Journal Name

Crossmark

RECEIVED
dd Month yyyy

REVISED
dd Month yyyy

PAPER

A localized consensus-based sampling algorithm

Arne Bouillon[1,*], Alexander Bodard[2], Panagiotis Patrinos[2], Dirk Nuyens[1] and Giovanni Samaey[1]

[1]KU Leuven, Department of Computer Science (NUMA), Celestijnenlaan 200A, 3001 Leuven, Belgium
[2]KU Leuven, Department of Electrical Engineering (ESAT-STADIUS), Kasteelpark Arenberg 10, 3001 Leuven, Belgium
[*]Author to whom any correspondence should be addressed

E-mail: arne.bouillon@kuleuven.be

Keywords: Consensus-based sampling, Langevin diffusion, Bayesian inverse problems

Abstract

We propose a localized consensus-based method for sampling from non-Gaussian distributions, a task that frequently arises when solving Bayesian inverse problems. Our method arises from an alternative derivation of consensus-based sampling (CBS). Starting from ensemble-preconditioned Langevin dynamics, we replace the potential by its Moreau envelope—a smoother approximation—in order to replace the gradient in the Langevin equation with a proximal operator. We then approximate this operator by a weighted mean. In the limit of infinitely smoothing the potential to a quadratic function, this procedure recovers the standard CBS dynamics. In addition, outside this limit, we retrieve a refined variant of polarized CBS. We call the resulting algorithm *localized consensus-based sampling*, since particles interact more with nearby particles than with faraway ones. Our method is affine-invariant, exact for Gaussian targets in the mean-field limit, and demonstrates improved robustness over polarized CBS in numerical experiments. Like other consensus-based methods, localized CBS is gradient-free and easily parallelizable.

1 Introduction

We consider the problem of sampling a d-dimensional probability distribution with density

$$\pi(u) = \frac{\exp(-V(u))}{\int \exp(-V(v))\,\mathrm{d}v} =: \frac{\widehat{\pi}(u)}{Z}, \tag{1}$$

where $V\colon \mathbb{R}^d \to \mathbb{R}$ is a potential function with $\int \exp(-V(v))\mathrm{d}v < \infty$. Here, we implicitly take integrals over $\mathbb{R}^d$. In many applications V is known but the integral Z may be very expensive to compute, such that π is available only up to a multiplicative constant.

1.1 Bayesian inverse problems

Equation (1) commonly arises in Bayesian inversion (see, e.g., [33, 59]). Consider a model G with unknown parameters u and perturbed observation

$$y = G(u) + \eta, \tag{2}$$

where η is some unknown measurement noise. The goal is characterizing likely values of u, based on y and G. Denoting by $\pi_{\text{prior}}(u)$ and $\pi_{\text{noise}}(\eta)$ our prior beliefs about u and η, Bayes' rule describes our belief about u after the observation by the *posterior* distribution with density

$$\pi_{\text{post}}(u \mid y) = \frac{\pi_{\text{noise}}(y - G(u))\,\pi_{\text{prior}}(u)}{\int \pi_{\text{noise}}(y - G(v))\,\pi_{\text{prior}}(v)\,\mathrm{d}v} =: \frac{\widehat{\pi}_{\text{post}}(u \mid y)}{\pi_{\text{ev}}(y)}. \tag{3}$$

One approach to solve the Bayesian inverse problem (2) is to find the *maximum a posteriori* (MAP) parameter u_{MAP} that maximizes (3) as discussed in, e.g., [10]:

$$u_{\text{MAP}} := \arg\max_u \widehat{\pi}_{\text{post}}(u \mid y). \tag{4}$$

Optimization algorithms can solve the inverse problem in this fashion. However, u_{MAP} gives no information about the uncertainty on u captured by π_{post}. When uncertainty quantification is

required, other characteristics of the posterior must be inferred. In that case, one commonly wants to integrate some function with respect to π_{post}, e.g., to compute statistical moments of the distribution. Numerical integration techniques, such as [28], which uses quasi-Monte Carlo, may provide a sufficient approximation even in high dimensions. Alternatively, it is common to use algorithms that produce a sequence of Monte Carlo samples distributed approximately according to π_{post}. Then solving (2) corresponds to drawing samples from (1), with the *evidence* $\pi_{\text{ev}}(y)$ as the intractable Z and with $\widehat{\pi}_{\text{post}}(\cdot \mid y)$ as the unnormalized density $\widehat{\pi}$.

1.2 Related literature

Sampling from distributions with density (1) with unknown Z is historically done with Markov chain Monte Carlo (MCMC) algorithms, which originate in [27, 41] and are reviewed in detail in [2]. MCMC remains a very active research topic with recent developments such as the popular pCN [8] and DILI [9] proposal moves that are effective in high dimensions and multilevel MCMC versions [16, 39] to exploit cheap approximations to G.

An alternative method is sequential Monte Carlo (SMC) [11], which evolves samples from an easy-to-sample distribution (e.g., π_{prior} in (3)) into samples from π gradually over multiple steps. Many other methods are based on evolving SDEs that have π (or an approximation to π) as an invariant distribution. An example is Langevin diffusion (see subsection 2.1); after time discretization, the unadjusted Langevin algorithm (ULA) [53] arises as an approximate sampler of π. Using ULA steps as MCMC proposals—or, equivalently, adding Metropolis–Hastings rejection to ULA—results in the Metropolis-adjusted Langevin algorithm (MALA) [1].

These Langevin methods use gradients of the potential V, which can be undesirable [17, 35]. In the case of non-differentiable potentials, gradients can be approximated using the proximal operator [18, 37, 48, 61] or similar techniques [36] to form new, gradient-free algorithms. However, unless the potential has certain structure, this introduces the need to solve an embedded optimization problem. Various methods adapt Langevin dynamics to be gradient-free by using an *ensemble* of multiple *particles* (instances of the SDE) that exchange information, which can be harnessed to approximate gradients. These are often formulated in the infinite-particle limit, as McKean–Vlasov SDEs. Examples include ensemble Kalman sampling (EKS) [22] and the closely related ALDI method [21], derived from interacting Langevin diffusions and inspired by the ensemble Kalman inversion (EKI) optimizer [31, 35, 56]. EKI is related to iterating the ensemble Kalman filter (EnKF) [19]; an EKI-based sampler is analyzed in [14]. Other ensemble variants of Langevin dynamics include [13] and the multiscale sampler in [47]. In addition, the consensus-based sampling (CBS) method [4], described in subsection 2.2, was developed independently of Langevin dynamics. It was inspired by consensus-based optimization (CBO) [50].

Interaction in particle methods can be *localized*: instead of computing a global interaction term, each particle interacts more with particles that are nearby than with those that are far away. The aim of localization is to avoid the linearity assumption on G that is implicit in these methods. Localized algorithms show improved performance for nonlinear G, since the adapted dynamics rely on *local* instead of global linearity. The paper [51] proposed to use localized interaction in the EKS/ALDI algorithm—details are given in subsection 5.3. Similar ideas were proposed for EKI and the EnKF [63], for an ensemble Kalman method for rare-event sampling [64], and for CBS as a variant named *polarized CBS* [3]—see subsection 5.2. Our paper develops a different CBS variant with localized interaction, named *localized CBS*.

Remark 1 (Another concept of localization)**.** *We stress that localization as discussed here is unrelated to the notion of localization in [29, 46] for the EnKF, in [24, 62] for EKI, and in [43] for MCMC. There, the empirical covariance matrix is corrected for spurious correlations, caused by a finite ensemble size, between parameter* indices *that should be uncorrelated—e.g., because they correspond to elements at a large distance* in the underlying problem*. This technique can also increase the rank of the covariance and prevent EKI's* subspace property *[62]. In contrast, we adopt the term* localized *to denote localized interaction—where particles interact more with nearby particles than with faraway ones—following related literature such as [51, 64].*

1.3 Objectives and contributions

There are several aspects to consider when assessing sampling algorithms, both computationally and in terms of accuracy.

(i) *The need for gradients.* Methods that require gradients of V cannot be used when gradients are unavailable.

(ii) *Parallelizability.* Standard MCMC methods are inherently sequential due to the need to overcome a burn-in phase and are hard to parallelize effectively [58, 65]. SMC and McKean–Vlasov-based methods use an ensemble of particles in parameter space that exchange information through interaction. This means that any work that is local to a single particle, such as evaluating G, can be trivially parallelized over all particles.

(iii) *Non-Gaussian performance.* Many sampling methods use approximations that hold only for Gaussian distributions, leading to poor sampling performance when the target distribution is non-Gaussian or multimodal. Non-Gaussian distributions frequently arise in practice, for instance when solving Bayesian inverse problems (2) with a nonlinear forward model G.

(iv) *Affine-invariance.* Some algorithms produce poor samples when the different dimensions of π are poorly scaled. The property of affine-invariance (detailed in Definition 2 in the following section) ensures that affine transformations of the parameters do not influence sampling performance.

This paper focuses on gradient-free, parallelizable sampling from poorly scaled, non-Gaussian distributions. We make the following contributions:

(i) Section 3 proposes a novel way to connect interacting Langevin diffusions [22], which use gradients, to consensus-based sampling [4], which does not. We rely on two approximations: replacing V by a Moreau envelope and approximating the proximal operator by a weighted mean. CBS arises in the limit when the Moreau envelope reduces to a quadratic.

(ii) Subsections 4.1 and 4.2 look at these approximations outside the CBS limit and tailor them to non-Gaussian problems. This results in an algorithm we name *localized consensus-based sampling*. We perform an analysis of localized CBS with a broad class of covariance-based preconditioners. We derive ordinary differential equations (ODEs) for the evolution of its mean and covariance when applied to Gaussian problems in subsection 4.3. More generally, subsection 4.4 shows that localized CBS is affine-invariant.

(iii) Section 5 compares the existing CBS, polarized CBS, and localized ALDI algorithms to our localized CBS method. We argue that our algorithm is an improved version of polarized CBS [3], with our comparative benefits (such as affine-invariance and non-Gaussian performance) emerging from the connection to interacting Langevin diffusions.

(iv) Section 6 discusses approximations of the mean-field equations with finite particle systems and other computational considerations. Section 7 performs numerical experiments that showcase the method's affine-invariance and non-Gaussian performance.

We start our discussion with a brief review of interacting Langevin diffusions and consensus-based sampling in section 2.

1.4 Notation

We denote by e_i the ith column of an identity matrix and by $\mathbf{1}$ a vector of ones, both of implicit length. In addition, we define the Mahalanobis norm of a vector $v \in \mathbb{R}^n$ weighted by the inverse of a positive definite matrix $A \in \mathbb{R}^{n\times n}$ as

$$\|v\|_A \coloneqq \sqrt{v^T A^{-1}\, v}. \tag{5}$$

2 Interacting-particle samplers

We now recall the interacting-Langevin and consensus-based samplers, and how they are used to sample in accordance with (1).

2.1 Interacting-Langevin sampling

Let W_t be a standard Brownian motion in $\mathbb{R}^d$. The Langevin equation reads

$$\mathrm{d}U_t = \nabla \log \widehat{\pi}(U_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t \tag{6}$$

and, like all SDEs in this paper, should be interpreted in the Itô sense. Preconditioning (6) with a positive definite matrix K leads to a different SDE, the preconditioned Langevin equation

$$\mathrm{d}U_t = K\nabla \log \widehat{\pi}(U_t)\,\mathrm{d}t + \sqrt{2K}\,\mathrm{d}W_t. \tag{7}$$

Denote by ρ_t the law of U_t. Both (6) and (7) have the property that, under certain conditions on $\widehat{\pi}$, they transform arbitrary laws ρ_0 of the initial value U_0 into a law $\rho_\infty \coloneqq \lim_{t\to\infty} \rho_t$ that is exactly the desired distribution π. This leads to a Monte Carlo algorithm for approximately sampling from π: simulate some number J of discretized paths $\{U_t^j\}_{j=1}^J$ of (6) or (7) until some sufficiently large time T and use $\{U_T^j\}_{j=1}^J$ as the approximate samples. If K is chosen well, (7) can converge faster than (6).

In [22], the constant matrix K is replaced by a time-varying matrix $\mathcal{C}(\rho_t)$, where ρ_t is the law of U_t and $\mathcal{C}$ denotes the covariance operator. This change results in

$$\mathrm{d}U_t = \mathcal{C}(\rho_t)\nabla \log \widehat{\pi}(U_t)\,\mathrm{d}t + \sqrt{2\mathcal{C}(\rho_t)}\,\mathrm{d}W_t. \tag{8}$$

The covariance operator $\mathcal{C}$ uses the mean μ in its definition:

$$\mu(\rho) \coloneqq \int v\,\mathrm{d}\rho(v) \qquad \text{and} \qquad \mathcal{C}(\rho) \coloneqq \int (v-\mu(\rho))\otimes(v-\mu(\rho))\,\mathrm{d}\rho(v). \tag{9}$$

Equation (8) is not a special case of (7); rather, it is an extension, and it avoids having to choose a suitable (problem-dependent) K. The corresponding Fokker–Planck equation is

$$\partial_t \rho_t(u) = \nabla\cdot\big(\rho_t(u)\,\mathcal{C}(\rho_t)\,(-\nabla\log\widehat{\pi}(u) + \nabla\log\rho(u))\big). \tag{10}$$

It is shown in [22] that (8) evolves to $\rho_\infty = \pi$ at an exponential rate when V is strongly convex and $\mathcal{C}(\rho_t)$ is bounded from below. In the non-convex case, π is still an invariant distribution. In addition, [21] shows that (8) has the important property of *affine-invariance*.

Definition 2 (Affine-invariance)**.** *Consider an SDE that depends on the unnormalized target distribution $\widehat{\pi}$ (or, equivalently, on the potential V). The SDE is called affine-invariant if it is invariant under affine transformations of the state variables and $\widehat{\pi}$. That is, define*

$$\tilde{\pi}(z) = \widehat{\pi}(Mz+b) \qquad \textit{and} \qquad \tilde{\rho}_0(z) = |M|\,\rho_0(Mz+b) \tag{11}$$

for an arbitrary invertible matrix M and vector b. Denote by ρ_t the probability density at time t of solutions to the SDE for $\widehat{\pi}$ and with initial distribution ρ_0, and by $\tilde{\rho}_t$ the equivalent for $\tilde{\pi}$ and $\tilde{\rho}_0$. Then, the SDE is affine-invariant if

$$\tilde{\rho}_t(z) = |M|\,\rho_t(Mz+b) \qquad \textit{for all } t. \tag{12}$$

Similar definitions are found in [21, 38] for SDEs and, earlier, in [25, 26] for discrete-time samplers. The importance of affine-invariance for samplers is widely acknowledged: it guarantees that the sampler will perform as well on poorly scaled distributions as on isotropic distributions. See [25] for an extensive discussion and examples.

Note that the SDE (8) is of McKean–Vlasov type[1]: its drift and diffusion depend on the law ρ_t through $\mathcal{C}(\rho_t)$. It is also possible to use any positive definite preconditioner $\mathbf{P}(\rho_t; U_t)$ that depends on the state U_t as well as on its law ρ_t. The resulting SDE is

$$\mathrm{d}U_t = \mathbf{P}(\rho_t;U_t)\nabla\log\widehat{\pi}(U_t)\,\mathrm{d}t + \sqrt{2\mathbf{P}(\rho_t;U_t)}\,\mathrm{d}W_t \tag{13}$$

and, similarly to what was noted in [44, 51], has as its Fokker–Planck equation

$$\begin{aligned}\partial_t\rho_t(u) &= \nabla\cdot\left[-\rho_t(u)\,\mathbf{P}(\rho_t;u)\nabla\log\widehat{\pi}(u) + \mathbf{P}(\rho_t;u)\nabla\rho_t(u) + \rho_t(u)\nabla\cdot\mathbf{P}(\rho_t;u)\right]\\ &= \nabla\cdot\left[\rho_t(u)\,\mathbf{P}(\rho_t;u)\,(-\nabla\log\widehat{\pi}(u) + \nabla\log\rho_t(u))\right] + \nabla\cdot\left[\rho_t(u)\nabla\cdot\mathbf{P}(\rho_t;u)\right].\end{aligned}$$

Without the last term, π would again be invariant under this equation. Hence, one can introduce a correction term into the dynamics to cancel it out; this results in

$$\mathrm{d}U_t = \mathbf{P}(\rho_t;U_t)\nabla\log\widehat{\pi}(U_t)\,\mathrm{d}t + \nabla_u\cdot\mathbf{P}(\rho_t;U_t)\,\mathrm{d}t + \sqrt{2\mathbf{P}(\rho_t;U_t)}\,\mathrm{d}W_t. \tag{14}$$

Ideas related to (14) (with an added skew-symmetric matrix in the drift; see also [40]) have been discussed in [38]. Furthermore, (14) underlies the finite-ensemble [44] and localized [51] versions of EKS/ALDI through *statistical linearization* (see [6]).

[1] These McKean–Vlasov equations can be simulated by evolving an ensemble of many time-discretized solution paths in parallel and using their empirical distribution instead of the law ρ_t. This is justified by the principle of *propagation of chaos* [60]; for the ensemble Kalman sampler, this was studied in detail in [15]. Until section 6, we study non-discretized McKean–Vlasov SDEs.

2.2 Consensus-based sampling

Consensus-based sampling [4] follows the dynamics[2]

$$\mathrm{d}U_t = -(U_t - \mu_\alpha(\rho_t))\,\mathrm{d}t + \sqrt{2(\alpha+1)\,\mathcal{C}_\alpha(\rho_t)}\,\mathrm{d}W_t. \tag{15}$$

Here, $\alpha > 0$ is a method parameter, ρ_t is the law of U_t, and W_t is a standard Brownian motion in $\mathbb{R}^d$. The weighted mean and covariance μ_α and $\mathcal{C}_\alpha$ are defined as

$$\mu_\alpha(\rho) \coloneqq \frac{\int v w_\alpha(v)\,\mathrm{d}\rho(v)}{\int w_\alpha(v)\,\mathrm{d}\rho(v)}, \tag{16a}$$

$$\mathcal{C}_\alpha(\rho) \coloneqq \frac{\int (v-\mu_\alpha(\rho)) \otimes (v-\mu_\alpha(\rho))\, w_\alpha(v)\,\mathrm{d}\rho(v)}{\int w_\alpha(v)\,\mathrm{d}\rho(v)}, \tag{16b}$$

where the weight function is given by

$$w_\alpha(v) \coloneqq \widehat{\pi}(v)^\alpha. \tag{17}$$

The parameter α determines the importance of $\widehat{\pi}$ in weighting the particles; when $\alpha = 0$, we get $(\mu_0, \mathcal{C}_0) = (\mu, \mathcal{C})$. The SDE (15) is of McKean–Vlasov type, similarly to (14). In [4] it is shown that if ρ_0 and π are Gaussian, then $\lim_{t\to\infty} \rho_t = \pi$ and the mean and covariance of ρ_t converge exponentially in t. Similarly to interacting Langevin diffusions, CBS is affine-invariant. The mean-field limit of (15) was studied in [23] and [34].

3 Consensus-based sampling as approximate interacting Langevin

We now propose a new connection between interacting-Langevin and consensus-based samplers, arriving at the CBS dynamics (15) starting from (14) by making three specific approximations. Throughout, we assume that the potential function V is differentiable and L-weakly convex (meaning that $V(\cdot) + \frac{L}{2}\|\cdot\|_2^2$ is convex for some $L \in \mathbb{R}_0^+$) and that $\mathcal{C}_\alpha(\rho_t)$ is positive definite for all t.

First step: Moreau envelope of V. For a symmetric positive definite matrix A, denote by

$$V^A(u) \coloneqq \inf_v \big(V(v) + \frac{1}{2}\|v-u\|_A^2\big) \tag{18}$$

the A-Moreau envelope of the potential V. It approximates V from below and coincides at the minimizers, with $V^A(u) \nearrow V(u)$ for all u as $\|A\| \searrow 0$ [54, §1G]. If V is L-weakly convex and $\|A\|_2 < L^{-1}$, then V^A is known to have gradient

$$\nabla V^A(u) = A^{-1}(u - \mathbf{prox}_V^A(u)) \tag{19}$$

(see [7, Example 3.9] and [54, Theorem 2.26]), where the *proximal operator* denotes the singleton

$$\mathbf{prox}_V^A(u) \coloneqq \arg\min_v \big(V(v) + \frac{1}{2}\|v-u\|_A^2\big). \tag{20}$$

Let us now take (14) with $\mathbf{P}(\rho_t; U_t) = \mathcal{C}_\alpha(\rho_t)$ as a starting point. We introduce a scalar parameter $\kappa > 0$ and approximate $V = -\log\widehat{\pi}$ in (14) by its A-Moreau envelope, with $A = \kappa\mathcal{C}_\alpha(\rho_t)$, rescaled by a scalar $\gamma > 0$: $V \approx \gamma V^{\kappa\mathcal{C}_\alpha(\rho_t)}$. We will later choose the parameter γ to counteract the approximation error of the Moreau envelope, with typically $\gamma \approx 1$. As opposed to ∇V, which is unknown, $\nabla(\gamma V^{\kappa\mathcal{C}_\alpha(\rho_t)})$ can be rewritten using (19), expressing our approximation to (14) as

$$\mathrm{d}U_t = -\frac{\gamma}{\kappa}\Big(U_t - \mathbf{prox}_V^{\kappa\mathcal{C}_\alpha(\rho_t)}(U_t)\Big)\,\mathrm{d}t + \sqrt{2\mathcal{C}_\alpha(\rho_t)}\,\mathrm{d}W_t. \tag{21}$$

Instead of a gradient, we now must evaluate the proximal operator.

Second step: Weighted mean as approximate proximal operator. Under mild conditions on a probability measure ρ and function f, including f having a unique minimizer, we have

$$\lim_{\beta\to\infty} \frac{\int v \exp(-\beta f(v))\,\mathrm{d}\rho(v)}{\int \exp(-\beta f(v))\,\mathrm{d}\rho(v)} = \arg\min_v f(v), \tag{22}$$

[2] CBS also has an *optimization mode* where the factor $(\alpha+1)$ is omitted, which results in an optimization method instead of a sampling one. This paper only considers CBS in sampling mode.

as implied by the Laplace principle [12, 42] (similarly to, e.g., [5, 20]). Related is the *quantitative Laplace principle* in [20], which was used in conjunction with the proximal operator to study convergence of the consensus-based optimization (CBO) algorithm in [52]. Equation (22) implies

$$\mathbf{prox}_{V}^{\kappa\mathcal{C}_\alpha(\rho_t)}(U_t) = \lim_{\beta\to\infty} \mu_{\beta,\kappa\mathcal{C}_\alpha(\rho_t)}(\rho_t; U_t), \tag{23}$$

where we used the weighted mean

$$\mu_{\beta,A}(\rho; u) \coloneqq \frac{\int v w_{\beta,A}(v; u)\,\mathrm{d}\rho(v)}{\int w_{\beta,A}(v; u)\,\mathrm{d}\rho(v)}, \qquad w_{\beta,A}(v; u) \coloneqq \exp\Big(-\beta\Big[V(v) + \frac{1}{2}\|v-u\|_A^2\Big]\Big). \tag{24}$$

As the smallest eigenvalue of A tends to ∞, $\mu_{\beta,A}$ approaches μ_β from (16). The choice $\rho = \rho_t$ in (23) was made to enable particle approximations in section 6.

A weighted mean was also used to approximate the proximal operator in [45] in the context of optimization methods. Motivated by (23), we choose some finite $\beta > 0$ and replace the proximal operator by a weighted mean as a second approximation:

$$\mathrm{d}U_t = -\frac{\gamma}{\kappa}\Big(U_t - \mu_{\beta,\kappa\mathcal{C}_\alpha(\rho_t)}(\rho_t; U_t)\Big)\,\mathrm{d}t + \sqrt{2\mathcal{C}_\alpha(\rho_t)}\,\mathrm{d}W_t. \tag{25}$$

Note that neither the proximal operator in (21) nor the weighted mean in (25) may seem inherently easier to compute than the gradient in (14). The eventual reason for these approximations is that weighted means are suitable to efficient gradient-free particle discretizations, such as the ones in [4] and section 6. A direct particle discretization of (14) would clearly require gradients. For now we will study the dynamics in McKean–Vlasov form, but one should keep in mind this motivation.

Recovering CBS in the limit. If we take the limit $\kappa \to \infty$ in (25) and allow γ to depend on κ, then the localized part of the weighted mean vanishes and we get, with $c = \lim_{\kappa\to\infty} \gamma/\kappa$,

$$\mathrm{d}U_t = -c\,\Big(U_t - \mu_\beta(\rho_t)\Big)\,\mathrm{d}t + \sqrt{2\mathcal{C}_\alpha(\rho_t)}\,\mathrm{d}W_t. \tag{26}$$

Choosing $\alpha = \beta$ and $\gamma = (\alpha+1)^{-1}\kappa$ then gives exactly the CBS dynamics (15).

To interpret this link between (25) and CBS, we study $V^{\mathrm{CBS}} \coloneqq \gamma V^{\kappa\mathcal{C}_\alpha(\rho_t)} = (\alpha+1)^{-1}\kappa V^{\kappa\mathcal{C}_\alpha(\rho_t)}$:

$$V^{\mathrm{CBS}}(u) = (\alpha+1)^{-1} \inf_v \big(\kappa V(v) + \frac{1}{2}\|v-u\|^2_{\mathcal{C}_\alpha(\rho_t)}\big). \tag{27}$$

As $\kappa \to \infty$, the infimum can only be reached by choosing $v \in \arg\min V$, so

$$\lim_{\kappa\to\infty} V^{\mathrm{CBS}}(u) - \kappa \min V = (\alpha+1)^{-1}\,\frac{1}{2}\,\mathrm{dist}_{\mathcal{C}_\alpha(\rho_t)}(u, \arg\min V)^2, \tag{28}$$

where $\mathrm{dist}_A(u, S) \coloneqq \inf_{v\in S}\|u-v\|_A$. When V has a unique minimizer u^*, this simplifies to

$$\lim_{\kappa\to\infty} V^{\mathrm{CBS}}(u) - \kappa \min V = (\alpha+1)^{-1}\,\frac{1}{2}\|u-u^*\|^2_{\mathcal{C}_\alpha(\rho_t)}. \tag{29}$$

In other words, in the limit corresponding to CBS (i.e., $\kappa\to\infty$ with $\alpha=\beta$ and $\gamma = (\alpha+1)^{-1}\kappa$), the first approximation step in this section "infinitely smooths" the potential V into a quadratic centered at the minimizer of V for the purpose of approximating its gradient.

Remark 3 ((Weak) convexity of V)**.** *The above connection between Langevin dynamics and CBS requires convexity of V such that $\nabla V^{\kappa\mathcal{C}_\alpha(\rho_t)}$ exists for any κ; the derivation leading up to the approximate Langevin dynamics (25) also allows V to be weakly convex with sufficiently small κ.*

Weakly convex potentials are a natural setting for sampling methods. In addition, since the approximations in this section are only a motivation for CBS and (25), those dynamics might also be effective outside those assumptions. This should be investigated in future work.

4 Localized consensus-based sampling

Armed with the novel interpretation of CBS in section 3, this section derives a new gradient-free sampling method. Our proposed localized consensus-based sampling algorithm retains affine-invariance, is derivative-free, and performs well on non-Gaussian distributions. We study localized CBS in time-continuous, mean-field form; section 6 will then consider discretizations.

4.1 Consensus-based sampling without Gaussian assumption

Recall that our goal is a gradient-free sampling method with good performance on non-Gaussian distributions. We consider the approximations in section 3 that transform interacting Langevin diffusions to CBS, and adapt them to suit this new objective.

The $\kappa \to \infty$ limit of (25), leading to CBS, smooths V to a quadratic. Since we want to sample from non-Gaussian π—i.e., non-quadratic V—we keep κ finite. In addition, the Moreau envelope approximation (18) to V suggests that κ should be chosen small, while the weighted-mean approximation (23) to the proximal operator implies that β should be large. Finally, we note that the steps leading up to (25) are not specific to the preconditioner $\mathcal{C}_\alpha(\rho_t)$ and allow a general positive definite preconditioner $\mathbf{P}_t \coloneqq \mathbf{P}(\rho_t)$. This results in our novel localized CBS dynamics

$$\mathrm{d}U_t = -\frac{\gamma}{\kappa}\Big(U_t - \mu_{\beta,\kappa\mathbf{P}_t}(\rho_t; U_t)\Big)\,\mathrm{d}t + \sqrt{2\mathbf{P}_t}\,\mathrm{d}W_t. \tag{30}$$

Here, κ controls the accuracy of the Moreau envelope approximation (i.e., the amount of "smoothing" of V) and β controls the accuracy of the weighted-mean approximation to the proximal operator (i.e., the accuracy of the gradient estimation). The preconditioner $\mathbf{P}_t$ can be chosen freely. Suitable choices include

(i) a constant matrix $\mathbf{P}_t = K$;

(ii) an unweighted covariance $\mathbf{P}_t = \mathcal{C}(\rho_t)$; and

(iii) a weighted covariance $\mathbf{P}_t = \mathcal{C}_\alpha(\rho_t)$, which is used in classical CBS. Larger α puts more weight on particles with low potential values. This pushes $\mathbf{P}_t$ more towards the target covariance, rather than to the covariance of ρ_t, which may be beneficial when those differ significantly in their scaling.

Finally, the parameter γ could simply be set to 1, but other choices may be able to partially mitigate the bias that is due to $\beta < \infty$ and $\kappa > 0$. We study this bias in more detail in subsection 4.3 as a function of γ, κ, and the choice of $\mathbf{P}_t$. Notably, when π is Gaussian and $\mathbf{P}_t$ is an unweighted or weighted covariance—choices (ii) and (iii) above—the bias can be removed fully. The value of γ that accomplishes this, given by (39) and (40), can then be used in (30) even in non-Gaussian cases as a heuristic.

4.2 Localized preconditioners

This subsection discusses more general preconditioners for the proposed localized CBS method, at the cost of a correction term and additional computational effort in the resulting algorithm. Concretely, we allow a preconditioner $\mathbf{P}_{U,t} \coloneqq \mathbf{P}(\rho_t; U_t)$ that depends on the variable U_t as well as its law ρ_t[3]. As discussed in subsection 2.1, this requires an additional drift term $\nabla_u \cdot \mathbf{P}_{U,t}$:

$$\mathrm{d}U_t = -\frac{\gamma}{\kappa}\Big(U_t - \mu_{\beta,\kappa\mathbf{P}_{U,t}}(\rho_t; U_t)\Big)\,\mathrm{d}t + \nabla_u \cdot \mathbf{P}_{U,t} + \sqrt{2\mathbf{P}_{U,t}}\,\mathrm{d}W_t. \tag{31}$$

Computation of $\nabla_u \cdot \mathbf{P}_{U,t}$ is discussed in section 6. The corresponding Fokker–Planck equation is

$$\partial_t \rho_t(u) = \nabla \cdot \Big(\frac{\gamma}{\kappa}\rho_t(u)\,\big(u - \mu_{\beta,\kappa\mathbf{P}_{U,t}}(\rho_t; u)\big) + \mathbf{P}_{U,t}\nabla\rho_t(u)\Big). \tag{32}$$

A particular choice of interest for the space-dependent preconditioner is

(iv) a weighted covariance $\mathbf{P}_{U,t} = \mathcal{C}_{\alpha,\alpha\lambda\mathcal{C}(\rho_t)}(\rho_t; U_t)$ with scalars $\alpha \geq 0$ and $\lambda > 0$, where

$$\mathcal{C}_{a,A}(\rho; u) \coloneqq \frac{\int (v - \mu_{a,A}(\rho; u)) \otimes (v - \mu_{a,A}(\rho; u))\, w_{a,A}(v; u)\,\mathrm{d}\rho(v)}{\int w_{a,A}(v; u)\,\mathrm{d}\rho(v)}, \tag{33}$$

Equation (33) uses the weighted mean and weight function (24). With this preconditioner, the bias of localized CBS can be removed for Gaussian π, similarly to the covariances in subsection 4.1, with the γ value given by (41). This weighted covariance is interesting for comparison to the *polarized CBS* method discussed in subsection 5.2, since $\mathcal{C}_{\alpha,\alpha\lambda\mathcal{C}(\rho_t)}(\rho; u)$ corresponds to the polarized covariance (55b) with a Gaussian kernel[4] $k(u, v) = \exp(-(2\lambda)^{-1}\|u - v\|^2_{\mathcal{C}(\rho_t)})$. Using such a weighted covariance can incur a significant extra computational cost and, in our experiments, does not seem necessary for good performance of localized CBS.

[3] When the preconditioner can depend on U_t, a motivation similar to that in section 3 is still possible. In this case, the property (19) gains a quadratic term. As $\|A\| \to 0$, this term vanishes, which justifies the use of (19) as an approximation.

[4] Polarized CBS does not include $\mathcal{C}(\rho_t)$ in the distance norm, so the correspondence is not perfect. We add this anisotropy in our covariance weights to ensure affine-invariance (see subsection 4.4).

4.3 Analysis for Gaussian distributions

While the localized CBS dynamics (31) are designed to sample from non-Gaussian distributions, we will now analyze their behavior in the case where the target distribution is Gaussian, i.e., when $\pi(u) \propto \widehat{\pi}(u) = \exp(-\frac{1}{2}\|u-m\|_\Sigma^2)$ for some m and Σ. This can be seen as verifying a minimum requirement: we can hardly expect the method to work well for non-Gaussian distributions if it cannot handle Gaussian ones. Moreover, this analysis results in values for γ that we will later use even for non-Gaussian π.

Lemma 4. *Assume that both the target distribution $\pi = \mathcal{N}(m, \Sigma)$ and the initial condition $U_0 \sim \mathcal{N}(m_0, \Sigma_0)$ are Gaussian. Then, the localized CBS dynamics (31) admit a Gaussian solution $U_t \sim \mathcal{N}(m_t, \Sigma_t)$ with mean and covariance that satisfy*

$$\frac{\mathrm{d}}{\mathrm{d}t} m_t = -\frac{\gamma\beta}{\kappa} Q_t \Sigma^{-1}(m_t - m), \tag{34a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} \Sigma_t = 2\Big[\mathbf{P}_{*,t} - \frac{\gamma}{\kappa}\Sigma_t + \frac{\gamma\beta}{2\kappa^2}\left(Q_t \mathbf{P}_{*,t}^{-1}\Sigma_t + \Sigma_t \mathbf{P}_{*,t}^{-1} Q_t\right)\Big], \tag{34b}$$

with $Q_t := (\beta\Sigma^{-1} + \Sigma_t^{-1} + \frac{\beta}{\kappa}\mathbf{P}_{*,t}^{-1})^{-1}$, if the preconditioner $\mathbf{P}_{*,t} := \mathbf{P}(\rho_t; U_t)$ does not depend on U_t when $\rho_t = \mathcal{N}(m_t, \Sigma_t)$. This is the case for the preconditioners considered before:

$$\mathbf{P}(\rho_t; U_t) = K, \tag{35a}$$

$$\mathbf{P}(\rho_t; U_t) = \mathcal{C}(\rho_t) = \Sigma_t, \tag{35b}$$

$$\mathbf{P}(\rho_t; U_t) = \mathcal{C}_\alpha(\rho_t) = (\alpha\Sigma^{-1} + \Sigma_t^{-1})^{-1}, \tag{35c}$$

$$\mathbf{P}(\rho_t; U_t) = \mathcal{C}_{\alpha,\alpha\lambda\mathcal{C}(\rho_t)}(\rho_t; U_t) = (\alpha\Sigma^{-1} + (1+\lambda^{-1})\Sigma_t^{-1})^{-1}, \tag{35d}$$

with parameters $\alpha \geq 0$, $\lambda > 0$, and $K \succ 0$.

Proof. Assume $\rho_t = \mathcal{N}(m_t, \Sigma_t)$. From the condition that $\mathbf{P}_{*,t}$ not depend on U_t follows that $\nabla_u \cdot \mathbf{P}_{*,t} = 0$. Appendix A derives the rightmost terms in (35b)–(35d) and

$$\mu_{\beta,\kappa\mathbf{P}_{*,t}}(\rho_t; U_t) = (\beta\Sigma^{-1} + \Sigma_t^{-1} + \frac{\beta}{\kappa}\mathbf{P}_{*,t}^{-1})^{-1}(\beta\Sigma^{-1}m + \Sigma_t^{-1}m_t + \frac{\beta}{\kappa}\mathbf{P}_{*,t}^{-1}U_t). \tag{36}$$

In this case, (31) becomes a *linear SDE*: the drift term depends on U_t in an affine way and the diffusion term is U_t-independent. Following, e.g., [55, Section 6.1], the SDE admits a Gaussian solution with mean and covariance that satisfy

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} m_t &= -\frac{\gamma}{\kappa}(I - \frac{\beta}{\kappa}Q_t\mathbf{P}_{*,t}^{-1})m_t + \frac{\gamma}{\kappa}Q_t\left(\beta\Sigma^{-1}m + \Sigma_t^{-1}m_t\right) \\ &= -\frac{\gamma}{\kappa}\Big[(I - Q_t\left(Q_t^{-1} - \beta\Sigma^{-1}\right))m_t - Q_t\beta\Sigma^{-1}m\Big] = -\frac{\gamma\beta}{\kappa}Q_t\Sigma^{-1}\left(m_t - m\right),\end{aligned} \tag{37}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\Sigma_t = -\frac{\gamma}{\kappa}(I - \frac{\beta}{\kappa}Q_t\mathbf{P}_{*,t}^{-1})\Sigma_t - \frac{\gamma}{\kappa}\Sigma_t\left(I - \frac{\beta}{\kappa}\mathbf{P}_{*,t}^{-1}Q_t\right) + 2\mathbf{P}_{*,t}. \tag{38}$$

□

Corollary 5. *Assume that the target distribution $\pi = \mathcal{N}(m, \Sigma)$ is Gaussian. The localized CBS dynamics admit π as a stationary distribution in the following cases.*

(i) The preconditioner is the unweighted covariance $\mathbf{P}_{U,t} = \mathcal{C}(\rho_t)$ and

$$\gamma = \kappa + \frac{\beta}{\beta+1}. \tag{39}$$

(ii) The preconditioner is the weighted covariance $\mathbf{P}_{U,t} = \mathcal{C}_\alpha(\rho_t)$ with $\alpha \geq 0$ and

$$\gamma = (\alpha+1)^{-1}\kappa + \frac{\beta}{\beta+1}. \tag{40}$$

(iii) The preconditioner is the weighted covariance $\mathbf{P}_{U,t} = \mathcal{C}_{\alpha,\alpha\lambda\mathcal{C}(\rho_t)}(\rho_t; U_t)$ with $\alpha \geq 0$, $\lambda > 0$, and

$$\gamma = \left(\frac{1}{\lambda} + \alpha + 1\right)^{-1}\kappa + \frac{\beta}{\beta+1}. \tag{41}$$

Proof. If π is a stationary distribution of localized CBS, then $\partial_t \rho_t(u) = 0$ when $\rho_t = \pi$. Since then both ρ_t and π are Gaussian, we can use Lemma 4, fill in an expression for the preconditioner such as (35), and require $\frac{\mathrm{d}}{\mathrm{d}t} m_t = 0$ and $\frac{\mathrm{d}}{\mathrm{d}t}\Sigma_t = 0$ when $m_t = m$ and $\Sigma_t = \Sigma$. Clearly, $\frac{\mathrm{d}}{\mathrm{d}t} m_t$ as given by (34a) is zero irrespective of γ and $\mathbf{P}_{*,t}$ when $m_t = m$. We also have

$$\frac{\mathrm{d}}{\mathrm{d}t}\Sigma_t = 2\Big[\mathbf{P}_{*,t} - \frac{\gamma}{\kappa}\Sigma + \frac{\gamma\beta}{2\kappa^2}\Big(Q_t \mathbf{P}_{*,t}^{-1}\Sigma + \Sigma\mathbf{P}_{*,t}^{-1} Q_t\Big)\Big]. \tag{42}$$

When $\mathbf{P}_{*,t}$ and $\Sigma_t = \Sigma$ commute—which is the case for the three preconditioners considered in this corollary, as seen in (35)—we can simplify (42) to

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\Sigma_t &= 2\Big[\mathbf{P}_{*,t} + \frac{\gamma}{\kappa}\Big(\big(\kappa\frac{\beta+1}{\beta}\mathbf{P}_{*,t}\Sigma^{-1} + I\big)^{-1} - I\Big)\Sigma\Big] \\ &= 2\Big[\mathbf{P}_{*,t} - \frac{\gamma}{\kappa}\big(\kappa\frac{\beta+1}{\beta}\mathbf{P}_{*,t}\Sigma^{-1} + I\big)^{-1}\big(\kappa\frac{\beta+1}{\beta}\mathbf{P}_{*,t}\Sigma^{-1}\big)\Sigma\Big] \\ &= 2\Big[\mathbf{P}_{*,t} - \gamma\big(\kappa\Sigma^{-1} + \frac{\beta}{\beta+1}\mathbf{P}_{*,t}^{-1}\big)^{-1}\Big],\end{aligned} \tag{43}$$

which, after some algebraic manipulations, can be seen to equal zero when

$$\kappa I = \Big(\gamma - \frac{\beta}{\beta+1}\Big)\Sigma\mathbf{P}_{*,t}^{-1}. \tag{44}$$

Equation (35d) becomes $\mathbf{P}_{*,t} = (\lambda^{-1} + \alpha + 1)\Sigma$ when $\Sigma_t = \Sigma$, which means that (44) holds when γ is given by (41). The other cases correspond to specific choices for λ and α. □

A statement such as Corollary 5 about the preconditioner $\mathbf{P}_{*,t} = K$ is not possible, since there does not exist a γ for which (42) is satisfied for all Σ in this case.

Remark 6 (Non-Gaussian performance)**.** *In the non-Gaussian case, localized CBS with finite parameters (β, κ) will not sample from the exact target distribution π in general. Characterizing the well-posedness of (31) and the relation between π and a possible stationary distribution ρ_∞ of localized CBS is an open question.*

4.4 Affine-invariance

We now study affine-invariance of localized CBS.

Lemma 7. *The SDE (31) is affine-invariant (see Definition 2) when the preconditioner is given by $\mathbf{P}(\rho_t; U_t) = \mathcal{C}(\rho_t)$, $\mathbf{P}(\rho_t; U_t) = \mathcal{C}_\alpha(\rho_t)$, or $\mathbf{P}(\rho_t; U_t) = \mathcal{C}_{\alpha,\alpha\lambda\mathcal{C}(\rho_t)}(\rho_t; U_t)$.*

Proof. To prove affine-invariance following Definition 2, we use the Fokker–Planck equation (32) and make the dependence on $\widehat{\pi}$ explicit:

$$\partial_t \rho_t(u) = \nabla \cdot \Big(\frac{\gamma}{\kappa}\rho_t(u)\big(u - \mu^{\widehat{\pi}}_{\beta,\kappa\mathbf{P}^{\widehat{\pi}}(\rho_t;u)}(\rho_t; u)\big) + \mathbf{P}^{\widehat{\pi}}(\rho_t; u)\nabla\rho_t(u)\Big). \tag{45}$$

With the transformed $\tilde{\pi}(z) = \widehat{\pi}(Mz + b)$, the Fokker–Planck equation looks like

$$\partial_t \tilde{\rho}_t(z) = \nabla \cdot \Big(\frac{\gamma}{\kappa}\tilde{\rho}_t(z)\big(z - \mu^{\tilde{\pi}}_{\beta,\kappa\mathbf{P}^{\tilde{\pi}}(\tilde{\rho}_t;z)}(\tilde{\rho}_t; z)\big) + \mathbf{P}^{\tilde{\pi}}(\tilde{\rho}_t; z)\nabla\tilde{\rho}_t(z)\Big). \tag{46}$$

Now define

$$\bar{\rho}_t(z) = |M|\,\rho_t(Mz + b); \tag{47}$$

then, by Definition 2, localized CBS is affine-invariant if (46) holds for $\tilde{\rho}_t = \bar{\rho}_t$. As a first step, one can check that these preconditioners satisfy (with $u = Mz + b$)

$$\mathbf{P}^{\widehat{\pi}}(\rho_t; u) = M\mathbf{P}^{\tilde{\pi}}(\bar{\rho}_t; z)M^T \tag{48}$$

and, thus, that

$$\mu^{\widehat{\pi}}_{\beta,\kappa\mathbf{P}^{\widehat{\pi}}(\rho_t;u)}(\rho_t; u) = M\mu^{\tilde{\pi}}_{\beta,\kappa\mathbf{P}^{\tilde{\pi}}(\bar{\rho}_t;z)}(\bar{\rho}_t; z) + b. \tag{49}$$

We can then write

$$\begin{aligned}\partial_t\bar{\rho}_t(z) &= |M|\,\partial_t\rho_t(u) \\ &= |M|\,\nabla_u \cdot \Big(\frac{\gamma}{\kappa}\rho_t(u)\big(u - \mu^{\widehat{\pi}}_{\beta,\kappa\mathbf{P}^{\widehat{\pi}}(\rho_t;u)}(\rho_t; u)\big) + \mathbf{P}^{\widehat{\pi}}(\rho_t; u)\nabla_u\rho_t(u)\Big) \\ &= \nabla_u \cdot \Big(\frac{\gamma}{\kappa}M\bar{\rho}_t(z)\big(z - \mu^{\tilde{\pi}}_{\beta,\kappa\mathbf{P}^{\tilde{\pi}}(\bar{\rho}_t;z)}(\bar{\rho}_t; z)\big) + M\mathbf{P}^{\tilde{\pi}}(\bar{\rho}_t; z)\nabla_z\bar{\rho}_t(z)\Big) \\ &= \nabla_z \cdot \Big(\frac{\gamma}{\kappa}\bar{\rho}_t(z)\big(z - \mu^{\tilde{\pi}}_{\beta,\kappa\mathbf{P}^{\tilde{\pi}}(\bar{\rho}_t;z)}(\bar{\rho}_t; z)\big) + \mathbf{P}^{\tilde{\pi}}(\bar{\rho}_t; z)\nabla_z\bar{\rho}_t(z)\Big).\end{aligned}$$

This proves affine-invariance. □

5 Comparison to alternative methods

This section compares the dynamics (31) of localized CBS to those of various other McKean–Vlasov-based samplers. Subsection 5.1 discusses classical CBS, after which subsections 5.2 and 5.3 compares our method to two algorithms that use localization in a similar way. Various other dynamics are briefly discussed in subsection 5.4. A detailed numerical comparison to some of these methods follows in section 7; here, we focus on different aspects of the dynamics and how these may affect the properties of the resulting algorithms.

5.1 Consensus-based sampling

We derived localized CBS in section 4 as a generalization of classical CBS to non-Gaussian sampling problems. We will now show that, when both the target distribution π and the initial particle distribution ρ_0 are Gaussian, the localized CBS dynamics (31) with a certain weighted covariance as preconditioner admit the same solution as the CBS dynamics (15).

Lemma 8. *Let $U_t \sim \mathcal{N}(m_t, \Sigma_t)$ follow the localized CBS dynamics (31) preconditioned with the weighted covariance $\mathbf{P}(\rho; u) = \mathcal{C}_{\alpha,\alpha\lambda\mathcal{C}(\rho)}(\rho; u)$ and let $U'_t \sim \mathcal{N}(m'_t, \Sigma'_t)$ follow the CBS dynamics (15). Let both the target density $\pi = \mathcal{N}(m, \Sigma)$ and the initial densities of both dynamics $\mathcal{N}(m_0, \Sigma_0) = \mathcal{N}(m'_0, \Sigma'_0)$ be Gaussian.*

For any choice of the CBS parameter[5] *$\alpha' > 0$ and with localized CBS parameters*

$$\alpha > 0, \qquad \kappa = \alpha, \qquad \beta = \alpha', \qquad \text{and} \qquad \lambda = \frac{\alpha'}{\alpha - \alpha'}, \tag{50}$$

it holds that $m_t = m'_{ct}$ and $\Sigma_t = \Sigma'_{ct}$ with

$$c = \frac{\alpha'}{\alpha\,(\alpha' + 1)}. \tag{51}$$

Proof. In [4], it is shown that

$$\frac{\mathrm{d}}{\mathrm{d}t}\Sigma'_t = -2\alpha'\Sigma'_t\,(\Sigma + \alpha'\Sigma'_t)^{-1}(\Sigma'_t - \Sigma). \tag{52}$$

On the other hand, filling in (50) into (34b) with (35d) results in

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\Sigma_t &= 2\Big[\mathbf{P} - 2\frac{\alpha'}{\alpha(\alpha'+1)}\Sigma_t + \frac{(\alpha')^2}{\alpha^2(\alpha'+1)}\left(Q_t\mathbf{P}^{-1}\Sigma_t + \Sigma_t\mathbf{P}^{-1}Q_t\right)\Big] \\
&= 2\Big[\mathbf{P} - 2c\Sigma_t + c\frac{\alpha'}{\alpha}\big(\frac{\alpha}{2\alpha'}I\Sigma_t + \Sigma_t\frac{\alpha}{2\alpha'}I\big)\Big] \\
&= 2\Big[\big(\alpha\Sigma^{-1} + \frac{\alpha}{\alpha'}\Sigma_t^{-1}\big)^{-1} - c\Sigma_t\Big] \\
&= 2\Big[c\big(\Sigma_t - \alpha'\Sigma_t\,(\Sigma + \alpha'\Sigma_t)^{-1}(\Sigma_t - \Sigma)\big) - c\Sigma_t\Big] \\
&= -2c\alpha'\Sigma_t\,(\Sigma + \alpha'\Sigma_t)^{-1}(\Sigma_t - \Sigma),
\end{aligned}$$

where we noted that $\gamma = 2\frac{\alpha'}{\alpha'+1}$. This is (52) accelerated with a factor c; hence, $\Sigma_t = \Sigma'_{ct}$.

Again from [4], we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}m'_t = -\alpha'\Sigma'_t\,(\Sigma + \alpha'\Sigma'_t)^{-1}(m'_t - m). \tag{53}$$

For localized CBS, (34a) becomes

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}m_t &= -2\alpha' c Q_t\Sigma^{-1}\,(m_t - m) = -c\alpha'\Sigma_t\,(\Sigma + \alpha'\Sigma_t)^{-1}(m_t - m) \\
&= -c\alpha'\Sigma'_{ct}\,(\Sigma + \alpha'\Sigma'_{ct})^{-1}(m_t - m).
\end{aligned}$$

This is the ODE (53) accelerated with a factor c. We conclude that $m_t = m'_{ct}$. □

For non-Gaussian distributions, this configuration of localized CBS likely performs quite poorly due to the large κ value, which would result in excessive smoothing of the potential function V.

As it concerns only Gaussian distributions, Lemma 8 is mainly of theoretical interest. A more general link between localized CBS and CBS is the $\kappa \to \infty$ limit from section 3.

[5] We write α' for the CBS parameter to distinguish it from the α parameter of localized CBS.

5.2 Polarized consensus-based sampling

A recently proposed modification of CBS, called polarized CBS [3], makes the weight function dependent on the position U_t through a *polarization kernel* $k(\cdot,\cdot)$. This results in the dynamics

$$\mathrm{d}U_t = -(U_t - \tilde{\mu}_{\alpha,k}(\rho_t; U_t))\,\mathrm{d}t + \sqrt{2(\alpha+1)\,\tilde{\mathcal{C}}_{\alpha,k}(\rho_t; U_t)}\,\mathrm{d}W_t, \tag{54}$$

where we use the weights $\tilde{w}_{\alpha,k}(v;u) = k(v,u)\,\widehat{\pi}(v)^\alpha$ to define a polarized mean and covariance

$$\tilde{\mu}_{\alpha,k}(\rho;u) \coloneqq \frac{\int v\tilde{w}_{\alpha,k}(v;u)\,\mathrm{d}\rho(v)}{\int \tilde{w}_{\alpha,k}(v;u)\,\mathrm{d}\rho(v)}, \tag{55a}$$

$$\tilde{\mathcal{C}}_{\alpha,k}(\rho;u) \coloneqq \frac{\int (v-\tilde{\mu}_{\alpha,k}(\rho;u))\otimes(v-\tilde{\mu}_{\alpha,k}(\rho;u))\,\tilde{w}_{\alpha,k}(v;u)\,\mathrm{d}\rho(v)}{\int \tilde{w}_{\alpha,k}(v;u)\,\mathrm{d}\rho(v)}. \tag{55b}$$

The main kernel used in [3] is Gaussian: for positive definite matrix D and scalar λ,

$$k(v,u) = \exp\Big(-(2\lambda)^{-1}\|v-u\|_D^2\Big). \tag{56}$$

This choice of k introduces a localization effect: the polarized mean $\tilde{\mu}_{\alpha,k}(\rho_t;U_t)$—the "target" of the drift in (54)—tends to produce a value that is close to U, as well as having a large density in the posterior $\pi \propto \widehat{\pi}$. When π is Gaussian and (56) is used for the kernel, [3, Proposition 1] proves that π is a stationary distribution of (54) with (56). More generally, a localized kernel attempts to lessen the Gaussian assumption present in CBS in order to allow multimodal sampling.

Equation (54) is very reminiscent of the localized CBS dynamics (31). We now compare the two, highlighting three main differences.

(i) Localized CBS includes the correction term $\nabla_u \cdot \mathbf{P}_{U,t}$. As $\beta\to\infty$ and $\kappa\to 0$, we expect convergence to the interacting Langevin dynamics, under which even non-Gaussian and multimodal π are invariant. In polarized CBS, we do not expect a similar convergence.

(ii) By including the preconditioner $\mathbf{P}_{U,t}$ in the distance weighting for the localized mean, localized CBS has affine-invariant dynamics when $\mathbf{P}_{U,t}$ is a weighted or unweighted covariance. The kernels proposed for polarized CBS in [3] are not affine-invariant as they weight the distance between two particles by a fixed matrix.

One could modify the kernel in polarized CBS to make it affine-invariant, e.g., by replacing D by $\mathcal{C}(\rho_t)$. While interesting, this would not change the other benefits of localized CBS.

(iii) Using $\widehat{\pi}$-based and distance-based weights in the (relatively cheap) mean computation is crucial in both methods, since that drives evolution towards high-probability regions and allows multimodal sampling. On the other hand, we find that the benefits of using weights in the (relatively expensive) covariance preconditioner are significantly less clear. Yet, this can be costly: distance-weighted statistics need to be computed for each of the J particles, instead of once, globally. Polarized CBS uses the same weighting for both the mean and the covariance, while localized CBS decouples these two elements and allows to use a cheap, unweighted covariance as a preconditioner. This can make localized CBS the computationally cheaper method in higher-dimensional problems.

5.3 Localized ALDI

The gradient-free McKean–Vlasov dynamics underlying the closely related EKS and ALDI algorithms [22, 21] are described by

$$\mathrm{d}U_t = -\mathcal{C}^G(\rho_t)\Gamma^{-1}\,(G(U_t)-y)\,\mathrm{d}t + \mathcal{C}(\rho_t)\Gamma_0^{-1}\,(U_t-u_0)\,\mathrm{d}t + \sqrt{2\mathcal{C}(\rho_t)}\,\mathrm{d}W_t, \tag{57}$$

where

$$\mu^G(\rho) \coloneqq \int G(v)\,\mathrm{d}\rho(v) \quad \text{and} \quad \mathcal{C}^G(\rho) \coloneqq \int (v-\mu(\rho))\otimes(G(v)-\mu^G(\rho))\,\mathrm{d}\rho(v). \tag{58}$$

Equation (57) approaches the sampling problem (1) explicitly for a Bayesian inverse problem (2), estimating the parameters of a model G. It assumes that $\pi_{\text{prior}} = \mathcal{N}(u_0,\Gamma_0)$ and $\pi_{\text{noise}} = \mathcal{N}(0,\Gamma)$ are Gaussians. When G is linear, (57) is equivalent to (8) and, hence, has π as an invariant distribution [22]. For nonlinear G, this is generally not the case.

In [51], localization with a scalar λ and positive definite matrix D is added to (57), together with a correction term similar to the one in (14). This can be seen as Langevin diffusions where the gradients are approximated using the method in [63]. The resulting dynamics are

$$\begin{aligned} \mathrm{d}U_t = &-\tilde{\mathcal{C}}^G_{\lambda D}(\rho_t; U_t)\Gamma^{-1}\left(G(U_t) - y\right)\mathrm{d}t - \tilde{\mathcal{C}}_{\lambda D}(\rho_t; U_t)\Gamma_0^{-1}\left(U_t - u_0\right)\mathrm{d}t \\ &+ \nabla_u \cdot \tilde{\mathcal{C}}_{\lambda D}(\rho_t; U_t) + \sqrt{2\tilde{\mathcal{C}}_{\lambda D}(\rho_t; U_t)}\,\mathrm{d}W_t, \end{aligned} \tag{59}$$

where we use $\tilde{w}_{\lambda D}(v; u) = \exp(-(2\lambda)^{-1}\|v - u\|_D^2)$ to define

$$\tilde{\mu}_{\lambda D}(\rho; u) \coloneqq \frac{\int v\tilde{w}_{\lambda D}(v; u)\,\mathrm{d}\rho(v)}{\int \tilde{w}_{\lambda D}(v; u)\,\mathrm{d}\rho(v)}, \tag{60a}$$

$$\tilde{\mathcal{C}}_{\lambda D}(\rho; u) \coloneqq \frac{\int (v - \tilde{\mu}_{\lambda D}(\rho; u)) \otimes (v - \tilde{\mu}_{\lambda D}(\rho; u))\,\tilde{w}_{\lambda D}(v; u)\,\mathrm{d}\rho(v)}{\int \tilde{w}_{\lambda D}(v; u)\,\mathrm{d}\rho(v)}. \tag{60b}$$

The definition of $\tilde{\mathcal{C}}^G_{\lambda D}$ is analogous. Localization in ALDI (which, we stress again, differs from the concept of localization discussed in Remark 1) improves performance for nonlinear G. However, we point out two potential comparative advantages of the localized CBS dynamics.

(i) Similarly to polarized CBS, when the matrix D in localized ALDI is fixed, the dynamics are not affine-invariant. In [51], affine-invariance is proven if $D = \mathcal{C}(\rho_0)$, such that D is also subject to any affine transformation. This can be considered a slightly weaker form of affine-invariance: in practice, the scaling of the parameters is often not fully known and ρ_0 cannot be chosen with optimal scaling. After all, if the scaling was known exactly, we could rescale the problem manually and would not need affine-invariant samplers. Whereas the effect of the initial distribution may diminish over time in localized CBS (see subsection 7.2 and Figure 3 in particular), choosing $D = \mathcal{C}(\rho_0)$ in localized ALDI ensures that an incorrectly scaled initial distribution persists.

Analogously to subsection 5.2 for polarized CBS, it is conceivable that replacing D by $\mathcal{C}(\rho_t)$ could improve the localized ALDI algorithm by making it affine-invariant.

(ii) As in polarized CBS, localized ALDI uses weighted covariance matrices, which can be more expensive to compute than unweighted ones in localized CBS.

5.4 Other methods

The localized CBS dynamics (31) arise from a specific approximation of the gradient in preconditioned Langevin dynamics, as detailed in section 3. Several other gradient-free Langevin-type variants have been proposed in the literature. Besides localized ALDI, discussed in subsection 5.3, notable examples include the following methods.

(i) The method of [13] approximates gradients using only particles within a fixed neighborhood of the target particle. The authors of [13] report that this approach can become unstable because isolated particles may experience unrealistically large approximate gradients. To mitigate this, they propose evaluating the true gradient at such particles while the remaining ones use the ensemble-based approximation. Although conceptually interesting—and potentially adaptable to localized CBS—this algorithm is not fully gradient-free and is therefore applicable in a more limited setting than localized CBS.

(ii) The multiscale sampler of [47] introduces two types of particles: a main (*macro*) particle that represent the distribution and a local swarm of auxiliary (*micro*) particles used to estimate gradients. This construction can handle non-Gaussian targets and may scale favorably with problem dimension because the particles stay tightly clustered. However, only the macro particle contributes to sampling; the computational effort invested in the auxiliary particles serves solely for gradient estimation. Moreover, since all particles in the swarm remain close, the method can explore only one mode at a time, which leads to metastability in multimodal problems. In contrast, every particle in localized CBS contributes to the sample ensemble, and the method can explore multiple modes simultaneously.

(iii) There exist many Langevin-type samplers using the proximal operator (e.g., [18, 37, 48, 61]), often with the goal of sampling from non-differentiable distributions. These are not ensemble methods and they are neither affine-invariant nor gradient-free. However, they typically scale very well to high dimensions, which is useful in applications such as Bayesian imaging.

Related to these algorithms, [36] uses *Tweedie's identity*

$$\nabla \widetilde{V}^\kappa(u) = \kappa^{-1}(u - \widetilde{D}^\kappa(u)) \tag{61}$$

with

$$\widetilde{V}^\kappa(u) = -\log\left(\int \exp(-\|u-v\|_2^2/(2\kappa))\exp(-V(v))\,\mathrm{d}v\right) \tag{62}$$

and the *denoiser*

$$\widetilde{D}^\kappa(u) = \frac{\int v \exp(-\|u-v\|_2^2/(2\kappa))\exp(-V(v))\,\mathrm{d}v}{\int \exp(-\|u-v\|_2^2/(2\kappa))\exp(-V(v))\,\mathrm{d}v} \tag{63}$$

to replace the gradient of a nondifferentiable part of the target potential by $\kappa^{-1}(u - \widetilde{D}^\kappa(u))$. Tweedie's identity offers a second two-step approximation scheme to retrieve (25) from interacting Langevin: approximate V by $\widetilde{V}^\kappa$ (with certain adaptations for weighting the exponential by β and the distance by $\mathbf{P}_t$) to apply Tweedie's identity, and then approximate the integrals over the Lebesgue measure in (63) by integrals over ρ_t.

6 Discretizations and computational considerations

We now bridge the gap between the theoretical localized CBS dynamics (31) and practical algorithms to sample from (1).

Particle discretization. The dynamics (31) of localized CBS are defined as a McKean–Vlasov SDE: the law of U_t factors into its evolution. Of course, this law is not available in practice and (31) does not yet constitute an algorithm. We now introduce particle approximations, i.e., schemes where J approximate instances of (31) are run in parallel. They use their own empirical distribution instead of their unknown exact law to drive their evolution. We denote the ith particle by U_t^i and the entire ensemble by the $d \times J$ matrix $\boldsymbol{U}_t = (U_t^1, U_t^2, \dots, U_t^J)$.

When $\mathbf{P}(\rho_t; U_t)$ is an unweighted or weighted covariance, we replace it by a particle approximation: a (weighted) sample covariance of all particles defined as

$$\widehat{\mathcal{C}}(\boldsymbol{U}) \coloneqq \frac{1}{J}\sum\nolimits_{j=1}^{J}(U^j - \widehat{\mu}(\boldsymbol{U})) \otimes (U^j - \widehat{\mu}(\boldsymbol{U})) \qquad \text{with} \qquad \widehat{\mu}(\boldsymbol{U}) \coloneqq \frac{1}{J}\sum\nolimits_{j=1}^{J} U^j \tag{64}$$

for the unweighted covariance and, for the weighted covariance,

$$\widehat{\mathcal{C}}_{\alpha,\lambda\widehat{\mathcal{C}}(\boldsymbol{U})}(\boldsymbol{U}; U^i) \coloneqq \sum\nolimits_{j=1}^{J} \omega^{ij}_{\alpha,\lambda\widehat{\mathcal{C}}(\boldsymbol{U})}\,(U^j - \widehat{\mu}_{\alpha,\lambda\widehat{\mathcal{C}}(\boldsymbol{U})}(\boldsymbol{U}; U^i)) \otimes (U^j - \widehat{\mu}_{\alpha,\lambda\widehat{\mathcal{C}}(\boldsymbol{U})}(\boldsymbol{U}; U^i)) \tag{65}$$

with

$$\widehat{\mu}_{\alpha,A}(\boldsymbol{U}; U^i) \coloneqq \sum\nolimits_{j=1}^{J} \omega^{ij}_{\alpha,A} U^j, \tag{66}$$

where $\omega^{ij}_{\alpha,A} = w^{ij}_{\alpha,A} / \sum_{k=1}^{J} w^{ik}_{\alpha,A}$ with $w^{ij}_{\alpha,A} = w_{\alpha,A}(U_j; U_i)$. In general, we denote the particle approximation to $\mathbf{P}(\rho; U^i)$ by $\widehat{\mathbf{P}}(\boldsymbol{U}; U^i)$. As in the mean-field setting of section 4, we assume that this preconditioner is invertible. Notably, this requires that the J particles span the entire space $\mathbb{R}^d$. Particle-based methods such as ALDI have been used with $J \ll d$ to sample in a linear subspace [21]; we leave the application of localized CBS to those settings to future work.

We also replace the weighted mean in (31) by a particle approximation. The most straightforward option is a weighted sample mean of all particles, given in (66). Later in this section, we will also explore a random-batch modification of this scheme.

Remark 9 (Scaling of the particle discretization to higher dimensions)**.** *The performance of localized CBS in high dimensions has not been theoretically analyzed. Since the method relies on local interactions between particles distributed throughout parameter space for gradient estimations, increasing dimensionality may increase the typical distance between particles and hence negatively affect performance. Clarifying this requires dedicated study.*

Time discretization. After particle discretization, we are left with a system of interacting SDEs. Since time discretization is not a main point of focus of this paper, we use the standard Euler–Maruyama discretization with N timesteps of length Δt in our numerical tests, although others are possible[6]. We denote the ith particle at timestep n by U_n^i and the ensemble by $\boldsymbol{U}_n$.

[6]Classical CBS uses a non-standard timestepping scheme, which helps them prove properties in the time-discrete setting. Polarized CBS adopts this same method. While EKS uses a linearly implicit split-step time discretization, both ALDI and localized ALDI use standard Euler–Maruyama.

Random-batch interaction. For the weighted mean in the drift term of localized CBS, we introduce a particle approximation that forbids self-to-self interaction (which would have large weights when particles are sparse) and randomly subsamples particles with a factor $\nu \in (0, 1]$. Consider

$$\widehat{\mu}^{(\nu,i)}_{\beta,\kappa\widehat{\mathbf{P}}}(\boldsymbol{U}_n; U^i_n) \coloneqq \frac{\sum_{j=1}^J \iota(i,j,n)\, U^j_n\, w^{ij}_{\alpha,\lambda\widehat{\mathbf{P}}}}{\sum_{j=1}^J \iota(i,j,n)\, w^{ij}_{\alpha,\lambda\widehat{\mathbf{P}}}}, \tag{67}$$

where ι selects whether to count the contribution of particle j:

$$\iota(i,j,n) \coloneqq \begin{cases} 0, & \text{if } i = j \text{ or } \theta^{i,j}_n > \nu, \text{ with } \theta^{i,j}_n \sim \mathcal{U}(0,1) \text{ i.i.d.,} \\ 1, & \text{otherwise.} \end{cases} \tag{68}$$

Here we write $\widehat{\mathbf{P}} \coloneqq \widehat{\mathbf{P}}(\boldsymbol{U}_n; U^i_n)$ for notational convenience. This idea is related to the random-batch method of [32], where interacting-particle systems are divided into small, random batches for a reduction in computational cost. We will choose relatively large values of ν (e.g., $\nu = 0.5$), so the cost difference is moderate; instead, we are motivated by our observation that random-batch interaction can improve sampling performance for higher-dimensional, multimodal problems. This is illustrated in section 7.

Square-root-free formulation. Similarly to what is proposed in [21], it is possible to replace the symmetric positive definite square root $\sqrt{\widehat{\mathbf{P}}} \in \mathbb{R}^{d\times d}$ by a (rectangular) Cholesky factor $\widehat{\mathbf{P}}^{1/2} \in \mathbb{R}^{d\times n}$, for any n, with the property that

$$\widehat{\mathbf{P}} = \widehat{\mathbf{P}}^{1/2}\,(\widehat{\mathbf{P}}^{1/2})^T. \tag{69}$$

By using $\widehat{\mathbf{P}}^{1/2}$ and simultaneously replacing the d-dimensional Brownian motion W_t by an n-dimensional one, the Fokker–Planck equation remains the same. Henceforth, let us define $\bar{U} \coloneqq \widehat{\mu}(\boldsymbol{U})$, $\bar{U}^{(i)} \coloneqq \widehat{\mu}_{\alpha,\lambda\widehat{\mathcal{C}}(\boldsymbol{U})}(\boldsymbol{U}; U^i)$, $w^{ij} \coloneqq w^{ij}_{\alpha,\lambda\widehat{\mathcal{C}}(\boldsymbol{U})}$, and $\omega^{ij} \coloneqq \omega^{ij}_{\alpha,\lambda\widehat{\mathcal{C}}(\boldsymbol{U})}$ for conciseness. The unweighted and weighted covariance have factors

$$\widehat{\mathcal{C}}^{1/2}(\boldsymbol{U}) = \frac{1}{\sqrt{J}}\begin{bmatrix}(U^1 - \bar{U}) & \dots & (U^J - \bar{U})\end{bmatrix} \tag{70}$$

$$\widehat{\mathcal{C}}^{1/2}_{\alpha,\lambda\widehat{\mathcal{C}}(\boldsymbol{U})}(\boldsymbol{U}; U^i) = \begin{bmatrix}\sqrt{\omega^{i1}}\,(U^1 - \bar{U}^{(i)}) & \dots & \sqrt{\omega^{iJ}}\,(U^J - \bar{U}^{(i)})\end{bmatrix}, \tag{71}$$

respectively. These quantities are readily available and avoid computing the square root of $\widehat{\mathbf{P}}$. This should offer computational advantages unless $J \gg d$.

Covariance-weighted norms. Within the weight function (24), localized CBS uses squared norms $\|U^i - U^j\|^2_{\widehat{\mathbf{P}}} = (U^i - U^j)^T\widehat{\mathbf{P}}^{-1}(U^i - U^j)$ weighted by the preconditioner. This quantity can be evaluated as follows: precompute

$$M \coloneqq \boldsymbol{U}^T\widehat{\mathbf{P}}^{-1}\boldsymbol{U}$$

and use

$$\left\|U^i - U^j\right\|^2_{\widehat{\mathbf{P}}} = M_{ii} + M_{jj} - 2M_{ij}.$$

If $\widehat{\mathbf{P}}$ depends on i, then so does M.

Correction term. The discretized dynamics contain the correction term $\nabla_{u^i}\cdot\widehat{\mathbf{P}}$, where we write $\widehat{\mathbf{P}} \coloneqq \widehat{\mathbf{P}}(\boldsymbol{U}_n; U^i_n)$. We now study this term for the preconditioners we consider in this article. When $\widehat{\mathbf{P}} = K$, clearly $\nabla_{u^i}\cdot\widehat{\mathbf{P}} = 0$; when $\widehat{\mathbf{P}} = \widehat{\mathcal{C}}_{\alpha,\lambda\widehat{\mathcal{C}}(\boldsymbol{U})}(\boldsymbol{U}; U)$, the correction term is given in Lemma 10. The preconditioners $\widehat{\mathcal{C}}_\alpha(\boldsymbol{U})$ and $\widehat{\mathcal{C}}(\boldsymbol{U})$ correspond to Lemma 10 with $\lambda = \infty$ and, in the latter case, $\alpha = 0$ as well. Note that the correction term contains a derivative of $\widehat{\pi}$ when $\alpha \neq 0$, which is problematic in our otherwise derivative-free method. In this paper we always choose $\alpha = 0$. If situations are identified where $\alpha \neq 0$ is preferred, one could neglect the $\widehat{\pi}$ derivatives (see Remark 11) or construct a particle approximation (e.g., based on the proximal operator such as in section 3 or on the technique in [57]). This is left for future work.

Lemma 10. *When* $\widehat{\mathbf{P}} = \widehat{\mathcal{C}}_{\alpha,\lambda\widehat{\mathcal{C}}(\boldsymbol{U})}(\boldsymbol{U}; U^i)$*, it holds that*

$$\begin{aligned}\nabla_{u^i}\cdot\widehat{\mathbf{P}} &= \omega^{ii}\,(d+1)U_c^{(i),i} + \frac{1}{\lambda}\boldsymbol{U}_c^{(i)}\left(\omega^i \odot \operatorname{diag}\big((\boldsymbol{U}_c^{(i)})^T\widehat{\mathcal{C}}^{-1}\boldsymbol{U}_c^{(i)}\big)\right)\\ &- \frac{1}{\lambda J}\widehat{\mathbf{P}}\widehat{\mathcal{C}}^{-1}\left(U_c^{(i),i}(U_c^{(i),i})^T + \widehat{\mathbf{P}}\right)\widehat{\mathcal{C}}^{-1}U_c^i\\ &+ \frac{1}{\lambda J}\boldsymbol{U}_c^{(i)}\left(\omega^i \odot \operatorname{diag}\big((\boldsymbol{U}_c^{(i)})^T\widehat{\mathcal{C}}^{-1}(\boldsymbol{U}-U^i\mathbf{1}^T)\big)\odot\big((U_c^i)^T\widehat{\mathcal{C}}^{-1}(\boldsymbol{U}-U^i\mathbf{1}^T)\big)^T\right)\\ &+ \alpha\omega^{ii}\left(U_c^{(i),i}(U_c^{(i),i})^T - \widehat{\mathbf{P}}\right)\frac{\nabla_{u^i}\widehat{\pi}(U^i)}{\widehat{\pi}(U^i)},\end{aligned} \tag{72}$$

where we denote by $\odot$ *the Hadamard (elementwise) product. Furthermore,* $\boldsymbol{U}_c \coloneqq \boldsymbol{U} - \bar{U}\mathbf{1}^T$ *and* $\boldsymbol{U}_c^{(i)} \coloneqq (\boldsymbol{U} - \bar{U}^{(i)}\mathbf{1}^T)$*. We write* U_c^i *for the* i*th column of* $\boldsymbol{U}_c$*,* $U_c^{(i),j}$ *for the* j*th column of* $\boldsymbol{U}_c^{(i)}$*,* $\widehat{\mathcal{C}}$ *as a shorthand for* $\widehat{\mathcal{C}}(\boldsymbol{U})$*, and* $\omega^i = (\omega^{i1},\dots,\omega^{iJ})^T$*.*

Proof. This is proven in Appendix A. □

Remark 11. *All terms in (72) except the second go down in magnitude as the number of particles* J *increases (through either* $1/J$ *or* $\omega^{ii} = w^{ii}/\sum_j w^{ij}$*, which depends on* J *due to the normalization). It may be computationally interesting to omit some of these terms in practice when* J *is large.*

Localized CBS algorithm. Combining the points discussed in this section leads to the localized CBS algorithm

$$U_{n+1}^i = U_n^i + \left[-\frac{\gamma}{\kappa}\left(U_n^i - \widehat{\mu}^{(\nu,i)}_{\beta,\kappa\widehat{\mathbf{P}}}(\boldsymbol{U}_n; U_n^i)\right) + \nabla_{u^i}\cdot\widehat{\mathbf{P}}\right]\Delta t + \sqrt{2\Delta t}\,\widehat{\mathbf{P}}^{1/2}\xi_n^i, \tag{73}$$

with $\xi_n^i \sim \mathcal{N}(0, I)$ i.i.d. random variables. Here, (β, κ) are free parameters and γ is chosen based on them (e.g., through Corollary 5). The choice of preconditioner $\widehat{\mathbf{P}} \coloneqq \widehat{\mathbf{P}}(\boldsymbol{U}_n; U_n^i)$ is free; examples are a constant matrix, an unweighted covariance as in ALDI, a weighted covariance as in CBS, or a localized weighted covariance. The correction term $\nabla_{u^i}\cdot\widehat{\mathbf{P}}$ is detailed in Lemma 10 and Remark 11; for many choices of $\widehat{\mathbf{P}}$, it is negligible or zero.

Remark 12 (Parallelizability)**.** *Earlier, we motivated particle methods such as localized CBS by their potential for parallelization. The algorithm (73) is indeed parallelizable across particles:* $\{V(U_n^i)\}_{i=1}^J$ *can be computed independently. To compute the interaction terms* $\widehat{\mu}^{(\nu,i)}_{\beta,\kappa\widehat{\mathbf{P}}}$ *and (potentially)* $\widehat{\mathbf{P}}$*, communication between particles is required. However, in practical applications,* V *is typically an expensive model (e.g., a PDE solver) whose evaluations are the main bottleneck.*

7 Numerical results

This section applies the localized CBS algorithm to a variety of problems, illustrating its properties. We compare our method to several existing alternatives: CBS (to illustrate that it is not suited for non-Gaussian problems) on the one hand and polarized CBS and localized ALDI on the other (to compare our algorithm to the two most similar options). Other algorithms are often challenging to compare to in a fair way: they either use gradients, are less parallelizable, or produce fewer samples per evaluation of V.

For both localized CBS and alternative methods, we will use the Euler–Maruyama time discretizations discussed in section 6. Unless mentioned otherwise, localized CBS uses timestep $\Delta t = 0.01$ and preconditioner $\mathcal{C}(\rho_t)$, an unweighted covariance. Except for Figure 1, we always use the γ value implied by the preconditioner through Corollary 5. The default random-batch parameter is $\nu = 1$, such that a particle interacts with all other particles. In addition, except in subsection 7.3, we sample by running the specified algorithm 16 times and combining the particle positions at all time steps within the final quarter of the time interval in each run. Samples are plotted with the `density` function from the Julia package `StatsPlots`, which performs kernel density estimation (KDE) to smooth the samples into an approximate distribution. When the dimension d is larger than one, we plot the marginal distribution in the first dimension unless specified otherwise. The initial distribution for all sampling methods is a zero-centered Gaussian $\mathcal{N}(0, \Sigma_0)$; by default, $\Sigma_0 = I/2$. Our code is located at

https://gitlab.kuleuven.be/numa/public/paper-code-lcbs.

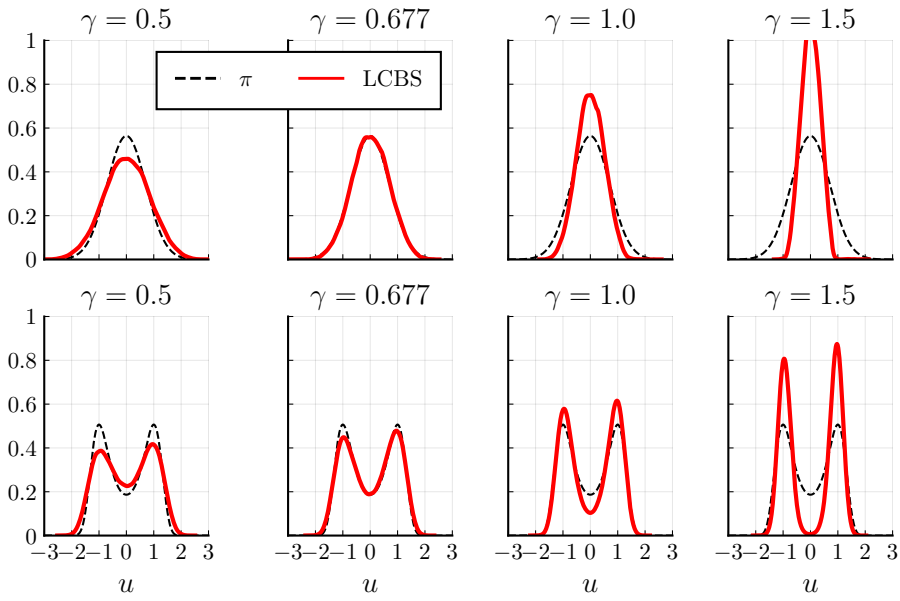


Figure 1. Localized CBS samples for a one-dimensional Gaussian target distribution (top row) and a one-dimensional multimodal target distribution (bottom row)

7.1 Gaussian distributions

We apply localized CBS to a Gaussian distribution, to verify the values of γ we obtained in subsection 4.3 and to assess the influence of an incorrect γ on the sampled distribution. We consider the one-dimensional potential function

$$V(u) = u^2 \tag{74}$$

and use $(\beta, \kappa) = (2.0, 0.01)$, $J = 500$ particles, and $N = 200$ timesteps. The top row of Figure 1 shows[7] the resulting samples with $\gamma \in \{0.5, \bar{\gamma}, 1.0, 1.5\}$, where $\bar{\gamma} \approx 0.677$ is the value given by (39)—since we use our default preconditioner $\mathcal{C}(\rho_t)$, this should ensure correct sampling for Gaussians. When $\gamma < \bar{\gamma}$, there is insufficient drift towards the weighted mean and the samples are too spread out. When $\gamma > \bar{\gamma}$, the opposite effect occurs. As predicted by the theory, the distribution is sampled correctly with $\gamma = \bar{\gamma}$.

7.2 Multimodal distributions

One of the design objectives of localized CBS is to handle non-Gaussian, potentially even multimodal, distributions. The first multimodal distribution that we consider corresponds to the potential function

$$V(u) = \|u^2 - 1\|_2^2, \tag{75}$$

where u^2 applies the exponent elementwise. The resulting distribution π is bimodal when the dimension d is 1 and a tensor product of bimodal distributions otherwise.

We first apply localized CBS with the same parameters from subsection 7.1 to this potential in dimension $d = 1$, with the results shown in the bottom row of Figure 1. Even though π is not sampled exactly, which is expected due to the small β value, using the $\gamma = \bar{\gamma}$ that is correct for Gaussian distributions seems to be a good heuristic even in this non-Gaussian case.

For this potential with $d \in \{1, 10\}$, we compare localized CBS (with our default parameters, along with $\beta = 10$ and $\kappa = 0.01$ when $d = 1$ or $\kappa = 0.03$ when $d = 10$) to CBS (with $\alpha = 10$ and $\Delta t = 0.01$), polarized CBS (with $\alpha = 10$, $\lambda = 0.005$ when $d = 1$ or $\lambda = 0.1$ when $d = 10$, $D = I$, and $\Delta t = 0.01$), and localized ALDI (with $\lambda = 0.02$ when $d = 1$ or $\lambda = 0.4$ when $d = 10$, $D = I$, and $\Delta t = 0.05$). These parameters were chosen to the best of our abilities to give good results for each method for the chosen values of $J = 200$ particles and $N = 1000$ timesteps. For localized CBS, we additionally compare $\nu = 1$ to $\nu = 0.5$. Figure 2 shows the resulting samples. Localized CBS accurately finds the distribution when $d = 1$; when $d = 10$, we get a good approximation only with random-batch interaction turned on ($\nu = 0.5$). Localized ALDI and polarized CBS each give relatively close approximations, but are unable to exactly fit π when $d = 10$. We also noted that for

[7] In the figures of this section, we abbreviate localized CBS, polarized CBS, and localized ALDI as *LCBS*, *PCBS*, and *LALDI*, respectively.

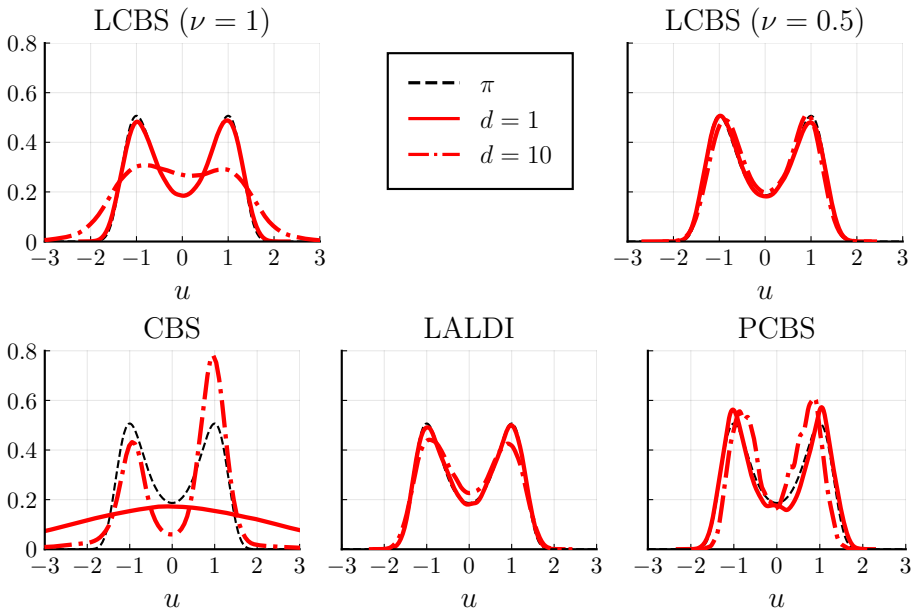


Figure 2. Marginalized samples for a 1- or 10-dimensional multimodal target distribution. Recall that we aggregate the samples from 16 runs, which explains the non-Gaussian CBS samples.

these two methods, the performance was significantly more sensitive to the choice of parameters than for localized CBS. Classical CBS finds the multiple modes only when $d = 10$, and fails to approximate the density in between.[8]

Secondly, we study affine-invariance of the three best performing algorithms: localized CBS, polarized CBS, and localized ALDI. To this end, consider

$$V(u) = \|(\sqrt{\Lambda}\, u)^2 - 1\|_2^2, \tag{76}$$

where $d = 2$ and the square again applies elementwise. The matrix $\Lambda = \mathrm{diag}(1, 10^4)$ changes the scale of the second parameter. Suppose that we may not know this scaling before solving the problem. Without knowing the form of the potential (76), we assume that the relative scaling of the two parameters is given by a matrix $\tilde{\Lambda}$ that can either be the correct $\tilde{\Lambda} = \Lambda$ or the incorrect $\tilde{\Lambda} = I$. We use this assumption to choose an initial covariance $\Sigma_0 = \tilde{\Lambda}^{-1}/2$. In addition,

(i) in polarized CBS and localized ALDI, we set $D = \tilde{\Lambda}^{-1}$;

(ii) in our localized CBS algorithm, we simply use the preconditioner $\mathcal{C}(\rho_t)$, which does not need *a priori* scaling information.

We use $(\beta, \kappa) = (10, 0.03)$ for localized CBS, $\lambda = 0.02$ and $\Delta t = 0.05$ for localized ALDI, and $\lambda = 0.005$ for polarized CBS. Figure 3 shows the samples by these algorithms for both $\tilde{\Lambda}$ cases. Polarized CBS and localized ALDI perform as expected when the scaling is guessed correctly, but generate clearly incorrect samples when it is not. This is because they depend on $\tilde{\Lambda}$ not only for their initial condition, but also for weighting the distance between particles throughout the algorithm. Localized CBS, on the other hand, proves itself able to recover from this poor initial guess and produces good—and almost identical—samples for both $\tilde{\Lambda}$s.

Now consider a different potential function with $d = 1$:

$$V(u) = 2(u\exp(u))^4 - 4(u\exp(u))^2 - 2(u/3)^5 + 2. \tag{77}$$

This again produces a multimodal distribution, now with one wide and one narrow peak. We first compare preconditioners for localized CBS with parameters $(\beta, \kappa) = (10, 0.02)$, $N = 1000$ timesteps, and $J = 200$ particles. We use preconditioner $\mathcal{C}_{\alpha,\alpha\lambda\mathcal{C}(\rho_t)}(\rho_t; U_t)$ for $\lambda \in \{0.1, 0.5, \infty\}$ and $\alpha \to 0$; when $\lambda = \infty$, this equals the unweighted covariance $\mathcal{C}(\rho_t)$. The initial distribution has covariance $\Sigma_0 \in \{I/2, 2I\}$. Resulting samples are shown in Figure 4. When $\lambda = \infty$, the global nature of the preconditioner means that the presence of the wide peak causes the narrow one to be

[8] Note that the particles in CBS always follow a Gaussian distribution; in our experiment, CBS captures multiple modes as the results from 16 runs (which may find different modes) are aggregated.

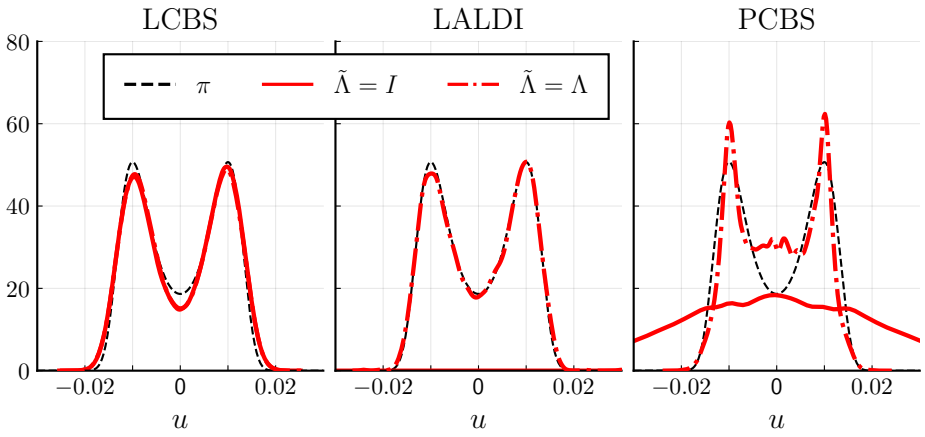


Figure 3. Marginalized samples for the poorly scaled distribution with potential (76). In the left figure, the two red lines are visually indistinguishable; in the middle figure, one line coincides with the horizontal axis.

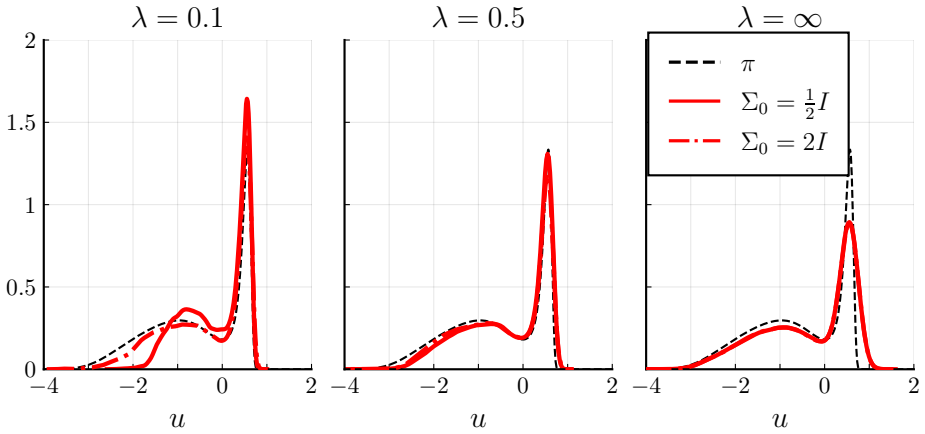


Figure 4. Localized CBS samples for (77) with various preconditioners and initial distributions

estimated poorly. When $\lambda = 0.1$, on the other hand, localized CBS struggles to find the full distribution when the initial distribution is too narrow. A moderate $\lambda = 0.5$ seems to combine the benefits of both extremes. Note that, as discussed in subsection 4.2, $\lambda = \infty$ has the computational advantage that only one covariance needs to be computed instead of J different ones.

Next, we illustrate the importance of the correction term $\nabla_u \cdot \mathbf{P}_{U,t}$ in the localized CBS dynamics (31) on the sampling problem (77). We consider the preconditioner $\mathcal{C}_{\alpha,\alpha\lambda\mathcal{C}(\rho_t)}(\rho_t; U_t)$ with $\lambda = 0.5$ and $\alpha \to 0$ and compare the normal localized CBS algorithm to a modified version where the correction term is omitted. We also apply polarized CBS, which does not have a similar correction term, with $\alpha = 10$ and $\lambda \in \{0.001, 0.002, 0.005\}$. We use $N = 1000$ timesteps, $J = 200$ particles, and an initial covariance $\Sigma_0 = 2I$ for all samplers. Figure 5 illustrates that only localized CBS with the correction term manages to accurately retrieve π. As before, polarized CBS only samples from a distribution approximating π.

7.3 A tent-shaped distribution

In the next example we study convergence of localized CBS for the one-dimensional target distribution corresponding to

$$\pi(u) = \widehat{\pi}(u) = \max(0, 1 - |u|), \tag{78}$$

a tent-shaped distribution with finite support. Our goal is not to compare to other methods, but rather to gain insight into the choice of (β, κ). We will answer the question why, even though choosing $\beta \to \infty$ and $\kappa \to 0$ seems desirable, it may still make sense to choose moderate values for both parameters in practice. Figure 6 shows the Wasserstein-2 distance between the exact distribution π and the empirical distribution of localized CBS ensembles with a given number J of particles. This is compared for several (β, κ) combinations. In this subsection, we find the

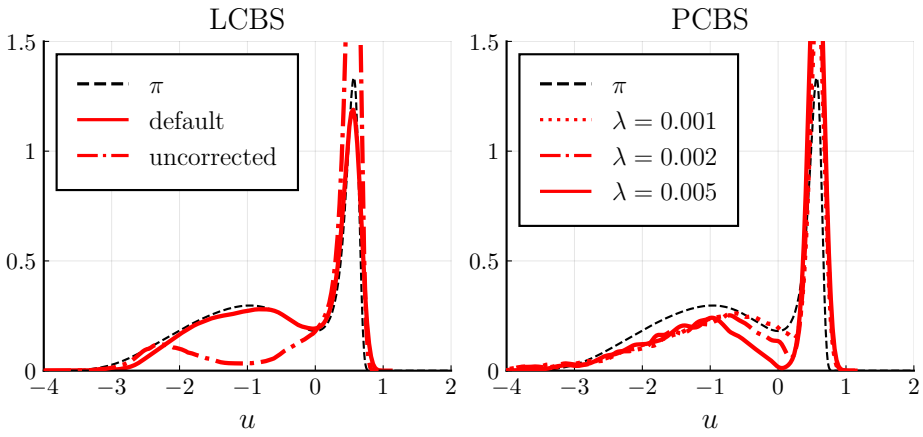


Figure 5. Localized and polarized CBS samples for (77)

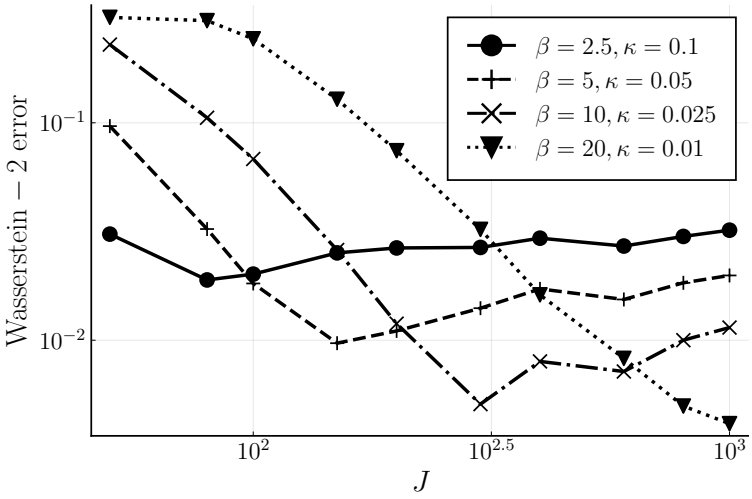


Figure 6. Wasserstein-2 error between the localized CBS samples and the target distribution (78)

empirical distribution for any (β, κ, J) by simulating localized CBS 480 times and, for each simulation, using the positions of 50 randomly chosen particles at the final time step as samples.

After a certain value, increasing J no longer improves accuracy. This is caused by the discrepancy between the mean-field distribution and π, which decreases with large β and small κ. However, these more accurate mean-field dynamics become increasingly difficult to approximate with a finite number of particles, as fewer particles get to contribute substantially to the weighted means. Hence, when large numbers of particles are computationally feasible, a high accuracy can be obtained by choosing $\beta \gg 0$ and $\kappa \approx 0$. When relatively few particles can be used, on the other hand, the accuracy is actually highest for more moderate β and κ choices.

Remark 13. *As J keeps increasing, the errors in Figure 6 become slightly larger again. This may be caused by a combination of two factors: (i) the Moreau envelope in section 3 approximates π by a distribution with longer tails and (ii) when J is small, localized CBS tends to undersample the tails. While the first factor is a source of error present for all J, the second factor may partially offset it for small J, resulting in an error smaller than that for $J \to \infty$.*

7.4 A Darcy flow inverse problem

Our final example is meant to assess how localized CBS scales to higher dimensions. We choose a more complex model than in earlier experiments, but the resulting posterior is near-Gaussian and does not require localization. We mainly use this example since it is slightly more realistic than the previous ones, and its dimension can be varied straightforwardly. A near-Gaussian distribution is a best-case scenario: scaling well here is likely a prerequisite for scaling well for non-Gaussian distributions.

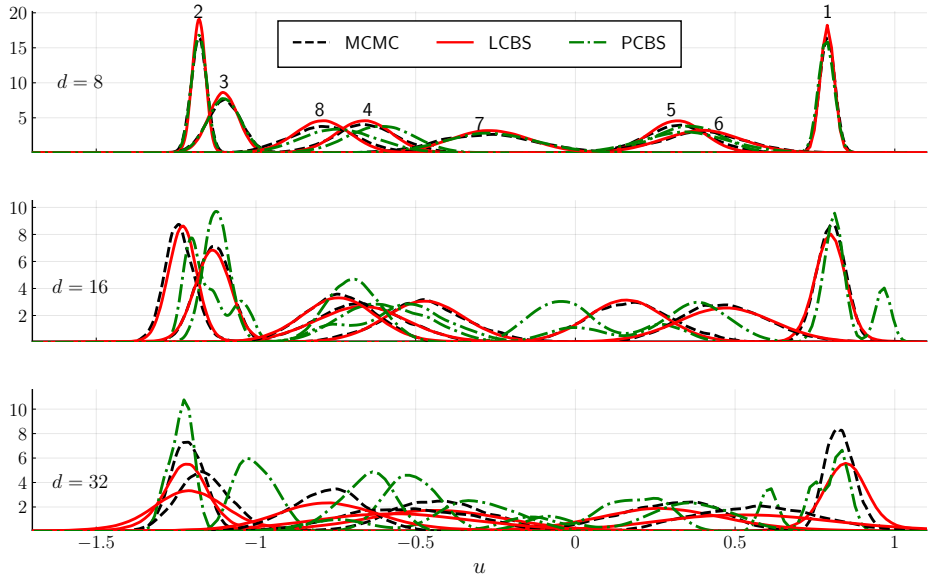


Figure 7. MCMC, localized CBS, and polarized CBS samples of the first 8 KL terms in the d-dimensional Darcy flow problem, with $d \in \{8, 16, 32\}$. The $d = 8$ plot labels the modes in order of decreasing eigenvalue λ_k.

Consider the spatial domain $\Omega := [0,1]^2$. The Darcy flow inverse problem is to recover parameters u of the permeability field $a(x,u)$ from noisy measurements of the pressure field $p(x)$ satisfying

$$-\nabla \cdot \big(a(x,u)\nabla p(x)\big) = f(x), \qquad x \in \Omega, \tag{79a}$$

$$p(x) = 0, \qquad x \in \partial\Omega. \tag{79b}$$

Here, $f(x) = c$ is a fluid source that we set to a constant value.

The permeability $a(x,u)$ is modeled as a log-normal random field with mean zero and covariance operator $(-\Delta + \tau^2)^{-s}$ of $\log a$, where $-\Delta$ represents the Laplacian with homogeneous Neumann boundary conditions. This corresponds to a standard normal prior on the parameters u_k in the Karhunen–Loève (KL) expansion

$$\log a(x,u) = \sum\nolimits_{k \in \mathbb{N}^2 \setminus \{(0,0)\}} u_k \sqrt{\lambda_k}\, \phi_k(x), \tag{80}$$

with eigenpairs $\lambda_k = (\pi^2 \|k\|_2^2 + \tau^2)^{-s}$ and $\phi_k = c_k \cos(\pi k_1 x_1)\cos(\pi k_2 x_2)$, where $c_k = \sqrt{2}$ if $k_1 k_2 = 0$ and $c_k = 2$ otherwise (see, e.g., [30]). We choose parameters $(\tau, s) = (3, 2)$ and truncate the expansion (80) after the d terms with largest eigenvalues λ_k. The coefficients u_k of these terms are the unknown parameters u with prior $\pi_{\text{prior}} = \mathcal{N}(0, \Gamma_0)$, with $\Gamma_0 = I$, in the Bayesian inverse problem (2). The data y consists of measurements of p, taken at 49 equispaced points and subject to noise known to be distributed according to $\pi_{\text{noise}} = \mathcal{N}(0, \Gamma)$ with $\Gamma = 10^{-4} I$. The model G solves (79) with central finite differences on a grid with stepsize $h = 2^{-5}$. An observation is generated by sampling $\{u_k\}_k$ from the prior, solving (79) with stepsize $h = 2^{-9}$, and perturbing that solution with a random sample from π_{noise}. As described in subsection 1.1, this Bayesian inverse problem corresponds to (1) with

$$V(u) = \frac{1}{2}\|y - G(u)\|_\Gamma^2 + \frac{1}{2}\|u\|_{\Gamma_0}^2. \tag{81}$$

Since the resulting posterior is unknown, we need to compute a sensible reference solution. For this, we use random-walk MCMC: we run 32 chains of 10^5 steps each, of which the first quarter is discarded as burn-in. For localized CBS, we tuned parameters to obtain good results for $d = 8$: we use $N = 400$ timesteps of length $\Delta t = 0.03$, $J = 1000$ particles, a large initial covariance $\Sigma_0 = 9I$ (similarly to [4]), parameters $(\beta, \kappa) = (3, 0.2)$, and our default preconditioner $\mathcal{C}(\rho_t)$.

We then test whether these parameters can be scaled up to higher dimensions. Figure 7 compares the reference samples to the localized CBS samples. It shows the marginalized distributions for the first 8 modes, for the number of dimensions $d \in \{8, 16, 32\}$. The modes are successfully located for each d. When $d \in \{8, 16\}$, the uncertainty around the mode is captured

well; when $d = 32$, it is slightly overestimated. This matches the argument in Remark 9 that scaling localized CBS to high dimensions may prove difficult, although further research is needed.

For comparison, we conduct the same experiment with polarized CBS[9], using $J = 1000$ particles over $N = 400$ timesteps of length $\Delta t = 0.02$. We chose the parameters $D = I$ and $\lambda = 0.25$ to work well for $d = 8$. The results, shown in Figure 7 as well, display a similar trend as for localized CBS (but more pronounced): in higher dimensions, these parameters no longer succeed in capturing the target distribution. Increasing the number of particles with the dimension may help, but becomes intractable quickly as d increases. As for localized CBS, further research is needed to determine whether polarized CBS can be scaled up efficiently to higher dimensions.

8 Conclusions

We have derived localized consensus-based sampling (localized CBS) as an alternative formulation of consensus-based sampling (CBS) obtained from ensemble-preconditioned Langevin dynamics through a sequence of approximations involving the Moreau envelope and a proximal operator. By retaining only the steps that remain valid beyond the Gaussian setting, we obtained a variant closely related to polarized CBS but endowed with full affine invariance. In the mean-field limit, localized CBS is exact for Gaussian targets.

Numerical experiments demonstrate that localized CBS achieves comparable or improved performance relative to polarized CBS and localized ALDI, particularly in non-Gaussian and moderately multimodal problems. The method is fully gradient-free and easily parallelizable.

Several open research directions remain. First, the theoretical understanding of localized CBS—including its non-Gaussian convergence, well-posedness, and mean-field limit—is still incomplete. As a first step, the ODEs (34) may allow deriving an explicit convergence rate and comparing different (α, λ) choices in the Gaussian case. Second, as discussed in Remark 9, the efficiency of localized CBS in high-dimensional settings has not been analyzed and likely depends on problem structure and localization scale. Third, localized CBS may be suitable for sampling from non-differentiable potentials, similarly to other prox-based samplers [37, 48, 61]. Finally, adapting localized CBS to exploit additional problem information—for example, hybrid strategies that use gradient evaluations only for selected particles—may help extend its applicability to more complex sampling tasks.

Acknowledgments

We are grateful to Kathrin Hellmuth for her comments on an earlier draft of this paper. We also wish to thank the anonymous referees for providing valuable feedback and suggestions, which greatly improved this paper.

Funding

Our work was partially funded by the Research Foundation – Flanders (FWO) (grants 1169725N, G081222N, G033822N, and G0A0920N) and the KU Leuven Research Council (C1 projects C14/23/098 and C14/24/103).

A Various technical proofs

Proof of (35) and (36). Consider means (m, m_t) and covariances (Σ, Σ_t) such that $\pi \sim \mathcal{N}(m, \Sigma)$ and $\rho_t \sim \mathcal{N}(m_t, \Sigma_t)$. We prove (36) similarly to [3], by using the formula found in, e.g., [49, §8.1.8] to compute $\mu_{\beta,\kappa\mathbf{P}_{*,t}}(\rho_t; U_t)$ as

$$\begin{aligned}&\frac{\int v \exp(-\frac{\beta}{2\kappa}\|v - U_t\|^2_{\mathbf{P}_{*,t}}) \exp(-\frac{\beta}{2}\|v - m\|^2_{\Sigma}) \exp(-\frac{1}{2}\|v - m_t\|^2_{\Sigma_t})\,\mathrm{d}v}{\int \exp(-\frac{\beta}{2\kappa}\|v - U_t\|^2_{\mathbf{P}_{*,t}}) \exp(-\frac{\beta}{2}\|v - m\|^2_{\Sigma}) \exp(-\frac{1}{2}\|v - m_t\|^2_{\Sigma_t})\,\mathrm{d}v} \\ &= \Big(\beta\Sigma^{-1} + \Sigma_t^{-1} + \frac{\beta}{\kappa}\mathbf{P}_{*,t}^{-1}\Big)^{-1}\Big(\beta\Sigma^{-1}m + \Sigma_t^{-1}m_t + \frac{\beta}{\kappa}\mathbf{P}_{*,t}^{-1}U_t\Big).\end{aligned}$$

Equations (35a) and (35b) follow directly from the definitions. Equations (35c) and (35d) are derived similarly to $\mu_{\beta,\kappa\mathbf{P}_{*,t}}(\rho_t; U_t)$ above. □

Proof of Lemma 10. In [51, Lemma 3.11], it is shown that with general weights,

$$\nabla_{u^i} \cdot \widehat{\mathbf{P}} = \omega^{ii}\,(d+1)(U^i - \bar{U}^{(i)}) + \sum\nolimits_{j=1}^{J}\Big(U_c^{(i),j}(U_c^{(i),j})^T - \bar{U}^{(i)}(\bar{U}^{(i)})^T\Big)\nabla_{u^i}\omega^{ij}. \tag{82}$$

[9] Localized ALDI diverges with this initial distribution.

With our weight function, we find that

$$\begin{aligned}
\nabla_{u^i}\omega^{ij} &= \nabla_{u^i}\Big(w^{ij}\Big/\sum\nolimits_{k=1}^J w^{ik}\Big) = \Big(\nabla_{u^i}w^{ij} - \omega^{ij}\nabla_{u^i}\Big(\sum\nolimits_{k=1}^J w^{ik}\Big)\Big)\Big/\sum\nolimits_{k=1}^J w^{ik} \\
&= \Big[-\frac{1}{2\lambda}\Big(2\widehat{\mathcal{C}}^{-1}(U^i-U^j) - X^{ij}\Big) + \delta^{ij}\alpha\frac{\nabla_{u^i}\widehat{\pi}(U^i)}{\widehat{\pi}(U^i)}\Big]\omega^{ij} \\
&\quad + \omega^{ij}\sum\nolimits_{k=1}^J\Big[\frac{1}{2\lambda}\Big(2\widehat{\mathcal{C}}^{-1}(U^i-U^k) - X^{ik}\Big) - \delta^{ik}\alpha\frac{\nabla_{u^i}\widehat{\pi}(U^i)}{\widehat{\pi}(U^i)}\Big]\omega^{ik} \\
&= \frac{\omega^{ij}}{\lambda}\Big(\big[\widehat{\mathcal{C}}^{-1}(U^j-U^i) + \frac{X^{ij}}{2}\big] + \big[\widehat{\mathcal{C}}^{-1}U_c^{(i),i} - \sum\nolimits_{k=1}^J\omega^{ik}\frac{X^{ik}}{2}\big]\Big) + (\delta^{ij}-\omega^{ij})\alpha\omega^{ii}\frac{\nabla_{u^i}\widehat{\pi}(U^i)}{\widehat{\pi}(U^i)} \\
&= \frac{\omega^{ij}}{\lambda}\Big(\widehat{\mathcal{C}}^{-1}U_c^{(i),j} - \sum\nolimits_{k=1}^J\omega^{ik}\frac{X^{ik}}{2} + \frac{X^{ij}}{2}\Big) + (\delta^{ij}-\omega^{ij})\alpha\omega^{ii}\frac{\nabla_{u^i}\widehat{\pi}(U^i)}{\widehat{\pi}(U^i)},
\end{aligned}$$

with $(X^{ik})_\ell = (U^i-U^k)^T\widehat{\mathcal{C}}^{-1}\left(\partial_{u^i_\ell}\widehat{\mathcal{C}}\right)\widehat{\mathcal{C}}^{-1}(U^i-U^k)$. We have $\partial_{u^i_\ell}\widehat{\mathcal{C}} = J^{-1}\big(U_c^i e_\ell^T + e_\ell(U_c^i)^T\big)$ and

$$X^{ij}/2 = J^{-1}\big((U_c^i)^T\widehat{\mathcal{C}}^{-1}(U^i-U^j)\big)\widehat{\mathcal{C}}^{-1}(U^i-U^j), \tag{83a}$$

$$\sum\nolimits_{k=1}^J X^{ik}\omega^{ik}/2 = J^{-1}\widehat{\mathcal{C}}^{-1}\big(U_c^{(i),i}(U_c^{(i),i})^T + \widehat{\mathbf{P}}\big)\widehat{\mathcal{C}}^{-1}U_c^i. \tag{83b}$$

We split $\nabla_{u^i}\cdot\widehat{\mathbf{P}} = A + (B_1+B_2+B_3-C_1-C_2-C_3)/\lambda + D$ with

$$\begin{aligned}
A &\coloneqq \omega^{ii}(d+1)(U^i-\bar{U}^{(i)}), \\
B_1 &\coloneqq \sum\nolimits_{j=1}^J U_c^{(i),j}(U_c^{(i),j})^T\omega^{ij}\big[\widehat{\mathcal{C}}^{-1}U_c^{(i),j}\big] = \boldsymbol{U}_c^{(i)}\big(\omega^i\odot\operatorname{diag}((\boldsymbol{U}_c^{(i)})^T\widehat{\mathcal{C}}^{-1}\boldsymbol{U}_c^{(i)})\big), \\
B_2 &\coloneqq \sum\nolimits_{j=1}^J U_c^{(i),j}(U_c^{(i),j})^T\omega^{ij}\big[-J^{-1}\widehat{\mathcal{C}}^{-1}\big(U_c^{(i),i}(U_c^{(i),i})^T + \widehat{\mathbf{P}}\big)\widehat{\mathcal{C}}^{-1}U_c^i\big] \\
&= -J^{-1}\widehat{\mathbf{P}}\widehat{\mathcal{C}}^{-1}(U_c^{(i),i}(U_c^{(i),i})^T + \widehat{\mathbf{P}})\widehat{\mathcal{C}}^{-1}U_c^i, \\
B_3 &\coloneqq \sum\nolimits_{j=1}^J U_c^{(i),j}(U_c^{(i),j})^T\omega^{ij}\big[J^{-1}\big((U_c^i)^T\widehat{\mathcal{C}}^{-1}(U^i-U^j)\big)\widehat{\mathcal{C}}^{-1}(U^i-U^j)\big] \\
&= J^{-1}\boldsymbol{U}_c^{(i)}(\omega^i\odot\operatorname{diag}((\boldsymbol{U}_c^{(i)})^T\widehat{\mathcal{C}}^{-1}(\boldsymbol{U}-U^i\mathbf{1}^T))\odot((U_c^i)^T\widehat{\mathcal{C}}^{-1}(\boldsymbol{U}-U^i\mathbf{1}^T))^T), \\
C_1 &\coloneqq \sum\nolimits_{j=1}^J \bar{U}^{(i)}(\bar{U}^{(i)})^T\omega^{ij}\big[\widehat{\mathcal{C}}^{-1}U_c^{(i),j}\big] = \bar{U}^{(i)}(\bar{U}^{(i)})^T\widehat{\mathcal{C}}^{-1}\Big(\sum\nolimits_{j=1}^J\omega^{ij}U_c^{(i),j}\Big) = 0, \\
C_2 &\coloneqq \sum\nolimits_{j=1}^J \bar{U}^{(i)}(\bar{U}^{(i)})^T\omega^{ij}\Big[-\sum\nolimits_{k=1}^J\omega^{ik}\frac{X^{ik}}{2}\Big] = -\bar{U}^{(i)}(\bar{U}^{(i)})^T\sum\nolimits_{k=1}^J\omega^{ik}\frac{X^{ik}}{2}, \\
C_3 &\coloneqq \sum\nolimits_{j=1}^J \bar{U}^{(i)}(\bar{U}^{(i)})^T\omega^{ij}\Big[\frac{X^{ij}}{2}\Big] = \bar{U}^{(i)}(\bar{U}^{(i)})^T\sum\nolimits_{j=1}^J\omega^{ij}\frac{X^{ij}}{2} = -C_2, \\
D &\coloneqq \sum\nolimits_{j=1}^J\Big(U_c^{(i),j}(U_c^{(i),j})^T - \bar{U}^{(i)}(\bar{U}^{(i)})^T\Big)(\delta^{ij}-\omega^{ij})\alpha\omega^{ii}\frac{\nabla_{u^i}\widehat{\pi}(U^i)}{\widehat{\pi}(U^i)} \\
&= \alpha\omega^{ii}\Big(U_c^{(i),i}(U_c^{(i),i})^T - \widehat{\mathbf{P}}\Big)\frac{\nabla_{u^i}\widehat{\pi}(U^i)}{\widehat{\pi}(U^i)}.
\end{aligned}$$

□

References

[1] Julian Besag. "Discussion of the Paper by Grenander and Miller". In: *J. R. Stat. Soc. B* 56.4 (1994), pp. 591–592.

[2] Steve Brooks et al., eds. *Handbook of Markov Chain Monte Carlo*. New York: Chapman and Hall/CRC, 2011.

[3] Leon Bungert, Tim Roith, and Philipp Wacker. "Polarized Consensus-Based Dynamics for Optimization and Sampling". In: *Math. Program.* (2024).

[4] J. A. Carrillo et al. "Consensus-based Sampling". In: *Stud. Appl. Math.* 148.3 (2022), pp. 1069–1140.

[5] José A. Carrillo et al. "An analytical framework for consensus-based global optimization method". In: *Math. Models Methods Appl. Sci.* 28.06 (2018), pp. 1037–1066.

[6] Neil K. Chada, Yuming Chen, and Daniel Sanz-Alonso. "Iterative Ensemble Kalman Methods: A Unified Perspective with Some New Variants". In: *Found. Data Sci.* 3.3 (2021), pp. 331–369.

[7] Patrick L. Combettes and Băng C. Vũ. "Variable metric forward–backward splitting with applications to monotone inclusions in duality". In: *Optimization* 63.9 (2014), pp. 1289–1318.

[8] S. L. Cotter et al. "MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster". In: *Stat. Sci.* 28.3 (2013), pp. 424–446.

[9] Tiangang Cui, Kody J. H. Law, and Youssef M. Marzouk. "Dimension-Independent Likelihood-Informed MCMC". In: *J. Comput. Phys.* 304 (2016), pp. 109–137.

[10] M. Dashti et al. "MAP Estimators and Their Consistency in Bayesian Nonparametric Inverse Problems". In: *Inverse Probl.* 29.9 (2013), p. 095017.

[11] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. "Sequential Monte Carlo Samplers". In: *J. R. Stat. Soc. B* 68.3 (2006), pp. 411–436.

[12] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Vol. 38. Stochastic Modelling and Applied Probability. Berlin, Heidelberg: Springer, 2010.

[13] Zhiyan Ding and Qin Li. "Constrained Ensemble Langevin Monte Carlo". In: *Found. Data Sci.* 4.1 (2022), p. 37.

[14] Zhiyan Ding and Qin Li. "Ensemble Kalman Inversion: Mean-Field Limit and Convergence Analysis". In: *Stat. Comput.* 31.1 (2021), p. 9.

[15] Zhiyan Ding and Qin Li. "Ensemble Kalman Sampler: Mean-field Limit and Convergence Analysis". In: *SIAM J. Math. Anal.* 53.2 (2021), pp. 1546–1578.

[16] T. J. Dodwell et al. "A Hierarchical Multilevel Markov Chain Monte Carlo Algorithm with Applications to Uncertainty Quantification in Subsurface Flow". In: *SIAM/ASA J. Uncertainty Quantif.* 3.1 (2015), pp. 1075–1108.

[17] Oliver R. A. Dunbar et al. "Ensemble Inference Methods for Models With Noisy and Expensive Likelihoods". In: *SIAM J. Appl. Dyn. Syst.* 21.2 (2022), pp. 1539–1572.

[18] Alain Durmus, Éric Moulines, and Marcelo Pereyra. "Efficient Bayesian Computation by Proximal Markov Chain Monte Carlo: When Langevin Meets Moreau". In: *SIAM J. Imag. Sci.* 11.1 (2018), pp. 473–506.

[19] Geir Evensen. "Sequential Data Assimilation with a Nonlinear Quasi-Geostrophic Model Using Monte Carlo Methods to Forecast Error Statistics". In: *J. Geophys. Res. Oceans* 99.C5 (1994), pp. 10143–10162.

[20] Massimo Fornasier, Timo Klock, and Konstantin Riedl. "Consensus-Based Optimization Methods Converge Globally". In: *SIAM J. Optim.* 34.3 (2024), pp. 2973–3004.

[21] Alfredo Garbuno-Inigo, Nikolas Nüsken, and Sebastian Reich. "Affine Invariant Interacting Langevin Dynamics for Bayesian Inference". In: *SIAM J. Appl. Dyn. Syst.* 19.3 (2020), pp. 1633–1658.

[22] Alfredo Garbuno-Inigo et al. "Interacting Langevin Diffusions: Gradient Structure and Ensemble Kalman Sampler". In: *SIAM J. Appl. Dyn. Syst.* 19.1 (2020), pp. 412–441.

[23] Nicolai Jurek Gerber, Franca Hoffmann, and Urbain Vaes. "Mean-Field Limits for Consensus-Based Optimization and Sampling". In: *arXiv:2312.07373* (2024).

[24] Omar Al-Ghattas and Daniel Sanz-Alonso. "Non-Asymptotic Analysis of Ensemble Kalman Updates: Effective Dimension and Localization". In: *Inf. Inference J. IMA* 13.1 (2024), iaad043.

[25] Jonathan Goodman and Jonathan Weare. "Ensemble Samplers with Affine Invariance". In: *Commun. Appl. Math. Comput. Sci.* 5.1 (2010), pp. 65–80.

[26] Philip Greengard. "An Ensemblized Metropolized Langevin Sampler". In: *Master's thesis, Courant Institute, New York University* (2015).

[27] W. K. Hastings. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". In: *Biometrika* 57.1 (1970), pp. 97–109.

[28] Lukas Herrmann, Magdalena Keller, and Christoph Schwab. "Quasi-Monte Carlo Bayesian Estimation under Besov Priors in Elliptic Inverse Problems". In: *Math. Comput.* 90.330 (2021), pp. 1831–1860.

[29] P. L. Houtekamer and Herschel L. Mitchell. "Data Assimilation Using an Ensemble Kalman Filter Technique". In: *Mon. Weather Rev.* 126.3 (1998), pp. 796–811.

[30] Daniel Zhengyu Huang, Tapio Schneider, and Andrew M. Stuart. "Iterated Kalman Methodology for Inverse Problems". In: *J. Comput. Phys.* 463 (2022), p. 111262.

[31] Marco A. Iglesias, Kody J. H. Law, and Andrew M. Stuart. "Ensemble Kalman Methods for Inverse Problems". In: *Inverse Probl.* 29.4 (2013), p. 045001.

[32] Shi Jin, Lei Li, and Jian-Guo Liu. "Random Batch Methods (RBM) for Interacting Particle Systems". In: *J. Comput. Phys.* 400 (2020), p. 108877.

[33] Jari P. Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*. Ed. by S.S. Antman, J.E. Marsden, and L. Sirovich. Vol. 160. Applied Mathematical Sciences. New York, NY: Springer, 2005.

[34] Marvin Koß, Simon Weissmann, and Jakob Zech. "On the Mean Field Limit of Consensus Based Methods". In: *arXiv:2409.03518* (2024).

[35] Nikola B. Kovachki and Andrew M. Stuart. "Ensemble Kalman Inversion: A Derivative-Free Technique for Machine Learning Tasks". In: *Inverse Probl.* 35.9 (2019), p. 095005.

[36] Rémi Laumont et al. "Bayesian Imaging Using Plug & Play Priors: When Langevin Meets Tweedie". In: *SIAM J. Imag. Sci.* 15.2 (2022), pp. 701–737.

[37] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. "Structured Logconcave Sampling with a Restricted Gaussian Oracle". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. PMLR, 2021, pp. 2993–3050.

[38] Benedict Leimkuhler, Charles Matthews, and Jonathan Weare. "Ensemble Preconditioning for Markov Chain Monte Carlo Simulation". In: *Stat. Comput.* 28.2 (2018), pp. 277–290.

[39] M. B. Lykkegaard et al. "Multilevel Delayed Acceptance MCMC". In: *SIAM/ASA J. Uncertainty Quantif.* 11.1 (2023), pp. 1–30.

[40] Yi-An Ma, Tianqi Chen, and Emily B. Fox. "A Complete Recipe for Stochastic Gradient MCMC". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, pp. 2917–2925.

[41] Nicholas Metropolis et al. "Equation of State Calculations by Fast Computing Machines". In: *J. Chem. Phys.* 21.6 (1953), pp. 1087–1092.

[42] Peter David Miller. *Applied Asymptotic Analysis*. Vol. 75. Graduate studies in mathematics. American Mathematical Soc., 2006.

[43] M. Morzfeld, X. T. Tong, and Y. M. Marzouk. "Localization for MCMC: Sampling High-Dimensional Posterior Distributions with Local Structure". In: *J. Comput. Phys.* 380 (2019), pp. 1–28.

[44] Nikolas Nüsken and Sebastian Reich. "Note on Interacting Langevin Diffusions: Gradient Structure and Ensemble Kalman Sampler by Garbuno-Inigo, Hoffmann, Li and Stuart". In: *arXiv:1908.10890* (2019).

[45] Stanley Osher, Howard Heaton, and Samy Wu Fung. "A Hamilton-Jacobi-based Proximal Operator". In: *Proc. Natl. Acad. Sci. U.S.A.* 120.14 (2023), e2220469120.

[46] Edward Ott et al. "A Local Ensemble Kalman Filter for Atmospheric Data Assimilation". In: *Tellus A: Dyn. Meteorol. Oceanogr.* 56.5 (2004), pp. 415–428.

[47] G. A. Pavliotis, A. M. Stuart, and U. Vaes. "Derivative-Free Bayesian Inversion Using Multiscale Dynamics". In: *SIAM J. Appl. Dyn. Syst.* 21.1 (2022), pp. 284–326.

[48] Marcelo Pereyra. "Proximal Markov Chain Monte Carlo Algorithms". In: *Stat. Comput.* 26.4 (2016), pp. 745–760.

[49] Kaare Brandt Petersen and M. S. Pedersen. "The Matrix Cookbook". In: 2006.

[50] René Pinnau et al. "A Consensus-Based Model for Global Optimization and Its Mean-Field Limit". In: *Math. Models Methods Appl. Sci.* 27.01 (2017), pp. 183–204.

[51] Sebastian Reich and Simon Weissmann. "Fokker–Planck Particle Systems for Bayesian Inference: Computational Approaches". In: *SIAM/ASA J. Uncertainty Quantif.* 9.2 (2021), pp. 446–482.

[52] Konstantin Riedl et al. "Gradient Is All You Need?" In: *arXiv:2306.09778* (2023).

[53] Gareth O. Roberts and Richard L. Tweedie. "Exponential Convergence of Langevin Distributions and Their Discrete Approximations". In: *Bernoulli* 2.4 (1996), pp. 341–363.

[54] R. Tyrrell Rockafellar and Roger J. B. Wets. *Variational Analysis*. Ed. by M. Berger et al. Vol. 317. Grundlehren der mathematischen Wissenschaften. Berlin, Heidelberg: Springer, 1998.

[55] Simo Särkkä and Arno Solin. *Applied Stochastic Differential Equations*. Institute of Mathematical Statistics Textbooks. Cambridge: Cambridge University Press, 2019.

[56] Claudia Schillings and Andrew M. Stuart. "Analysis of the Ensemble Kalman Filter for Inverse Problems". In: *SIAM J. Numer. Anal.* 55.3 (2017), pp. 1264–1290.

[57] Claudia Schillings, Claudia Totzeck, and Philipp Wacker. "Ensemble-Based Gradient Inference for Particle Methods in Optimization and Sampling". In: *SIAM/ASA J. Uncertainty Quantif.* 11.3 (2023), pp. 757–787.

[58] Linus Seelinger et al. "High Performance Uncertainty Quantification with Parallelized Multilevel Markov Chain Monte Carlo". In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. SC '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 1–15.

[59] A. M. Stuart. "Inverse Problems: A Bayesian Perspective". In: *Acta Numer.* 19 (2010), pp. 451–559.

[60] Alain-Sol Sznitman. "Topics in Propagation of Chaos". In: *Ecole d'Eté de Probabilités de Saint-Flour XIX — 1989*. Ed. by Donald L. Burkholder et al. Berlin, Heidelberg: Springer, 1991, pp. 165–251.

[61] Michalis K. Titsias and Omiros Papaspiliopoulos. "Auxiliary Gradient-Based Sampling Algorithms". In: *J. R. Stat. Soc. B* 80.4 (2018), pp. 749–767.

[62] Xin T. Tong and Matthias Morzfeld. "Localized ensemble Kalman inversion". In: *Inverse Probl.* 39.6 (2023), p. 064002.

[63] Philipp Wacker. "Perspectives on Locally Weighted Ensemble Kalman Methods". In: *arXiv:2402.00027* (2024).

[64] Fabian Wagner, I. Papaioannou, and E. Ullmann. "The Ensemble Kalman Filter for Rare Event Estimation". In: *SIAM/ASA J. Uncertainty Quantif.* 10.1 (2022), pp. 317–349.

[65] Shihao Yang et al. "On Parallelizable Markov Chain Monte Carlo Algorithms with Waste-Recycling". In: *Stat. Comput.* 28.5 (2018), pp. 1073–1081.