PROVABLE ACCURACY BOUNDS FOR HYBRID DYNAMICAL OPTIMIZATION AND SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Analog dynamical accelerators (DXs) are a growing sub-field in computer architecture research, offering order-of-magnitude gains in power efficiency and latency over traditional digital methods in several machine learning, optimization, and sampling tasks. However, limited-capacity accelerators require hybrid analog/digital algorithms to solve real-world problems, commonly using large-neighborhood local search (LNLS) frameworks. Unlike fully digital algorithms, hybrid LNLS has no non-asymptotic convergence guarantees and no principled hyperparameter selection schemes, particularly limiting cross-device training and inference.

In this work, we provide non-asymptotic convergence guarantees for hybrid LNLS by reducing to block Langevin Diffusion (BLD) algorithms. Adapting tools from classical sampling theory, we prove exponential KL-divergence convergence for randomized and cyclic block selection strategies using ideal DXs. With finite device variation, we provide explicit bounds on the 2-Wasserstein bias in terms of step duration, noise strength, and function parameters. Our BLD model provides a key link between established theory and novel computing platforms, and our theoretical results provide a closed-form expression linking device variation, algorithm hyperparameters, and performance.

1 INTRODUCTION

Computing research has long borrowed from the physical sciences. Sampling and optimization algorithms such as simulated annealing (Kirkpatrick, 1984), parallel tempering (J. Earl & W. Deem, 2005), and Langevin Monte Carlo (LMC) (Chewi et al., 2021) were directly inspired by physical processes observed in nature. Novel dynamical formulations of classical algorithms such as Nesterov accelerated gradient and Polyak's heavy-ball method (Kovachki & Stuart, 2021) and stochastic gradient descent (Orvieto & Lucchi, 2020) have provided optimized step-size schemes and insights into iterate behavior.

Drawing on the close connection between computation and physics, a growing computer architecture sub-field has proposed leveraging physical dynamics to accelerate computationally expensive workloads using "dynamical accelerators" (DXs). Originally, research focused on combinatorial optimization problems (Inagaki et al., 2016; Ushijima-Mwesigwa et al., 2017; Wang & Roychowdhury, 2019; Afoakwa et al., 2021; Mohseni et al., 2022) and matrix-vector multiplication Xiao et al. (2022). However, the field has expanded to sampling for energy-based model training and inference (Vengalam et al., 2023) and generative inference in graph neural networks (Wu et al., 2024; Song et al., 2024).

The interest in analog acceleration coincides with novel proposals for "local update" algorithms, where layer activations h are solutions to a minimization problem $h_\ell^* = \text{argmin}_h f(h)$ (Scellier & Bengio, 2017; Stern et al., 2021; Millidge et al., 2022; Scellier et al., 2023). While costly in digital systems, stochastic analog optimizers can effectively solve $\text{argmin}_h f(h)$ in minimal time and energy (Wu et al., 2024), making them suitable candidates for local-update learning implementations.

However, real-world problems are typically too large for dynamical accelerators to optimize in their entirety, requiring routines to partition and iteratively sample/optimize subspaces (Booth et al., 2017; Sharma et al., 2022; Song et al., 2024), most commonly using hybrid "large-neighborhood local search" (LNLS) frameworks (Ahuja et al., 2002; Booth et al., 2017). In hybrid LNLS, the DX is used to perform alternating sampling/minimization over within-capacity subproblems. However, hybrid LNLS has undergone little theoretical examination. No non-asymptotic convergence bounds yet exist, limiting the appeal of hybrid LNLS compared with more well-understood digital algorithms. Moreover, the effect of algorithm hyperparameters on convergence and their interplay with device non-idealities is unclear. Models trained on one DX may require hyperparameter adjustment, if not outright device-specific retraining, prior to inference on another (He et al., 2019; Long et al., 2019). Without non-asymptotic analysis linking device variation and accelerator convergence, accelerator adaptation reduces to trial-and-error.

In this work, we provide the first explicit probabilistic convergence guarantees for hybrid LNLS algorithms in activation sampling and optimization: a crucial first step in optimizing and analyzing hybrid DX frameworks. We start by reducing hybrid LNLS to block sampling with continuous-time, Langevin diffusion-based sub-samplers, to which we can apply tools from classical sampling analysis. Two block selection rules for "block Langevin diffusion" (BLD) are examined, randomized and cyclic, using ideal (Secs. 3.2 and 3.3) and finite-variation (Sec 3.4) analog components. Under a log-Sobolev inequality (LSI), we prove that ideal accelerators converge to the target distribution exponentially fast. However, we show that finite device variation incurs a bias in W_2 distance, proportional to step duration and dependent on variation magnitude. We illustrate our findings with numerical experiments on a toy Gaussian sampling problem, demonstrating the effect of device variation and hyperparameter choice on W_2 convergence.

Our contributions can be summarized as follows:

1. We provide novel bounds on randomized block diffusions using explicit constants (Theorem 1), strengthening the results of Ding et al. (2020)
2. We provide completely novel bounds for cyclic block diffusions (Theorem 2) by proving a novel conditional sampling lemma for Kullback-Liebler divergence (Lemma 1)
3. Using a Talagrand transportation inequality, we combine our ideal results with analysis following Raginsky et al. (2017) to provide non-asymptotic guarantees for DXs with analog non-idealities (Theorem 3), applicable to both sampling and optimization tasks.

1.1 RELATED WORKS

Ding & Li (2021) and Ding et al. (2021) proposed and analyzed "randomized coordinate Langevin Monte Carlo" (RCLMC) methods for sampling tasks using over and underdamped Langevin dynamics. Their methodology used Wasserstein coupling arguments akin to Dalalyan (2016), in contrast to our interpolation arguments following Vempala & Wibisono (2019). Accordingly, the authors assumed a strongly-log concave target distribution: a much stronger assumption than an LSI. Moreover, Ding et al. (2021) provided insufficient analysis for the continuous-time case, focusing primarily on the discrete RCLMC algorithm. DX algorithm analysis required continuous-time bounds with explicit constants, necessitating our contributions.

Two algorithms related to BLD garnering recent interest are "coordinate ascent variational inference" (CAVI), which performs variational inference over factorized "mean-field" distributions (Bhattacharya et al., 2023; Arnese & Lacker, 2024), and the split Gibbs sampler (SGS), which alternates sampling over problem variables with augmented priors (Vono et al., 2019; 2022). CAVI is similar to BLD, and indeed the information theoretic analysis by Lee (2022) has a similar structure to our proof of Lemma 1. SGS has been likened to the ADMM opimization algorithm Vono et al. (2022), indicating there may be an equivalence to BLD akin to classical block optimization Tibshirani (2017).

A related class of works have analyzed the accuracy of analog matrix-vector multiplication (MVM) accelerators in neural network inference (Klachko et al., 2019; Xiao et al., 2022). MVM accelerators are a restricted class of DXs minimizing $\min_{y\in\mathbb{R}^d} ||y-Wx||^2$: equivalent to performing MVM in the analog domain. Our analysis generalizes MVM analysis and is applicable in more complex analog settings such as generative sampling (Vengalam et al., 2023; Melanson et al., 2023; Wu et al., 2024).

Optimization-based convergence analyses of specific DX architectures were carried out by Erementchouk et al. (2022); Pramanik et al. (2023). Asymptotic convergence in expectation to the global minimizer was proved by Pramanik et al. (2023) in the zero-temperature limit with decreasing stepsize, echoing our results in Sec. 3.4. However, neither work accounted for the effect of device variation or problem partitioning, and both focused on specific DX modalities (nonlinear electronic/optical oscillators) rather than a general model of DX behavior. Information-theoretic analysis conducted by Dambre et al. (2012); Hu et al. (2023) have bounded the asymptotic computational capabilities of DX systems, but not their probabilistic convergence.

2 BACKGROUND

2.1 DYNAMICAL ACCELERATORS

The first wave of dynamics-accelerated optimizers primarily targeted the Ising Spin Glass (ISG) Hamiltonian from statistical physics, earning the appellation "Ising Machines". The ISG Hamiltonian describes quadratic interactions between binary "spins", which can be used to solve intractable combinatorial problems (Lucas, 2014). Ising machines have been implemented using quantum spins (Ushijima-Mwesigwa et al., 2017), electronic (Wang & Roychowdhury, 2019; Albertsson & Rusu, 2023) and optical (Inagaki et al., 2016; Honjo et al., 2021) oscillator phases, resistively-coupled capacitors (Afoakwa et al., 2021), and many more besides (Mohseni et al., 2022). These initial prototypes successfully optimized binary target functions, however recent architectures have broader applications domains: with support for non-quadratic cost functions (Sharma et al., 2023; Bashar & Shukla, 2023; Bybee et al., 2023) and continuous values (Brown et al., 2024; Wu et al., 2024; Song et al., 2024). Since these designs have moved beyond the ISG Hamiltonian, we term this broader class simply as "dynamical accelerators" (DXs).

While the physical implementation differs between DXs, several proposals can be described by a Langevin stochastic differential equation (SDE)

$$dx_t = -\nabla h(x_t, t)dt + \sqrt{2\beta(t)^{-1}}dW_t \tag{1}$$

where $x_t \triangleq x(t)$ is the system state, dW_t is a Brownian noise term, $h(x, t)$ is the deterministic system potential, and $\beta(t) = 1/T(t)$ is the (also potentially time dependent) inverse pseudo-temperature of the system.

$x(t)$ represents the continuous, physical degrees of freedom of the optimizer/sampler, such as capacitor voltage (Afoakwa et al., 2021) or oscillator phase (Inagaki et al., 2016; Wang & Roychowdhury, 2019). Several DX prototypes have been shown to follow forms of Equation (1), either intentionally to escape local minima (Wang & Roychowdhury, 2019; Sharma et al., 2023; Aifer et al., 2023) or unintentionally to model dynamic environment noise (Wang et al., 2013). The potential $h(x, t)$ includes the target function $f(x)$ along with optional time-dependent terms, such as a sub-harmonic injection locking potential for binary applications (Wang & Roychowdhury, 2019).

DXs are also prone to static "device variation" owing to analog non-idealities. Unlike the Brownian term dW_t, static non-idealities are not self-averaging, and result in a biased estimate $g_\delta(x)$ of the gradient $\nabla f(x)$. In a quadratic function $f(x) = x^T W x$, for instance, the gradient estimate can be described as $g_\delta(x) = (W + W^T)x + \delta x$, where $\delta_{ij} \sim \mathcal{N}(0, \Delta^2)$ are fixed non-idealities in device components. Previous studies have examined the impact of static variation on binary optimization (Albash et al., 2019) and matrix-vector multiplication (Xiao et al., 2022), but have not extended to non-asymptotic convergence analysis for more general functions over $\mathbb{R}^d$.

2.2 LANGEVIN DIFFUSION

If we restrict our analysis to the time-homogeneous case where $h(x, t) = f(x)$, $\beta(t) = \beta$, the dynamics are Markovian with a constant stationary distribution

$$\pi_\beta(x) \propto e^{-\beta f(x)}.$$

The Langevin SDE

$$dx_t = -\nabla f(x_t)dt + \sqrt{2\beta^{-1}}dW_t$$

produces a continuous sample path $x(t)$ with each $x(\tau), \tau \geq 0$ acting as a random variable. The law of x_t, μ_t (denoted $\mu_t = \mathcal{L}(x_t)$), is described by the *Fokker-Planck* equation (FPE)

$$\partial_t \mu_t = \beta^{-1}\nabla^2 \cdot \mu_t + \nabla \cdot [\mu_t \nabla f(x_t)].$$

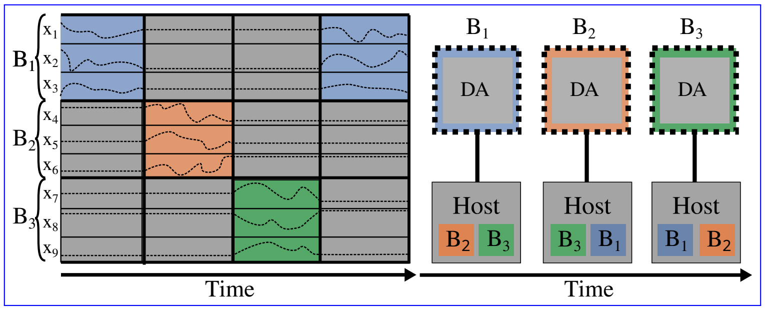


Figure 1: Illustration of the LNLS algorithm on a 3-block, 9-variable problem. [Left] An illustration of the variable sample paths during algorithm execution. When a block is not being actively evolved, the constituent variables remain fixed (gray). [Right] Logical partition of variables in an LNLS framework, where one block is being actively evolved by the DX with the others resident in digital memory. The digital host performs the control operations needed to read the block state, write back to memory, and begin the next block evolution.

The Langevin SDE describes the physical evolution of x_t, while the FPE describes the change in the sample distribution μ_t in measure space. If π_β satisfies a log-Sobolev inequality (LSI, see Sec. 3), then μ_t converges to π_β exponentially fast in measure space (Theorem 1 Vempala & Wibisono, 2019)

$$\mathrm{D}_{\mathrm{KL}}(\mu_t\|\pi) \leq e^{-2\gamma\beta t}\,\mathrm{D}_{\mathrm{KL}}(\mu_0\|\pi) \tag{2}$$

where $\mathrm{D}_{\mathrm{KL}}(\mu_t\|\pi) \triangleq \int \mu_t(x)\log\frac{\mu_t(x)}{\pi(x)}dx \triangleq \int \mu_t \log\frac{d\mu_t}{d\pi}$ is the Kullback-Leibler (KL) divergence between two probability measures and $1/\gamma$ is the log-Sobolev constant.

Recalling the Otto-Villani theorem (Theorem 1 Otto & Villani, 2000), an LSI inequality further implies a Talagrand transportation inequality

$$W_2(\mu_t,\pi) \leq \left(\frac{2}{\gamma}\right)^{1/2}\sqrt{\mathrm{D}_{\mathrm{KL}}(\mu_t\|\pi)}$$

where $W_2(\mu_t,\pi) = \inf_{\nu\in\mathcal{C}(\mu_t,\pi)}\left(\int ||x-y||_2^2\,\nu(x,y)dxdy\right)^{1/2}$ is the 2-Wasserstein distance between μ_t and π and $\nu \in \mathcal{C}(\mu_t,\pi)$ is a *coupling* over μ_t, π. Convergence in D_{KL} under an LSI therefore implies convergence in W_2, allowing us to state bounds in both. Crucially for our purposes, the 2-Wasserstein distance is a metric over probability distributions, allowing use of the triangle inequality (Raginsky et al., 2017).

As $\beta \to \infty$, $\pi_\beta(x)$ concentrates around the minimizer(s) of f. This observation permits us to unite optimization and sampling using annealing schemes (Kirkpatrick, 1984; Chiang et al., 1987; Chak et al., 2023) which gradually increase β to escape early local minima and (hopefully) find the global minimum, indicating a direction for future work extending BLD. Previous works have also used bounds on convergence to π_β at constant β to bound optimizer hitting times Zhang et al. (2017) and expected excess risk (Raginsky et al., 2017; Xu et al., 2020; Farghly & Rebeschini, 2021; Zhang et al., 2023) in non-convex optimization.

3 MAIN RESULTS

3.1 LNLS AS BLOCK SAMPLING

DXs have a finite capacity. To solve problems exceeding that capacity, hybrid analog/digital algorithms are necessary. A popular candidate for hybrid optimization/sampling is the Large-Neighborhood Local Search (LNLS) framework (Raymond et al., 2023; Booth et al., 2017; Ahuja et al., 2002; Sharma et al., 2022), where a local solver (the DX) is used to optimize/sample blocks of variables $\{B_1, B_2, ..., B_b\}$ conditioned on the rest of the problem state, illustrated in Fig. 1.

We can formalize LNLS by borrowing notation from classical coordinate descent (Nesterov, 2012; Beck & Tetruashvili, 2013). We assume the Cartesian product decomposition $\mathbb{R}^d = \times_{i=1}^{b} B_i$ sat-

Algorithm 1 Block Langevin Diffusion (BLD)

```
procedure BLD($x_0 \in \text{dom}(f)$, Decomposition $\{B_1, ..., B_b\}$, Step Size Set $\lambda \in \mathbb{R}^b_+$)
    for $k \geq 0$ do
        Choose $B_k$ (Random or Deterministic)
        Sample:
            $x_{k+1} = x_k - \int_0^{\lambda_k} U_k \nabla f(x_k) dt + \int_0^{\lambda_k} U_k \sqrt{2\beta^{-1}} dW_t$
    end for
end procedure
```

isfies $B_i \cap B_j = \emptyset$ for $i \neq j$ and each block subspace B_i has dimension d_i. LNLS frameworks essentially perform block Gibbs sampling from each conditional distribution $\mu_{B_i|B_1 \ldots B_b}$, where each block is chosen at random or in a deterministic order.

To express updates in $\mathbb{R}^d$, we decompose $I^d = \sum_{i=1}^b U_i$ where each $U_i \in \mathbb{R}^{d \times d}$ has ones along diagonal indices corresponding to unit vectors in B_i and zeros elsewhere. Then $\sum_{i=1}^b U_i \nabla f(x) = \nabla f(x)$ and we can express the SDE for a single block B_i diffusion as

$$dx = -U_i \nabla f(x) dt + U_i \sqrt{2\beta^{-1}} dW_t. \tag{3}$$

Equation (3) leaves the conditioned dimensions $\overline{B}_i \triangleq \{j \in \{1, ..., d\} : j \notin B_i\}$ invariant. Each block diffusion occurs in continuous time, but the blocks are swapped at discrete steps. Accordingly, we denote x_t^k as the iterate at time t in block step k and μ_t^k as its associated probability distribution. When each block is evolved at constant β according to Equation (3), LNLS becomes equivalent to a block sampling algorithm, *Block Langevin Diffusion* (BLD), shown in Algorithm 1. BLD is a continuous-time generalization of "randomized coordinate Langevin Monte Carlo" (RCLMC) studied in Ding et al. (2021); Ding & Li (2021). By reducing LNLS to a block Langevin algorithm, we can tractably analyze algorithm performance using well-developed tools from stochastic process analysis. The BLD framework given in Algorithm 1 leaves open the choice of block selection. Here we consider randomized and cyclic selection rules, denoted *Randomized Block Langevin Diffusion* (RBLD) and *Cyclic Block Langevin Diffusion* (CBLD) respectively.

Throughout our analysis, we make the following assumptions on f.

Assumption 1. f is continuously differentiable

Assumption 2. $\pi_\beta \propto \exp[-\beta f(x)]$ satisfies a log-Sobolev inequality (LSI) with $C_{\text{LSI}} = \frac{1}{\gamma}$ if, for all distributions μ with finite second moment

$$\mathrm{D}_{\mathrm{KL}}(\mu \| \pi_\beta) \triangleq \int \mu(x) \log \frac{\mu(x)}{\pi(x)} dx \leq \frac{1}{2\gamma} \overbrace{\int \mu(x) \|\nabla \log \frac{\mu(x)}{\pi(x)}\|^2 dx}^{\mathrm{FI}(\mu\|\pi_\beta)}$$

where $\mathrm{FI}(\mu\|\pi_\beta)$ is the (relative) *Fisher information*. An LSI condition is the sampling equivalent of a Polyak-Łojasiewicz (PL) "gradient domination" inequality in optimization, where $\mathrm{D}_{\mathrm{KL}}(\mu\|\pi_\beta)$ is our objective function(al) and $\mathrm{FI}(\mu\|\pi_\beta)$ is a "gradient norm". An LSI can hold even in non-log-concave distributions, making it a more general assumption than the strong log-concavity presumed by Ding & Li (2021). Examples include globally strongly log-concave measures with bounded regions of non-log concavity (Raginsky et al., 2017; Ma et al., 2019), high-temperature spin systems (Bauerschmidt & Bodineau, 2019) and heavy-tailed distributions which are not *strongly* log-concave.

3.2 RANDOMIZED BLOCK LANGEVIN DIFFUSION

In the randomized case, we select the next variable block according to the probability distribution $\phi \in \mathbb{R}^b$. Ding et al. (2021) analyzed RCLMC using Wasserstein coupling arguments, however our analysis builds on the traditional proof of Equation (2) which relies on the *de Bruijin identity*

$$\partial_t \mathrm{D}_{\mathrm{KL}}(\mu_t\|\pi_\beta) = -\beta^{-1} \mathrm{FI}(\mu_t\|\nu)$$

which, when combined with the LSI, proves exponential convergence since $-\operatorname{FI}(\mu_t\|\nu) \leq -2\gamma \operatorname{D_{KL}}(\mu_t\|\pi_\beta)$. In the same vein, we use probabilistic arguments in Appendix B.2 to prove a de Bruijin *in*equality

$$\partial_t \operatorname{D_{KL}}(\mu_t\|\pi_\beta) \leq -\phi_{\min}\beta^{-1} \operatorname{FI}(\mu_t\|\nu)$$

where $\phi_{\min}$ is the minimum block probability in ϕ.

By integrating and expanding the inequality, we easily obtain convergence in $\operatorname{D_{KL}}(\mu_t^k\|\pi)$, expressed in Theorem 1. We also prove convergence for a discrete-time variant (RBLMC) in Appendix B

Theorem 1 (RBLD $\operatorname{D_{KL}}(\mu_t^k\|\pi)$ Convergence)**.** *Let $\theta = (B_1, ..., B_b)$ be a given block permutation and let $\lambda = (\lambda_1, ..., \lambda_b) \in \mathbb{R}^b$ be the sampling times for each block, $\lambda_i > 0$. For any $\pi_\beta \propto \exp[-\beta f(x)]$ satisfying Assumptions 1 and 2, and any $\beta > 0$, the sample distribution after k steps of RBLD (μ^k) satisfies*

$$\operatorname{D_{KL}}(\mu^k\|\pi) \leq e^{-2\gamma\beta^{-1}\phi_{\min}\lambda_{\min}k} \operatorname{D_{KL}}(\mu^0\|\pi).$$

3.3 CYCLIC BLOCK LANGEVIN DIFFUSION

While our randomized results tighten existing theory, real-world instances of LNLS often use cyclic orderings Sharma et al. (2022); Song et al. (2024); Wu et al. (2024), as they are more amenable to direct hardware and software optimization and are easier to implement in practice.

However, unlike RBLD, we cannot easily prove a “de Bruijin inequality” for CBLD. Instead, we make extensive use of the *chain lemma* for $\operatorname{D_{KL}}$

$$\operatorname{D_{KL}}(\mu\|\nu) = \mathbb{E}_{\mu_B}[\operatorname{D_{KL}}(\mu_{A|B}\|\nu_{A|B})] + \operatorname{D_{KL}}(\mu_B\|\nu_B).$$

where A, B are disjoint subspaces of $\mathbb{R}^d$, $A \cup B = \mathbb{R}^d$, and μ, ν are measures supported on $\mathbb{R}^d$ with $\mu_{A|B}$ denotes the measure over A conditioned on $B = b$ for arbitrary b. Note that if we set $A = B_i$, $B = \overline{B}_i$, the CBLD diffusion will result in exponential contraction in $\mathbb{E}_{\mu_{\overline{B}_i}}[\operatorname{D_{KL}}(\mu_{B_i|\overline{B}_i}\|\nu_{B_i|\overline{B}_i})]$ while leaving $\operatorname{D_{KL}}(\mu_{\overline{B}_i}\|\nu_{\overline{B}_i})$ constant. CBLD then trivially results in non-increasing $\operatorname{D_{KL}}(\mu_t^k\|\pi_\beta)$, however expressing descent across iterations is more subtle due to the sub-additivity of KL-divergence.

Taking inspiration from Beck & Tetruashvili (2013), we bound descent across b steps, an entire “cycle” over the problem space, expressed in a general lemma for $\operatorname{D_{KL}}(\mu^k\|\pi)$ (proved in Appendix C).

Lemma 1 (Cyclic KL Contraction)**.** *Let the set $C = \{C_1, ..., C_b\} \in \mathbb{R}_+^b$ satisfy $0 < C_i < 1$, and let $D_i \in \mathbb{R}$ be arbitrary constants $D_i \geq 0$ and let π be an arbitrary distribution with finite second moment. Suppose $(\mu^0, \mu^1, ...)$ is a sequence of measures satisfying for $k \geq 1$ and $n = k \bmod b$*

$$\operatorname{D_{KL}}(\mu^k\|\pi_\beta) \leq C_n \operatorname{D_{KL}}(\mu^{k-1}\|\pi) + (1 - C_n) \operatorname{D_{KL}}(\mu^{k-1}_{\overline{B}_n}\|\pi_{\overline{B}_n}) + D_n.$$

Then we can bound

$$\operatorname{D_{KL}}(\mu^{kb}\|\pi) \leq C_{\max}^k \operatorname{D_{KL}}(\mu^{(k-1)b}\|\pi) + \sum_{i=1}^{b} D_i$$

where $C_{\max} = \max\{C_1, ..., C_b\}$.

When $D_i = 0$, Lemma 1 can be seen as an information-theoretic bound on the change in global KL-divergence from the application of factorized noise channels on μ, π_β. Lee (2022) *lower* bounded the KL-divergence in Bayesian coordinate ascent variational inference by similarly comparing the change in $\operatorname{D_{KL}}$ across conditioned steps. However, their focus was on inference over mean-field parametric distributions rather than the broader class of LSI Gibbs measures, making Lemma 1 a stronger result.

The convergence of CBLD follows by choosing $D_i = 0$, $C_{\max} = e^{-2\gamma\beta^{-1}\lambda_{\min}}$:

Theorem 2 (CBLD $\operatorname{D_{KL}}(\mu_t^k\|\pi)$ Convergence)**.** *Let $\theta = (B_1, ..., B_b)$ be a given block permutation and let $\lambda = (\lambda_1, ..., \lambda_b) \in \mathbb{R}^d$ be the sampling times for each block, $\lambda_i > 0$. For any $\pi_\beta \propto \exp[-\beta f(x)]$ satisfying Assumptions 1 and 2, and any $\beta > 0$, the sample distribution after kb steps of CBLD (μ^{kb}) satisfies*

$$\operatorname{D_{KL}}(\mu^{kb}\|\pi) \leq e^{-2\gamma\beta^{-1}\lambda_{\min}k} \operatorname{D_{KL}}(\mu^0\|\pi).$$

When $D_i \neq 0$, Lemma 1 accounts for biased sampling algorithms, such as Langevin Monte Carlo (LMC). Accordingly, we combine Lemma 1 with existing LSI bounds for LMC from Chewi et al. (2021) to prove convergence for a discrete time "cyclic block Langevin Monte Carlo" in Appendix C.

For RBLD and CBLD, the convergence is limited by the shortest step duration $\lambda_{\min}$ and minimal block probability $\phi_{\min}$. For constant block sizes, the optimal choice for both CBLD and RBLD is therefore constant $\lambda_i = \lambda_j = \lambda$ and uniform $\phi_i = 1/b$. This contrasts discrete-time block optimization, where distinct step sizes/probability distributions provide advantage on ill-conditioned problems (Nesterov, 2012; Beck & Tetruashvili, 2013; Ding et al., 2021) due to the effect of varying Lipschitz constants in discretization error terms. In the case of constant λ with uniform ϕ, RBLD and CBLD have identical descent bounds, as we numerically demonstrate in Section 4. This considerably simplifies hyperparameter selection *for ideal devices*, reducing from $O(b)$ parameters to 1 (λ). In the following section we continue to assume a constant step duration λ for simplicity, though future analyses may reveal more optimized step size selections for finite-variation devices.

3.4 FINITE VARIATION

Theorems 1 and 2 provide optimistic lower bounds for DX sampling, however a real machine will have analog errors perturbing the target function (Albash et al., 2019; Melanson et al., 2023). As a generalization of Albash et al. (2019), we model a DX with analog variation with a "perturbed" gradient oracle $g_\delta(x) : \mathbb{R}^d \to \mathbb{R}^d$, where $\delta \in \boldsymbol{D}$ denotes a fixed perturbation from arbitrary domain $\boldsymbol{D}$. Unlike stochastic optimization, which assumes that the perturbation changes with each gradient evaluation, DX perturbations are fixed for each device. To provide guarantees under device variation, we need to restrict the perturbations and functions permitted:

Assumption 3. For fixed $\delta \in \boldsymbol{D}$, there exist constants $M, B \geq 0$ such that

$$\|\nabla f(x(t)) - g_\delta(x(t))\|^2 \leq M^2\|x(t)\|^2 + B^2.$$

Assumption 4. f is L-smooth and, for fixed $\delta \in \boldsymbol{D}$, g_δ is G-Lipschitz continuous. That is, for all $x, y \in \mathbb{R}^d$

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|,$$
$$\|g_\delta(x) - g_\delta(y)\| \leq G\|x - y\|.$$

Assumption 5. f and g_δ are (m, c)-dissipative and $(\mathfrak{m}, \mathfrak{c})$-dissipative respectively, i.e., there exists positive constants $m > 0$, c, $\mathfrak{m}$, $\mathfrak{c} \geq 0$ such that for all $x \in \mathbb{R}^d$:

$$\langle \nabla f(x), x\rangle \geq m\|x\|^2 - c,$$
$$\langle \nabla g_\delta(x), x\rangle \geq \mathfrak{m}\|x\|^2 - \mathfrak{c}.$$

Assumption 3 limits the Euclidean distance between ∇f and g_δ, with the constants M and B appearing in later bounds. Assumption 5 is a common assumption in analyses of stochastic gradient sampling algorithms (Raginsky et al., 2017; Li & Wang, 2022; Zhang et al., 2023). Specifically, it enables us to bound the ideal Langevin second moment $\mathbb{E}\|y^k(t)\|^2$ in the proof of Theorem 3. Assumption 4 is not directly used in our proofs, but is required for a Girsanov change of measure. Assumptions 3 and 5 both restrict the type of perturbation with Assumption 5 also limiting the magnitude. Assumptions 3 and 5 are both reasonable, as DX variation typically manifests as additive or multiplicative perturbations in analog components implementing ∇f Xiao et al. (2022); Aifer et al. (2023).

Take the example of a Gaussian potential $f(x) = \frac{1}{2}x^T\Sigma^{-1}x$ with $g_\delta(x) = \Sigma^{-1}\circ(1+\delta)x$, where $\delta \in \mathbb{R}^{d\times d}$ is a "perturbation matrix" with $\delta_{ij} \sim \mathcal{N}(0, \Delta^2)$ and $\circ$ denotes a component-wise Hadamard product. Regardless of the standard deviation Δ, we satisfy Assumptions 3 and 4 with M and L both equal to the maximal magnitude eigenvalue of δ and Σ^{-1} respectively with $B = 0$. However, if $\Sigma^{-1}(1+\delta)$ has negative eigenvalues there is no $\mathfrak{m} > 0$ satisfying Assumption 5, placing an upper limit on the perturbation strength.

Assumption 6.

The density of the initial law μ_0 satisfies

$$\kappa_0 \triangleq \log \int_{\mathbb{R}^d} e^{\|x\|^d} d\mu_0 < \infty.$$

In practice, dynamical accelerators typically operate over bounded domains, such as the unit hypercube (Afoakwa et al., 2021) or unit circle (Wang & Roychowdhury, 2019; Inagaki et al., 2016), hence the iterate magnitude is bounded in any case. However, bounding over the entire space would provide insufficiently tight upper bounds and our methodology assumes that the measures are supported on $\mathbb{R}^d$. We leave consideration of domains with bounded support to future work, potentially applying methods from reflected Langevin diffusion theory (Bubeck et al., 2018).

We begin by stating the following bound on the distance between the measures of ideal and perturbed BLD, proved in Appendix D:

Lemma 2 (Finite Variation Block Langevin W_2 Distance)**.** *Let $x_k(t)$, $y_k(t)$ be the states of non-ideal and ideal block Langevin systems respectively, with associated probability laws μ_t^k, ν_t^k. For any $\pi_\beta \propto \exp[-\beta f(x)]$ satisfying Assumptions 1, 2, and 5 with $\beta > \frac{2}{m}$, stochastic gradient oracle $g(x)$ satisfying Assumption 3, initial distribution μ^0 satisfying Assumption 6, and $k\lambda > 1$ we have the following bound*

$$W_2(\mu^k, \nu^k) \leq \sqrt{C_0\left[(C_1 + \sqrt{C_1}) + (C_2 + \sqrt{C_2})\sqrt{\lambda}\right]k\lambda}$$

where C_0, C_1, and C_2 are given in Appendix D.

From previous discussions, setting $\phi_i = 1/b$, $\lambda_i = \lambda$ unifies the bounds for RBLD and CBLD. In this regime, we can prove the following statement as a simple consequence of the triangle inequality $W_2(\mu, \nu) \leq W_2(\mu, \eta) + W_2(\eta, \nu)$ and the Otto-Villani theorem

Theorem 3 (Finite-Variation BLD W_2^2 Convergence)**.**

$$W_2(\mu^{bk}, \pi_\beta) \leq \left(\frac{2}{\gamma}\right)^{1/2} e^{-\gamma\beta^{-1}\lambda bk}\sqrt{\mathrm{D}_{\mathrm{KL}}(\mu^0\|\pi)} \\ + \sqrt{C_0\left[(C_1 + \sqrt{C_1}) + (C_2 + \sqrt{C_2})\sqrt{\lambda}\right]bk\lambda}.$$

Following Raginsky et al. (2017), if we choose $k\lambda = \frac{\beta}{b\gamma}\log\frac{2\sqrt{2\,\mathrm{D}_{\mathrm{KL}}(\mu^0\|\pi)}}{\varepsilon\sqrt{\gamma}}$ and set $\lambda \leq \left(\frac{\varepsilon\gamma}{\beta\log[2\sqrt{2\,\mathrm{D}_{\mathrm{KL}}(\mu^0\|\pi)}/(\sqrt{\gamma}\varepsilon)]}\right)^4$, we have

$$W_2(\mu^{bk}, \pi_\beta) \leq \frac{\varepsilon}{2} + \sqrt{C_0}\left[\sqrt{C_1 + \sqrt{C_1}}\frac{\beta}{\gamma}\log\frac{2\sqrt{2\,\mathrm{D}_{\mathrm{KL}}(\mu^0\|\pi)}}{\varepsilon\sqrt{\gamma}} + \varepsilon\sqrt{C_2 + \sqrt{C_2}}\right]. \tag{4}$$

We thereby obtain a total bound on the Wasserstein error $\mathcal{O}(\log\frac{1}{\varepsilon} + \varepsilon)$ for arbitrary $\varepsilon > 0$. Our Wasserstein bound has a finite lower bound with respect to epsilon: non-ideal devices introduce *bias*. Unlike discrete LMC, the bias in Equation (4) does not result from a forward-flow discretization (Wibisono, 2018; Chewi et al., 2021). Instead, the constants C_0, C_1, C_2 are solely due to finite analog variation. For $M = 0$, $B = 0$, we recover exponential, unbiased convergence in W_2. However, akin to LMC, practitioners can select the step size λ and the injected noise β to control the bias. Higher temperatures (lower β) result in a lower bias, as expected from the application of a noisy channel in measure space. Moreover, DX users/designers typically characterize M, B during device calibration: simultaneously lowering the impact of analog non-ideality and allowing for a rough bound on the distribution bias (See Section C.2.a of Melanson et al., 2023).

A Wasserstein bound suffices as a performance guarantee in sampling tasks such as Boltzmann machine inference Hinton et al. (2006) or statistical physics simulation (Hamerly et al., 2019; Ng et al., 2022; Inaba et al., 2023). For optimization, assuming quadratic function growth with $\beta \geq \frac{2}{m}$ and a dissapative gradient oracle (see Appendix D for discussion) allows the use of a continuity inequality (Lemma 6 of Raginsky et al., 2017) and second moment bound (Proposition 11 of Raginsky et al., 2017) to bound $\mathbb{E}_{\mu^k}[f(x)] - \mathbb{E}_{\pi_\beta}[f(x)]$ and $\mathbb{E}_{\pi_\beta}[f(x)] - \min_{x\in\mathbb{R}^d} f(x)$ respectively

$$\mathbb{E}_{\mu^k}[f(x)] - \mathbb{E}_{\pi_\beta}[f(x)] \leq (M\sigma + B)W_2(\mu^{bk}, \pi_\beta), \tag{5}$$

$$\mathbb{E}_{\pi_\beta}[f(x)] - \min_{x\in\mathbb{R}^d} f(x) \leq \frac{d}{2\beta}\log\left(\frac{eL}{m}\left(\frac{c\beta}{d} + 1\right)\right) \tag{6}$$

where $\sigma^2 = \max\{\mathbb{E}_{\mu^k}[x^2], \mathbb{E}_{\pi_\beta}[x^2]\}$ (given in Appendix D). Combining Equations (5) and (6), we obtain

$$\begin{aligned}\mathbb{E}_{\mu^k}[f(x)] - \min_{x\in\mathbb{R}^d} f(x) &\leq \frac{d}{2\beta}\log\left(\frac{eL}{m}\left(\frac{c\beta}{d}+1\right)\right) + (M\sigma + B)W_2(\mu^{bk}, \pi_\beta) \\ &= \mathcal{O}(\frac{d}{\beta}\log\beta d + (M+B)(\varepsilon + \log\varepsilon^{-1})).\end{aligned}$$

Controlling the first term requires increasing β (lower-magnitude Brownian noise) in tandem with problem dimension. Conversely, controlling the second term requires $\lambda k \propto \beta$, $\lambda \propto \frac{1}{\beta^4}$, i.e., more iterations with lower step duration with increasing β. In digital algorithms, we are free to choose λ arbitrarily small to meet given precision requirements (though the program convergence may be impractically slow). Dynamical accelerators typically have a practical lower bound on λ (e.g., a digital clock period), translating into an effective upper bound on β.

4 NUMERICAL EXPERIMENTS

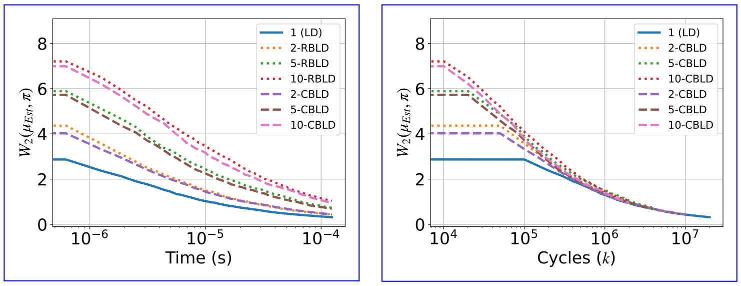
As an illustrative example, we simulated CBLD and RBLD behavior in Gaussian sampling. Gaussian distributions permit closed-form solutions for $W_2(\mathcal{N}(x_1, \Sigma_1), \mathcal{N}(x_2, \Sigma_2))$, allowing for a quantitative estimate of convergence. Moreover, several proposed use cases for DXs rely on Gaussian sampling, including matrix inversion (Aifer et al., 2023) and uncertainty quantification (Melanson et al., 2023). Other works have also proposed using DXs to optimize strongly-convex functions of the form $f(x) = (x-\mu)^T W (x-\mu)$ (Wu et al., 2024; Song et al., 2024), making Gaussian analysis practical as well as tractable. As discussed in the preceding section, our bounds provide expected function gap guarantees from sampling $\pi = \mathcal{N}(0, 2\beta^{-1}W^{-1})$, where optimization would occur in the $\beta \to \infty$ limit.

We simulate DX sampling a $d = 50$ Gaussian with zero mean and a random covariance matrix Σ. Simulation parameters are based on the analog electronic DX proposed in Afoakwa et al. (2021); Sharma et al. (2023); Song et al. (2024), with time determined by the device $R \cdot C$ constant (6.2 ns). For each datapoint, we compute the empirical mean and covariance of the recorded sample distribution. Appendix A gives more details on our experimental configuration.

We focus on the rates of convergence and their dependence on algorithm parameters (step duration λ, block count b, etc.) rather than the exact W_2 value. Fig. 2a shows the convergence in W_2 for ideal-component sampling using block counts $b \in \{1, 2, 5, 10\}$ compared to simulated time ($\mathcal{O}(kb)$) while Fig. 2b shows the same data relative to the number of “whole-space” cycles ($\mathcal{O}(k)$). As expected, BLD requires $\mathcal{O}(b)$ more time to match the W_2 decay of full-gradient LD, with RBLD and CBLD being roughly equivalent. However, normalizing by the block count demonstrates that each method is equally efficient relative to whole-problem cycles, as expected from a simple comparison of Equation (2) with Theorems 1 and 2. Figs. 2c and 2d compare varying step durations λ for $b = 5$ BCLD. While all step sizes lead to the same convergence rate with respect to time, larger step sizes lead to larger decay w.r.t. whole problem cycles, again as expected from Theorems 1 and 2. Finally, we perturb the similarity matrix Σ^{-1} with componentwise variation $\tilde{\Sigma}_{ij} = \Sigma_{ij}(1 + \delta_{ij})$, $\delta_{ij} \sim \mathcal{N}(0, \Delta)$. Fig. 2e shows the impact on W_2 convergence with increasing perturbation strength. For small perturbations, the deviation from ideal is minimal. However, the distribution bias is clear for $\delta = 0.3$. At $\delta = 0.4$, Σ is no longer positive-definite, causing the iterate to diverge (not shown on plot). $\delta \in \{0.1, 0.2, 0.3\}$ satisfied Assumption 5, but $\delta = 0.4$ did not, in line with discussion in the preceding section.

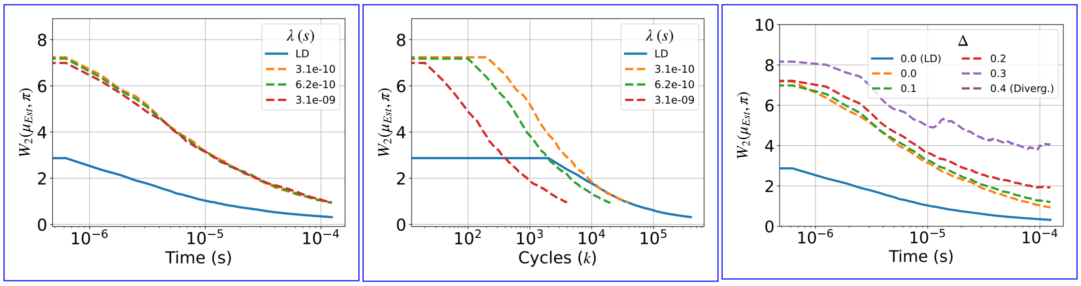
5 CONCLUSION

In this work, we provide novel bounds for hybrid dynamical/digital sampling algorithms leveraging continuous Langevin diffusion. Our bounds extend to both ideal and non-ideal components, with the latter providing an explicit trade-off between W_2 bias, step size, and device non-idealities. We analyze randomized and cyclic selection rules, finding them to be equivalent in D_{KL} contraction with iteration count held constant. Our findings are supported by numerical experiments on Gaussian sampling, observing the expected linear $O(b)$ slowdown in measure-space convergence compared to fully-dynamical LD. Our bounds imply concrete tradeoffs in convergence rate and bias in problem



(a) Convergence in W_2 for varying block counts b with for $b-$BCLD and $b-$RCLD (versus simulated time)

(b) Convergence in W_2 for varying block counts b with for $b-$BCLD and $b-$RCLD (versus cycles kb)



(c) Varying block duration λ (versus simulated time)

(d) Varying block duration λ (versus whole-problem cycles kb)

(e) Varying perturbation deviation δ

Figure 2: Estimated $W_2(\mu_t^k, \pi)$ for BLD sampling methods. (c)-(e) use BLCD, given the close similarity between BCLD and RCLD shown in (a) and (b)

and device constants, providing valuable information to device designers, potential DX adopters, and future analyses.

5.1 LIMITATIONS AND DIRECTIONS FOR FUTURE WORK

In this work, we focused on the continuous activation inference stage. However, even in DX systems model weights are updated in discrete time, either by the digital controller (Song et al., 2024) or by a discrete step in the analog domain (Vengalam et al., 2023). Developing a theoretical framework which provides convergence guarantees for both activation inference and model updates would be a boon to DX research.

We assumed that the inference takes place over $\mathbb{R}^d$, however DXs generally optimize over bounded subspaces such as the unit circle (Inagaki et al., 2016) or unit hypercube (Afoakwa et al., 2021). Previous work on projected (Bubeck et al., 2018) and mirror (Ahn & Chewi, 2021) Langevin dynamics successfully applied LMC methods to constrained sampling. Future work analyzing DX operation using projections could provide concrete bounds for capped-voltage optimizers Afoakwa et al. (2021) and insights from Mirror-LMC could provide insights for DX designers to increase sampling/optimization efficiency.

Assumptions 2 and 5 provide useful bounds for many ML and optimization problems over continuous domains. However, DX applications include discrete choice problems and/or significantly non-convex potentials, such as mixed integer programming. Future bounds necessarily involve more general assumptions than the γ-LSI class considered here. Analog accelerators also typically use low-precision ($< 8b$) DACs and ADCs for input/output (Xiao et al., 2022), making studies of quantizated convergence/expected function gap critical for real-world applications.

Finally, our work focuses on a simplified LNLS framework. While dynamics-accelerated LNLS is popular in literature (Sharma et al., 2022; Raymond et al., 2023; Wu et al., 2024), our work leaves open the question of whether additional digital steps, such as a Metropolis-Hastings filter or replica exchange, could improve the non-asymptotic accuracy or convergence rate.

REFERENCES

Richard Afoakwa, Yiqiao Zhang, Uday Kumar Reddy Vengalam, Zeljko Ignjatovic, and Michael Huang. BRIM: Bistable Resistively-Coupled Ising Machine. In *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pp. 749–760, February 2021. doi: 10.1109/HPCA51647.2021.00068. ISSN: 2378-203X.

Kwangjun Ahn and Sinho Chewi. Efficient constrained sampling via the mirror-Langevin algorithm. In *Advances in Neural Information Processing Systems*, volume 34, pp. 28405–28418. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/hash/ef1e491a766ce3127556063d49bc2f98-Abstract.html`.

Ravindra K. Ahuja, Özlem Ergun, James B. Orlin, and Abraham P. Punnen. A survey of very large-scale neighborhood search techniques. *Discrete Applied Mathematics*, 123(1):75–102, November 2002. ISSN 0166-218X. doi: 10.1016/S0166-218X(01)00338-9. URL `https://www.sciencedirect.com/science/article/pii/S0166218X01003389`.

Maxwell Aifer, Kaelan Donatella, Max Hunter Gordon, Thomas Ahle, Daniel Simpson, Gavin E. Crooks, and Patrick J. Coles. Thermodynamic Linear Algebra, August 2023. URL `http://arxiv.org/abs/2308.05660`. arXiv:2308.05660 [cond-mat, physics:quant-ph].

Tameem Albash, Victor Martin-Mayor, and Itay Hen. Analog errors in Ising machines. *Quantum Science and Technology*, 4(2):02LT03, April 2019. ISSN 2058-9565. doi: 10.1088/2058-9565/ab13ea. URL `https://dx.doi.org/10.1088/2058-9565/ab13ea`. Publisher: IOP Publishing.

Dagur I. Albertsson and Ana Rusu. Ising Machine Based on Bifurcations in a Network of Duffing Oscillators. In *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, Monterey, CA, USA, May 2023. IEEE. ISBN 978-1-66545-109-3. doi: 10.1109/ISCAS46773.2023.10181810. URL `https://ieeexplore.ieee.org/document/10181810/`.

Manuel Arnese and Daniel Lacker. Convergence of coordinate ascent variational inference for log-concave measures via optimal transport, April 2024. URL `http://arxiv.org/abs/2404.08792`. arXiv:2404.08792 [cs, math, stat].

Mohammad Khairul Bashar and Nikhil Shukla. Designing Ising machines with higher order spin interactions and their application in solving combinatorial optimization. *Scientific Reports*, 13 (1):9558, June 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-36531-4. URL `https://www.nature.com/articles/s41598-023-36531-4`. Publisher: Nature Publishing Group.

Roland Bauerschmidt and Thierry Bodineau. A very simple proof of the LSI for high temperature spin systems. *Journal of Functional Analysis*, 276(8):2582–2588, April 2019. ISSN 0022-1236. doi: 10.1016/j.jfa.2019.01.007. URL `https://www.sciencedirect.com/science/article/pii/S0022123619300278`.

Amir Beck and Luba Tetruashvili. On the Convergence of Block Coordinate Descent Type Methods. *SIAM Journal on Optimization*, 23(4):2037–2060, January 2013. ISSN 1052-6234. doi: 10.1137/120887679. URL `https://epubs.siam.org/doi/10.1137/120887679`. Publisher: Society for Industrial and Applied Mathematics.

Anirban Bhattacharya, Debdeep Pati, and Yun Yang. On the Convergence of Coordinate Ascent Variational Inference, June 2023. URL `http://arxiv.org/abs/2306.01122`. arXiv:2306.01122 [cs, math, stat].

François Bolley and Cédric Villani. Weighted csiszár-kullback-pinsker inequalities and applications to transportation inequalities. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 14, pp. 331–352, 2005.

Michael Booth, Steven P Reinhardt, and Aidan Roy. Partitioning Optimization Problems for Hybrid Classical/Quantum Execution. Technical report, D-Wave, 2017.

Robin Brown, Davide Venturelli, Marco Pavone, and David E. Bernal Neira. Accelerating Continuous Variable Coherent Ising Machines via Momentum, January 2024. URL `http://arxiv.org/abs/2401.12135`. arXiv:2401.12135 [quant-ph].

Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a Log-Concave Distribution with Projected Langevin Monte Carlo. *Discrete & Computational Geometry*, 59(4):757–783, June 2018. ISSN 1432-0444. doi: 10.1007/s00454-018-9992-1. URL `https://doi.org/10.1007/s00454-018-9992-1`.

Connor Bybee, Denis Kleyko, Dmitri E. Nikonov, Amir Khosrowshahi, Bruno A. Olshausen, and Friedrich T. Sommer. Efficient optimization with higher-order Ising machines. *Nature Communications*, 14(1):6033, September 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-41214-9. URL `https://www.nature.com/articles/s41467-023-41214-9`. Publisher: Nature Publishing Group.

Martin Chak, Nikolas Kantas, and Grigorios A. Pavliotis. On the Generalized Langevin Equation for Simulated Annealing. *SIAM/ASA Journal on Uncertainty Quantification*, 11(1):139–167, March 2023. doi: 10.1137/21M1462970. URL `https://epubs.siam.org/doi/full/10.1137/21M1462970`. Publisher: Society for Industrial and Applied Mathematics.

Sinho Chewi. *Log-Concave Sampling*. March 2024. URL `https://chewisinho.github.io/main.pdf`.

Sinho Chewi, Murat A. Erdogdu, Mufan Bill Li, Ruoqi Shen, and Matthew Zhang. Analysis of Langevin Monte Carlo from Poincaré to Log-Sobolev, December 2021. URL `http://arxiv.org/abs/2112.12662`. arXiv:2112.12662 [math, stat].

Tzuu-Shuh Chiang, Chii-Ruey Hwang, and Shuenn Jyi Sheu. Diffusion for Global Optimization in $\mathbb{R}^n$. *SIAM Journal on Control and Optimization*, 25(3):737–753, May 1987. ISSN 0363-0129. doi: 10.1137/0325042. URL `https://epubs.siam.org/doi/10.1137/0325042`. Publisher: Society for Industrial and Applied Mathematics.

Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities, December 2016. URL `http://arxiv.org/abs/1412.7392`. arXiv:1412.7392 [math, stat].

Joni Dambre, David Verstraeten, Benjamin Schrauwen, and Serge Massar. Information Processing Capacity of Dynamical Systems. *Scientific Reports*, 2(1):514, July 2012. ISSN 2045-2322. doi: 10.1038/srep00514. URL `https://www.nature.com/articles/srep00514`. Publisher: Nature Publishing Group.

Zhiyan Ding and Qin Li. Langevin Monte Carlo: random coordinate descent and variance reduction. *The Journal of Machine Learning Research*, 22(1):205:9312–205:9362, January 2021. ISSN 1532-4435.

Zhiyan Ding, Qin Li, Jianfeng Lu, and Stephen J. Wright. Random Coordinate Underdamped Langevin Monte Carlo, October 2020. URL `http://arxiv.org/abs/2010.11366`. arXiv:2010.11366 [cs, stat].

Zhiyan Ding, Qin Li, Jianfeng Lu, and Stephen J. Wright. Random Coordinate Langevin Monte Carlo. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pp. 1683–1710. PMLR, July 2021. URL `https://proceedings.mlr.press/v134/ding21a.html`. ISSN: 2640-3498.

Mikhail Erementchouk, Aditya Shukla, and Pinaki Mazumder. On computational capabilities of Ising machines based on nonlinear oscillators,. *Physica D: Nonlinear Phenomena*, 437:133334, September 2022. ISSN 01672789. doi: 10.1016/j.physd.2022.133334. URL `http://arxiv.org/abs/2105.07591`. arXiv:2105.07591 [cond-mat, physics:physics].

Tyler Farghly and Patrick Rebeschini. Time-independent Generalization Bounds for SGLD in Non-convex Settings, November 2021. URL `http://arxiv.org/abs/2111.12876`. arXiv:2111.12876 [cs, math, stat].

Ryan Hamerly, Takahiro Inagaki, Peter L. McMahon, Davide Venturelli, Alireza Marandi, Tatsuhiro Onodera, Edwin Ng, Carsten Langrock, Kensuke Inaba, Toshimori Honjo, Koji Enbutsu, Takeshi Umeki, Ryoichi Kasahara, Shoko Utsunomiya, Satoshi Kako, Ken-ichi Kawarabayashi, Robert L. Byer, Martin M. Fejer, Hideo Mabuchi, Dirk Englund, Eleanor Rieffel, Hiroki Takesue, and Yoshihisa Yamamoto. Experimental investigation of performance differences between coherent Ising machines and a quantum annealer. *Science Advances*, 5(5):eaau0823, May 2019. doi: 10.1126/sciadv.aau0823. URL `https://www.science.org/doi/full/10.1126/sciadv.aau0823`. Publisher: American Association for the Advancement of Science.

Zhezhi He, Jie Lin, Rickard Ewetz, Jiann-Shiun Yuan, and Deliang Fan. Noise Injection Adaption: End-to-End ReRAM Crossbar Non-ideal Effect Adaption for Neural Network Mapping. In *Proceedings of the 56th Annual Design Automation Conference 2019*, DAC '19, pp. 1–6, New York, NY, USA, June 2019. Association for Computing Machinery. ISBN 978-1-4503-6725-7. doi: 10.1145/3316781.3317870. URL `https://dl.acm.org/doi/10.1145/3316781.3317870`.

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

Toshimori Honjo, Tomohiro Sonobe, Kensuke Inaba, Takahiro Inagaki, Takuya Ikuta, Yasuhiro Yamada, Takushi Kazama, Koji Enbutsu, Takeshi Umeki, Ryoichi Kasahara, Ken-ichi Kawarabayashi, and Hiroki Takesue. 100,000-spin coherent Ising machine. *Science Advances*, 7(40):eabh0952, September 2021. doi: 10.1126/sciadv.abh0952. URL `https://www.science.org/doi/full/10.1126/sciadv.abh0952`. Publisher: American Association for the Advancement of Science.

Fangjun Hu, Gerasimos Angelatos, Saeed A. Khan, Marti Vives, Esin Türeci, Leon Bello, Graham E. Rowlands, Guilhem J. Ribeill, and Hakan E. Türeci. Tackling Sampling Noise in Physical Systems for Machine Learning Applications: Fundamental Limits and Eigentasks. *Physical Review X*, 13(4):041020, October 2023. ISSN 2160-3308. doi: 10.1103/PhysRevX.13.041020. URL `https://link.aps.org/doi/10.1103/PhysRevX.13.041020`.

Kensuke Inaba, Yasuhiro Yamada, and Hiroki Takesue. Thermodynamic quantities of two-dimensional Ising models obtained by noisy mean field annealing and coherent Ising machine, February 2023. URL `http://arxiv.org/abs/2302.01454`. arXiv:2302.01454 [cond-mat, physics:quant-ph].

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, November 2016. doi: 10.1126/science.aah4243. URL `https://www.science.org/doi/full/10.1126/science.aah4243`. Publisher: American Association for the Advancement of Science.

David J. Earl and Michael W. Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005. doi: 10.1039/B509983H. URL `https://pubs.rsc.org/en/content/articlelanding/2005/cp/b509983h`. Publisher: Royal Society of Chemistry.

Scott Kirkpatrick. Optimization by simulated annealing: Quantitative studies. *Journal of Statistical Physics*, 34(5):975–986, March 1984. ISSN 1572-9613. doi: 10.1007/BF01009452. URL `https://doi.org/10.1007/BF01009452`.

Michael Klachko, Mohammad Reza Mahmoodi, and Dmitri Strukov. Improving Noise Tolerance of Mixed-Signal Neural Networks. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, July 2019. doi: 10.1109/IJCNN.2019.8851966. URL `https://ieeexplore.ieee.org/abstract/document/8851966`. ISSN: 2161-4407.

Nikola B. Kovachki and Andrew M. Stuart. Continuous time analysis of momentum methods. *The Journal of Machine Learning Research*, 22(1):17:760–17:799, January 2021. ISSN 1532-4435.

Se Yoon Lee. Gibbs sampler and coordinate ascent variational inference: A set-theoretical review. *Communications in Statistics - Theory and Methods*, 51(6):1549–1568, March 2022. ISSN 0361-0926. doi: 10.1080/03610926.2021.1921214. URL `https://doi.org/10.1080/03610926.2021.1921214`. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03610926.2021.1921214.

Benedict Leimkuhler and Charles Matthews. Rational Construction of Stochastic Numerical Methods for Molecular Sampling. *Applied Mathematics Research eXpress*, 2013(1):34–56, January 2013. ISSN 1687-1200. doi: 10.1093/amrx/abs010. URL `https://doi.org/10.1093/amrx/abs010`.

Lei Li and Yuliang Wang. A sharp uniform-in-time error estimate for Stochastic Gradient Langevin Dynamics, October 2022. URL `http://arxiv.org/abs/2207.09304`. arXiv:2207.09304 [cs, math, stat].

Yun Long, Xueyuan She, and Saibal Mukhopadhyay. Design of Reliable DNN Accelerator with Un-reliable ReRAM. In *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1769–1774, March 2019. doi: 10.23919/DATE.2019.8715178. URL `https://ieeexplore.ieee.org/abstract/document/8715178`. ISSN: 1558-1101.

Andrew Lucas. Ising formulations of many NP problems. *Frontiers in Physics*, 2, 2014. ISSN 2296-424X. URL `https://www.frontiersin.org/articles/10.3389/fphy.2014.00005`.

Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I. Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42):20881–20885, October 2019. doi: 10.1073/pnas.1820003116. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1820003116`. Publisher: Proceedings of the National Academy of Sciences.

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, December 2023. URL `http://arxiv.org/abs/2312.04836`. arXiv:2312.04836 [cond-mat].

Beren Millidge, Tommaso Salvatori, Yuhang Song, Rafal Bogacz, and Thomas Lukasiewicz. Predictive Coding: Towards a Future of Deep Learning beyond Backpropagation?, February 2022. URL `http://arxiv.org/abs/2202.09467`. arXiv:2202.09467 [cs].

Naeimeh Mohseni, Peter L. McMahon, and Tim Byrnes. Ising machines as hardware solvers of combinatorial optimization problems. *Nature Reviews Physics*, 4(6):363–379, June 2022. ISSN 2522-5820. doi: 10.1038/s42254-022-00440-8. URL `https://www.nature.com/articles/s42254-022-00440-8`. Number: 6 Publisher: Nature Publishing Group.

Yu. Nesterov. Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems. *SIAM Journal on Optimization*, 22(2):341–362, January 2012. ISSN 1052-6234. doi: 10.1137/100802001. URL `https://epubs.siam.org/doi/abs/10.1137/100802001`. Publisher: Society for Industrial and Applied Mathematics.

Edwin Ng, Tatsuhiro Onodera, Satoshi Kako, Peter L. McMahon, Hideo Mabuchi, and Yoshihisa Yamamoto. Efficient sampling of ground and low-energy Ising spin configurations with a coherent Ising machine. *Physical Review Research*, 4(1):013009, January 2022. ISSN 2643-1564. doi: 10.1103/PhysRevResearch.4.013009. URL `https://link.aps.org/doi/10.1103/PhysRevResearch.4.013009`.

Antonio Orvieto and Aurelien Lucchi. Continuous-time Models for Stochastic Optimization Algorithms, March 2020. URL `http://arxiv.org/abs/1810.02565`. arXiv:1810.02565 [cs, math].

F. Otto and C. Villani. Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality. *Journal of Functional Analysis*, 173(2):361–400, June 2000. ISSN 0022-1236. doi: 10.1006/jfan.1999.3557. URL `https://www.sciencedirect.com/science/article/pii/S0022123699935577`.

Yury Polyanskiy and Yihong Wu. Wasserstein Continuity of Entropy and Outer Bounds for Interference Channels. *IEEE Trans. Inf. Theor.*, 62(7):3992–4002, July 2016. ISSN 0018-9448. doi: 10.1109/TIT.2016.2562630. URL `https://doi.org/10.1109/TIT.2016.2562630`.

Sayantan Pramanik, Sourav Chatterjee, and Harshkumar Oza. Convergence Analysis of Opto-Electronic Oscillator based Coherent Ising Machines, December 2023. URL `http://arxiv.org/abs/2312.04290`. arXiv:2312.04290 [quant-ph].

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, pp. 1674–1703. PMLR, June 2017. URL `https://proceedings.mlr.press/v65/raginsky17a.html`. ISSN: 2640-3498.

Jack Raymond, Radomir Stevanovic, William Bernoudy, Kelly Boothby, Catherine McGeoch, Andrew J. Berkley, Pau Farré, and Andrew D. King. Hybrid quantum annealing for larger-than-QPU lattice-structured problems. *ACM Transactions on Quantum Computing*, 4(3):1–30, September 2023. ISSN 2643-6809, 2643-6817. doi: 10.1145/3579368. URL `http://arxiv.org/abs/2202.03044`. arXiv:2202.03044 [quant-ph].

Benjamin Scellier and Yoshua Bengio. Equilibrium Propagation: Bridging the Gap between Energy-Based Models and Backpropagation. *Frontiers in Computational Neuroscience*, 11, May 2017. ISSN 1662-5188. doi: 10.3389/fncom.2017.00024. URL `https://www.frontiersin.org/articles/10.3389/fncom.2017.00024`. Publisher: Frontiers.

Benjamin Scellier, Maxence Ernoult, Jack Kendall, and Suhas Kumar. Energy-based learning algorithms for analog computing: a comparative study. *Advances in Neural Information Processing Systems*, 36:52705–52731, December 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/hash/a52b0d191b619477cc798d544f4f0e4b-Abstract-Conference.html`.

Anshujit Sharma, Richard Afoakwa, Zeljko Ignjatovic, and Michael Huang. Increasing ising machine capacity with multi-chip architectures. In *Proceedings of the 49th Annual International Symposium on Computer Architecture*, ISCA '22, pp. 508–521, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-8610-4. doi: 10.1145/3470496.3527414. URL `https://dl.acm.org/doi/10.1145/3470496.3527414`.

Anshujit Sharma, Matthew Burns, and Michael C. Huang. Combining Cubic Dynamical Solvers with Make/Break Heuristics to Solve SAT. In Meena Mahajan and Friedrich Slivovsky (eds.), *26th International Conference on Theory and Applications of Satisfiability Testing (SAT 2023)*, volume 271 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 25:1–25:21, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-286-0. doi: 10.4230/LIPIcs.SAT.2023.25. URL `https://drops.dagstuhl.de/opus/volltexte/2023/18487`. ISSN: 1868-8969.

Ruibing Song, Chunshu Wu, Chuan Liu, Ang Li, Michael Huang, and Tong Geng. DS-GL: Advancing Graph Learning via Harnessing the Power of Nature within Dynamic Systems. Technical Report PNNL-SA-196761, Pacific Northwest National Laboratory (PNNL), Richland, WA (United States), August 2024. URL `https://www.osti.gov/biblio/2426329`.

Menachem Stern, Daniel Hexner, Jason W. Rocks, and Andrea J. Liu. Supervised Learning in Physical Networks: From Machine Learning to Learning Machines. *Physical Review X*, 11(2):021045, May 2021. doi: 10.1103/PhysRevX.11.021045. URL `https://link.aps.org/doi/10.1103/PhysRevX.11.021045`. Publisher: American Physical Society.

Ryan J Tibshirani. Dykstra' s Algorithm, ADMM, and Coordinate Descent: Connections, Insights, and Extensions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/5ef698cd9fe650923ea331c15af3b160-Abstract.html`.

Hayato Ushijima-Mwesigwa, Christian F. A. Negre, and Susan M. Mniszewski. Graph Partitioning using Quantum Annealing on the D-Wave System. In *Proceedings of the Second International Workshop on Post Moores Era Supercomputing*, PMES'17, pp. 22–29, New York, NY, USA, November 2017. Association for Computing Machinery. ISBN 978-1-4503-5126-3. doi: 10.1145/3149526.3149531. URL `https://dl.acm.org/doi/10.1145/3149526.3149531`.

Santosh Vempala and Andre Wibisono. Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/65a99bb7a3115fdede20da98b08a370f-Abstract.html`.

Uday Kumar Reddy Vengalam, Yongchao Liu, Tong Geng, Hui Wu, and Michael Huang. SUPPORTING ENERGY-BASED LEARNING WITH AN ISING MACHINE SUBSTRATE: A CASE STUDY ON RBM. In *Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO '23, pp. 465–478, New York, NY, USA, December 2023. Association for Computing Machinery. ISBN 9798400703294. doi: 10.1145/3613424.3614315. URL `https://dl.acm.org/doi/10.1145/3613424.3614315`.

Maxime Vono, Nicolas Dobigeon, and Pierre Chainais. Split-and-Augmented Gibbs Sampler—Application to Large-Scale Inference Problems. *IEEE Transactions on Signal Processing*, 67(6):1648–1661, March 2019. ISSN 1941-0476. doi: 10.1109/TSP.2019.2894825. URL `https://ieeexplore.ieee.org/abstract/document/8625467`. Conference Name: IEEE Transactions on Signal Processing.

Maxime Vono, Daniel Paulin, and Arnaud Doucet. Efficient MCMC sampling with dimension-free convergence rate using ADMM-type splitting. *The Journal of Machine Learning Research*, 23 (1):25:1100–25:1168, January 2022. ISSN 1532-4435.

Tianshi Wang and Jaijeet Roychowdhury. OIM: Oscillator-based Ising Machines for Solving Combinatorial Optimisation Problems, March 2019. URL `http://arxiv.org/abs/1903.07163`. arXiv:1903.07163 [cs].

Zhe Wang, Alireza Marandi, Kai Wen, Robert L. Byer, and Yoshihisa Yamamoto. Coherent Ising machine based on degenerate optical parametric oscillators. *Physical Review A*, 88(6):063853, December 2013. doi: 10.1103/PhysRevA.88.063853. URL `https://link.aps.org/doi/10.1103/PhysRevA.88.063853`. Publisher: American Physical Society.

Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Proceedings of the 31st Conference On Learning Theory*, pp. 2093–3027. PMLR, July 2018. URL `https://proceedings.mlr.press/v75/wibisono18a.html`. ISSN: 2640-3498.

Chunshu Wu, Ruibing Song, Chuan Liu, Yunan Yang, Ang Li, Michael Huang, and Tong Geng. Extending power of nature from binary to real-valued graph learning in real world. In *The Twelfth International Conference on Learning Representations*, 2024.

T. Patrick Xiao, Ben Feinberg, Christopher H. Bennett, Venkatraman Prabhakar, Prashant Saxena, Vineet Agrawal, Sapan Agarwal, and Matthew J. Marinella. On the Accuracy of Analog Neural Network Inference Accelerators. *IEEE Circuits and Systems Magazine*, 22(4):26–48, 2022. ISSN 1558-0830. doi: 10.1109/MCAS.2022.3214409. URL `https://ieeexplore.ieee.org/document/10018023/?arnumber=10018023&tag=1`. Conference Name: IEEE Circuits and Systems Magazine.

Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global Convergence of Langevin Dynamics Based Algorithms for Nonconvex Optimization, October 2020. URL `http://arxiv.org/abs/1707.06618`. arXiv:1707.06618 [cs, math, stat].

Ying Zhang, Ömer Deniz Akyildiz, Theodoros Damoulas, and Sotirios Sabanis. Nonasymptotic Estimates for Stochastic Gradient Langevin Dynamics Under Local Conditions in Nonconvex Optimization. *Applied Mathematics & Optimization*, 87(2):25, January 2023. ISSN 1432-0606. doi: 10.1007/s00245-022-09932-6. URL `https://doi.org/10.1007/s00245-022-09932-6`.

Yiqiao Zhang, Uday Kumar Reddy Vengalam, Anshujit Sharma, Michael Huang, and Zeljko Ignjatovic. QuBRIM: A CMOS Compatible Resistively-Coupled Ising Machine with Quantized Nodal

Interactions. In *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design*, ICCAD '22, pp. 1–8, New York, NY, USA, December 2022. Association for Computing Machinery. ISBN 978-1-4503-9217-4. doi: 10.1145/3508352.3549443. URL `https://dl.acm.org/doi/10.1145/3508352.3549443`.

Yuchen Zhang, Percy Liang, and Moses Charikar. A Hitting Time Analysis of Stochastic Gradient Langevin Dynamics. In *Proceedings of the 2017 Conference on Learning Theory*, pp. 1980–2022. PMLR, June 2017. URL `https://proceedings.mlr.press/v65/zhang17b.html`. ISSN: 2640-3498.

SUPPLEMENTARY MATERIALS

Here we provide proofs and explanations of experimental methods. Additionally, we apply our analysis to bound D_{KL} for discrete-time variants of RBLD and CBLD (RBLMC and CBLMC).

A EXPERIMENTAL METHODS

A.1 DIFFUSION SIMULATION

We simulate Langevin SDEs using a second-order Leimkuhler-Matthews integration scheme (Leimkuhler & Matthews, 2013) with a time step size of 1×10^{-12} seconds. Block diffusions are simulated for a fixed number of steps (50, 100, or 500 for the plots generated), then the block is switched either cyclically or randomly, depending on the algorithm. The total CBLD/RBLD step counts are chosen to ensure that each simulation has the same total simulation time (2×10^7 time steps), allowing for a 1:1 comparison in total optimizer time.

We take the resistively coupled BRIM architecture from Afoakwa et al. (2021) with the Langevin perturbations proposed by Sharma et al. (2023) as our baseline DX. The BRIM architecture is more easily extensible to general classes of real-valued functions (Sharma et al., 2023; Song et al., 2024; Wu et al., 2024) than oscillator-based DXs (Wang & Roychowdhury, 2019; Inagaki et al., 2016), motivating the selection.

We model the device using 310 kOhm resistors and 20 fF capacitors, leading to an RC time constant of 6.2 ns and an effective step size of 6.2 psec, which we use to plot total estimated DX time. These circuit parameters are comparable to those proposed in literature (Afoakwa et al., 2021; Zhang et al., 2022), however different device parameters will simply rescale the x-axis.

A.2 TARGET POTENTIAL

As stated in the main text, we choose a Gaussian target measure to obtain a direct estimate of convergence rather than using proxy statistical observables, as done in Ding & Li (2021). The $d = 50$ Gaussian used to produce Fig. 2 was generated using the following procedure:

1. Generate a 50×50 matrix Σ^{-1} with elements $\sim \mathrm{Unif}[-5, 5]$
2. Make the matrix symmetric by setting $\Sigma^{-1} = \frac{1}{2}(\Sigma^{-1} + (\Sigma^{-1})^T)$
3. If the minimum eigenvalue $\lambda_{\min}$ is negative, set $\Sigma^{-1} = \Sigma^{-1} + 1.2\lambda_{\min} I_{50}$

The resulting matrix is symmetric and positive-definite, making it a valid similarity matrix. We then invert Σ^{-1} to obtain the target covariance matrix Σ. We choose $[-5, 5]$ as the distribution to test a larger range of perturbation strengths $\Delta \in [0.1, 0.4]$, as the W_2 diverged much earlier ($\Delta < 0.2$) with a uniform $[-1, 1]$ distribution.

As our focus is sampling rather than optimization, we set $\beta = 1$ for simplicity. We also assume the Gaussian mean is zero, making the target distribution:

$$\pi(x) \propto e^{-\frac{1}{2}x^T\Sigma^{-1}x}.$$

A.3 SAMPLING PROCEDURE

We recorded one sample after every 10 block updates (thinning parameter of 10) to reduce the impact of highly correlated samples and to make the subsequent steps more computationally efficient. We neglect any burn-in period, as the purpose of our experiment is to see the convergence in distribution (i.e., the necessary length of burn-in). The empirical mean $\overline{x}_k \in \mathbb{R}^d$ and covariance $\Sigma_k \in \mathbb{R}^{d\times d}$ after collecting k samples were computed by

$$\overline{x}_k = \frac{1}{k}\sum_{i=1}^{k} X_{:,i},$$

$$\Sigma_k = \frac{1}{k-1}(X - \overline{x}_k)(X - \overline{x}_k)^T$$

where $X \in \mathbb{R}^{d\times k}$ is the matrix of samples, $\sum_{i=1}^{k} X_{:,i}$ denotes a row(dimension)-wise summation, and $(X - \overline{x}_k)$ is a column (sample)-wise subtraction.

With our estimated distribution, we compute the W_2 distance from the target Gaussian using the closed-form expression

$$W_2^2(\mathcal{N}(\overline{x}_k, \Sigma_k), \mathcal{N}(0, \Sigma)) = ||\overline{x}_k||^2 + \text{Tr}\left[\Sigma_k + \Sigma - 2\sqrt{\sqrt{\Sigma}\Sigma_k\sqrt{\Sigma}}\right]. \tag{7}$$

We compute Equation (7) every 5000 samples to produce the plots in Fig. 2. The 5000 sample count was chosen to balance figure resolution (enough data points) with computational time, as the empirical covariance matrix calculation scaled $O(d^2 \cdot S(t)^2)$, where $S(t)$ is the total number of samples collected up to time t ($O(d \cdot S(T))$ to compute the empirical mean, then $O((d \cdot S(T))^2)$ to compute the outer product).

B RANDOMIZED BLOCK LANGEVIN DIFFUSION (RBLD)

In this section we provide proofs relating to Randomized Block Langevin Diffusion (RBLD, the focus of the main text) and a time-discretized version, Randomized Block Langevin Monte Carlo (RBLMC). RBLMC was previously introduced in Ding et al. (2021) as a coordinate-wise scheme, however we examine block partitions. Moreover, our results using γ-LSI target measures are more general than the strongly log-concave convergence results given in that work.

Algorithm 2 gives the structure of RBLD/RBLMC sampling, where $\phi = \{\phi_1, ..., \phi_b\}$ is a discrete probability mass function over coordinate block indices.

B.1 CONTINUOUS TIME ITERATION

Algorithm 2 Randomized Block Langevin Dynamics (RBLD)

```
1: procedure RBLD($x_0 \in \text{dom}(f)$, Block Distribution $\phi$ over $\{B_1, ..., B_b\}$, Step Size Set $\lambda \in \mathbb{R}_+^b$)
2:     for $k \geq 0$ do
3:         Choose $B_k \sim \phi$
4:         Sample:
               $$x_{k+1} = x_k - \int_0^{\lambda_k} U_k \nabla f(x_k)dt + \int_0^{\lambda_k} U_k\sqrt{2\beta^{-1}}dW_t$$
5:     end for
6: end procedure
```

We first consider the case when each diffusion occurs in continuous time. For a single iteration, we can formulate the evolution of the system by the following Itô SDE:

$$dx = -U_k\nabla\left(f(x)dt + \sqrt{2\beta^{-1}}dW_t\right)$$

To prove continuous-time descent in KL-divergence, we combine standard Langevin gradient flow arguments with methodology inspired by Ref. Vempala & Wibisono (2019) when considering expectation terms.

B.2 FOKKER-PLANCK EQUATION

Let μ_t be the law of x at time t, and let $\mu_{t|0}$ be the measure jointly conditioned ① on the state at time 0 and ② the choice of block B_k. Within a single step, $\mu_{t|0}$ will obey the Fokker-Planck continuity equation

$$\partial_t\mu_{t|0} = Tr[U_k\beta^{-1}\nabla^2\mu_{t|0}] + \text{div}(\mu_{t|0}U_k\nabla f(x_t)).$$

If we were tracking the diffusion over a single block, we would take expectation over the starting state x_0 while conditioning on the block index. However, as discussed in the main text, we take a "meta-Eulerian" perspective. Instead of tracking one block diffusion, our approach finds the average behavior of an ensemble of diffusion processes, each independently sampling their blocks according to ϕ. We therefore take expectation over both x_0 and B_k to derive the change in the "ensemble" measure μ_t.

Therefore we have

$$\partial_t \mu_t = Tr[\beta^{-1} U_\phi \nabla^2 \mu_t] + \mathbb{E}[\text{div}(\mu_{t|0} U_k \nabla f(x))].$$

Where we have defined $U_\phi \triangleq (\phi_1 U_1, ..., \phi_b U_b) \in \mathbb{R}^{d \times d}$.

Let ν be the joint law of (x_0, B_k). Note that

$$\begin{aligned}
\mu_t(x_t|x_0, B_k)\nu(x_0, B_k) &= \mu_t(x_t)\nu(x_0, B_k|x_t) \\
&= \mu_t(x_t)\nu(x_0|x_t, B_k)\nu(B_k|x) \\
&= \mu_t(x_t)\nu(x_0|x_t, B_k)\nu(B_k) \\
&= \mu_t(x_t)\nu(x_0|x_t, B_k)\phi_k.
\end{aligned}$$

Then we can express the second term as

$$\begin{aligned}
\mathbb{E}[\text{div}(\mu_{t|0} U_k \nabla f(x_t))] &= \text{div}(\sum_{i=1}^{b} \int \mu_t(x_t|x_0, i) U_k \nabla f(x_t) \nu(x_0, i) dx_0) \\
&= \text{div}(\sum_{i=1}^{b} \phi_i \int \mu_t(x_t) U_i \nabla f(x_t) \nu(x_0|x_t, i) dx_0) \\
&= \text{div}(\mu_t(x_t) U_\phi \nabla f(x_t))
\end{aligned}$$

since

$$\sum_{i=1}^{b} \phi_i U_i \nabla f(x_t) = U_\phi \nabla f(x_t).$$

Therefore, the FPE of the "meta-Eulerian" RBLD process is

$$\partial_t \mu_t = Tr[\beta^{-1} U_\phi \nabla^2 \mu_t] + \text{div}(\mu_t U_\phi \nabla f(x_t)).$$

Note that the we can use the identity $\nabla f(x) = \beta^{-1} \nabla \log \pi_\beta$ to re-express the FPE as

$$\partial_t \mu_t = \nabla \cdot \left(\beta^{-1} \mu_t U_\phi \nabla \log \frac{\mu_t}{\pi_\beta} \right). \tag{8}$$

B.3 KL-DIVERGENCE CONTRACTION

Lemma 3.

$$\mathrm{D_{KL}}(\mu_t \| \pi) \leq \mathrm{D_{KL}}(\mu_0 \| \pi) e^{-2\beta^{-1}\gamma\lambda_{\min}\phi_{\min}}.$$

Proof. The proof follows conventional analyses of Langevin diffusion processes, e.g., see Vempala & Wibisono (2019); Chewi (2024); Chewi et al. (2021). However, we complete the proof anew for completeness, as well as to show the differences with baseline LD.

With the time evolution of the measure, we can now express the time evolution of the KL-divergence

$$\begin{aligned}
\partial_t \mathrm{D_{KL}}(\mu_t \| \pi) &= \partial_t \int \mu_t(x) \log \frac{\mu_t(x)}{\pi(x)} dx \\
&= \int \partial_t [\mu_t(x) \log \frac{\mu_t(x)}{\pi(x)}] dx \\
&= -\int \partial_t \mu_t(x) \log \frac{\mu_t(x)}{\pi(x)} dx + \overbrace{\int \partial_t \mu_t(x) dx}^{=0}
\end{aligned}$$

where the second term is equal to zero since

$$\int \partial_t \mu_t(x) dx = \partial_t \int \mu_t(x) dx = \partial_t[1] = 0.$$

Using Eqn. (8), we then have

$$\partial_t \operatorname{D}_{\mathrm{KL}}(\mu_t \| \pi) = \int \{\nabla \cdot \left(\beta^{-1} \mu_t U_\phi \nabla \log \frac{\mu_t}{\pi_\beta}\right)\} \log \frac{\mu_t(x)}{\pi(x)} dx.$$

Through integration by parts, we obtain

$$\begin{aligned} \partial_t \operatorname{D}_{\mathrm{KL}}(\mu_t \| \pi) &= -\beta^{-1} \int \left\langle U_\phi \mu_t \nabla \log \frac{\mu_t}{\pi}, \nabla \log \frac{\mu_t(x)}{\pi(x)} \right\rangle dx \\ &= -\beta^{-1} \mathbb{E}_{\mu_t} \left[\left\langle U_\phi \nabla \log \frac{\mu_t}{\pi}, \nabla \log \frac{\mu_t(x)}{\pi(x)} \right\rangle \right]. \end{aligned}$$

U_ϕ is positive-definite with minimum eigenvalue $\phi_{\min}$, therefore

$$\begin{aligned} \partial_t \operatorname{D}_{\mathrm{KL}}(\mu_t \| \pi) =& - \beta^{-1} \mathbb{E}_{\mu_t} \left[\left\langle U_\phi \nabla \log \frac{\mu_t}{\pi}, \nabla \log \frac{\mu_t(x)}{\pi(x)} \right\rangle \right] \leq -\beta^{-1} \phi_{\min} \mathbb{E}_{\mu_t} \left[\left|\left| \nabla \log \frac{\mu_t}{\pi} \right|\right|^2 \right] \\ =& - \beta^{-1} \phi_{\min} FI(\mu_t \| \pi) \leq -2\beta^{-1} \gamma \phi_{\min} \operatorname{D}_{\mathrm{KL}}(\mu_t \| \pi) \end{aligned}$$

where the last inequality utilizes the γ-LSI. Here we highlight a principle difference between LD and RBLD analysis. In LD, we have the "de Brujin *identity*"

$$\partial_t \operatorname{D}_{\mathrm{KL}}(\mu_t \| \pi) = -2\beta^{-1} \gamma FI(\mu_t \| \pi).$$

However, for RBLD we have a "de Brujin *in*equality"

$$\partial_t \operatorname{D}_{\mathrm{KL}}(\mu_t \| \pi) \leq -\beta^{-1} \gamma \phi_{\min} FI(\mu_t \| \pi).$$

We now integrate up to λ_k. Since this step size depends on the choice of B_k, we take expectation of $\operatorname{D}_{\mathrm{KL}}(\mu_k \| \pi)$ over k

$$\mathbb{E}[\operatorname{D}_{\mathrm{KL}}(\mu_k \| \pi)] \leq \mathbb{E}[e^{-2\gamma\beta^{-1}\phi_{\min}\lambda_i}] \operatorname{D}_{\mathrm{KL}}(\mu_{k-1} \| \pi)$$

or deterministically

$$\operatorname{D}_{\mathrm{KL}}(\mu_k \| \pi) \leq e^{-2\gamma\beta^{-1}\phi_{\min}\lambda_{\min}} \operatorname{D}_{\mathrm{KL}}(\mu_{k-1} \| \pi).$$

Expanding the inequality k times yields the result. □

B.4 RCLMC: Euler-Maruyama Discretization

We now extend our analysis to discrete-time Randomized Block Langevin Monte Carlo (RBLMC), shown in Algorithm. 3. While the continuous-time diffusion can be implemented on dynamical hardware, digital applications require an error bound in the discrete-setting. The following derivation closely follows the methods of Vempala & Wibisono (2019) by modeling the divergence of the discrete scheme from a continuous-time interpolation.

Algorithm 3 Randomized Block Langevin Monte Carlo (RBLMC)

```
procedure RBLMC($x_0 \in \mathrm{dom}(f)$, Block Distribution $\phi$ over $\{B_1, ..., B_b\}$, Step Size Set $\lambda \in \mathbb{R}_+^b$)
    for $k \geq 0$ do
        Choose $B_k \sim \phi$, sample $\xi \sim \mathcal{N}(0, 1)$
        Set:
            $x_{k+1} = x_k - \lambda_k U_k \nabla f(x_k) + U_k \sqrt{2\beta^{-1}\lambda_k} \xi$
    end for
end procedure
```

We now consider the SDE

$$dx = U_k[-\nabla f(x_0)dt + \sqrt{2\beta^{-1}}dW_t]$$

where x_0 is the initial state. The SDE has the solution

$$x_t = x_0 + U_k[-\nabla f(x_0)t + \sqrt{2\beta^{-1}t}\xi_t]$$

for $t \in [0, \lambda_n]$ and $\xi_t \sim \mathcal{N}(0, I^{d_i})$. Conditioned on the initial state x_0 and the choice of i, we have the FPE

$$\partial_t \mu_{t|k,x_0} = \beta^{-1}\nabla^2 \cdot \mu_{t|k,x_0} + \nabla \cdot \mu_{t|k,x_0} U_i \nabla f(x_0).$$

Taking expectation over both sides (as previously) yields

$$\partial_t \mu = Tr[\beta^{-1}U_\phi \nabla^2 \mu_{t|k,x_0}] + \nabla \cdot \mathbb{E}[\mu_{t|k,x_0} U_i \nabla f(x_0)].$$

Again noting that the choice of block and the initial state x_0 are independent, we can express the expectation as

$$\mathbb{E}[\mu_{t|k,x_0} U_k \nabla f(x_0)] = \sum_{i=1}^{b} \phi_i \int \mu(x_t|i, x_0)\nu(x_0) U_i \nabla f(x_0) dx_0.$$

Note that while x_0 and ϕ_i are independent random variables, they are not independent when conditioned on x_t. We then have

$$\begin{aligned}
\phi_i \mu(x_t|i, x_0)\nu(x_0) &= \mu(x_t|i, x_0)\nu(x_0, i) \\
&= \mu(x_t)\nu(x_0, i|x_t) \\
&= \mu(x_t)\nu(x_0|x_t, i)\phi_{i|x_t} \\
&= \mu(x_t)\nu(x_0|x_t, i)\phi_i.
\end{aligned}$$

Then

$$\begin{aligned}
\sum_{i=1}^{b} \phi_i \int \mu(x_t|i, x_0)\nu(x_0) U_i \nabla f(x_0) dx_0 &= \sum_{i=1}^{b} \phi_i \int \mu(x_t)\nu(x_0|x_t, i) U_i \nabla f(x_0) dx_0 \\
&= \mu(x_t) \int \nu(x_0|x_t) U_\phi \nabla f(x_0) dx_0 \\
&= \mu_t U_\phi \mathbb{E}[\nabla f(x_0)].
\end{aligned}$$

We then have the following FPE

$$\begin{aligned}
\partial_t \mu_t &= Tr[\beta^{-1}U_\phi \nabla^2 \mu_t] + \nabla \cdot [\mu_t U_\phi \mathbb{E}[\nabla f(x_0)]] \\
&= \nabla \cdot [\beta^{-1} U_\phi \nabla \mu_t + \mu_t U_\phi \mathbb{E}[\nabla f(x_0)]].
\end{aligned}$$

Combining our previous argument with the analysis of Vempala & Wibisono (2019), we have

$$\begin{aligned}
\partial_t \operatorname{D_{KL}}(\mu_t \| \pi) &= \partial_t \int \mu_t(x) \log \frac{\mu_t(x)}{\pi(x)} dx \\
&= \int \partial_t \mu_t(x) \log \frac{\mu_t(x)}{\pi(x)} dx \\
&= \int \nabla \cdot [\beta^{-1} U_\phi \nabla \mu_t + \mu_t U_\phi \mathbb{E}[\nabla f(x_0)]] \log \frac{\mu_t(x)}{\pi(x)} dx \\
&= -\int \left\langle \beta^{-1} U_\phi \nabla \mu_t + \mu_t U_\phi \mathbb{E}[\nabla f(x_0)]], \nabla \log \frac{\mu_t(x)}{\pi(x)} \right\rangle dx \\
&= -\int \left\langle \beta^{-1} U_\phi \mu_t \nabla \log \mu_t + \beta^{-1} U_\phi \mu_t \nabla \log \pi - \beta^{-1} \mu_t \nabla \log \pi + \mu_t U_\phi \mathbb{E}[\nabla f(x_0)]], \nabla \log \frac{\mu_t(x)}{\pi(x)} \right\rangle dx \\
&= -\int \left\langle \beta^{-1} U_\phi \mu_t \nabla \log \frac{\mu_t}{\pi} + \mu_t U_\phi \mathbb{E}[\nabla f(x_0) - \nabla f(x_t)]], \nabla \log \frac{\mu_t(x)}{\pi(x)} \right\rangle dx \\
&= -\beta^{-1} \mathbb{E}[\| U_\phi^{1/2} \nabla \log \frac{\mu_t(x)}{\pi(x)} \|^2] + \mathbb{E}[\left\langle U_\phi^{1/2} \mathbb{E}[\nabla f(x_t) - \nabla f(x_0)], U_\phi^{1/2} \nabla \log \frac{\mu_t(x)}{\pi(x)} \right\rangle].
\end{aligned}$$

where we have used the fact that U_ϕ is a diagonal matrix with non-negative entries, so $U_\phi = U_\phi^{1/2}U_\phi^{1/2} = (U_\phi^{1/2})^T U_\phi^{1/2}$. Then we have (by Cauchy-Schwartz and Young's)

$$\begin{aligned}\mathbb{E}[\left\langle U_\phi^{1/2}\mathbb{E}[\nabla f(x_t) - \nabla f(x_0)], U_\phi^{1/2}\nabla \log\frac{\mu_t(x)}{\pi(x)}\right\rangle] &\leq \mathbb{E}[\|U_\phi^{1/2}\mathbb{E}[\nabla f(x_t) - \nabla f(x_0)]\|^2] + \frac{1}{4}\mathbb{E}\|U_\phi^{1/2}\nabla\log\frac{\mu_t(x)}{\pi(x)}\|^2\\ &= \mathbb{E}[\|U_\phi^{1/2}[\nabla f(x_t) - \nabla f(x_0)]\|^2] + \frac{1}{4}\mathbb{E}\|U_\phi^{1/2}\nabla\log\frac{\mu_t(x)}{\pi(x)}\|^2.\end{aligned}$$

We can decompose the first term as

$$\mathbb{E}[\|U_\phi^{1/2}[\nabla f(x_t) - \nabla f(x_0)]\|^2] = \sum_{i=1}^{b}\phi_i\|U_k\nabla f(x_t) - U_k\nabla f(x_0)\|^2.$$

In line with the presentation in the draft Chewi (2024) we apply Lemma 16 from Chewi et al. (2021), which only requires smoothness and L^2 integrability in the marginal potential:

Lemma 4 (Lemma 16 of Chewi et al. (2021))**.** *Assume probability measure $\pi \propto e^{-f(x)} \in \mathcal{P}^2(\mathbb{R}^d)$ has L-smooth potential f. Then for any probability measure μ*

$$\mathbb{E}_\mu[\|\nabla f\|^2] \leq FI(\mu\|\pi) + 2dL.$$

By the smoothness of f, we have:

$$\begin{aligned}\mathbb{E}\|U_k\nabla f(x_t) - U_k\nabla f(x_0)\|^2 &\leq 2L_i^2\mathbb{E}\|x_t - x_0\|^2 = 2L_i^2\mathbb{E}\|U_k t\nabla f(x_0) + U_k\sqrt{2}W_t\|^2\\ &\leq 2L_i^2t^2\mathbb{E}\|U_k\nabla f(x_0) + U_k\nabla f(x_t) - U_k\nabla f(x_t)\|^2 + \mathbb{E}[2d_iL_i^2t]\\ &\leq 2L_i^2t^2\mathbb{E}\|U_k\nabla f(x_0) - U_k\nabla f(x_t)\|^2 + 2L_i^2\mathbb{E}\|U_k\nabla f(x_t)\|^2 + \mathbb{E}[2d_iL_i^2t].\end{aligned}$$

Suppose $t \leq \lambda_i \leq \frac{1}{2L_i}$, then

$$\mathbb{E}\|U_k\nabla f(x_t) - U_k\nabla f(x_0)\|^2 \leq \frac{1}{2}\mathbb{E}\|U_k\nabla f(x_0) - U_k\nabla f(x_t)\|^2 + 2L_i^2\mathbb{E}\|U_k\nabla f(x_t)\|^2 + \mathbb{E}[2d_iL_i^2t].$$

Hence

$$\mathbb{E}\|U_k\nabla f(x_t) - U_k\nabla f(x_0)\|^2 \leq 4L_i^2\mathbb{E}\|U_k\nabla f(x_t)\|^2 + \mathbb{E}[4d_iL_i^2t].$$

Plugging in Lemma 4 yields

$$\mathbb{E}\|U_k\nabla f(x_t) - U_k\nabla f(x_0)\|^2 \leq 4L_i^2FI(\mu\|\pi) + \mathbb{E}[8td_iL_i^3 + 4d_iL_i^2t].$$

Assume $\lambda_i \leq \frac{\sqrt{\phi_{\min}}}{4L_i}$. Then

$$\begin{aligned}\mathbb{E}[\sum_{i=1}^{b}\phi_i[4t^2L_i^2FI(\mu_{B_i}\|\pi_{B_i}) + 8dL_i^3t^2 + 4d_iL_i^2t]] &\leq \mathbb{E}[\sum_{i=1}^{b}\frac{\phi_{\min}\phi_i}{4}FI(\mu_{B_i}\|\pi_{B_i}) + \sum_{i=1}^{b}\phi_i[8dL_i^3t^2 + 4d_iL_i^2t]]\\ &\leq FI(\mu_t\|\pi)\mathbb{E}[\sum_{i=1}^{b}\phi_i\frac{\phi_{\min}}{4} + \sum_{i=1}^{b}\phi_i[8d_iL_i^3t^2 + 4d_iL_i^2t]]\\ &= \frac{\phi_{\min}}{4}FI(\mu_t\|\pi) + \mathbb{E}[\sum_{i=1}^{b}\phi_i[8d_iL_i^3t^2 + 4d_iL_i^2t]]\\ &\leq \frac{\phi_{\min}}{4}FI(\mu_t\|\pi) + \mathbb{E}[6d_iL_i^2t].\end{aligned}$$

We then have

$$\partial_t\,\mathrm{D_{KL}}(\mu_t\|\pi) \leq -\frac{\phi_{\min}}{2}FI(\mu_t\|\pi) + 6\mathbb{E}[d_iL_i^2]t \leq -\phi_{\min}\gamma\,\mathrm{D_{KL}}(\mu_t\|\pi) + 6\mathbb{E}[d_iL_i^2t].$$

We start by multiplying both sides by $e^{-\phi_{\min}\gamma t}$ and integrating from $t = 0$ to λ_1

$$KL(\mu_{\lambda_i} \| \pi) \leq e^{-\beta^{-1}\phi_{\min}\lambda_i} \, \mathrm{D}_{\mathrm{KL}}(\mu_0 \| \pi) + 3\mathbb{E}[d_i L_i^2 \lambda_i^2].$$

Taking expectation over i then gives the result

$$\mathbb{E}_i[KL(\mu_{\lambda_i} \| \pi)] \leq \mathbb{E}_i[e^{-\gamma\beta^{-1}\phi_{\min}\lambda_i}] \, \mathrm{D}_{\mathrm{KL}}(\mu_0 \| \pi) + 3\mathbb{E}[d_i L_i^2 \lambda_i^2].$$

Iterating B.4 gives

$$\begin{aligned}\mathbb{E}[KL(\mu^k \| \pi)] &\leq \mathbb{E}[e^{-\beta^{-1}\gamma\phi_{\min}\lambda_{\min}}]^k \, \mathrm{D}_{\mathrm{KL}}(\mu_0 \| \pi) + 3\mathbb{E}[d_i L_i^2 \lambda_i^2] \sum_{i=0}^{k} \mathbb{E}[e^{-\beta^{-1}\gamma\phi_{\min}\lambda_{\min}}]^i \\ &\leq e^{-\gamma\phi_{\min}\lambda_i k} \, \mathrm{D}_{\mathrm{KL}}(\mu_0 \| \pi) + \frac{4\beta}{\gamma\phi_{\min}\lambda_{\min}} \mathbb{E}[d_i L_i^2 \lambda_i^2],\end{aligned}$$

where we first bound using the minimum step size, then apply the power series bound

$$\sum_{i=0}^{k} \mathbb{E}[e^{-\beta^{-1}\gamma\phi_{\min}\lambda_i}]^i \leq \sum_{i=0}^{k} e^{-\gamma\phi_{\min}\lambda_{\min}} \leq \frac{1}{1 - e^{-\gamma\phi_{\min}\lambda_{\min}}}$$

and then apply $\frac{1}{1-e^{-a}} \leq \frac{4}{3a}$ to obtain

$$\frac{1}{1 - e^{-\beta^{-1}\gamma\phi_{\min}\lambda_{\min}}} \leq \frac{4}{3\beta^{-1}\gamma\phi_{\min}\lambda_{\min}}.$$

C CYCLIC BLOCK LANGEVIN DIFFUSION

In this section we provide proofs relating to Cyclic Block Langevin Diffusion (CBLD, the focus of the main text) and a time-discretized version, Cyclic Block Langevin Monte Carlo (CBLMC).

The CBLD sampling algorithm is shown in Algorithm 4:

Algorithm 4 Cyclic Block Langevin Diffusion/Monte Carlo (CBLD)

```
procedure CBLD($x_0 \in \mathrm{dom}(f)$, Block Permutation $\sigma = \{B_1, ..., B_b\}$, Step Sizes $\lambda \in \mathbb{R}^b_+$)
    for $k \geq 0$ do
        Set $x_0^{k+1} = x^k$
        for $n = 1$ to $b$ do
            Choose $B_n = \sigma_n$
            Sample:
                $x_{n+1} = x_n - \int_0^{\lambda_n} U_n \nabla f(x_k) dt + \int_0^{\lambda_n} U_n \sqrt{2\beta^{-1}\lambda_n} dW_t$
        end for
        Set $x^{k+1} = x_b^{k+1}$
    end for
end procedure
```

A crucial identity used in our analysis is the "chain lemma" for KL-divergence. For any two distributions μ

$$\mathrm{D}_{\mathrm{KL}}(\mu_t \| \pi) = \mathbb{E}[\mathrm{D}_{\mathrm{KL}}(\mu_{t|B} \| \pi_{|B})] + \mathrm{D}_{\mathrm{KL}\ B}(\mu_t | \pi)$$

where B is a subspace of $\mathbb{R}^d$, $\mathrm{D}_{\mathrm{KL}}(\mu_{t|B} \| \pi_{|B})$ is the KL-divergence of μ_t and π conditioned on an element of B, and $\mathrm{D}_{\mathrm{KL}\ B}(\mu_t | \pi)$ is the KL-divergence of μ_t and π marginalized over $\mathbb{R} \setminus B$. We also state two trivial lemmas for any γ-LSI distribution ν. We first state an equivalent definition of Assumption 2.

Definition 1 (Alternative LSI)**.** $\pi_\beta \propto \exp[-\beta f(x)]$ satisfies a log-Sobolev inequality (LSI) with $C_{LSI} = \frac{1}{\gamma}$ if for all smooth g:

$$\mathbb{E}_{\pi}[g^2 \log g^2] - \mathbb{E}_{\pi_\beta}[g^2] \log \mathbb{E}_{\pi_\beta}[g^2] \le \frac{1}{2\gamma} \mathbb{E}_{\pi_\beta}[\|\nabla g\|^2]$$

where the equivalence with the previous statement follows by choosing $g^2(x) = \frac{\mu(x)}{\pi_\beta(x)}$.

Lemma 5. *Suppose A, B are disjoint subspaces of $\mathbb{R}^d$ with $A \cup B = \mathbb{R}^d$. Then the A marginal ν_A also satisfies γ-LSI.*

Proof. By the LSI, for any smooth $g : \mathbb{R}^d \to \mathbb{R}$

$$\mathbb{E}_\nu \left[g^2 \log g^2\right] - \mathbb{E}_\nu \left[g^2\right] \log \mathbb{E}_\nu \left[g^2\right] \le \mathbb{E}_\nu \left[\|\nabla g\|^2\right].$$

For $g : A \to \mathbb{R}$, we can re-express the terms as

$$\mathbb{E}_{\nu_{B|A}} \mathbb{E}_\nu A[\left[g^2 \log g^2\right] - \mathbb{E}_{\nu_{B|A}} \left(\mathbb{E}_\nu A \left[g^2\right] \log \mathbb{E}_\nu \left[g^2\right]\right) \le \mathbb{E}_{\nu_{B|A}} \mathbb{E}_\nu A \left[\|\nabla g\|^2\right].$$

Since $\mathbb{E}_{\nu_{B|A}}[g(z)] = g(z)$ for all $z \in A$, we simplify to

$$\mathbb{E}_\nu A \left[g^2 \log g^2\right] - \mathbb{E}_\nu A \left[g^2\right] \log \mathbb{E}_\nu \left[g^2\right] \le \mathbb{E}_\nu A \left[\|\nabla g\|^2\right].$$

□

Recall that the sub-step dynamics are described by the SDE

$$dx = U_n \left[-\nabla f(x) dt + \sqrt{2\beta} dW_t\right]. \tag{9}$$

We can then derive the coordinate Fokker-Planck equation:

Lemma 6. *Let $\mu_{t|x_0}$ be the law of x at time $t \in [0, \lambda_n]$ described by the SDE in Equation (9), where $\mu_{t|x_0}$ is conditioned on the starting state x_0. Then $\partial_t \mu_{t,\overline{B}_k|x_0} = 0$ and*

$$\partial_t \mu_{t,B_n|\overline{B}_k,x_0} = \beta^{-1} \nabla^2 \cdot \mu_{B_n|\overline{B}_k,x_0} + \nabla \cdot (\mu_{B_n|\overline{B}_k,x_0} \nabla f(x_t))$$

is the Fokker-Planck equation for the subspace diffusion.

Proof. The second claim is trivially shown using Itô's Lemma. Note that since $\mu_{t,B_n|\overline{B}_n,x_0}$ is only supported on B_n:

1. $Tr[\beta^{-1} U_n \nabla^2 \mu_{t,B_n|\overline{B}_n,x_0}] = \beta^{-1} \nabla^2 \cdot \mu_{t,B_n|\overline{B}_n,x_0}$
2. $\nabla \cdot \mu_{t,B_n|\overline{B}_n,x_0} \nabla f(x_t) = \nabla \cdot \mu_{t,B_n|\overline{B}_n,x_0} \nabla f(x_t)$

Then we have

$$\partial_t \mu_{t,B_n|\overline{B}_k,x_0} = \beta^{-1} \nabla^2 \cdot \mu_{B_n|\overline{B}_k,x_0} + \nabla \cdot (\mu_{B_n|\overline{B}_k,x_0} \nabla f(x_t)).$$

We now use this to prove the first claim.

Consider the law of x in sub-step n conditioned on the initial state x_0 given by $\mu_{t|x_0}$. Note that $\mu_{t|x_0} = \mu_{t,B_n|\overline{B}_n,x_0} \mu_{t,\overline{B}_n|x_0}$.

By the Fokker-Planck equation associated with the SDE and the product rule, we have:

$$\begin{aligned}
\partial_t \mu_{t|x_0} &= \beta^{-1} Tr[U_n^T \nabla^2 \mu_{t|x_0}] + \nabla \cdot (\mu_{t|x_0} U_n \nabla f(x)) \\
\mu_{t,\overline{B}_n|x_0} \partial_t \mu_{t,B_n|\overline{B}_n,x_0} + \mu_{t,B_n|\overline{B}_n,x_0} \partial_t \mu_{t,\overline{B}_n|x_0} &= \beta^{-1} Tr[U_n^T \nabla^2 \mu_{t,B_n|\overline{B}_n,x_0} \mu_{t,\overline{B}_n|x_0}] \\
&\quad + \nabla \cdot (\mu_{t,B_n|\overline{B}_n,x_0} \mu_{t,\overline{B}_n|x_0} U_n \nabla f(x)).
\end{aligned}$$

Note that

$$\beta^{-1}Tr[U_n^T\nabla^2\mu_{t,B_n|\overline{B}_n,x_0}\mu_{t,\overline{B}_n|x_0}] = \beta^{-1}\mu_{t,\overline{B}_n|x_0}\nabla^2\cdot\mu_{t,B_n|\overline{B}_n,x_0}$$

and

$$\nabla\cdot(\mu_{t,B_n|\overline{B}_n,x_0}\mu_{t,\overline{B}_n|x_0}U_n\nabla f(x)) = \mu_{t,\overline{B}_n|x_0}\nabla\cdot(\mu_{t,B_n|\overline{B}_n,x_0}\nabla f(x)).$$

We then have

$$\mu_{t,\overline{B}_n|x_0}\overbrace{(\partial_t\mu_{t,B_n|\overline{B}_n,x_0} - \beta^{-1}\nabla^2\cdot\mu_{t,B_n|\overline{B}_n,x_0} - \nabla\cdot(\mu_{t,B_n|\overline{B}_n,x_0}\nabla f(x)))}^{①} = \mu_{t,B_n|\overline{B}_n,x_0}\partial_t\mu_{t,\overline{B}_n|x_0}.$$

We assume that μ_t is supported on $\mathbb{R}^d$, therefore $\mu_{t|x_0} = \mu_{t,\overline{B}_n|x_0}\mu_{t,B_n|\overline{B}_n,x_0} > 0$.

As previously discussed, Itô's lemma implies ① is 0. For equality to hold, then, $\partial_t\mu_{t,\overline{B}_n} = 0$. ☐

We prove the following technical lemma for later use in the descent bound:

Lemma 7. *Suppose* A, B *are disjoint subspaces of* $\mathbb{R}^d$*. Then we have*

$$\mathrm{D}_{\mathrm{KL}}(\mu_A\|\pi_A) \leq \mathrm{D}_{\mathrm{KL}}(\mu_{A|B}\|\pi_{A|B}).$$

Proof. Note that for all $x \in A$

$$\mu_A(x) = \int_B \mu_{A,B}(x,y)dy = \int_B \mu_A(x|y)\mu_B(y)dy = \mathbb{E}_{y\in B}[\mu_A(x|y)] \triangleq \mathbb{E}_B[\mu_{A|B}].$$

By the convexity of the KL-divergence and Jensen's Inequality

$$\mathrm{D}_{\mathrm{KL}}(\mu_A\|\pi_A) = KL(\mathbb{E}_B[\mu_{A|B}]\|\mathbb{E}_B[\pi_{A|B}]) \leq \mathbb{E}_B[\mathrm{D}_{\mathrm{KL}}(\mu_{A|B}\|\pi_{A|B})] \triangleq \mathrm{D}_{\mathrm{KL}}(\mu_{A|B}\|\pi_{A|B}).$$

☐

Lemma 7 can be considered a restatement of the "data processing inequality". Removing the conditioning on subspace B effectively reduces the available information, akin to a noisy channel, decreasing the divergence between distributions.

Lemma 8.

$$\mathrm{D}_{\mathrm{KL}}(\mu_n\|\pi_\beta) \leq e^{-2\gamma\beta^{-1}\lambda_n}\left[\mathrm{D}_{\mathrm{KL}}(\mu_{n-1}\|\pi)\right] + (1 - e^{-2\gamma\beta^{-1}\lambda_n})\,\mathrm{D}_{\mathrm{KL}\,\overline{B}_1}(\mu_0\|\pi)$$

Proof. Using Lemma 6, we can show by standard arguments Vempala & Wibisono (2019); Chewi et al. (2021) that within sub-step n:

$$\mathrm{D}_{\mathrm{KL}}(\mu_{t|\overline{B}_1}\|\pi_{\overline{B}_1}) \leq \mathrm{D}_{\mathrm{KL}}(\mu_{0|\overline{B}_1}\|\pi_{\overline{B}_1})e^{-2\gamma\beta^{-1}t} \tag{10}$$

Using (10) and the chain rule for KL-divergence

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(\mu_n\|\pi) &= \mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}(\mu_{1|B}\|\pi_{|B})\right] + \mathrm{D}_{\mathrm{KL}\,\overline{B}_1}(\mu_0|\pi) \\
&\leq e^{-2\gamma\lambda_n\beta^{-1}}\mathbb{E}\left[\mathrm{D}_{\mathrm{KL}}(\mu_{n-1|B}\|\pi_{|B})\right] + \mathrm{D}_{\mathrm{KL}\,\overline{B}_1}(\mu_{n-1}|\pi) \\
&= e^{-2\gamma\lambda_n\beta^{-1}}\left[\mathrm{D}_{\mathrm{KL}}(\mu_{n-1}\|\pi) - \mathrm{D}_{\mathrm{KL}\,\overline{B}_1}(\mu_{n-1}|\pi)\right] + \mathrm{D}_{\mathrm{KL}\,\overline{B}_1}(\mu_{n-1}|\pi) \\
&= e^{-2\gamma\lambda_n\beta^{-1}}\left[\mathrm{D}_{\mathrm{KL}}(\mu_{n-1}\|\pi)\right] + (1 - e^{-2\gamma\lambda_n\beta^{-1}})\,\mathrm{D}_{\mathrm{KL}\,\overline{B}_1}(\mu_{n-1}\|\pi).
\end{aligned}$$

☐

An immediate consequence of Lemma 8 is that the KL-divergence is non-increasing, as stated in the following Corollary.

Corollary 1. *For all* $i \in \{1, ..., b\}$, $\mathrm{D}_{\mathrm{KL}}(\mu_i\|\pi) \leq \mathrm{D}_{\mathrm{KL}}(\mu_0\|\pi)$

C.1 PROOF OF LEMMA 1

Proof. We prove the claim by induction on b. The claim is immediately evident for $b = 1$ as a consequence of 8 with $C_{\max} = C_1$, since $\overline{B}_1 = \emptyset$.

Now we assume the inductive hypothesis for some $b - 1 \geq 1$ and prove the claim for $b \geq 2$ blocks.

We start by applying Lemma 8 twice to obtain terms relating to step $b - 2$, obtaining

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(\mu_b \| \pi) \leq & C_b \, \mathrm{D}_{\mathrm{KL}}(\mu_{b-1} \| \pi) + (1 - C_b) \, \mathrm{D}_{\mathrm{KL}}(\mu_{b-1,\overline{B}_b} \| \pi_{\overline{B}_b}) + D_b \\
\text{Second descent expansion: } \leq & C_b C_{b-1} \, \mathrm{D}_{\mathrm{KL}}(\mu_{b-2} \| \pi) + (1 - C_b) \, \mathrm{D}_{\mathrm{KL}}(\mu_{b-1,\overline{B}_b} \| \pi_{\overline{B}_b}) \\
& + C_b(1 - C_{b-1}) \, \mathrm{D}_{\mathrm{KL}}(\mu_{b-2,\overline{B}_b} \| \pi_{\overline{B}_{b-1}}) + D_b + C_b D_{b-1} \\
& .
\end{aligned}$$

From here, we note that $\mathrm{D}_{\mathrm{KL}}(\mu_{b-1,\overline{B}_b} \| \pi_{\overline{B}_b})$ satisfies the theorem conditions, since all blocks in $\overline{B}_b$ have been sampled. We can therefore apply the inductive hypothesis and obtain

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(\mu_b \| \pi) \leq & C_b C_{b-1} \, \mathrm{D}_{\mathrm{KL}}(\mu_{b-2} \| \pi) + C_{\max}(1 - C_b) \, \mathrm{D}_{\mathrm{KL}}(\mu_{0,\overline{B}_b} \| \pi_{\overline{B}_b}) + (1 - C_b) \sum_{i=1}^{b-1} D_i \\
& + C_b(1 - C_{b-1}) \, \mathrm{D}_{\mathrm{KL}}(\mu_{b-2,\overline{B}_{b-1}} \| \pi_{\overline{B}_{b-1}}) + D_b + C_b D_{b-1}.
\end{aligned}$$

Using Lemma 7, we can upper bound

$$\mathrm{D}_{\mathrm{KL}}(\mu_{b-2,\overline{B}_{b-1}} \| \pi_{\overline{B}_{b-1}}) \leq \mathrm{D}_{\mathrm{KL}}(\mu_{b-2,\overline{B}_{b-1}|B_{b-1}} \| \pi_{\overline{B}_{b-1}|B_{b-1}})$$

and then apply the chain lemma

$$\mathrm{D}_{\mathrm{KL}}(\mu_{b-2,\overline{B}_{b-1}|B_{b-1}} \| \pi_{\overline{B}_{b-1}|B_{b-1}}) = \mathrm{D}_{\mathrm{KL}}(\mu_{b-2} \| \pi) - \mathrm{D}_{\mathrm{KL}}(\mu_{b-2,B_{b-1}} \| \pi_{B_{b-1}})$$

to obtain

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(\mu_b \| \pi) \leq & C_b C_{b-1} \, \mathrm{D}_{\mathrm{KL}}(\mu_{b-2} \| \pi) \\
& + C_b(1 - C_{b-1}) \, \mathrm{D}_{\mathrm{KL}}(\mu_{b-2} \| \pi) - C_b(1 - C_{b-1}) \, \mathrm{D}_{\mathrm{KL}}(\mu_{b-2,B_{b-1}} \| \pi_{B_{b-1}}) \\
& + (1 - C_b) \sum_{i=1}^{b-1} D_i + C_{\max}(1 - C_b) \, \mathrm{D}_{\mathrm{KL}}(\mu_{0,\overline{B}_b} \| \pi_{\overline{B}_b}) + D_b + C_b D_{b-1}.
\end{aligned}$$

We can define $\overline{B}_{b,b-1} \triangleq \overline{B}_b \cap \overline{B}_{b-1}$ (all variable blocks except the last two) and apply the chain lemma

$$\mathrm{D}_{\mathrm{KL}}(\mu_{0,\overline{B}_b} \| \pi_{\overline{B}_b}) = \mathrm{D}_{\mathrm{KL}}(\mu_{0,\overline{B}_{b,b-1|B_{b-1}}} \| \pi_{\overline{B}_{b,b-1|B_{b-1}}}) + \mathrm{D}_{\mathrm{KL}}(\mu_{0,B_{b-1}} \| \pi_{B_{b-1}}).$$

to obtain the bound

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(\mu_b \| \pi) \leq & C_b C_{b-1} \, \mathrm{D}_{\mathrm{KL}}(\mu_{b-2} \| \pi) \\
& + C_b(1 - C_{b-1}) \, \mathrm{D}_{\mathrm{KL}}(\mu_{b-2} \| \pi) - C_b(1 - C_{b-1}) \, \mathrm{D}_{\mathrm{KL}}(\mu_{b-2,B_{b-1}} \| \pi_{B_{b-1}}) \\
& + (1 - C_b) \sum_{i=1}^{b-1} D_i \\
& + C_{\max}(1 - C_b) \, \mathrm{D}_{\mathrm{KL}}(\mu_{0,\overline{B}_{b,b-1|B_{b-1}}} \| \pi_{\overline{B}_{b,b-1|B_{b-1}}}) + C_{\max}(1 - C_b) \, \mathrm{D}_{\mathrm{KL}}(\mu_{0,B_{b-1}} \| \pi_{B_{b-1}}) \\
& + D_b + C_b D_{b-1}.
\end{aligned}$$

We can regroup the terms and cancel $C_B D_{b-1} - C_B D_{b-1} = 0$ yields

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(\mu_b \| \pi) \leq & C_b \, \mathrm{D}_{\mathrm{KL}}(\mu_{b-2} \| \pi) \\
& - C_b(C_{\max} \, \mathrm{D}_{\mathrm{KL}}(\mu_{0,\overline{B}_{b,b-1|B_{b-1}}} \| \pi_{\overline{B}_{b,b-1|B_{b-1}}}) + \sum_{i=1}^{b-2} D_i) \\
& - C_b(1 - C_{b-1}) \, \mathrm{D}_{\mathrm{KL}}(\mu_{b-2,B_{b-1}} \| \pi_{B_{b-1}}) \\
& + C_{\max} \, \mathrm{D}_{\mathrm{KL}}(\mu_{0,\overline{B}_{b,b-1|B_{b-1}}} \| \pi_{\overline{B}_{b,b-1|B_{b-1}}}) + C_{\max}(1 - C_b) \, \mathrm{D}_{\mathrm{KL}}(\mu_{0,B_{b-1}} \| \pi_{B_{b-1}}) \\
& + D_b + \sum_{i=1}^{b-1} D_i.
\end{aligned}$$

By applying the inductive hypothesis in reverse, we can show

$$-C_b(C_{\max} \operatorname{D_{KL}}(\mu_{0,\overline{B}_{b-1}|B_{b-1}} \| \pi_{\overline{B}_{b,b-1}|B_{b-1}}) + \sum_{i=1}^{b,b-1} D_i) \leq -C_b \operatorname{D_{KL}}(\mu_{b-1,\overline{B}_{b,b-1}|B_{b-1}} \| \pi_{\overline{B}_{b,b-1}|B_{b-1}}).$$

Substituting this into the second line, we have

$$\begin{aligned}
\operatorname{D_{KL}}(\mu_b \| \pi) \leq & C_b \operatorname{D_{KL}}(\mu_{b-2} \| \pi) \\
& - C_b \operatorname{D_{KL}}(\mu_{b-2,\overline{B}_{b,b-1|B_{b-1}}} \| \pi_{\overline{B}_{b,b-1}|B_{b-1}}) - C_b(1 - C_{b-1}) \operatorname{D_{KL}}(\mu_{b-2,B_{b-1}} \| \pi_{B_{b-1}}) \\
& + C_{\max} \operatorname{D_{KL}}(\mu_{0,\overline{B}_{b,b-1|B_{b-1}}} \| \pi_{\overline{B}_{b,b-1}|B_{b-1}}) + C_{\max}(1 - C_b) \operatorname{D_{KL}}(\mu_{0,B_{b-1}} \| \pi_{B_{b-1}}) \\
& + \sum_{i=1}^{b} D_i .
\end{aligned}$$

We can once again expand the terms

$$\begin{aligned}
\operatorname{D_{KL}}(\mu_{b-2,\overline{B}_{b,b-1|B_{b-1}}} \| \pi_{\overline{B}_{b,b-1}|B_{b-1}}) \\
C_{\max} \operatorname{D_{KL}}(\mu_{0,\overline{B}_{b,b-1|B_{b-1}}} \| \pi_{\overline{B}_{b,b-1}|B_{b-1}}).
\end{aligned}$$

Using the chain lemma and canceling the single block terms gives

$$\begin{aligned}
\operatorname{D_{KL}}(\mu_b \| \pi) \leq & C_b \operatorname{D_{KL}}(\mu_{b-2} \| \pi) \\
& - C_b \operatorname{D_{KL}}(\mu_{b-2,\overline{B}_b} \| \pi_{\overline{B}_b}) + C_b C_{b-1} \operatorname{D_{KL}}(\mu_{b-2,B_{b-1}} \| \pi_{B_{b-1}}) \\
& + C_{\max} \operatorname{D_{KL}}(\mu_{0,\overline{B}_b} \| \pi_{\overline{B}_b}) - C_b C_{\max} \operatorname{D_{KL}}(\mu_{0,B_{b-1}} \| \pi_{B_{b-1}}) \\
& + \sum_{i=1}^{b} D_i .
\end{aligned}$$

Since $C_{\max} \geq C_{b-1}$ by definition, we can disregard $C_b C_{b-1} \operatorname{D_{KL}}(\mu_{b-2,B_{b-1}} \| \pi_{B_{b-1}}) - C_b C_{\max} \operatorname{D_{KL}}(\mu_{b-2,B_{b-1}} \| \pi_{B_{b-1}}) \leq 0$.

We now add zero to the right hand side via

$$0 = C_b KL(\mu_{b-2,B_b|\overline{B}_b}) - C_b KL(\mu_{b-2,B_b|\overline{B}_b}).$$

We then have the three terms

$$\begin{aligned}
C_b \operatorname{D_{KL}}(\mu_{b-2} \| \pi) + \sum_{i=1}^{b} D_i, \\
-C_b \operatorname{D_{KL}}(\mu_{b-2,\overline{B}_b} \| \pi_{\overline{B}_b}) - C_b KL(\mu_{b-2,B_b|\overline{B}_b}), \qquad (11) \\
C_{\max} \operatorname{D_{KL}}(\mu_{0,\overline{B}_b} \| \pi_{\overline{B}_b}) + C_b KL(\mu_{b-2,B_b|\overline{B}_b}). \qquad (12)
\end{aligned}$$

Note that the previous time steps left $\mu_{b-2,B_b|\overline{B}_b}$ invariant, hence $KL(\mu_{b-2,B_b|\overline{B}_b}) = KL(\mu_{0,B_b|\overline{B}_b})$. Then by applying the chain lemma to (12) and (11), we obtain

$$\begin{aligned}
\operatorname{D_{KL}}(\mu_b \| \pi) \leq & C_b \operatorname{D_{KL}}(\mu_{b-2} \| \pi) - C_b \operatorname{D_{KL}}(\mu_{b-2} \| \pi) + C_{\max} \operatorname{D_{KL}}(\mu_0 \| \pi) + \sum_{i=1}^{b} D_i \\
= & C_{\max} \operatorname{D_{KL}}(\mu_0 \| \pi) + \sum_{i=1}^{b} D_i
\end{aligned}$$

which completes the proof. □

Algorithm 5 Cyclic Block Langevin Monte Carlo (CBLMC)

```
procedure CBLD($x_0 \in \mathrm{dom}(f)$, Block Permutation $\sigma = \{B_1, ..., B_b\}$, Step Sizes $\lambda \in \mathbb{R}_+^b$)
    for $k \geq 0$ do
        Set $x_0^{k+1} = x^k$
        for $n = 1$ to $b$ do
            Choose $B_n = \sigma_n$, sample $\xi \sim \mathcal{N}(0, I^d)$
                $x_{n+1} = x_n - \lambda_n U_n \nabla f(x_k) + U_n \sqrt{2\beta^{-1}\lambda_n}\xi$
        end for
        Set $x^{k+1} = x_b^{k+1}$
    end for
end procedure
```

As discussed in the main text, Lemma 1 can be used to trivially bound both the continuous ($C_i = e^{-2\gamma\lambda_i\beta^{-1}}$, $D_i = 0$) and discrete ($C_i = e^{-\gamma\lambda_i\beta^{-1}}$, $D_i = 3d_i L_i^2 \lambda_i^2$) cases. Discrete-time CBLMC is shown in Algorithm 5

For CBLMC, we additionally assume that the potential is L-smooth (Assumption 4). From Beck & Tetruashvili (2013), this implies each block has a separate smoothness constant $L_i \leq L$. From applying Vempala & Wibisono (2019) with the modification using Lemma 4 proposed in Chewi (2024), each block step has the descent

$$\mathrm{D_{KL}}(\mu^{kb}\|\pi) \leq e^{-\gamma\lambda_i\beta^{-1}}\,\mathrm{D_{KL}}(\mu^{b-1}\|\pi) + (1 - e^{-\gamma\lambda_i\beta^{-1}})\,\mathrm{D_{KL}}(\mu^{b-1}_{\overline{B_i}}\|\pi_{\overline{B_i}}) + 3L_i^2 d_i \lambda_i^2.$$

When iterated for kb cycles, we obtain the bound

$$\mathrm{D_{KL}}(\mu^{kb}\|\pi) \leq e^{-\gamma kb\lambda_{\min}\beta^{-1}kb}\,\mathrm{D_{KL}}(\mu^{0}\|\pi) + \frac{4}{\gamma\lambda_{\min}}\sum_{i=1} L_i^2 d_i \lambda_i^2.$$

D PROOF OF THEOREM 3

We begin by recalling the following Lemmas from literature:

Lemma 9 (Uniform L^2 bound on Langevin Diffusion (Lemma 3 of Raginsky et al. (2017)))**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function satisfying Assumption 5. For a random variable $x(t) = x(0) - \int_0^t \nabla f(x(s))ds + \int_0^t dW_s$, we have the bound*

$$\mathbb{E}[\|x(t)\|^2] \leq \mathbb{E}[\|x(0)\|^2]e^{-mt} + \frac{d/\beta + c}{m}(1 - e^{-2mt}).$$

Lemma 10 (Wasserstein bound from Relative Entropy (Corollary 2.3 of Bolley & Villani (2005)))**.** *Let μ, ν be two probability measures on some measurable space X equipped with measurable distance $\mathscr{D}$, and let $\phi : X \to \mathbb{R}^+$ be a non-negative measurable function. Assume that $\exists x_0 \in X$, $\alpha > 0$ such that $\int e^{\alpha\mathscr{D}(x_0,x)^p} d\nu(x)$ is finite. Then*

$$W_2 \leq C\left[\mathrm{D_{KL}}(\mu\|\nu)^{1/2} + \left(\frac{\mathrm{D_{KL}}(\mu\|\nu)}{2}\right)^{1/4}\right]$$

where

$$C \triangleq 2\inf_{x_0\in X}\left(\frac{1}{\alpha}(\frac{3}{2} + \log\int e^{\alpha\mathscr{D}(x_0,x)^p} d\nu(x))\right)^{1/p}.$$

In addition, we adapt the following Lemma from Raginsky et al. (2017)

Lemma 11 (Exponential L^2 Integrability of Block Langevin Diffusion)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function satisfying Assumption 5, and let $x^k(t) = x(0) - \int_0^t U_k\nabla f(x(s))ds + \int_0^t U_k dW_s$*

be a random variable in $\mathbb{R}^d$ across some number of iterations k, where $\sum_{i=1}^{b} U_i = I_d$. Suppose the initial state x_0 is drawn from some μ_0 satisfying Assumption 6 and $\beta > 2/m$. Then on iteration k

$$\log E\left[e^{\|x_\lambda^k\|^2}\right] \leq \kappa_0 + 2(c + \frac{d_{\max}}{\beta})k\lambda.$$

Proof. Define $G(x_t^k) \triangleq e^{\|x_t^k\|^2}$. By Itô's lemma, on iteration k of BLD we have

$$\begin{aligned} dG(x_t^k) =& -2\left\langle x_t^k, U_k \nabla f(x_t^k)\right\rangle e^{\|x_t^k\|^2} dt + \frac{2\beta^{-1}}{2} \mathrm{Tr}\left[U_k^2 (2e^{\|x_t^k\|^2} I + 4x^T x e^{\|x_t^k\|^2})\right] dt \\ &+ \sqrt{2\beta}\left\langle x_t^k, U_k\right\rangle e^{\|x_t^k\|^2} dW_t \\ =& -2\left\langle x_t^k, U_k \nabla f(x_t^k)\right\rangle G(t)dt \\ &+ 2d_k\beta^{-1}G(x_t^k)dt + 4\|U_k x_t^k\|^2\beta^{-1}G(x_t^k)dt + \sqrt{2\beta}\left\langle x_t^k, U_k\right\rangle G(x_t^k)dW_t. \end{aligned}$$

Integrating and summing across k steps, we obtain

$$\begin{aligned} G(x_\lambda^k) =& G(x^0) + \sum_{i=1}^{k}\left[2\int_0^\lambda \left[-\left\langle x_t^k, U_k\nabla f(x_t^k)\right\rangle + 2\beta^{-1}\|U_k x_t^k\|^2)\right] G(x_t^k)dt \right. \\ &\left. + \int_0^\lambda 2d_k\beta^{-1}G(x_t^k)dt + \int_0^\lambda \sqrt{2\beta}\left\langle x_t^k, U_k\right\rangle G(x_t^k)dW_t\right]. \end{aligned}$$

Applying the dissapativity condition and assuming $\beta > 2/m$, we can bound the first integrand as

$$-\left\langle x_t^k, U_k\nabla f(x_t^k)\right\rangle + 2\beta^{-1}\|U_k x_t^k\|^2 \leq (2\beta^{-1} - m)\sum_{j\in B_i}(x_{t,j}^k)^2 + c \leq c$$

which results in

$$G(x_\lambda^k) = G(x^0) + \sum_{i=1}^{k} 2(c + d_k\beta^{-1})\int_0^\lambda G(x_t^k)dt + \int_0^\lambda \sqrt{2\beta}G(x_t^k)\left\langle x_t^k, U_k dW_t\right\rangle.$$

As stated in Raginsky et al. (2017), each Itô integral $\int_0^\lambda \sqrt{2\beta}G(x_t^k)\left\langle x_t^k, U_k dW_t\right\rangle$ is a zero-mean Martingale. Taking expectations over both sides and applying Assumption 6 yields

$$\begin{aligned} \mathbb{E}[G(x_\lambda^k)] &= \mathbb{E}[G(x^0)] + \sum_{i=1}^{k} 2(c + d_k\beta^{-1})\int_0^\lambda \mathbb{E}[G(x_t^k)]dt \\ &\leq e^{\kappa_0} + 2(c + d_{\max}\beta^{-1})\int_0^{k\lambda} \mathbb{E}[G(x_t^k)]dt. \end{aligned}$$

where the integrability of $\mathbb{E}[G(x_t^k)]$ across block steps follows from the continuity of x_t^k across each block step k. By Grönwell's Lemma, we then have the result. ☐

Theorem 3 follows as a consequence of Lemma 2 by applying the Otto-Villani theorem coupled with the triangle inequality for W_2 as stated in the main text.

D.1 PROOF OF LEMMA 2

Proof. Let μ_t^k and ν_t^k be the laws of SGBLD and BLD at times t and iteration k respectively with iterates $x^k(s), y^k(s)$. We assume that each process selects the same variable blocks at each iteration, i.e. $B_x^k = B_y^k$.

Using the Girsanov formula, we can express the Radon-Nikodym derivative $\frac{d\nu_t^k}{d\mu_t^k}$ as

$$\begin{aligned}\frac{d\nu_t^k}{d\mu_t^k} =& \exp\Bigg[\frac{\beta}{4}\int_0^t \left\langle U_k\nabla f(y^k(s)) - U_k g_z(y^k(t)), -U_k\nabla f(y^k(s))ds + U_k dW_s\right\rangle \\ &+ \frac{\beta}{4}\int_0^t \left\langle U_k\nabla f(y^k(s)) - U_k g_z(y^k(t)), U_k\nabla f(y^k(s)) + U_k g_z(y^k(t))\right\rangle\Bigg] \\ =& \exp\left[\frac{\beta}{4}\int_0^t \left\langle U_k\nabla f(y^k(s)) - U_k g_z(y^k(s)), dW_s\right\rangle - \frac{\beta}{4}\int_0^t \|U_k\nabla f(y^k(s)) - U_k g_z(y^k(s))\|^2 ds\right].\end{aligned}$$

Setting $t = \lambda_k$, we can express $\mathrm{D}_{\mathrm{KL}}(\mu_t^k \| \nu_t^k)$ as

$$\mathrm{D}_{\mathrm{KL}}(\mu_t^k \| \nu^k) = -\int d\mu_t^k \log \frac{d\nu_t^k}{d\mu_t^k} = \sum_{i=1}^k \mathbb{E}\left[\frac{\beta}{4}\int_0^\lambda \|U_k\nabla f(y^k(s)) - U_k g_z(y^k(s))\|^2 ds\right].$$

Using Assumption 3, we obtain

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(\mu_t^k \| \nu_t^k) &= \sum_{i=1}^k \mathbb{E}\left[\frac{\beta}{4}\int_0^\lambda \|U_k\nabla f(y^i(s)) - U_k g_z(y^i(s))\|^2 ds\right] \\ &\leq \sum_{i=1}^k \left[\frac{\beta}{4}\int_0^\lambda M^2\mathbb{E}\|y^i(s)\|^2 + B^2 ds\right] \\ &\leq \sum_{i=1}^k \left[\frac{\beta}{4}\int_0^\lambda M^2(e^{-ms}\mathbb{E}\|y^i(0)\|^2 + \frac{d_i + c}{m}(1 - e^{-ms})) + B^2 ds\right].\end{aligned}$$

where we have applied Lemma 9 in the last line. Integrating, we obtain

$$\mathrm{D}_{\mathrm{KL}}(\mu_t^k \| \nu_t^k) \leq \sum_{i=1}^k \frac{\beta}{4}\mathbb{E}\left[\frac{M^2}{m}(1 - e^{-m\lambda})\mathbb{E}\|y^i(0)\|^2 + \frac{M^2(d_i/\beta + c)}{m^2}(mt + e^{-m\lambda} - 1)) + B^2\lambda\right].$$

Expanding $e^{-m\lambda}$ and leveraging that $m\lambda \geq 1 - e^{-m\lambda} \geq m\lambda - \frac{m^2\lambda^2}{2}$

$$\mathrm{D}_{\mathrm{KL}}(\mu_\lambda^k \| \nu_\lambda^k) \leq \sum_{i=1}^k \frac{\beta M^2 \lambda}{4}\mathbb{E}\|y^i(0)\|^2 + \frac{M^2\lambda^2(d_i + c\beta)}{4} + \frac{\beta B^2 t}{4}.$$

By repeatedly expanding Lemma 9, we obtain

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(\mu_t^k \| \nu_t^k) &\leq \sum_{i=0}^{k-1} \frac{M^2\beta\lambda}{4}\kappa_0 + e^{-m(i-1)\lambda}\frac{M^2\lambda^2(d_i + \beta c)}{8} + \frac{M^2\lambda^2(d_i + \beta c)}{8} + \frac{\beta B^2 \lambda k}{4} \\ &\leq \frac{M^2\beta\lambda k}{4}\kappa_0 + \frac{M^2\lambda^2(d_{\max} + \beta c)k}{4} + \frac{\beta B^2\lambda k}{4} \\ &\triangleq (C_1 + C_2\lambda)\lambda k.\end{aligned}$$

where we have defined for convenience

$$C_1 \triangleq \frac{M^2\beta\kappa_0}{4} + \frac{\beta B^2}{4}$$

and

$$C_2 \triangleq \frac{M^2(d_{\max} + \beta c)}{4}$$

By Lemma 10, we can bound $W_2^2(\mu_t^k, \nu_t^k)$ as

$$W_2^2(\mu_t^k, \nu_t^k) \leq 4C^2\left[\mathrm{D}_{\mathrm{KL}}(\mu\|\nu)^{1/2} + \left(\frac{\mathrm{D}_{\mathrm{KL}}(\mu\|\nu)}{2}\right)^{1/4}\right]^2.$$

Setting $\alpha = 1$, $d(x) = \|x\|^{1/2}$, and $p = 1/2$, we obtain from Lemma 11

$$4C^2 \leq (12 + 4\kappa_0 + 8(2c + \frac{d_{\max}}{\beta})k\lambda).$$

Note that for any $a \geq 0$, we have $(\sqrt{a} + (\frac{a}{2})^{1/4})^2 \leq 2a + 2\sqrt{a}$, since

$$(\sqrt{a} + (\frac{a}{2})^{1/4})^2 = a + 2^{3/4}a^{3/4} + \frac{a^{1/2}}{2^{1/2}} = a + (2^{1/4}a^{1/4})(2^{1/2}a^{1/2}) + \frac{a^{1/2}}{2^{1/2}} + \frac{a^{1/2}}{2^{1/2}}.$$

By Young's inequality, $(2^{1/4}a^{1/4})(2^{1/2}a^{1/2}) \leq \frac{\sqrt{a}}{2^{3/4}} + \frac{a}{2^{1/2}}$, hence

$$(\sqrt{a} + (\frac{a}{2})^{1/4})^2 = a + (2^{1/4}a^{1/4})(2^{1/2}a^{1/2}) + \frac{a^{1/2}}{2^{1/2}} \leq a(1 + \frac{1}{\sqrt{2}}) + \frac{2\sqrt{a}}{\sqrt{2}} \leq 2a + 2\sqrt{a}.$$

plugging in Lemma 11, and assuming $k\lambda \geq 1$, $k > \lambda$ we have

$$W_2^2(\mu_t^k, \nu_t^k) \leq 2C^2 \left[\mathrm{D}_{\mathrm{KL}}(\mu\|\nu) + \sqrt{\mathrm{D}_{\mathrm{KL}}(\mu\|\nu)}\right]$$

$$\begin{aligned}
&\leq (12 + 8(\kappa_0 + (2c + d_{\max}/\beta))) \left[(C_1 + C_2\lambda)k\lambda + \sqrt{(C_1 + C_2\lambda)k\lambda}\right](k\lambda) \\
&\leq (12 + 8(\kappa_0 + (2c + d_{\max}/\beta))) \left[(C_2 + \sqrt{C_2})\sqrt{\lambda k} + (C_1 + \sqrt{C_1})\right](k\lambda)^2 \\
&= C_0^2 \left[(C_2 + \sqrt{C_2})\sqrt{\lambda k} + (C_1 + \sqrt{C_1})\right](k\lambda)^2.
\end{aligned}$$

We thereby obtaining Lemma 2 with:

$$\begin{aligned}
C_0 &\triangleq (12 + 8(\kappa_0 + (2c + d_{\max}/\beta))), \\
C_1 &\triangleq \frac{M^2\beta\kappa_0}{4} + \frac{\beta B^2}{4}, \\
C_2 &\triangleq \frac{M^2(d_{\max} + \beta c)}{4}.
\end{aligned}$$

□

D.2 CONSTANTS IN EXPECTED FUNCTION GAP BOUNDS

We start by recalling the Lemma from Polyanskiy & Wu (2016):

Lemma 12 (Wasserstein Continuity for Quadratic-Growth Potentials)**.** *Let μ, π be probability distributions with finite second moments and let $f : \mathbb{R}^d \to \mathbb{R}^+$ be a continuously differentiable function satisfying $\|\nabla f(x)\|^2 \leq c_1\|x\|^2 + c_2$. Then we have*

$$\left|\int f(x)d\mu(x) - \int f(x)d\pi(x)\right| \leq (c_1\sigma + c_2)W_2(\mu, \pi)$$

where $\sigma = \sqrt{\max[\mathbb{E}_\mu[\|x\|^2],\ \mathbb{E}_\pi[\|x\|^2]]}$.

Raginsky et al. (2017) bound the constant $\sigma^2 = \max \mathbb{E}_{\mu^k}[x^2], \mathbb{E}_{\pi_\beta}[x^2]$ using an unbiased oracle. As discussed in the main text, DXs have fixed device variation from analog errors, precluding unbiased estimation. However, DX errors take the form of perturbations in the underlying function, i.e. the target function characteristics are intact. For instance, DXs with quadratic potential targets (Aifer et al., 2023; Song et al., 2024) are still optmizing/sampling quadratic functions. Accordingly, Assumptions 4 and 5, that the DX gradient retains both Lipschitz continuity and dissipativity, are reasonable. Assuming g_δ is $(\mathfrak{m}, \mathfrak{c})$-dissipative, we have from Lemma 3 of Raginsky et al. (2017):

$$\|x(t)\|^2 \leq \kappa_0 + \frac{\mathfrak{c} + d/\beta}{\mathfrak{m}}.$$

Then

$$\sigma^2 = \kappa_0 + \max\left[\frac{c + d/\beta}{m}, \frac{\mathfrak{c} + d/\beta}{\mathfrak{m}}\right].$$