QUENCHED ENSEMBLE SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Some of the sharpest challenges in sampling from the energy functions of physical systems arise at phase transitions, where the density of states changes abruptly and many sampling algorithms stall. Nested sampling is a particle method that traverses the density of states under a hard energy constraint and is known to be robust to such transitions, but its application in high dimension is limited by the difficulty of sampling under that constraint. In this work we introduce Quenched Ensemble Sampling, which generalises the hard constraint to a family of repulsive potentials at the energy boundary. This preserves the quenched path of monotonically decreasing energy while making the constrained target amenable to scalable gradient-based kernels. We demonstrate on synthetic models of phase transitions that our method estimates the marginal likelihood and draws posterior samples across a first-order transition where popular alternatives such as tempering fail. We apply the procedure to marginal likelihood estimation in Bayesian neural networks, enabling model comparison between network architectures. Finally, in a high-dimensional continuous lattice field theory, we show that this method traverses a first-order transition and estimates the partition function.

1 INTRODUCTION

Estimating the marginal likelihood is a central problem in Bayesian inference, and is particularly challenging for sampling algorithms. The marginal likelihood is defined as the integral, $Z = \int \mathcal{L}(x)\,\pi(x)\,\mathrm{d}x$, of a likelihood $\mathcal{L}$ over a prior π on parameters x. In Bayesian inference this is the normalising constant of the posterior, and is central to Bayesian model comparison (Llorente et al., 2023). Estimation of this integral is typically performed by introducing a series of interpolating distributions between the prior and posterior, where the integral can then be approximated by a series of ratios of normalising constants along this path. The most common approach is to use a temperature ladder, where the interpolating distributions are proportional to $\mathcal{L}^{\beta_k}(x)\,\pi(x)$ for a discrete schedule of inverse temperatures $0 = \beta_0 < \beta_1 < \cdots < \beta_K = 1$. This ladder is traversed sequentially in annealed importance sampling (AIS) (Neal, 2001) and sequential Monte Carlo (SMC) (Doucet et al., 2001), or held concurrently as exchanging replicas in parallel tempering (Swendsen & Wang, 1986). This makes the integral numerically tractable, and brings additional benefits for exploring the posterior landscape: at high temperatures the interpolating distributions flatten the energy barriers between modes, allowing transitions that are inaccessible at $\beta = 1$ (Syed et al., 2022).

Complementary to these approaches, it is possible to formulate a *vertical* path through the density of states, rather than a *horizontal* path through temperature (Polson & Scott, 2015). This is the basis of nested sampling (Skilling, 2006), which has been applied to a wide range of problems in physics (Pártay et al., 2010; Ashton et al., 2022). This approach writes the marginal likelihood as an integral over the prior volume at monotonically increasing values of the likelihood; we review this construction in detail in Section 2. The path of increasing likelihood is notably robust to first-order phase transitions, allowing the path of distributions to traverse regions of parameter space that hold a vanishing fraction of the prior mass, and which are therefore inaccessible to tempering methods. However, its practical implementation relies on sampling under a hard constraint, which degrades the efficiency of gradient-based MCMC kernels (Kroupa et al., 2025).

In this work we introduce *Quenched Ensemble Sampling*, a method that preserves the attractive features of the path of increasing likelihood level sets, while incorporating a family of repulsive potentials at the hard constraint that allow the particles to efficiently move in high-dimensional spaces

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

with standard gradient-based MCMC kernels (Fearnhead et al., 2025). This path follows a trajectory of monotonically decreasing energy in an extended state space, which when implemented with an ensemble of particles, allows for a Sequential Monte Carlo sampler to be constructed that is robust to first order phase transitions. We formulate this procedure using motivations from statistical physics, highlighting the importance of this quenched path in the context of sampling from Boltzmann distributions, and demonstrate its performance on a range of problems, including synthetic models of phase transitions, a high-dimensional lattice field theory problem, and marginal likelihood estimation in Bayesian neural networks.

The key contributions of this work are:

- We formulate an adaptive SMC sampler over a quenched family of distributions, and show that a family of repulsive potentials at the energy level defines a soft microcanonical ensemble, admitting gradient-based MCMC mutation kernels.
- We show that the softness parameter interpolates between classic nested sampling and tempering, generalising the hard constraint to a distribution of effective temperatures (Appendix B).
- We provide a practical implementation with a well-tuned ensemble-preconditioned MALA kernel (Roberts & Tweedie, 1996), and contrast the quenched and tempered paths directly on high-dimensional problems inaccessible to standard nested sampling.

2 Background

To motivate the quenched path, we review the formulation of Bayesian inference in the language of statistical mechanics, and the connection between the marginal likelihood and the partition function. Connecting the statistical mechanics and Bayesian inference perspectives, we write the canonical (Boltzmann) density—in Bayesian terms, the tempered posterior—at inverse temperature $\beta = 1/T$ as

$$\rho_\beta(x) = \frac{1}{Z(\beta)}\,\mathcal{L}^\beta(x)\,\pi(x), \qquad Z(\beta) = \int \mathcal{L}^\beta(x)\,\pi(x)\,\mathrm{d}x\,, \tag{1}$$

where the Boltzmann factor $e^{-\beta U(x)} = \mathcal{L}^\beta(x)$ is the tempered likelihood, with energy $U(x) = -\log \mathcal{L}(x)$, and the prior π plays the role of the phase-space measure (Table 5). Samples from $\rho_\beta(x)$ constitute the canonical ensemble, and the normalising constant $Z(\beta)$ is the partition function, a central quantity in statistical mechanics; at $\beta = 1$ it is the marginal likelihood, whose estimation underpins Bayesian model comparison. In a statistical mechanics context, this integral is often written in terms of the density of states,

$$g(E) = \int \delta\big(E - U(x)\big)\,\pi(x)\,\mathrm{d}x\,, \tag{2}$$

the prior mass of configurations (microstates) per unit energy at E—the analogue of the density of microstates, with the prior as the a priori measure—giving

$$Z(\beta) = \int \mathcal{L}^\beta(x)\,\pi(x)\,\mathrm{d}x = \int g(E)\,e^{-\beta E}\,\mathrm{d}E\,, \tag{3}$$

which expresses the partition function as a Laplace transform of the density of states: an integral on the energy axis. The insight of nested sampling (Skilling, 2006) lies in working with the *cumulative density of states*, $G(E) = \int_{-\infty}^{E} g(u)\,\mathrm{d}u$—a prior probability that can be estimated directly by sampling, where g itself cannot. Substituting $g(E)\,\mathrm{d}E = \mathrm{d}G$, with $E(G)$ the generalised inverse of $G(E)$, gives

$$Z(\beta) = \int e^{-\beta E}\,\mathrm{d}G(E) = \int_0^1 e^{-\beta E(G)}\,\mathrm{d}G\,, \tag{4}$$

the vertical representation of the partition function: all dependence on the D-dimensional parameter space is absorbed into the one-dimensional monotone curve $E(G)$, and estimation reduces to resolving this curve at exponentially small values of G.

2.1 NESTED SAMPLING AND ITS LIMITATIONS

The nested sampling procedure was proposed by Skilling (2006) as a generic particle method whose central computational requirement is efficient sampling from the constrained prior $\pi^*(x) = \pi(x)\,\mathbf{1}\{U(x) < E\}$. In practice this is met by rejection sampling in low dimensions (Feroz et al., 2009), slice sampling as a scalable black-box MCMC mutation kernel (Neal, 2003; Handley et al., 2015; Yallup, 2026), with some problem-specific constructions giving enhanced scalability (McEwen et al., 2023; Murray et al., 2005; Yallup et al., 2026). However, as a general-purpose method, it has seen limited application beyond $D \sim 10^2$, where slice sampling becomes computationally expensive. This limitation arises from the hard constraint: away from its boundary, the constrained target $\pi^*(x)$ inherits the score $\nabla \log \pi^*(x)$ from the prior, which typically carries no useful information about the energy landscape, and kernels engineered for the constraint have not meaningfully changed the scaling of the method (Appendix B.1). Instead of trying to construct an MCMC kernel that can efficiently sample under a hard constraint, we take a different approach and ask if we can soften the hard constraint whilst preserving the properties of the path of monotonically decreasing energy.

The work of Salomone et al. (2025) has addressed the consistency and bias of normalising constant estimation by framing nested sampling as a Sequential Monte Carlo method. We follow the volume accounting from this work, using incremental importance weights to measure the volume ratio between successive levels rather than assigning it from the order-statistic law of the classic nested sampling algorithm (Chopin & Robert, 2010). For a fixed schedule, this affords the unbiasedness results central to SMC (Del Moral, 2004). We nevertheless retain and recommend an inherently adaptive level schedule. This is both natural for the quenched path with its moving support and immensely practical for traversing phase transitions. As in adaptive SMC more generally, this introduces a finite-particle bias (Beskos et al., 2016). We separate the measured-volume accounting from the effect of adaptation in Appendices C.1 and C.2.

2.2 SCALABLE MCMC KERNELS

Setting aside distributions with hard constraints, there has been substantial progress in the robust implementation of scalable Markov chain Monte Carlo (MCMC) methods (Fearnhead et al., 2025). By incorporating the gradient of the energy function, $\nabla U(x)$, these methods suppress the random-walk behaviour of vanilla MCMC and improve its dimension scaling (Appendix B.1), and incorporating such kernels into tempered SMC has enabled applications to problems with thousands of dimensions (Buchholz et al., 2021). These developments do not transfer cleanly to hard-constrained targets; softening the constraint makes them available to the quenched path. In this work, we use the Metropolis-adjusted Langevin algorithm (MALA) as a proof of principle and develop a recipe for a well-tuned kernel, using the particle ensemble to precondition the proposal and permit efficient application across a variety of problems without hand-tuning. We apply the same kernel within adaptive tempered SMC as a control, providing the key comparison that isolates the effect of the quenched path from that of the MCMC kernel itself.

MALA uses the score of the target to add a drift to the random-walk proposal, $x' = x + \frac{\epsilon^2}{2} M^{-1} \nabla \log \rho(x) + \epsilon M^{-1/2} \eta$ with $\eta \sim \mathcal{N}(0, I)$, accepted with the standard Metropolis–Hastings correction. Here ϵ is the step size and M a positive-definite mass matrix, so the proposal covariance is $\epsilon^2 M^{-1}$; the step size is tuned towards the optimal acceptance rate, known to be around 0.574 for MALA in high dimensions (Roberts & Rosenthal, 1998).

3 METHOD

Quenching a configuration in molecular simulation commonly refers to rapidly relaxing it towards an energy minimum, typically performed with fast gradient descent methods (Thompson et al., 2022). We present Quenched Ensemble Sampling (QES), a Sequential Monte Carlo sampler that generalises this idea to a population of particles. The population is evolved down a ladder of decreasing energy levels by a combination of reweighting, resampling, and mutation, and the partition function is obtained from the measured level volumes. The resulting sequence of distributions that the particles follow simulates a series of thermodynamic quenches.

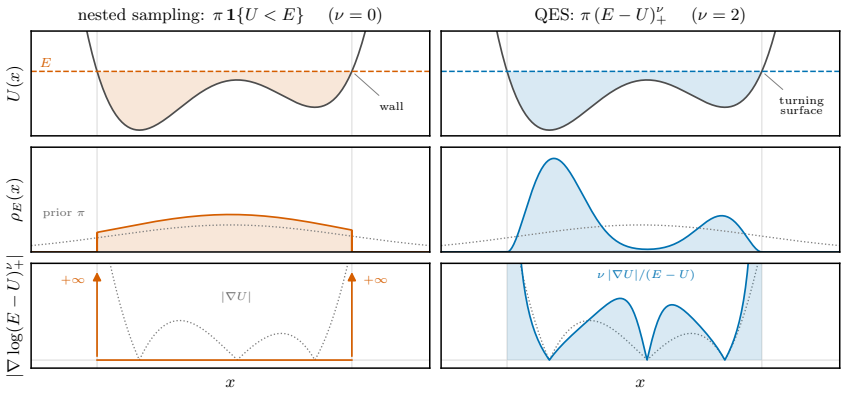


Figure 1: QES against nested sampling on a one-dimensional asymmetric double well (unit Gaussian prior): the energy landscape with the level E as a waterline (top), the profile of ρ_E (middle), and its score (bottom). For $\nu > 0$ the score is a smooth informative gradient throughout the level set, vanishing only at stationary points of U; dotted, $|\nabla U|$.

3.1 The softened level set

At a given energy level E, we define a target

$$\rho_E(x) \;\propto\; \pi(x)\,\big(E - U(x)\big)_+^{\nu}, \tag{5}$$

where $\nu > -1$ is a fixed parameter, and $(\cdot)_+ = \max(\cdot, 0)$. The soft microcanonical factor $(E-U)_+^\nu$ is hence a repulsive potential at the level, so that ρ_E vanishes continuously at the boundary of its support $\{U < E\}$. Setting $\nu = 0$ recovers the hard constrained prior at energy E, which is the target of nested sampling, and ν can be set arbitrarily large to increase the repulsive force from the boundary. Pragmatically we use $\nu = 2$ for the mobility and efficiency reasons detailed in Appendix B, and find this choice effective for most applications. The support of ρ_E is still confined to the level set $\{U < E\}$, but its boundary is softened such that the score is defined throughout so can inform MCMC dynamics (Figure 1).

This family is the conditional of an exact augmentation at fixed temperature. Fix $\nu > -1$ and consider the joint density on (x, E)

$$p(x, E) \;\propto\; \pi(x)\,(E - U(x))_+^{\nu}\, e^{-E}. \tag{6}$$

Integrating out E with the substitution $s = E - U(x)$ gives $\int_0^\infty s^\nu e^{-(s+U)}\,\mathrm{d}s = \Gamma(\nu+1)\,e^{-U(x)}$, so the x-marginal of Eq. (6) is exactly the posterior, with $\Gamma(\nu+1)$ the only constant. Its conditionals are

$$x \mid E \;\sim\; \rho_E, \qquad E \mid x \;\sim\; U(x) + \mathrm{Gamma}(\nu+1, 1). \tag{7}$$

Where E is styled as an auxiliary slice variable. The E conditional also informs the meaning of ν, $\nu + 1$ is the mean energy headroom, in nats, between the level and a particle's own energy. As this is a scale defined in energy (or log-likelihood) space, we see that picking a fixed ν is justifiable, and this headroom need not necessarily scale with dimension.

Closely related power-law targets arise in microcanonical Monte Carlo (Ray, 1991) and in nested sampling "demon" and total-enthalpy constructions (Habeck, 2015; Baldock et al., 2017), which tie ν to the degrees of freedom of the augmented system. QES instead fixes ν independently of both problem and dimension, measures successive level-volume ratios by SMC importance weights rather than nested-sampling order statistics, and admits any mutation kernel that can consume the differentiable log density of Eq. (5).

3.2 Partition function from level volumes

The cumulative density of states defined by Eq. (2) gives the level volume $G(E) = \int_{-\infty}^{E} g(u)\,\mathrm{d}u$. This extends naturally to the soft microcanonical family of Eq. (5), whose level volume is

$$G_\nu(E) \;=\; \mathbb{E}_\pi\big[(E-U)_+^\nu\big] \;=\; \int_{-\infty}^{E} g(u)\,(E-u)^\nu\,\mathrm{d}u. \tag{8}$$

The same Gamma identity used above expresses the partition function as the one-dimensional integral

$$Z \;=\; \frac{1}{\Gamma(\nu+1)}\int_{-\infty}^{\infty} G_\nu(E)\,e^{-E}\,\mathrm{d}E, \tag{9}$$

valid for every $\nu > -1$.

Consecutive levels $E' < E$ are related by the ratio

$$r(x) \;=\; \left(\frac{(E'-U(x))_+}{(E-U(x))_+}\right)^\nu, \qquad \frac{G_\nu(E')}{G_\nu(E)} \;=\; \mathbb{E}_{\rho_E}\big[r\big]. \tag{10}$$

Given a weighted particle population $\{(x_i, w_i)\}_{i=1}^N$ targeting ρ_E, with $\sum_i w_i = 1$, the incremental ratio is estimated before mutation by

$$\widehat{r}_{E\to E'} \;=\; \sum_{i=1}^{N} w_i\, r(x_i), \qquad \widehat{G}_\nu(E') \;=\; \widehat{G}_\nu(E)\,\widehat{r}_{E\to E'}. \tag{11}$$

The same reweighting moves the population to the next target, since

$$\rho_E(x)\,r(x) \;=\; \frac{G_\nu(E')}{G_\nu(E)}\,\rho_{E'}(x). \tag{12}$$

After a resampling or branching step, an MCMC kernel invariant for $\rho_{E'}$ is applied to rejuvenate the population and move duplicated particles apart. Repeating these steps down the level ladder estimates G_ν through a telescoping product of importance-weight averages. This is the standard SMC normalising constant estimator (Doucet et al., 2001), applied here to softened level-set volumes as the path to the marginal likelihood.

3.3 Algorithm

With the definitions of the previous sections, we can now compose the complete algorithm. We assume that the reference distribution π is easy to sample and that the energy U is differentiable. The algorithm is initialised by drawing an anchor population from π and choosing E_0 by bisection until the importance weights $(E_0 - U)_+^\nu$ hold an effective sample size (ESS) of N; their empirical mean initialises $G_\nu(E_0)$, and their normalised values are resampled into the N-particle working population (Appendix A.1).

Given N particles approximately distributed as ρ_{E_k}, Algorithm 1 transitions to the next level $E_{k+1} < E_k$ in three steps: dissect the population to choose E_{k+1}, by bisection until the prospective reweighted population reaches a prescribed ESS; reweight and branch for the new level; and mutate with an MCMC kernel K_E whose construction is unrestricted provided it leaves $\rho_{E_{k+1}}$ invariant. The process repeats until the active particles' contribution to the marginal likelihood is negligible against the accumulated estimate, at which point Z is computed by quadrature on the ladder of levels.

Choice of mutation kernel. The estimator requires an invariant mutation kernel, where there are a vast array of choices. We employ Metropolis-adjusted Langevin dynamics as detailed in Section 2.2, with a kernel pre-tuning recipe defined in this section. This proves to be a relatively robust choice, and the automated tuning we implement is effective for a variety of problems covered in Section 4, this also defines a strong baseline for tempered SMC which we provide as well in the code implementation. Following the pretuning pattern of Buchholz et al. (2021), we use the accumulated gradient information from the previous level to construct a diagonal preconditioner. Writing this score

$s^{(i)} = \nabla \log \rho\big(x^{(i)}\big)$ for the score at particle i of the ensemble, we set $M^{-1} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_D^2)$ with

$$\sigma_j^{-1} \;=\; \underset{i}{\mathrm{median}} \big\| s^{(i)} \big\| \left\langle \frac{s_j^2}{\|s\|^2} \right\rangle^{1/2}, \tag{13}$$

where $\langle\cdot\rangle$ averages over mutation steps and the weighted ensemble, and the scale is the ensemble median of each walker's step-averaged norm. This is a natural choice: the anisotropy $\langle s_j^2/\|s\|^2\rangle$ is the direction normalised Fisher information metric of the target, frozen across the mutation chain (Girolami & Calderhead, 2011). Rather than using the mean of the score norm as a scale we use the median since the additional boundary proximity term $(E-U)^{-1}$ that appears in this particular score can diverge for particles near the boundary, and the median is robust to this. The step size ϵ is adapted across levels by constant-gain stochastic approximation (Robbins & Monro, 1951), driving the acceptance rate observed at the previous level towards 0.574, the gain is held constant rather than decayed, since E moves at every level and the optimal step size drifts with it.

Two properties of Eq. (13) are important for embedding in a particle method. First, the metric is read from the score rather than from the spread of the particles. The particle cloud covariance, the nested sampling default (Yallup et al., 2026), estimates the extent of the ensemble. On a multimodal target that extent is the separation between modes rather than the width of any one of them, so the metric degrades on the type of problem that particle methods are commonly invoked on. Using the averaged local score of Eq. (13) composes only the local property of the target, so is well defined even on massively multimodal targets such as Bayesian neural networks. Second, neither M nor ϵ changes while the mutation kernel is applied to the level set. Both are constants of a level, set from the ensemble at the level above, maintaining the required invariance for an unbiased estimator. The measured acceptance stays in the range 0.53–0.57 throughout Section 4.

The adaptive schedule. Choosing E_{k+1} from the current population makes the schedule adaptive, and adaptive schedules cost exact unbiasedness. The estimator stays consistent, but $\mathbb{E}[\hat{Z}] = Z$ no longer holds at finite N. This is a property of adaptive SMC generally, and of nested sampling in particular, whose energy threshold is likewise read off the current particles—Salomone et al. (2025) show that the fixed-threshold form of nested sampling is unbiased while the adaptive form is only consistent. The nature of the moving maximum E is that progressing to E_{k+1} pushes the target fraction of the population out of the support of $\rho_{E_{k+1}}$, this form of progression naturally benefits from an adaptive schedule. If too large a jump is taken in tempering the population collapses to a single particle, in the case of nested sampling it could cause all particles to be outside the support of the next target, and the algorithm would fail. Hence, initially running with an adaptive schedule is a practical necessity, and we find that it does not introduce a significant bias in practice; its finite-particle consequence is detailed in Appendix C.2 and measured along the longest ladder of the benchmark in Appendix C.5. A second ESS target controls full systematic resampling of the population weights, allowing many levels to be crossed without resampling; Appendix C.4 details this cadence.

Termination. Unlike SMC, where the invariant target is the canonical ensemble, quenched sampling would theoretically continue quenching until the energy of each particle is at the global minimum. One can either truncate at a chosen energy level, or use the standard criterion of Skilling (2006), estimating the remaining integral mass from $G_\nu(E)\, e^{-U_{\min}}$, where $U_{\min}$ is the current lowest particle energy, and stopping when this estimate falls below e^{dlogz} times the accumulated integral. We adopt $\mathrm{dlogz} = -3$ as a default, which is standard for typical Bayesian posterior targets (Handley et al., 2015). For targets with a first-order transition, the threshold must be deepened so that termination occurs beyond the transition rather than in the broad phase, as detailed in Appendix A.2. Finally Eq. (9) can be evaluated by quadrature on the ladder of quenched transitions. Equally weighted posterior draws can be obtained from the augmentation in Eq. (6) by sampling a level with probability proportional to $G_\nu(E)e^{-E}\mathrm{d}E$, then subsequently sampling a particle from that level's cloud. The full weighted-pool construction is detailed in Appendix C.3.

4 Experiments

We demonstrate the method on three sets of experiments.

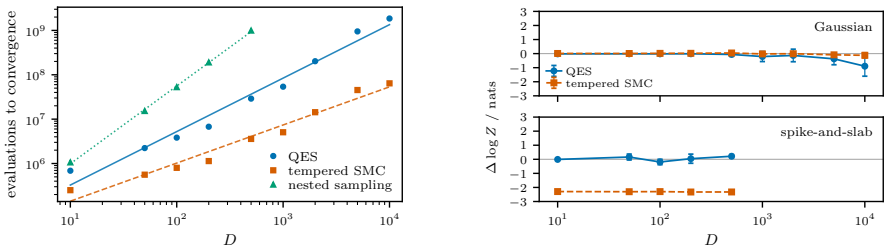


Figure 2: Cost and accuracy comparison across the three methods. *Left:* evaluations to convergence on the Gaussian, with fitted power laws CD^{α}: $\alpha = 0.86$ (tempered SMC), 1.21 (QES), 1.75 (nested sampling, counted in likelihood evaluations rather than gradients). *Right:* Integral bias on a shared symmetric axis, Gaussian and spike–slab.

(i) Analytic targets, including a Gaussian with no transition and a spike–slab mixture with a first-order transition, to test scalability and accuracy against exact results (Section 4.1).

(ii) Bayesian neural network inference and model comparison on UCI classification benchmarks (Section 4.2).

(iii) A lattice ϕ^6 field theory with a first-order transition, to test traversal of a discontinuous phase transition at scale in a physical model (Section 4.3).

Across the examples we primarily compare against tempered SMC using the same inner kernel and tuning. For the BNN tasks we also compare against scalable MCMC methods, including NUTS (Hoffman & Gelman, 2014) and MCLMC (Robnik et al., 2023), and on the analytic targets we include a gradient-free nested sampling baseline (Yallup, 2026). Experimental configurations are detailed in Appendix A.2. All experiments are implemented in `JAX` (Bradbury et al., 2018) using `blackjax` (Cabezas et al., 2024) and run on a Google Cloud (2024) TPU v6e (Trillium) accelerator. All sampling uses single precision `float32` numerics, enabling efficient computation on a variety of accelerator backends.

4.1 Analytic targets

We summarise QES against tempered SMC using the same inner MCMC kernel, and against classic nested sampling using slice sampling, in Table 3; populations, budgets and thresholds are matched throughout (Appendix A.2).

The Gaussian target combines a unit Gaussian prior with a narrower likelihood, $\sigma = 0.5$, and tests scaling over $D \in \{10, 100, 1000, 10000\}$. QES remains accurate well beyond the range reached by the nested sampling baseline, with a sub-nat residual bias emerging only at $D = 10^4$ (Figure 2), whose origin is traced in Appendix C.5; the cost scaling and pooled effective sample sizes of the three methods are analysed in Appendix A. The spike–slab benchmark—a diffuse slab beneath a narrow spike carrying 90% of the posterior mass—provides a controlled first-order transition (Figure 3, left). Across the tested dimensions and schedules, tempered SMC remains approximately 2.3 nats low, whereas the QES estimates are consistent with the analytic value within the seed variation. Finally, a strongly anisotropic bimodal mixture tests mode balance: losing a symmetric mode leaves the marginal likelihood unchanged, so we report mode occupancy (exact value 0.5), and both methods recover the balance on average (Appendix C.4).

4.2 Bayesian neural networks

A Bayesian treatment of neural networks places a prior on the network weights and samples the posterior over parameters given a training dataset (Izmailov et al., 2021): a high-dimensional, highly multimodal task whose training can exhibit phase-transition properties (Montanari & Zhong, 2022; Power et al., 2022). We fit five UCI classification tasks (Kelly et al.) in the benchmark setting of Sommer et al. (2024), with fully connected 2×16 $\tanh$ networks of 528–1280 parameters, and compare QES and tempered SMC at matched population, kernel and budget against ensembles of NUTS (Hoffman & Gelman, 2014) and MCLMC (Robnik et al., 2023) chains; architectures, priors,

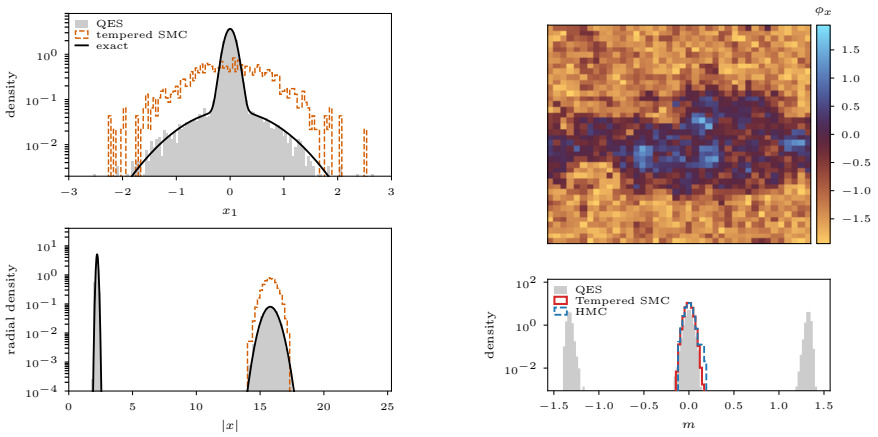


Figure 3: Posterior reconstruction across the two first-order transitions. *Left:* the spike–slab target at $D = 500$ against closed form, the marginal in the first coordinate (top) and the radial density (bottom): the phases are disjoint shells and the coexistence gap is empty over four decades—tempered SMC samples only the slab. *Right:* the lattice ϕ^6 model at coexistence ($L = 40$): a retained ladder state at the free-energy barrier (top) and the unsmoothed magnetisation histogram (bottom)—QES resolves all three m branches, while tempered SMC and an HMC chain occupy only the disordered branch.

budgets and ESS conventions are detailed in Appendices A.2 and A.3. As summarised in Table 1, all methods give very similar predictive NLL, and at comparable accuracy QES returns between 1.4 and 2.0 times the pooled Kish ESS per gradient of tempered SMC across the five tasks, while successfully exploring these BNN posteriors from prior initialisation.

Secondly, we use the marginal likelihood estimates to compare activation functions on the same networks, adding a Laplace approximation built from a MAP estimate with generalised Gauss–Newton curvature (Kingma & Ba, 2015; Duffield et al., 2024) as an approximate baseline. The two particle methods agree closely on every dataset while the Laplace approximation reports inflated factors and picks inconsistent activations (Table 4), validating the QES marginal likelihood for model comparison in Bayesian neural networks at this scale.

Lastly, we search for phase transitions in the full batch likelihood. For this we extend the convergence criterion from $\mathrm{dlogz} = -3$ to the deeper values used in the spike–slab experiments, and reconstruct the level marginal $P(E) \propto G_\nu(E)e^{-E}$—the integrand of Eq. (9), read directly off the quenched ladder. Bimodality in $P(E)$ identifies a phase transition, and a barrier growing with system size identifies it as first order (Lee & Kosterlitz, 1991). Figure 4 shows this profile for the five classification tasks against the spike–slab target of Section 4.1 at comparable dimensionality: the spike–slab barrier grows extensively, while the network posteriors show no evidence of first-order structure at the resolution probed (Appendix A). Studying these phenomena empirically in finite-width networks on real data remains an open question, and quenched sampling provides a unique tool for this investigation.

4.3 Lattice ϕ^6: A first-order transition

Finally, we consider the lattice ϕ^6 model, a two-dimensional scalar field theory with a first-order transition on a continuous state space (Makhankov, 1990; Sanati & Saxena, 1999). The field is defined on a periodic $L \times L$ lattice, with $D = L^2$. We use the exactly samplable free field at reference mass m_0^2 as the prior $\pi_0(\phi)$, and define the remaining interaction action—the energy function passed to the sampler—as $U(\phi) = \sum_x \left[\frac{1}{2}(\mu^2 - m_0^2)\phi_x^2 + \lambda\,\phi_x^4 + \eta\,\phi_x^6\right]$, where x runs over the lattice sites, so the posterior is $\propto \pi_0(\phi)e^{-U(\phi)}$. With $\lambda < 0$ and $\eta > 0$, the on-site potential has a pair of minima that exchange stability with the one at the origin *discontinuously*, giving a first-order transition rather than the second-order one of ordinary ϕ^4. Mean field puts coexistence at $\mu_\star^2 = \lambda^2/2\eta$; at $L = 40$, $\lambda = -1$, $\eta = 0.2$, $m_0^2 = 1$ the located coexistence point is $\mu^2 = 2.28347$ (Appendix A.3).

Table 1: Posterior comparison on the UCI classification tasks, averaged over three seeds: held-out posterior-predictive NLL and ESS per 10^6 gradient evaluations (including warmup and tuning). ESS conventions for the particle and chain methods are given in Appendix A.2; the corresponding model comparison is in Table 4.

		test NLL				ESS per 10^6 gradients			
	D	QES	SMC	NUTS	MCLMC	QES	SMC	NUTS	MCLMC
sonar	1280	0.458 ± 0.002	0.456 ± 0.001	0.457 ± 0.003	0.472 ± 0.002	3184 ± 81	2195 ± 152	8996 ± 42	64 ± 1
glass	528	0.990 ± 0.002	0.991 ± 0.002	0.991 ± 0.002	1.002 ± 0.001	2367 ± 19	1362 ± 84	11631 ± 38	1837 ± 500
heart	528	0.464 ± 0.000	0.465 ± 0.001	0.463 ± 0.000	0.463 ± 0.001	3017 ± 23	1836 ± 45	10556 ± 68	442 ± 23
australian	544	0.319 ± 0.000	0.319 ± 0.001	0.319 ± 0.001	0.319 ± 0.001	2270 ± 8	1422 ± 34	10804 ± 18	1296 ± 138
wine	560	0.909 ± 0.000	0.909 ± 0.001	0.909 ± 0.001	0.911 ± 0.000	1189 ± 10	582 ± 32	4375 ± 17	208 ± 14

We report two order parameters: $m = V^{-1} \sum_x \phi_x$, which identifies the Z_2 parity symmetry sector, and $q = V^{-1} \sum_x \phi_x^2$, which distinguishes the phases, with $V = L^2$; configurations are classified by the mean-field cut $q_{\text{cut}} = 1.25$, and at coexistence the two phases carry equal probability.

Table 2: Lattice ϕ^6 at coexistence, $L = 40$ ($D = 1600$), $\mu^2 \approx 2.28$, averaged over three seeds; configurations in Appendix A.2. SMC is shown at three adaptive schedules and at 10,000 equally spaced temperatures. The QES seeds agree to 0.01 nats in $\log \hat{Z}$, and only QES resolves both phases.

	schedule	levels	$\log \hat{Z}$	ordered share	$\langle \lvert m \rvert \rangle$	$\langle q \rangle$
QES	0.95	11502 ± 372	-53.76 ± 0.01	0.543 ± 0.009	0.733 ± 0.011	1.112 ± 0.015
tempered SMC	0.95	20	-54.54	0.000	0.027	0.184
	0.99	44	-54.54	0.000	0.028	0.184
	0.999	137	-54.54	0.000	0.028	0.184
	fixed	10,000	-54.54	0.000	0.028	0.184

The QES ladder resolves both phases (Figure 3, right), with an ordered share of 0.543 ± 0.009 over three seeds against the coexistence value of 0.5. Tempered SMC does not enter the ordered phase in any tested run, even when its target ESS is increased to 0.999, and a long preconditioned HMC chain (window-adapted HMC, Appendix A.2) remains in the disordered phase. Resolving both phases requires a termination threshold deep enough to span the energy barrier and therefore carries a substantial computational cost. Flat-histogram methods offer another route across such transitions (Wang & Landau, 2001; Dayal et al., 2004); classic nested sampling resembles an optimal energy binning discovered on the fly (Pártay et al., 2010), and QES inherits this property.

5 Conclusion

In this work, we isolated the most attractive feature of the classical nested sampling algorithm for estimating marginal likelihoods: the quenched path of monotonically decreasing energy. Rather than trying to construct an MCMC mutation kernel that can efficiently follow this path in high dimensions, which is notoriously challenging, we demonstrate that the hard energy constraint can be replaced with a family of soft microcanonical potentials that preserves the quenched path while admitting efficient use of standard gradient-based mutation kernels. The resulting Quenched Ensemble Sampling method is a sequential Monte Carlo algorithm that performs at scale and is robust to first-order phase transitions, where tempering is not. In thermodynamic terms, the method is simulated quenching where tempering is simulated annealing, and the ability to follow this path of distributions at scale is unique.

Reproducibility statement

An anonymised version of the code developed for this work is available at https://anonymous.4open.science/r/quenched_sampling-EB52/README.md, the full public repository will replace this link upon acceptance.

AI use statement

In this work, we used generative AI tools (Claude Fable 5 and GPT 5.6) for the development and testing of the research code, and for editing and improving text originally written by the authors. We have not used generative AI tools for research ideation or the execution of research ideas—the design of the method and of the experiments, and the interpretation of results—nor to draft sections of the paper or for retrieval and discovery of related work; the remaining disclosure categories are not applicable to this work. Additionally, we used generative AI tools to execute the author-designed experiments on the compute cluster and to prepare figures. We have reviewed all AI-assisted work: AI-generated code was verified against the closed-form analytic benchmarks reported in the paper, and all AI-edited text was reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims or artefacts produced with the aid of generative AI.

References

William Steinbrunn Andras Janosi. Heart Disease, 1989. URL https://archive.ics.uci.edu/dataset/45.

Greg Ashton, Noam Bernstein, Johannes Buchner, Xi Chen, Gábor Csányi, Andrew Fowlie, Farhan Feroz, Matthew Griffiths, Will Handley, Michael Habeck, Edward Higson, Michael Hobson, Anthony Lasenby, David Parkinson, Livia B. Pártay, Matthew Pitkin, Doris Schneider, Joshua S. Speagle, Leah South, John Veitch, Philipp Wacker, David J. Wales, and David Yallup. Nested sampling for physical scientists. *Nat Rev Methods Primers*, 2(1):39, May 2022. ISSN 2662-8449. doi: 10.1038/s43586-022-00121-x. URL https://www.nature.com/articles/s43586-022-00121-x.

B. German. Glass Identification, 1987. URL https://archive.ics.uci.edu/dataset/42.

Robert J. N. Baldock, Noam Bernstein, K. Michael Salerno, Lívia B. Pártay, and Gábor Csányi. Constant-pressure nested sampling with atomistic dynamics. *Phys. Rev. E*, 96(4):043311, October 2017. doi: 10.1103/PhysRevE.96.043311. URL https://link.aps.org/doi/10.1103/PhysRevE.96.043311.

Alexandros Beskos, Natesh Pillai, Gareth Roberts, Jesus-Maria Sanz-Serna, and Andrew Stuart. Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, 19(5A):1501–1534, 2013. ISSN 1350-7265. URL https://www.jstor.org/stable/42919328.

Alexandros Beskos, Ajay Jasra, Nikolas Kantas, and Alexandre Thiery. On the Convergence of Adaptive Sequential Monte Carlo Methods. *The Annals of Applied Probability*, 26(2):1111–1146, 2016. ISSN 1050-5164. URL https://www.jstor.org/stable/43859623.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Alexander Buchholz, Nicolas Chopin, and Pierre E Jacob. Adaptive tuning of hamiltonian monte carlo within sequential monte carlo. *Bayesian Analysis*, 16(3):745–771, 2021.

Alberto Cabezas, Adrien Corenflos, Junpeng Lao, Rémi Louf, Antoine Carnec, Kaustubh Chaudhari, Reuben Cohn-Gordon, Jeremie Coullon, Wei Deng, Sam Duffield, Gerardo Durán-Martín, Marcin Elantkowski, Dan Foreman-Mackey, Michele Gregori, Carlos Iguaran, Ravin Kumar, Martin Lysy, Kevin Murphy, Juan Camilo Orduz, Karm Patel, Xi Wang, and Rob Zinkov. BlackJAX: Composable Bayesian inference in JAX, 2024. URL https://arxiv.org/abs/2402.10797.

Nicolas Chopin and Christian P. Robert. Properties of nested sampling. *Biometrika*, 97(3):741–755, 2010. ISSN 0006-3444. URL https://www.jstor.org/stable/25734120.

P. Dayal, S. Trebst, S. Wessel, D. Würtz, M. Troyer, S. Sabhapandit, and S. N. Coppersmith. Performance Limitations of Flat-Histogram Methods. *Phys. Rev. Lett.*, 92(9):097201, March 2004. doi: 10.1103/PhysRevLett.92.097201. URL https://link.aps.org/doi/10.1103/PhysRevLett.92.097201.

Pierre Del Moral. *Feynman-Kac Formulae*. Probability and its Applications. Springer, New York, NY, 2004. ISBN 978-1-4419-1902-1 978-1-4684-9393-1. doi: 10.1007/978-1-4684-9393-1. URL http://link.springer.com/10.1007/978-1-4684-9393-1.

Arnaud Doucet, Nando de Freitas, and Neil Gordon. An Introduction to Sequential Monte Carlo Methods. In Arnaud Doucet, Nando de Freitas, and Neil Gordon (eds.), *Sequential Monte Carlo Methods in Practice*, pp. 3–14. Springer, New York, NY, 2001. ISBN 978-1-4757-3437-9. doi: 10.1007/978-1-4757-3437-9_1. URL https://doi.org/10.1007/978-1-4757-3437-9_1.

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Phys. Lett. B*, 195: 216–222, 1987. doi: 10.1016/0370-2693(87)91197-X.

Samuel Duffield, Kaelan Donatella, Johnathan Chiu, Phoebe Klett, and Daniel Simpson. Scalable Bayesian Learning with posteriors. *arXiv preprint arXiv:2406.00104*, 2024.

Paul Fearnhead, Christopher Nemeth, Chris J. Oates, and Chris Sherlock. *Scalable Monte Carlo for Bayesian Learning*. Institute of Mathematical Statistics Monographs. Cambridge University Press, Cambridge, 2025. ISBN 978-1-009-28844-6. doi: 10.1017/9781009288460. URL https://www.cambridge.org/core/books/scalable-monte-carlo-for-bayesian-learning/7FDC9891DDA086B85AD8AB386F6F3BD3.

F. Feroz, M. P. Hobson, and M. Bridges. MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics. *Mon. Not. Roy. Astron. Soc.*, 398:1601–1614, 2009. doi: 10.1111/j.1365-2966.2009.14548.x.

E Fong and C C Holmes. On the marginal likelihood and cross-validation. *Biometrika*, 107(2): 489–496, June 2020. ISSN 0006-3444. doi: 10.1093/biomet/asz077. URL https://doi.org/10.1093/biomet/asz077.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2010.00765.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2010.00765.x.

Google Cloud. TPU v6e (Trillium), 2024. URL https://docs.cloud.google.com/tpu/docs/v6e.

Michael Habeck. Nested sampling with demons. *AIP Conf. Proc.*, 1641(1):121–129, January 2015. ISSN 0094-243X. doi: 10.1063/1.4905971. URL https://doi.org/10.1063/1.4905971.

W. J. Handley, M. P. Hobson, and A. N. Lasenby. polychord: next-generation nested sampling. *Mon. Not. Roy. Astron. Soc.*, 453(4):4385–4399, 2015. doi: 10.1093/mnras/stv1911.

Matthew Hoffman and Andrew Gelman. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, January 2014. ISSN 1532-4435. URL https://dl.acm.org/doi/10.5555/2627435.2638586.

Pavel Izmailov, Sharad Vikram, Matthew D. Hoffman, and Andrew Gordon Gordon Wilson. What Are Bayesian Neural Network Posteriors Really Like? In *Proceedings of the 38th International Conference on Machine Learning*, pp. 4629–4640. PMLR, July 2021. URL https://proceedings.mlr.press/v139/izmailov21a.html.

Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The UCI Machine Learning Repository. URL https://archive.ics.uci.edu. Published: UCI Machine Learning Repository.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

Jere Koskela, Paul A. Jenkins, Adam M. Johansen, and Dario Spanò. Asymptotic genealogies of interacting particle systems with an application to sequential Monte Carlo. *The Annals of Statistics*, 48(1):560–583, February 2020. ISSN 0090-5364, 2168-8966. doi: 10.1214/19-AOS1823. URL https://projecteuclid.org/journals/annals-of-statistics/volume-48/issue-1/Asymptotic-genealogies-of-interacting-particle-systems-with-an-application-to/10.1214/19-AOS1823.full.

Namu Kroupa, Gábor Csányi, and Will Handley. Resonances in reflective Hamiltonian Monte Carlo. *Phys. Rev. E*, 111(4):045308, April 2025. doi: 10.1103/PhysRevE.111.045308. URL https://link.aps.org/doi/10.1103/PhysRevE.111.045308.

Thi Le Thu Nguyen, Francois Septier, Gareth W. Peters, and Yves Delignon. Improving SMC sampler estimate by recycling all past simulated particles. In *2014 IEEE Workshop on Statistical Signal Processing (SSP)*, pp. 117–120, June 2014. doi: 10.1109/SSP.2014.6884589. URL https://ieeexplore.ieee.org/document/6884589. ISSN: 2373-0803.

Jooyoung Lee and J. M. Kosterlitz. Finite-size scaling and Monte Carlo simulations of first-order phase transitions. *Phys. Rev. B*, 43(4):3265–3277, February 1991. doi: 10.1103/PhysRevB.43.3265. URL https://link.aps.org/doi/10.1103/PhysRevB.43.3265.

Pablo Lemos, Nikolay Malkin, Will Handley, Yoshua Bengio, Yashar Hezaveh, and Laurence Perreault-Levasseur. Improving Gradient-Guided Nested Sampling for Posterior Inference. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 27230–27253. PMLR, July 2024. URL https://proceedings.mlr.press/v235/lemos24a.html.

Samuel Livingstone and Giacomo Zanella. The Barker proposal: Combining robustness and efficiency in gradient-based MCMC. *J R Stat Soc Series B Stat Methodol*, 84(2):496–523, April 2022. ISSN 1369-7412. doi: 10.1111/rssb.12482. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9303935/.

Fernando Llorente, Luca Martino, David Delgado, and Javier Lopez-Santiago. Marginal likelihood computation for model selection and hypothesis testing: an extensive review. *SIAM Rev.*, 65 (1):3–58, February 2023. ISSN 0036-1445, 1095-7200. doi: 10.1137/20M1310849. URL http://arxiv.org/abs/2005.08334.

Vladimir G. Makhankov. ϕ^6 Theory and Bose Drops. In Vladimir G. Makhankov (ed.), *Soliton Phenomenology*, pp. 211–221. Springer Netherlands, Dordrecht, 1990. ISBN 978-94-009-2217-4. doi: 10.1007/978-94-009-2217-4_8. URL https://doi.org/10.1007/978-94-009-2217-4_8.

Jason D. McEwen, Tobías I. Liaudat, Matthew A. Price, Xiaohao Cai, and Marcelo Pereyra. Proximal Nested Sampling with Data-Driven Priors for Physical Scientists. *Physical Sciences Forum*, 9(1):13, 2023. ISSN 2673-9984. doi: 10.3390/psf2023009013. URL https://www.mdpi.com/2673-9984/9/1/13.

Andrea Montanari and Yiqiao Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *The Annals of Statistics*, 50(5):2816–2847, October 2022. ISSN 0090-5364, 2168-8966. doi: 10.1214/22-AOS2211. URL https://projecteuclid.org/journals/annals-of-statistics/volume-50/issue-5/The-interpolation-phase-transition-in-neural-networks--Memorization-and/10.1214/22-AOS2211.full.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. On adaptive resampling strategies for sequential Monte Carlo methods. *Bernoulli*, 18(1), February 2012. ISSN 1350-7265. doi: 10.3150/10-BEJ335. URL http://arxiv.org/abs/1203.0464.

Iain Murray, David MacKay, Zoubin Ghahramani, and John Skilling. Nested sampling for Potts models. In *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005. URL https://proceedings.neurips.cc/paper/2005/hash/9dc372713683fd865d366d5d9ee810ba-Abstract.html.

Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, April 2001. ISSN 1573-1375. doi: 10.1023/A:1008923215028. URL https://doi.org/10.1023/A:1008923215028.

Radford M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, June 2003. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1056562461. URL https://projecteuclid.org/journals/annals-of-statistics/volume-31/issue-3/Slice-sampling/10.1214/aos/1056562461.full.

Radford M. Neal. *MCMC using Hamiltonian dynamics*. May 2011. doi: 10.1201/b10905. URL http://arxiv.org/abs/1206.1901.

Christopher Nemeth and Paul Fearnhead. Stochastic gradient markov chain monte carlo. *Journal of the American Statistical Association*, 116(533):433–450, 2021.

A. Cerdeira Paulo Cortez. Wine Quality, 2009. URL https://archive.ics.uci.edu/dataset/186.

Nicholas G. Polson and James G. Scott. Vertical-likelihood Monte Carlo, June 2015. URL http://arxiv.org/abs/1409.3601.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets. *CoRR*, abs/2201.02177, 2022. URL https://arxiv.org/abs/2201.02177. arXiv: 2201.02177.

Lívia B. Pártay, Albert P. Bartók, and Gábor Csányi. Efficient Sampling of Atomic Configurational Spaces. *J. Phys. Chem. B*, 114(32):10502–10512, August 2010. ISSN 1520-6106. doi: 10.1021/jp1012973. URL https://dx.doi.org/10.1021/jp1012973.

Ross Quinlan. Statlog (Australian Credit Approval), 1987. URL https://archive.ics.uci.edu/dataset/143.

John R. Ray. Microcanonical ensemble Monte Carlo method. *Phys. Rev. A*, 44(6):4061–4064, September 1991. doi: 10.1103/PhysRevA.44.4061. URL https://link.aps.org/doi/10.1103/PhysRevA.44.4061.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400–407, September 1951. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177729586. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-22/issue-3/A-Stochastic-Approximation-Method/10.1214/aoms/1177729586.full.

Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998. ISSN 1467-9868. doi: 10.1111/1467-9868.00123. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00123.

Gareth O. Roberts and Jeffrey S. Rosenthal. Complexity bounds for Markov chain Monte Carlo algorithms via diffusion limits. *Journal of Applied Probability*, 53(2):410–420, June 2016. ISSN 0021-9002, 1475-6072. doi: 10.1017/jpr.2016.9. URL https://www.cambridge.org/core/journals/journal-of-applied-probability/article/abs/complexity-bounds-for-markov-chain-monte-carlo-algorithms-via-diffusion-limits/AE8D620D7114F27ED889705735C162D7.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, December 1996. ISSN 1350-7265. URL https://projecteuclid.org/journals/bernoulli/volume-2/issue-4/Exponential-convergence-of-Langevin-distributions-and-their-discrete-approximations/bj/1178291835.full.

Jakob Robnik, G. Bruno De Luca, Eva Silverstein, and Uroš Seljak. Microcanonical Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 24(311):1–34, 2023. ISSN 1533-7928. URL http://jmlr.org/papers/v24/22-1450.html.

Robert Salomone, Leah F South, Christopher Drovandi, Dirk P Kroese, and Adam M Johansen. Unbiased and consistent nested sampling via sequential Monte Carlo. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 87(4):1221–1238, September 2025. ISSN 1369-7412. doi: 10.1093/jrsssb/qkaf015. URL https://doi.org/10.1093/jrsssb/qkaf015.

M. Sanati and A. Saxena. Half-kink lattice solution of the phi**6 model. *J. Phys. A*, 32:4311–4320, 1999. doi: 10.1088/0305-4470/32/23/309.

John Skilling. Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1(4):833–859, December 2006. ISSN 1936-0975, 1931-6690. doi: 10.1214/06-BA127. URL https://projecteuclid.org/journals/bayesian-analysis/volume-1/issue-4/Nested-sampling-for-general-Bayesian-computation/10.1214/06-BA127.full.

John Skilling. Galilean and Hamiltonian Monte Carlo. *Proceedings*, 33(1):19, 2019. ISSN 2504-3900. doi: 10.3390/proceedings2019033019. URL https://www.mdpi.com/2504-3900/33/1/19.

Emanuel Sommer, Jakob Robnik, Giorgi Nozadze, Uros Seljak, and David Rügamer. Microcanonical Langevin Ensembles: Advancing the Sampling of Bayesian Neural Networks. October 2024. URL https://openreview.net/forum?id=QMtrW8Ej98.

Robert H. Swendsen and Jian-Sheng Wang. Replica Monte Carlo Simulation of Spin-Glasses. *Phys. Rev. Lett.*, 57(21):2607–2609, November 1986. doi: 10.1103/PhysRevLett.57.2607. URL https://link.aps.org/doi/10.1103/PhysRevLett.57.2607.

Saifuddin Syed, Alexandre Bouchard-Côté, George Deligiannidis, and Arnaud Doucet. Non-Reversible Parallel Tempering: a Scalable Highly Parallel MCMC Scheme. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):321–350, April 2022. ISSN 1369-7412, 1467-9868. doi: 10.1111/rssb.12464. URL http://arxiv.org/abs/1905.02939.

R. Gorman Terry Sejnowski. Connectionist Bench (Sonar, Mines vs. Rocks), 1988. URL https://archive.ics.uci.edu/dataset/151.

Aidan P. Thompson, H. Metin Aktulga, Richard Berger, Dan S. Bolintineanu, W. Michael Brown, Paul S. Crozier, Pieter J. in 't Veld, Axel Kohlmeyer, Stan G. Moore, Trung Dac Nguyen, Ray Shan, Mark J. Stevens, Julien Tranchida, Christian Trott, and Steven J. Plimpton. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications*, 271:108171, February 2022. ISSN 0010-4655. doi: 10.1016/j.cpc.2021.108171. URL https://www.sciencedirect.com/science/article/pii/S0010465521002836.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, 16(2):667–718, 2021. doi: 10.1214/20-BA1221.

Fugao Wang and D. P. Landau. An efficient, multiple range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.*, 86(10):2050–2053, March 2001. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.86.2050. URL http://arxiv.org/abs/cond-mat/0011174.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 681–688, Madison, WI, USA, June 2011. Omnipress. ISBN 978-1-4503-0619-5.

Florian Wenzel, Kevin Roth, Bastiaan Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How Good is the Bayes Posterior in Deep Neural Networks Really? In *Proceedings of the 37th International Conference on Machine Learning*, pp. 10248–10259. PMLR, November 2020. URL https://proceedings.mlr.press/v119/wenzel20a.html.

David Yallup. Nested Sampling with Slice-within-Gibbs: Efficient Evidence Calculation for Hierarchical Bayesian Models, February 2026. URL http://arxiv.org/abs/2602.17414.

David Yallup, Namu Kroupa, and Will Handley. Nested Slice Sampling: Vectorized Nested Sampling for GPU-Accelerated Inference. *Transactions on Machine Learning Research*, pp. arXiv:2601.23252, 2026. ISSN 2835-8856. doi: 10.48550/arXiv.2601.23252. URL https://openreview.net/forum?id=5mF2eRl3gt.

A Additional experimental details

For reference, Algorithm 1 lists one complete level transition of QES as described in Section 3.3. Here K_E denotes the complete inner MCMC mutation block: its construction and number of internal steps are unrestricted, provided that it leaves ρ_E invariant. The transition consists of three main steps: dissecting the current population to choose the next energy level, reweighting and branching the population to account for the new level, and mutating the particles with a kernel invariant to $\rho_{E_{k+1}}$.

Algorithm 1 One quenched level transition

Require: Weighted ensemble $\{(x_j, w_j)\}_{j=1}^N$ targeting ρ_{E_k}, with $\sum_j w_j = 1$

Require: Current level volume $\widehat{G}_{\nu,k}$ and an MCMC kernel K_E that leaves ρ_E invariant

1: **Dissect.** Choose $E_{k+1} < E_k$ by bisection until the ESS of $\{w_j r(x_j)\}_{j=1}^N$ reaches its target.
2: **Measure.** At the pre-mutation positions, compute $\widehat{r}_k \leftarrow \sum_j w_j r(x_j)$ and $\widehat{G}_{\nu,k+1} \leftarrow \widehat{G}_{\nu,k}\widehat{r}_k$.
3: **Reweight.** Set $w_j \leftarrow w_j r(x_j)/\widehat{r}_k$ for every particle.
4: **Branch.** Replace each zero-weight particle by a copy of a survivor drawn by weight; assign half the donor weight to each copy.
5: **Resample.** When required, based on a second ESS threshold, resample the ensemble and reset all $w_j \leftarrow 1/N$.
6: **Mutate.** Apply $K_{E_{k+1}}$ independently to every particle.

Ensure: $\{(x_j, w_j)\}_{j=1}^N$, E_{k+1}, and $\widehat{G}_{\nu,k+1}$

A.1 Importance-sampling initialisation

Both QES and tempered SMC may be initialised by an initial sampling step from the prior. Drawing N samples and setting E_0 to the maximum energy in that population is a valid starting point, but it exposes the initial importance step to extreme weights in the tail of the prior-energy distribution. For QES, we instead draw an anchor population of $N_\mathrm{a} = 10N$ independent samples $\widetilde{x}_i \sim \pi$ and choose the initial energy threshold E_0 by bisection until the weights $\omega_i^{(0)} = (E_0 - U(\widetilde{x}_i))_+^\nu$ have ESS N. Their empirical mean initialises the level volume,

$$\widehat{G}_\nu(E_0) = \frac{1}{N_\mathrm{a}} \sum_{i=1}^{N_\mathrm{a}} \omega_i^{(0)}, \tag{14}$$

and their normalised values define an importance-sampling approximation to ρ_{E_0}, from which the working population is resampled. This stabilizes the initial importance step and prevents isolated outlying density evaluations from determining the initialisation.

Levels above E_0 may be estimated from the same prior draws and included in the final quadrature. Tempered SMC can use the same construction by replacing $\omega_i^{(0)}$ with the importance ratio from the prior to its chosen first tempered target.

A.2 Sampler configurations

The configurations below are grouped by experiment because the population size, target ESS and mutation budget are selected at the scale of each problem. Within an experiment, QES and tempered SMC use the same population, diagonally preconditioned MALA kernel, mutation budget and target

ESS; only the sequence of intermediate distributions differs. This ensures the experiments have a strong control that can isolate the effect of the quenched path. Unless stated otherwise, reported values are averaged over three independent sampler seeds. QES uses softness $\nu = 2$ throughout, and triggers full resampling when the lineage-grouped ESS falls below $0.5N$ (Appendix C.4). The diagonal preconditioner and scalar step size are estimated from the preceding population and then held fixed during each level's mutation, with the MALA acceptance rate targeted at 0.574.

Analytic targets. All methods use $N = 1000$ particles. QES and tempered SMC choose successive levels at a default target ESS of 0.95. The mutation budget follows $\sqrt{D}$, the rate at which a random walk crosses a D-dimensional shell, rounded up to a power of two so that ladder cost stays predictable and floored at the value calibrated on the smallest targets,

$$n_{\text{steps}} = \max\left(16,\ 2^{\lceil \log_2 \sqrt{D} \rceil}\right),$$

MALA steps per level. Over the suite this takes four values: 16 for $D \leq 200$, 32 at $D = 500$ and 1000, 64 at $D = 2000$, and 128 at $D = 5000$ and 10^4. The Gaussian run terminates at $\mathrm{dlogz} = -3$. For the spike–slab and bimodal targets, the threshold is deepened to $-2.5D - 20$, where D is the dimensionality of the target, this ensures that termination occurs beyond the transition rather than in the broad phase. Whilst this seems like a problem-specific tuning, it is in reality a general requirement for any sampler to reach the low-energy phase of a first-order transition. When scanning for phase change behaviour in e.g. Section 4.2, the depth can be set arbitrarily low. On physical targets, the depth required can often be motivated from ground state energy estimates. As such $\mathrm{dlogz} = -3$ is more of a rule to match the efficiency of tempering on standard Bayesian posterior targets, rather than a sensitive tuning parameter, and in practice quenching can continue until all particles are at the ground state, although this would be inefficient for many problems.

The nested sampling baseline uses 1000 particles, removes 10% of the live set per iteration, and applies $2D$ slice-sampling steps per replacement. Its cost is reported in likelihood evaluations rather than gradients; consequently, only scaling exponents, not absolute intercepts, are compared across that boundary.

On the Gaussian benchmark of Table 3, the number of quenched levels grows linearly with dimension at fixed target ESS, against $\sqrt{D}$ for tempering, so QES requires more gradient evaluations; however, the number of levels carrying appreciable posterior weight grows as $\sqrt{D}$, while it remains roughly constant for tempering. From $D = 100$ onward the larger QES pool outweighs the longer path and gives a higher pooled Kish ESS per gradient evaluation (Appendix C.3). The residual bias at $D = 10^4$ is attributed to the increased length of the QES path and accumulated finite-N effects (Appendix C); tempered SMC is more accurate there, as expected on a unimodal target. On the spike–slab runs, the large nominal SMC Kish ESS measures weight concentration within the sampled phase and does not certify convergence of the path.

Bayesian neural networks. QES and tempered SMC use $N = 1000$, target ESS 0.90, and 96 MALA steps per level, with three sampler seeds for each fixed data split. NUTS uses 1000 independent chains initialised from the prior, 250 window-adaptation steps, and 1000 retained transitions. MCLMC follows the tuning procedure of Sommer et al. (2024), using 1000 prior-initialised chains, 2000 tuning steps, and 10^4 sampling steps. The deep ensemble optimiser initialisation is omitted as we are interested in purely sampling performance. MCLMC traces are thinned by a factor of ten only when computing the rank-normalised ESS; all simulated steps remain included in the gradient count. The Laplace–GGN baseline uses a MAP estimate obtained with 2×10^4 full-batch Adam steps at learning rate 10^{-3}, followed by a generalised Gauss–Newton approximation. This positive-semidefinite curvature is preferable to the raw Hessian for the non-convex network posterior; the resulting Bayes factors are reported in Table 4. In Table 1, each particle run reports one Kish calculation pooling the per-particle weights from every rung of the ladder, while the chain methods report the rank-normalised bulk ESS of Vehtari et al. (2021) over all 1000 chains jointly—not strictly comparable to the pooled Kish ESS, but usable as a qualitative comparison. Strict efficiency comparison between particle methods and chain ensembles is difficult due to implementation differences; under the tested budgets NUTS is the most efficient method, and MCLMC trails slightly in predictive NLL on two of the five tasks.

In terms of the free-energy profile $F(E) = -\log P(E)$ shown in Figure 4, a first-order transition appears as two peaks in $P(E)$ separated by a valley of suppressed interfacial states; the barrier ΔF

Table 3: The analytic test problems (target definitions in Appendix A.3): $\log Z$ bias against the analytic value and pooled Kish ESS per 10^6 gradient evaluations. Gaussian $\log Z$ errors use ten seeds; the ESS values and remaining results use three. The bimodal test problem tests against *mode occupancy*, the diagnostic that detects mode loss. The ESS convention is one Kish calculation after pooling the full per-particle weight vectors from every rung, as defined in Appendix C.3, divided by that run's own gradient count. NS uses slice sampling and counts likelihood evaluations as opposed to gradients, dashes mark cells not run due to computational cost. Tempering is consistent with the analytic value where there is no transition and 2.3 nats low wherever there is one, while matching QES on mode balance. QES starts to exhibit a residual bias on the Gaussian at $D = 10^4$.

		$\Delta \log Z$			ESS per 10^6 gradients		
target	D	QES	SMC	NS	QES	SMC	NS
Gaussian	10	-0.02 ± 0.04	$+0.01 \pm 0.03$	$+0.01 \pm 0.08$	21519	26507	126
	100	-0.02 ± 0.08	$+0.02 \pm 0.05$	$+0.66 \pm 0.07$	13384	7933	2.47
	1000	-0.21 ± 0.36	-0.02 ± 0.12	—	2972	1228	—
	10000	-0.89 ± 0.71	-0.12 ± 0.18	—	266	101	—
spike–slab	10	-0.01 ± 0.03	-2.29 ± 0.00	$+0.09 \pm 0.09$	6564	48393	19
	50	$+0.16 \pm 0.21$	-2.30 ± 0.00	$+0.19 \pm 0.13$	3707	26233	0.94
	200	$+0.04 \pm 0.32$	-2.32 ± 0.01	—	2698	13168	—
	500	$+0.22 \pm 0.12$	-2.32 ± 0.05	—	702	3957	—
		mode occupancy (true 0.5)			ESS per 10^6 gradients		
	D	QES	SMC	NS	QES	SMC	NS
bimodal	10	0.49 ± 0.24	0.51 ± 0.02	0.51 ± 0.22	4494	7080	7.77
	50	0.60 ± 0.19	0.49 ± 0.04	0.61 ± 0.16	2679	2916	0.46
	100	0.46 ± 0.13	0.45 ± 0.05	—	2044	1984	—
	200	0.63 ± 0.25	0.52 ± 0.20	—	1479	1448	—

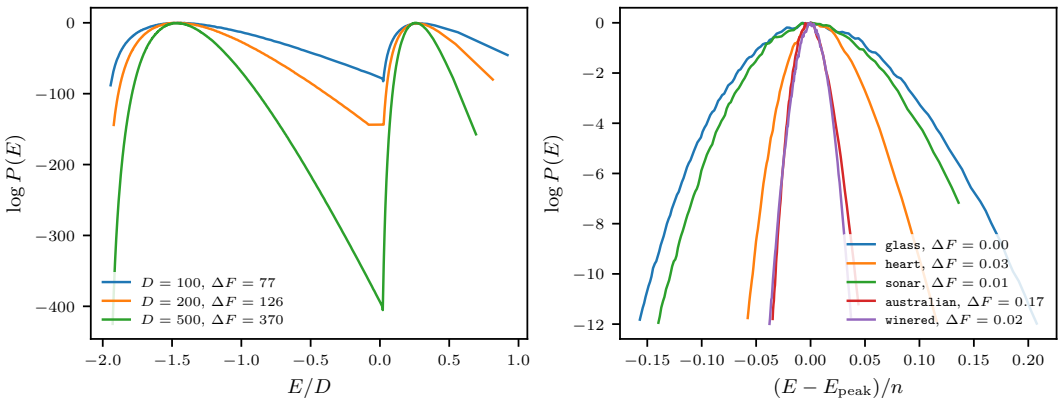


Figure 4: Scan for phase transitions in the level marginal $P(E) \propto G_\nu(E)\, e^{-E}$, the integrand of Eq. (9), read from a single ladder per curve. *Left:* the spike–slab posteriors of Section 4.1 are bimodal with a free-energy barrier ΔF that grows extensively with dimension — a first-order transition. *Right:* the UCI classification posteriors at comparable dimension show no evidence of first-order phase structure at the resolution probed, across a deep energy range. The energy axis is made intensive, by dimensionality for the spike–slab and by training set size for the networks, with the network profiles additionally centred on the posterior modal energy E_{peak} so that they occupy a similar range.

is the depth of that valley below the lower peak, in nats. By Eq. (6), $E = U(x) + s$ with independent slack $s \sim \Gamma(\nu+1, 1)$, so $P(E)$ is the $\beta = 1$ posterior energy distribution under a fixed smoothing of

Table 4: Model comparison on the UCI classification tasks, fixed network dimensions comparing $\tanh$ against ReLU activations. $\log \mathrm{BF} = \log \hat{Z}(\tanh) - \log \hat{Z}(\mathrm{ReLU})$ is computed over three repeated seeds and used to compare the two activations where shading marks the selected model. QES and tempered SMC agree on every dataset and their Bayes factors agree to within 0.6 nats; every factor is resolved (the smallest lies 3.5 seed standard deviations from zero) while on `glass`, `sonar` and `winered` the test NLL weakly favours the activation the Bayes factors reject. The Laplace–GGN approximation reports factors two to four times larger and selects the opposite activation on some tasks.

			test NLL			log BF		
		D	QES	SMC	Laplace	QES	SMC	Laplace
glass	tanh	528	0.990 ± 0.002	0.991 ± 0.002	1.231 ± 0.001	-6.0 ± 0.46	-6.2 ± 0.11	$+2.3 \pm 0.39$
	ReLU		0.993 ± 0.005	0.993 ± 0.001	1.181 ± 0.024			
heart	tanh	528	0.464 ± 0.000	0.465 ± 0.001	0.503 ± 0.002	$+7.7 \pm 0.46$	$+7.7 \pm 0.03$	$+19.6 \pm 0.49$
	ReLU		0.465 ± 0.001	0.467 ± 0.002	0.495 ± 0.004			
sonar	tanh	1280	0.458 ± 0.002	0.456 ± 0.001	0.534 ± 0.007	$+5.3 \pm 0.11$	$+5.2 \pm 0.06$	$+15.7 \pm 0.45$
	ReLU		0.444 ± 0.002	0.444 ± 0.001	0.505 ± 0.004			
australian	tanh	544	0.319 ± 0.000	0.319 ± 0.001	0.428 ± 0.001	$+10.5 \pm 0.22$	$+10.6 \pm 0.05$	$+36.9 \pm 1.55$
	ReLU		0.324 ± 0.001	0.324 ± 0.001	0.448 ± 0.027			
winered	tanh	560	0.909 ± 0.000	0.909 ± 0.001	1.005 ± 0.003	-37.3 ± 0.10	-37.9 ± 0.16	$+15.9 \pm 4.50$
	ReLU		0.930 ± 0.001	0.929 ± 0.000	0.985 ± 0.007			

Table 5: Dictionary between Bayesian inference and statistical mechanics. The prior plays the role of the (usually implicit) a priori phase-space measure, and consequently enters the statistical-mechanics description only through the density of states, which is the prior mass per unit energy rather than a microstate count.

Bayesian inference	**Statistical mechanics**
Prior $\pi(x)\,\mathrm{d}x$	Phase-space measure $\mathrm{d}\mu(x)$
Negative log-likelihood $-\log \mathcal{L}(x)$	Energy $U(x)$
Tempered likelihood $\mathcal{L}^{\beta}(x)$	Boltzmann factor $e^{-\beta U(x)}$
Tempered posterior $\rho_{\beta}(x)$	Canonical (Boltzmann) distribution
Marginal likelihood $Z = Z(1)$	Partition function $Z(\beta)$
Likelihood level set $\{\mathcal{L}(x) > \lambda\}$	Energy level set $\{U(x) < E\}$
Prior density of $-\log \mathcal{L}$	Density of states $g(E)$
Cumulative prior mass $G(E)$ (prior volume X)	Cumulative density of states

width $\sqrt{\nu + 1}$ nats—invisible to extensive first-order structure, but blurring features narrower than a few nats.

Lattice field theory. At the 40×40 coexistence point, QES and tempered SMC use $N = 8000$, target ESS 0.95, and 64 MALA steps per level. The termination threshold $\mathrm{dlogz} = -850$ is chosen to extend the quenched ladder beyond the approximately 741-nat (0.463 per site) uniform-field barrier. Tempered SMC is additionally run at target ESS 0.99 and 0.999 to test whether schedule refinement recovers the ordered phase. The single-chain control uses window-adapted NUTS with 500 warmup steps and 2000 retained states from a prior draw.

All samplers are implemented in JAX and BlackJAX and run in `float32` on a single TPU v6e. Reported gradient totals include adaptation and tuning; nested sampling instead counts every likelihood evaluation.

A.3 Problem definitions

Table 5 collects the correspondence between the Bayesian and statistical-mechanics vocabularies used throughout the paper.

Analytic targets. All three targets use the standard Gaussian prior $\pi = \mathcal{N}(0, I_D)$ and a Gaussian-mixture likelihood, so their marginal likelihood and posterior moments are available in closed form. The Gaussian and spike–slab targets share the form

$$e^{-U(x)} = \sum_{j=1}^{J} h_j \exp\left(-\frac{\|x\|^2}{2\sigma_j^2}\right). \tag{15}$$

Writing $\tau_j^2 = \sigma_j^2/(1+\sigma_j^2)$, the marginal likelihood and posterior component weights are

$$Z = \sum_{j=1}^{J} h_j \tau_j^D, \qquad \omega_j = \frac{h_j \tau_j^D}{Z}. \tag{16}$$

Conditional on component j, the posterior is $\mathcal{N}(0, \tau_j^2 I_D)$. This gives an exact reference for both the marginal likelihood and posterior marginals.

Gaussian a single isotropic component ($J = 1$, $\sigma = 0.5$) over all D coordinates. This smooth, unimodal target provides a scaling control without a phase transition.

Spike–slab a broad component $\sigma_1 = 1$ and a narrow component $\sigma_2 = 0.1$. The narrow component's height is chosen so that it carries 90% of the integral mass at every D. The posterior phases occupy disjoint radial shells, producing a controlled first-order transition whose depth grows with dimension.

Bimodal two equal-height anisotropic components centred at $\pm\mu$, with coordinate widths logarithmically spaced from 10^{-2} to 1. The posterior modes are related by symmetry and therefore each carry exactly half of the integral mass, making mode occupancy an exact diagnostic of population balance.

For the first two targets, U depends on x only through $s = \|x\|^2$, with $s \sim \chi^2_D$ under the prior. Their softened level volume therefore reduces to the one-dimensional integral

$$G_\nu(E) = \int_0^{s_E} p_{\chi^2_D}(s)\,[E - U(s)]^\nu \,\mathrm{d}s, \tag{17}$$

where s_E solves $U(s_E) = E$. Numerical quadrature of this expression provides an independent reference for the level-volume estimate at every rung, not only for the final Z.

Bayesian neural networks. The UCI classification tasks use fully connected networks with two hidden layers of width 16. Unless an activation comparison is stated explicitly, the activation is $\tanh$. Every parameter has an independent standard normal prior; each layer's weight contribution is scaled by $1/\sqrt{d_{\text{in}}}$ in the forward pass and hidden-layer biases are scaled by 0.1. Thus the effective weights have variance $1/d_{\text{in}}$ while the sampler retains isotropic standard-normal coordinates. Inputs are standardised using training-set statistics from a fixed 90–10 split. Classification uses a categorical likelihood on the network logits f_k,

$$p(y \mid x, \theta) = \frac{\exp f_y(x,\theta)}{\sum_{k=1}^{K} \exp f_k(x,\theta)}, \tag{18}$$

where K is the number of classes. Reported test negative log-likelihoods are the posterior-predictive mixture, evaluated over S posterior draws $\{\theta_s\}$ and the held out test set $\{(x_i, y_i)\}$ of size n_{test},

$$\text{NLL} = -\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \log \frac{1}{S} \sum_{s=1}^{S} p(y_i \mid x_i, \theta_s). \tag{19}$$

The datasets chosen correspond to the Red Wine Quality (Paulo Cortez, 2009), Heart Disease (Andras Janosi, 1989), Australian Credit Approval (Quinlan, 1987), Sonar (Terry Sejnowski, 1988) and Glass identification (B. German, 1987) tasks from the UCI repository. These are small classification tasks, with a mixture of continuous and discrete inputs from input dimension 9-60, and a mixture of binary and multi-class outputs. The dataset sizes range from 208-1599, generally small enough to allow for rapid experimentation with a full batch likelihood. Scaling to datasets with larger number

of inputs typically requires employing Stochastic Gradient MCMC techniques (Nemeth & Fearnhead, 2021) such as SGLD (Welling & Teh, 2011). How well these approaches perform relative to full-batch methods is an open question (Wenzel et al., 2020), and employing a quenched path in a mini-batch likelihood settings is conceptually challenging. Comparison of marginal likelihood between full-batch quenched path estimation and exhaustive leave n-out cross-validation would be an interesting future direction (Fong & Holmes, 2020).

Lattice field theory. The lattice target is the periodic $L = 40$ scalar field defined in Section 4.3, with an exactly sampled Gaussian free-field prior at $m_0^2 = 1$. We set $\lambda = -1$, $\eta = 0.2$, and use the finite-volume coexistence point $\mu^2 = 2.28347$, located by bracketing pilot QES ladders at $L = 40$ on equal phase weight; the equal-weight points measured this way at $L = 8$–40 drift with the expected $2 \ln 2/(V\,\Delta q)$ multiplicity correction, extrapolating to $\mu^2_\infty \approx 2.283$. Both particle methods include the exact Z_2 transformation $\phi \mapsto -\phi$ as an invariant move, removing sign switching as a confound. The phase coordinate $q = L^{-2} \sum_x \phi_x^2$ is invariant under this move, so crossing between the ordered and disordered phases must still be achieved by the sampling path.

B Mobility in the soft microcanonical family

The exponent ν controls the strength at which the barrier repulsion appears. As illustrated in Figure 5, the effective gradient of the boundary term becomes a pronounced steep force close to the boundary, a valid concern is then how *mobile* a particle can be in practice under this force. Writing the slack from the boundary as $s = E - U(x)$, the contribution of the soft constraint to the score is

$$\nabla \log(E - U(x))_+^\nu = -\frac{\nu}{s} \nabla U(x). \tag{20}$$

As $\nu \to 0$, the target approaches the hard indicator used by nested sampling. At the opposite extreme, setting $E = \nu/\beta$ and letting $\nu \to \infty$ gives $(E - U)_+^\nu = E^\nu\,(1 - \beta U/\nu)_+^\nu \to E^\nu e^{-\beta U}$, so a joint scaling of level and softness recovers the canonical ensemble at inverse temperature β: the family interpolates exactly between the nested-sampling constraint and tempering. We aim to motivate the specific choice $\nu = 2$ as a compromise between the two extremes.

Firstly, we note that the choice $\nu = 2$ has a useful boundary interpretation. Let $\beta_E = \partial_E \log g(E)$ denote the local slope of the density of states. When $\log g$ is approximately linear over the typical slack, $g(E - s) \simeq g(E) e^{-\beta_E s}$ and hence

$$s \mid E \;\simeq\; \mathrm{Gamma}(\nu + 1, \text{rate } \beta_E), \qquad \mathbb{E}\left[\frac{\nu}{s}\right] \simeq \beta_E, \qquad \mathrm{Var}\left[\frac{\nu}{s}\right] \simeq \frac{\beta_E^2}{\nu - 1}. \tag{21}$$

The variance is finite for $\nu > 1$, making $\nu = 2$ the smallest integer for which both the mean and variance of the boundary contribution are finite. At the same time, the target density vanishes quadratically at $U = E$, replacing the hard wall by a continuously differentiable turning surface. Whilst there may be stronger arguments for other values of ν, we primarily motivate the choice empirically, using the spike–slab target to illustrate the trade-off between bias and cost.

Selecting the softness. We sweep ν on the $D = 200$ spike–slab target, where a genuine first-order transition makes excessive departure from the nested sampling limit to tempering directly observable. All other settings are held fixed, and each point is repeated over three seeds.

The accurate region is broad, every point in $0.75 \lesssim \nu \lesssim 2.25$ is consistent with the analytic value within the seed variation. At the hard end, the gradient cost increases by a factor of 3.6 between $\nu = 4$ and $\nu = 0.5$, and the estimate at $\nu = 0.5$ is $-0.78\,{\scriptstyle \pm 0.24}$ nats low. Above $\nu \simeq 2.5$, the bias moves steadily toward the tempered-SMC deficit, reaching $-2.24\,{\scriptstyle \pm 0.09}$ nats at $\nu = 4$. The default $\nu = 2$ is consequently near the least expensive edge of the accurate region while remaining separated from the tempering-like regime. This gives justification for a single choice that requires no target-specific adjustment in our experiments.

Mobility near the boundary. We measure actual particle movement on the isotropic Gaussian, defining a proposal as near the boundary when its slack is below one fifth of the mean slack at that level. With $\nu = 2$, only 3–7% of proposals fall in this region. Their move rate is lower than in the bulk, but remains substantial: at $D = 1000$ it is 0.289, compared with 0.598 in the full population.

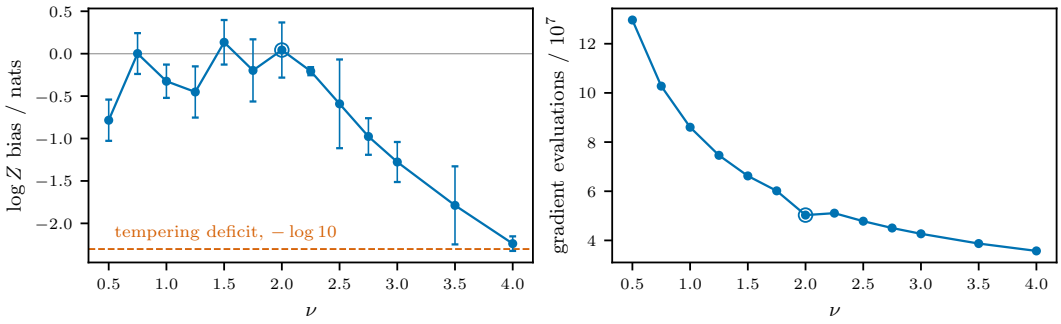


Figure 5: Softness sweep on the spike–slab target at $D = 200$, with three seeds per point and the default $\nu = 2$ circled. *Left:* $\log Z$ bias; the dashed line marks the $-\log 10$ deficit of tempered SMC, corresponding to loss of the spike carrying 90% of the integral mass. *Right:* total gradient evaluations. Small ν approaches the hard-constraint limit and lengthens the ladder, whereas large ν lowers the cost but becomes tempering-like and misses the spike.

The two-dimensional case is the least mobile, and the near-boundary to overall ratio increases from 0.31 at $D = 2$ to 0.65 at $D = 3$ before remaining of order one half at larger dimensions. Thus the soft boundary slows a small fraction of walkers rather than creating a frozen layer, and the effect does not worsen toward zero mobility as dimension increases.

The local Gamma approximation in Eq. (21) predicts that 2.31% of an equilibrated $\nu = 2$ population lies below this slack threshold. The larger measured fraction reflects finite relaxation after each level change: lowering E reduces every particle's slack simultaneously, after which the mutation kernel moves the population back into the interior.

D	near the wall	moves there	moves overall	ratio
2	3.6%	0.126	0.403	0.31
3	3.2%	0.282	0.434	0.65
5	3.5%	0.327	0.464	0.70
10	4.3%	0.366	0.497	0.74
50	6.1%	0.286	0.554	0.52
1000	7.1%	0.289	0.598	0.48

Table 6: Mobility near the level-set boundary on the isotropic Gaussian, with $N = 200$, $n_{\text{steps}} = 10$, and $\nu = 2$. "Near the wall" denotes slack below one fifth of the level mean. The two move columns report the fraction of proposals that change the particle position, measured directly from the trajectories; the final column is their ratio.

B.1 Kernels under the hard constraint

The dominant paradigm for extending nested sampling to high dimensions in general problems has been to employ reflective Hamiltonian dynamics (Skilling, 2019; Lemos et al., 2024). However, these methods have been unable to meaningfully change the scaling of the method and have been shown to be biased in high dimensions (Kroupa et al., 2025). More broadly, many recent developments in scalable and efficient MCMC build on gradient information and do not transfer cleanly to hard-constrained targets. Softening the constraint makes innovations in gradient-based MCMC kernels such as the Barker proposal MCMC (Livingstone & Zanella, 2022) and isokinetic samplers (Robnik et al., 2023) available to the quenched path.

The ability to utilise gradient based samplers has a clear advantage in high-dimensional spaces. Under standard assumptions, random-walk methods exhibit $\mathcal{O}(D)$ scaling (Roberts & Rosenthal, 2016), whereas Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Neal, 2011) and the Metropolis-adjusted Langevin algorithm (Roberts & Tweedie, 1996; Roberts & Rosenthal, 1998)

can achieve $\mathcal{O}(D^{1/4})$ and $\mathcal{O}(D^{1/3})$ scaling, respectively (Roberts & Rosenthal, 2016; Beskos et al., 2013).

C ACCOUNTING AND POSTERIOR RECONSTRUCTION

The normalizing constant estimate and posterior samples are the two primary outputs of the sampling process. How these are computed is a product of how the algorithm implements the sequential outer steps, in this section we define the choices that are distinct to QES, contrasted against typical implementations of both nested sampling and SMC. The central distinction is between quantities that are *measured* from the weighted population and operations that only change how that population is represented.

C.1 ASSIGNED AND MEASURED LEVEL VOLUMES

Both nested sampling and SMC estimate the normalising constant from a series of incremental ratios along a path. Classic nested sampling assigns the compression $t_k = G_k/G_{k-1}$ from an order-statistic law: if the m live points are independent draws from the constrained prior, their enclosed masses are independent $\mathrm{Uniform}[0,1]$ variables, and discarding the worst point gives $t_k \sim \mathrm{Beta}(m,1)$. The energy thresholds are measured, but the volumes are estimated from this law (Skilling, 2006; Chopin & Robert, 2010).

QES instead measures every soft-volume increment from the population present at the level, following a more standard SMC approach. For particles with normalised weights w_j targeting ρ_{E_k}, the pointwise ratio r of Eq. (10), evaluated for the pair (E_k, E_{k+1}), gives

$$\widehat{r}_k = \sum_{j=1}^{N} w_j\, r(x_j), \qquad \widehat{G}_{\nu,k+1} = \widehat{G}_{\nu,k}\widehat{r}_k, \tag{22}$$

where the ratio is evaluated before mutation. The ratio is of unnormalised densities, the normalised ratio $\rho_{E_{k+1}}/\rho_{E_k}$ has expectation one under ρ_{E_k}, since the normalisation is the quantity being estimated. This is the standard SMC normalising constant construction applied to the soft level family (Salomone et al., 2025).

This distinction changes the assumptions and resulting guarantees of the sampling process. For a fixed, externally specified schedule, exact incremental weights and mutation kernels invariant for their current targets give an unbiased normalising constant estimator for any particle count and any amount of mutation (Del Moral, 2004). Poor mutation results in increased variance, without adding a mixing bias to that estimator. Normalised posterior expectations and $\log \widehat{Z}$ still have their usual finite-particle bias. The classic nested sampling assignment of volumes carries a stricter requirement, namely that the live particles must also be mutually independent. A replacement chain that remains correlated with its surviving start violates the order-statistic premise, so the assigned compressions need not describe the realised population. In practice the bias is often small, but practical implementations of nested sampling often use relatively long mutation chains to keep this in check. As a result, when it comes to efficient scaling with a matched comparison to tempered SMC, as in this work, we find the more standard SMC accounting to be more appropriate.

Killing a particle at a new level gives it incremental weight zero, so its mass remains in the denominator of Eq. (22) and is automatically banked in $\widehat{G}_{\nu,k} - \widehat{G}_{\nu,k+1}$. Replacing a particle of weight w by two coincident copies of weight $w/2$ also leaves the empirical measure exactly unchanged,

$$\frac{w}{2}\delta_x + \frac{w}{2}\delta_x = w\delta_x. \tag{23}$$

Mutation subsequently moves the copies apart. We find that while the population diversity is sufficiently high, the simple particle splitting is efficient, and the systematic resampling step is only triggered when the ESS of the carried weights falls below a threshold. This is discussed in more detail in Appendix C.4.

The marginal likelihood quadrature. The integral of Eq. (9) is evaluated from the stored descending ladder $(E_k, \widehat{G}_\nu(E_k))$: between adjacent levels $\log \widehat{G}_\nu$ is taken to vary linearly in E, and

the contributions are accumulated in log space. Taking into account the chosen termination, the contribution of the final particles once the criteria is reached can be appended to the ladder, however in practice this additional contribution is often negligible (1.3×10^{-5} nats at the termination of the $D = 10^4$ Gaussian ladder). Computing Z by numerical quadrature over the ladder is standard practice in nested sampling (Skilling, 2006), unlike in tempered SMC, the quenched path never targets the canonical ensemble at $\beta = 1$, so the marginal likelihood is carried by the path as a whole rather than read off any terminal target. The quadrature introduces some discretisation error, however due to the adaptive nature of the schedule, and the fine resolution of the quenched ladder that results, the quadrature error is generally far below the finite-N effects that dominate the error budget (3×10^{-6} nats when the exact integrand is evaluated on the ladder of the $D = 10^4$ Gaussian example in Appendix C.5).

C.2 Adaptive levels and finite-particle bias

The next energy level is selected from the current population by solving for the value of $E_{k+1} < E_k$ at which the prospective incremental weights reach the target ESS. This is an adaptive schedule: it avoids specifying the unknown shape of $E(G)$ in advance and automatically shortens the step when the population would otherwise lose too much support. Adaptive SMC estimators remain consistent and satisfy a central limit theorem (Beskos et al., 2016). The method therefore preserves the practical adaptivity of nested sampling without claiming exact finite-N unbiasedness; we demonstrate that the observed bias is dominated by the finite-N effects in Appendix C.5.

This level-selection ESS is distinct from the second ESS used to decide when to redraw the population. The first sets the distance between adjacent targets; the second controls accumulated weight degeneracy across many such steps. Conflating them would force a full resampling at every rung of a path whose length grows rapidly with dimension, which leads to an unnecessary particle degeneracy.

C.3 Pooling the path into a posterior sample

In order to reconstruct the posterior, samples from the entire path are combined. Whilst this is possible in tempered SMC (Le Thu Nguyen et al., 2014), it is a necessity for the quenched path as none of the series of distributions are the canonical Boltzmann distribution at $\beta = 1$. From Eq. (9), level k carries quadrature weight

$$b_k \propto \widehat{G}_\nu(E_k) e^{-E_k} \Delta E_k. \tag{24}$$

Here ΔE_k is the positive finite-difference width computed from the stored descending ladder. If w_{kj} is particle j's normalised carried weight at that level, then the pooled posterior weight is $q_{kj} \propto b_k w_{kj}$. We report the Kish quantity

$$N_{\mathrm{Kish}} = \frac{\left(\sum_{k,j} q_{kj}\right)^2}{\sum_{k,j} q_{kj}^2} \tag{25}$$

over every particle of every rung. For tempered SMC, let $\beta_1, \ldots, \beta_K$ denote the recorded post-mutation stages. Because each stage has the same population size, a particle at stage β_k is reweighted against the equally weighted deterministic mixture of stages,

$$q_{kj} \propto \frac{e^{-U(x_{kj})}}{\sum_{\ell=1}^{K} \widehat{Z}_{\beta_\ell}^{-1} e^{-\beta_\ell U(x_{kj})}}, \tag{26}$$

where $\widehat{Z}_{\beta_\ell}$ is the normalising-constant estimate accumulated by that SMC run. The prior density and the common mixture factor cancel from this ratio. This means both methods are scored in efficiency using all particles in all recorded clouds. We form this complete weight vector and evaluate Eq. (25) once per run; ESS per gradient is computed for that run before results are averaged over repeats.

For the particle methods, pooled Kish ESS measures the concentration of this full posterior-weight vector across particles and levels. It does not account for shared ancestry, mutation correlation or dependence between neighbouring levels, and is therefore a nominal rather than correlation-adjusted sample size. For NUTS and MCLMC, bulk ESS accounts for within-chain correlation. Both diagnostics indicate how efficiently computation is converted into posterior samples and are qualitatively comparable, but their values should not be interpreted as exact ratios of independent draws.

The extensive path length of QES. The large QES pool follows directly from the geometry of the energy path. For the isotropic Gaussian, let $\pi = \mathcal{N}(0, I_D)$ and $U(x) = \|x\|^2/(2\sigma^2)$. The posterior is $\mathcal{N}(0, \sigma_\star^2 I_D)$ with $\sigma_\star^2 = \sigma^2/(1+\sigma^2)$. Under the exact augmentation, $E = U + s$, where $s \sim \mathrm{Gamma}(\nu+1, 1)$ is independent of x, so

$$\mathrm{Var}(E) = \frac{D}{2(1+\sigma^2)^2} + \nu + 1. \tag{27}$$

The posterior-weighted range of levels therefore has width $\Theta(\sqrt{D})$.

At a typical level the local form $g(E-s) \simeq g(E)e^{-\beta_E s}$ implies that the moments of the incremental weight depend on the level drop through $\beta_E \Delta E$. A fixed incremental ESS consequently fixes $\beta_E \Delta E$; for this target $\beta_E \to 1$, so $\Delta E = \Theta(1)$. The full QES path spans an $\Theta(D)$ energy range and therefore contains $\Theta(D)$ rungs, of which $\Theta(\sqrt{D})$ carry posterior weight.

Tempering has the complementary scaling. Its incremental ESS fixes $\Delta\beta = \Theta(D^{-1/2})$, giving $\Theta(\sqrt{D})$ stages across $\beta \in [0,1]$. A stage can be reweighted to the posterior without weight collapse only when $1-\beta = O(D^{-1/2})$, so only $\Theta(1)$ stages carry posterior weight. Hence, before accounting for within-level weight variation, the two pooled Kish counts scale as

$$N_{\mathrm{Kish}}^{\mathrm{QES}} = \Theta(N\sqrt{D}), \qquad N_{\mathrm{Kish}}^{\mathrm{SMC}} = \Theta(N). \tag{28}$$

This is a path-level sample-size advantage, not a free computational advantage: QES also uses $\sqrt{D}$ more stages, cancelling the additional $\sqrt{D}$ in its pooled Kish. At fixed mutation cost per stage, both methods therefore scale as $\Theta(D^{-1/2})$ Kish per gradient. Under the benchmark budget $n_{\mathrm{steps}} = \Theta(\sqrt{D})$, both acquire the same additional $D^{-1/2}$ cost and scale as $\Theta(D^{-1})$.

Table 7 verifies these counts. Across $D = 10$ to 10^4, the number of QES rungs carrying posterior weight grows from 20.0 to 674.9, while the tempered count remains between 9.5 and 9.9. The full QES ladder grows from 43 to 14,493 rungs, whereas the tempered path grows from 16 to 502 stages. The number of ladder rungs "carrying weight" is the effective sample size of the posterior ladder weights.

Table 7: Path length and posterior-weight-carrying stages on the isotropic Gaussian, with $N = 1000$, three seeds and incremental ESS target 0.95. The "carrying weight" column shows the effective sample size of the posterior ladder weights. For QES this count grows as $\sqrt{D}$ while the full ladder grows as D; for tempered SMC the full path grows as $\sqrt{D}$ while the weight-carrying count remains bounded.

	QES		tempered SMC	
D	rungs	carrying weight	stages	carrying weight
10	43	20.0	16	9.9
50	139	48.8	35	9.5
100	238	68.9	50	9.7
200	416	93.7	71	9.8
500	892	144.4	112	9.5
1000	1684	217.3	158	9.7
2000	3144	303.6	224	9.5
5000	7446	488.6	354	9.6
10000	14493	674.9	502	9.9

C.4 Resampling cadence

The long QES ladder makes resampling cadence more of an issue, particularly for multimodal targets. Redrawing at every rung gives fresh uniform weights, but it also repeatedly resamples neutral mode labels (Koskela et al., 2020). On a symmetric multimodal target with no local transitions, occupancy then performs a random walk with absorbing states at zero and one. At $D = 100$, four of ten runs fixate during the approximately 1,800 per-level redraws, even though the measured death rates are identical in the two modes (0.0040 in each).

At the other extreme, branching can replace zero-weight particles while all survivor weights are carried indefinitely. This preserves the represented measure, but eventually concentrates it on few ancestral lineages, increasing variance and the finite-N bias of nonlinear summaries such as $\log \widehat{Z}$. We therefore branch at every level and fully resample only when the carried-weight ESS falls below fN (Moral et al., 2012). The trigger is computed after grouping weights by lineage, splitting a donor into two children halves its contribution to the ordinary sum of squared weights and would spuriously increase the ungrouped ESS, although it has created no new independent lineage. With the default $f = 0.5$, full resampling occurs about every 14 levels at an incremental ESS target of 0.95, roughly fourteen times less often than per-level resampling (measured across the analytic problems).

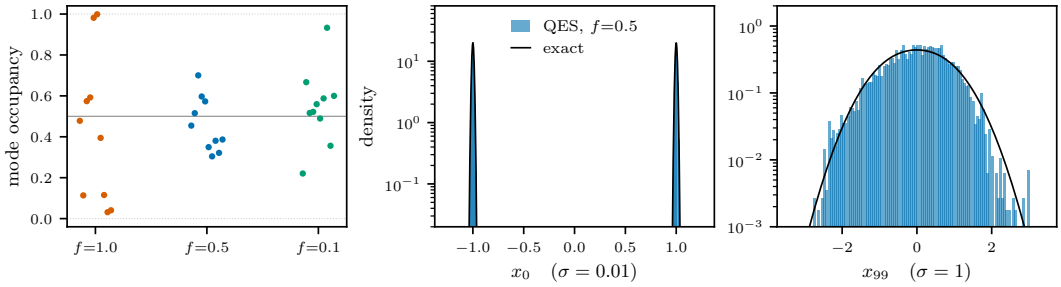


Figure 6: *Left:* mode occupancy for every run at each resampling threshold, ten seeds per setting on the bimodal target at $D = 100$; a fixated run lies on a dotted boundary. Per-level resampling ($f = 1$) drives neutral occupancy toward fixation, whereas the default $f = 0.5$ preserves both modes in every run. *Middle, right:* posterior marginals under the default against the exact density at the two ends of the anisotropy, $\sigma = 0.01$ and $\sigma = 1$.

Figure 6 illustrates this trade-off on the bimodal target at $D = 100$. Resampling every level ($f = 1.0$) causes the occupancy to collapse into one mode more often than the default ($f = 0.5$), while the default cadence preserves both modes in every run and reconstructs the marginals at both ends of the anisotropy. Resampling less often than the default lets the occupancy drift back toward imbalance, with one of ten runs approaching fixation. Accounting for this drift is a challenge for any interacting particle method, and motivates global moves that cross between modes; constructing such moves is itself hard in high dimension without an *a priori* known symmetry.

C.5 Mutation budget

The residual Gaussian bias at $D = 10^4$ (Table 3) is probed by sweeping the number of MALA steps per level around the working point of $n_{\text{steps}} = 128$, with every other setting of the benchmark held fixed. Below the default the bias grows steadily as the population lags its moving target, while doubling the budget changes nothing within the seed variation (Table 8). The default therefore sits at the knee of the curve, and we attribute the residual that remains there to the accumulated finite-N effects of Appendix C rather than to incomplete mutation. Doubling the particle count at the default budget instead shrinks the bias and contracts the seed scatter by the predicted $\sqrt{2}$, consistent with this attribution. Figure 7 shows the reconstructed radial posterior at the default working point.

Table 8: Budget sweeps on the Gaussian at $D = 10^4$: $\log Z$ bias against the analytic value and gradient evaluations per run, ten seeds per setting. The default is $N = 1000$, $n_{\text{steps}} = 128$; the final row doubles the particle count at the default mutation budget.

N	n_{steps}	gradients	$\Delta \log Z$
1000	32	0.47×10^9	-1.91 ± 0.68
1000	64	0.94×10^9	-1.62 ± 0.93
1000	128	1.86×10^9	-0.89 ± 0.71
1000	256	3.69×10^9	-1.10 ± 0.77
2000	128	3.74×10^9	-0.62 ± 0.45

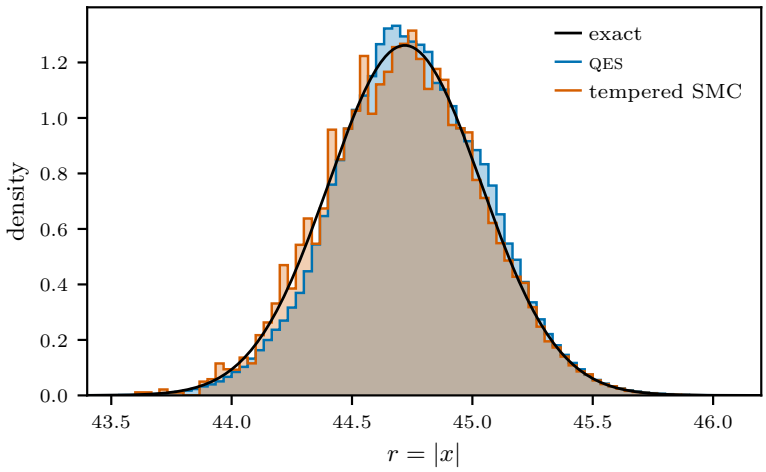


Figure 7: Reconstructed radial posterior at $D = 10^4$ against the exact shell $r = \tau\chi_D$, for a single run at the default working point. Both methods pool every weighted particle from every level of the path: QES with the level-volume quadrature weights of Appendix C.3, tempered SMC with deterministic-mixture balance weights over its temperatures. The pooled Kish ESS is 4.8×10^5 for QES and 6.9×10^3 for tempered SMC.