أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Toward Scalable Bayesian Uncertainty for Machine Learning with the Ray Tracing Sampler

Yasir Alsugair, **Joshua Speagle**

yasiralsugair@gmail.com, j.speagle@utoronto.ca

Abstract

Minibatch gradients make Bayesian sampling of neural networks affordable; their noise biases what the sampler draws. Behroozi's ray tracing sampler replaces HMC's unbounded momentum with a constant-speed direction on a sphere, so gradient noise can rotate the direction but cannot heat it. I checked that claim in stages: exact full-batch chains on MNIST-scale networks (D = 50,890 and 12,810) gave every shortcut an exact reference; noise experiments measured the tolerance laws ($\sigma_c \propto h^{-1}$ for ray tracing, $h^{-1/2}$ for HMC); and a minibatch campaign on Gaia XP spectra (126,156 giants, an 11,394 parameter network) climbed a noise-model ladder that took held-out calibration from 6.88 to 1.08, cut the $|z| > 4$ tail from 108 to 14 stars, and reached the best held-out NLL of every method (−1.973). Held-out stars were on average 27% more probable under the posterior average than under the single best fit. A companion analysis by Speagle explains the measured scalings and predicts the published tolerances within 30% with no fitted quantities. Robustness to gradient noise is validated; the cost saving is partial (about 1.5x at matched accuracy, batch 256, one problem); posterior width, anisotropic noise, and modern network scale remain open.

Date: August 22, 2026 (Final Report)

Code: https://github.com/YasirAlsugair/ray-tracing-sampler-experiments

In partnership with

العقيل الخيرية
مؤسسة أبناء عبدالرحمن ناصر العقيل الخيرية
ALAGIL FOUNDATION

رزم للاستثمار
RZM INVESTMENT

Contents

1 Introduction **2**

2 The sampler and the three targets **3**

3 Results **4**

3.1 High acceptance is not convergence: ground truth on MNIST 5

3.2 Noise tolerance: one power of the step size better 6

3.3 The ladder fixes calibration 6

3.4 The exactness stamp is on the location 7

3.5 SGHMC runs hot; the ray cannot 7

3.6 The gain is in the error bars 8

3.7 Sanity checks 9

4 Why the tolerance differs: the geometry of the auxiliary variable **10**

5 Discussion **12**

6 Limitations and future work **13**

7 Conclusion **13**

References **14**

1 Introduction

Sampling a network posterior runs on gradient-based MCMC: to average predictions over it, a sampler such as Hamiltonian Monte Carlo [1] needs the gradient of the log-posterior, and on modern datasets computing it exactly means a full pass over the data at every step. The practical replacement estimates the gradient from a minibatch of B of the N training examples: unbiased, N/B times cheaper per step, and noisy. That noise is not a cosmetic problem. The Metropolis test that makes trajectory samplers exact compares full-data densities, and naive subsampling breaks it [2]. The standard stochastic-gradient samplers each give something up: SGLD [3] drops the test and accepts a step-size-dependent bias, while SGHMC [4] adds a friction term that must be tuned against the gradient-noise scale it is meant to absorb.

The ray tracing sampler of Behroozi [5] changes the geometry of the auxiliary variable instead. Where HMC carries an unbounded momentum $p \in \mathbb{R}^D$, the ray carries a unit direction u on the sphere S^{D-1} and moves at constant speed through a medium with refractive index $n(x) = \pi(x)^{1/(D-1)}$; Snell's law bends the path toward higher probability. The noisy kick is projected onto the tangent plane of the sphere, so gradient noise can rotate the direction but cannot change its norm. Heating a momentum makes it hotter without bound; heating a direction only makes it point elsewhere. Section 2 sets out the mechanism. The claim to check is quantitative: Behroozi's own experiment at $D = 10{,}000$ reports a critical noise variance of $\sigma_c^2 \approx 3.2$ for ray tracing against 0.012 for HMC, a variance ratio above 250 [5].

A claim this size should not be taken on trust, so the summer was arranged to check it in stages. The first stage built exact ground truth: small MNIST networks (an MLP with $D = 50{,}890$ parameters, a CNN with $D = 12{,}810$) where exact full-batch chains are still feasible, so every later shortcut had something exact to compare with. The second stage measured how much gradient noise each sampler tolerates before its chain breaks down, first under idealized isotropic noise and then under real minibatch noise from a Bayesian linear regression (UCI superconductivity, $D = 81$, $N = 17{,}010$), together with the cost in dataset passes per effective sample. The third stage moved to a real scientific target: regressing $[\alpha/\mathrm{M}]$ from Gaia XP spectra, the data behind Laroche and Speagle [6], with a $D = 11{,}394$ network, 126,156 stars, and minibatch ray tracing carried through calibration, baselines, and an exactness certificate. The last stage is theory, and the theory is not mine: the companion analysis by Speagle [7] explains why the tolerance differs and makes quantitative predictions, which I confront with measurement in Section 4. Ray tracing is also not the only sampler with this geometry: the microcanonical family likewise replaces the momentum with a constant-speed direction [8–10], so the tolerance question is broader than one method.

This report contributes:

- Exact ground truth on MNIST-scale networks: a $T = 1$ posterior target where full-batch chains are feasible, so every minibatch chain in this report is checked against an exact reference (Section 3).
- Measured noise-tolerance laws and an honest cost frontier. The critical noise level σ_c scales as h^{-1} for ray tracing against $h^{-1/2}$ for HMC under idealized isotropic noise ($D = 16$ to 1024); under real minibatch noise the advantage survives but the clean scaling weakens, and the gate-passing saving at matched accuracy was about 1.5x at batch 256, on one problem (Section 3).
- A ray tracing campaign on a real survey regression, carried through a calibration ladder, matched baselines, and an exactness certificate to one measured gain: held-out stars were on average 27% more probable under the posterior average than under the single best fit (25,232 test stars; Section 3).

- The empirical confrontation of the companion analysis by Speagle [7]. The theory is Speagle's; the test is mine: its predicted scalings and absolute tolerances are compared against Behroozi's published experiment and this report's measurements (Section 4).

2 The sampler and the three targets

Everything measured this summer ran on three targets, ordered by how much could be checked exactly: MNIST posteriors with exact ground truth, a linear regression where the gradient noise is under my control, and a real survey posterior where it is not. This section fixes the sampler and the setup of each; results follow in Section 3.

The sampler. The ray tracing sampler [5] is a gradient-based MCMC that propagates a constant-speed ray through a medium with refractive index $n(x) = \pi(x)^{1/(D-1)}$, where π is the target density; Snell's law bends the ray toward higher probability. The contrast that matters is with HMC [1]: HMC carries an unbounded momentum $p \in \mathbb{R}^D$, ray tracing carries a unit direction u on the sphere S^{D-1}. A noisy minibatch gradient enters HMC's momentum and raises its expected kinetic energy, which is heating; for the ray the same kick is projected onto the tangent plane of the sphere, so noise can rotate the direction but cannot change its norm. Heating a momentum makes it hotter without bound; heating a direction only makes it point elsewhere. Figure 1 draws the two responses to the same noise.

Minibatch gradients. For a posterior over N data points, one exact gradient touches all N points while a minibatch of size B touches only B, so each step is N/B times cheaper. The estimate is unbiased and its covariance is proportional to $1/B$; the per-step direction is another matter. On a Bayesian linear regression on the UCI superconductivity data ($D = 81$, $N = 17{,}010$), the minibatch gradient at batch 2,048 sits 63 degrees off the exact gradient on average, and 90 degrees would be a random direction (Figure 2). The average is right; no individual step uses it.

Ground truth first. Before trusting any minibatch shortcut I built targets where the exact answer is available. The MNIST study (exp6 in the repository) defines a $T = 1$ posterior over two small classifiers, an MLP with $D = 50{,}890$ parameters and a CNN with $D = 12{,}810$, small enough that exact full-batch chains are still feasible, so every later approximation has something to be checked against. Each network was sampled two ways: with exact full-batch chains, and with the minibatch ray tracer under the noise-softened acceptance gate of Behroozi [5] (his Eq. 33). A 10-seed deep ensemble [11] serves as the non-Bayesian reference. What these chains found is in Section 3.

Measuring tolerance. Noise tolerance has a clean version and a real one, and I measured both. The clean version injects idealized isotropic gradient noise at a controlled amplitude, at dimensions $D = 16$ to 1024, and records the largest noise level σ_c each sampler tolerates at a given step size h. The real version runs the same comparison under actual minibatch noise on the superconductivity regression, where the noise is anisotropic and not mine to choose. The first isolates the scaling law; the second checks whether it survives a real dataset.

The Gaia campaign. The main target (exp7) is a real survey regression: Gaia XP spectral coefficients (55 BP plus 55 RP, 110 per star) [12] to the alpha abundance [α/M], the dataset behind Laroche and Speagle [6]. I reproduced the paper's pristine cut at 126,156 giants with APOGEE labels (the paper reports 123,804, within 1.9%), split 80/20 at seed 2003 into 100,924 training and 25,232 held-out stars. The network is an MLP (110-64-64-1, tanh) under a Glorot-style prior [13], with $D = 11{,}329$ parameters growing to 11,395 as the noise model gains its own. That noise model is

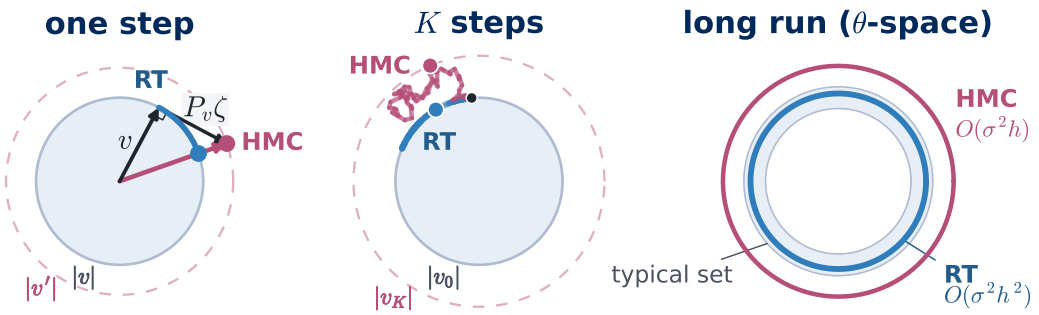


Figure 1 One noise kick, two auxiliary variables. Left: a single noisy kick in velocity space; HMC adds it to the momentum and lands outside the unit circle, while the ray's kick is projected onto the tangent plane and the endpoint stays on the circle. Middle: both samplers fed the identical noise sequence over many steps; the HMC speed wanders while the ray's stays at 1 and only its heading moves. Right: the long run in parameter space; the heated chain settles off the typical set at first order in the step size, the ray at second order. The quantity to watch is the norm: noise moves it for HMC and cannot move it for the ray.

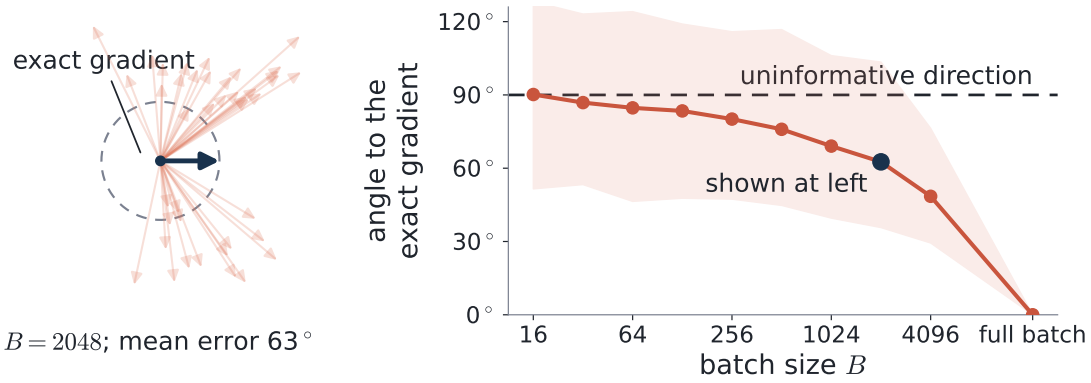


Figure 2 Minibatch gradients against the exact one. Left: a fan of minibatch gradient directions around the exact gradient for the superconductivity regression (D = 81, N = 17,010). Right: mean angle to the exact gradient as the batch grows; at batch 2,048 the average offset is still 63 degrees, where 90 degrees would be a random direction. Every sampler step uses one arrow from the fan.

the experimental variable: a ladder of nested likelihood rungs, from survey label errors taken at face value up to a per-star Student-t. Every chain ran Behroozi's released minibatch ray tracer with the Eq. 33 gate; the whole campaign cost roughly $85 to $90 of cloud GPU time.

Certificates and baselines. A chain is declared converged only by a drift-plus-acceptance rule: every monitored series (data misfit, weighted norm, noise-model parameters) must be level within its own noise band, with the acceptance rate always read next to the drift. A chain that rejects everything is frozen, and a frozen chain passes any stationarity test. A converged endpoint is then stamped by an exact finisher, full-batch gradients with the exact Metropolis test and no minibatch approximation anywhere, so the final location carries an exact certificate rather than an appeal to the gate. Three baselines fit the identical heteroscedastic target: a MAP fit with Adam, a 10-member deep ensemble, and a hand-tuned SGHMC [4] whose kinetic temperature is monitored, so any heating is measured rather than inferred.

3 Results

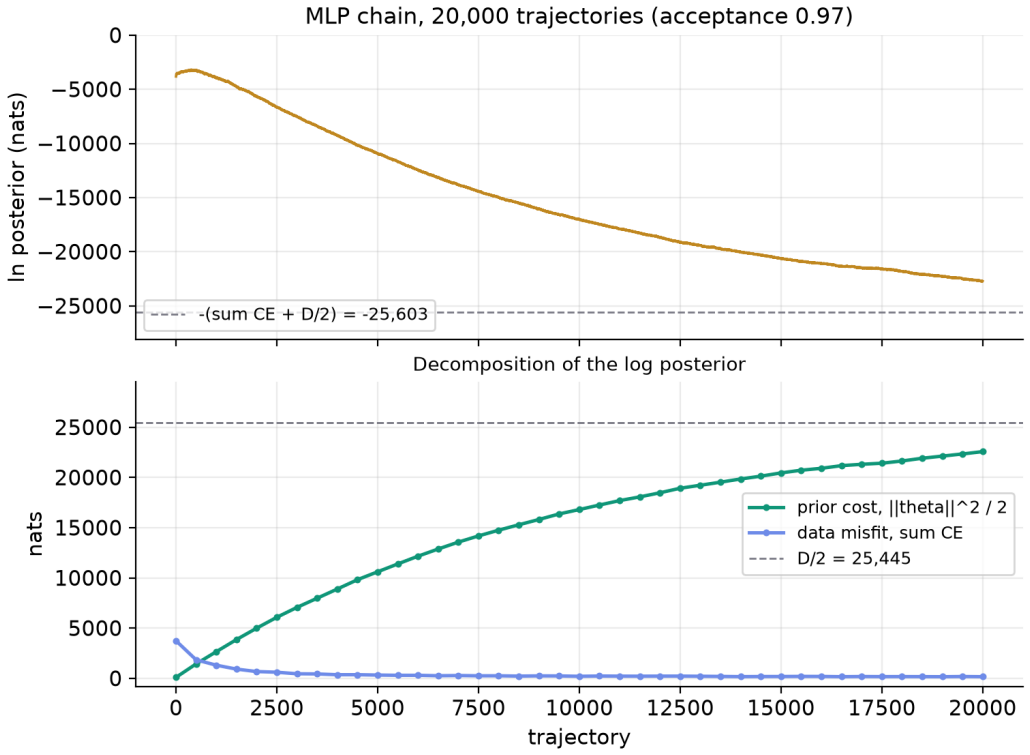


Figure 3 The march to the shell (MLP chain, 20,000 trajectories, acceptance 0.97). Top: the log posterior, still descending at the end of the run. Bottom: its decomposition; the data misfit (blue) flattens early while the prior cost $\|\theta\|^2/2$ (green) keeps climbing toward the $D/2$ line. Read the panels together: the budget is spent reaching the right norm, not fitting the data.

3.1 High acceptance is not convergence: ground truth on MNIST

Exact full-batch chains on the MNIST networks ran at 97 to 99% acceptance and never passed the stationarity checks; the gated minibatch chains passed on both architectures, the MLP ($D = 50{,}890$) at 5.6M minibatch steps and the CNN ($D = 12{,}810$) at 14.6M. Acceptance measures whether the step size is small enough, not whether the chain has arrived. The noise-softened acceptance gate (Eq. 33 of Behroozi [5]) that makes the minibatch runs valid is a time tax, not a bias: gated and ungated runs agree.

Where the budget went is visible in Figure 3. Chains start at low-norm states (the Adam point, random initializations) and spend the run marching out toward the shell at $\|\theta\|^2 \approx D$, where the posterior lives: the data misfit flattens almost immediately while the prior term keeps climbing. A prior-aware MAP-transit start, suggested by my mentor, pre-pays that burn-in. Chains launched from transit states accepted 92 to 96% at the full tuned step; raw bad-fit starts needed 12x to 350x smaller steps.

Table 1 compares the converged CNN chain with a 10-seed deep ensemble [11]. The chain essentially matches the ensemble on accuracy (0.9892 against 0.9904) and is well calibrated (ECE 0.0033); an NLL gap remains (0.0854 against 0.0310). The more practical head-to-head fixes a matched review budget: flag the 11% of test images each method is least sure about and count the errors caught. The converged chain catches 90% of its own errors at 75x enrichment; the single-network point estimate catches 86% at 52x; the ensemble catches 92% at 93x. The three flagged sets only half-overlap (Jaccard 0.32 to 0.52), so the methods are not ranking the same images. On fake images, per-sample disagreement is the most consistent detector: confidence

Table 1 Converged CNN chain against a 10-seed deep ensemble on the MNIST test set. The chain matches the ensemble's accuracy and is well calibrated; an NLL gap remains.

method	accuracy	NLL	ECE
ray tracing chain (converged)	0.9892	0.0854	0.0033
deep ensemble (10 seeds)	0.9904	0.0310	

and disagreement decouple there.

3.2 Noise tolerance: one power of the step size better

Ray tracing tolerates one extra power of the step size in gradient noise. Under idealized isotropic noise the critical level σ_c (the largest injected noise at which the sampler still passes its accuracy check) scales as h^{-1} for ray tracing and as $h^{-1/2}$ for HMC, and the measured slopes match those exponents across $D = 16$ to 1024 (Figure 4a). Halving the step doubles ray tracing's tolerance; HMC gains only a factor of $\sqrt{2}$.

Real minibatch noise keeps the ranking and spoils the clean law. On the superconductivity regression ($D = 81$, $N = 17{,}010$) the local step-size exponent falls from about 1 toward 0.6 as the step shrinks (Figure 4b). The advantage survives; the idealized scaling does not. The reason is that real minibatch noise is anisotropic, and anisotropy enters at first order in the step size; the companion analysis [7] makes this quantitative, and Section 4 returns to it.

The cost frontier prices the difference in dataset passes per effective sample (Figure 4c). Minibatch HMC needs roughly 30 passes per effective sample and gets worse as the batch shrinks; minibatch ray tracing sits below the full-batch reference line at small batches, which is the regime minibatching exists for.

Two numbers keep this honest. First, configurations at the smallest batches looked much faster still, but they fail the study's own accuracy gate, so I do not quote them; at matched accuracy the gate-passing saving was about 1.5x at batch 256, on this one problem. Second, the large version of the effect is Behroozi's own measurement, not mine: his $D = 10{,}000$ experiment reports a critical noise variance σ_c^2 of about 3.2 for ray tracing against 0.012 for HMC, a variance ratio above 250 [5].

3.3 The ladder fixes calibration

On the Gaia target [6] the sampler was never the calibration problem; the noise model was. The standardized residuals on held-out stars have standard deviation 6.88 taking the survey's label errors at face value, 6.23 after adding the posterior's member disagreement, 0.95 after adding one global intrinsic scatter s, and 1.08 with a per-star scatter $\sigma(x)$ predicted by the network (Figure 5). A calibrated model gives 1. Every rung is the same sampler pointed at a richer likelihood.

The heteroscedastic rung ($D = 11{,}394$) is the chain of record. Its calibration is flat where the global-s rung's was not: across the four label-error bins the z standard deviation runs 0.97 / 1.07 / 1.07 / 1.07, against a spread of 0.39 to 1.43 under the global scatter, so precise stars are no longer padded and noisy stars are no longer overconfident. Overall SD(z) is 1.08 at RMSE 0.0494, with 108 held-out stars past $|z| = 4$.

The Student-t rung ($D = 11{,}395$) fixes the tail shape. Certified at 8.5M steps with every drift check level, it lands at $\nu = 5.48$, improves RMSE to 0.0480, and cuts the tail count from 108 to 14 under the fitted t's own tail probability, against a chance count of 1.6. Its test NLL of -1.973 is the best of every method in this report. The width is still not fully honest: the chain under-covers by about 4 to 5% in PIT, and that is the open calibration item.

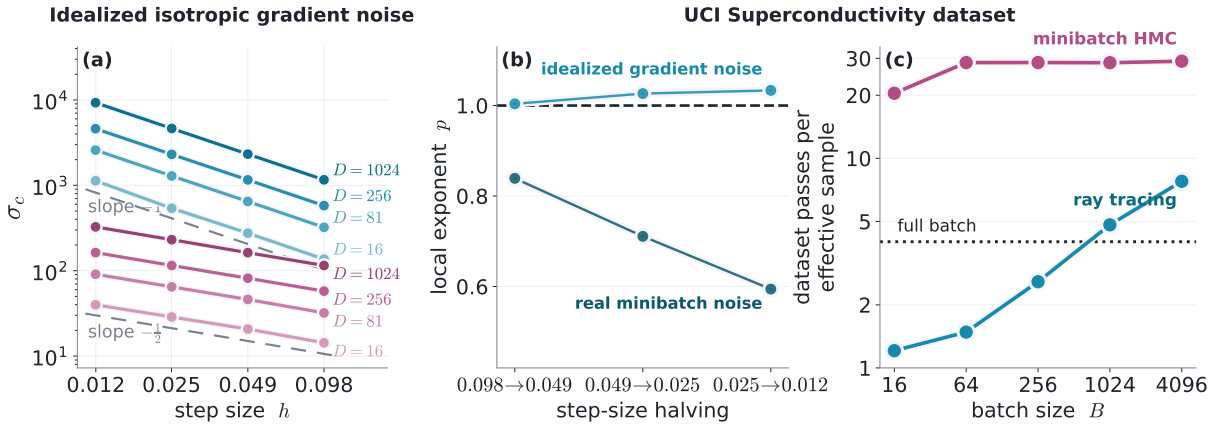


Figure 4 Noise tolerance and cost. (a) Critical noise σ_c against step size h under idealized isotropic noise, $D = 16$ to 1024: ray tracing follows the slope −1 guide, HMC the slope −1/2 guide. (b) The local step-size exponent on the superconductivity regression: near 1 under idealized noise, falling toward 0.6 under real minibatch noise. (c) Dataset passes per effective sample against batch size: minibatch HMC needs roughly 30 passes per effective sample, and more at smaller batches; minibatch ray tracing is below the full-batch line at small batches.

3.4 The exactness stamp is on the location

An exact full-batch test accepts a converged minibatch endpoint at 0.98: where these chains settle is a genuine posterior location, not a minibatch artifact. Getting there was the hard part.

Convergence was a campaign. At batch 1,024 the chain of record's weighted norm marched indefinitely; escalating to batch 4,096 halved the gate's measured noise scale σ_{sto}, and the march then decelerated, plateaued, and reversed; every monitored drift series went level at 16M total steps. The procedural rule the campaign left behind: never read a drift check without the acceptance rate next to it.

The finisher itself is strict. Launched from the converged endpoint of the global-scatter rung, the full-batch Metropolis test rejects everything down the step ladder until $dt = 2 \times 10^{-6}$, which accepts at 1.00; the production run of 5,000 trajectories at $L = 30$ accepts at 0.98 and reproduces that chain's predictions at RMSE 0.0469. The stamp is on the location, not the spread, and it belongs to the rung it ran on: an accepted endpoint certifies where that chain is, not how well any chain's draws cover the posterior's width.

Table 2 puts every method on the same heteroscedastic target. The cleanest single summary is the maximum bin deviation from 1: the ray tracing chain 0.07, the 10-seed ensemble 0.21, the Adam MAP fit 0.26 (Figure 6). The z kurtosis separates the same groups: 15.9 for the MAP fit and 10.3 for tuned SGHMC against 8.1 for the Gaussian chain and 6.6 for the ensemble. Across its learning-rate sweep the MAP fit's calibration anti-correlates with its training loss: the deeper the optimizer sits in the mode, the worse the calibration. Only the sampled posterior distributes uncertainty correctly star by star.

3.5 SGHMC runs hot; the ray cannot

SGHMC [4] holds its temperature only as well as its friction is tuned, and on this target it never held it exactly. At friction 30 the kinetic temperature heats from 1.38 to 9.62 across the step-size grid. At friction 3000 it stays near 1 but is still 8 to 9% hot even hand-tuned, and its weighted norm was still marching at 2M steps, so it carries no convergence certificate at that budget (Figure 7). Ray tracing holds temperature by construction: the direction lives on the sphere, the speed is

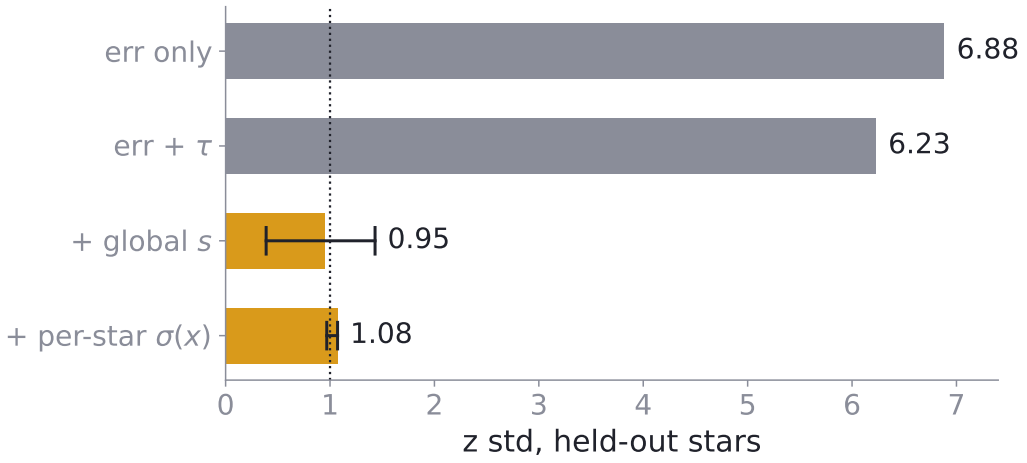


Figure 5 The noise-model ladder on 25,232 held-out stars. A calibrated model has z standard deviation 1 (dotted). Survey errors alone leave the posterior overconfident by a factor of about seven (6.88); member disagreement barely moves it (6.23); a global scatter s fixes the average (0.95); the per-star $\sigma(x)$ fixes it while staying flat across label-error bins (1.08). Read left to right: the sampler is fixed, the likelihood climbs.

Table 2 All methods on the same heteroscedastic likelihood, 25,232 held-out stars. Columns: RMSE, test NLL, the standard deviation of the standardized residuals, its value in each of the four label-error bins, and stars past $|z| = 4$. The Student-t row's bracketed SD(z) is the variance-based analog, and its tail count uses the fitted t's own tail probability.

method	RMSE	test NLL	SD(z)	z std by bin	$\lvert z\rvert > 4$
MAP (Adam)	0.0489	−1.733	1.180	1.26 / 1.24 / 1.13 / 1.04	209
deep ensemble (10 seeds)	0.0473	−1.933	0.962	0.79 / 0.91 / 0.97 / 1.05	61
SGHMC (tuned)	0.0475	−1.875	1.155	1.06 / 1.18 / 1.13 / 1.08	170
ray tracing, Gaussian	0.0494	−1.941	1.08	0.97 / 1.07 / 1.07 / 1.07	108
ray tracing, Student-t	0.0480	−1.973	(1.12)		14

fixed, and there is no noise-matched knob to set.

The heated arm is the warning. It posts RMSE 0.0474 despite boiling: point accuracy is no evidence of a correct posterior. Two caveats belong right here rather than in the limitations: the SGHMC budget was 2M steps against the ray tracing campaign's 16M, and this is one target.

3.6 The gain is in the error bars

Held-out stars were on average 27% more probable under the posterior average than under the single best fit. The provenance: over the 25,232 test stars the Student-t chain's marginal predictive (a logsumexp mixture over 50 members) scores a mean NLL of −1.9731 against −1.7332 for the MAP plug-in; the difference is 0.2399 nats, and $e^{0.2399} = 1.271$.

The gain decomposes cleanly. Marginalization alone is worth 0.208 nats (23%): the Gaussian chain's marginal already scores −1.9409. The Student-t tails add about 3% more. About two thirds of the stars improve individually, with a median improvement near 21%; this is not a few outliers rescuing an average.

RMSE does not move: 0.0489 for the MAP fit against 0.0494 for the chain, a tie. The gain is in the error bars, not the point predictions. What the MAP fit lacks is honest width: it is overconfident at z standard deviation 1.18, and the mixture is what gets the width right. Figure 8a

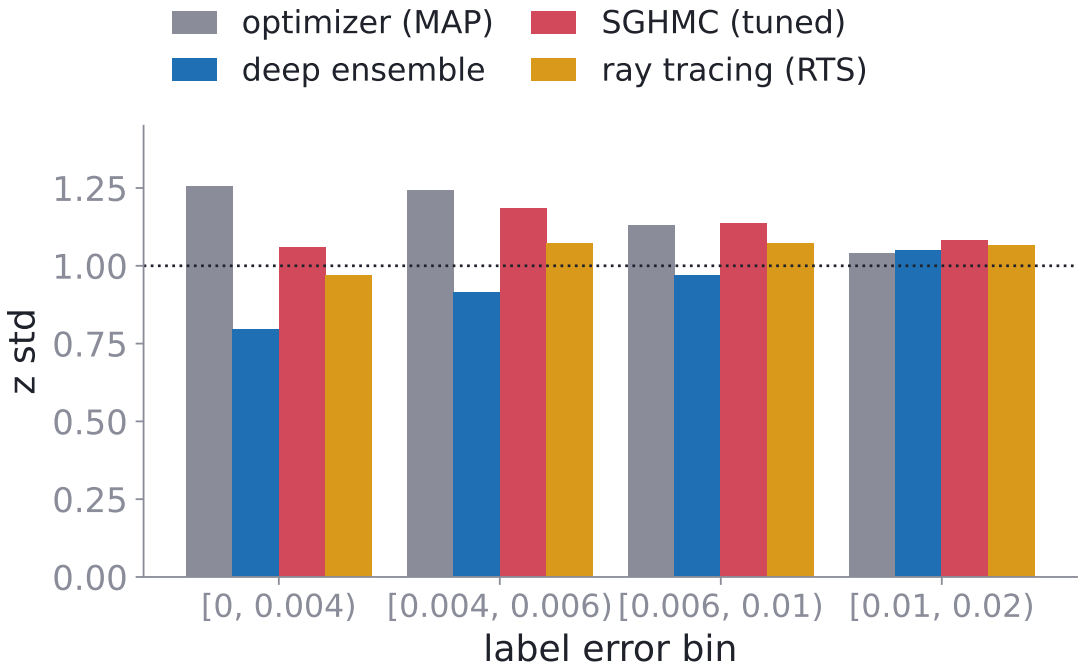


Figure 6 Per-bin calibration on the shared heteroscedastic likelihood. A calibrated method sits at 1 in every label-error bin. The MAP fit is an overconfident staircase (1.26 down to 1.04), the ensemble an inverted one (0.79 up to 1.05); the ray tracing chain is flat (0.97 to 1.07). Maximum deviation from 1: chain 0.07, ensemble 0.21, MAP 0.26.

shows the 50-member prediction cloud for a typical star and for a star the members disagree on; Figure 8b shows the resulting predictive densities for the same two stars.

3.7 Sanity checks

The chain's spread detects what it should. Its member-disagreement alarm scores AUC 0.932 on dwarf-star imposters and 0.998 on hot stars; on survey-flagged stars it scores 0.588, near chance, and that failure is a label-side problem no spectrum-based alarm can see. One framing rule matters here: out-of-distribution detection separates methods with genuine member spread from methods whose spread has collapsed, and it does not rank ray tracing against SGHMC. The contribution of ray tracing to this panel is the certificate that its spread is genuine.

The second check is physics. At fixed effective temperature and metallicity, high-alpha giants are 20 to 30% brighter below 450 nm, and the network's calibrated saliency concentrates in the same blue region of the spectrum. The network is using the part of the spectrum where the physics says the signal is.

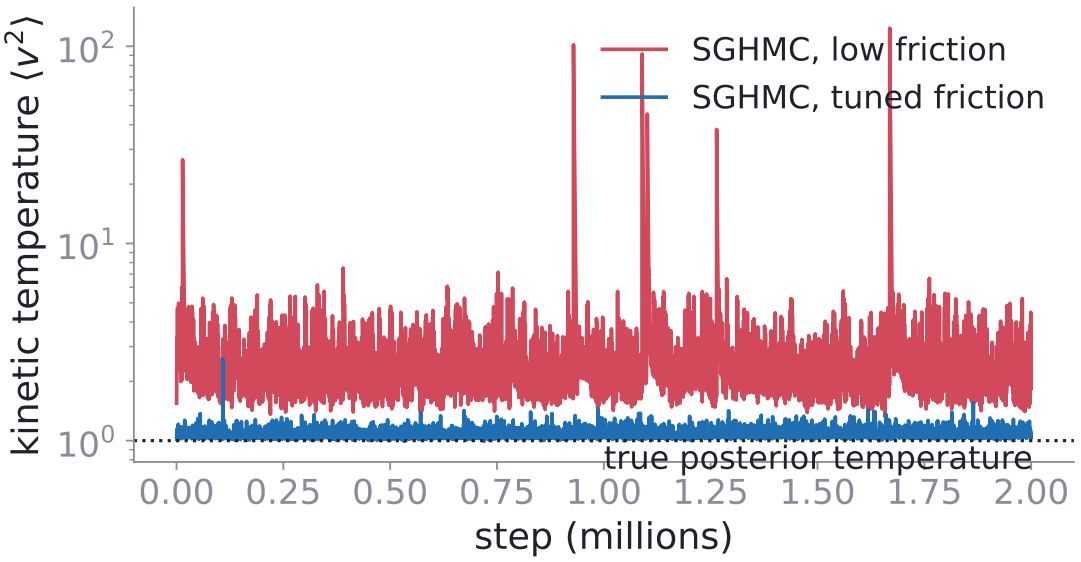
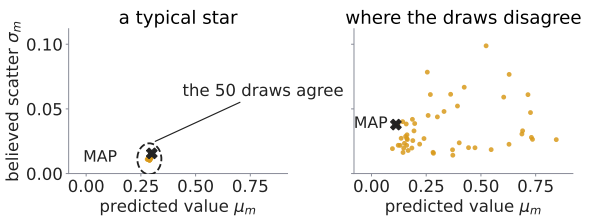
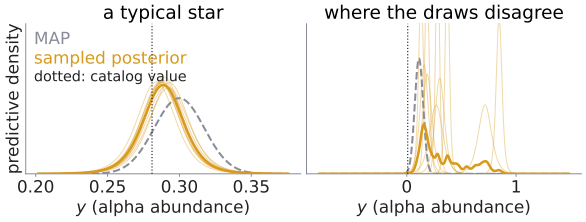


Figure 7 Kinetic temperature along the run for the two SGHMC frictions; the true posterior temperature is the dotted line at 1. The low-friction arm boils, with excursions orders of magnitude above the target; the tuned arm holds near 1 yet sits 8 to 9% hot throughout. Ray tracing has no version of this plot: its speed is constant by construction.



(a) The 50-draw cloud: a typical star and a disagreement star.



(b) Predictive densities for the same two stars.

Figure 8 The posterior average against the single best fit on two held-out stars. The mixture widens exactly where the members disagree; a plug-in density cannot.

4 Why the tolerance differs: the geometry of the auxiliary variable

The measurements above say that the two samplers fail at very different noise levels and with different step-size laws; they do not say why. The explanation is the subject of a companion manuscript by my mentor [7], written in August 2026 alongside this project. The derivations are his, and I summarize them here without proofs. What this section reports is the part that belongs to the summer: the theory made quantitative predictions, and the measurements were already in hand to test them.

Two exposures to one noise source. Minibatching supplies a noisy estimate of the log-density, and a sampler can be exposed to that noise in two places. Differentiating the noisy log-density gives score noise, which perturbs the dynamics of an unadjusted sampler; differencing it between two states gives acceptance noise, which perturbs a Metropolis test. Acceptance noise admits an exact penalty correction [14, 15], but the cost of that correction grows exponentially in the noise level. Score noise has no correction and costs nothing extra. An unadjusted sampler takes the cheap exposure and pays for it with a bias in its stationary distribution; the analysis computes

that bias.

One identity for the stationary bias. For a sampler that alternates deterministic transport with a refresh, the stationary bias is injection over relaxation: $b = J/(1-\lambda)$, the noise injected per step divided by the fraction the refresh removes. The identity recovers the classical results. Unadjusted Langevin gives $b = \sigma^2 h/2$ to leading order, the known SGLD bias. Unadjusted HMC gives $b = c(S)\,\sigma^2 h$ with $c(S) = (2S - \sin 2S)/(4\sin^2 S)$, which diverges at the half-orbit $S = \pi$: the refresh stops contracting exactly where the noise keeps arriving. Minibatch HMC does not just carry a bias, it carries a resonance.

The ray tracing result. For ray tracing the same identity gives $b = \sigma^2 h^2/[12(D-1)]$: independent of trajectory length, one order higher in the step size than HMC, and with no resonance at any length, because the noise forces exactly the mode that the refresh relaxes, so injection and relaxation vanish together where the resonance would sit. The mechanism in words: the direction lives on a compact sphere whose equilibrium law is uniform, and isotropic noise acts as a random rotation, which leaves a uniform law invariant, so the leading effect of the noise cancels. What survives is the failure of the noise to commute with the deterministic bending inside a single step, and that is second order in h.

Anisotropy, the caveat that matches the measurement. Real minibatch noise is not isotropic, and the h^2 law is the isotropic result. With noise covariance C, the sampled covariance picks up a distortion of $\sigma^2 h S(C - I)/(3D)$: first order in h, so a smaller step cannot refine it away; growing with trajectory length; and traceless, so mean-squared-radius diagnostics never see it. This is the term behind the measurement in Section 3.2, where ray tracing's local step-size exponent weakened from about 1 toward 0.6 once idealized isotropic noise was replaced by real minibatch noise. The practical advice follows directly: estimate the noise covariance and whiten it, and the h^2 law is bought back.

Confrontation with measurement. The predicted step-size exponents ($\sigma_c \propto h^{-1/2}$ for HMC, $\sigma_c \propto h^{-1}$ for ray tracing) match both the slopes measured here under idealized noise (Section 3.2) and the slopes in Behroozi's $D = 10{,}000$ experiment [5]. The predicted absolute tolerances, with no fitted quantities, land within 30% of that experiment: theory $\sigma_c^2 = 2.67$ against a reported 3.2 for ray tracing; 0.0085 against 0.012 for HMC; a predicted variance ratio of 314 against the reported ratio above 250. For a first-principles calculation with nothing tuned, that is close.

Batch economics. Since $\sigma^2 \propto 1/B$, holding accuracy fixed forces $h \propto B^{-1/2}$ for HMC, and total compute is then flat in B: shrinking the batch buys nothing. Ray tracing needs only $h \propto B^{-1/4}$, so total compute scales as $\sqrt{B}$, and a hundredfold smaller batch costs ten times less. This is the theoretical version of the cost frontier measured in Section 3.2.

Two design warnings. The analysis also warns against two plausible modifications. Splitting the noise out of the kick, applied instead as an independent rotation of matched variance, triples the bias; and at matched gradient cost the minimal-norm integrator carries about 20% more noise bias than plain leapfrog, despite its smaller clean discretization error. Matching the variance of a noise source is not the same as matching its effect. On the constructive side, the analysis supports Behroozi's conjecture that a constant-speed variant of HMC would behave like ray tracing: the proofs use only the compactness of the auxiliary domain and the rotational invariance of its equilibrium law, and nothing in them refers to the optical construction.

What exact adjustment changes. The second companion manuscript, Speagle [16], asks what the advantage is worth once the Metropolis test itself is corrected for minibatch noise, and the answer

sharpens every claim in this report. With an exact minibatch adjustment in place, ray tracing and HMC tolerate the same noise variance to within 0.05% at matched decorrelation and gradient budget: the factor above 250 is a property of *unadjusted* sampling, not of the dynamics as such. What the adjustment removes differs sharply between the two samplers. At the operating point, ray tracing's unadjusted bias is smaller than HMC's by a factor that grows with trajectory length, about $6T$, so adjustment takes almost nothing away from ray tracing; per squared step size the residual bias is $1/[12(D-1)T]$ for ray tracing against $1/(2D)$ for HMC. A lower bound on the product of batch size and trajectory length, $BT \geq c\,\Lambda^2/s_*^2$, adds the economic half: minibatching and exact adjustment are alternatives, not complements, since paying for one removes the reason to pay for the other. The honest summary of the whole report follows. The large unadjusted advantage is real, but it is an advantage of unadjusted sampling; adjusted, the two samplers reach tolerance parity; ray tracing is the sampler for which skipping the adjustment is nearly free.

5 Discussion

Four things came out of the summer that are worth stating separately from the numbers.

The geometry argument holds, and it is now quantitative. At the start of the project the claim was a picture: a bounded direction cannot heat the way an unbounded momentum can. That picture is now attached to measured exponents (h^{-1} against $h^{-1/2}$), to a closed-form bias ($\sigma^2 h^2/[12(D-1)]$), and to agreement between the two within 30% on an experiment neither was fitted to. It also held up on a real problem: the Gaia chains converged under minibatch gradients at $D = 11{,}394$, and an exact full-batch test accepted the global-scatter rung's converged endpoint at 0.98.

The sampler was rarely the hard part. On the Gaia campaign the largest single improvement, by a wide margin, came from fixing the noise model, not from anything about sampling: z standard deviation 6.88 to 1.08. On the MNIST networks the corresponding cost was where the chains started, because a bad start spends the budget on burn-in. Reading drift diagnostics correctly was its own discipline: no drift rule was read without the acceptance rate next to it. A sampler with better noise tolerance removes one obstacle. It does not remove the modelling work, and on this problem the modelling work was larger.

The payoff shows up in the error bars, not the predictions. Across five methods on the same Gaia target, RMSE spans 0.0473 to 0.0494, which is close to a tie. Calibration spans a maximum bin deviation of 0.07 (the chain) to 0.26 (the MAP fit), which is not. The 27% mean improvement in held-out predictive density is almost entirely marginalization: 23 of the 27 points come from averaging over the posterior, and only about 3 from the heavier tails. This is the practical case for paying for a sampler, and it is also the reason accuracy comparisons are a poor way to evaluate one. The SGHMC run that was visibly boiling still reached RMSE 0.0474.

On comparison claims. Two comparisons that would have been convenient to make are not supported by the evidence collected here. The out-of-distribution results do not separate ray tracing from SGHMC in any statistically meaningful way; what they separate is genuine spread from collapsed spread, and both samplers have genuine spread. And the speed comparison is weaker than the noise-tolerance comparison: at matched accuracy the gate-passing saving was about 1.5× at batch 256 on one problem. The published variance ratio above 250 at $D = 10{,}000$ [5] is a noise-tolerance number, not a wall-clock speedup, and it belongs to that paper rather than to this work.

6 Limitations and future work

Not every claim in this report is equally settled. This section sorts them into validated, partial, and open.

Validated. Robustness to stochastic-gradient noise. This has a mechanism (bounded auxiliary variable on a compact sphere), measured scalings on both idealized and real noise, an independent theoretical derivation [7] that agrees with the tolerances published by Behroozi [5] to within 30%, and a working application on a real survey dataset at $D = 11{,}394$, whose global-scatter rung passed an exact full-batch check.

Partial. The cost story. The honest statement is that at matched accuracy the gate-passing saving was about 1.5× at batch 256, on one problem. The measured cost frontier used an assumed full-batch reference rather than a measured one, and the theory's prediction of $O(\sqrt{B})$ economics has not been tested end to end on a real target. Any larger speedup figure from these runs comes from configurations that fail their own accuracy gate and should not be quoted.

Open.

- Width honesty: the Student-t chain still shows about 4 to 5% PIT under-coverage. The exact finisher certifies where the chain sits, not how wide it is, so this is unresolved by the diagnostics available here.
- Anisotropy: the theory says the surviving first-order bias term is traceless and that whitening the noise covariance should recover the h^2 law. Neither the whitening nor a direct measurement of that traceless distortion was attempted here. This is the most concrete next experiment.
- Every production result comes from one chain per configuration: there is no multi-chain $\hat{R}$, so the convergence evidence is drift-based and within-chain.
- Scale: $D = 11{,}394$ is a real network but far below the scale where Bayesian deep learning is usually argued about [17, 18]. Whether the h^2 law survives at that scale with real anisotropic noise is untested.
- Mode coverage: nothing in this project validates behaviour on multimodal or disconnected targets. The published method has a separate emission rule for that case [5]; it was not exercised here and no claim is made about it.
- Learning the weighting function: the sampler's family is parameterized by a weighting function W that recovers HMC, microcanonical HMC, Gibbs, and tempering as special cases, and fair sampling holds for any choice. Learning W for a given target, suggested by Prof. Ricardo Baptista, is the natural research direction out of this work and was not attempted.

One disclosure. The rung-1 production gate ran roughly 49× too soft, because σ_{sto} had been measured at a warm start that a bug had scrambled. This was caught on July 30, 2026. The validity of the results reported here rests on the exact full-batch finisher and the drift checks, which are unaffected, and the chain of record is the rung-2 heteroscedastic chain rather than rung 1.

7 Conclusion

The question this project set out to answer was whether replacing an unbounded momentum with a bounded direction makes minibatch MCMC practical. On the evidence collected here the answer is yes for the noise problem specifically, and the size of the effect is now known rather than argued.

Three results carry that conclusion. Under isotropic gradient noise the critical noise level scales as h^{-1} for ray tracing against $h^{-1/2}$ for HMC, measured directly. The companion analysis [7] derives a stationary bias of $\sigma^2 h^2/[12(D-1)]$, one order higher in h, with no resonance at any trajectory length, and it reproduces the tolerances that Behroozi [5] reports at $D = 10{,}000$ to within 30% with nothing fitted. And on a real target, 126,156 Gaia giants and an 11,394 parameter network, minibatch chains converged, an exact full-batch test accepted the global-scatter rung's endpoint, and the chain of record stayed flat in calibration across every label-error bin where MAP, a deep ensemble, and SGHMC are all visibly off; with a Student-t noise model, held-out stars were on average 27% more probable under the posterior average than under the single best fit.

The rest of the summer was a reminder of where the difficulty actually sits. Fixing the noise model took the held-out z standard deviation from 6.88 to 1.08; on the fixed target, going from the best fit to the sampled posterior moved it only from 1.18 to 1.08. Diagnostics had to be read in pairs, because a drift series without an acceptance rate beside it says nothing. And point accuracy turned out to be almost useless as a signal: the SGHMC arm that was demonstrably overheating still reached RMSE 0.0474, in the same range as the converged methods.

What remains open is the width of the posterior, the effect of anisotropic noise in practice, and whether any of this holds at the scale of a modern network. The most direct next step is the one the theory points at: estimate the minibatch noise covariance, whiten it, and check whether the second-order law comes back. Every number in this report traces to the campaign records in the repository linked on the cover.

Acknowledgements

I thank Prof. Joshua Speagle for mentoring this project, for the MAP-transit suggestion that made the MNIST chains affordable, and for sharing the two companion manuscripts that this report's theory section draws on [7, 16]. I thank Prof. Ricardo Baptista for the questions that shaped how the method was framed, and for the weighting-function direction. The SUDS poster presenting part of this work was joint work with Chuxuan Ai. This project was carried out in the KAUST Academy Summer School 2026, hosted at the University of Toronto Data Sciences Institute, and I thank both programs and their sponsors for the opportunity and the compute.

References

[1] Radford M. Neal. "MCMC using Hamiltonian dynamics". In: *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011.

[2] Michael Betancourt. "The fundamental incompatibility of scalable Hamiltonian Monte Carlo and naive data subsampling". In: *Proceedings of the 32nd International Conference on Machine Learning*. 2015.

[3] Max Welling and Yee Whye Teh. "Bayesian learning via stochastic gradient Langevin dynamics". In: *Proceedings of the 28th International Conference on Machine Learning*. 2011.

[4] Tianqi Chen, Emily B. Fox, and Carlos Guestrin. "Stochastic gradient Hamiltonian Monte Carlo". In: *Proceedings of the 31st International Conference on Machine Learning*. 2014.

[5] Peter Behroozi. "The Ray Tracing Sampler". In: *arXiv preprint arXiv:2510.25824* (2025).

[6] Alexander Laroche and Joshua S. Speagle. "Closing the stellar labels gap: stellar-label-independent evidence for [α/M] information in Gaia BP/RP spectra". In: *The Astrophysical Journal* 979 (2025), p. 5.

[7] Joshua S. Speagle. "The geometry of the auxiliary variable controls stochastic-gradient sampling bias". Manuscript in preparation, University of Toronto. 2026.

[8] Jakob Robnik et al. "Microcanonical Hamiltonian Monte Carlo". In: *Journal of Machine Learning Research* 24 (2023).

[9] Jakob Robnik and Uroš Seljak. "Fluctuation without dissipation: microcanonical Langevin Monte Carlo". In: *arXiv preprint arXiv:2303.18221* (2024).

[10] E. Sommer et al. "Can microcanonical Langevin dynamics leverage mini-batch gradient noise?" In: *Proceedings of the 43rd International Conference on Machine Learning*. 2026.

[11] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Advances in Neural Information Processing Systems*. 2017.

[12] Gaia Collaboration, A. Vallenari, A. G. A. Brown, et al. "Gaia Data Release 3: summary of the content and survey properties". In: *Astronomy & Astrophysics* 674 (2023), A1.

[13] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feed-forward neural networks". In: *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*. 2010.

[14] D. M. Ceperley and M. Dewing. "The penalty method for random walks with uncertain energies". In: *The Journal of Chemical Physics* 110 (1999), pp. 9812–9820.

[15] Geoff K. Nicholls, Colin Fox, and Alexis Muir Watt. "Coupled MCMC with a randomized acceptance probability". In: *arXiv preprint arXiv:1205.6857* (2012).

[16] Joshua S. Speagle. "Exact Mini-Batch Sampling on the Sphere, and What It Costs". Manuscript in preparation, University of Toronto. 2026.

[17] Pavel Izmailov et al. "What are Bayesian neural network posteriors really like?" In: *Proceedings of the 38th International Conference on Machine Learning*. 2021.

[18] Florian Wenzel et al. "How good is the Bayes posterior in deep neural networks really?" In: *Proceedings of the 37th International Conference on Machine Learning*. 2020.