

Thermodynamic Bayesian Inference

Maxwell Aifer, Samuel Duffield, Kaelan Donatella, Denis Melanson, Phoebe Klett,
Zach Belateche, Gavin Crooks, Antonio J. Martinez, Patrick J. Coles
Normal Computing Corporation, New York, New York, USA

Abstract—A fully Bayesian treatment of complicated predictive models (such as deep neural networks) would enable rigorous uncertainty quantification and the automation of higher-level tasks including model selection. However, the intractability of sampling Bayesian posteriors over many parameters inhibits the use of Bayesian methods where they are most needed. Thermodynamic computing has emerged as a paradigm for accelerating operations used in machine learning, such as matrix inversion, and is based on the mapping of Langevin equations to the dynamics of noisy physical systems. Hence, it is natural to consider the implementation of Langevin sampling algorithms on thermodynamic devices. In this work we propose electronic analog devices that sample from Bayesian posteriors by realizing Langevin dynamics physically. Circuit designs are given for sampling the posterior of a Gaussian-Gaussian model and for Bayesian logistic regression, and are validated by simulations. It is shown, under reasonable assumptions, that the Bayesian posteriors for these models can be sampled in time scaling with $\ln(d)$, where d is dimension. For the Gaussian-Gaussian model, the energy cost is shown to scale with $d \ln(d)$. These results highlight the potential for fast, energy-efficient Bayesian inference using thermodynamic computing.

1. Introduction

Bayesian statistics has proved an effective framework for making predictions under uncertainty [1], [2], [3], [4], [5], [6], and it is central to proposals for automating machine learning [7]. Bayesian methods enable uncertainty quantification by incorporating prior knowledge and modeling a distribution over the parameters of interest. Popular machine learning methods that employ this approach include Bayesian linear and non-linear regression [8], Kalman filters [9], Thompson sampling [2], continual learning [10], [11], and Bayesian neural networks [3], [12].

Unfortunately, computing the posterior distribution in these settings is often intractable [13]. Methods such as the Laplace approximation [14] and variational inference [15] may be used to approximate the posterior in these cases, however their accuracy struggles for complicated posteriors, such as those of a Bayesian neural network [13]. Regardless, sampling accurately from such posteriors requires enormous computing resources [13].

Computational bottlenecks in Bayesian inference motivate the need for novel hardware accelerators. Physics-

based sampling hardware has been proposed for this purpose, including Ising machines [16], [17], [18], [19], [20], probabilistic bit computers [21], [22], [23], and thermodynamic computers [24], [25], [26], [27], [28], [29], [30], [31], [32], [33]. Continuous-variable hardware is particularly suited to Bayesian inference since continuous distributions are typically used in probabilistic machine learning [27]. However, a rigorous treatment of how such hardware can perform Bayesian inference with scalable circuits has not yet been given.

The most computationally tractable algorithms for exact Bayesian inference are Monte Carlo sampling algorithms. The Langevin sampling algorithm [34], [35] is an elegant example inspired by statistical physics, based on the dynamics of a damped system in contact with a heat bath. What we propose in this work is to build a physical realization of the system that is simulated by the Langevin algorithm. The system must be designed to have a potential energy such that the Gibbs distribution $p(x) \propto e^{-\beta U(x)}$ is the desired posterior distribution which is reached at thermodynamic equilibrium. We present circuit schematics for electronic implementations of such devices for Bayesian inference for two special cases. The first is a Gaussian-Gaussian model (where the prior and the likelihood are both multivariate normal, as found in linear regression and Kalman filtering), and the second is logistic regression (where the prior is Gaussian and the likelihood is Bernoulli parameterized by a logistic function). In each case, the parameters of the prior and likelihood are encoded in the values of components of the circuit, and then voltages or currents are measured to sample the random variable.

While thermodynamic algorithms have been proposed for linear algebra [29] and neural network training [33], our work can be viewed as the first thermodynamic algorithm for sampling from Bayesian posteriors. Moreover, our work provides the first concrete proposal for non-Gaussian sampling with thermodynamic hardware. Overall, our work opens up a new field of rigorous Bayesian inference with thermodynamic computers and lays the groundwork for scalable CMOS-based chips for probabilistic machine learning.

We show that in theory the devices proposed for sampling the Gaussian-Gaussian model and logistic regression posteriors can obtain N samples in d dimensions in time scaling with $O(N \ln d)$. This is a significant speedup over typical methods used digitally for the same problems; for example sampling the Gaussian-Gaussian posterior digitally

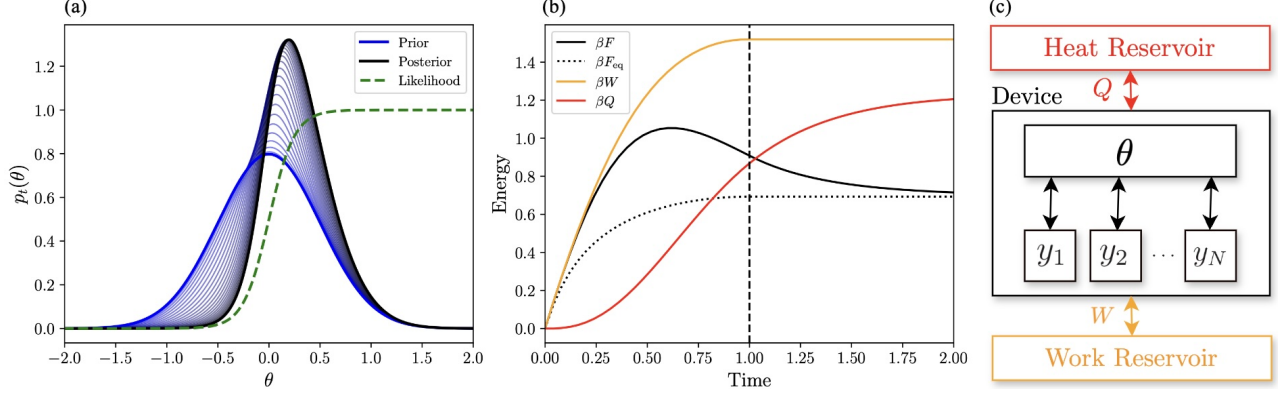


Figure 1. Overview of Thermodynamic Bayesian Inference. (a) Time-evolution of the probability density $p_t(\theta)$ under the Fokker-Planck equation for a time-dependent potential energy $U_t(\theta)$. Initially the potential corresponds to the prior $\beta U_0(\theta) = -\ln p_\theta(\theta)$, and a logistic likelihood term is gradually introduced via a quadratic ramp-up until $t = 1$, with $U_1(\theta) = -\ln p_\theta(\theta) - \ln p_{y|\theta}(y|\theta)$, after which the potential does not change. The probability density is initially a Gaussian prior (blue) and approaches the Bayesian posterior (black) over time. This is the posterior for logistic regression with a single data point. (b) Thermodynamic quantities during Fokker-Planck evolution. The free energy $\beta F = \beta \langle U \rangle - S$ (solid black) is computed at all times as well as F_{eq} (dotted black), the equilibrium free energy associated with potential U_t . Their difference $F - F_{\text{eq}}$ can be seen as the degree to which the system is out of equilibrium, and equals the KL divergence to the equilibrium distribution. The latter quantity approaching zero at the end of the protocol signifies that $p_t(\theta)$ approaches the true posterior. The convention for the signs of work and heat is $dE = dW - dQ$. The change in free energy over time (dotted black) lower bounds the work (orange), and the gap between the two is the dissipated work. Heat continues to flow after time $t = 1$ (the end of the control protocol), while work is done only for $t < 1$. (c) Thermodynamic interpretation of Bayesian inference protocol. The parameter θ and the data points $y_1 \dots y_N$ are physical degrees of freedom, where $y_1 \dots y_N$ are fixed and θ is free to vary. Initially θ is decoupled from $y_1 \dots y_N$, but a coupling is turned on which requires work to be done by a work reservoir. The system exchanges heat with a heat reservoir, coming to thermal equilibrium in the posterior distribution $p_{\theta|y}$.

involves matrix inversions taking time scaling with $O(d^\omega)$ where $2 < \omega < 3$. This speedup is larger than the linear (in dimension) speedups found in previous work on thermodynamic algorithms for linear algebra primitives [29], where the goal was simply to accelerate standard computations, while not fundamentally changing which problems are considered tractable. In contrast, the more significant speedups found in this work have the potential to make computations possible which previously were not. This is particularly true for the sampling of non-Gaussian posteriors, such as the case of Bayesian logistic regression.

2. Thermodynamic Bayesian Inference

Suppose that we have samples of a random vector y , and would like to estimate a random vector θ on which y depends somehow. The Bayesian approach is to assume a prior distribution on θ given by a density function $p_\theta(\theta)$, and a likelihood function $p_{y|\theta}(y|\theta)$. The posterior distribution for θ is then given by Bayes's theorem $p_{\theta|y}(\theta|y) = p_{y|\theta}(y|\theta)p_\theta(\theta)/p_y(y)$. To sample from the posterior using the Langevin algorithm, one first computes the score

$$\nabla_\theta \ln p_{\theta|y}(\theta|y) = \nabla_\theta \ln p_{y|\theta}(y|\theta) + \nabla_\theta \ln p_\theta(\theta). \quad (1)$$

Then the score is used as the drift term in the following stochastic differential equation (SDE)

$$d\theta = \nabla_\theta \ln p_{\theta|y}(\theta|y) dt_c + \mathcal{N}[0, 2 dt_c], \quad (2)$$

where t_c denotes a dimensionless time coordinate. After this SDE is evolved for a sufficient time T , the value of θ will be

a sample from $p_{\theta|y}$. This algorithm is equivalent to the equilibration of an overdamped system, as we will now describe. First let r be a vector of the same dimension as θ describing the state of a physical system, and satisfying $r = \theta \tilde{r}$ for some constant \tilde{r} (this factor is necessary because θ is unitless while the physical quantity r has units). Now we define the potential energy function $\beta U(r) = -\ln p_{\theta|y}(r/\tilde{r} | y)$. The dynamics of an overdamped system with potential energy U in contact with a heat bath at inverse temperature β can be modeled by the overdamped Langevin equation

$$dr = -\gamma^{-1} \nabla_r U(r) dt + \mathcal{N}[0, 2\gamma^{-1} \beta^{-1} dt], \quad (3)$$

where γ is a damping constant and t is a physical time coordinate (i.e. having units of time). Note that this implies that γ has dimensions of energy \cdot time/ $[r]^2$. If we introduce a time constant $\tau = \gamma \beta \tilde{r}^2$, Eq. (3) can be written

$$d\theta = \nabla_\theta \ln p_{\theta|y}(\theta|y) \tau^{-1} dt + \mathcal{N}[0, 2 \tau^{-1} dt], \quad (4)$$

which has the same form as Eq. (2), except with the time constant τ . It is clear that if Eq. (2) must be run for a dimensionless duration T to achieve convergence, then the physical system must be allowed to evolve for a physical time duration τT to achieve the same result. While we have addressed the case of conditioning on a single sample y above, the generalization of these ideas to the case of conditioning on multiple I.I.D. samples is given in Appendix F. In what follows we will present designs for circuits whose potential energy results in an overdamped Langevin equation that yields samples from Bayesian posteriors.

The process just described is illustrated in Figure 1, for the example of Bayesian logistic regression with a single

data point and a Gaussian prior. In this case $p_\theta(\theta) \propto \frac{1}{2}\sigma^{-2}\theta^2$ and $p_{y|\theta}(y|\theta) = 1/(1 + e^{-\theta xy})$, where $y \in \{-1, 1\}$ and x is a constant representing a dependent variable. In order to solve a problem using a physical device, we first must map the parameters of the problem onto physical properties of the device, and this is modeled by a control protocol which changes the potential energy over time, specifically¹

$$U_t(\theta) = -\ln p_\theta(\theta) - \lambda(t)p_{y|\theta}(y|\theta). \quad (5)$$

The control parameter λ is smoothly transitioned from zero to one using a quadratic ramp-up

$$\lambda(t) = \begin{cases} -t(t-2) & 0 \leq t < 1 \\ 1 & t \leq 1. \end{cases} \quad (6)$$

The probability density $p_t(\theta)$ evolves according to the Fokker-Planck equation [36] (see Appendix A)

$$\dot{p}_t = \beta \nabla \cdot (p_t \nabla U_t) + \nabla^2 p_t, \quad (7)$$

with initial condition set by the prior, $p_0 = p_\theta$. The solution to this initial value problem is shown in Fig. 1 (a), and we see the smooth interpolation between the prior (blue) at time $t = 0$ and the posterior (black) at time $t = 2$.

In theory, the energetic cost of implementing the algorithm is entirely due to the work done during the control protocol for $0 \leq t < 1$, as the heat comes from a reservoir which may be taken to be available naturally in the environment.² The work and heat are quantified using the formulas [37]

$$dW = \int d\theta p_t(\theta) dU_t(\theta), \quad dQ = - \int d\theta U_t(\theta) dp_t(\theta), \quad (8)$$

so the first law $dE = dW - dQ$ is satisfied. The equilibrium free energy F_{eq} is defined as the free energy the system would have it were in equilibrium with the potential $U_t(\theta)$, and is given as

$$\beta F_{\text{eq}}(t) = -\ln \left(\int d\theta e^{-\beta U_t(\theta)} \right). \quad (9)$$

We also may compute the non-equilibrium free energy $\beta F = \beta E - S$. As described in Appendix A, the difference between F and F_{eq} is simply the KL divergence to the equilibrium distribution [38]

$$\beta(F - F_{\text{eq}}) = \text{KL}(p||p_{\text{eq}}), \quad (10)$$

where $p_{\text{eq}}(\theta) \propto e^{-\beta U(\theta)}$ is the Boltzmann distribution and $\text{KL}(p||q) = \int d\theta p(\theta) \ln p(\theta)/q(\theta)$.

In Fig. 1 (b), we see that initially $F = F_{\text{eq}} = 0$, so the system is at equilibrium in the Gaussian prior distribution. However, as the potential changes over time, the system is brought out of equilibrium, and F increases faster than F_{eq} .

1. Note that we abuse notation by writing $U_t(\theta)$ instead of $U_t(r)$ for simplicity of presentation.

2. In practice, however, the reservoir will be implemented as a random number generator in our proposed devices, and so the power consumed by the random number generators must be included in estimates of energy cost.

For $t \geq 1$ the potential is constant and so the system begins to come back to equilibrium, which can be seen from the fact that F approaches F_{eq} for $t \geq 1$. In fact, the approach to equilibrium can be interpreted using the framework of Wasserstein gradient flows (see Appendix A). In general, Eq. (7) is equivalent to a gradient descent in the space of probability distributions, where the objective function is $F - F_{\text{eq}}$ and the step size is measured using the Wasserstein 2 metric.

It is important to note the significance of the gap between F and W in Fig. 1, which represents dissipated work [39]. For an idealized device run in the quasistatic limit (ie with a very slowly changing potential), we would have $F = W = F_{\text{eq}}$ at all times, so the dissipation would be zero and the system would always be at equilibrium. In this case, it would be possible to reverse the protocol and recover an amount of work equal to the work spent during the forward protocol. However, if dissipated work is nonzero, when the protocol is reversed the dissipated work cannot be recovered. Therefore the dissipated work can be seen as the fundamental lower limit on the amount of energy needed to carry out Bayesian inference using this method. This results in a tradeoff between energy and time cost; the dissipated work (and thus the energy cost) can be reduced at the cost of increasing the protocol's time duration.

In Fig 1 (c), the protocol is interpreted as the operation of a kind of thermodynamic machine. The parameter θ and data $y_1 \dots y_N$ are encoded in physical degrees of freedom of a system, with θ allowed to vary and $y_1 \dots y_N$ fixed. Work is drawn from a reservoir (i.e., a battery) in order to modulate control parameters that couple the θ subsystem to the subsystems $y_1 \dots y_N$. Heat is exchanged with a heat reservoir, bringing the system towards equilibrium. It follows from the non-negativity of mutual information that the entropy of a Bayesian posterior is (on average) less than or equal to the entropy of the prior, which intuitively means that our certainty increases as we gather more data. For example, in the case of logistic regression, we can see that the entropy of the system decreases as the distribution becomes more sharply peaked (see Fig. 1 (a)). Therefore this machine can be seen as an ‘‘entropy pump’’, requiring work in order to reduce the entropy of a system while dissipating heat to its environment. Interestingly, it is hypothesized that a similar entropy pumping mechanism is integral to the maintenance of homeostasis in biological systems [40].

2.1. Gaussian-Gaussian model

A particularly simple special case of Bayesian inference is a when both the prior and the likelihood are multivariate normal, and we address this simple model first in order to illustrate our approach more clearly. Specifically, let $\theta \in \mathbb{R}^d$ have prior distribution $p_\theta(\theta) = \mathcal{N}[\mu_\pi, \Sigma_\pi]$, and let the likelihood be $p_{y|\theta}(y|\theta) = \mathcal{N}[\theta, \Sigma_\ell]$, where $y \in \mathbb{R}^d$ is an observed sample. In this case the posterior $p_{\theta|y}$ is also multivariate normal, with parameters [12]

$$\mu_{\theta|y} = \mu_\pi + \Sigma_\pi (\Sigma_\pi + \Sigma_\ell)^{-1} (y - \mu_\pi), \quad (11)$$

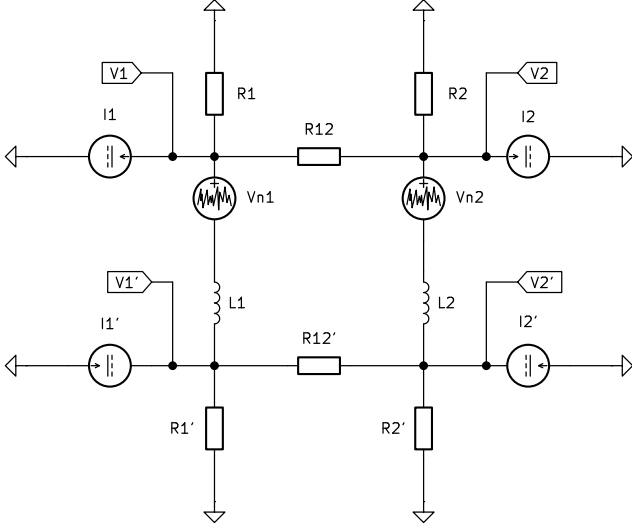


Figure 2. **Circuit schematic for the Gaussian-Gaussian model posterior sampling device.** Two resistor networks are coupled via inductors, with the currents through the inductors given by (16).

$$\Sigma_{\theta|y} = \hat{\Sigma}_{\pi} - \Sigma_{\pi}(\Sigma_{\pi} + \Sigma_{\ell})^{-1}\Sigma_{\pi}. \quad (12)$$

For this model, the posterior is tractable and can be computed on digital computers relatively efficiently, however for very large dimensions the necessary matrix inversion and matrix-matrix multiplications can still create a costly computational bottleneck. As we will see, the thermodynamic approach provides a means to avoid the costly inversion and matrix products in the computation, and therefore to accelerate Bayesian inference for this model.

We begin by deriving the Langevin equation for sampling this posterior. For this prior and likelihood, the score of the posterior Eq. (1) is

$$\nabla_{\theta} \ln p_{\theta|y}(\theta|y) = -\Sigma_{\pi}^{-1}(\theta - \mu_{\pi}) - \Sigma_{\ell}^{-1}(\theta - y), \quad (13)$$

and so Eq. (4) becomes

$$d\theta = -\Sigma_{\pi}^{-1}(\theta - \mu_{\pi})\tau^{-1}dt - \Sigma_{\ell}^{-1}(\theta - y)\tau^{-1}dt + \mathcal{N}[0, 2\mathbb{I}\tau^{-1}dt]. \quad (14)$$

In fact, this SDE can be implemented by a circuit consisting of two resistor networks coupled by inductors, shown in Fig. 2 for the two-dimensional case.

The full analysis of the circuit in Fig. 2 is given in Appendix B, but a few remarks are made here to explain its operation. First, we define the conductance matrices \mathcal{G} as

$$\mathcal{G} = \begin{pmatrix} R_{11}^{-1} + R_{12}^{-1} & -R_{12}^{-1} \\ -R_{12}^{-1} & R_{22} + R_{12}^{-1} \end{pmatrix}, \quad (15)$$

and \mathcal{G}' is defined in the same way for the primed resistors $R'_1, R'_2,$ and R'_{12} . By applying Kirchoff's current law (KCL), the voltages across the resistors can be eliminated. Then the

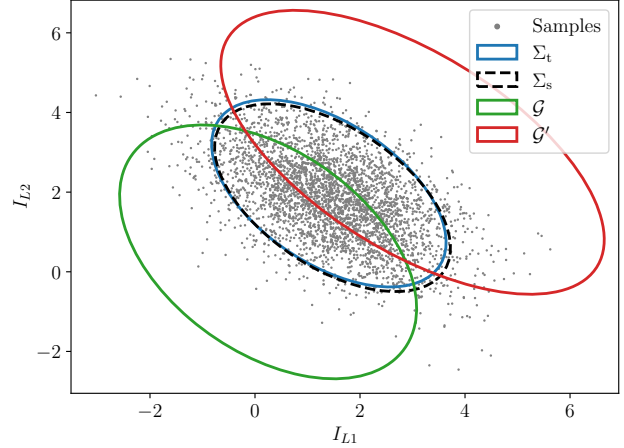


Figure 3. **SPICE simulations of proposed Gaussian-Gaussian circuit.** The gray points represent the simulated circuit's sampled inductor currents (normalized). The dashed black and solid blue ellipses represent the empirical sample covariance matrix and the target posterior covariance matrix from a Gaussian Bayesian update, respectively. The red and green ellipses represent the prior and likelihood covariance matrices, respectively.

equation $V = LI$ is used to derive the following stochastic differential equation for the currents through the inductors

$$dI_L = -L^{-1}\mathcal{G}^{-1}(I_L - I)dt - L^{-1}\mathcal{G}'^{-1}(I_L - I')dt + L^{-1}\sqrt{S}\mathcal{N}[0, \mathbb{I}dt], \quad (16)$$

where $I_L = (I_{L1} \ I_{L2})^T$ and S is the power spectral density of each noise source. This equation has the same form as Eq. (14), so it is only necessary to determine an appropriate mapping of distributional parameters to physical properties of the circuit's components (see Appendix B). We verify the behavior of this circuit by running SPICE simulations. The results are shown in Fig. 3 and discussed in greater detail in Section 4.

By including more inductors and coupling resistors (as well as current and voltage sources), the design can be generalized to arbitrary dimension. We note that because of the negative sign in the off-diagonal elements of Eq. (15), this specific architecture can only implement matrices with negative off-diagonal elements since passive components cannot achieve a negative conductance. This limitation can be overcome by modifying the architecture and including inductive transformers to change the polarity of the interaction [28] or a differential design where symmetry can be exploited to change the direction of the interaction.

The energy and time costs of this algorithm are analyzed in Appendix C and presented in Section 3. Numerical simulations are provided in Section 4.

2.2. Bayesian linear regression and Kalman filtering

A generalization of the Gaussian-Gaussian model is that of Bayesian linear regression [8] (or equivalently a Kalman

filter update step [9], [12]). In full generality we have

$$p_{\theta}(\theta) = \mathcal{N}[\mu_{\pi}, \Sigma_{\pi}], \quad (17)$$

$$p_{y|\theta}(y | \theta) = \mathcal{N}[H\theta, \Sigma_{\ell}], \quad (18)$$

Then the overdamped Langevin SDE becomes

$$\begin{aligned} d\theta &= -\Sigma_{\pi}^{-1}(\theta - \mu_{\pi})\tau^{-1} dt - H^{\top}\Sigma_{\ell}^{-1}(y - H\theta)\tau^{-1} dt \\ &\quad + \mathcal{N}[0, 2\mathbb{I}\tau^{-1}dt], \\ &= -(A\theta - b)\tau^{-1} + \mathcal{N}[0, 2\mathbb{I}\tau^{-1}dt], \quad (19) \\ &\quad \text{for } A = \Sigma_{\pi}^{-1} + H^{\top}\Sigma_{\ell}^{-1}H \text{ and } b = \mu_{\pi} + H^{\top}\Sigma_{\ell}^{-1}y. \end{aligned}$$

The form of the SDE (Ornstein-Uhlenbeck process) in (19) is exactly that of the thermodynamic device in [29] which if given input A and b above will produce samples from the Gaussian Bayesian posterior $p_{\theta|y}(\theta | y)$. Compared to the simpler Gaussian-Gaussian model above, a disadvantage of this approach is that the covariances Σ and Σ_{ℓ} have to be inverted prior to input as A . However, for linear regression, these matrices are often assumed to be diagonal and otherwise they can be efficiently inverted using the thermodynamic procedures in [29] as preprocessing. Additionally, the formulation of A requires matrix-matrix multiplications which can be costly (even in the case of diagonal covariances). Although, this can be accelerated with parallelization.

On the other hand, the generality of (17-18) makes the approach highly practical. Encompassing Bayesian linear regression [41] and the update step of the Kalman filter [9]. Moreover in the setting of Kalman filtering, the matrices Σ and Σ_{ℓ} are typically shared across time points and thus only need to be inverted once in comparison to the Bayesian posterior update which is applied at every time step (and typically represents the computation bottleneck due to the required matrix inversion).

2.3. Bayesian logistic regression

Logistic regression is a method for classification tasks (both binary and multiclass) that models the dependence of class probabilities on independent variables using a logistic function. In the Bayesian setting, a prior can be assumed on the parameters of a logistic regression model, for example it is common to assume a Gaussian prior. However, after conditioning on observed data a posterior distribution is produced that has no analytical closed form, making Bayesian logistic regression far less efficient than obtaining a point estimate of the parameters. In this section we present a thermodynamic hardware architecture capable of sampling the posterior for binary logistic regression, and show some preliminary evidence that this architecture can do so more efficiently than existing methods.

Given a parameter vector $\theta \in \mathbb{R}^d$ and an independent variable vector $x \in \mathbb{R}^d$, binary logistic regression outputs a class probability $p_{y|\theta,x}(y|\theta, x)$, where $y \in \{-1, 1\}$ (often $y \in \{0, 1\}$ is written instead but we choose this notation to simplify the presentation). The likelihood is $p_{y|\theta,x}(y|\theta, x) = L(y\theta^{\top}x)$ where $L(z) = 1/(1 + e^{-z})$

is the standard logistic function [42]. Note that we will first consider the case of conditioning on a single sample, and in this case the likelihood will be denoted $p_{y|\theta}(y|\theta)$ as x is constant. Additionally, a multivariate normal prior is assumed for the parameters $\theta \sim \mathcal{N}[\mu_{\pi}, \Sigma_{\pi}]$. The Langevin equation for sampling the posterior is therefore:

$$\begin{aligned} d\theta &= -\Sigma_{\pi}^{-1}(\theta - \mu_{\pi})\tau^{-1}dt + L(-y\theta^{\top}x)yx\tau^{-1}dt \\ &\quad + \mathcal{N}[0, 2\mathbb{I}\tau^{-1}dt]. \quad (20) \end{aligned}$$

A circuit implementing Eq. (20) is shown in Fig. 8, and the detailed analysis of this circuit is given in Appendix D. Equation (20) is valid for a single data sample, however, as mentioned, in practice we generally take gradients over a larger number of examples such that the gradients are less noisy. This can be done by enlarging the hardware, resulting in the second term of Eq. (20) being replaced by a sum $\sum_{i=1}^N L(-y_i\theta^{\top}x_i)y_ix_idt$, with N the number of data points. One may also consider minibatches, and the sum is only over a batch of size b . This is achievable by summing currents, which is detailed in the circuit implementation in Appendix D. At a high-level, implementing this protocol in hardware is very simple in the case of a full batch, since the data only needs to be sent once onto the hardware. The following steps are taken to collect the samples: (1) Map the data labels to $\{+1, -1\}$. (2) Map the data (X, Y) onto the hardware (full batch setting). (3) Initialize the state of the system, set the mean and the covariance matrix of the prior. (4) At every interval t_s (the sampling time), measure the state of the system $\theta(t)$ to collect samples.

3. Complexity

Compared to previously derived thermodynamic algorithms (such as those in [29] and [33]), the algorithms presented here differ in that they do not require the estimation of moments of the equilibrium distribution. For example, in the the algorithm for inverting a matrix [29], only a small fraction of the time is spent allowing the system to come to equilibrium, and most of the time is spent collecting samples from this distribution and estimating second moments. Therefore we expect thermodynamic Bayesian inference algorithms can achieve a larger advantage than algorithms based on moment estimation, and we will see that this is indeed the case.

3.1. Time Complexity

It has been noted that the concept of time-complexity is somewhat ambiguous for analog computing devices, and generally the energy cost should be accounted for as well [43]. It is still interesting to consider the physical time necessary to perform an analog computation, and how this scales with the size of the input. In this work, the output is a sample from a probability distribution, and so an appropriate error metric must be used to define the criteria for a successful computation. Here we use the Wasserstein 2 distance between the sampled distribution and the target

distribution (see Appendix A.1), normalized by the norm of the target covariance.

3.1.1. Gaussian-Gaussian model. For the Gaussian-Gaussian model, we make the following assumptions

- $\|\Sigma_\pi\| \leq 1$ and $\|\Sigma_\ell\| \leq 1$.
- $\mu_\pi^\top \Sigma_\pi^{-1} \mu_\pi \leq \mathcal{M}_{\max}$ and $y^\top \Sigma_\ell^{-1} y \leq \mathcal{M}_{\max}$.

The first assumption reflects the fact that in order to solve a problem on an analog computing device, the problem must be rescaled to ensure an appropriate signal range for physical dynamical quantities. The second assumption defines how well-conditioned the problem is, and we include the scaling of resources in the parameter \mathcal{M}_{\max} in our complexity analysis. Subject to these assumptions, it is shown in Appendix C that in order to have the error bounded as $\mathcal{W}_T^2 / \|\Sigma_{\theta|y}\| \leq \varepsilon_W^2$ it is sufficient to allow time T before sampling, given as

$$T = \tau \ln((d + 2\mathcal{M}_{\max})\varepsilon_W^{-2}). \quad (21)$$

In order to gather N I.I.D. samples from the posterior one simply repeats the protocol N times, resulting in an overall time of $N\tau \ln((d + 2\mathcal{M}_{\max})\varepsilon_W^{-2})$. Interestingly, the parameter \mathcal{M}_{\max} may be allowed to increase linearly with dimension, which preserves the logarithmic scaling in dimension. We thus quote the time complexity for collecting N samples from the posterior as

$$T = O(N \ln(d\varepsilon_W^{-2})), \quad (22)$$

when $\mathcal{M}_{\max} \leq O(d)$. It should be noted that this is a worst-case complexity, and the average-case complexity has not yet been fully investigated.

3.1.2. Logistic Regression. For our Bayesian logistic regression algorithm, a study of the required energy would require a much more detailed analysis than for the Gaussian-Gaussian model, and is outside the scope of this work. However, it is possible to derive bounds on the required time with less effort, which is done in Appendix E. In particular, assuming $\|\Sigma_\pi\| \leq 1$, we find that it suffices to allow time T before sampling, given as

$$T = \tau \ln((d + \mathcal{M}_{\theta|y})\varepsilon^{-2}), \quad (23)$$

where $\mathcal{M}_{\theta|y} = \mu_{\theta|y}^\top \Sigma_{\theta|y} \mu_{\theta|y}$. This result leaves something to be desired, as it involves the posterior mean and covariance, and as of yet we have no results constraining the scaling of these parameters with dimension. However, if we introduce an ad-hoc assumption that $\mathcal{M}_{\theta|y} \leq O(d)$ then we obtain

$$T = O(N \ln(d\varepsilon^{-2})). \quad (24)$$

3.2. Energy Complexity

Some skepticism is warranted of treatments that provide the scaling of time required for analog computation alone. This is because the analog device itself will have to grow with dimension, possibly affording some additional parallelism; in other words, it is possible that dimensional

scalings appear too favorable because the computational resources available also grow with dimension. It is therefore essential to also investigate the scaling of energy with dimension, which allows for a fairer comparison to other computational paradigms where the computational resources do not grow with problem size.

The energy required for the Gaussian-Gaussian protocol is derived in Appendix C, where we use the first law of thermodynamics $\Delta E = W - Q$ to compute the work done by the voltage and current sources in the circuit. In this case, the change in internal energy is associated with the inductors and the dissipated heat with the resistors. The same assumptions are made as were used for the time analysis, leading to the following expression for a sufficient amount of work

$$W = 2\tilde{I}^2 L \mathcal{M}_{\max} \ln((d + 2\mathcal{M}_{\max})\varepsilon_W^{-2}) + \frac{1}{2} \tilde{I}^2 L (d + 2\mathcal{M}_{\max}). \quad (25)$$

Once again, this result is simply multiplied by N in order to collect N I.I.D. samples. Assuming, as before, that $\mathcal{M}_{\max} \leq O(d)$, the following scaling is found

$$E = O(Nd \ln(d\varepsilon_W^{-2})). \quad (26)$$

Once again, this is a worst-case result, and the treatment of the average case remains for future work.

4. Experiments

4.1. Gaussian-Gaussian model

To verify that the proposed circuit in Fig. 2 does indeed evolve according to the correct SDE, we ran SPICE circuit simulations. Figure 3 shows the results of such a simulation where a 2-dimensional Gaussian prior and a 2-dimensional Gaussian likelihood are encoded into the conductances while the current in each inductor is measured to determine the resulting posterior. The circuit simulations show strong agreement between the theoretical prediction of the posterior and the simulated current distribution.

The system is simulated for 100 μs with a sampling rate of 2.0 MHz and a burn-in period of 10 μs . For clarity, only a small portion of the 180 000 simulated current samples used for the empirical covariance calculation are plotted in Fig. 3. The conductance matrices are

$$\mathcal{G} = \begin{pmatrix} 2.0 & -1.0 \\ -1.0 & 2.5 \end{pmatrix} \text{ and } \mathcal{G}' = \begin{pmatrix} 3.3 & -2.0 \\ -2.0 & 3.2 \end{pmatrix}. \quad (27)$$

The current sources are set to $I = (0.3, 0.5)^\top$ and $I' = (3.0, 3.0)^\top$. Finally, the scales involved are as follows: $\tilde{I} = 1.0$ mA, $\tilde{R} = 1.0$ k Ω , and $L = 1.0$ μH .

In Figure 4(a), we report the convergence of simulated thermodynamic samples for the Gaussian-Gaussian model with zero prior mean and covariances Σ, Σ_ℓ randomly sampled from a Wishart distribution with $2d$ degrees of freedom. We see fast $O(\log(d))$ convergence in Wasserstein distance between $p(\theta_t | \theta_0)$ and the true posterior, supporting our theoretical claims (22).

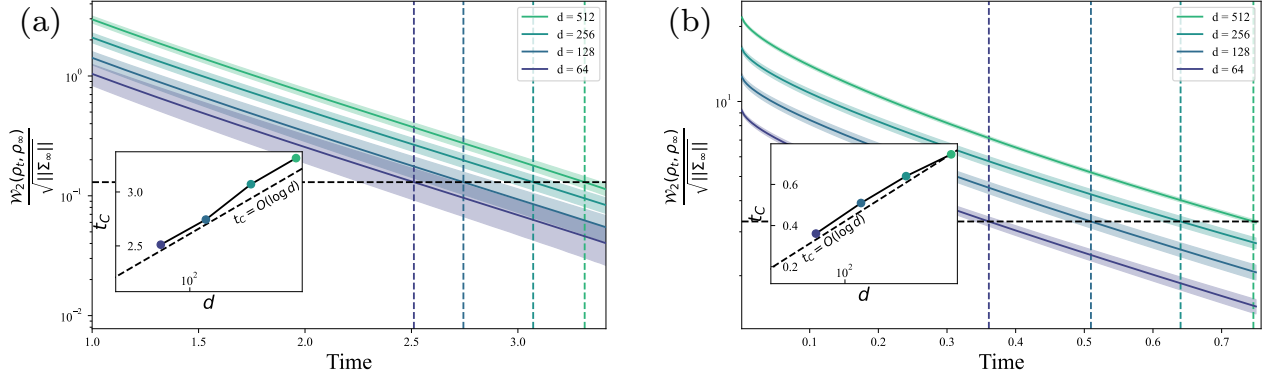


Figure 4. **Convergence in Wasserstein distance for simulations of Thermodynamic Bayesian Inference.** Here, we consider the Wasserstein distance between simulated thermodynamic samples and the true Gaussian posterior as a function of the number of samples (sampling time). All results are simulated exactly with `thermoX` [44] and averaged over 50 random seeds with one standard deviation shown. Panel (a): Gaussian-Gaussian model with zero prior mean and covariances sampled from a Wishart distribution. Panel (b): Bayesian linear regression with randomly sampled data and design matrix.

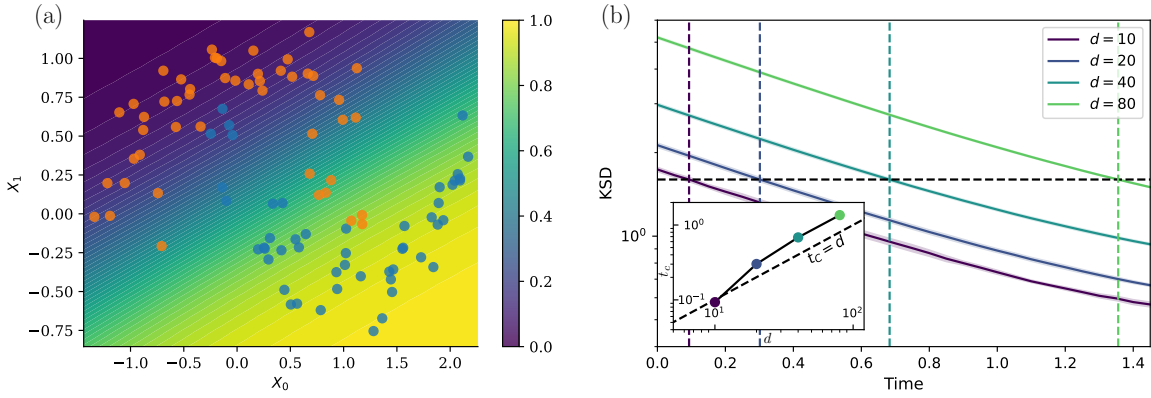


Figure 5. **Bayesian logistic regression on a two-moons dataset.** Panel (a): Contour plot of the probability for a point in the (x_1, x_0) plane to belong to class 1 (blue points). The dataset is also shown, where class 0 (blue points) and class 2 (orange points) are arranged in two intersecting moons. This corresponds to a logistic regression problem with $d = 2$ features. Panel (b): Kernel Stein discrepancy (KSD) of samples as a function of time with an ideal thermodynamic sampler for varying dimensions. Inset: extracted scaling of the crossing time as a function of dimension for a chosen crossing value of the KSD, similarly to other experiments. The color of each point matches the color of vertical dashed lines in the main figure which indicates the crossing time for each considered dimension. The KSD is averaged over five different runs, and for each time 1000 samples are collected.

4.2. Bayesian linear regression

In Figure 4(b), we simulate the evaluation of the thermodynamic linear algebra device [29] for a Bayesian linear regression task. We sample synthetic data $y \in \mathbb{R}^n$ from $p(\theta) = \mathcal{N}(0, \mathbb{I})$ and $p(y | \theta) = \mathcal{N}(n^{-\frac{1}{2}}H\theta, \mathbb{I})$ with random elements of the design matrix $[H]_{ij} \sim \mathcal{N}(0, 1)$ and the $n^{-\frac{1}{2}}$ scaling ensures data y is normalized (a standard practice in machine learning). We fix $n = 500$ and vary the dimension of θ . We observe that the Wasserstein distance to the posterior converges rapidly, matching the logarithmic convergence in the numerics of the Gaussian-Gaussian model.

4.3. Bayesian logistic regression

In Fig. 5(a), we present results for a Bayesian logistic regression on a two-moons dataset, made of points separated

in two classes that are arranged in intersecting moons in the 2D planes, as shown in Fig. 5(a). These results are obtained by running the SDE of Eq. (20) for $d = 2$, hence corresponds to an ideal simulation of the thermodynamic hardware. In this scenario, there are 3 parameters to sample, and $N = 100$ points were considered. In Fig. 5(a), we see that even for such a simple model, only a few points are misclassified. As mentioned, previously, this setting also gives access to better methods to estimate uncertainty in predictions.

In Fig. 5(b), the Kernel Stein discrepancy (KSD) [45] of samples as a function of time is shown for varying dimensions. For these experiments, a dataset made of random points in the d -dimensional hyperplane were generated randomly, belonging to two classes as in the two-moons dataset. The figure displays an exponential scaling of the KSD towards its final value (which is not necessarily zero,

see [45]), similarly to the Wasserstein distance in other experiments. The inset of Fig. 5(b) shows the extracted scaling of the crossing time t_c as a function of dimension, which was obtained by fixing a KSD threshold and extracting after how much time this threshold was reached for various dimensions. While our Eq. (24) predicts a logarithmic scaling of time with dimension for this problem, we observe an approximately linear scaling of the time to reach a given KSD. This discrepancy could be explained by a number of factors; for one, it is not clear that our simulations go to high enough dimension to reveal asymptotic behavior. Also, due to the intractability of the true posterior we use the KSD rather than Wasserstein 2 distance. The KSD requires a kernel specification with a bandwidth that we keep constant across dimension, which may influence dimensional scaling. Finally, unlike our simulations for Gaussian sampling, this system cannot be simulated exactly, so it is possible that there is a dimension-dependent bias due to time discretization (i.e., finite step size).

5. Conclusion

The connection between Bayesian inference and thermodynamics has been highlighted previously [46], [47], [48], [49], although largely in an abstract sense. In this work, we proposed a concrete approach to sampling Bayesian posteriors based on thermodynamics. In Thermodynamic Bayesian Inference (TBI), observed data is encoded in constraints on a physical system, whose degrees of freedom represent the variables we would like to learn about. The process of learning from data is accomplished via the natural equilibration of the system, or equivalently the minimization of its free energy with respect to the imposed constraints. The device that accomplishes this can be viewed as an “entropy pump”, which requires work to be done in order to reduce the entropy of a system while emitting heat to its environment. Interestingly, it has been put forward that similar mechanisms are used in biological systems (in particular, the brain) for maintaining homeostasis and learning from experience [40], [47].

Beyond these conceptual insights, our work has direct practical relevance. We provided explicit constructions of CMOS-compatible analog circuits to implement our TBI algorithms with scalable silicon chips. Our circuit for performing logistic regression represents the first concrete proposal for non-Gaussian sampling with a thermodynamic computer. It is widely acknowledged that non-Gaussian sampling is difficult for digital computers, and often avoided digitally by imposing Gaussian approximations. Thus, our non-Gaussian sampling approach could open up qualitatively new algorithms that otherwise would be avoided due to their difficulty.

In the cases of Gaussian Bayesian inference (Gaussian prior, Gaussian likelihood) and logistic regression, our analysis showed a sublinear complexity in d , leading to a speedup over standard digital methods that is greater than linear. This is an even larger speedup than those previously observed for thermodynamic linear algebra [29], suggesting

that Bayesian inference is an ideal application for thermodynamic computers. Our work lays the foundation for accelerating Bayesian inference, a key component of probabilistic machine learning, with physics-based hardware.

Given that the use of thermodynamic computing for Bayesian inference has not been previously explored, many open questions remain. Immediate extensions of our results include designing a circuit realization of our algorithm for Bayesian linear regression, and quantifying the energy consumption of our Bayesian logistic regression protocol. A farther reaching goal is to understand the fundamental limits on Bayesian inference imposed by thermodynamics, in terms of resources including energy and time. In general, we believe that the impact of this work is not only in providing a new means of implementing Bayesian computations, but additionally our results yield a new perspective on Bayesian inference through the lens of thermodynamics.

References

- [1] Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C Margosian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020.
- [2] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.
- [3] Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.
- [4] Samuel Duffield, Kaelan Donatella, Johnathan Chiu, Phoebe Klett, and Daniel Simpson. Scalable bayesian learning with posteriors. *arXiv preprint arXiv:2406.00104*, 2024.
- [5] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.
- [6] Theodore Papamarkou, Maria Skoularidou, Konstantina Palla, Laurence Aitchison, Julyan Arbel, David Dunson, Maurizio Filippone, Vincent Fortuin, Philipp Hennig, José Miguel Hernández-Lobato, et al. Position: Bayesian deep learning is needed in the age of large-scale ai. In *Forty-first International Conference on Machine Learning*, 2024.
- [7] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated machine learning: methods, systems, challenges*. Springer Nature, 2019.
- [8] Christopher M Bishop and Michael E Tipping. Bayesian regression and classification. In *Advances in learning theory: methods, models and applications*, pages 267–285. IOS Press, 2003.
- [9] Simo Särkkä and Lennart Svensson. *Bayesian filtering and smoothing*, volume 17. Cambridge university press, 2023.
- [10] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.
- [11] Cuong V. Nguyen, Yingzhen Li, Thang D. Bui, and Richard E. Turner. Variational continual learning. In *International Conference on Learning Representations*, 2018.

- [12] Samuel Duffield and Sumeetpal S Singh. Ensemble kalman inversion for general likelihoods. *Statistics & Probability Letters*, 187:109523, 2022.
- [13] Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. What are bayesian neural network posteriors really like? In *International conference on machine learning*, pages 4629–4640. PMLR, 2021.
- [14] Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux – effortless bayesian deep learning, 2022.
- [15] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, April 2017.
- [16] Takahiro Inagaki, Yoshitaka Haribara, Koji Igarashi, Tomohiro Sonobe, Shuhei Tamate, Toshimori Honjo, Alireza Marandi, Peter L. McMahon, Takeshi Umeki, Koji Enbutsu, Osamu Tadanaga, Hirokazu Takenouchi, Kazuyuki Aihara, Ken-ichi Kawarabayashi, Kyo Inoue, Shoko Utsunomiya, and Hiroki Takesue. A coherent ising machine for 2000-node optimization problems. *Science*, 354(6312):603–606, 2016.
- [17] Jeffrey Chou, Suraj Bramhavar, Siddhartha Ghosh, and William Herzog. Analog coupled oscillator based weighted ising machine. *Scientific reports*, 9(1):14786, 2019.
- [18] Naeimeh Mohseni, Peter L. McMahon, and Tim Byrnes. Ising machines as hardware solvers of combinatorial optimization problems. *Nat. Rev. Phys.*, 4(6):363–379, 2022.
- [19] Y Yamamoto, T Leleu, S Ganguli, and H Mabuchi. Coherent ising machines—quantum optics and neural network perspectives. *Applied Physics Letters*, 117(16):160501, 2020.
- [20] Tianshi Wang and Jaijeet Roychowdhury. Oim: Oscillator-based ising machines for solving combinatorial optimisation problems. In *Unconventional Computation and Natural Computation: 18th International Conference, UCNC 2019, Tokyo, Japan, June 3–7, 2019, Proceedings 18*, pages 232–256. Springer, 2019.
- [21] J. Kaiser, S. Datta, and B. Behin-Aein. Life is probabilistic—why should all our computers be deterministic? computing with p-bits: Ising solvers and beyond. In *2022 International Electron Devices Meeting (IEDM)*, pages 21–4. IEEE, 2022.
- [22] Kerem Y. Camsari, Brian M. Sutton, and Supriyo Datta. p-bits for probabilistic spin logic. *Appl. Phys. Rev.*, 6(1):011305, 2019.
- [23] Navid Anjum Aadit, Andrea Grimaldi, Mario Carpentieri, Luke Theogarajan, John M. Martinis, Giovanni Finocchio, and Kerem Y. Camsari. Massively parallel probabilistic computing with sparse Ising machines. *Nat. Electron.*, 5(7):460–468, 2022.
- [24] Tom Conte, Erik DeBenedictis, Natesh Ganesh, Todd Hylton, John Paul Strachan, R Stanley Williams, Alexander Alemi, Lee Altenberg, Gavin E. Crooks, James Crutchfield, et al. Thermodynamic computing. *arXiv preprint arXiv:1911.01968*, 2019.
- [25] Todd Hylton. Thermodynamic neural network. *Entropy*, 22(3):256, 2020.
- [26] Natesh Ganesh. A thermodynamic treatment of intelligent systems. In *2017 IEEE International Conference on Rebooting Computing (ICRC)*, pages 1–4, 2017.
- [27] Patrick J Coles, Collin Szczepanski, Denis Melanson, Kaelan Donatella, Antonio J Martinez, and Faris Sbahi. Thermodynamic ai and the fluctuation frontier. In *2023 IEEE International Conference on Rebooting Computing (ICRC)*, pages 1–10. IEEE, 2023.
- [28] Denis Melanson, Mohammad Abu Khater, Maxwell Aifer, Kaelan Donatella, Max Hunter Gordon, Thomas Ahle, Gavin Crooks, Antonio J Martinez, Faris Sbahi, and Patrick J Coles. Thermodynamic computing system for AI applications. *arXiv preprint arXiv:2312.04836*, 2023.
- [29] Maxwell Aifer, Kaelan Donatella, Max Hunter Gordon, Samuel Duffield, Thomas Ahle, Daniel Simpson, Gavin E Crooks, and Patrick J Coles. Thermodynamic linear algebra. *arXiv preprint arXiv:2308.05660*, 2023.
- [30] Samuel Duffield, Maxwell Aifer, Gavin Crooks, Thomas Ahle, and Patrick J Coles. Thermodynamic matrix exponentials and thermodynamic parallelism. *arXiv preprint arXiv:2311.12759*, 2023.
- [31] Maxwell Aifer, Denis Melanson, Kaelan Donatella, Gavin Crooks, Thomas Ahle, and Patrick J Coles. Error mitigation for thermodynamic computing. *arXiv preprint arXiv:2401.16231*, 2024.
- [32] Patryk Lipka-Bartosik, Martí Perarnau-Llobet, and Nicolas Brunner. Thermodynamic computing via autonomous quantum thermal machines. *arXiv preprint arXiv:2308.15905*, 2023.
- [33] Kaelan Donatella, Samuel Duffield, Maxwell Aifer, Denis Melanson, Gavin Crooks, and Patrick J Coles. Thermodynamic natural gradient descent. *arXiv preprint arXiv:2405.13817*, 2024.
- [34] Radford M Neal. Mcmc using hamiltonian dynamics. *arXiv preprint arXiv:1206.1901*, 2012.
- [35] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.
- [36] Crispin W Gardiner. Handbook of stochastic methods for physics, chemistry and the natural sciences. *Springer series in synergetics*, 1985.
- [37] Udo Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on progress in physics*, 75(12):126001, 2012.
- [38] Takahiro Sagawa. Entropy, divergence, and majorization in classical and quantum thermodynamics. *arXiv preprint arXiv:2007.09974*, 2020.
- [39] Luca Peliti and Simone Pigolotti. *Stochastic thermodynamics: an introduction*. Princeton University Press, 2021.
- [40] Yuri M Svirzhev. Thermodynamics and ecology. *Ecological Modelling*, 132(1-2):11–22, 2000.
- [41] Thomas Minka. Bayesian linear regression. Technical report, Citeseer, 2000.
- [42] Kevin P Murphy. *Probabilistic machine learning: an introduction*. MIT press, 2022.
- [43] Gregory Valiant. Matrix multiplication in quadratic time and energy? towards a fine-grained energy-centric church-turing thesis. *arXiv preprint arXiv:2311.16342*, 2023.
- [44] Samuel Duffield, Kaelan Donatella, and Denis Melanson. thermox: Exact ou processes with jax. <https://github.com/normal-computing/thermox>, 2024.
- [45] Jackson Gorham and Lester Mackey. Measuring sample quality with stein’s method. *Advances in neural information processing systems*, 28, 2015.
- [46] Colin H LaMont and Paul A Wiggins. Correspondence between thermodynamics and inference. *Physical Review E*, 99(5):052140, 2019.
- [47] Karl Friston, James Kilner, and Lee Harrison. A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1):70–87, 2006. Theoretical and Computational Neuroscience: Understanding Brain Functions.
- [48] Susanne Still, David A Sivak, Anthony J Bell, and Gavin E Crooks. Thermodynamics of prediction. *Physical review letters*, 109(12):120604, 2012.
- [49] Anthony Bartolotta, Sean M Carroll, Stefan Leichenauer, and Jason Pollack. Bayesian second law of thermodynamics. *Physical Review E*, 94(2):022102, 2016.

- [50] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.
- [51] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.
- [52] François Bolley, Ivan Gentil, and Arnaud Guillin. Convergence to equilibrium in wasserstein distance for fokker–planck equations. *Journal of Functional Analysis*, 263(8):2430–2457, 2012.

Appendix A. Preliminaries

Here we introduce some mathematical concepts that will be useful for understanding the derivations in the Appendix.

A.1. Wasserstein distance

When analyzing the time evolution of probability distributions with Langevin dynamics, the Wasserstein distance plays a key role. While it actually a continuous family of distances with hyperparameter p , here we focus on the Wasserstein distance with $p = 2$, which can be defined as:

$$\mathcal{W}_2(\rho_1, \rho_2) = \inf(\mathbb{E}|X - Y|^2)^{1/2} \quad (28)$$

where the infimum is taken over all random variables X and Y whose respective probability densities are ρ_1 and ρ_2 . We will often omit the subscript 2 (and use the subscript to denote something else), as in this work we always use $p = 2$. This distance invites alternative interpretations in terms of optimal transport or fluid mechanics. In the latter case, one can write

$$\mathcal{W}_2(\rho_1, \rho_2) = \min_v \int_0^1 \int_{\mathbb{R}^n} \|v(x, t)\|^2 \rho(x, t) dx dt \quad (29)$$

where the density ρ and velocity v satisfy a continuity equation: $\dot{\rho} + \nabla \cdot (\rho v) = 0$ with boundary conditions: $\rho(t = 0) = \rho_1$ and $\rho(t = 1) = \rho_2$.

Below we discuss how the Wasserstein distances enables one to reinterpret Langevin dynamics as a gradient descent process.

A.2. Wasserstein gradient flows

There is an elegant reinterpretation of Langevin dynamics in terms of gradient descent, which is often called Wasserstein gradient flows [50]. In this formulation, we consider a recursive stepping rule, where the probability density $\rho^{(k-1)}$ is updated to a new density $\rho^{(k)}$ given by

$$\rho^{(k)} = \min_{\rho \in K} \left(\frac{1}{2} \mathcal{W}_2(\rho^{(k-1)}, \rho)^2 + hF(\rho) \right). \quad (30)$$

Here, K is the set of all probability densities with finite second moment and h can be viewed as a step size (analogous to the learning rate in gradient descent). The function

$$F(\rho) = E(\rho) - S(\rho)/\beta \quad (31)$$

is the free energy, where the energy and entropy functions are given by

$$E(\rho) = \int_{\mathbb{R}^n} U(x)\rho(x)dx, \quad S(\rho) = - \int_{\mathbb{R}^n} \rho(x) \log \rho(x)dx. \quad (32)$$

The key connection is that the recursive evolution of $\rho^{(k)}$ given by Eq. (30) exactly matches the time evolution of ρ given by the following Fokker-Planck equation:

$$\dot{\rho} = \nabla \cdot (\nabla U(x)\rho) + (1/\beta)\nabla^2 \rho, \quad (33)$$

where $U(x)$ is identified as the potential energy function and β as the inverse temperature. Of course, this Fokker-Planck equation can be alternatively written in terms of its associated Langevin dynamics:

$$dX_t = -\nabla U(X_t)dt + \sqrt{2/\beta}dW_t \quad (34)$$

where W_t is the standard Weiner process, and X_t is a random variable whose probability density is ρ . Thus, Ref. [50] found a deep connection between Langevin dynamics and the recursive process in (30), which can be viewed as a gradient descent process. Specifically, (30) involves gradient descent on the free energy function $F(\rho)$, in a space where the Wasserstein distance is the metric of choice (hence the name Wasserstein gradient flow). This shows that Langevin dynamics attempt to minimize the free energy function.

A.3. Connection between free energy and KL divergence

We remark that there is a close connection between the free energy function and the Kullback-Leibler (KL) divergence. Specifically, suppose that we consider the normalized stationary state of the Fokker-Planck equation, $\rho_\infty(x) = Z^{-1}e^{-\beta U(x)}$, where $Z = \int dx e^{-\beta U(x)}$. Then the KL divergence to this state is given by:

$$\text{KL}(\rho||\rho_\infty) = \int_{\mathbb{R}^n} dx \rho(x) \ln \frac{\rho(x)}{\rho_\infty(x)} dx \quad (35)$$

$$= -S[\rho] - \int_{\mathbb{R}^n} dx \rho(x)(-\beta U(x) - \ln(Z)) dx \quad (36)$$

$$= \beta F[\rho] + \ln(Z) \quad (37)$$

$$= \beta(F[\rho] - F[\rho_\infty]), \quad (38)$$

where in the last line we used the equilibrium thermodynamic identity $\beta F = -\ln(Z)$. So we see that the KL divergence to the stationary state is proportional to the free energy. In light of this connection, we can reinterpret the Wasserstein gradient flow from Eq. (30) as gradient descent on the KL divergence to the stationary state. Therefore, time evolution under the Fokker-Planck equation in (33) progressively minimizes this KL divergence.

A.4. Convergence to the stationary distribution

Under certain conditions, any initial probability density ρ_0 will evolve over time under Langevin dynamics such that it approaches the stationary state ρ_∞ exponentially in time. One can capture this notion quantitatively, for example, with either the KL divergence or the Wasserstein distance.

With the KL divergence, suppose we assume that the stationary state ρ_∞ satisfies the log-Solobev inequality (LSI) with constant $1/\alpha$. Then it follows that the state at time t , ρ_t , satisfies:

$$\text{KL}(\rho_t||\rho_\infty) \leq e^{-2\alpha t} \text{KL}(\rho_0||\rho_\infty). \quad (39)$$

This implies that the KL divergence to ρ_∞ shrinks monotonically with time t , since we can choose ρ_0 to be $\rho_{t-\Delta t}$ where Δt is small shift in time. The LSI condition that is required for (39) is an example of isoperimetry, see for example Ref. [51] for a discussion of isoperimetry. A simpler condition that is sufficient for convergence is positive curvature of the potential [52]. Specifically, suppose it holds that

$$\nabla^2 U(x) \geq (\alpha/\beta)\mathbb{I}, \quad (40)$$

where ∇^2 denotes the Hessian, \mathbb{I} is the identity matrix and the stationary state of the Fokker-Planck equation is $\rho_\infty(x) \propto e^{-\beta U(x)}$. Then as shown in Ref. [52], it follows that the ($p = 2$) Wasserstein distance contracts exponentially in time:

$$\mathcal{W}_2(\rho_t, \sigma_t) \leq e^{-\alpha t} \mathcal{W}_2(\rho_0, \sigma_0), \quad (41)$$

for any two initial states ρ_0 and σ_0 , and their corresponding states at time t , ρ_t and σ_t . Because we can always choose σ_0 to be the stationary state, it follows that the Wasserstein distance to the stationary state shrinks exponentially in time:

$$\mathcal{W}_2(\rho_t, \rho_\infty) \leq e^{-\alpha t} \mathcal{W}_2(\rho_0, \rho_\infty). \quad (42)$$

Thus, positive curvature guarantees exponential convergence in time.

A.5. First law of thermodynamics

The above preliminaries are useful when analyzing the time dynamics (i.e., the runtime cost) of thermodynamic algorithms. However, we are also interested in analyzing the energy dynamics (i.e., the energy cost) of such algorithms.

For the purpose of analyzing energy dynamics, it is useful to review the first law of thermodynamics. This law is essentially a statement of the law of conservation of energy. Whenever no matter is exchanged between a system and its environment, the change in internal energy ΔE of the system is the work W done on the system and minus the heat Q that is lost (i.e., dissipated) from the system:

$$\Delta E = W - Q. \quad (43)$$

Below we will employ this law when analyzing the energy cost of our thermodynamic algorithms (see Appendix C).

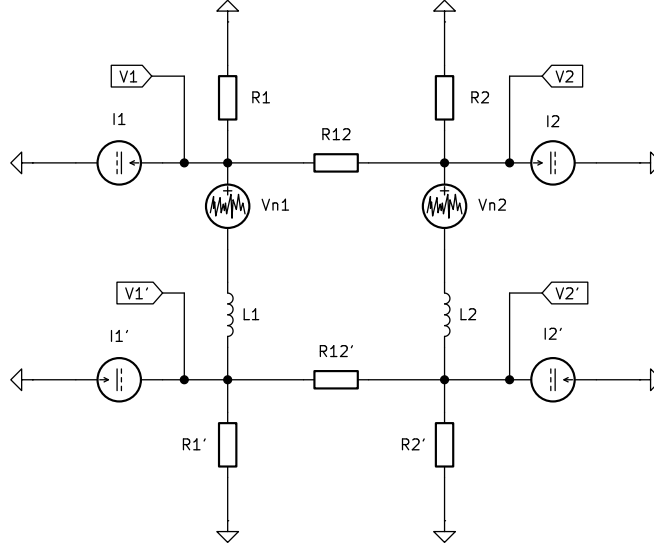


Figure 6. **Circuit schematic for the Gaussian-Gaussian model posterior sampling device.** Note: This figure has the same content as Fig. 2, but is repeated here for convenience.

Appendix B. Analysis of Gaussian Bayesian Inference Circuit

In Figure 6, positive current goes up through the two inductors, left to right through R_{12} and R'_{12} , and towards ground in the other resistors. The two inductors have the same inductance L . KCL gives

$$I_{L1} - I_1 = I_{R1} + I_{12} \quad (44)$$

$$I_{L2} - I_2 = I_{R2} - I_{12} \quad (45)$$

$$-I_{L1} + I'_1 = I'_{R1} + I'_{12} \quad (46)$$

$$-I_{L2} + I'_2 = I'_{R2} - I'_{12}. \quad (47)$$

Using Ohm's law,

$$I_{L1} - I_1 = R_1^{-1}V_1 + R_{12}^{-1}(V_1 - V_2) = (R_1^{-1} + R_{12}^{-1})V_1 - R_{12}^{-1}V_2 \quad (48)$$

$$I_{L2} - I_2 = R_2^{-1}V_2 - R_{12}^{-1}(V_1 - V_2) = (R_2^{-1} + R_{12}^{-1})V_2 - R_{12}^{-1}V_1. \quad (49)$$

These can be written as a single vector equation as follows

$$I_L - I = \mathcal{G}V, \quad (50)$$

where $I_L = (I_{L1} \ I_{L2})^\top$, $I = (I_1 \ I_2)^\top$, and

$$\mathcal{G} = \begin{pmatrix} R_1^{-1} + R_{12}^{-1} & -R_{12}^{-1} \\ -R_{12}^{-1} & R_2^{-1} + R_{12}^{-1} \end{pmatrix}. \quad (51)$$

Similarly, for the lower subcircuit we have

$$-I_L + I' = \mathcal{G}'V'. \quad (52)$$

The inductors obey the equations

$$L_1 \dot{I}_{L1} = V'_1 - (V_1 - V_{n1}) \quad (53)$$

$$L_2 \dot{I}_{L2} = V'_2 - (V_2 - V_{n2}), \quad (54)$$

or in vector notation

$$L \dot{I}_L = V' - V + V_n. \quad (55)$$

Substituting in the expressions for V and V' derived before, we have

$$L\dot{I}_L = \mathcal{G}'^{-1}(I' - I_L) - \mathcal{G}^{-1}(I_L - I) + V_n, \quad (56)$$

or

$$\dot{I}_L = -L^{-1}\mathcal{G}^{-1}(I_L - I) - L^{-1}\mathcal{G}'^{-1}(I_L - I') + L^{-1}V_n. \quad (57)$$

$$dI_L = -L^{-1}\mathcal{G}^{-1}(I_L - I) dt - L^{-1}\mathcal{G}'^{-1}(I_L - I') dt + L^{-1}\sqrt{S}\mathcal{N}[0, \mathbb{I} dt]. \quad (58)$$

We now proceed to non-dimensionalize the above equation. Let $\mathcal{G} = \tilde{R}^{-1}\Sigma_\pi$, $\mathcal{G}' = \tilde{R}^{-1}\Sigma_\ell$, $I_L = \tilde{I}\theta$, $I = \tilde{I}\mu_\pi$, and $I' = \tilde{I}y$. Then

$$\tilde{I}d\theta = -\tilde{I}\tilde{R}L^{-1}\Sigma_\pi^{-1}(\theta - \mu_\pi) dt - \tilde{I}\tilde{R}L^{-1}\Sigma_\ell^{-1}(\theta - y) dt + L^{-1}\sqrt{S}\mathcal{N}[0, \mathbb{I} dt]. \quad (59)$$

Define $\tau = L/\tilde{R}$, giving

$$d\theta = -\Sigma_\pi^{-1}(\theta - \mu_\pi)\tau^{-1} dt - \Sigma_\ell^{-1}(\theta - y)\tau^{-1} dt + \tilde{I}^{-1}L^{-1}\sqrt{S}\mathcal{N}[0, \mathbb{I} dt]. \quad (60)$$

If we set $S = 2\tilde{I}^2L\tilde{R}$, then we have

$$d\theta = -\Sigma_\pi^{-1}(\theta - \mu_\pi)\tau^{-1} dt - \Sigma_\ell^{-1}(\theta - y)\tau^{-1} dt + \mathcal{N}[0, 2\mathbb{I}\tau^{-1} dt], \quad (61)$$

which is Eq. (14)

Appendix C. Time and Energy Cost of Gaussian-Gaussian Posterior Sampling

When the circuit in Fig. 6 is operated, work is done by the voltage and current sources, energy is stored in the inductors, and heat is dissipated by the resistors. The total energy must be conserved in accordance with the first law of thermodynamics, as noted in Eq. (43),

$$W = \Delta E + Q, \quad (62)$$

where W is the work done by the voltage and current sources, ΔE is the change in internal energy of the inductors, and Q is the heat dissipated by the resistors. All of these quantities are with respect to a time interval $[0, T]$, during which the circuit is allowed to evolve before a sample is taken at time T . At time $t = 0$ we set $I_L = 0$, and so

$$\Delta E = \frac{1}{2}L |I_L(T)|^2 \quad (63)$$

Meanwhile, each resistor dissipates heat at power V^2/R , so the total heat flow from the resistors is

$$\dot{Q} = V^\top \mathcal{G}V + V'^\top \mathcal{G}'V', \quad (64)$$

or using the relations $V = \mathcal{G}^{-1}(I_L - I)$ and $V' = \mathcal{G}'^{-1}(-I_L + I')$,

$$\dot{Q} = (I_L - I)^\top \mathcal{G}^{-1}(I_L - I) + (I_L - I')^\top \mathcal{G}'^{-1}(I_L - I'). \quad (65)$$

Note that Eq. (14) can be written $d\theta = -\frac{1}{2}\tilde{I}^{-2}\tilde{R}^{-1}\nabla\dot{Q}\tau^{-1}dt + \mathcal{N}[0, 2\mathbb{I}\tau^{-1}dt]$, so

$$d\langle\dot{Q}\rangle = \langle\nabla\dot{Q} \cdot dI_L\rangle = -\frac{1}{4}\tilde{R}^{-1} \left\langle \left| \nabla\dot{Q} \right|^2 \right\rangle \tau^{-1}dt, \quad (66)$$

establishing that $\langle\dot{Q}\rangle$ is decreasing. Completing the square, we find

$$\tilde{I}^{-2}\tilde{R}^{-1}\dot{Q} = (\theta - \mu_{\theta|y})^\top \Sigma_{\theta|y}^{-1}(\theta - \mu_{\theta|y}) + \mu_\pi^\top \Sigma_\pi^{-1}\mu_\pi + y^\top \Sigma_\ell^{-1}y - \mu_{\theta|y}^\top \Sigma_{\theta|y}^{-1}\mu_{\theta|y}, \quad (67)$$

and because $\langle\dot{Q}\rangle$ is decreasing we have an upper bound

$$\tilde{I}^{-2}\tilde{R}^{-1}\langle\dot{Q}\rangle \leq \mu_\pi^\top \Sigma_\pi^{-1}\mu_\pi + y^\top \Sigma_\ell^{-1}y. \quad (68)$$

Because $I_L(0) = 0$, $\langle\Delta E\rangle$ is increasing. Therefore we get an upper bound by computing $\langle\Delta E\rangle$ in the stationary distribution

$$\tilde{I}^{-2}L^{-1}\langle\Delta E\rangle \leq \frac{1}{2}(\text{tr}\{\Sigma_{\theta|y}\} + \mu_{\theta|y}^\top \mu_{\theta|y}). \quad (69)$$

We now find the time needed to achieve a desired error in the squared Wasserstein distance. As the Wasserstein distance itself behaves like an absolute error rather than a relative error, we will later consider the Wasserstein distance normalized by $\|\Sigma_{\mu|\theta}\|^{1/2}$. The squared Wasserstein distance satisfies

$$\mathcal{W}_t^2 \leq e^{-\lambda_{\min}(\Sigma_{\theta|y}^{-1})t/\tau} \mathcal{W}_0^2 \quad (70)$$

$$\leq e^{-t/\|\Sigma_{\theta|y}\|\tau} \mathcal{W}_0^2. \quad (71)$$

Note that for $\mu_0 = 0$ and $\Sigma_0 = 0\mathbb{I}$,

$$\mathcal{W}_0^2 = \text{tr}\{\Sigma_{\theta|y}\} + \mu_{\theta|y}^\top \mu_{\theta|y}. \quad (72)$$

To achieve $\mathcal{W}_T^2 \leq \varepsilon_W^2 \|\Sigma_{\theta|y}\|$, we can set

$$T = \|\Sigma_{\theta|y}\|\tau \ln(\|\Sigma_{\theta|y}\|^{-1} \mathcal{W}_0^2 \varepsilon_W^{-2}). \quad (73)$$

We arrive at the following upper bound on the work

$$W \leq \tilde{I}^2 L \|\Sigma_{\theta|y}\| (\mu_\pi^\top \Sigma_\pi^{-1} \mu_\pi + y^\top \Sigma_\ell^{-1} y) \ln(\|\Sigma_{\theta|y}\|^{-1} \mathcal{W}_0^2 \varepsilon_W^{-2}) + \frac{1}{2} \tilde{I}^2 L \mathcal{W}_0^2 \quad (74)$$

We now must obtain an upper bound on \mathcal{W}_0^2 . Assume that for some $\mathcal{M}_{\max} > 0$, we have $\mu_\pi^\top \Sigma_\pi^{-1} \mu_\pi \leq \mathcal{M}_{\max}$ and $y^\top \Sigma_\ell^{-1} y \leq \mathcal{M}_{\max}$. Note that the term $\mu_\pi^\top \Sigma_\pi^{-1} \mu_\pi + y^\top \Sigma_\ell^{-1} y - \mu_{\theta|y}^\top \Sigma_{\theta|y}^{-1} \mu_{\theta|y}$ which appeared when completing the square must be positive to preserve the positivity of the original quadratic. Therefore

$$\begin{aligned} \mathcal{W}_0^2 &= \text{tr}\{\Sigma_{\theta|y}\} + \mu_{\theta|y}^\top \mu_{\theta|y} \\ &\leq \text{tr}\{\Sigma_{\theta|y}\} + \lambda_{\min}(\Sigma_{\theta|y}^{-1})^{-1} \mu_{\theta|y}^\top \Sigma_{\theta|y}^{-1} \mu_{\theta|y} \\ &\leq \|\Sigma_{\theta|y}\| (d + 2\mathcal{M}_{\max}). \end{aligned}$$

We then have

$$W \leq 2\tilde{I}^2 L \|\Sigma_{\theta|y}\| \mathcal{M}_{\max} \ln((d + 2\mathcal{M}_{\max}) \varepsilon_W^{-2}) + \frac{1}{2} \tilde{I}^2 L \|\Sigma_{\theta|y}\| (d + 2\mathcal{M}_{\max}) \quad (75)$$

Finally we assume that $\|\Sigma_\pi\| \leq 1$ and $\|\Sigma_\ell\| \leq 1$ (which can always be assured by rescaling the original problem). In this case

$$W \leq 2\tilde{I}^2 L \mathcal{M}_{\max} \ln((d + 2\mathcal{M}_{\max}) \varepsilon_W^{-2}) + \frac{1}{2} \tilde{I}^2 L (d + 2\mathcal{M}_{\max}) \quad (76)$$

Using these assumptions, we also get an upper bound on time

$$T \leq \tau \ln((d + 2\mathcal{M}_{\max}) \varepsilon_W^{-2}). \quad (77)$$

Appendix D. Analysis of Bayesian Logistic Regression Circuit

We now analyze the circuit in Figure 8. The boxes labeled Diff. Pair represent differential pairs of NPN bipolar junction transistors (BJTs), as shown in Fig. 7. To achieve a working implementation, additional circuitry is needed to support the differential pair and assure that it is appropriately biased, including a power source and possibly current mirrors.

The following conventions for current flow will be used

- I_C is the current *into* the collector of a transistor. I_B is the current *into* the base of a transistor. I_E is the current *out of* the emitter of a transistor.
- The output current I_o of a differential pair is the current that flows *into* the collector of the BJT labeled Q_a .
- Positive current flows in the direction of the arrow through all current sources.
- Positive current flows downwards through C_1 and C_2 and from left to right through R_{12} .
- Through resistors R_{A11} , R_{B11} , etc. positive current always flows towards the base of the transistor.

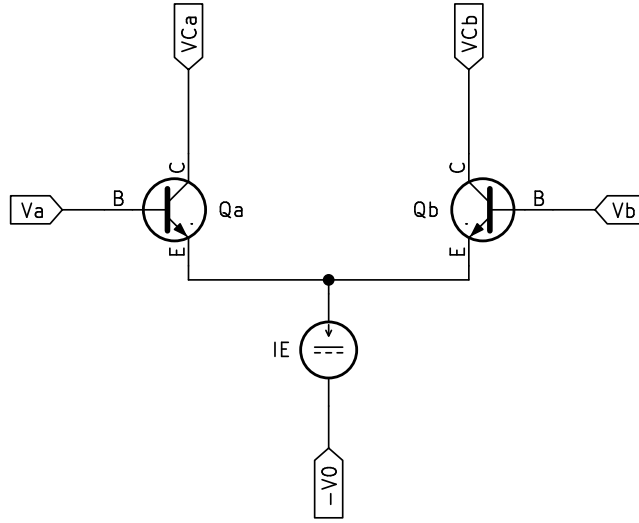


Figure 7. Circuit schematic for the BJT differential pair.

D.1. Analysis of the BJT differential pair

We first consider the behavior of differential pair subcircuit, which can be explained using the Ebers-Moll model. The Ebers-Moll model describes the BJT in active mode, meaning when $V_E < V_B < V_C$, and the circuit must be appropriately biased at all times to ensure the device is always in active mode. According to this model, in active mode the following relations are satisfied

$$I_C = I_S \left(e^{(V_B - V_E)/V_T} - 1 \right), \quad (78)$$

$$I_C = \alpha I_E, \quad (79)$$

where I_S is the saturation current, V_T the thermal voltage, and α is the common-base current gain. I_S is typically on the order of 10^{-15} to 10^{-12} Amps, and at room temperature $V_T = 25.3\text{mV}$. The parameter α is between 0.98 and 1. It follows from Kirchoff's current law (KCL) that $I_B = (1 - \alpha)I_E$. For these typical values of the parameters appearing in Eq. (78) the subtraction of unity in parentheses can safely be ignored, which we will do in what follows. In order for the Ebers-Moll model to be valid, the voltage V_0 should be determined such that $V_C > V_B > V_E$ for all transistors at all times, but the value of V_0 is otherwise unimportant.

To analyze the differential pair of transistors Q_a and Q_b , observe that (by KCL)

$$I_{Ea} + I_{Eb} = I_E. \quad (80)$$

We must distinguish between the two base voltages V_a and V_b , but the two emitter voltages are the same, so we write $V_E = V_{Ea} = V_{Eb}$. Using Eqs. (78) and (79) then,

$$I_E = \frac{I_S}{\alpha} e^{-V_E/V_T} \left(e^{V_a/V_T} + e^{V_b/V_T} \right), \quad (81)$$

where we have dropped the -1 as explained earlier. Now the emitter current I_{Ea} can be written as

$$I_{Ea} = \frac{I_S}{\alpha} e^{(V_a - V_E)/V_T} \quad (82)$$

$$= \frac{I_E e^{V_a/V_T}}{e^{V_a/V_T} + e^{V_{Eb}/V_T}} \quad (83)$$

$$= \frac{I_E}{1 + e^{-(V_a - V_b)/V_T}}, \quad (84)$$

and similarly

$$I_{Eb} = \frac{I_E}{1 + e^{(V_a - V_b)/V_T}}. \quad (85)$$

Equation (79) is then used to find the collector currents

$$I_{Ca} = \frac{\alpha I_E}{1 + e^{-(V_a - V_b)/V_T}}, \quad (86)$$

$$I_{Cb} = \frac{\alpha I_E}{1 + e^{(V_a - V_b)/V_T}}. \quad (87)$$

The base voltages V_a and V_b are still undetermined. However, we will assume the limit $\alpha \rightarrow 1$, where the base current goes to zero. In this limit, the two transistor bases may be connected to nodes in an external circuit to set their voltages. As there is no base current, these connections do not affect the voltages in the external circuit. In what follows, we will consider I_{Ca} the output of the differential pair, and label this current I_o . Again taking the limit $\alpha \rightarrow 1$, we have

$$I_o = \frac{I_E}{1 + e^{(V_a - V_b)/V_T}} = I_E L(-(V_a - V_b)/V_T), \quad (88)$$

where $L(z) = 1/(1 + e^{-z})$ is the standard logistic function. Note that the support circuitry may include a current mirror that inverts the sign of the output current. As this formally has the same effect as a negative value of I_E , we will allow I_E to be negative in what follows.

D.2. Analysis of the logistic regression circuit

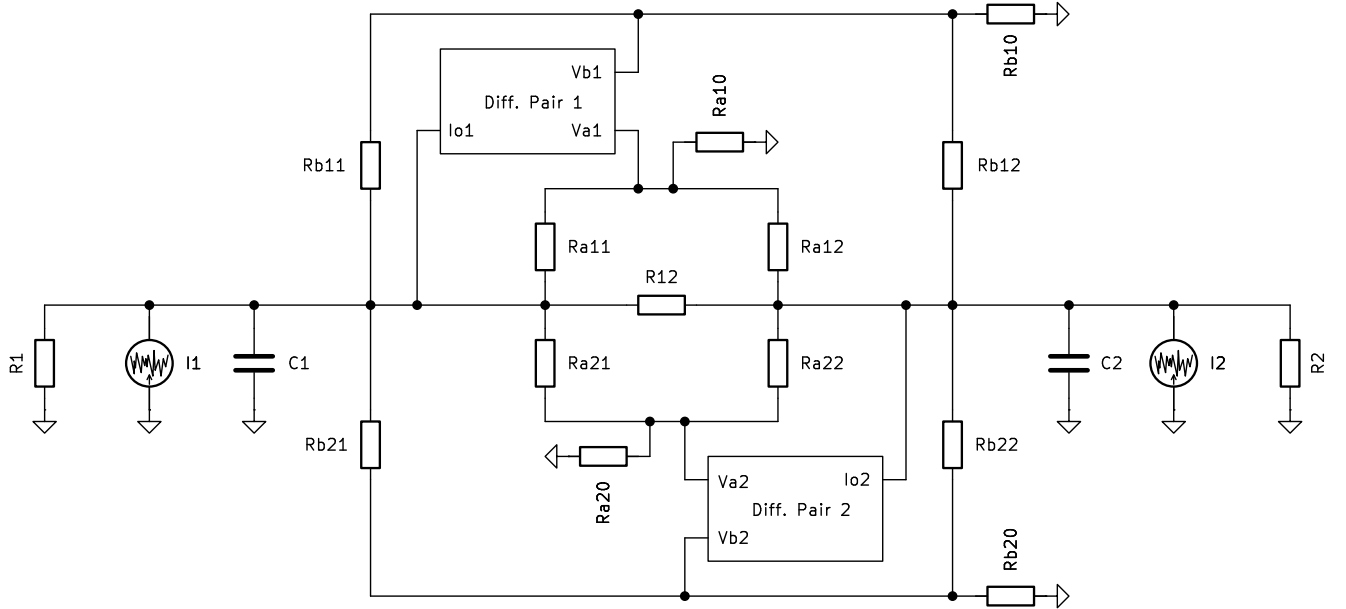


Figure 8. Circuit schematic for the logistic regression posterior sampling device.

As the BJT bases draw negligible current, the voltages V_{a1} , V_{b1} , V_{a2} , and V_{b2} in the circuit can be determined by considering the circuit in the absence of the differential pairs. In this case, we see that (by KCL)

$$R_{a11}^{-1}(V_{C1} - V_{a1}) + R_{a12}^{-1}(V_{C2} - V_{a1}) - R_{a10}^{-1}V_{a1} = 0, \quad (89)$$

and solving for V_{a1} gives

$$V_{a1} = \frac{R_{a11}V_{C1} + R_{a12}V_{C2}}{R_{a11} + R_{a12} + R_{a11}R_{a12}R_{a10}^{-1}} = \frac{R_{a11}^{-1}V_{C1} + R_{a12}^{-1}V_{C2}}{R_{a10}^{-1} + R_{a11}^{-1} + R_{a12}^{-1}}. \quad (90)$$

The same reasoning applies for V_{b1} , resulting in

$$V_{b1} = \frac{R_{b11}^{-1}V_{C1} + R_{b12}^{-1}V_{C2}}{R_{b10}^{-1} + R_{b11}^{-1} + R_{b12}^{-1}}, \quad (91)$$

so

$$V_{a1} - V_{b1} = \frac{g_{a11}V_{C1} + g_{a12}V_{C2}}{g_{a10} + g_{a11} + g_{a12}} - \frac{g_{b11}V_{C1} + g_{b12}V_{C2}}{g_{b10} + g_{b11} + g_{b12}}, \quad (92)$$

where we have written the previous results in terms of the conductance $g = R^{-1}$. The above can be written more conveniently by defining the vectors $\hat{g}_{a1} = (g_{a10} + g_{a11} + g_{a12})^{-1}(g_{a11}, g_{a12})^\top$ and $\hat{g}_{b1} = (g_{b10} + g_{b11} + g_{b12})^{-1}(g_{b11}, g_{b12})^\top$, in terms of which we have

$$V_{a1} - V_{b1} = (\hat{g}_a - \hat{g}_b)^\top V_C. \quad (93)$$

or, defining $\hat{g}_1 = \hat{g}_{a1} - \hat{g}_{b1}$, we simply have

$$V_{a1} - V_{b1} = \hat{g}_1^\top V_C. \quad (94)$$

The latter result can be plugged into Eq. (88) to get I_{o1} ,

$$I_{o1} = I_{E1}L(-\hat{g}_1^\top V_C/V_T), \quad (95)$$

where, as before, $L(z) = 1/(1 + e^{-z})$ is the standard logistic function. By an identical derivation to the one above, a similar relation holds for the lower subcircuit

$$I_{o2} = I_{E2}L(-\hat{g}_2^\top V_C/V_T). \quad (96)$$

We also assume that all resistors R_{aij} , R_{bij} are very large compared to R_{12} so the current flowing through these resistors can be treated as negligible. This assumption does not affect the function of resistors R_{aij} , R_{bij} because only the ratios of these resistances determine the voltages V_{ai} , V_{bi} . Next, we apply KCL to the nodes at the top of capacitors C_1 and C_2

$$-I_{C1} + I_1 - R_1^{-1}V_{C1} + R_{12}^{-1}(V_{C2} - V_{C1}) - I_{o1} = 0, \quad (97)$$

Similarly, KCL for the node above capacitor C_2 reads

$$-I_{C2} + I_2 - R_2^{-1}V_{C2} + R_{12}^{-1}(V_{C1} - V_{C2}) - I_{o2} = 0. \quad (98)$$

Substituting in the expressions derived for the collector currents, we then have

$$-I_{C1} + I_1 - R_1^{-1}V_{C1} + R_{12}^{-1}(V_{C2} - V_{C1}) - I_{E1}L(-\hat{g}_1^\top V_C/V_T) = 0, \quad (99)$$

$$-I_{C2} + I_2 - R_2^{-1}V_{C2} + R_{12}^{-1}(V_{C1} - V_{C2}) - I_{E2}L(-\hat{g}_2^\top V_C/V_T) = 0. \quad (100)$$

Next we define the conductance matrix

$$\mathcal{G} = \begin{pmatrix} R_1^{-1} + R_{12}^{-1} & -R_{12}^{-1} \\ -R_{12}^{-1} & R_2^{-1} + R_{12}^{-1} \end{pmatrix}, \quad (101)$$

allowing us to write a single vector equation

$$-I_C + I - \mathcal{G}V_C - I_E L(-\hat{g}^\top V_C/V_T) = 0, \quad (102)$$

where we have also set $\hat{g}_1 = \hat{g}_2$. Now using the fact that $dV_C/dt = C^{-1}I_C$, we have the following vector differential equation

$$C \frac{dV_C}{dt} = -\mathcal{G}V_C - I_E L(-\hat{g}^\top V_C/V_T) + I. \quad (103)$$

We assume the current vector I has a DC component I_{DC} and a noise component I_{noise} . The noise component is assumed to be an ideal white noise process of infinite bandwidth and power spectral density S , which we write $I_{\text{noise}} = \sqrt{S}\xi(t)$. Altogether, we get the stochastic differential equation

$$dV_C = -C^{-1}\mathcal{G}V_C dt + C^{-1}I_{DC} dt - C^{-1}\alpha I_E L(\hat{g}^\top V_C/V_T) dt + C^{-1}\sqrt{S}\xi(t) dt. \quad (104)$$

Using the identity $\xi(t) dt = \mathcal{N}[0, dt]$, this becomes

$$dV_C = -C^{-1}\mathcal{G}V_C dt + C^{-1}I_{DC} dt - C^{-1}\alpha I_E L(\hat{g}^\top V_C/V_T) dt + C^{-1}\sqrt{S}\mathcal{N}[0, dt]. \quad (105)$$

At this point it is convenient to define dimensionless quantities which are mapped to the physical parameters of the circuit. Define $\theta = V_C/\tilde{V}$, $\Sigma^{-1} = \tilde{R}\mathcal{G}$, $\Sigma^{-1}\mu = I_{DC}/\tilde{I}$, and $yx = -I_E/\tilde{I}$. Our equation now takes the form

$$\tilde{V}dx = -\tilde{V}\tilde{R}^{-1}C^{-1}\Sigma^{-1}x dt + C^{-1}\tilde{I}\Sigma^{-1}\mu dt + C^{-1}\tilde{I}L(-\hat{g}^\top V_C/V_T)yx dt + C^{-1}\sqrt{S}\mathcal{N}[0, \mathbb{I} dt]. \quad (106)$$

Next, let $\tau = \tilde{R}C$, and set $\tilde{I} = \tilde{V}C/\tau$ and $S = 2\tilde{V}^2C^2/\tau$. In this case,

$$d\theta = -\Sigma^{-1}\theta\tau^{-1}dt + \Sigma^{-1}\mu\tau^{-1}dt + L(-\hat{g}^\top V_C/V_T)yx\tau^{-1}dt + \mathcal{N}[0, 2\mathbb{I}\tau^{-1}dt]. \quad (107)$$

Finally, we set $\hat{g} = yxV_T/\tilde{V}$ to obtain

$$d\theta = -\Sigma^{-1}(\theta - \mu)\tau^{-1}dt + L(-yx\theta^\top x)yx\tau^{-1}dt + \mathcal{N}[0, 2\mathbb{I}\tau^{-1}dt], \quad (108)$$

which is identical to Eq. (20).

Appendix E. Time Cost of Logistic Regression

We again use the fact that when $-\nabla^2 \ln p_{\theta|y} \geq \alpha \mathbb{I}$ everywhere, we have

$$\mathcal{W}_t^2 \leq e^{\alpha t/\tau} \mathcal{W}_0^2. \quad (109)$$

For logistic regression we have

$$-\ln p_{\theta|y} = \frac{1}{2}(\theta - \mu_\pi)^\top \Sigma_\pi^{-1}(\theta - \mu_\pi) - \ln(L(y\theta^\top x)). \quad (110)$$

Taking the Hessian of the second term gives

$$-\partial_{\theta_i} \partial_{\theta_j} \ln(L(y\theta^\top x)) = \frac{e^{y\theta^\top x}}{(e^{y\theta^\top x} + 1)^2} x_i x_j, \quad (111)$$

using the fact that $y = \pm 1$ so $y^2 = 1$. Therefore the Hessian is

$$-\nabla^2 \ln(L(y\theta^\top x)) = \frac{e^{y\theta^\top x}}{(e^{y\theta^\top x} + 1)^2} x x^\top. \quad (112)$$

Obviously all of the eigenvalues of the above matrix are non-negative, so overall we have

$$-\nabla^2 \ln p_{\theta|y} \geq \lambda_{\min}(\Sigma_\pi^{-1}) \mathbb{I}. \quad (113)$$

This results on the following bound on the Wasserstein distance to the posterior

$$\mathcal{W}_t^2 \leq e^{-\lambda_{\min}(\Sigma_\pi^{-1})t/\tau} \mathcal{W}_0^2 \quad (114)$$

$$\leq e^{-t/\|\Sigma_\pi\|\tau} \mathcal{W}_0^2. \quad (115)$$

To achieve $\mathcal{W}_T^2 \leq \varepsilon_W^2 \|\Sigma_{\theta|y}\|$, we can set

$$T = \|\Sigma_\pi\| \tau \ln(\|\Sigma_{\theta|y}\|^{-1} \mathcal{W}_0^2 \varepsilon_W^{-2}). \quad (116)$$

The next task is to identify an upper bound on $\|\Sigma_{\theta|y}\|^{-1} \mathcal{W}_0^2$ and . Recall that the definition of the Wasserstein 2 distance is a minimization problem over the set of joint distributions having the correct marginals

$$\mathcal{W}(\rho_1, \rho_2) = \inf(\mathbb{E}|X_1 - X_2|^2)^{1/2}, \quad (117)$$

and so any joint distribution with the correct marginals can be used to give an upper bound. In particular, we consider the case where X and Y are independent, with $X_1 = 0$ and $X_2 \sim p_{\theta|y}$, and find that

$$\mathcal{W}_0^2 \leq \langle X_2^\top X_2 \rangle \quad (118)$$

$$= \text{tr}\{\Sigma_{\theta|y}\} + \mu_{\theta|y}^\top \mu_{\theta|y} \quad (119)$$

$$\leq \text{tr}\{\Sigma_{\theta|y}\} + \lambda_{\min}(\Sigma_{\theta|y}^{-1})^{-1} \mu_{\theta|y}^\top \Sigma_{\theta|y}^{-1} \mu_{\theta|y} \quad (120)$$

$$\leq \|\Sigma_{\theta|y}\| (d + \mathcal{M}_{\theta|y}), \quad (121)$$

where $\mathcal{M}_{\theta|y} = \mu_{\theta|y}^\top \Sigma_{\theta|y}^{-1} \mu_{\theta|y}$. Unfortunately it is less clear how to bound $\mathcal{M}_{\theta|y}$ in terms of the prior and likelihood parameters, and we leave this task for future work. At any rate, we arrive at

$$T \leq \|\Sigma_\pi\| \tau \ln((d + \mathcal{M}_{\theta|y}) \varepsilon_W^{-2}). \quad (122)$$

As in the case of the Gaussian-Gaussian model, without loss of generality we may again assume $\|\Sigma_\pi\| \leq 1$ to obtain

$$T \leq \tau \ln((d + \mathcal{M}_{\theta|y}) \varepsilon_W^{-2}). \quad (123)$$

Appendix F. Conditioning on multiple I.I.D. samples

When conditioning on a single sample y , the energy U can be separated into two terms, one mapping to the prior and the other to the likelihood:

$$U(r) = U_\pi(r) + U_\ell(r), \quad (124)$$

where $\beta U_\pi(r) = -\ln p_\theta(r/\tilde{r})$ and $\beta U_\ell(r) = -\ln p_{y|\theta}(y|r/\tilde{r})$. In general we may have a number of I.I.D. samples $Y = (y_1, \dots, y_N)$, and would like to sample from $p_{\theta|Y}(\theta|Y)$. Because the samples of y are I.I.D., we have

$$p_{Y|\theta}(Y|\theta) = \prod_{i=1}^N p_{y|\theta}(y_i|\theta). \quad (125)$$

In this case the likelihood part of the potential energy takes the form

$$\beta U_\ell(r) = -\sum_{i=1}^N \ln p_{y|\theta}(y_i|r/\tilde{r}), \quad (126)$$

while the prior part is the same as in the single-sample case, $\beta U_\pi(r) = -\ln p_\theta(r/\tilde{r})$. This form of the potential energy has a convenient physical interpretation: the function U_π can be interpreted as the self-energy of the system in state r (that is when it is decoupled from an external system), while the function $U_\ell(\theta)$ can be viewed as an interaction energy between the state r and the state y of an external system. When there are multiple I.I.D. samples, this is analogous to the state r interacting with a collection of external systems in states $Y = (y_1 \dots y_N)$, and each such interaction contributes its own term to the interaction energy. This provides a framework for building a physical device to sample from the posterior conditioned on multiple I.I.D. samples; one must simply couple a collection of external systems in states $Y = (y_1 \dots y_N)$ to the system in such a way that each interaction contributes an energy of $-\ln p_{y|\theta}(y|r/\tilde{r})$.

We will now describe another approach to building a physical device that samples from the posterior conditioned on multiple I.I.D. samples of y . We first observe that the Langevin equation for the device in this case must be

$$d\theta = \nabla_\theta \ln p_\theta(\theta) \tau^{-1} dt + \sum_{i=1}^N \nabla_\theta \ln p_{y|\theta}(y_i|\theta) dt + \mathcal{N}[0, 2\tau^{-1} dt], \quad (127)$$

As discussed above, if we have a device that can implement the N likelihood drift terms simultaneously then the problem is solved. However, suppose that we have a device that is only capable of implementing a single likelihood term at a time, but y may be varied as a function of time. Additionally, we make the interaction energy for this device larger by a factor of N for reasons that will become clear. That is, we have a device that implements an SDE of the form

$$d\theta = \nabla_\theta \ln p_\theta(\theta) \tau^{-1} dt + N \nabla_\theta \ln p_{y|\theta}(y(t)|\theta) dt + \mathcal{N}[0, 2\tau^{-1} dt]. \quad (128)$$

We may choose a short time duration Δt , and set

$$y(t) = y_{\lfloor t/\Delta t \rfloor \bmod N+1}. \quad (129)$$

So for $0 \leq t \leq \Delta t$ we set $y(t) = y_1$, for $\Delta t < t \leq 2\Delta t$ we set $y(t) = y_2$, and so on. Once $t > N\Delta t$ we start over at y_1 and continue cycling over all of the I.I.D. samples. Suppose that Δt is short enough that all of the samples are cycled over before the state θ changes significantly. We may then average drift term $N \nabla_\theta \ln p_{y|\theta}$ over a period of time $N\Delta t$ and consider θ constant within this average. Carrying out this time average, we find

$$\frac{1}{N\Delta t} \sum_{i=1}^N \Delta t N \nabla_\theta \ln p_{y|\theta}(y_i|\theta) = \sum_{i=1}^N \nabla_\theta p_{y|\theta}(y_i|\theta), \quad (130)$$

resulting in the correct form of the Langevin equation.