discrete model that describes partially observed population dynamics of fish and animals in ecology. The latent population $N_{i,t}$ follows

$$\log N_{i,t+1} = \log r + \log N_{i,t} - N_{i,t} + \sigma \varepsilon_{i,t}, \qquad \varepsilon_{i,t} \sim N(0,1),$$

where r denotes the intrinsic growth rate and σ is the dispersion of innovations. The index t represents time and runs through 1 to $T = 20$. The index i represents independent observations and runs through 1 to $n = 300$. The initial population $N_{i,0}$ may be set as 1 or set randomly after some burn-in period. We observe $X_{i,t}$ such that

$$X_{i,t} \mid N_{i,t} \sim \text{Poisson}(\varphi N_{i,t}),$$

where φ is a scale parameter. The objective is to make inference on $\theta := (\log r, \sigma^2, \varphi)$. Each time sequence $X_i := (X_{i,1}, \ldots, X_{i,T})$ constitutes an observation, where i runs through n. In our notation, we can define the underlying data-generating process as $\widetilde{X}_{i,t} := (U_{i,t}, \varepsilon_{i,t})$ for $U_{i,t} \sim U[0,1]$ and set the function T_θ to map ε_i to N_i and then (U_i, N_i) to X_i through the Poisson inverse transform sampling of $U_{i,t}$ into $X_{i,t}$. We set the true parameter as $(\log r_0, \sigma_0^2, \varphi_0) = (3.8, 1, 10)$ and employ an improper, flat prior. Note that our method can accommodate an improper prior, unlike ABC.

There is no obvious sufficient statistic for this model, and the likelihood is intractable due to the nontrivial time dependence of $N_{i,t}$. We use an average of neural network discriminators to adapt to the unknown likelihood ratio. First, we estimate D_θ by a neural network with one hidden layer with 50 nodes, each of which is equipped with the hyperbolic tangent sigmoid activation function. Then, we compute the log likelihood of the data $\sum_i \log \frac{1-\hat{D}_\theta}{\hat{D}_\theta}$. We repeat this for 20 times with independently drawn $\widetilde{X}$ and take

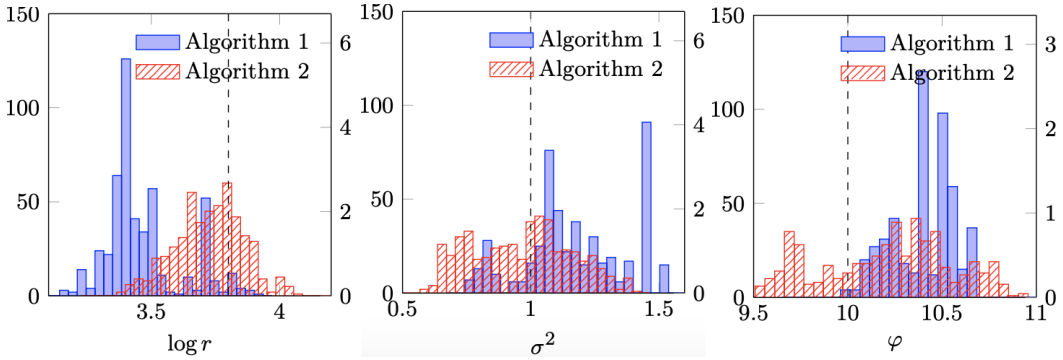


Fig 5: MHC samples for the Ricker model.

the average of the log likelihood. This specification produces approximately quadratic likelihood-ratio curves (Figure 4). Unlike the location-scale normal model, the fixed design does not produce entirely smooth curves due to the averaging aspect over many discriminators. The quadratic shape is nevertheless recovered here, implying that the differentiability assumptions from Section 4.1 are not entirely objectionable.

Figure 5 shows the marginal histograms of the MHC samples (500 MCMC iterations). The proposal distribution is independent across parameters; $\log r$ uses the normal distribution, σ^2 the inverse-gamma distribution, and φ the gamma distribution; each of them has the mean equal to the previous draw and variance $1/n$. The vertical dashed lines indicate the true parameter θ_0. Note that the posterior is asymptotically centered at the MLE, not θ_0. However, the blue histograms on the left (Algorithm 1) seem too far away from θ_0 relative to the widths of the histograms. On the other hand, the red histograms (Algorithm 2) are more dispersed but located closer to θ_0. These observations confirm our theoretical findings. Histograms of Algorithm 3 (Figure 6) look reasonable as a posterior sample, center around the true values.

Figure 6 compares our method with the MCWM pseudo-marginal Metropolis-Hastings algorithm [1]. The pseudo-marginal method deploys an average of conditional likelihoods for X_i, given N_i,

$$\hat{p}(X_i) = \frac{1}{K}\sum_{k=1}^{K}\prod_{t=1}^{T} p(X_{i,t} \mid N_{i,t,k}) = \frac{1}{K}\sum_{k=1}^{K}\prod_{t=1}^{T} \frac{(\varphi N_{i,t,k})^{X_{i,t}} e^{-\varphi N_{i,t,k}}}{X_{i,t}!}$$

as the likelihood approximation, where K is some positive integer and where $N_{i,t,k}$ are independently drawn across $k = 1, \ldots, K$. In our comparisons, we let $K = 20n$. Figure 6 shows that the two methods produce posterior draws that are located at similar places, and the widths of the histograms are also comparable. We would like to point out, again, that our method does not require that a tractable conditional likelihood is available nor that a user-specified summary statistic is supplied.

5.2. *Bayesian Model Selection.* The performance of summary statistic-based methods is sensitive to the quality of the summary statistic, which can sometimes be very subtle. [36] show that when model selection is concerned, ABC may fail even when the summary statistic is sufficient for *each* of the models considered. Our method *does not* require a summary statistic but a sieve of discriminators that can adapt to the oracle discriminator in the limit. This creates hope that our method can tackle model selection problems.

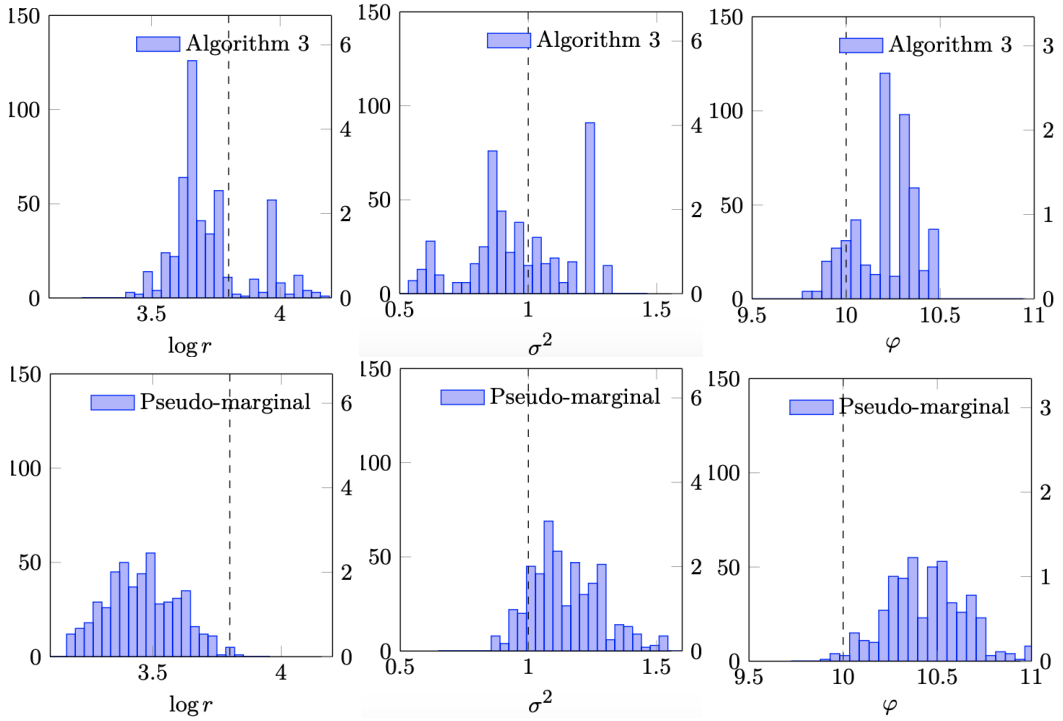


Fig 6: (Top) MHC samples for the Ricker model (Algorithm 3), (Bottom) Pseudo-marginal method

To illustrate this point, we consider the following model choice problem. The actual data follows $X_i \sim N(0,1)$ for $i = 1, \ldots, n = 500$. We have two candidate models $P_{1,\mu} = N(\mu, 1)$ and $P_{2,\mu} = N(\mu, 1+3/\sqrt{n})$ to choose from. We let the parameters be $\theta := (m, \mu)$, where $m \in \{1, 2\}$ is the model indicator and μ is unknown mean with a prior $N(0,1)$. The model is assigned a uniform prior, i.e. $P(m=1) = P(m=2) = 0.5$. Following the traditional Bayesian model selection formalism, we collect evidence for model $m = 1$ with a Bayes factor

$$B_{12} := \frac{\pi_n(m=1 \mid X)}{\pi_n(m=2 \mid X)}.$$

The Bayes factor is the ratio of the marginal likelihoods (or posterior probabilities) of $m = 1$ over $m = 2$. The actual Bayes factor value is $B_{12} = 9$, indicating strong evidence in favor of $m = 1$. The Bayes factor will be estimated by the ratio of the frequencies of the posterior samples given by ABC or our method. Since our parameter of interest m is discrete, there is no de-biasing for this example.

[36] in their Lemma 2 show that when the summary statistic is $\sum_i X_i$, the Bayes factor estimated by ABC asymptotes to 1. This is equivalent to choosing the model with a coin toss. For our method, we use the logistic regression on regressors $(1, X_i, X_i^2)$, which can mimic the oracle discriminator.

The trace plots of sampled models for exact MH, MHC and ABC are provided in Figure 7. Table 1 summarizes the posterior model frequencies. The true posterior probabilities are $\pi_n(m=1 \mid X) \approx 0.9$ and $\pi_n(m=2 \mid X) \approx 0.1$, so the Bayes factor is 9. The "Oracle MH" is the Metropolis-Hastings with the true likelihood, in which 84.4% of the posterior draws choose model 1. Algorithms 1 and 2 choose model 1 respectively 93.2% and 70% of the times. ABC based on the sum, on the other hand, chooses

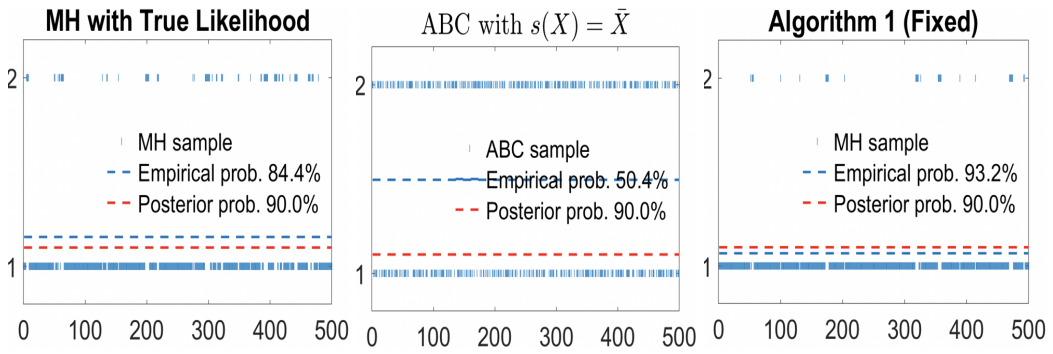
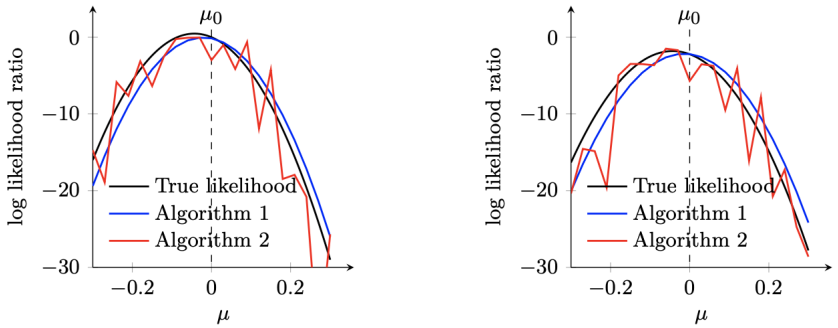


Fig 7: Trace plots of sampled models using: (Left) MH with the true likelihood ratio, (Middle) ABC with $s(X^{(n)}) = \bar{X}_n$ and (Right) fixed generator MHC.



(a) Estimated log likelihood for $m = 1$. The vertical dashed line indicates $\mu_0 = 0$.

(b) Estimated log likelihood for $m = 2$. The vertical dashed line indicates $\mu_0 = 0$.

Fig 8: Estimated log likelihood for models 1 and 2. The figures indicate that it is smooth in μ and have the same curvature as the true log likelihood.

	Posterior	Oracle MH	Algorithm 1	Algorithm 2	ABC
Model 1	90%	422	466	350	252
Model 2	10%	78	34	150	248
Bayes factor	9.00	5.41	13.71	2.33	1.02

TABLE 1

"Posterior" column gives the posterior probability of each model, $\pi_n(m = j \mid X)$. Other columns give the frequencies of the corresponding sample of size 500. "Oracle MH" refers to the Metropolis-Hastings algorithm with the true likelihood. "ABC" is based on the summary statistics $s(X) = \bar{X}_n$.

the model randomly. Finally, Figure 8 gives the estimated log-likelihood ratio for each model. In terms of μ, we again see that Algorithm 1 is slightly biased with the correct shape and Algorithm 2 is less biased but more dispersed on average.

6. Discussion. This paper develops an approximate Metropolis-Hastings (MH) posterior sampling method for when the likelihood is not tractable. By deploying a Generator and a Classifier (similarly as in Generative Adversarial Networks [18]), likelihood ratio estimators are obtained which are then plugged into the MH sampling routine. We consider two variants: (1) a fixed generator design yielding biased samples,

and (2) a random generator yielding more dispersed samples. We provide a thorough frequentist characterization of the stationary distribution including convergence rates and asymptotic normality. Under suitable differentiability assumptions, we conclude that correct shape and location can be recovered by deploying a debiasing combination of the fixed and random generator variants.

7. Proofs.

7.1. *Proof of Theorem 4.2.* The following lemma bounds the Kullback-Leibler divergence and variation by possibly non-diverging multiples of the Hellinger distance.[8] This can be used to derive sharper rates of posterior contraction in models with unbounded likelihood ratios [see 17, p. 199 and Appendix B].

LEMMA 7.1. *For probability measures P and P_0 such that $P_0(p_0/p) < \infty$, let $M := \inf_{c\geq 1} cP_0(\frac{p_0}{p} \mid \frac{p_0}{p} \geq [1+\frac{1}{2c}]^2)$ where $P_0(\cdot \mid A) = 0$ if $P_0(A) = 0$. For $k \geq 2$, the following hold.*

(i) $-P_0 \log \frac{p}{p_0} \leq (3+M)h(p,p_0)^2$.
(ii) $P_0|\log \frac{p}{p_0}|^k \leq 2^{k-1}\Gamma(k+1)(2+M)h(p,p_0)^2$.
(iii) $P_0|\log \frac{p}{p_0} - P_0 \log \frac{p}{p_0}|^k \leq 2^{2k-1}\Gamma(k+1)(2+M)h(p,p_0)^2$.
(iv) $\|\frac{1}{2}\log \frac{p}{p_0}\|^2_{P_0,B} \leq (2+M)h(p,p_0)^2$.
(v) $\|\frac{1}{4}(\log \frac{p}{p_0} - P_0 \log \frac{p}{p_0})\|^2_{P_0,B} \leq (2+M)h(p,p_0)^2$.

Here, $\|f\|_{P,B} := \sqrt{2P(e^{|f|} - 1 - |f|)}$ is the Bernstein "norm".

PROOF. (iv) Using $e^{|x|} - 1 - |x| \leq (e^x - 1)^2$ for $x \geq -\frac{1}{2}$ and $e^{|x|} - 1 - |x| < e^x - \frac{3}{2}$ for $x > \frac{1}{2}$,

$$\Big\|\log\sqrt{\tfrac{p}{p_0}}\Big\|^2_{P_0,B} \leq 2P_0\Big(\sqrt{\tfrac{p}{p_0}} - 1\Big)^2 \mathbb{1}\{\tfrac{p}{p_0} \geq \tfrac{1}{e}\} + 2P_0\Big(\sqrt{\tfrac{p_0}{p}} - \tfrac{3}{2}\Big)\mathbb{1}\{\tfrac{p_0}{p} > e\}.$$

The first term is bounded by $2h(p,p_0)^2$. For every $c \geq 1$,

$$\begin{aligned} P_0\Big(\sqrt{\tfrac{p_0}{p}} - \tfrac{3}{2}\Big)\mathbb{1}\{\tfrac{p_0}{p} > e\} &\leq P_0\Big(\sqrt{\tfrac{p_0}{p}} - 1 - \tfrac{1}{2c}\Big)\mathbb{1}\Big\{\sqrt{\tfrac{p_0}{p}} \geq 1 + \tfrac{1}{2c}\Big\} \\ &= P_0\Big(\sqrt{\tfrac{p_0}{p}} \geq 1 + \tfrac{1}{2c}\Big)\Big[P_0\Big(\sqrt{\tfrac{p_0}{p}} - 1 \,\Big|\, \sqrt{\tfrac{p_0}{p}} \geq 1 + \tfrac{1}{2c}\Big) - \tfrac{1}{2c}\Big]. \end{aligned}$$

Since $x - \frac{1}{2c} \leq \frac{c}{2}x^2$ for every x,

$$\begin{aligned} P_0\Big(\sqrt{\tfrac{p_0}{p}} - 1 \,\Big|\, \sqrt{\tfrac{p_0}{p}} \geq 1 + \tfrac{1}{2c}\Big) - \tfrac{1}{2c} &\leq \tfrac{c}{2}\Big[P_0\Big(\sqrt{\tfrac{p_0}{p}} - 1 \,\Big|\, \sqrt{\tfrac{p_0}{p}} \geq 1 + \tfrac{1}{2c}\Big)\Big]^2 \\ &\leq \tfrac{c}{2}P_0\Big(\tfrac{p_0}{p} \,\Big|\, \sqrt{\tfrac{p_0}{p}} \geq 1 + \tfrac{1}{2c}\Big)P_0\Big(\Big[1 - \sqrt{\tfrac{p}{p_0}}\Big]^2 \,\Big|\, \sqrt{\tfrac{p_0}{p}} \geq 1 + \tfrac{1}{2c}\Big) \end{aligned}$$

by the Cauchy-Schwarz inequality. Then the result follows.

(i) Write $-P_0 \log \frac{p}{p_0} = P_0(\frac{p}{p_0} - 1 - \log \frac{p}{p_0}) + P(p_0 = 0)$. With $\frac{1}{x} - 1 - \log \frac{1}{x} < 2(\sqrt{x} - \frac{3}{2})$ for every $x \geq 3$,

$$P_0(\tfrac{p}{p_0} - 1 - \log \tfrac{p}{p_0}) \leq P_0(\tfrac{p}{p_0} - 1 - \log \tfrac{p}{p_0})\mathbb{1}\{\tfrac{p}{p_0} > \tfrac{1}{3}\} + 2P_0\Big(\sqrt{\tfrac{p_0}{p}} - \tfrac{3}{2}\Big)\mathbb{1}\{\tfrac{p_0}{p} \geq 3\}.$$

[8] Lemma 7.1 (iv) first appeared in Kaji, Manresa and Pouliot [23, Lemma 5]. We reproduce the proof here as it is used to prove other statements.

The second term is bounded as above. For the first term, observe that

$$\frac{P_0(\frac{p}{p_0}-1-\log\frac{p}{p_0})\mathbb{1}\{\frac{p}{p_0}>\frac{1}{3}\}}{P_0(1-\sqrt{p/p_0})^2\mathbb{1}\{\frac{p}{p_0}>\frac{1}{3}\}}\le\sup_{p/p_0>1/3}\frac{\frac{p}{p_0}-1-\log\frac{p}{p_0}}{(1-\sqrt{p/p_0})^2}<3.$$

With $P(p_0=0)=\int(\sqrt{p}-\sqrt{p_0})^2\mathbb{1}\{p_0=0\}$ follows the result.

(ii) Since $e^x-1-x\ge x^k/\Gamma(k+1)$ for $k\ge 2$ and $x\ge 0$,[9] $P_0|\log\frac{p}{p_0}|^k\le 2^{k-1}\Gamma(k+1)\|\frac{1}{2}\log\frac{p}{p_0}\|^2_{P_0,B}$. Then, apply (iv).

(iii) By the triangle and Hölder's inequalities, for $k\ge 1$, $P_0|\log\frac{p}{p_0}-P_0\log\frac{p}{p_0}|^k\le[(P_0|\log\frac{p}{p_0}|^k)^{1/k}+P_0\log\frac{p}{p_0}]^k\le 2^kP_0|\log\frac{p}{p_0}|^k$. Then, use (ii).

(v) By the convexity of $e^{|x|}-1-|x|$ and Jensen's inequality, $\|\frac{1}{4}(\log\frac{p}{p_0}-P_0\log\frac{p}{p_0})\|^2_{P_0,B}\le\frac{1}{2}\|\frac{1}{2}\log\frac{p}{p_0}\|^2_{P_0,B}+\frac{1}{2}\|P_0\frac{1}{2}\log\frac{p}{p_0}\|^2_{P_0,B}\le\|\frac{1}{2}\log\frac{p}{p_0}\|^2_{P_0,B}$. With (iv) follows the result. □

Proof of Theorem 4.2. For $D\in\mathcal{D}$, write $\mathbb{P}_n(\log\frac{1-D}{1-D_\theta}-\log\frac{D}{D_\theta})$ as

$$P_0\log\tfrac{1-D}{1-D_\theta}-P_0\log\tfrac{D}{D_\theta}+(\mathbb{P}_n-P_0)\log\tfrac{1-D}{1-D_\theta}-(\mathbb{P}_n-P_0)\log\tfrac{D}{D_\theta}.$$

Let $W_1:=\sqrt{\frac{D}{D_\theta}}-1$, $W_2:=\sqrt{\frac{1-D}{1-D_\theta}}-1$, and $\delta:=d_\theta(D,D_\theta)$. By Taylor's theorem, $\log(1+x)=x-\frac{1}{2}x^2+\frac{1}{2}x^2R(x)$ where $R(x)=O(x)$ as $x\to 0$. Therefore, $P_0\log\frac{D}{D_\theta}=2P_0W_1-P_0W_1^2+P_0W_1^2R(W_1)$. Note that $P_0W_1^2=P_0(\sqrt{D/D_\theta}-1)^2=h_\theta(D,D_\theta)^2$. Since $W_1^2\ge 0$, this implies that $W_1(X_i)^2=O_P(\delta^2)$ and $W_1(X_i)=o_P(1)$. Also,

$$\begin{aligned}2P_0W_1&=\Big[2P_0\tfrac{\sqrt{D(p_0+p_\theta)}}{\sqrt{p_0}}-\int D(p_0+p_\theta)-\int p_0\Big]+(P_0+P_\theta)(D-D_\theta)\\&=-h_\theta(D,D_\theta)^2+(P_0+P_\theta)(D-D_\theta).\end{aligned}$$

Note that $(P_0+P_\theta)|D-D_\theta|\le(P_0+P_\theta)(\sqrt{D}+\sqrt{D_\theta})|\sqrt{D}-\sqrt{D_\theta}|\le 2\sqrt{2}h_\theta(D,D_\theta)$ by the Cauchy-Schwarz inequality. Next, for $1/5\le c<1$,

$$\begin{aligned}|P_0W_1^2R(W_1)|&\le P_0W_1^2|R(W_1)|\mathbb{1}\{W_1\le -c\}+P_0W_1^2|R(W_1)|\mathbb{1}\{W_1>-c\}\\&\le P_0(-R(W_1)\mathbb{1}\{W_1\le -c\})+P_0W_1^2|R(-c)\vee R(W_1)|.\end{aligned}$$

Since $R(x)<1$ and $R(W_1)=o_P(1)$, the second term is $o(\delta^2)$ for every c by the dominated convergence theorem. By the diagonal argument, there exists a sequence $c\to 1$ for given $D\to D_\theta$ such that the second term remains $o(\delta^2)$. Since $0<-R(x)<-2\log(1+x)$ for $x\le-\frac{1}{5}$,

$$\begin{aligned}&P_0(-R(W_1)\mathbb{1}\{W_1\le -c\})\le P_0(\log\tfrac{D_\theta}{D}\mathbb{1}\{W_1\le -c\})\\&\quad=P_0(\tfrac{D}{D_\theta}\log\tfrac{D_\theta}{D}\cdot\tfrac{D_\theta}{D}\mathbb{1}\{W_1\le -c\})\le\sup_{x\ge(1-c)^{-2}}|\tfrac{1}{x}\log x|\cdot P_0(\tfrac{D_\theta}{D}\mathbb{1}\{W_1\le -c\}).\end{aligned}$$

The first term is $o(1)$ as $c\to 1$. The second term is bounded by $P_0(\frac{D_\theta}{D}\mathbb{1}\{W_1\le-\frac{1}{5}\})=P_0(W_1\le-\frac{1}{5})P_0(\frac{D_\theta}{D}\mid\frac{D_\theta}{D}\ge\frac{25}{16})\le P_0(W_1\le-\frac{1}{5})M$ by Assumption 2. By Markov's inequality, $P_0(W_1\le-\frac{1}{5})\le 25P_0W_1^2=O(\delta^2)$. Thus, $|P_0W_1^2R(W_1)|=o(\delta^2)$. Altogether, we have $P_0\log\frac{D}{D_\theta}=O(\delta)$.

[9] $\Gamma(k-1)\ge\int_x^\infty y^{k-2}e^{-y}dy\ge x^{k-2}e^{-x}$ implies $\frac{d^2}{dx^2}(e^x-1-x)\ge\frac{d^2}{dx^2}x^k/\Gamma(k+1)$.

Next, write $P_0 \log \frac{1-D}{1-D_\theta} = 2P_0 W_2 - P_0 W_2^2 + P_0 W_2^2 R(W_2)$. By the Cauchy-Schwarz inequality,

$$P_0 W_2 \le \sqrt{P_0 \tfrac{p_0}{p_\theta}} \cdot h_\theta(1-D, 1-D_\theta) \le \sqrt{M}\delta,$$

$$P_0 W_2^2 \le \sqrt{(P_0 + P_\theta)\big(\tfrac{p_0}{p_\theta}\big)^2 (\sqrt{1-D} - \sqrt{1-D_\theta})^2} \cdot h_\theta(1-D, 1-D_\theta).$$

Since D and D_θ are bounded by 0 and 1,

$$(P_0 + P_\theta)\big(\tfrac{p_0}{p_\theta}\big)^2 (\sqrt{1-D} - \sqrt{1-D_\theta})^2 \le P_0 \big(\tfrac{p_0}{p_\theta}\big)^2 + P_0 \tfrac{p_0}{p_\theta} \le 2M.$$

Therefore, by the dominated convergence theorem, the LHS is $o(1)$ and hence $P_0 W_2^2 = o(\delta)$. This also implies $W_2(X_i) = o_P(1)$, $W_2^2(X_i) = o_P(\delta)$, and $R(W_2(X_i)) = o_P(1)$. Next, similarly as before, for $1/5 \le c < 1$,

$$|P_0 W_2^2 R(W_2)| \le P_0(-R(W_2)\mathbb{1}\{W_2 \le -c\}) + P_0 W_2^2 |R(-c) \vee R(W_2)|.$$

There exists a sequence $c \to 1$ such that the second term is $o(\delta)$. Also,

$$P_0(-R(W_2)\mathbb{1}\{W_2 \le -c\}) \le \sup_{x \ge (1-c)^{-2}} |\tfrac{1}{x} \log x| \cdot P_0(\tfrac{1-D_\theta}{1-D}\mathbb{1}\{W \le -c\}).$$

The first term is $o(1)$ as $c \to 1$. The second term is bounded by $P_0(W_2 \le -\frac{1}{5})M$ by Assumption 2. By Markov's inequality, $P_0(W_2 \le -\frac{1}{5}) \le 25 P_0 W_2^2 = O(\delta)$. Thus, $|P_0 W_2^2 R(W_2)| = o(\delta)$. Altogether, we have $P_0 \log \frac{1-D}{1-D_\theta} = O(\delta)$.

Next, we bound $\mathbb{E}^* \sup_{D \in \mathcal{D}^\theta_{n,\delta_n}} |\sqrt{n}(\mathbb{P}_n - P_0) \log \frac{D}{D_\theta}|$. Under Assumption 2, an analogous argument as Lemma 7.1 (iv) yields that $\|\frac{1}{2} \log \frac{D}{D_\theta}\|_{P_0,B}^2 \le 2(1+M) h_\theta(D, D_\theta)^2 = O(\delta^2)$. By van der Vaart and Wellner [42, Lemma 3.4.3], we have

$$\mathbb{E}^* \sup_{D \in \mathcal{D}^\theta_{n,\delta_n}} |\sqrt{n}(\mathbb{P}_n - P_0) \log \tfrac{D}{D_\theta}| \lesssim J\Big(1 + \tfrac{J}{\delta^2 \sqrt{n}}\Big).$$

for $J = J_{[]}(\delta, \{\log \frac{D}{D_\theta} : D \in \mathcal{D}^\theta_{n,\delta_n}\}, \|\cdot\|_{P_0,B})$. Note that a δ-bracket in $\mathcal{D}^\theta_{n,\delta_n}$ induces a $C\delta$-bracket in $\{\log \frac{D}{D_\theta}\}$ for some constant C since $\|\log \frac{u}{D_\theta} - \log \frac{\ell}{D_\theta}\|_{P_0,B}^2 \le 4P_0\big(\sqrt{u/\ell} - 1\big)^2 = O(d_\theta(u,\ell)^2)$ by Assumption 2. Therefore, $J \le J_{[]}(\delta, \mathcal{D}^\theta_{n,\delta_n}, d_\theta)$ and hence $J(1 + \frac{J}{\delta^2\sqrt{n}}) \lesssim \delta^2 \sqrt{n}$ by Assumption 1.

Finally, we bound $\mathbb{E}^* \sup_{D \in \mathcal{D}^\theta_{n,\delta_n}} |\sqrt{n}(\mathbb{P}_n - P_0) \log \frac{1-D}{1-D_\theta}|$. As in Lemma 7.1 (iv), we obtain $\rho^2 := \|\frac{1}{2} \log \frac{1-D}{1-D_\theta}\|_{P_0,B}^2 \le 2(1+M) P_0 W_2^2 = o(\delta)$. Therefore, by van der Vaart and Wellner [42, Lemma 3.4.3], $\mathbb{E}^* \sup_{D \in \mathcal{D}^\theta_{n,\delta_n}} |\sqrt{n}(\mathbb{P}_n - P_0) \log \frac{1-D}{1-D_\theta}| \lesssim J(1 + \frac{J}{\delta^2\sqrt{n}})$ for $J = J_{[]}(\rho, \{\log \frac{1-D}{1-D_\theta} : D \in \mathcal{D}^\theta_{n,\delta_n}\}, \|\cdot\|_{P_0,B})$. With a δ-bracket in $\mathcal{D}^\theta_{n,\delta_n}$, Assumption 2 implies $\|\log \frac{1-\ell}{1-D_\theta} - \log \frac{1-u}{1-D_\theta}\|_{P_0,B}^2 \le 4P_0\big(\sqrt{(1-\ell)/(1-u)} - 1\big)^2 = o(\delta)$. Therefore, the expectation of the supremum is of order $o(\delta\sqrt{n})$. □

7.2. *Proof of Theorem 4.3.* Let h_n be a bounded sequence and $\theta_n := \theta_0 + \frac{h_n}{\sqrt{n}}$ and $W_n := \sqrt{\hat{p}_{\theta_n}/\hat{p}_{\theta_0}} - 1$. Since $\log(1+x) = x - \frac{1}{2}x^2 + \frac{1}{2}x^2 R(x)$ for $R(x) = O(x)$, $n\mathbb{P}_n \log \frac{\hat{p}_{\theta_n}}{\hat{p}_{\theta_0}} = 2n\mathbb{P}_n W_n - n\mathbb{P}_n W_n^2 + n\mathbb{P}_n W_n^2 R(W_n)$. By Assumption 4 (ii) and $P_{\theta_0}\dot{\ell}_{\theta_0} = 0$, $2n\mathbb{P}_n W_n - n\mathbb{P}_n W_n^2 = 2nP_{\theta_0} W_n + \sqrt{n}\mathbb{P}_n h_n' \dot{\ell}_{\theta_0} - nP_{\theta_0} W_n^2 + o_P(1)$. Observe that $nP_{\theta_0} W_n^2 = \frac{1}{4} h_n' I_{\theta_0} h_n + o_P(1)$ and

$$2nP_{\theta_0} W_n = 2n\hat{P}_{\theta_0} W_n + 2n(P_{\theta_0} - \hat{P}_{\theta_0}) W_n$$

$$= -n\int\Big(\sqrt{\hat{p}_{\theta_n}} - \sqrt{\hat{p}_{\theta_0}}\Big)^2 + n(c_{\theta_n} - c_{\theta_0}) - \sqrt{n}\hat{P}_{\theta_0}h_n'\dot{\ell}_{\theta_0}$$
$$+ 2n\int\Big(\sqrt{p_{\theta_0}} - \sqrt{\hat{p}_{\theta_0}}\Big)\Big(\sqrt{p_{\theta_0}} + \sqrt{\hat{p}_{\theta_0}}\Big)\Big(W_n - \tfrac{h_n'\dot{\ell}_{\theta_0}}{2\sqrt{n}}\Big).$$

By Assumption 4 (i) and $\int(\sqrt{\hat{p}_{\theta_0}} - \sqrt{p_{\theta_0}})^2 = O_P(\delta_n^2)$, $n\int(\sqrt{\hat{p}_{\theta_n}} - \sqrt{\hat{p}_{\theta_0}})^2 = \frac{1}{4}h_n'I_{\theta_0}h_n + o_P(1)$. By the Cauchy-Schwarz inequality,

$$\left|\int\Big(\sqrt{p_{\theta_0}} - \sqrt{\hat{p}_{\theta_0}}\Big)\Big(\sqrt{p_{\theta_0}} + \sqrt{\hat{p}_{\theta_0}}\Big)\Big(W_n - \tfrac{h_n'\dot{\ell}_{\theta_0}}{2\sqrt{n}}\Big)\right|$$
$$\leq \left[\int\Big(\sqrt{p_{\theta_0}} - \sqrt{\hat{p}_{\theta_0}}\Big)^2\int\Big(\sqrt{p_{\theta_0}} + \sqrt{\hat{p}_{\theta_0}}\Big)^2\Big(W_n - \tfrac{h_n'\dot{\ell}_{\theta_0}}{2\sqrt{n}}\Big)^2\right]^{1/2},$$

which is $O_P(\delta_n n^{-3/4}) = o_P(n^{-1})$ under Assumption 4 (i) and $\delta_n = o(n^{-1/4})$.

Since $|n\mathbb{P}_n W_n^2 R(W_n)| \leq |n\mathbb{P}_n W_n^2|\max_{1\leq i\leq n}|R(W_n(X_i))|$, it remains to show that the maximum is $o_P(1)$. Write $V_n := W_n - \frac{h_n'\dot{\ell}_{\theta_0}}{2\sqrt{n}}$. Then,

$$\max_i |W_n(X_i)| \leq \max_i |\tfrac{1}{2\sqrt{n}}h_n'\dot{\ell}_{\theta_0}(X_i)| + \max_i |V_n(X_i)|.$$

By Markov's inequality,

$$P\Big(\max_{1\leq i\leq n}|\tfrac{1}{\sqrt{n}}h_n'\dot{\ell}_{\theta_0}(X_i)| > \varepsilon\Big) \leq nP(|\tfrac{1}{\sqrt{n}}h_n'\dot{\ell}_{\theta_0}(X_i)| > \varepsilon)$$
$$\leq \varepsilon^{-2}P_{\theta_0}((h_n'\dot{\ell}_{\theta_0})^2\mathbb{1}\{(h_n'\dot{\ell}_{\theta_0})^2 > n\varepsilon^2\}),$$

which converges to zero as $n\to\infty$ for every $\varepsilon > 0$. Thus, $\max_i|\frac{1}{\sqrt{n}}h_n'\dot{\ell}_{\theta_0}(X_i)|$ converges to zero in probability. Since Assumption 4 (ii) and (i) imply that $n\mathbb{P}_n V_n^2 = nP_{\theta_0}V_n^2 + o_P(1) = o_P(1)$, we have $\max_i V_n^2(X_i) = o_P(1)$ and hence $\max_i|V_n(X_i)| = o_P(1)$. Conclude that $\max_i|W_n(X_i)|$ converges to zero in probability and so does $\max_i|R(W_n(X_i))|$.

7.3. *Proof of Theorem 4.4.* We will prove the Theorem under the weaker assumption in Remark 2. We will use the following Lemma (an analogue of Lemma 10 [16]).

LEMMA 7.2. *Recall the definition $I_n(\Pi_n, X^{(n)}, \epsilon)$ in Remark 2 and define $q_\theta^{(n)} = p_\theta^{(n)}/p_{\theta_0}^{(n)}\mathrm{e}^{u_\theta}$. Then we have for any $C, \varepsilon > 0$*

$$P_{\theta_0}^{(n)}\left(\int_{B(\theta_0,\varepsilon)} q_\theta^{(n)}\mathrm{d}\Pi_n(\theta) \leq \mathrm{e}^{-(1+C)n\varepsilon^2}\times I_n(\Pi_n, X^{(n)}, \epsilon)\right) \leq \frac{1}{C^2 n\varepsilon^2}.$$

PROOF. Define a changed prior measure $\Pi_n^\star(\cdot)$ through $\mathrm{d}\Pi_n^\star(\theta) = \frac{\mathrm{e}^{u_\theta(X^{(n)})}}{\int \mathrm{e}^{u_\theta(X^{(n)})}\mathrm{d}\theta}\mathrm{d}\Pi_n(\theta)$. Lemma 10 of [16] then yields

$$P_{\theta_0}^{(n)}\left(\int_{B(\theta_0,\varepsilon)} q_\theta^{(n)}\mathrm{d}\Pi_n(\theta) \leq \mathrm{e}^{-(1+C)n\varepsilon^2} I_n(\Pi_n, X^{(n)}, \epsilon)\right)$$
$$= P_{\theta_0}^{(n)}\left(\int_{B(\theta_0,\varepsilon)} p_\theta^{(n)}/p_{\theta_0}^{(n)}\mathrm{d}\Pi_n^\star(\theta) \leq \Pi_n^\star(B(\theta_0,\varepsilon))\mathrm{e}^{-(1+C)n\varepsilon^2}\right) \leq \frac{1}{C^2 n\epsilon^2}. \qquad \square$$

Recall the definition $I_n(\Pi_n, X^{(n)}, \varepsilon_n) = \int_{B(\theta_0,\varepsilon_n)} \mathrm{e}^{u_\theta(X^{(n)})} \mathrm{d}\Pi_n(\theta)$ and define an event

$$\mathcal{A}_n = \left\{ X^{(n)} : \int_{B(\theta_0,\varepsilon_n)} q_\theta^{(n)} \mathrm{d}\Pi_n(\theta) > \mathrm{e}^{-2n\varepsilon_n^2} I_n(\Pi_n, X^{(n)}, \varepsilon_n) \right\}$$

where $q_\theta^{(n)} = p_\theta^{(n)} / p_{\theta_0}^{(n)} \mathrm{e}^{u_\theta}$. From assumptions in Remark 2, there exists a sequence $\widetilde{C}_n > 0$ such that the complement of the set

$$\mathcal{B}_n = \left\{ X^{(n)} : I_n(\Pi_n, X^{(n)}, \varepsilon_n) > \mathrm{e}^{-\widetilde{C}_n n\varepsilon_n^2} \text{ and } \sup_{\Theta_n^c \cup d_n(\theta,\theta_0) > \varepsilon_n} |u_\theta(X^{(n)})| \le \widetilde{C}_n n\varepsilon_n^2 \right\}$$

has a vanishing probability. Lemma 7.2 then yields $P_{\theta_0}^{(n)}[\mathcal{A}_n^c \cup \mathcal{B}_n^c] = o(1)$ as $n \to \infty$. The following calculations are thus conditional on the set $\mathcal{A}_n \cap \mathcal{B}_n$. On this set, we can lower-bound the denominator of (4.2) as follows

$$\int_\Theta q_\theta^{(n)} \mathrm{d}\Pi_n(\theta) > \int_{B(\theta_0,\varepsilon_n)} q_\theta^{(n)} \mathrm{d}\Pi_n(\theta) > \mathrm{e}^{-2n\varepsilon_n^2} I_n(\Pi_n, X^{(n)}, \varepsilon_n) \ge \mathrm{e}^{-(2+\widetilde{C}_n)n\varepsilon_n^2}.$$

We first show that $P_{\theta_0}^{(n)}[\Pi_n^\star(\Theta \backslash \Theta_n \,|\, X^{(n)})] = o(1)$ as $n \to \infty$. On the set $\mathcal{A}_n \cap \mathcal{B}_n$ we have from (4.4) and from the Fubini's theorem

$$\begin{aligned} P_{\theta_0}^{(n)} \left[\Pi_n^\star(\Theta \backslash \Theta_n \,|\, X^{(n)}) \right] &= P_{\theta_0}^{(n)} \left[\frac{\int_{\Theta \backslash \Theta_n} q_\theta^{(n)} \mathrm{d}\Pi_n(\theta)}{\int_\Theta q_\theta^{(n)} \mathrm{d}\Pi_n(\theta)} \right] \le \mathrm{e}^{2n\varepsilon_n^2} \frac{\Pi_n^\star(\Theta \backslash \Theta_n)}{\Pi_n^\star(B_n(\theta_0, \varepsilon_n))} \\ &= \mathrm{e}^{2(1+\widetilde{C}_n)n\varepsilon_n^2} \frac{\Pi_n(\Theta \backslash \Theta_n)}{\Pi_n(B_n(\theta_0, \varepsilon_n))} = o(1). \end{aligned}$$

For some $J > 0$ (to be determined later) we define the complement of the ball around the truth as a union of shells

$$U_n = \{\theta \in \Theta_n : d_n(\theta, \theta_0) > MJ\varepsilon_n\} = \bigcup_{j \ge J} \Theta_{n,j}$$

where each shell equals

$$\Theta_{n,j} = \{\theta \in \Theta_n : Mj\varepsilon_n < d_n(\theta, \theta_0) \le M(j+1)\varepsilon_n\}.$$

We now invoke the local entropy Assumption (3.2) in [16] which guarantees (according to Lemma 9 in [16]) that there exist tests ϕ_n (for each n) such that

$$P_{\theta_0}^{(n)} \phi_n \lesssim \mathrm{e}^{n\varepsilon_n^2 - nM^2\varepsilon_n/2} \quad \text{and} \quad P_\theta^{(n)}(1 - \phi_n) \le \mathrm{e}^{-nM^2\varepsilon_n^2 j^2/2} \tag{7.1}$$

for all $\theta \in \Theta_n$ such that $d_n(\theta, \theta_0) > M\varepsilon_n j$ and for every $j \in \mathbb{N} \backslash \{0\}$ and $M > 0$. One can then write

$$\begin{aligned} P_{\theta_0}^{(n)} \Pi \left(\theta \in \Theta : d(\theta, \theta_0) > MJ\varepsilon_n \,|\, X^{(n)} \right) &\le P_{\theta_0}^{(n)} \Pi(\Theta_n^c \,|\, X^{(n)}) + P_{\theta_0}^{(n)} \phi_n + P_{\theta_0}^{(n)}(\mathcal{A}_n^c) + P_{\theta_0}^{(n)}(\mathcal{B}_n^c) \\ &\quad + \sum_{j \ge J} P_{\theta_0}^{(n)} [\Pi(\Theta_{n,j} \,|\, X^{(n)})(1 - \phi_n) \mathbb{I}(\mathcal{A}_n \cap \mathcal{B}_n)] \end{aligned}$$

For the last term above, we recall that $\Pi(\Theta_{n,j} \,|\, X^{(n)}) = \frac{\int_{\Theta_{n,j}} q_\theta^{(n)} \mathrm{d}\Pi_n(\theta)}{\int_\Theta q_\theta^{(n)} \mathrm{d}\Pi_n(\theta)}$. We bound the denominator as before. Regarding the numerator, on the event $\mathcal{B}_n$ we have from (7.1) and from the Fubini's theorem

$$P_{\theta_0}^{(n)} \int_{\Theta_{n,j}} q_\theta^{(n)} \mathrm{d}\Pi_n(\theta)(1 - \phi_n) \le \mathrm{e}^{-nM^2\varepsilon_n^2 j^2/2 + \widetilde{C}_n n\varepsilon_n^2} \Pi_n(\Theta_{n,j}) \tag{7.2}$$

Putting the pieces together, we obtain

$$P_{\theta_0}^{(n)}[\Pi(\Theta_{n,j} \mid X^{(n)})(1-\phi_n)\mathrm{I}(\mathcal{A}_n \cap \mathcal{B}_n)] \leq \mathrm{e}^{-nM^2\varepsilon_n^2 j^2/2+2(1+\widetilde{C}_n)n\varepsilon_n^2} \frac{\Pi_n(\Theta_{n,j})}{\Pi_n[B_n(\theta_0,\varepsilon_n)]}.$$

Assumption (3.4) of [16] writes as

$$\frac{\Pi_n(\Theta_{n,j})}{\Pi_n[B_n(\theta_0,\varepsilon_n)]} \leq \mathrm{e}^{nM^2\varepsilon_n^2 j^2/4} \tag{7.3}$$

which yields

$$P_{\theta_0}^{(n)}\Pi\left(\theta \in \Theta : d(\theta,\theta_0) > MJ\varepsilon_n \mid X^{(n)}\right) \leq o(1) + \sum_{j \geq J} \mathrm{e}^{-n\varepsilon_n^2(M^2j^2/4-2-2\widetilde{C}_n)}.$$

The right hand side converges to zero as long as $J = J_n \to \infty$ fast enough so that $\widetilde{C}_n = o(J_n)$ and $n\varepsilon_n^2$ is bounded away from zero. □

7.4. *Proof of Theorem 4.5.* We define the event

$$\mathcal{A} = \left\{ X^{(n)} \in \mathcal{X} : \int \frac{\widetilde{p}_\theta^{(n)}}{\widetilde{p}_{\theta^*}^{(n)}} \mathrm{d}\widetilde{\Pi}_n(\theta) > \mathrm{e}^{-(1+C)n\epsilon^2} \widetilde{\Pi}_n[B(\epsilon, \widetilde{P}_{\theta^*}^{(n)}, P_{\theta_0}^{(n)})] \right\}.$$

The following lemma shows that $P_{\theta_0}^{(n)}[\mathcal{A}^c] = o(1)$ as $n \to \infty$.

LEMMA 7.3. *For $k \geq 2$, every $\epsilon > 0$ and a prior measure $\widetilde{\Pi}_n(\theta)$ on Θ, we have for every $C > 0$*

$$P_{\theta_0}^{(n)}\left(\int \frac{\widetilde{p}_\theta^{(n)}}{\widetilde{p}_{\theta^*}^{(n)}} \mathrm{d}\widetilde{\Pi}_n(\theta) \leq \mathrm{e}^{-(1+C)n\epsilon^2} \widetilde{\Pi}_n[B(\epsilon, \widetilde{P}_{\theta^*}^{(n)}, P_{\theta_0}^{(n)})] \right) \leq \frac{1}{C^2 n\epsilon^2}.$$

PROOF. This follows directly from Lemma 10 in [16].

We now define $U_n(\epsilon) = \Pi_n(\theta \in \Theta : d(\widetilde{P}_\theta^{(n)}, \widetilde{P}_{\theta^*}^{(n)}) > \epsilon \mid X^{(n)})$. For every $n \geq 1$ and $J \in \mathbb{N}\backslash\{0\}$, we can decompose

$$\begin{aligned} P_{\theta_0}^{(n)} U_n(JM\varepsilon_n) =& P_{\theta_0}^{(n)}[U_n(JM\varepsilon_n)\phi_n] + P_{\theta_0}^{(n)}[U_n(JM\varepsilon_n)(1-\phi_n)\mathbb{I}(\mathcal{A}^c)] \\ &+ P_{\theta_0}^{(n)}[U_n(JM\varepsilon_n)(1-\phi_n)\mathbb{I}(\mathcal{A})]. \end{aligned}$$

The first term is bounded (from the assumption (4.9)) as

$$P_{\theta_0}^{(n)}[U_n(JM\varepsilon_n)\phi_n] \leq P_{\theta_0}^{(n)}\phi_n \lesssim \mathrm{e}^{-n\varepsilon_n^2 J^2 M^2}.$$

The second term can be bounded by $P_{\theta_0}^{(n)}[\mathbb{I}(\mathcal{A}^c)] \leq \frac{1}{C^2 J^2 M^2 n\varepsilon_n^2}$ which converges to zero as $n\varepsilon_n^2 \to \infty$. The last term satisfies

$$\begin{aligned} P_{\theta_0}^{(n)}[U_n(JM\varepsilon_n)(1-\phi_n)\mathbb{I}(\mathcal{A})] &= P_{\theta_0}^{(n)}\left[(1-\phi_n)\mathbb{I}(\mathcal{A}) \frac{\int_{\theta : d(\widetilde{P}_\theta^{(n)}, \widetilde{P}_{\theta^*}^{(n)}) > JM\varepsilon_n} \frac{\widetilde{p}_\theta^{(n)}}{\widetilde{p}_{\theta^*}^{(n)}} \widetilde{\Pi}_n(\theta)\mathrm{d}\theta}{\int_\Theta \frac{\widetilde{p}_\theta^{(n)}}{\widetilde{p}_{\theta^*}^{(n)}} \widetilde{\Pi}_n(\theta)\mathrm{d}\theta} \right] \\ &\leq \frac{\mathrm{e}^{(1+C)n\epsilon^2}}{\widetilde{\Pi}_n[B(\epsilon, \widetilde{P}_{\theta^*}^{(n)}, P_{\theta_0}^{(n)})]} \int_{\theta : d(\widetilde{P}_\theta^{(n)}, \widetilde{P}_{\theta^*}^{(n)}) > JM\varepsilon_n} \left[\int_{\mathcal{X}} (1-\phi_n) p_{\theta_0}^{(n)} \frac{\widetilde{p}_\theta^{(n)}}{\widetilde{p}_{\theta^*}^{(n)}} \right] \widetilde{\Pi}_n(\theta)\mathrm{d}\theta \\ &\leq \frac{\mathrm{e}^{(1+C)n\epsilon^2}}{\widetilde{\Pi}_n[B(\epsilon, \widetilde{P}_{\theta^*}^{(n)}, P_{\theta_0}^{(n)})]} \sum_{j \geq J} \int_{U_{n,j}} Q_\theta^{(n)}(1-\phi_n)\mathrm{d}\widetilde{\Pi}_n(\theta), \end{aligned}$$

where $U_{n,j} = \{\theta : jM\varepsilon_n < d(\widetilde{P}_\theta^{(n)}, \widetilde{P}_{\theta^*}^{(n)}) \le (j+1)M\varepsilon_n)\}$. The tests (from the assumption (4.9)) satisfy $Q_\theta^{(n)}(1-\phi_n) \le \mathrm{e}^{-nj^2M^2\varepsilon_n^2/4}$ uniformly on $U_{n,j}$. Then we find (using the assumption (4.10))

$$P_{\theta_0}^{(n)}[U_n(JM\varepsilon_n)(1-\phi_n)\mathbb{I}(\mathcal{A})] \le \mathrm{e}^{(1+C)n\varepsilon_n^2} \sum_{j\ge J} \mathrm{e}^{-nj^2M^2\varepsilon_n^2/4 + nj^2M^2\varepsilon_n^2/8}.$$

The sum converges to zero when $n\varepsilon_n^2$ is bounded away from zero and $J \to \infty$. □

7.5. *Proof of Lemma 4.7.* We can write

$$\log \frac{\widetilde{p}_{\theta^*+\varepsilon_n h}^{(n)}}{\widetilde{p}_{\theta^*}^{(n)}} = \log \frac{C_{\theta^*+\varepsilon_n h}}{C_{\theta^*}} + \log \frac{p_{\theta^*+\varepsilon_n h}^{(n)}}{p_{\theta^*}^{(n)}} + u_{\theta^*+\varepsilon_n h} - u_{\theta^*}. \tag{7.4}$$

This yields, from Lemma 19.31 in [41], that

$$\mathbb{G}_n\left(\sqrt{n}\log\frac{p_{\theta^*+h/\sqrt{n}}}{p_\theta^*} - h'\dot{\ell}_{\theta^*}\right) \to 0 \quad \text{in } P_0,$$

where $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_{\theta_0})$ is the empirical process. Assuming that

$$P_{\theta_0}\log\left(\frac{p_\theta}{p_{\theta^*}}\right) = P_{\theta_0}\dot{\ell}'_{\theta^*}(\theta-\theta^*) + \frac{1}{2}(\theta-\theta^*)'V_{\theta^*}(\theta-\theta^*) + o(\|\theta-\theta^*\|^2) \quad \text{as } \theta\to\theta^* \tag{7.5}$$

one obtains

$$\begin{aligned}\log\frac{p_{\theta^*+h/\sqrt{n}}^{(n)}}{p_{\theta^*}^{(n)}} = n\mathbb{P}_n\log\frac{p_{\theta^*+h/\sqrt{n}}}{p_{\theta^*}} &= o_P(1) + \mathbb{G}_n h'\dot{\ell}_{\theta^*} + nP_{\theta_0}\log\frac{p_{\theta^*+h/\sqrt{n}}}{p_{\theta^*}} \\ &= o_P(1) + \mathbb{G}_n h'\dot{\ell}_{\theta^*} + \frac{h_n'V_{\theta^*}h}{2} + \sqrt{n}P_{\theta_0}h'\dot{\ell}_\theta\end{aligned}$$

If we assume asymptotic linearity of u_θ around θ^*, i.e.

$$u_{\theta^*+h/\sqrt{n}}(X^{(n)}) - u_{\theta^*}(X^{(n)}) = \frac{1}{\sqrt{n}}h'u^\star(X^{(n)}) + o_P(1) \tag{7.6}$$

for some $u^\star(X^{(n)})$ and

$$\log\frac{C_{\theta^*+h_n/\sqrt{n}}}{C_{\theta^*}} = \frac{\dot{C}'_{\theta^*}h_n}{\sqrt{n}} + o(1)$$

then (4.14) holds with (4.15). □

Acknowledgements. Tetsuya Kaji acknowledges the support from the Richard N. Rosett Faculty Fellowship and the Liew Family Faculty Fellowship at the University of Chicago Booth School of Business. Veronika Rockova gratefully acknowledges the support from James S. Kemper Faculty Scholarship and the National Science Foundation (DMS: 1944740).

REFERENCES

[1] Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics* **37** 697–725.
[2] Andrieu, C. and Vihola, M. (2015). Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *The Annals of Applied Probability* **25** 1030–1077.
[3] Azadi, S., Olsson, C., Darrell, T., Goodfellow, I. and Odena, A. (2019). Discriminator Rejection Sampling. In *International Conference on Learning Representations*.

[4] Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics* **164** 1139–1160.
[5] Beaumont, M. A., Zhang, W. and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162** 2025–2035.
[6] Belloni, A. and Chernozhukov, V. (2009). On the Computational Complexity of MCMC-based Estimators in Large Samples. *The Annals of Statistics* **37** 2011–2055.
[7] De Blasi, P. and Walker, S. (2013). Bayesian Estimation of the Discrepancy with Misspecified Parametric Models. *Bayesian Analysis* **8** 781 –800.
[8] Deligiannidis, G., Doucet, A. and Pitt, M. K. (2018). The correlated pseudomarginal method. *Journal of the Royal Statistical Society (Series B)* **80** 839–870.
[9] Diggle, P. J. and Gratton, R. J. (1984). Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society (Series B)* **46** 193–212.
[10] Donnet, S., Rivoirard, V., Rousseau, J. and Scricciolo, C. (2018). Posterior concentration rates for empirical Bayes procedures with applications to Dirichlet process mixtures. *Bernoulli* **24** 231–256.
[11] Fearnhead, P. and Prangle, D. (2011). Constructing ABC summary statistics: semi-automatic ABC. *Nature Precedings* 1–10.
[12] Frazier, D. and Drovandi, C. (2021). Robust Approximate Bayesian Inference with Synthetic Likelihood. *Journal of Computational and Graphical Statistics*.
[13] Frazier, D., Robert, C. and Rousseau, J. (2020). Model misspecification in approximate Bayesian computation: consequences and diagnostics. *Journal of the Royal Statistical Society (Series B)* **82** 421–444.
[14] Frazier, D., Martin, G., Robert, C. and Rousseau, J. (2018). Asymptotic properties of approximate Bayesian computation. *Biometrika* **105** 593–607.
[15] Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence Rates of Posterior Distributions. *The Annals of Statistics* **28** 500–531.
[16] Ghosal, S. and van der Vaart, A. W. (2007). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics* **35** 192–223.
[17] Ghosal, S. and van der Vaart, A. W. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, Cambridge.
[18] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D. and Ozair, S. (2014). Generative adversarial nets. *Proceedings of the 27th International Conference on Neural Information Processing Systems* **2** 2672–2680.
[19] Gourieroux, C., Monfort, A. and Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics* **8** 85–118.
[20] Gutmann, M. U. and Corander, J. (2016). Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research* **17** 4256–4302.
[21] Gutmann, M. U., Dutta, R., Kaski, S. and Corander, J. (2018). Likelihood-free inference via classification. *Statistics and Computing* **28** 411–425.
[22] Jiang, B., Wu, T.-y., Zheng, C. and Wong, W. H. (2017). Learning summary statistic for approximate Bayesian computation via deep neural network. *Statistica Sinica* 1595–1618.
[23] Kaji, T., Manresa, E. and Pouliot, G. (2020). An Adversarial Approach to Structural Estimation. arxiv:2007.06169.
[24] Kleijn, B. and van der Vaart, A. (2006). Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics* **34** 837–877.
[25] Kleijn, B. and van der Vaart, A. (2012). The Bernstein-Von-Mises theorem under misspecification. *Electronic Journal of Statistics* **6** 354 – 381.
[26] Lovasz, L. and Simonovits, M. (1993). Random walks in a convex body and an improved volume algorithm. *Random Structures and Algorithms* **4** 359–412.
[27] Marin, J., Pillai, N., Robert, C. and Rousseau, J. (2014). Relevant statistics for Bayesian model choice. *Journal of the Royal Statistiscial Society (Series B)* **76** 833–859.
[28] Mengersen, K. and Tweedie, R. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics* **24** 101–121.
[29] O'Neill, P. D., Balding, D. J., Becker, N. G., Eerola, M. and Mollison, D. (2000). Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society (Series C)* **49** 517–542.
[30] O'Ryan, C., Harley, E. H., Bruford, M. W., Beaumont, M., Wayne, R. K. and Cherry, M. I. (1998). Microsatellite analysis of genetic diversity in fragmented South African buffalo populations. *Animal Conservation* **1** 85–94.

[31] Patilea, V. (2001). Convex Models, MLS and Misspecification. *The Annals of Statistics* **20** 94–123.
[32] Price, L., Drovandi, C., Lee, A. and Nott, D. (2018). Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics* **27** 1–119.
[33] Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. and Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular biology and evolution* **16** 1791–1798.
[34] Ricker, W. E. (1954). Stock and recruitment. *Journal of the Fisheries Board of Canada* **11** 559–623.
[35] Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, second ed. Springer, New York.
[36] Robert, C. P., Cornuet, J.-M., Marin, J.-M. and Pillai, N. S. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences* **108** 15112–15117.
[37] Rousseau, J. and Szabo, B. (2017). Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator. *The Annals of Statistics* **45** 833–865.
[38] Tavaré, S., Balding, D. J., Griffiths, R. C. and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics* **145** 505–518.
[39] Tsvetkov, D., Hristov, L. and Angelova-Slavova, R. (2017). On the Convergence of the Metropolis-Hastings Markov Chains. *Serdica Math. J.* **43** 93–110.
[40] Turner, R., Hung, J., Frank, E., Saatchi, Y. and Yosinski, J. (2019). Metropolis-Hastings generative adversarial networks. In *International Conference on Machine Learning* 6345–6353. PMLR.
[41] van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.
[42] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York.
[43] Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466** 1102–1104.