How Consensus-Based Optimization can be Interpreted as a Stochastic Relaxation of Gradient Descent

Konstantin Riedl [*1,2] **Timo Klock** [3] **Carina Geldhauser** [1,2] **Massimo Fornasier** [*1,2,4]

Abstract

We provide a novel analytical perspective on the theoretical understanding of gradient-based learning algorithms by interpreting consensus-based optimization (CBO), a recently proposed multi-particle derivative-free optimization method, as a stochastic relaxation of gradient descent. Remarkably, we observe that through communication of the particles, CBO exhibits a stochastic gradient descent (SGD)-like behavior despite solely relying on evaluations of the objective function. The fundamental value of such link between CBO and SGD lies in the fact that CBO is provably globally convergent to global minimizers for ample classes of nonsmooth and nonconvex objective functions. Hence, on the one side, we offer a novel explanation for the success of stochastic relaxations of gradient descent by furnishing useful and precise insights that explain how problem-tailored stochastic perturbations of gradient descent (like the ones induced by CBO) overcome energy barriers and reach deep levels of nonconvex functions. On the other side, and contrary to the conventional wisdom for which derivative-free methods ought to be inefficient or not to possess generalization abilities, our results unveil an intrinsic gradient descent nature of heuristics.

1. Introduction

Gradient-based learning algorithms, such as stochastic gradient descent (SGD), AdaGrad (Duchi et al., 2011) and Adam (Kingma & Ba, 2015), just to name a few, have undoubtedly been one of the cornerstones of the astounding successes of machine learning (Collobert & Weston, 2008; Graves et al., 2013; Krizhevsky et al., 2017) in the last decades. Despite an ever-growing relevance of advancing our mathematical understanding concerning the behavior of gradient-based learning algorithms, the fundamental reasons behind their empirical successes largely defy our theoretical understanding (Zhang et al., 2021; Mei et al., 2018).

In this work, we consider the more generic, ubiquitous problem of finding a global minimizer of a potentially nonsmooth and nonconvex objective function $\mathcal{E} : \mathbb{R}^d \to \mathbb{R}$, i.e., solving

$$x^* \in \arg\min_{x \in \mathbb{R}^d} \mathcal{E}(x). \tag{1}$$

Supported by illustrative numerical experiments, see Figure 1 below, we shall provide a novel analytical perspective on the theoretical understanding of gradient-based learning algorithms for such general global optimization problem by interpreting a recently proposed multi-particle metaheuristic derivative-free (zero-order) optimization method, called consensus-based optimization (CBO) (Pinnau et al., 2017; Bailo et al., 2024), as a stochastic relaxation of gradient descent (GD), see Theorem 3.1 below for the statement of our main result. The essential benefit of establishing such link between CBO and (S)GD lies in the fact that CBO is provably capable of achieving global convergence towards global minimizers for rich classes of nonsmooth and nonconvex objective functions (Carrillo et al., 2018, 2021; Fornasier et al., 2021a,b, 2022, 2023). Hence, such up to now largely unexplored connection between mathematically explainable derivative-free optimization methods and gradient-based learning algorithms discloses, on the one side, a novel and complementary perspective on why stochastic relaxations of GD are so successful, and, conversely, but no less surprising, unveils an intrinsic GD nature of heuristics on the other.

Contributions. In view of the overwhelming empirical evidence that gradient-based learning algorithms exceed in a variety of machine learning tasks what is mathematically rigorously justified, we provide in this work a novel and surprising analytical perspective on their theoretical understanding by interpreting consensus-based optimization (CBO), which is guaranteed to globally converge to global minimizers of potentially nonsmooth and nonconvex loss functions (Fornasier et al., 2021b, 2022), as a stochastic

[*]Equal contribution [1]School of Computation, Information and Technology, Technical University of Munich, Munich, Germany [2]Munich Center for Machine Learning, Munich, Germany [3]Deeptech Consulting, Oslo, Norway [4]Munich Data Science Institute, Munich, Germany. Correspondence to: Konstantin Riedl <konstantin.riedl@ma.tum.de>.

Published at the 2^{nd} Differentiable Almost Everything Workshop at the 41^{st} International Conference on Machine Learning, Vienna, Austria. July 2024.

relaxation of gradient descent (GD). Specifically, we show that in suitable scalings of its parameters, CBO — despite being a derivative-free (zero-order) optimization method — naturally approximates a stochastic gradient flow dynamics, hence implicitly behaves like a gradient-based (first-order) method, see Theorem 3.1 and Figure 1. Our results furnish useful and precise insights that explain the mechanisms which enable stochastic perturbations of GD to overcome energy barriers and to reach deep levels of nonconvex objective functions, even allowing for global optimization. While the usual approach to a global analysis of (stochastic) GD requires the loss to be L-smooth and to obey the Polyak-Łojasiewicz condition, for the global convergence of CBO merely local Lipschitz continuity and a certain growth condition around the global minimizer are required (Fornasier et al., 2021b, 2022). By establishing such link between stochastic GD on the one hand and metaheuristic black-box optimization algorithms such as CBO on the other, we not just allow for complementing our theoretical understanding of successfully deployed optimization algorithms in machine learning and beyond, but we also widen the scope of applications of methods which — in one way or another, be it explicitly or implicitly — estimate and exploit gradients.

2. Characterization of the class of objective functions

The theoretical findings of this work hold for objectives satisfying the following conditions.

Assumption 2.1. Throughout we consider $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$,

- A1 for which there exists $x^* \in \mathbb{R}^d$ such that $\mathcal{E}(x^*) = \inf_{x\in\mathbb{R}^d} \mathcal{E}(x) =: \underline{\mathcal{E}}$,
- A2 for which there exist $C_1, C_2 > 0$ such that $|\mathcal{E}(x)-\mathcal{E}(x')| \leq C_1(1+\|x\|_2+\|x'\|_2)\,\|x-x'\|_2$ for all $x, x' \in \mathbb{R}^d$ as well as $|\mathcal{E}(x) - \underline{\mathcal{E}}| \leq C_2(1 + \|x\|_2^2)$ for all $x \in \mathbb{R}^d$,
- A3 for which either $\bar{\mathcal{E}} := \sup_{x\in\mathbb{R}^d} \mathcal{E}(x) < \infty$, or for which there exist $C_3, C_4 > 0$ such that $\mathcal{E}(x) - \underline{\mathcal{E}} \geq C_3 \|x\|_2^2$ for all $x \in \mathbb{R}^d$ with $\|x\|_2 \geq C_4$,
- A4 which are semi-convex (Λ-convex for some $\Lambda \in \mathbb{R}$), i.e., $\mathcal{E}(\bullet) - \frac{\Lambda}{2}\|\bullet\|_2^2$ is convex.

A detailed discussion may be found in Appendix B.

3. Main results

Inspired by particle swarm optimization (PSO) (Kennedy & Eberhart, 1995), CBO methods employ an interacting stochastic system of N particles $X^1, \dots, X^N$ to explore the domain and to form consensus about the global minimizer x^* over time. More concretely, given a finite number of time steps K, a discrete time step size $\Delta t > 0$ and denoting the position of the i-th particle at time step $k \in \{0, \dots, K\}$ by X_k^i, this position is computed for user-specified parameters $\alpha, \lambda, \sigma > 0$ according to the iterative update rule

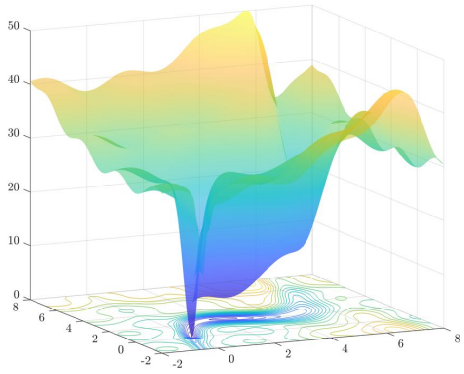
$$\begin{aligned} X_k^i = X_{k-1}^i &- \Delta t\lambda \left(X_{k-1}^i - x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^N)\right) \\ &+ \sigma D\left(X_{k-1}^i - x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^N)\right) B_k^i, \end{aligned} \tag{2}$$

where $\widehat{\rho}_k^N$ denotes the empirical measure of the particles at time step k, i.e., $\widehat{\rho}_k^N = \frac{1}{N}\sum_{i=1}^N \delta_{X_k^i}$. In the spirit of the exploration-exploitation philosophy of evolutionary computation techniques (Holland, 1975; Bäck et al., 1997; Fogel, 2000), the dynamics (2) of each particle is governed by two competing terms, one being stochastic, the other deterministic in nature. The first of the two terms on the right-hand side of (2) imposes a deterministic drift towards the so-called consensus point $x_\alpha^{\mathcal{E}}$, which is defined for a measure $\varrho \in \mathcal{P}(\mathbb{R}^d)$ by

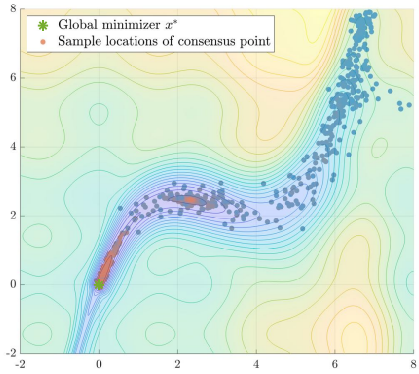
$$x_\alpha^{\mathcal{E}}(\varrho) := \int x \frac{\omega_\alpha^{\mathcal{E}}(x)}{\|\omega_\alpha^{\mathcal{E}}\|_{L^1(\varrho)}} d\varrho(x), \tag{3}$$

with $\omega_\alpha^{\mathcal{E}}(x) := \exp(-\alpha\mathcal{E}(x))$. Notice that in the case $\varrho = \widehat{\rho}_k^N$, Formula (3) is just a weighted (exploiting the particles' knowledge of their objective function values) convex combination of the positions X_k^i. To be precise, owed to the particular choice of Gibbs weights $\omega_\alpha^{\mathcal{E}}$, larger mass is attributed to particles with comparably low objective value, whereas only little mass is given to particles whose value is undesirably high. This facilitates the interpretation that $x_\alpha^{\mathcal{E}}(\widehat{\rho}_k^N)$ is an approximation to $\arg\min_{i=1,\dots,N} \mathcal{E}(X_k^i)$, which improves as $\alpha \to \infty$ and which can be regarded as a proxy for the global minimizer x^*, based on the information currently available to the particles. Theoretically, this is justified by the log-sum-exp trick or the Laplace principle (Dembo & Zeitouni, 1998; Miller, 2006). Let us further remark that the particles communicate and exchange information amongst each other exclusively through sharing the consensus point $x_\alpha^{\mathcal{E}}$. The other term in (2) is a stochastic diffusion injecting randomness into the dynamics, thereby encoding its explorative nature. Given i.i.d. Gaussian random vectors B_k^i in $\mathbb{R}^d$ with zero mean and covariance matrix $\Delta t\mathrm{Id}$, each particle is subject to anisotropic noise, i.e., $D(\bullet) = \mathrm{diag}(\bullet)$, which favors exploration the farther a particle is away from the consensus point in a certain direction. System (2) is complemented with independent initial data x_0^i distributed according to a common probability measure $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$, i.e., $X_0^i = x_0^i \sim \rho_0$.

An insightful theoretical understanding of the behavior of CBO is to be gained, as we are about to show, by tracing the dynamics of the consensus point $x_\alpha^{\mathcal{E}}$ of the CBO algorithm (2). For this purpose, let us introduce the CBO scheme



(a) A noisy Canyon function $\mathcal{E}$ with a valley shaped as a third degree polynomial.



(b) The CBO scheme (4) (sampled over several runs) follows on average the valley of $\mathcal{E}$ while passing over local minima.

Figure 1: An illustration of the intuition that the CBO scheme (4) can be regarded as a stochastic derivative-free relaxation of GD. To find the global minimizer x^* of the nonconvex objective function $\mathcal{E}$ depicted in (a), we run the CBO algorithm (2) for $K = 250$ iterations with parameters $\Delta t = 0.1$, $\alpha = 100$, $\lambda = 1$ and $\sigma = 1.6$, and $N = 200$ particles, initialized i.i.d. according to $\rho_0 = \mathcal{N}\big((8,8), 0.5\mathrm{Id}\big)$. This experiment is performed 50 times. For each run we depict in (b) the positions of the consensus points computed during the CBO algorithm (2), i.e., the iterates of the CBO scheme (4) for $k = 1, \dots, K$. The color of the individual points corresponds to time, i.e., iterates at the beginning of the scheme are plotted in blue, whereas later iterates are colored orange. We observe that, after starting close to the initial position, the trajectories of the consensus points follow the path of the valley leading to the global minimizer x^*, until it is reached. In particular, unlike GD (cf. Figure C.1b), the scheme (4) has the capability of jumping over locally deeper passages. Such desirable behavior is observed also for the Langevin dynamics (6) (see Figure C.1c), which can be regarded as a stochastic (noisy) version of GD.

as the iterates $(x_k^{\mathrm{CBO}})_{k=0,\dots,K}$ defined according to

$$\begin{aligned} x_k^{\mathrm{CBO}} &= x_\alpha^{\mathcal{E}}(\widehat{\rho}_k^N), \quad \text{with} \quad \widehat{\rho}_k^N = \frac{1}{N}\sum_{i=1}^N \delta_{X_k^i}, \\ x_0^{\mathrm{CBO}} &= x_0 \sim \rho_0, \end{aligned} \tag{4}$$

where the particles' positions X_k^i are given by Equation (2). The main theoretical finding of this work is concerned with the observation that the iterates of the CBO scheme (4), i.e., the trajectory of the consensus point $x_\alpha^{\mathcal{E}}$, follow, with high probability, a stochastically perturbed GD. This is illustrated in Figure 1 above and made rigorous in the following Theorem 3.1, whose proof is deferred to Section C.

Theorem 3.1 (CBO is a stochastic relaxation of GD)**.** *Let $\mathcal{E} \in \mathcal{C}^1(\mathbb{R}^d)$ be L-smooth and satisfy minimal assumptions (summarized in Assumption 2.1 above). Then, for $\tau > 0$ (satisfying $\tau < 1/(-2\Lambda)$ if $\Lambda < 0$) and with parameters $\alpha, \lambda, \sigma, \Delta t > 0$ such that $\alpha \gtrsim \frac{1}{\tau} d \log d$, the iterates $(x_k^{\mathrm{CBO}})_{k=0,\dots,K}$ of the CBO scheme (4) follow a stochastically perturbed GD, i.e., they obey*

$$x_k^{\mathrm{CBO}} = x_{k-1}^{\mathrm{CBO}} - \tau \nabla \mathcal{E}(x_{k-1}^{\mathrm{CBO}}) + g_k, \tag{5}$$

where g_k is stochastic noise fulfilling for each $k = 1, \dots, K$ with high probability the quantitative estimate $\|g_k\|_2 = \mathcal{O}\big(|\lambda - 1/\Delta t| + \sigma\sqrt{\Delta t} + \sqrt{\tau/\alpha} + N^{-1/2}\big) + \mathcal{O}(\tau)$.

Let us now comment on technical aspects of Theorem 3.1, describe its interpretation and discuss its implications.

Concerning the assumptions, it shall be mentioned that, in particular compared to Polyak-Łojasiewicz-like conditions (Karimi et al., 2016) or certain families of log-Sobolev inequalities (Chizat & Bach, 2018) that are required to analyze the dynamics of gradient-based methods such as (S)GD or the Langevin dynamics, the assumptions under which our statement holds are rather weak and complementary. Combined with similar assumptions being sufficient to prove global convergence of CBO, this extends the class of functions, for which SGD-like methods are successful in global optimization.

The statement of Theorem 3.1 has to be read with a twofold interpretation. First, in view of the capability of CBO to converge to global minimizers for rich classes of nonsmooth and nonconvex objectives (Fornasier et al., 2021b, 2022; Riedl, 2023), Theorem 3.1 states that there exist stochastic relaxations of GD that are provably able to robustly and reliably overcome energy barriers and reach deep levels of nonconvex functions. Such relaxations may even be derivative-free and do not require smoothness of the objective, as in CBO. Second, and conversely, against the common wisdom that derivative-free optimization heuristics search the domain mainly by random exploration and therefore ought to be inefficient, we provide evidence that such heuristics in fact

work successfully in finding benign optima (Duchi et al., 2015; Nesterov & Spokoiny, 2017; Chen et al., 2017; Nikolakakis et al., 2022), precisely because they are suitable stochastic relaxations of gradient-based methods.

The interpretation of the CBO scheme (4) as a stochastic relaxation of GD is substantiated visually, analytically and numerically as follows.
While the trajectories of (4) are to be seen in Figure 1b, we depict for comparison in Figure C.1c the discretized dynamics of the annealed Langevin dynamics (Chiang et al., 1987; Roberts & Tweedie, 1996; Durmus & Moulines, 2017),

$$dX_t = -\nabla \mathcal{E}(X_t)\, dt + \sqrt{2\beta_t^{-1}}\, dB_t. \tag{6}$$

Both stochastic methods are capable of global minimization while overcoming energy barriers and escaping local minima. For analyses of the (annealed) Langevin dynamics we refer to (Gelfand & Mitter, 1991; Chizat, 2022).
The stochastic perturbations g_k in (5) are meaningful and not generic as they obey precise scalings thanks to the established estimate in Theorem 3.1. In particular, as reflected by the first term of the bound on the error $\|g_k\|_2$, they become tighter as soon as the discrete CBO time step size $\Delta t \ll 1$, the drift parameter $\lambda \approx 1/\Delta t$, the noise parameter σ becomes smaller, the weight parameter α is sufficiently large, and the number of employed particles N becomes larger. This behavior is confirmed numerically in Figure 2 by measuring the closeness between the trajectories of the CBO scheme (4) and GD. More precisely, better approximation is

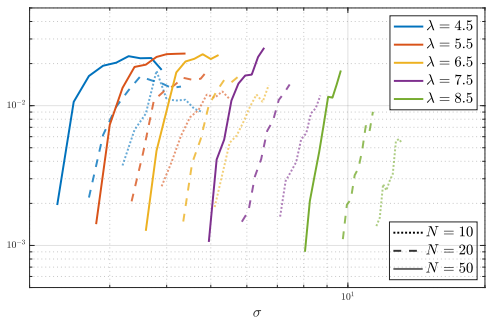


Figure 2: Numerical analysis of the approximation error between the trajectories of the CBO scheme (4) and GD, i.e., the stochastic noise g_k in (5). In the setting of the Canyon function $\mathcal{E}$ from Figure 1a but without a local minimum in the valley, we measure the distance between the two trajectories and plot the resulting approximation error for different values of λ (different colors), σ (horizontal axis), and N (different line styles). The other parameters of the CBO scheme (4) are $K = 1000$, $\Delta t = 0.1$ and $\alpha = 10^{16}$ with the remaining setting being as in Figure 1.

achieved for the values of λ closer to $1/\Delta t$ (compare lines with different colors but same line style, and notice that smaller error can be obtained for larger λ), larger choices of N (compare different line styles within a color), and σ as small as possible (each line decreases as σ decreases). For fixed λ and N, however, σ needs to be sufficiently large (in particular in case of a fixed number of iterates K) to allow the CBO scheme (4) to iteratively explore the energy landscape within the time horizon. As visible from Figure 2, a larger number of particles N is needed to pass to smaller σ and thus better approximation. Regarding the second term of the bound on the error $\|g_k\|_2$, we conjecture a potential amelioration of the estimate by refining the quantitative Laplace principle from (Fornasier et al., 2021b) involved in the proof of Proposition C.2, which would allow to remove the order $\mathcal{O}(\tau)$ dependence of the bound. Yet, as it stands, this term is about a *deterministic bounded* perturbation of the gradient, which is possibly of smaller magnitude than the gradient. Such bounded perturbation alone does not allow to overcome local energy barriers in general (just think of a local minimizer, around which the magnitude of gradients grows faster than the displacement: any movement from the minimizer ought necessarily to get reverted). Hence, it is the *stochastic* part of the perturbation that enables the convergence to global minimizers. In fact, for a moderate time step size $\Delta t > 0$, a drift parameter $\lambda > 0$ relatively small compared to $1/\Delta t$, a non-insignificant noise parameter $\sigma > 0$, a moderate value of the weight parameter $\alpha > 0$ and a modest number N of particles, CBO is factually a stochastic relaxation of GD with strong noise.

Apart from gaining primarily theoretical insights from this link, let us conclude this section by mentioning a further, more practical aspect of establishing such a connection. In several real-world applications, including various machine learning settings, using gradients may be undesirable or even not feasible. This can be due to the black-box nature or nonsmoothness of the objective, memory limitations constraining the use of automatic differentiation, a substantial presence of spurious local minima, or the fact that gradients carry relevant information about data, which one may wish to keep private. In machine learning, the problems of hyperparameter tuning (Bergstra et al., 2011; Rapin & Teytaud, 2018), convex bandits (Agarwal et al., 2011; Shamir, 2017), reinforcement learning (Sutton & Barto, 1998), the training of sparse and pruned neural networks (Hoefler et al., 2021), and federated learning (Shokri & Shmatikov, 2015; McMahan et al., 2017) stimulate interest in methods alternative to gradient-based ones. In such situations, if one still wishes to rely on a GD-like optimization behavior, Theorem 3.1 suggests the use of CBO (or related methods such as PSO (Cipriani et al., 2022; Huang et al., 2023)), which will be reliable and efficient, with linear complexity in the number of deployed particles. We report, for instance, recent ideas in the setting of clustered federated learning (Carrillo et al., 2023), where CBO is leveraged to avoid reverse engineering of private data through exchange of gradients.

References

Agarwal, A., Foster, D. P., Hsu, D. J., Kakade, S. M., and Rakhlin, A. Stochastic convex optimization with bandit feedback. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

Anitescu, M. Degenerate nonlinear programming with a quadratic growth condition. *SIAM J. Optim.*, 10(4):1116–1135, 2000.

Bäck, T., Fogel, D. B., and Michalewicz, Z. (eds.). *Handbook of evolutionary computation*. Institute of Physics Publishing, Bristol; Oxford University Press, New York, 1997.

Bailo, R., Barbaro, A., Gomes, S. N., Riedl, K., Roith, T., Totzeck, C., and Vaes, U. Cbx: Python and julia packages for consensus-based interacting particle methods. *arXiv preprint arXiv:2403.14470*, 2024.

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. Algorithms for hyper-parameter optimization. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Bolte, J., Nguyen, T. P., Peypouquet, J., and Suter, B. W. From error bounds to the complexity of first-order descent methods for convex functions. *Math. Program.*, 165(2, Ser. A):471–507, 2017.

Bungert, L., Roith, T., and Wacker, P. Polarized consensus-based dynamics for optimization and sampling. *Math. Program.*, pp. 1–31, 2024.

Carrillo, J. A., Choi, Y.-P., Totzeck, C., and Tse, O. An analytical framework for consensus-based global optimization method. *Math. Models Methods Appl. Sci.*, 28 (6):1037–1066, 2018.

Carrillo, J. A., Jin, S., Li, L., and Zhu, Y. A consensus-based global optimization method for high dimensional machine learning problems. *ESAIM Control Optim. Calc. Var.*, 27(suppl.):Paper No. S5, 22, 2021.

Carrillo, J. A., Trillos, N. G., Li, S., and Zhu, Y. FedCBO: Reaching group consensus in clustered federated learning through consensus-based optimization. *J. Mach. Learn. Res. (to appear), arXiv preprint arXiv:2305.02894*, 2023.

Chen, P., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C. ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In Thuraisingham, B., Biggio, B., Freeman, D. M., Miller, B., and Sinha, A. (eds.), *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017*, pp. 15–26. ACM, 2017.

Chiang, T.-S., Hwang, C.-R., and Sheu, S. J. Diffusion for global optimization in $\mathbb{R}^n$. *SIAM J. Control Optim.*, 25 (3):737–753, 1987. ISSN 0363-0129.

Chizat, L. Mean-field langevin dynamics: Exponential convergence and annealing. *Trans. Mach. Learn. Res.*, 2022, 2022.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Chow, Y. S. and Teicher, H. *Probability theory*. Springer Texts in Statistics. Springer-Verlag, New York, third edition, 1997. ISBN 0-387-98228-0. Independence, interchangeability, martingales.

Cipriani, C., Huang, H., and Qiu, J. Zero-inertia limit: from particle swarm optimization to consensus-based optimization. *SIAM J. Math. Anal.*, 54(3):3091–3121, 2022. ISSN 0036-1410,1095-7154.

Collobert, R. and Weston, J. A unified architecture for natural language processing: deep neural networks with multitask learning. In Cohen, W. W., McCallum, A., and Roweis, S. T. (eds.), *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pp. 160–167. ACM, 2008.

De Giorgi, E. New problems on minimizing movements. In *Boundary value problems for partial differential equations and applications*, volume 29 of *RMA Res. Notes Appl. Math.*, pp. 81–98. Masson, Paris, 1993.

Dembo, A. and Zeitouni, O. *Large deviations techniques and applications*, volume 38 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 1998.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, 2011. ISSN 1532-4435,1533-7928.

Duchi, J. C., Jordan, M. I., Wainwright, M. J., and Wibisono, A. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Trans. Inf. Theory*, 61(5):2788–2806, 2015.

Durmus, A. and Moulines, E. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 2017. ISSN 1050-5164,2168-8737.

Fogel, D. B. *Evolutionary computation. Toward a new philosophy of machine intelligence*. IEEE Press, Piscataway, NJ, second edition, 2000.

Fornasier, M., Huang, H., Pareschi, L., and Sünnen, P. Consensus-based optimization on hypersurfaces: Well-posedness and mean-field limit. *Math. Models Methods Appl. Sci.*, 30(14):2725–2751, 2020.

Fornasier, M., Huang, H., Pareschi, L., and Sünnen, P. Consensus-based optimization on the sphere: convergence to global minimizers and machine learning. *J. Mach. Learn. Res.*, 22:Paper No. 237, 55, 2021a.

Fornasier, M., Klock, T., and Riedl, K. Consensus-based optimization methods converge globally. *SIAM J. Optim. (to appear), arXiv preprint arXiv:2103.15130*, 2021b.

Fornasier, M., Klock, T., and Riedl, K. Convergence of anisotropic consensus-based optimization in mean-field law. In Laredo, J. L. J., Hidalgo, J. I., and Babaagba, K. O. (eds.), *Applications of Evolutionary Computation - 25th European Conference, EvoApplications 2022, Held as Part of EvoStar 2022, Madrid, Spain, April 20-22, 2022, Proceedings*, volume 13224 of *Lecture Notes in Computer Science*, pp. 738–754. Springer, 2022.

Fornasier, M., Richtárik, P., Riedl, K., and Sun, L. Consensus-based optimization with truncated noise. *Eur. J. Appl. Math. (special issue "From integro-differential models to data-oriented approaches for emergent phenomena")*, pp. 1–24, 2023.

Fournier, N. and Guillin, A. On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields*, 162(3-4):707–738, 2015.

Gelfand, S. B. and Mitter, S. K. Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$. *SIAM J. Control Optim.*, 29(5):999–1018, 1991. ISSN 0363-0129.

Graves, A., Mohamed, A., and Hinton, G. E. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, pp. 6645–6649. IEEE, 2013.

Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., and Peste, A. Sparsity in deep learning: pruning and growth for efficient inference and training in neural networks. *J. Mach. Learn. Res.*, 22:Paper No. 241, 124, 2021. ISSN 1532-4435,1533-7928.

Holland, J. H. *Adaptation in natural and artificial systems. An introductory analysis with applications to biology, control, and artificial intelligence*. University of Michigan Press, Ann Arbor, Mich., 1975.

Huang, H., Qiu, J., and Riedl, K. On the global convergence of particle swarm optimization methods. *Appl. Math. Optim.*, 88(2):Paper No. 30, 44, 2023. ISSN 0095-4616,1432-0606.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In Frasconi, P., Landwehr, N., Manco, G., and Vreeken, J. (eds.), *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I*, volume 9851 of *Lecture Notes in Computer Science*, pp. 795–811. Springer, 2016.

Kennedy, J. and Eberhart, R. Particle swarm optimization. In *Proceedings of International Conference on Neural Networks (ICNN'95), Perth, WA, Australia, November 27 - December 1, 1995*, pp. 1942–1948. IEEE, 1995.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, 2017.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Singh, A. and Zhu, J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci. USA*, 115(33):E7665–E7671, 2018. ISSN 0027-8424,1091-6490.

Miller, P. D. *Applied asymptotic analysis*, volume 75 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2006.

Necoara, I., Nesterov, Y., and Glineur, F. Linear convergence of first order methods for non-strongly convex optimization. *Math. Program.*, 175(1-2, Ser. A):69–107, 2019.

Nesterov, Y. and Spokoiny, V. Random gradient-free minimization of convex functions. *Found. Comput. Math.*, 17 (2):527–566, 2017.

Nikolakakis, K., Haddadpour, F., Kalogerias, D., and Karbasi, A. Black-box generalization: Stability of zeroth-order learning. volume 35, pp. 31525–31541, 2022.

Pinnau, R., Totzeck, C., Tse, O., and Martin, S. A consensus-based model for global optimization and its mean-field limit. *Math. Models Methods Appl. Sci.*, 27(1):183–204, 2017.

Rapin, J. and Teytaud, O. Nevergrad — a gradient-free optimization platform, 2018.

Riedl, K. Leveraging memory effects and gradient information in consensus-based optimisation: On global convergence in mean-field law. *Eur. J. Appl. Math.*, pp. 1–32, 2023.

Roberts, G. O. and Tweedie, R. L. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996. ISSN 1350-7265.

Santambrogio, F. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bull. Math. Sci.*, 7(1):87–154, 2017.

Shamir, O. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *J. Mach. Learn. Res.*, 18:Paper No. 52, 11, 2017. ISSN 1532-4435,1533-7928.

Shokri, R. and Shmatikov, V. Privacy-preserving deep learning. In *53rd Annual Allerton Conference on Communication, Control, and Computing, Allerton 2015, Allerton Park & Retreat Center, Monticello, IL, USA, September 29 - October 2, 2015*, pp. 909–910. IEEE, 2015.

Sutton, R. S. and Barto, A. G. *Reinforcement learning - an introduction*. Adaptive computation and machine learning. MIT Press, 1998. ISBN 978-0-262-19398-6.

Villani, C. *Optimal transport: Old and new*, volume 338 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009.

Xu, Y., Lin, Q., and Yang, T. Adaptive svrg methods under error bound conditions with unknown growth parameter. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, 2021.

Appendices

Appendices for the paper: "**How Consensus-Based Optimization can be Interpreted as a Stochastic Relaxation of Gradient Descent**" authored by **Konstantin Riedl**, **Timo Klock**, **Carina Geldhauser**, and **Massimo Fornasier**.

In the GitHub repository CBOGlobalConvergenceAnalysis/CBOstochasticGD, we provide the implementation of the algorithms analyzed in this work and the code used to create the visualizations.

A. Introductory facts

Notation. To keep the notation concise, we hide generic constants, i.e., we write $a \lesssim b$ for $a \leq cb$, if c is a constant independent of problem-dependent constants. Moreover, since we work with random variables in several instances, many equalities and inequalities hold almost surely without being mentioned explicitly. We abbreviate with i.i.d. independently and identically distributed.

We write $\|\bullet\|_2$ and $\langle \bullet, \bullet \rangle$ for the Euclidean norm and scalar product on $\mathbb{R}^d$, respectively. Euclidean balls are denoted by $B_r(x) := \{z \in \mathbb{R}^d : \|z - x\|_2 \leq r\}$. Moreover, we write $\|\bullet\|_\infty$ for the ℓ^∞-norm and denote the associated ℓ^∞-balls by $B_r^\infty(x) := \{z \in \mathbb{R}^d : \|z - x\|_\infty \leq r\}$.

For the space of continuous functions $f : X \to Y$ we write $\mathcal{C}(X, Y)$, with $X \subset \mathbb{R}^n$ and a suitable topological space Y. For an open set $X \subset \mathbb{R}^n$ and for $Y = \mathbb{R}^m$ the space $\mathcal{C}^k(X, Y)$ contains functions $f \in \mathcal{C}(X, Y)$ that are k-times continuously differentiable. We omit Y in the real-valued case, i.e., $\mathcal{C}(X) = \mathcal{C}(X, \mathbb{R})$ and $\mathcal{C}^k(X) = \mathcal{C}^k(X, \mathbb{R})$.

A function $f \in \mathcal{C}^1(\mathbb{R}^d)$ is L-smooth if $\|\nabla f(x) - \nabla f(x')\|_2 \leq L \|x - x'\|_2$ for all $x, x' \in \mathbb{R}^d$.

The operator ∇ denotes the gradient of a function on $\mathbb{R}^d$.

The operator $\mathrm{diag} : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ denotes the operator mapping a vector onto a diagonal matrix with the vector as its diagonal.

Convex analysis. For a convex function $f \in \mathcal{C}(\mathbb{R}^d)$ the subdifferential $\partial f(x)$ at a point $x \in \mathbb{R}^d$ is the set

$$\partial f(x) = \left\{ p \in \mathbb{R}^d : f(y) \geq f(x) + \langle p, y - x \rangle \text{ for all } y \in \mathbb{R}^d \right\}.$$

In the setting $f \in \mathcal{C}(\mathbb{R}^d)$, $\partial f(x)$ is closed, convex, nonempty and bounded. If $f \in \mathcal{C}^1(\mathbb{R}^d)$, $\partial f(x) = \{\nabla f(x)\}$. Moreover, it is straightforward to verify that for $x_1, x_2, p_1, p_2 \in \mathbb{R}^d$ with $p_1 \in \partial f(x_1)$ and $p_2 \in \partial f(x_2)$ it holds $\langle p_1 - p_2, x_1 - x_2 \rangle \geq 0$.

Probability measures. The set of all Borel probability measures over $\mathbb{R}^d$ is denoted by $\mathcal{P}(\mathbb{R}^d)$. For $p > 0$, we collect measures $\varrho \in \mathcal{P}(\mathbb{R}^d)$ with finite p-th moment $\int \|x\|_2^p \, d\varrho(x)$ in $\mathcal{P}_p(\mathbb{R}^d)$. $\mathcal{P}_p(\mathbb{R}^d)$ is metrized by the Wasserstein-p distance W_p, see, e.g., (Ambrosio et al., 2008; Villani, 2009) and the subsequent paragraph.

The Dirac delta δ_x for a point $x \in \mathbb{R}^d$ is a measure satisfying $\delta(B) = 1$ if $x \in B$ and $\delta(B) = 0$ if $x \notin B$ for any measurable set $B \subset \mathbb{R}^d$.

$\mathcal{N}(m, \Sigma)$ denotes a Gaussian distribution with mean m and covariance matrix Σ.

Wasserstein distance. For any $1 \leq p < \infty$, the Wasserstein-p distance between two Borel probability measures $\varrho, \varrho' \in \mathcal{P}_p(\mathbb{R}^d)$ is defined by

$$W_p(\varrho, \varrho') = \left(\inf_{\gamma \in \Pi(\varrho, \varrho')} \int \|x - x'\|_2^p \, d\gamma(x, x') \right)^{1/p}, \tag{7}$$

where $\Pi(\varrho, \varrho')$ denotes the set of all couplings of (a.k.a. transport plans between) ϱ and ϱ', i.e., the collection of all Borel probability measures over $\mathbb{R}^d \times \mathbb{R}^d$ with marginals ϱ and ϱ' on the first and second component, respectively, see, e.g., (Ambrosio et al., 2008; Villani, 2009). $\mathcal{P}_p(\mathbb{R}^d)$ endowed with the Wasserstein-p distance W_p is a complete separable metric space (Ambrosio et al., 2008, Proposition 7.1.5).

A generalized triangle-type inequality. It holds for $p, J \in \mathbb{N}$ by Hölder's inequality

$$\left| \sum_{j=1}^{J} a_j \right|^p \leq J^{p-1} \sum_{j=1}^{J} |a_j|^p \,. \tag{8}$$

A discrete variant of Grönwall's inequality. If $z_k \leq a z_{k-1} + b$ with $a, b \geq 0$ for all $k \geq 1$, then

$$z_k \leq a^k z_0 + b \sum_{\ell=0}^{k-1} a^\ell \leq a^k z_0 + b \prod_{\ell=1}^{k-1} (1 + a) \leq a^k z_0 + b e^{a(k-1)} \tag{9}$$

for all $k \geq 1$. Notice that, while the first inequality in (9) is as sharp as the initial estimates, the remaining two inequalities are rather rough upper bounds.

B. Discussion of Assumption 2.1

Assumption A1 requires that the continuous objective function $\mathcal{E}$ attains its globally minimal value $\underline{\mathcal{E}}$ at some $x^* \in \mathbb{R}^d$. This does in particular not exclude objectives with multiple global minimizers.

Remark B.1. For the global convergence results (Fornasier et al., 2021b, 2022) of CBO, however, uniqueness of the global minimizer x^* is required and implied by an additional local coercivity condition of the form

$$\begin{aligned} \|x - x^*\|_\infty &\leq \frac{1}{\eta} (\mathcal{E}(x) - \underline{\mathcal{E}})^\nu \quad \text{for all } x \in B_{R_0}^\infty(x^*) \\ \mathcal{E}(x) - \underline{\mathcal{E}} &> \mathcal{E}_\infty \quad \text{for all } x \in \left(B_{R_0}^\infty(x^*)\right)^c \end{aligned}$$

with constants $\eta, \nu, \mathcal{E}_\infty, R_0 > 0$. It can be regarded as a tractability condition of the energy landscape of $\mathcal{E}$ and is also known as the inverse continuity property from (Fornasier et al., 2021a) or as the error bound condition from (Anitescu, 2000; Xu et al., 2017; Bolte et al., 2017; Necoara et al., 2019).

To deploy CBO in the setting of objective functions with several global minima, Bungert et al. (2024) propose a polarized variant of CBO, which localizes the dynamics by integrating a kernel in the computation of the consensus point (3). This ensures that each particle is primarily influenced by particles close to it, allowing for the creation of clusters.

Assumptions A2 and A3 can be regarded as regularity conditions on the objective landscape of $\mathcal{E}$. The first part of A2 is a local Lipschitz condition, which ensures that the objective function does not change too quickly, assuring that the information obtained when evaluating the function is informative within a region around the point of evaluation. The second part of A2 controls and limits the growth of the objective in the farfield. In combination with the second option in A3 this forces the objective to grow quadratically in the farfield. However, note that one can always redefine the objective outside a sufficiently large ball such that both conditions are met while the other assumptions are preserved. Alternatively, the first option in A3 allows for bounded functions.

Assumption A4 requires the objective $\mathcal{E}$ to be semi-convex with parameter $\Lambda \in \mathbb{R}$. For $\Lambda > 0$, Λ-convexity is stronger than convexity (strong convexity with parameter Λ). For $\Lambda < 0$, semi-convexity is weaker, i.e., potentially nonconvex functions $\mathcal{E}$ are included in the definition. The class of semi-convex functions is typical in the literature of gradient flows, since their general theory extends from the convex to this more general setting (Santambrogio, 2017). One particular property, which

we shall exploit in this work, is that for such functions the time discretization of a gradient flow, potentially for a small step size, defined through an iterated scheme, called minimizing movement scheme (De Giorgi, 1993), is well-defined. However, while semi-convexity is useful to ensure the well-posedness of gradient flows, it is not sufficient to obtain convergence to global minimizers. Other properties such as the Polyak-Łojasiewicz condition (Karimi et al., 2016) or the log-Sobolev inequalities governing the flow of the Langevin dynamics (Chizat & Bach, 2018) may be necessary.

C. Consensus-based optimization is a stochastic relaxation of gradient descent

In this section we present the technical details behind the main theoretical result of this work, Theorem 3.1, i.e., we explain how to establish a connection between the CBO scheme (4), which captures the flow of the derivative-free CBO dynamics (2), and GD.

From CBO to consensus hopping. Let us envision for the moment the movement of the particles during the CBO dynamics (2). At every time step k, after having computed $x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^N)$, each particle moves a $\Delta t\lambda$ fraction of its distance towards this consensus point, before being perturbed by stochastic noise. As we let $\lambda \to 1/\Delta t$, the particles' velocities increase, until, in the case $\lambda = 1/\Delta t$, each of them hops directly to the previously computed consensus point, followed by a random fluctuation. Put differently, we are left with a numerical scheme, which, at time step k, samples N particles around the old iterate in order to subsequently compute as new iterate the consensus point (3) of the empirical measure of the samples. Such algorithm is precisely a Monte Carlo approximation of the consensus hopping (CH) scheme with iterates $(x_k^{\mathrm{CH}})_{k=0,\dots,K}$ defined by

$$\begin{aligned} x_k^{\mathrm{CH}} &= x_\alpha^{\mathcal{E}}(\mu_k), \quad \text{with} \\ \mu_k &= \mathcal{N}\big(x_{k-1}^{\mathrm{CH}}, \widetilde{\sigma}^2 \mathrm{Id}\big), \\ x_0^{\mathrm{CH}} &= x_0. \end{aligned} \tag{10}$$

Theorem C.1 in Appendix C.2 makes this intuition rigorous by quantifying the approximation quality between the CBO and the CH scheme in terms of the parameters of the two schemes. Sample trajectories of the CH scheme are depicted in Figure C.1a.

From CH to GD. With the sampling measure μ_k assigning (in particular for small $\widetilde{\sigma}$) most mass to the region close to the old iterate, the CH scheme (10) improves at every time step k its objective function value while staying near the previous iterate. A conceptually analogous behavior to such localized sampling can be achieved through penalizing the length of the step taken at time step k. This gives rise to an implicit version of the CH scheme with iterates $(\widetilde{x}_k^{\mathrm{CH}})_{k=0,\dots,K}$ given as

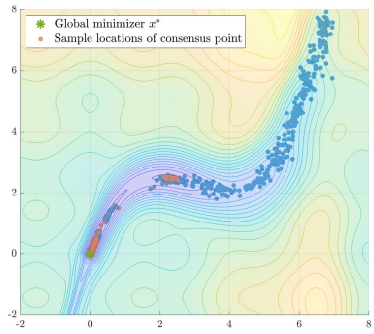
$$\begin{aligned} \widetilde{x}_k^{\mathrm{CH}} &= \operatorname*{arg\,min}_{x\in\mathbb{R}^d} \widetilde{\mathcal{E}}_k(x), \quad \text{with} \\ \widetilde{\mathcal{E}}_k(x) &:= \frac{1}{2\tau}\left\|x_{k-1}^{\mathrm{CH}} - x\right\|_2^2 + \mathcal{E}(x), \\ \widetilde{x}_0^{\mathrm{CH}} &= x_0. \end{aligned} \tag{11}$$

Actually, the modulated objective $\widetilde{\mathcal{E}}_k$ defined in (11) naturally appears when writing out the expression of $x_\alpha^{\mathcal{E}}(\mu_k)$ from (10) using that μ_k is a Gaussian. This creates a link between the sampling width $\widetilde{\sigma}$ and the step size τ. The fact that the parameter τ can be seen as the step size of (11) becomes apparent when observing that the optimality condition of the k-th iterate of (11) reads $\widetilde{x}_k^{\mathrm{CH}} = x_{k-1}^{\mathrm{CH}} - \tau\nabla\mathcal{E}(\widetilde{x}_k^{\mathrm{CH}})$, which is an implicit gradient step. Proposition C.2 in Appendix C.2 estimates the discrepancy between x_k^{CH} and $\widetilde{x}_k^{\mathrm{CH}}$ employing the quantitative Laplace principle (Fornasier et al., 2021b, Proposition 18).

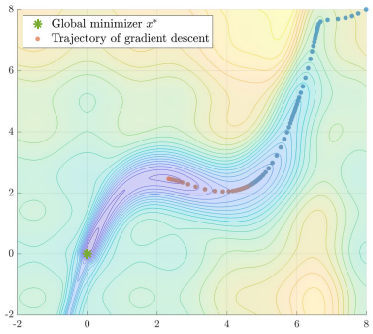
Let us conclude this discussion by remarking that the scheme (11) itself is not self-consistent but requires the computation of the iterates of the CH scheme (10). For this reason we introduce the minimizing movement scheme (MMS) (De Giorgi, 1993) as the iterates $(x_k^{\mathrm{MMS}})_{k=0,\dots,K}$ given according to

$$\begin{aligned} x_k^{\mathrm{MMS}} &= \operatorname*{arg\,min}_{x\in\mathbb{R}^d} \mathcal{E}_k(x), \quad \text{with} \\ \mathcal{E}_k(x) &:= \frac{1}{2\tau}\left\|x_{k-1}^{\mathrm{MMS}} - x\right\|_2^2 + \mathcal{E}(x), \\ x_0^{\mathrm{MMS}} &= x_0, \end{aligned} \tag{12}$$

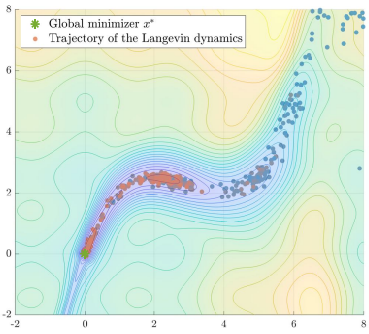
which is known to be the discrete-time implicit Euler of the gradient flow $\frac{d}{dt}x(t) = -\nabla\mathcal{E}(x(t))$ (Santambrogio, 2017).



(a) The CH scheme (10) (sampled over several runs) follows on average the valley of $\mathcal{E}$ and can occasionally escape local minima.



(b) GD gets stuck in a local minimum of $\mathcal{E}$.



(c) The Langevin dynamics (6) (sampled over several runs) follows on average the valley of $\mathcal{E}$ and escapes local minima.

Figure C.1: An illustrative comparison between the algorithms discussed in this work. While GD (obtained as an explicit Euler time discretization of $\frac{d}{dt}x(t) = -\nabla\mathcal{E}(x(t))$ with time step size $\Delta t = 0.01$ and ran for $K = 10^4$ iterations) gets stuck in a local minimum along the valley of $\mathcal{E}$ (see (b)), the stochastic algorithms in (a) and (c) as well as Figure 1b have the capability of escaping local minima. In (a) we depict the positions of the consensus hopping scheme (10) for $K = 250$ iterations with parameters $\alpha = 100$ and $\widetilde{\sigma} = 0.6$, and where we approximate the underlying measure μ_k at each step k using 200 samples. The ability of the CH scheme to escape local minima improves with larger $\widetilde{\sigma}$, see Figure G.1 in Appendix G. In (c) we depict the trajectory of the overdamped Langevin dynamics (6) with $\beta_t = 0.02\log(t+1)$ (obtained as an Euler-Maruyama time discretization of (6) with time step size $\Delta t = 0.001$ and ran for $K = 10^4$ iterations). The remaining setting is as in Figure 1, in particular, 50 individual runs of the experiment are plotted in (a) and (c).

C.1. Proof of the main result, Theorem 3.1

Proof of Theorem 3.1. From the optimality condition of the scheme $(\widetilde{x}_k^{\mathrm{CH}})_{k=1,\dots,K}$ in (11) and with the iterations $(x_k^{\mathrm{CH}})_{k=1,\dots,K}$ as in (10), we get $\big(\widetilde{x}_k^{\mathrm{CH}} - x_{k-1}^{\mathrm{CH}}\big) + \tau\nabla\mathcal{E}(\widetilde{x}_k^{\mathrm{CH}}) = 0$. Using this we decompose

$$\begin{aligned} x_k^{\mathrm{CBO}} &= \widetilde{x}_k^{\mathrm{CH}} + \big(x_k^{\mathrm{CBO}} - \widetilde{x}_k^{\mathrm{CH}}\big) \\ &= x_{k-1}^{\mathrm{CH}} - \tau\nabla\mathcal{E}(\widetilde{x}_k^{\mathrm{CH}}) + \big(x_k^{\mathrm{CBO}} - \widetilde{x}_k^{\mathrm{CH}}\big). \end{aligned}$$

Since $x_{k-1}^{\mathrm{CH}} = x_{k-1}^{\mathrm{CBO}} + \big(x_{k-1}^{\mathrm{CH}} - x_{k-1}^{\mathrm{CBO}}\big)$ and $\nabla\mathcal{E}(\widetilde{x}_k^{\mathrm{CH}}) = \nabla\mathcal{E}(x_{k-1}^{\mathrm{CBO}}) + \big(\nabla\mathcal{E}(\widetilde{x}_k^{\mathrm{CH}}) - \nabla\mathcal{E}(x_{k-1}^{\mathrm{CBO}})\big)$ we can continue the former to obtain

$$\begin{aligned} x_k^{\mathrm{CBO}} = x_{k-1}^{\mathrm{CBO}} &- \tau\nabla\mathcal{E}(x_{k-1}^{\mathrm{CBO}}) + \big(x_{k-1}^{\mathrm{CH}} - x_{k-1}^{\mathrm{CBO}}\big) \\ &- \tau\big(\nabla\mathcal{E}(\widetilde{x}_k^{\mathrm{CH}}) - \nabla\mathcal{E}(x_{k-1}^{\mathrm{CBO}})\big) \\ &+ \big(x_k^{\mathrm{CBO}} - \widetilde{x}_k^{\mathrm{CH}}\big), \end{aligned}$$

where it remains to control the stochastic error term g_k from (5), which is comprised of the terms $g_k^1 := x_{k-1}^{\mathrm{CH}} - x_{k-1}^{\mathrm{CBO}}$, $g_k^2 := \tau\big(\nabla\mathcal{E}(\widetilde{x}_k^{\mathrm{CH}}) - \nabla\mathcal{E}(x_{k-1}^{\mathrm{CBO}})\big)$ and $g_k^3 := x_k^{\mathrm{CBO}} - \widetilde{x}_k^{\mathrm{CH}}$. By Theorem C.1,

$$\big\|g_k^1\big\|_2 = \mathcal{O}\big(|\lambda - 1/\Delta t| + \sigma\sqrt{\Delta t} + \widetilde{\sigma} + N^{-1/2}\big)$$

with high probability. For g_k^2, first notice that $\frac{1}{2\tau}\big\|\widetilde{x}_k^{\mathrm{CH}} - x_{k-1}^{\mathrm{CH}}\big\|_2^2 + \mathcal{E}(\widetilde{x}_k^{\mathrm{CH}}) \leq \mathcal{E}(x_{k-1}^{\mathrm{CH}})$ by definition of $\widetilde{x}_k^{\mathrm{CH}}$, which facilitates a bound on $\big\|\widetilde{x}_k^{\mathrm{CH}} - x_{k-1}^{\mathrm{CH}}\big\|_2$ of order $\mathcal{O}(\tau)$ with high probability under A2 and by means of Remark D.7. Since $\mathcal{E}$ is L-smooth, with the latter derivations and Theorem C.1,

$$\begin{aligned} \big\|g_k^2\big\|_2 &\leq \tau L\,\big\|\widetilde{x}_k^{\mathrm{CH}} - x_{k-1}^{\mathrm{CBO}}\big\|_2 \\ &\leq \tau L\,\big(\big\|\widetilde{x}_k^{\mathrm{CH}} - x_{k-1}^{\mathrm{CH}}\big\|_2 + \big\|x_{k-1}^{\mathrm{CH}} - x_{k-1}^{\mathrm{CBO}}\big\|_2\big) \\ &= \mathcal{O}(\tau^2) + \mathcal{O}\big(\tau\big(|\lambda - 1/\Delta t| + \sigma\sqrt{\Delta t} + \widetilde{\sigma} + N^{-1/2}\big)\big) \end{aligned}$$

with high probability. Eventually, by Theorem C.1 and Proposition C.2 (hence, the quantitative Laplace principle (Fornasier et al., 2021b, Proposition 18), see Proposition F.2), it holds for a sufficiently large choice of α that

$$\begin{aligned}\left\|g_k^3\right\|_2 &\leq \left\|x_k^{\mathrm{CBO}} - x_k^{\mathrm{CH}}\right\|_2 + \left\|x_k^{\mathrm{CH}} - \widetilde{x}_k^{\mathrm{CH}}\right\|_2 \\ &= \mathcal{O}\big(|\lambda - 1/\Delta t| + \sigma\sqrt{\Delta t} + \widetilde{\sigma} + N^{-1/2}\big) + \mathcal{O}(\tau)\end{aligned}$$

with high probability, which concludes the proof recalling that $\widetilde{\sigma}^2 = \tau/(2\alpha)$ as of Proposition C.2. □

C.2. Technical details connecting CBO with GD via the CH scheme (10)

We now make rigorous what was described colloquially at the beginning of Appendix C. The proofs of the results below, which are the central technical tools that we utilized to prove Theorem 3.1 in Appendix C.1, are presented in Appendices E and F, respectively. $\mathcal{M}$ is the moment bound from Remark D.7.

CBO is a stochastic relaxation of CH. Theorem C.1 explains how the CBO scheme (4) can be interpreted as a stochastic relaxation of the CH scheme (10).

Theorem C.1 (CBO relaxes CH). *Fix $\varepsilon > 0$ and $\delta \in (0, 1/2)$. Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A3. We denote by $(x_k^{\mathrm{CBO}})_{k=0,\dots,K}$ the iterates of the CBO scheme (4) and by $(x_k^{\mathrm{CH}})_{k=0,\dots,K}$ the ones of the CH scheme (10). Then, with probability larger than $1 - (\delta + \varepsilon)$, it holds for all $k = 1, \dots, K$ that*

$$\left\|x_k^{\mathrm{CBO}} - x_k^{\mathrm{CH}}\right\|_2^2 \leq \varepsilon^{-1} C\big(|\lambda - 1/\Delta t|^2 + \sigma^2 \Delta t + \widetilde{\sigma}^2 + N^{-1}\big) \tag{13}$$

with $C = C(\delta^{-1}, \Delta t, d, \alpha, \lambda, \sigma, b_1, b_2, C_1, C_2, K, \mathcal{M})$.

The proof of Theorem C.1 is presented in Appendix E.4 with auxiliary results provided in Appendix E.

CH behaves like a gradient-based method. Since by definition of the iterates $\widetilde{x}_k^{\mathrm{CH}}$ in (11), it holds $\widetilde{x}_k^{\mathrm{CH}} = x_{k-1}^{\mathrm{CH}} - \tau \nabla \mathcal{E}(\widetilde{x}_k^{\mathrm{CH}})$, Proposition C.2 constitutes that (granted a sufficiently large choice of α and a suitably small choice of $\widetilde{\sigma}$) the CH scheme (10) performs a gradient step at every time step k.

Proposition C.2 (CH performs gradient steps). *Fix $\varepsilon > 0$ and $\delta \in (0, 1/2)$. Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A4. We denote by $(x_k^{\mathrm{CH}})_{k=0,\dots,K}$ the iterations of the CH scheme (10) and by $(\widetilde{x}_k^{\mathrm{CH}})_{k=0,\dots,K}$ the ones of the scheme (11). Moreover, assume that the parameters α, τ and $\widetilde{\sigma}$ are such that $\tau < 1/(-2\Lambda)$ if $\Lambda < 0$, $\alpha \gtrsim \frac{1}{\tau} d \log d$ is sufficiently large and $\widetilde{\sigma}^2 = \tau/(2\alpha)$. Then, with probability larger than $1 - (\delta + \varepsilon)$, it holds for all $k = 1, \dots, K$ that*

$$\left\|x_k^{\mathrm{CH}} - \widetilde{x}_k^{\mathrm{CH}}\right\|_2^2 \leq \varepsilon^{-1} c \tau^2 \tag{14}$$

with $c = c(\delta^{-1}, C_1, \mathcal{M})$.

The proof of Proposition C.2 is based on the quantitative Laplace principle (Fornasier et al., 2021b, Proposition 18) (see also Proposition F.2). We conjecture that a refinement thereof may allow to control the error in (14) just through α and $\widetilde{\sigma}$ without creating a dependence on τ. Nevertheless, the bound is sufficient to suggest a gradient-like behavior of the CH scheme (10) (see the discussion after Theorem 3.1).

Combining Proposition C.2 with a stability argument for the MMS and applying Grönwall's inequality allows to control in Theorem C.3 the divergence between the CH scheme (10) and the MMS (12).

Theorem C.3 (CH relaxes a gradient flow). *Fix $\varepsilon > 0$ and $\delta \in (0, 1/2)$. Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A4. We denote by $(x_k^{\mathrm{CH}})_{k=0,\dots,K}$ the iterations of the CH scheme (10) and by $(x_k^{\mathrm{MMS}})_{k=0,\dots,K}$ the ones of the MMS (12). Moreover, assume that the parameters α, τ and $\widetilde{\sigma}$ are such that $\tau < 1/(-2\Lambda)$ if $\Lambda < 0$, $\alpha \gtrsim \frac{1}{\tau} d \log d$ is sufficiently large and $\widetilde{\sigma}^2 = \tau/(2\alpha)$. Then, with probability larger than $1 - (\delta + \varepsilon)$, it holds for all $k = 1, \dots, K$ that*

$$\left\|x_k^{\mathrm{CH}} - x_k^{\mathrm{MMS}}\right\|_2^2 \leq \varepsilon^{-1} c (1 + \vartheta^{-1}) \tau^2 \sum_{\ell=0}^{k-1} \left(\frac{1+\vartheta}{(1+\tau\Lambda)^2}\right)^{\ell} \tag{15}$$

for any $\vartheta \in (0, 1)$ and with $c = c(\delta^{-1}, C_1, \mathcal{M})$.

Corollary C.4. *Fix $\varepsilon > 0$ and $\delta \in (0, 1/2)$. Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A4 with $\Lambda > 0$. Then, in the setting of Theorem C.3 and with probability larger than $1 - (\delta + \varepsilon)$, it holds for all $k = 1, \dots, K$ that*

$$\left\|x_k^{\mathrm{CH}} - x_k^{\mathrm{MMS}}\right\|_2^2 \leq \varepsilon^{-1} c(1+\vartheta^{-1})\tau^2 \frac{(1+\tau\Lambda)^2}{(1+\tau\Lambda)^2 - (1+\vartheta)}. \tag{16}$$

The proofs of Proposition C.2 and Theorem C.3 are presented in Appendices F.3 and F.4, respectively, with auxiliary results provided in Appendix F.

D. Boundedness of the numerical schemes

Before showing the boundedness in expectation of the numerical schemes (4), (10), (12) and (11) over time in Sections D.1–D.4, respectively, let us first recall from (Carrillo et al., 2018, Lemma 3.3) an estimate on the consensus point (3), which facilitates the subsequent proofs.

Lemma D.1 (Boundedness of consensus point $x_\alpha^{\mathcal{E}}$)**.** *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A3. Moreover, let $\varrho \in \mathcal{P}_2(\mathbb{R}^d)$. Then it holds*

$$\left\|x_\alpha^{\mathcal{E}}(\varrho)\right\|_2^2 \leq b_1 + b_2 \int \|x\|_2^2 \, d\varrho(x)$$

with constants $b_1 = 0$ and $b_2 = b_2(\alpha, \underline{\mathcal{E}}, \bar{\mathcal{E}}) > 0$ in case the first condition of A3 holds and with $b_i = b_i(\alpha, C_2, C_3, C_4) > 0$ for $i = 1, 2$ as given in (17) in case of the second condition of A3.

Proof. In case the first condition of A3 holds, we have by definition of the consensus point $x_\alpha^{\mathcal{E}}$ in (3) and Jensen's inequality

$$\left\|x_\alpha^{\mathcal{E}}(\varrho)\right\|_2^2 \leq \int \|x\|_2^2 \frac{\omega_\alpha^{\mathcal{E}}(x)}{\|\omega_\alpha^{\mathcal{E}}\|_{L_1(\varrho)}} d\varrho(x) \leq e^{\alpha(\bar{\mathcal{E}} - \underline{\mathcal{E}})} \int \|x\|_2^2 \, d\varrho(x).$$

In case of the second condition of A3, the statement follows from (Carrillo et al., 2018, Lemma 3.3) with constants

$$b_1 = C_4^2 + b_2 \quad \text{and} \quad b_2 = 2\frac{C_2}{C_3}\left(1 + \frac{1}{\alpha C_3}\frac{1}{C_4^2}\right), \tag{17}$$

which concludes the proof. □

With this estimate we have all necessary tools at hand to prove the boundedness of the numerical schemes investigated in this paper.

D.1. Boundedness of the consensus-based optimization (CBO) dynamics (2) and (4)

Let us remind the reader that the iterates $(x_k^{\mathrm{CBO}})_{k=0,\dots,K}$ of the consensus-based optimization (CBO) scheme (4) are defined by

$$\begin{aligned} x_k^{\mathrm{CBO}} &= x_\alpha^{\mathcal{E}}(\widehat{\rho}_k^N), \quad \text{with} \quad \widehat{\rho}_k^N = \frac{1}{N}\sum_{i=1}^N \delta_{X_k^i}, \\ x_0^{\mathrm{CBO}} &= x_0 \sim \rho_0, \end{aligned}$$

where the iterates $\left((X_k^i)_{k=0,\dots,K}\right)_{i=1,\dots,N}$ are given as in (2) by

$$\begin{aligned} X_k^i &= X_{k-1}^i - \Delta t \lambda \left(X_{k-1}^i - x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^N)\right) + \sigma D\left(X_{k-1}^i - x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^N)\right) B_k^i, \\ X_0^i &= x_0^i \sim \rho_0 \end{aligned}$$

with B_k^i being i.i.d. Gaussian random vectors in $\mathbb{R}^d$ with zero mean and covariance matrix $\Delta t \mathrm{Id}$ for $k = 0, \dots, K$ and $i = 1, \dots, N$, i.e., $B_k^i \sim \mathcal{N}(0, \Delta t \mathrm{Id})$.

Lemma D.2 (Boundedness of the CBO dynamics (2) and the CBO scheme (4)). *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A3. Moreover, let $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$. Then, for the empirical random measures $(\widehat{\rho}_k^N)_{k=0,\dots,K}$ and the iterates $(X_k^i)_{k=0,\dots,K}$ of (2) it holds*

$$\mathbb{E} \max_{k=0,\dots,K} \int \|x\|_2^4 \, d\widehat{\rho}_k^N(x) \leq \mathcal{M}^{\mathrm{CBO}} \quad \text{and} \quad \max_{i=1,\dots,N} \mathbb{E} \max_{k=0,\dots,K} \left\|X_k^i\right\|_2^4 \leq \mathcal{M}^{\mathrm{CBO}}$$

with a constant $\mathcal{M}^{\mathrm{CBO}} = \mathcal{M}^{\mathrm{CBO}}(\lambda, \sigma, d, b_1, b_2, K\Delta t, K, \rho_0) > 0$. Moreover, for the iterates $(x_k^{\mathrm{CBO}})_{k=0,\dots,K}$ of (4) it holds

$$\mathbb{E} \max_{k=0,\dots,K} \left\|x_k^{\mathrm{CBO}}\right\|_2^4 \leq \mathcal{M}^{\mathrm{CBO}}.$$

Proof. We first note that X_k^i as defined iteratively in (2) satisfies

$$X_k^i = X_0^i - \Delta t \lambda \sum_{\ell=1}^k \left(X_{\ell-1}^i - x_\alpha^{\mathcal{E}}(\widehat{\rho}_{\ell-1}^N)\right) + \sigma \sum_{\ell=1}^k D\left(X_{\ell-1}^i - x_\alpha^{\mathcal{E}}(\widehat{\rho}_{\ell-1}^N)\right) B_\ell^i$$

and that for any $k = 1, \dots, K$ by means of the standard inequality (8) for $p = 4$ and $J = 3$ we have

$$\begin{aligned} \max_{\ell=0,\dots,k} \left\|X_\ell^i\right\|_2^4 \lesssim \left\|X_0^i\right\|_2^4 + (\Delta t \lambda)^4 \max_{\ell=1,\dots,k} \left\| \sum_{m=1}^\ell \left(X_{m-1}^i - x_\alpha^{\mathcal{E}}(\widehat{\rho}_{m-1}^N)\right) \right\|_2^4 \\ + \sigma^4 \max_{\ell=1,\dots,k} \left\| \sum_{m=1}^\ell D\left(X_{m-1}^i - x_\alpha^{\mathcal{E}}(\widehat{\rho}_{m-1}^N)\right) B_m^i \right\|_2^4. \end{aligned} \tag{18}$$

Noticing that the random process $Y_\ell^i := \sum_{m=1}^\ell D\left(X_{m-1}^i - x_\alpha^{\mathcal{E}}(\widehat{\rho}_{m-1}^N)\right) B_m^i$, $\ell = 0, \dots, k$ is a martingale w.r.t. the filtration $\left\{\mathcal{F}_\ell = \sigma\left(\{X_0^i\} \cup \{B_m^i, m = 1, \dots, \ell\}\right)\right\}_{\ell=0}^{k-1}$ since it satisfies $\mathbb{E}\left[Y_\ell^i \mid \mathcal{F}_{\ell-1}\right] = Y_{\ell-1}^i$ for $\ell = 1, \dots, k$, we can apply a discrete version of the Burkholder-Davis-Gundy inequality (Chow & Teicher, 1997, Corollary 11.2.1) yielding

$$\mathbb{E} \max_{\ell=1,\dots,k} \left\| \sum_{m=1}^\ell D\left(X_{m-1}^i - x_\alpha^{\mathcal{E}}(\widehat{\rho}_{m-1}^N)\right) B_m^i \right\|_2^4 \lesssim d\, \mathbb{E} \sum_{j=1}^d \left(\sum_{\ell=1}^k \left(D\left(X_{\ell-1}^i - x_\alpha^{\mathcal{E}}(\widehat{\rho}_{\ell-1}^N)\right)\right)_{jj}^2 (B_\ell^i)_j^2 \right)^2.$$

Thus, when taking the expectation on both sides of (18) and employing Jensen's inequality, we can use the latter to obtain

$$\begin{aligned} \mathbb{E} \max_{\ell=0,\dots,k} \left\|X_\ell^i\right\|_2^4 &\lesssim \mathbb{E} \left\|X_0^i\right\|_2^4 + (\Delta t \lambda)^4 K^3\, \mathbb{E} \sum_{\ell=1}^k \left\|X_{\ell-1}^i - x_\alpha^{\mathcal{E}}(\widehat{\rho}_{\ell-1}^N)\right\|_2^4 \\ &\quad + \sigma^4 dK\, \mathbb{E} \sum_{j=1}^d \sum_{\ell=1}^k \left(D\left(X_{\ell-1}^i - x_\alpha^{\mathcal{E}}(\widehat{\rho}_{\ell-1}^N)\right)\right)_{jj}^4 (B_\ell^i)_j^4 \\ &\lesssim \mathbb{E} \left\|X_0^i\right\|_2^4 + (\Delta t \lambda)^4 K^3\, \mathbb{E} \sum_{\ell=1}^k \left(\left\|X_{\ell-1}^i\right\|_2^4 + \left\|x_\alpha^{\mathcal{E}}(\widehat{\rho}_{\ell-1}^N)\right\|_2^4\right) \\ &\quad + (\Delta t)^2 \sigma^4 dK\, \mathbb{E} \sum_{j=1}^d \sum_{\ell=1}^k \left(\left(X_{\ell-1}^i\right)_j^4 + \left(x_\alpha^{\mathcal{E}}(\widehat{\rho}_{\ell-1}^N)\right)_j^4\right) \\ &\lesssim \left(1 + (\Delta t \lambda)^4 K^3 + (\Delta t \sigma^2 d)^2 K\right) \mathbb{E} \sum_{\ell=1}^k \left(\left\|X_{\ell-1}^i\right\|_2^4 + \left\|x_\alpha^{\mathcal{E}}(\widehat{\rho}_{\ell-1}^N)\right\|_2^4\right) \\ &\lesssim \left(1 + \lambda^4 (K\Delta t)^4 + \sigma^4 d^2 (K\Delta t)^2\right) \mathbb{E} \max_{\ell=1,\dots,k} \left(\left\|X_{\ell-1}^i\right\|_2^4 + \left\|x_\alpha^{\mathcal{E}}(\widehat{\rho}_{\ell-1}^N)\right\|_2^4\right) \\ &\leq C\, \mathbb{E} \max_{\ell=1,\dots,k} \left(\left\|X_{\ell-1}^i\right\|_2^4 + b_1^2 + b_2^2 \int \|x\|_2^4 \, d\widehat{\rho}_{\ell-1}^N(x)\right) \end{aligned} \tag{19}$$

with a constant $C = C(\lambda, \sigma, d, K\Delta t)$. In the second step we made use of the standard inequality (8) for $p = 4$ and $J = 2$, exploited that B_ℓ^i is independent from $D\left(X_{\ell-1}^i - x_\alpha^{\mathcal{E}}(\widehat{\rho}_{\ell-1}^N)\right)$ for any $\ell = 1, \dots, k$ and used that the fourth moment of a

Gaussian random variable $B \sim \mathcal{N}(0,1)$ is $\mathbb{E}B^4 = 3$ (e.g., by recalling that $\mathbb{E}B^4 = \frac{d^4}{dx^4} M_B(x)\big|_{x=0}$, where M_B denotes the moment-generating function of B). Moreover, recall that $K\Delta t$ denotes the final time horizon, and note that the last step is due to Lemma D.1. Averaging (19) over i allows to bound

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{E} \max_{\ell=0,\dots,k} \left\|X_\ell^i\right\|_2^4 \leq \widetilde{C}\left(1 + \frac{1}{N}\sum_{i=1}^{N} \mathbb{E} \max_{\ell=1,\dots,k} \left\|X_{\ell-1}^i\right\|_2^4\right) \tag{20}$$

with a constant $\widetilde{C} = \widetilde{C}(\lambda, \sigma, d, b_1, b_2, K\Delta t)$. Since $\mathbb{E}\int \|x\|_2^4 \, d\widehat{\rho}_0^N(x) = \frac{1}{N}\sum_{i=1}^N \mathbb{E}\,\|x_0^i\|_2^4$, an application of the discrete variant of Grönwall's inequality (9) yields the second inequality in

$$\begin{aligned}\mathbb{E} \max_{\ell=0,\dots,k} \int \|x\|_2^4 \, d\widehat{\rho}_\ell^N(x) &\leq \frac{1}{N}\sum_{i=1}^{N} \mathbb{E} \max_{\ell=0,\dots,k} \left\|X_\ell^i\right\|_2^4 \\ &\leq \widetilde{C}^k \, \mathbb{E}\int \|x\|_2^4 \, d\widehat{\rho}_0^N(x) + \widetilde{C} e^{\widetilde{C}(k-1)},\end{aligned} \tag{21}$$

showing that the left-hand side is bounded independently of N, which gives the first bound in the first part of the statement. Making use thereof in (19) also yields the second part after another application of Grönwall's inequality. The second part of the statement follows by noting that an application of Lemma D.1 gives

$$\begin{aligned}\mathbb{E} \max_{\ell=1,\dots,k} \left\|x_\ell^{\mathrm{CBO}}\right\|_2^4 &= \mathbb{E} \max_{\ell=1,\dots,k} \left\|x_\alpha^{\mathcal{E}}(\widehat{\rho}_\ell^N)\right\|_2^4 \\ &\leq 2b_1^2 + 2b_2^2 \, \mathbb{E} \max_{\ell=1,\dots,k} \int \|x\|_2^4 \, d\widehat{\rho}_\ell^N(x),\end{aligned}$$

where the last expression is bounded as in (21). Recalling that $x_0^{\mathrm{CBO}} = x_0 \sim \rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$ and choosing the constant $\mathcal{M}^{\mathrm{CBO}}$ large enough for all three estimates to hold with $k = K$ concludes the proof. □

D.2. Boundedness of the consensus hopping scheme (10)

Let us recall that the iterates $(x_k^{\mathrm{CH}})_{k=0,\dots,K}$ of the consensus hopping (CH) scheme (10) are defined by

$$\begin{aligned}&x_k^{\mathrm{CH}} = x_\alpha^{\mathcal{E}}(\mu_k), \quad \text{with} \quad \mu_k = \mathcal{N}\big(x_{k-1}^{\mathrm{CH}}, \widetilde{\sigma}^2 \mathrm{Id}\big), \\ &x_0^{\mathrm{CH}} = x_0.\end{aligned}$$

Lemma D.3 (Boundedness of the CH scheme (10))**.** *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A3. Moreover, let $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$. Then, for the random measures $(\mu_k)_{k=1,\dots,K}$ in (10) it holds*

$$\mathbb{E} \max_{k=1,\dots,K} \int \|x\|_2^4 \, d\mu_k(x) \leq \mathcal{M}^{\mathrm{CH}}$$

with a constant $\mathcal{M}^{\mathrm{CH}} = \mathcal{M}^{\mathrm{CH}}(\widetilde{\sigma}, d, b_1, b_2, K, \rho_0) > 0$. Moreover, for the iterates $(x_k^{\mathrm{CH}})_{k=0,\dots,K}$ of (10) it holds

$$\mathbb{E} \max_{k=0,\dots,K} \left\|x_k^{\mathrm{CH}}\right\|_2^4 \leq \mathcal{M}^{\mathrm{CH}}.$$

Proof. According to the definition of the scheme (10) and with the standard inequality (8) for $p = 4$ and $J = 2$, we observe that for any $k = 2, \dots, K$ it holds

$$\begin{aligned}\int \|x\|_2^4 \, d\mu_k(x) &= \int \|x\|_2^4 \, d\mathcal{N}\big(x_{k-1}^{\mathrm{CH}}, \widetilde{\sigma}^2 \mathrm{Id}\big)(x) \\ &\lesssim \left\|x_{k-1}^{\mathrm{CH}}\right\|_2^4 + \int \|x\|_2^4 \, d\mathcal{N}\big(0, \widetilde{\sigma}^2 \mathrm{Id}\big)(x) \\ &= \left\|x_\alpha^{\mathcal{E}}(\mu_{k-1})\right\|_2^4 + (d^2 + 2d)\,\widetilde{\sigma}^4 \\ &\lesssim b_1^2 + b_2^2 \int \|x\|_2^4 \, d\mu_{k-1}(x) + d^2\widetilde{\sigma}^4,\end{aligned}$$

where for the third step we explicitly computed that for the fourth moment of a multivariate Gaussian distribution it holds $\int \|x\|_2^4 \, d\mathcal{N}(0, \mathrm{Id})(x) = d^2 + 2d$. Moreover, in the final step we employed Lemma D.1 together with Jensen's inequality. Along the same lines we have $\int \|x\|_2^4 \, d\mu_1(x) \lesssim \|x_0\|_2^4 + d^2\widetilde{\sigma}^4$. An application of the discrete variant of Grönwall's inequality (9) therefore allows to obtain

$$\int \|x\|_2^4 \, d\mu_k(x) \lesssim b_2^{2k} \|x_0\|_2^4 + \left(b_1^2 + d^2\widetilde{\sigma}^4\right) e^{cb_2^2(k-1)}$$

with a generic constant $c > 0$. Taking the maximum over the iterations k and the expectation w.r.t. the initial condition ρ_0 gives the first part of the statement. Recalling that $x_0^{\mathrm{CH}} = x_0 \sim \rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$, the second part follows after an application of Lemma D.1, since

$$\begin{aligned}\mathbb{E} \max_{\ell=1,\dots,k} \left\|x_\ell^{\mathrm{CH}}\right\|_2^4 &= \mathbb{E} \max_{\ell=1,\dots,k} \left\|x_\alpha^{\mathcal{E}}(\mu_\ell)\right\|_2^4 \\ &\leq 2b_1^2 + 2b_2^2 \, \mathbb{E} \max_{\ell=1,\dots,k} \int \|x\|_2^4 \, d\mu_\ell(x).\end{aligned}$$

Choosing the constant $\mathcal{M}^{\mathrm{CH}}$ large enough for either estimate to hold with $k = K$ concludes the proof. □

Lemma D.4. *Let $Y_k^i \sim \mu_k$ for $i = 1, \dots, N$ and let $\widehat{\mu}_k^N = \frac{1}{N}\sum_{i=1}^N \delta_{Y_k^i}$. Then, under the assumptions of Lemma D.3, for the empirical random measures $(\widehat{\mu}_k^N)_{k=1,\dots,K}$ it holds*

$$\mathbb{E} \max_{k=1,\dots,K} \int \|x\|_2^4 \, d\widehat{\mu}_k^N(x) \leq \widehat{\mathcal{M}}^{\mathrm{CH}}$$

with a constant $\widehat{\mathcal{M}}^{\mathrm{CH}} = \widehat{\mathcal{M}}^{\mathrm{CH}}(\widetilde{\sigma}, d, b_1, b_2, K, \rho_0) > 0$.

Proof. By definition of the empirical measure $\widehat{\mu}_k^N$ it holds

$$\mathbb{E} \max_{k=1,\dots,K} \int \|x\|_2^4 \, d\widehat{\mu}_k^N(x) = \mathbb{E} \max_{k=1,\dots,K} \frac{1}{N}\sum_{i=1}^N \left\|Y_k^i\right\|_2^4 \leq \frac{1}{N}\sum_{i=1}^N \mathbb{E} \max_{k=1,\dots,K} \left\|Y_k^i\right\|_2^4. \tag{22}$$

Since $Y_k^i \sim \mu_k = \mathcal{N}\left(x_{k-1}^{\mathrm{CH}}, \widetilde{\sigma}^2\mathrm{Id}\right)$ for any $k = 1, \dots, K$ and $i = 1, \dots, N$, we can write $Y_k^i = x_{k-1}^{\mathrm{CH}} + \widetilde{\sigma} B_{Y,k}^i$, where $B_{Y,k}^i$ is a standard Gaussian random vector, i.e., $B_{Y,k}^i \sim \mathcal{N}(0, \mathrm{Id})$. By means of the standard inequality (8) for $p = 4$ and $J = 2$ we thus have

$$\begin{aligned}\mathbb{E} \max_{k=1,\dots,K} \left\|Y_k^i\right\|_2^4 &\lesssim \mathbb{E} \max_{k=1,\dots,K} \left\|x_{k-1}^{\mathrm{CH}}\right\|_2^4 + \widetilde{\sigma}^4 \mathbb{E} \max_{k=1,\dots,K} \left\|B_{Y,k}^i\right\|_2^4 \\ &\leq \mathcal{M}^{\mathrm{CH}} + K\widetilde{\sigma}^4(d^2 + 2d),\end{aligned} \tag{23}$$

where in the last step we employed Lemma D.3 for the first term and bounded the maximum by the sum in the second term before using again that $\mathbb{E}\|B\|_2^4 = d^2 + 2d$ for $B \sim \mathcal{N}(0, \mathrm{Id})$. Inserting (23) into (22) yields the claim. □

D.3. Boundedness of the minimizing movement scheme (12)

We recall that the iterates $(x_k^{\mathrm{MMS}})_{k=0,\dots,K}$ of the minimizing movement scheme (MMS) (12) are defined by

$$\begin{aligned}x_k^{\mathrm{MMS}} &= \operatorname*{arg\,min}_{x \in \mathbb{R}^d} \mathcal{E}_k(x), \quad \text{with} \quad \mathcal{E}_k(x) := \frac{1}{2\tau}\left\|x_{k-1}^{\mathrm{MMS}} - x\right\|_2^2 + \mathcal{E}(x), \\ x_0^{\mathrm{MMS}} &= x_0.\end{aligned}$$

Lemma D.5 (Boundedness of the MMS (12))**.** *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A2. Moreover, let $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$. Then, for the iterates $(x_k^{\mathrm{MMS}})_{k=0,\dots,K}$ of (12) it holds*

$$\mathbb{E} \max_{k=0,\dots,K} \left\|x_k^{\mathrm{MMS}}\right\|_2^4 \leq \mathcal{M}^{\mathrm{MMS}}$$

with a constant $\mathcal{M}^{\mathrm{MMS}} = \mathcal{M}^{\mathrm{MMS}}(K\tau, C_2, \rho_0) > 0$.

Proof. Since x_k^{MMS} is the minimizer of $\mathcal{E}_k$, see (12), a comparison with the old iterate x_{k-1}^{MMS} yields

$$\frac{1}{2\tau}\left\|x_{k-1}^{\mathrm{MMS}} - x_k^{\mathrm{MMS}}\right\|_2^2 + \mathcal{E}(x_k^{\mathrm{MMS}}) \leq \mathcal{E}(x_{k-1}^{\mathrm{MMS}})$$

for any $k = 1, \dots, K$. Using the standard inequality (8) for $p = 2$ and $J = k$, this can be utilized to obtain

$$\begin{aligned}
\left\|x_k^{\mathrm{MMS}}\right\|_2^2 &\leq 2\left\|x_0^{\mathrm{MMS}}\right\|_2^2 + 2K\sum_{\ell=1}^{k}\left\|x_\ell^{\mathrm{MMS}} - x_{\ell-1}^{\mathrm{MMS}}\right\|_2^2 \\
&\leq 2\left\|x_0^{\mathrm{MMS}}\right\|_2^2 + 4K\tau\sum_{\ell=1}^{k}\left(\mathcal{E}(x_{\ell-1}^{\mathrm{MMS}}) - \mathcal{E}(x_\ell^{\mathrm{MMS}})\right) \\
&= 2\left\|x_0^{\mathrm{MMS}}\right\|_2^2 + 4K\tau\left(\mathcal{E}(x_0^{\mathrm{MMS}}) - \mathcal{E}(x_k^{\mathrm{MMS}})\right) \\
&\leq 2\left\|x_0\right\|_2^2 + 4K\tau\left(\mathcal{E}(x_0) - \underline{\mathcal{E}}\right) \\
&\leq 2\left\|x_0\right\|_2^2 + 4K\tau C_2(1 + \left\|x_0\right\|_2^2) \\
&= 2\left(1 + 2K\tau C_2\right)\left\|x_0\right\|_2^2 + 4K\tau C_2,
\end{aligned}$$

which trivially also holds for $k = 0$. Taking the square and expectation w.r.t. the initial condition ρ_0 on both sides concludes the proof. □

D.4. Boundedness of the implicit version of the CH scheme (11)

Let us recall that the iterates $(\widetilde{x}_k^{\mathrm{CH}})_{k=0,\dots,K}$ of the scheme (11) are defined by

$$\begin{aligned}
&\widetilde{x}_k^{\mathrm{CH}} = \operatorname*{arg\,min}_{x\in\mathbb{R}^d} \widetilde{\mathcal{E}}_k(x), \quad \text{with} \quad \widetilde{\mathcal{E}}_k(x) := \frac{1}{2\tau}\left\|x_{k-1}^{\mathrm{CH}} - x\right\|_2^2 + \mathcal{E}(x), \\
&\widetilde{x}_0^{\mathrm{CH}} = x_0.
\end{aligned}$$

Lemma D.6 (Boundedness of the implicit version of the CH scheme (11))**.** *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A3. Moreover, let $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$. Then, for the iterates $(\widetilde{x}_k^{\mathrm{CH}})_{k=0,\dots,K}$ of (11) it holds*

$$\mathbb{E}\max_{k=0,\dots,K}\left\|\widetilde{x}_k^{\mathrm{CH}}\right\|_2^4 \leq \widetilde{\mathcal{M}}^{\mathrm{CH}}$$

with a constant $\widetilde{\mathcal{M}}^{\mathrm{CH}} = \widetilde{\mathcal{M}}^{\mathrm{CH}}(\tau, C_2, \mathcal{M}^{\mathrm{CH}}) > 0$.

Proof. Since $\widetilde{x}_k^{\mathrm{CH}}$ is the minimizer of $\widetilde{\mathcal{E}}_k$, see (11), a comparison with x_{k-1}^{CH} yields

$$\frac{1}{2\tau}\left\|x_{k-1}^{\mathrm{CH}} - \widetilde{x}_k^{\mathrm{CH}}\right\|_2^2 + \mathcal{E}(\widetilde{x}_k^{\mathrm{CH}}) \leq \mathcal{E}(x_{k-1}^{\mathrm{CH}}).$$

This can be utilized to obtain

$$\begin{aligned}
\left\|\widetilde{x}_k^{\mathrm{CH}}\right\|_2^2 &= 2\left\|\widetilde{x}_k^{\mathrm{CH}} - x_{k-1}^{\mathrm{CH}}\right\|_2^2 + 2\left\|x_{k-1}^{\mathrm{CH}}\right\|_2^2 \\
&\leq 4\tau\left(\mathcal{E}(x_{k-1}^{\mathrm{CH}}) - \mathcal{E}(\widetilde{x}_k^{\mathrm{CH}})\right) + 2\left\|x_{k-1}^{\mathrm{CH}}\right\|_2^2 \\
&\leq 4\tau\left(\mathcal{E}(x_{k-1}^{\mathrm{CH}}) - \underline{\mathcal{E}}\right) + 2\left\|x_{k-1}^{\mathrm{CH}}\right\|_2^2 \\
&\leq 4\tau C_2\left(1 + \left\|x_{k-1}^{\mathrm{CH}}\right\|_2^2\right) + 2\left\|x_{k-1}^{\mathrm{CH}}\right\|_2^2 \\
&= 2\left(1 + 2\tau C_2\right)\left\|x_{k-1}^{\mathrm{CH}}\right\|_2^2 + 4\tau C_2.
\end{aligned}$$

Taking the square and expectation w.r.t. the initial condition ρ_0 on both sides concludes the proof by virtue of Lemma D.3. □

D.5. Boundedness of all numerical schmemes

Remark D.7 (Boundedness of the schemes (4), (10), (11) and (12)). To keep the notation of the main body of the paper concise, we denote by $\mathcal{M}$ the collective moment bound

$$\mathcal{M} = \max\left\{\mathcal{M}^{\mathrm{CBO}}, \widetilde{\mathcal{M}}^{\mathrm{CBO}}, \mathcal{M}^{\mathrm{CH}}, \widehat{\mathcal{M}}^{\mathrm{CH}}, \widehat{\mathcal{M}}^{\mathrm{MMS}}, \widetilde{\mathcal{M}}^{\mathrm{CH}}\right\}, \tag{24}$$

where $\mathcal{M}^{\mathrm{CBO}}$, $\mathcal{M}^{\mathrm{CH}}$, $\widehat{\mathcal{M}}^{\mathrm{CH}}$,$\widehat{\mathcal{M}}^{\mathrm{MMS}}$, and $\widetilde{\mathcal{M}}^{\mathrm{CH}}$ are as defined in Lemmas D.2, D.3, D.4, D.5, and D.6, respectively. Moreover, $\widetilde{\mathcal{M}}^{\mathrm{CBO}} = \mathcal{M}^{\mathrm{CBO}}(1/\Delta t, \sigma, d, b_1, b_2, K\Delta t, K, \rho_0)$.

E. Proof details for Theorem C.1

Theorem C.1 is centered around the observation that, as $\lambda \to 1/\Delta t$ in the CBO dynamics (2), the CBO scheme (4) resembles an implementation of the CH scheme (10) via sampling from the underlying distribution μ_k and computing the associated weighted empirical average. Accordingly, the proof of Theorem C.1 consists of three ingredients. First, a stability estimate for the CBO dynamics (2) w.r.t. the parameter λ, see Lemma E.2. Second, a quantification of the structural difference in the noise component between the CBO scheme (4) and the CH scheme (10), and third a large deviation bound to control the sampling error associated with the Monte Carlo approximation of the CH scheme (10), see Lemma E.3.

E.1. Stability of the consensus point (3) w.r.t. the underlying measure

We first recall from (Carrillo et al., 2018, Lemma 3.2) in a slightly modified form a stability estimate for the consensus point (3) w.r.t. the measure from which it is computed. Loosely speaking, we show that the mapping $x_\alpha^{\mathcal{E}} : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^d$ is Lipschitz-continuous in the Wasserstein-2 metric.

Lemma E.1 (Stability of the consensus point $x_\alpha^{\mathcal{E}}$)**.** *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A2. Moreover, let $\varrho, \varrho' \in \mathcal{P}(\mathbb{R}^d)$ be random measures and define the cutoff function (random variable)*

$$\bar{\mathcal{I}}_M^1 = \begin{cases} 1, & \text{if } \max\left\{\int \|\bullet\|_2^4 \, d\varrho, \int \|\bullet\|_2^4 \, d\varrho'\right\} \leq M^4, \\ 0, & \text{else.} \end{cases}$$

Then it holds

$$\left\|x_\alpha^{\mathcal{E}}(\varrho) - x_\alpha^{\mathcal{E}}(\varrho')\right\|_2 \bar{\mathcal{I}}_M^1 \leq c_0 W_2(\varrho, \varrho') \bar{\mathcal{I}}_M^1$$

with a constant $c_0 = c_0(\alpha, C_1, C_2, M) > 0$.

Proof. To start with, we note that under A2 and with Jensen's inequality it holds

$$\begin{aligned} \frac{e^{-\alpha \underline{\mathcal{E}}} \, \bar{\mathcal{I}}_M^1}{\|\omega_\alpha^{\mathcal{E}}\|_{L_1(\varrho)}} &= \frac{\bar{\mathcal{I}}_M^1}{\int \exp\left(-\alpha(\mathcal{E}(x) - \underline{\mathcal{E}})\right) d\varrho(x)} \leq \frac{\bar{\mathcal{I}}_M^1}{\int \exp\big(-\alpha C_2(1 + \|x\|_2^2)\big) \, d\varrho(x)} \\ &\leq \frac{\bar{\mathcal{I}}_M^1}{\exp\big(-\alpha C_2(1 + \int \|x\|_2^2 \, d\varrho(x))\big)} \leq \exp\big(\alpha C_2(1 + M^2)\big) =: c_M. \end{aligned} \tag{25}$$

An analogous statement can be obtained for the measure ϱ'.

By definition of the consensus point $x_\alpha^{\mathcal{E}}$ in (3), it holds for any coupling $\gamma \in \Pi(\varrho, \varrho')$ between ϱ and ϱ' by Jensen's inequality

$$\begin{aligned} \left\|x_\alpha^{\mathcal{E}}(\varrho) - x_\alpha^{\mathcal{E}}(\varrho')\right\|_2 \bar{\mathcal{I}}_M^1 &\leq \iint \left\| x \frac{\omega_\alpha^{\mathcal{E}}(x)}{\|\omega_\alpha^{\mathcal{E}}\|_{L_1(\varrho)}} - x' \frac{\omega_\alpha^{\mathcal{E}}(x')}{\|\omega_\alpha^{\mathcal{E}}\|_{L_1(\varrho')}} \right\|_2 d\gamma(x, x') \, \bar{\mathcal{I}}_M^1 \\ &\leq \iint \left(\|T_1(x, x')\|_2 + \|T_2(x, x')\|_2 + \|T_3(x, x')\|_2\right) d\gamma(x, x') \, \bar{\mathcal{I}}_M^1, \end{aligned} \tag{26}$$

where the terms T_1, T_2 and T_3 are defined implicitly and bounded as follows. For the first term T_1 we have

$$\|T_1(x, x')\|_2 \bar{\mathcal{I}}_M^1 = \|x - x'\|_2 \frac{\omega_\alpha^{\mathcal{E}}(x)}{\|\omega_\alpha^{\mathcal{E}}\|_{L_1(\varrho)}} \bar{\mathcal{I}}_M^1 \leq c_M \|x - x'\|_2 \bar{\mathcal{I}}_M^1, \tag{27}$$

where we utilized (25) in the last step. For the second term T_2, with A2 and again (25) we obtain

$$\begin{aligned}\|T_2(x,x')\|_2 \bar{\mathcal{I}}_M^1 &= \|x'\|_2 \frac{\left|\omega_\alpha^{\mathcal{E}}(x) - \omega_\alpha^{\mathcal{E}}(x')\right|}{\|\omega_\alpha^{\mathcal{E}}\|_{L_1(\varrho)}} \bar{\mathcal{I}}_M^1 \\ &\leq \|x'\|_2 \frac{\alpha e^{-\alpha \mathcal{E}} C_1 (1 + \|x\|_2 + \|x'\|_2) \|x - x'\|_2}{\|\omega_\alpha^{\mathcal{E}}\|_{L_1(\varrho)}} \bar{\mathcal{I}}_M^1 \\ &\leq \alpha c_M C_1 \|x'\|_2 (1 + \|x\|_2 + \|x'\|_2) \|x - x'\|_2 \bar{\mathcal{I}}_M^1.\end{aligned} \tag{28}$$

Eventually, for the third therm T_3 it holds by following similar steps

$$\begin{aligned}\|T_3(x,x')\|_2 \bar{\mathcal{I}}_M^1 &= \|x'\|_2 \omega_\alpha^{\mathcal{E}}(x') \frac{\left| \|\omega_\alpha^{\mathcal{E}}\|_{L_1(\varrho')} - \|\omega_\alpha^{\mathcal{E}}\|_{L_1(\varrho)} \right|}{\|\omega_\alpha^{\mathcal{E}}\|_{L_1(\varrho)} \|\omega_\alpha^{\mathcal{E}}\|_{L_1(\varrho')}} \bar{\mathcal{I}}_M^1 \\ &\leq c_M \|x'\|_2 \frac{\iint \alpha e^{-\alpha \mathcal{E}} C_1 (1 + \|x\|_2 + \|x'\|_2) \|x - x'\|_2 \, d\pi(x,x')}{\|\omega_\alpha^{\mathcal{E}}\|_{L_1(\varrho)}} \bar{\mathcal{I}}_M^1 \\ &\leq \alpha c_M^2 C_1 \|x'\|_2 \iint (1 + \|x\|_2 + \|x'\|_2) \|x - x'\|_2 \, d\pi(x,x') \, \bar{\mathcal{I}}_M^1.\end{aligned} \tag{29}$$

Collecting the estimates (27) – (29) in (26), we obtain with Cauchy-Schwarz inequality and by exploiting the definition of $\bar{\mathcal{I}}_M^1$ that

$$\left\| x_\alpha^{\mathcal{E}}(\varrho) - x_\alpha^{\mathcal{E}}(\varrho') \right\|_2 \bar{\mathcal{I}}_M^1 \leq c_M \left(1 + 3\alpha C_1 (1 + c_M) M (1 + 3M)\right) \sqrt{\iint \|x - x'\|_2^2 \, d\gamma(x,x') \, \bar{\mathcal{I}}_M^1}. \tag{30}$$

Squaring both sides and optimizing over all couplings $\gamma \in \Pi(\varrho, \varrho')$ concludes the proof. □

E.2. Stability of the CBO dynamics (2) w.r.t. the parameters λ and σ

Let us now show the stability of the CBO dynamics (2) w.r.t. its parameters, in particular, the drift and noise parameters λ and σ. For this we control in Lemma E.2 below the mismatch of the iterates of the CBO dynamics (2) for different parameters, however, provided coinciding initialization and discrete Brownian motion paths.

Lemma E.2 (Stability of the CBO dynamics (2))**.** *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1 – A3. Moreover, let $\rho_0 \in \mathcal{P}_4(\mathbb{R}^d)$. We denote by $\left((X_k^{i,1})_{k=0,\dots,K}\right)_{i=1,\dots,N}$ and $\left((X_k^{i,2})_{k=0,\dots,K}\right)_{i=1,\dots,N}$ solutions to (2) with parameters λ_1, σ_1 and λ_2, σ_2, respectively. Furthermore, we write $(\widehat{\rho}_k^{N,1})_{k=0,\dots,K}$ and $(\widehat{\rho}_k^{N,2})_{k=0,\dots,K}$ for the associated empirical measures and introduce the cutoff function (random variable)*

$$\bar{\mathcal{I}}_{M,k}^1 = \begin{cases} 1, & \text{if } \max\left\{ \int \|\bullet\|_2^4 \, d\widehat{\rho}_k^{N,1}, \int \|\bullet\|_2^4 \, d\widehat{\rho}_k^{N,2} \right\} \leq M^4, \\ 0, & \text{else.} \end{cases} \tag{31}$$

Then, under the assumption of coinciding initial conditions $X_0^{i,1} = X_0^{i,2}$ for all $i = 1, \dots, N$ as well as Gaussian random vectors B_k^i for all $k = 1, \dots, K$ and all $i = 1, \dots, N$, it holds

$$\frac{1}{N} \sum_{i=1}^N \mathbb{E} \left\| X_k^{i,1} - X_k^{i,2} \right\|_2^2 \bar{\mathcal{I}}_{M,k}^1 \leq c_1 \left(|\lambda_1 - \lambda_2|^2 + |\sigma_1 - \sigma_2|^2 \right) e^{c_2 (k-1)}$$

with constants $c_1 = c_1(\Delta t, d, b_1, b_2, M) > 0$ and $c_2 = c_2(\Delta t, d, \alpha, \lambda_2, \sigma_2, C_1, C_2, M) > 0$ for all $k \geq 1$.

Proof. Let us first remark that the cutoff function $\bar{\mathcal{I}}_{M,k}^1$ defined in (31) is adapted to the natural filtration $\{\mathcal{F}_k\}_{k=0,\dots,K}$, where $\mathcal{F}_k$ denotes the sigma algebra generated by $\{B_\ell^i,\ \ell = 1, \dots, k,\ i = 1, \dots, N\}$. Now, using the iterative update rule (2) for $X_k^{i,1}$ and $X_k^{i,2}$ with parameters λ_1, σ_1 and λ_2, σ_2, respectively, we obtain, by employing the standard inequality (8) for

$p = 2$ and $J = 5$, for their squared norm difference the upper bound

$$\begin{aligned}\left\|X_k^{i,1} - X_k^{i,2}\right\|_2^2 \lesssim{}& \left\|X_{k-1}^{i,1} - X_{k-1}^{i,2}\right\|_2^2 + (\Delta t\,|\lambda_1-\lambda_2|)^2 \left(\left\|X_{k-1}^{i,1}\right\|_2^2 + \left\|x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,1})\right\|_2^2\right)\\ &+ (\Delta t\lambda_2)^2 \left(\left\|X_{k-1}^{i,1} - X_{k-1}^{i,2}\right\|_2^2 + \left\|x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,1}) - x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,2})\right\|_2^2\right)\\ &+ |\sigma_1-\sigma_2|^2 \left(\left\|X_{k-1}^{i,1}\right\|_2^2 + \left\|x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,1})\right\|_2^2\right) \left\|B_k^i\right\|_2^2\\ &+ \sigma_2^2 \left(\left\|X_{k-1}^{i,1} - X_{k-1}^{i,2}\right\|_2^2 + \left\|x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,1}) - x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,2})\right\|_2^2\right) \left\|B_k^i\right\|_2^2\\ \lesssim{}& \left(1+(\Delta t\lambda_2)^2+\sigma_2^2\left\|B_k^i\right\|_2^2\right) \left(\left\|X_{k-1}^{i,1}-X_{k-1}^{i,2}\right\|_2^2+\left\|x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,1})-x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,2})\right\|_2^2\right)\\ &+ \left((\Delta t\,|\lambda_1-\lambda_2|)^2 + |\sigma_1-\sigma_2|^2\left\|B_k^i\right\|_2^2\right) \left(\left\|X_{k-1}^{i,1}\right\|_2^2+\left\|x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,1})\right\|_2^2\right).\end{aligned} \tag{32}$$

Since $\bar{\mathcal{I}}_{M,k}^1$ satisfies $\bar{\mathcal{I}}_{M,k}^1 = \bar{\mathcal{I}}_{M,k}^1\bar{\mathcal{I}}_{M,\ell}^1$ for all $\ell \le k$ and $\bar{\mathcal{I}}_{M,k}^1 \le 1$, we obtain from (32) that

$$\begin{aligned}&\left\|X_k^{i,1} - X_k^{i,2}\right\|_2^2 \bar{\mathcal{I}}_{M,k}^1\\ &\lesssim \left(1+(\Delta t\lambda_2)^2+\sigma_2^2\left\|B_k^i\right\|_2^2\right) \left(\left\|X_{k-1}^{i,1} - X_{k-1}^{i,2}\right\|_2^2 + \left\|x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,1}) - x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,2})\right\|_2^2\right) \bar{\mathcal{I}}_{M,k-1}^1\\ &\quad + \left((\Delta t\,|\lambda_1-\lambda_2|)^2 + |\sigma_1-\sigma_2|^2\left\|B_k^i\right\|_2^2\right) \left(\left\|X_{k-1}^{i,1}\right\|_2^2 + \left\|x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,1})\right\|_2^2\right) \bar{\mathcal{I}}_{M,k-1}^1.\end{aligned}$$

With the random variables $X_{k-1}^{i,1}$, $X_{k-1}^{i,2}$, $x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,1})$, $x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,2})$ and $\bar{\mathcal{I}}_{M,k-1}^1$ being $\mathcal{F}_{k-1}$-measurable, taking the expectation w.r.t. the sampling of the random vectors B_k^i, $i = 1, \dots, N$, i.e., the conditional expectation $\mathbb{E}_k = \mathbb{E}[\,\bullet\,|\,\mathcal{F}_{k-1}]$, yields

$$\begin{aligned}&\mathbb{E}_k\left\|X_k^{i,1} - X_k^{i,2}\right\|_2^2 \bar{\mathcal{I}}_{M,k}^1\\ &\lesssim \left(1+(\Delta t\lambda_2)^2+d\Delta t\sigma_2^2\right) \left(\left\|X_{k-1}^{i,1} - X_{k-1}^{i,2}\right\|_2^2 + \left\|x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,1}) - x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,2})\right\|_2^2\right) \bar{\mathcal{I}}_{M,k-1}^1\\ &\quad + \left((\Delta t\,|\lambda_1-\lambda_2|)^2 + d\Delta t\,|\sigma_1-\sigma_2|^2\right) \left(\left\|X_{k-1}^{i,1}\right\|_2^2 + \left\|x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,1})\right\|_2^2\right) \bar{\mathcal{I}}_{M,k-1}^1,\end{aligned}$$

where we used the fact that $\mathbb{E}_k\|B_k^i\|_2^2 = d\Delta t$. Taking now the total expectation $\mathbb{E}$ on both sides, we have by tower property (law of total expectation)

$$\begin{aligned}&\mathbb{E}\left\|X_k^{i,1} - X_k^{i,2}\right\|_2^2 \bar{\mathcal{I}}_{M,k}^1\\ &\lesssim \left(1+(\Delta t\lambda_2)^2+d\Delta t\sigma_2^2\right) \left(\mathbb{E}\left\|X_{k-1}^{i,1} - X_{k-1}^{i,2}\right\|_2^2 \bar{\mathcal{I}}_{M,k-1}^1 + \mathbb{E}\left\|x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,1}) - x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,2})\right\|_2^2 \bar{\mathcal{I}}_{M,k-1}^1\right)\\ &\quad + \left((\Delta t\,|\lambda_1-\lambda_2|)^2 + d\Delta t\,|\sigma_1-\sigma_2|^2\right) \left(\mathbb{E}\left\|X_{k-1}^{i,1}\right\|_2^2 \bar{\mathcal{I}}_{M,k-1}^1 + \mathbb{E}\left\|x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,1})\right\|_2^2 \bar{\mathcal{I}}_{M,k-1}^1\right).\end{aligned} \tag{33}$$

As a consequence of the stability estimate for the consensus point, Lemma E.1, it holds for a constant $c_0 = c_0(\alpha, C_1, C_2, M) > 0$ that

$$\begin{aligned}\mathbb{E}\left\|x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,1}) - x_\alpha^{\mathcal{E}}(\widehat{\rho}_{k-1}^{N,2})\right\|_2^2 \bar{\mathcal{I}}_{M,k-1}^1 &\le c_0\mathbb{E}W_2^2\left(\widehat{\rho}_{k-1}^{N,1}, \widehat{\rho}_{k-1}^{N,2}\right) \bar{\mathcal{I}}_{M,k-1}^1\\ &\le c_0\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left\|X_{k-1}^{i,1} - X_{k-1}^{i,2}\right\|_2^2 \bar{\mathcal{I}}_{M,k-1}^1,\end{aligned}$$

where we chose $\pi = \frac{1}{N}\sum_{i=1}^N \delta_{X_{k-1}^{i,1}} \otimes \delta_{X_{k-1}^{i,2}}$ as viable transportation plan in Definition (7) to upper bound the Wasserstein distance in the second step. Utilizing this when averaging (33) over i gives

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left\|X_k^{i,1}-X_k^{i,2}\right\|_2^2 \bar{\mathcal{I}}_{M,k}^1 \lesssim{}& (1+c_0)\left(1+(\Delta t\lambda_2)^2+d\Delta t\sigma_2^2\right) \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left\|X_{k-1}^{i,1}-X_{k-1}^{i,2}\right\|_2^2 \bar{\mathcal{I}}_{M,k-1}^1\\ &+ \left((\Delta t\,|\lambda_1-\lambda_2|)^2 + d\Delta t\,|\sigma_1-\sigma_2|^2\right)\left(b_1 + (1+b_2)M^2\right),\end{aligned} \tag{34}$$

where we employed Lemma D.1 together with the definition of the cutoff function $\bar{\mathcal{I}}^1_{M,k-1}$ to obtain the bound in the second line of (34). Exploiting that $X_0^{i,1} = X_0^{i,2}$ for $i = 1, \dots, N$ by assumption, we conclude the proof by an application of the discrete variant of Grönwall's inequality (9), which proves that for all $k \geq 1$ it holds

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}\|X_k^{i,1} - X_k^{i,2}\|_2^2 \bar{\mathcal{I}}^1_{M,k} \leq c_1 \left((\Delta t\, |\lambda_1 - \lambda_2|)^2 + d\Delta t\, |\sigma_1 - \sigma_2|^2 \right) e^{c_2(k-1)}$$

with constants $c_1 = c_1(b_1, b_2, M) > 0$ and $c_2 = c_2(c_0, \Delta t, d, \lambda_2, \sigma_2) > 0$. □

E.3. A large deviation bound for the consensus point (3)

For a given measure $\varrho \in \mathcal{P}(\mathbb{R}^d)$ and a set of N i.i.d. random variables $Y^i \sim \varrho$ with empirical random measure $\widehat{\varrho}^N = \frac{1}{N}\sum_{i=1}^N \delta_{Y^i}$, one expects that under certain regularity assumptions it holds by the law of large numbers

$$x_\alpha^{\mathcal{E}}(\widehat{\varrho}^N) \xrightarrow{\text{a.s.}} x_\alpha^{\mathcal{E}}(\varrho) \quad \text{as } N \to \infty.$$

This is made rigorous in the subsequent lemma, which is based on arguments from (Fornasier et al., 2020, Lemma 3.1) and (Fornasier et al., 2021b, Lemma 23).

Lemma E.3 (Large deviation bound for the consensus point $x_\alpha^{\mathcal{E}}$)**.** *Let $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfy A1–A2. Moreover, for $k = 1, \dots, K$, let $\mu_k \in \mathcal{P}(\mathbb{R}^d)$ be a random measure, let $(Y_k^i)_{i=1,\dots,N}$ be N i.i.d. random variables distributed according to μ_k, denote by $\widehat{\mu}_k^N$ the empirical random measure $\widehat{\mu}_k^N = \frac{1}{N}\sum_{i=1}^N \delta_{Y_k^i}$ and define the cutoff function (random variable)*

$$\bar{\mathcal{I}}^2_{M,k} = \begin{cases} 1, & \text{if } \max\left\{ \int \|\bullet\|_2^4 \, d\widehat{\mu}_k^N, \int \|\bullet\|_2^4 \, d\mu_k \right\} \leq M^4, \\ 0, & \text{else.} \end{cases} \tag{35}$$

Then it holds

$$\max_{k=1,\dots,K} \mathbb{E}\left\| x_\alpha^{\mathcal{E}}(\widehat{\mu}_k^N) - x_\alpha^{\mathcal{E}}(\mu_k) \right\|_2^2 \bar{\mathcal{I}}^2_{M,k} \leq c_3 N^{-1}$$

with a constant $c_3 = c_3(\alpha, b_1, b_2, C_2, M) > 0$.

Proof. To start with, we note that under A2 and with Jensen's inequality it holds

$$\begin{aligned} \frac{e^{-\alpha\underline{\mathcal{E}}}\,\bar{\mathcal{I}}^2_{M,k}}{\frac{1}{N}\sum_{j=1}^N \omega_\alpha^{\mathcal{E}}(Y_k^j)} &= \frac{\bar{\mathcal{I}}^2_{M,k}}{\frac{1}{N}\sum_{j=1}^N \exp\big(-\alpha(\mathcal{E}(Y_k^j) - \underline{\mathcal{E}})\big)} \leq \frac{\bar{\mathcal{I}}^2_{M,k}}{\frac{1}{N}\sum_{j=1}^N \exp\big(-\alpha C_2(1 + \|Y_k^j\|_2^2)\big)} \\ &\leq \frac{\bar{\mathcal{I}}^2_{M,k}}{\exp\big(-\alpha C_2(1 + \frac{1}{N}\sum_{j=1}^N \|Y_k^j\|_2^2)\big)} \leq \exp\big(\alpha C_2(1 + M^2)\big) =: c_M. \end{aligned} \tag{36}$$

By definition of the consensus point $x_\alpha^{\mathcal{E}}$ in (3), it holds

$$\begin{aligned} \left\| x_\alpha^{\mathcal{E}}(\widehat{\mu}_k^N) - x_\alpha^{\mathcal{E}}(\mu_k) \right\|_2 \bar{\mathcal{I}}^2_{M,k} &= \left\| \sum_{i=1}^N Y_k^i \frac{\omega_\alpha^{\mathcal{E}}(Y_k^i)}{\sum_{j=1}^N \omega_\alpha^{\mathcal{E}}(Y_k^j)} - \int x \frac{\omega_\alpha^{\mathcal{E}}(x)}{\|\omega_\alpha^{\mathcal{E}}\|_{L_1(\mu_k)}} d\mu_k(x) \right\|_2 \bar{\mathcal{I}}^2_{M,k} \\ &\leq \big(\|T_1\|_2 + \|T_2\|_2 \big) \bar{\mathcal{I}}^2_{M,k}, \end{aligned} \tag{37}$$

where the terms T_1 and T_2 are defined implicitly and bounded as follows. For the first term T_1 we have

$$\begin{aligned} \|T_1\|_2 \bar{\mathcal{I}}^2_{M,k} &= \left\| \sum_{i=1}^N Y_k^i \frac{\omega_\alpha^{\mathcal{E}}(Y_k^i)}{\sum_{j=1}^N \omega_\alpha^{\mathcal{E}}(Y_k^j)} - \int x \frac{\omega_\alpha^{\mathcal{E}}(x)}{\frac{1}{N}\sum_{j=1}^N \omega_\alpha^{\mathcal{E}}(Y_k^j)} d\mu_k(x) \right\|_2 \bar{\mathcal{I}}^2_{M,k} \\ &= \frac{\bar{\mathcal{I}}^2_{M,k}}{\frac{1}{N}\sum_{j=1}^N \omega_\alpha^{\mathcal{E}}(Y_k^j)} \left\| \frac{1}{N}\sum_{i=1}^N Y_k^i \omega_\alpha^{\mathcal{E}}(Y_k^i) - \int x \omega_\alpha^{\mathcal{E}}(x)\, d\mu_k(x) \right\|_2 \\ &\leq c_M e^{\alpha\underline{\mathcal{E}}} \left\| \frac{1}{N}\sum_{i=1}^N Y_k^i \omega_\alpha^{\mathcal{E}}(Y_k^i) - \int x \omega_\alpha^{\mathcal{E}}(x)\, d\mu_k(x) \right\|_2, \end{aligned} \tag{38}$$

where we utilized (36) in the last step. Similarly, for the second term T_2 we have

$$\begin{aligned}\|T_2\|_2 \bar{\mathcal{I}}_{M,k}^2 &= \left\| \int x \frac{\omega_\alpha^{\mathcal{E}}(x)}{\frac{1}{N}\sum_{j=1}^N \omega_\alpha^{\mathcal{E}}(Y_k^j)} d\mu_k(x) - \int x \frac{\omega_\alpha^{\mathcal{E}}(x)}{\|\omega_\alpha^{\mathcal{E}}\|_{L_1(\mu_k)}} d\mu_k(x) \right\|_2 \bar{\mathcal{I}}_{M,k}^2 \\ &= \frac{\bar{\mathcal{I}}_{M,k}^2}{\frac{1}{N}\sum_{j=1}^N \omega_\alpha^{\mathcal{E}}(Y_k^j)} \left\| x_\alpha^{\mathcal{E}}(\mu_k) \right\|_2 \left| \frac{1}{N}\sum_{j=1}^N \omega_\alpha^{\mathcal{E}}(Y_k^j) - \int \omega_\alpha^{\mathcal{E}}(x)\, d\mu_k(x) \right|_2 \\ &\leq c_M e^{\alpha\underline{\mathcal{E}}} \left(b_1 + b_2 M\right) \left| \frac{1}{N}\sum_{j=1}^N \omega_\alpha^{\mathcal{E}}(Y_k^j) - \int \omega_\alpha^{\mathcal{E}}(x)\, d\mu_k(x) \right|_2,\end{aligned} \tag{39}$$

where the last step involved additionally Lemma D.1. Let us now introduce the random variables

$$Z_k^i := Y_k^i \omega_\alpha^{\mathcal{E}}(Y_k^i) - \int x \omega_\alpha^{\mathcal{E}}(x)\, d\mu_k(x) \quad \text{and} \quad z_k^i := \omega_\alpha^{\mathcal{E}}(Y_k^i) - \int \omega_\alpha^{\mathcal{E}}(x)\, d\mu_k(x),$$

respectively, which have zero expectation, and are i.i.d. for $i = 1, \dots, N$. With these definitions as well as the bounds (38) and (39) we obtain

$$\begin{aligned}\mathbb{E}\, \|T_1\|_2^2 \bar{\mathcal{I}}_{M,k}^2 &\leq c_M^2 e^{2\alpha\underline{\mathcal{E}}} \mathbb{E} \left\| \frac{1}{N}\sum_{i=1}^N Z_k^i \right\|_2^2 \bar{\mathcal{I}}_{M,k}^2 = c_M^2 e^{2\alpha\underline{\mathcal{E}}} \frac{1}{N^2} \mathbb{E} \sum_{i=1}^N \sum_{j=1}^N \left\langle Z_k^i, Z_k^j \right\rangle \bar{\mathcal{I}}_{M,k}^2 \\ &= c_M^2 e^{2\alpha\underline{\mathcal{E}}} \frac{1}{N^2} \mathbb{E} \sum_{i=1}^N \left\| Z_k^i \right\|_2^2 \bar{\mathcal{I}}_{M,k}^2 \leq 4 c_M^2 M^2 \frac{1}{N}\end{aligned} \tag{40}$$

and, analogously,

$$\mathbb{E}\, \|T_2\|_2^2 \bar{\mathcal{I}}_{M,k}^2 \leq c_M^2 e^{2\alpha\underline{\mathcal{E}}} \left(b_1 + b_2 M\right)^2 \frac{1}{N^2} \mathbb{E} \sum_{i=1}^N \left\| z_k^i \right\|_2^2 \bar{\mathcal{I}}_{M,k}^2 \leq 4 c_M^2 \left(b_1 + b_2 M\right)^2 \frac{1}{N}. \tag{41}$$

The last inequalities of (40) and (41) are due to the estimates

$$\begin{aligned}\mathbb{E} \frac{1}{N}\sum_{i=1}^N \left\| Z_k^i \right\|_2^2 \bar{\mathcal{I}}_{M,k}^2 &\leq 2\mathbb{E} \frac{1}{N}\sum_{i=1}^N \left\| Y_k^i \omega_\alpha^{\mathcal{E}}(Y_k^i) \right\|_2^2 \bar{\mathcal{I}}_{M,k}^2 + 2\mathbb{E} \left\| \int x \omega_\alpha^{\mathcal{E}}(x)\, d\mu_k(x) \right\|_2^2 \bar{\mathcal{I}}_{M,k}^2 \\ &\leq 2e^{-2\alpha\underline{\mathcal{E}}} \mathbb{E} \frac{1}{N}\sum_{i=1}^N \left\| Y_k^i \right\|_2^2 \bar{\mathcal{I}}_{M,k}^2 + 2e^{-2\alpha\underline{\mathcal{E}}} \mathbb{E} \int \|x\|_2^2\, d\mu_k(x)\, \bar{\mathcal{I}}_{M,k}^2 \\ &\leq 4e^{-2\alpha\underline{\mathcal{E}}} M^2\end{aligned}$$

and, similarly,

$$\mathbb{E} \left| z_k^1 \right|_2^2 \bar{\mathcal{I}}_{M,k}^2 \leq 4e^{-2\alpha\underline{\mathcal{E}}}.$$

Combining (40) and (41) concludes the proof. □

Remark E.4. Alternatively to the explicit computations of Lemma E.3, the stability estimate for the consensus point, Lemma E.1, would allow to obtain

$$\max_{k=1,\dots,K} \mathbb{E} \left\| x_\alpha^{\mathcal{E}}(\widehat{\mu}_k^N) - x_\alpha^{\mathcal{E}}(\mu_k) \right\|_2^2 \bar{\mathcal{I}}_{M,k}^2 \leq c_0 \max_{k=1,\dots,K} \mathbb{E} W_2^2(\widehat{\mu}_k^N, \mu_k)\, \bar{\mathcal{I}}_{M,k}^2,$$

where $\mathbb{E} W_2^2(\widehat{\mu}_k^N, \mu_k)$ can be controlled by employing (Fournier & Guillin, 2015, Theorem 1). This, however, only gives a quantitative convergence rate of order $\mathcal{O}(N^{-2/d})$, which is affected by the curse of dimensionality. The convergence rate $\mathcal{O}(N^{-1})$ obtained in Lemma E.3 matches the one to be expected from Monte Carlo sampling.

E.4. Proof of Theorem C.1

We now have all necessary tools at hand to present the detailed proof of Theorem C.1.

Proof of Theorem C.1. We notice that for the choice $\lambda = 1/\Delta t$ the iterative update rule of the particles of the CBO dynamics (2) becomes

$$\widetilde{X}_k^i = x_\alpha^{\mathcal{E}}(\widetilde{\rho}_{k-1}^N) + \sigma D\big(\widetilde{X}_{k-1}^i - x_\alpha^{\mathcal{E}}(\widetilde{\rho}_{k-1}^N)\big)\, B_k^i, \tag{42}$$

where $\widetilde{\rho}_k^N = \frac{1}{N}\sum_{i=1}^N \delta_{\widetilde{X}_k^i}$. In this case, the associated CBO scheme (4) reads

$$\begin{aligned} &\widetilde{x}_k^{\mathrm{CBO}} = x_\alpha^{\mathcal{E}}(\widetilde{\rho}_k^N) \quad \text{with } \widetilde{\rho}_k^N = \frac{1}{N}\sum_{i=1}^N \delta_{\widetilde{X}_k^i}, \text{ where } \widetilde{X}_k^i \sim \mathcal{N}\Big(\widetilde{x}_{k-1}^{\mathrm{CBO}}, \Delta t \sigma^2 D\big(\widetilde{X}_{k-1}^i - \widetilde{x}_{k-1}^{\mathrm{CBO}}\big)^2\Big), \\ &\widetilde{x}_0^{\mathrm{CBO}} = x_0, \end{aligned} \tag{43}$$

which resembles the CH dynamics (10) with the difference in the underlying measure on which basis the consensus point (3) is computed. Let us further denote by $\widehat{\mu}_k^N$ the empirical measure $\widehat{\mu}_k^N = \frac{1}{N}\sum_{i=1}^N \delta_{Y_k^i}$, where $Y_k^i \sim \mu_k = \mathcal{N}\big(x_{k-1}^{\mathrm{CH}}, \widetilde{\sigma}^2 \mathrm{Id}\big)$ for $i = 1, \dots, N$, i.e., $Y_k^i = x_{k-1}^{\mathrm{CH}} + \widetilde{\sigma} B_{Y,k}^i$ with $B_{Y,k}^i$ being a standard Gaussian random vector.

To obtain the probabilistic formulation of the statement, let us denote the underlying probability space over which all considered random variables get their realizations by $(\Omega, \mathcal{F}, \mathbb{P})$ and introduce the subset Ω_M of Ω of suitably bounded random variables according to

$$\Omega_M := \left\{ \omega \in \Omega : \max_{k=0,\dots,K} \max \left\{ \int \|\bullet\|_2^4 \, d\widehat{\rho}_k^N, \int \|\bullet\|_2^4 \, d\widetilde{\rho}_k^N, \int \|\bullet\|_2^4 \, d\mu_k, \int \|\bullet\|_2^4 \, d\widehat{\mu}_k^N \right\} \leq M^4 \right\}.$$

For the associated cutoff function (random variable) we write $\mathbb{1}_{\Omega_M}$. Moreover, let us define the cutoff functions

$$\mathcal{I}_{M,k} = \begin{cases} 1, & \text{if } \max\left\{ \int \|\bullet\|_2^4 \, d\widehat{\rho}_k^N, \int \|\bullet\|_2^4 \, d\widetilde{\rho}_k^N, \int \|\bullet\|_2^4 \, d\mu_k, \int \|\bullet\|_2^4 \, d\widehat{\mu}_k^N \right\} \leq M^4 \text{ for all } \ell \leq k, \\ 0, & \text{else}, \end{cases} \tag{44}$$

which are adapted to the natural filtration and satisfy $\mathbb{1}_{\Omega_M} \leq \mathcal{I}_{M,k}$ as well as $\mathcal{I}_{M,k} = \mathcal{I}_{M,k}\mathcal{I}_{M,\ell}$ for all $\ell \leq k$.

We can decompose the expected squared discrepancy $\mathbb{E}\left\|x_k^{\mathrm{CBO}} - x_k^{\mathrm{CH}}\right\|_2^2 \mathbb{1}_{\Omega_M}$ between the CBO scheme (4) and the CH scheme (10) as

$$\mathbb{E}\left\|x_k^{\mathrm{CBO}} - x_k^{\mathrm{CH}}\right\|_2^2 \mathcal{I}_{M,k} \leq 2\mathbb{E}\left\|x_k^{\mathrm{CBO}} - \widetilde{x}_k^{\mathrm{CBO}}\right\|_2^2 \mathcal{I}_{M,k} + 2\mathbb{E}\left\|\widetilde{x}_k^{\mathrm{CBO}} - x_k^{\mathrm{CH}}\right\|_2^2 \mathcal{I}_{M,k}. \tag{45}$$

In what follows we individually bound the two terms on the right-hand side of (45).

First term: Let us start with the term $\mathbb{E}\left\|x_k^{\mathrm{CBO}} - \widetilde{x}_k^{\mathrm{CBO}}\right\|_2^2 \mathcal{I}_{M,k}$, which we bound by combining the stability estimate for the consensus point, Lemma E.1, with Lemma E.2, a stability estimate for the underlying CBO dynamics (2) w.r.t. its parameters λ and σ. Denoting the auxiliary cutoff function defined in (31) in the setting $\widehat{\rho}_k^{N,1} = \widehat{\rho}_k^N$ and $\widehat{\rho}_k^{N,2} = \widetilde{\rho}_k^N$ by $\bar{\mathcal{I}}_{M,k}^1$, we have due to Lemma E.1 the estimate

$$\begin{aligned} \mathbb{E}\left\|x_k^{\mathrm{CBO}} - \widetilde{x}_k^{\mathrm{CBO}}\right\|_2^2 \mathcal{I}_{M,k} &= \mathbb{E}\left\|x_\alpha^{\mathcal{E}}(\widehat{\rho}_k^N) - x_\alpha^{\mathcal{E}}(\widetilde{\rho}_k^N)\right\|_2^2 \mathcal{I}_{M,k} \\ &\leq \mathbb{E}\left\|x_\alpha^{\mathcal{E}}(\widehat{\rho}_k^N) - x_\alpha^{\mathcal{E}}(\widetilde{\rho}_k^N)\right\|_2^2 \bar{\mathcal{I}}_{M,k}^1 \leq c_0 \mathbb{E} W_2^2(\widehat{\rho}_k^N, \widetilde{\rho}_k^N)\, \bar{\mathcal{I}}_{M,k}^1 \end{aligned} \tag{46}$$

with a constant $c_0 = c_0(\alpha, C_1, C_2, M) > 0$. In the first inequality of (46) we exploited $\mathcal{I}_{M,k} \leq \bar{\mathcal{I}}_{M,k}^1$. The Wasserstein distance appearing on the right-hand side of (46) can be upper bounded by choosing $\pi = \frac{1}{N}\sum_{i=1}^N \delta_{X_k^i} \otimes \delta_{\widetilde{X}_k^i}$ as viable transportation plan in Definition (7). This constitutes the first inequality in the estimate

$$\begin{aligned} \mathbb{E} W_2^2(\widehat{\rho}_k^N, \widetilde{\rho}_k^N)\, \bar{\mathcal{I}}_{M,k}^1 &\leq \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left\|X_k^i - \widetilde{X}_k^i\right\|_2^2 \bar{\mathcal{I}}_{M,k}^1 \\ &\leq c_1 \left(|\lambda_1 - \lambda_2|^2 + |\sigma_1 - \sigma_2|^2\right) e^{c_2(k-1)} \leq c_1 \left|\lambda - \frac{1}{\Delta t}\right|^2 e^{c_2(k-1)}, \end{aligned} \tag{47}$$

whereas the second step is a consequence of Lemma E.2 applied with $\lambda_1 = \lambda, \sigma_1 = \sigma$ and $\lambda_2 = 1/\Delta t, \sigma_2 = \sigma$ as exploited in the third step. Hence, the constants are $c_1 = c_1(\Delta t, d, b_1, b_2, M) > 0$ and $c_2 = c_2(\Delta t, d, \alpha, \lambda, \sigma, C_1, C_2, M) > 0$.

Second term: To control the term $\mathbb{E}\left\|\widetilde{x}_k^{\mathrm{CBO}} - x_k^{\mathrm{CH}}\right\|_2^2 \mathcal{I}_{M,k}$ we start by decomposing it according to

$$\mathbb{E}\left\|\widetilde{x}_k^{\mathrm{CBO}} - x_k^{\mathrm{CH}}\right\|_2^2 \mathcal{I}_{M,k} \leq 2\mathbb{E}\left\|\widetilde{x}_k^{\mathrm{CBO}} - x_\alpha^{\mathcal{E}}(\widehat{\mu}_k^N)\right\|_2^2 \mathcal{I}_{M,k} + 2\mathbb{E}\left\|x_\alpha^{\mathcal{E}}(\widehat{\mu}_k^N) - x_k^{\mathrm{CH}}\right\|_2^2 \mathcal{I}_{M,k}, \tag{48}$$

where $\widehat{\mu}_k^N$ is as introduced at the beginning of the proof. For the first summand in (48) the stability estimate for the consensus point, Lemma E.1, gives

$$\begin{aligned}\mathbb{E}\left\|\widetilde{x}_k^{\mathrm{CBO}} - x_\alpha^{\mathcal{E}}(\widehat{\mu}_k^N)\right\|_2^2 \mathcal{I}_{M,k} &= \mathbb{E}\left\|x_\alpha^{\mathcal{E}}(\widetilde{\rho}_k^N) - x_\alpha^{\mathcal{E}}(\widehat{\mu}_k^N)\right\|_2^2 \mathcal{I}_{M,k} \\ &\leq c_0 \mathbb{E} W_2^2(\widetilde{\rho}_k^N, \widehat{\mu}_k^N)\, \mathcal{I}_{M,k}\end{aligned} \tag{49}$$

with a constant $c_0 = c_0(\alpha, C_1, C_2, M) > 0$. By choosing $\pi = \frac{1}{N}\sum_{i=1}^N \delta_{\widetilde{X}_k^i} \otimes \delta_{Y_k^i}$ as viable transportation plan in Definition (7), we can further bound

$$\mathbb{E} W_2^2(\widetilde{\rho}_k^N, \widehat{\mu}_k^N)\, \mathcal{I}_{M,k} \leq \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left\|\widetilde{X}_k^i - Y_k^i\right\|_2^2 \mathcal{I}_{M,k} \tag{50}$$

and since $\widetilde{X}_k^i \sim \mathcal{N}\big(\widetilde{x}_{k-1}^{\mathrm{CBO}}, \Delta t \sigma^2 D(\widetilde{X}_{k-1}^i - \widetilde{x}_{k-1}^{\mathrm{CBO}})^2\big)$ and $Y_k^i \sim \mathcal{N}\big(x_{k-1}^{\mathrm{CH}}, \widetilde{\sigma}^2 \mathrm{Id}\big)$ we have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left\|\widetilde{X}_k^i - Y_k^i\right\|_2^2 \mathcal{I}_{M,k} &\leq 2\mathbb{E}\left\|\widetilde{x}_{k-1}^{\mathrm{CBO}} - x_{k-1}^{\mathrm{CH}}\right\|_2^2 \mathcal{I}_{M,k-1} \\ &\quad + \frac{4}{N}\sum_{i=1}^N \left(\sigma^2 \mathbb{E}\left\|D(\widetilde{X}_{k-1}^i - \widetilde{x}_{k-1}^{\mathrm{CBO}}) B_k^i\right\|_2^2 \mathcal{I}_{M,k-1} + \widetilde{\sigma}^2 \mathbb{E}\left\|B_{Y,k}^i\right\|_2^2\right) \\ &\leq 2\mathbb{E}\left\|\widetilde{x}_{k-1}^{\mathrm{CBO}} - x_{k-1}^{\mathrm{CH}}\right\|_2^2 \mathcal{I}_{M,k-1} + 8\sigma^2 \Delta t \left(b_1 + (1+b_2)M^2\right) + 4\widetilde{\sigma}^2.\end{aligned} \tag{51}$$

Note that in the last step we exploited the definition of the cutoff function $\mathcal{I}_{M,k}$, which allowed to derive the bound

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left\|D(\widetilde{X}_{k-1}^i - \widetilde{x}_{k-1}^{\mathrm{CBO}}) B_k^i\right\|_2^2 \mathcal{I}_{M,k-1} &\leq \frac{2}{N}\sum_{i=1}^N \mathbb{E}\left(\left\|\widetilde{X}_{k-1}^i\right\|_2^2 + \left\|\widetilde{x}_{k-1}^{\mathrm{CBO}}\right\|_2^2\right) \left\|B_k^i\right\|_2^2 \mathcal{I}_{M,k-1} \\ &\leq 2\mathbb{E}\left\|\widetilde{x}_{k-1}^{\mathrm{CBO}}\right\|_2^2 \mathcal{I}_{M,k-1} + \frac{2}{N}\sum_{i=1}^N \mathbb{E}\left\|\widetilde{X}_{k-1}^i\right\|_2^2 \mathcal{I}_{M,k-1} \\ &\leq 2\left(b_1 + (1+b_2)M^2\right)\end{aligned}$$

by using Lemma D.1 and the fact that $B_k^i \sim \mathcal{N}(0, \Delta t \mathrm{Id})$ is independent from $\widetilde{X}_{k-1}^i$ and $\widetilde{x}_{k-1}^{\mathrm{CBO}}$. Inserting (51) into (50) and this into (49) afterwards, we are left with

$$\mathbb{E}\left\|\widetilde{x}_k^{\mathrm{CBO}} - x_\alpha^{\mathcal{E}}(\widehat{\mu}_k^N)\right\|_2^2 \mathcal{I}_{M,k} \leq c\left(\mathbb{E}\left\|\widetilde{x}_{k-1}^{\mathrm{CBO}} - x_{k-1}^{\mathrm{CH}}\right\|_2^2 \mathcal{I}_{M,k-1} + \sigma^2 \Delta t + \widetilde{\sigma}^2\right) \tag{52}$$

with a constant $c = c(c_0, b_1, b_2, M) > 0$. For the second summand in (48) we have by Lemma E.3

$$\begin{aligned}\mathbb{E}\left\|x_\alpha^{\mathcal{E}}(\widehat{\mu}_k^N) - x_k^{\mathrm{CH}}\right\|_2^2 \mathcal{I}_{M,k} &\leq \mathbb{E}\left\|x_\alpha^{\mathcal{E}}(\widehat{\mu}_k^N) - x_\alpha^{\mathcal{E}}(\mu_k)\right\|_2^2 \overline{\mathcal{I}}_{M,k}^2 \\ &\leq c_3 N^{-1},\end{aligned} \tag{53}$$

with $c_3 = c_3(\alpha, b_1, b_2, C_2, M) > 0$ and where $\overline{\mathcal{I}}_{M,k}^2$ is an auxiliary cutoff function as defined in (35). Combining (52) with (53) we arrive for (48) at

$$\mathbb{E}\left\|\widetilde{x}_k^{\mathrm{CBO}} - x_k^{\mathrm{CH}}\right\|_2^2 \mathcal{I}_{M,k} \leq c\mathbb{E}\left\|\widetilde{x}_{k-1}^{\mathrm{CBO}} - x_{k-1}^{\mathrm{CH}}\right\|_2^2 \mathcal{I}_{M,k-1} + c\sigma^2 \Delta t + c\widetilde{\sigma}^2 + c_3 N^{-1}. \tag{54}$$

An application of the discrete variant of Grönwall's inequality (9) shows that

$$\mathbb{E}\left\|\widetilde{x}_k^{\mathrm{CBO}} - x_k^{\mathrm{CH}}\right\|_2^2 \mathcal{I}_{M,k} \leq c^k \mathbb{E}\left\|\widetilde{x}_0^{\mathrm{CBO}} - x_0^{\mathrm{CH}}\right\|_2^2 + \left(c\sigma^2\Delta t + c\widetilde{\sigma}^2 + c_3 N^{-1}\right) e^{c(k-1)}, \tag{55}$$

where the first term vanishes as both schemes are initialized with x_0.

Concluding step: Collecting the estimates (46) combined with (47), and (55) yields for (45) the bound

$$\begin{aligned}\mathbb{E}\left\|x_k^{\mathrm{CBO}} - x_k^{\mathrm{CH}}\right\|_2^2 \mathbb{1}_{\Omega_M} &\lesssim c_0 c_1 \left|\lambda - \frac{1}{\Delta t}\right|^2 e^{c_2(k-1)} + \left(c\sigma^2\Delta t + c\widetilde{\sigma}^2 + c_3 N^{-1}\right) e^{c(k-1)} \\ &\leq C\left(\left|\lambda - \frac{1}{\Delta t}\right|^2 + \sigma^2\Delta t + \widetilde{\sigma}^2 + c_3 N^{-1}\right),\end{aligned} \tag{56}$$

with a constant $C = C(\Delta t, d, \alpha, \lambda, \sigma, b_1, b_2, C_1, C_2, K, M) > 0$. Observe that we additionally used $\mathbb{1}_{\Omega_M} \leq \mathcal{I}_{M,k}$ as observed at the beginning.

Probabilistic formulation: We first note that with Markov's inequality we have the estimate

$$\begin{aligned}\mathbb{P}(\Omega_M^c) &= \mathbb{P}\left(\max_{k=0,\dots,K} \max\left\{\int \|\bullet\|_2^4 \, d\widehat{\rho}_k^N, \int \|\bullet\|_2^4 \, d\widetilde{\rho}_k^N, \int \|\bullet\|_2^4 \, d\mu_k, \int \|\bullet\|_2^4 \, d\widehat{\mu}_k^N\right\} > M^4\right) \\ &\leq \frac{1}{M^4}\left(\mathbb{E}\max_{k=0,\dots,K}\int \|\bullet\|_2^4 \, d\widehat{\rho}_k^N + \mathbb{E}\max_{k=0,\dots,K}\int \|\bullet\|_2^4 \, d\widetilde{\rho}_k^N \right. \\ &\qquad \left. + \mathbb{E}\max_{k=0,\dots,K}\int \|\bullet\|_2^4 \, d\mu_k + \mathbb{E}\max_{k=0,\dots,K}\int \|\bullet\|_2^4 \, d\widehat{\mu}_k^N\right) \\ &\leq \frac{1}{M^4}\left(\mathcal{M}^{\mathrm{CBO}} + \widetilde{\mathcal{M}}^{\mathrm{CBO}} + \mathcal{M}^{\mathrm{CH}} + \widehat{\mathcal{M}}^{\mathrm{CH}}\right),\end{aligned}$$

where the last inequality is due to Lemmas D.2, D.3 and D.4. Here, $\widetilde{\mathcal{M}}^{\mathrm{CBO}}$ represents the constant $\mathcal{M}^{\mathrm{CBO}}$ from Lemma D.2 in the setting where $\lambda = 1/\Delta t$, i.e., $\widetilde{\mathcal{M}}^{\mathrm{CBO}} = \mathcal{M}^{\mathrm{CBO}}(1/\Delta t, \sigma, d, b_1, b_2, K\Delta t, K, \rho_0)$. Thus, for any $\delta \in (0, 1/2)$, a sufficiently large choice $M = M(\delta^{-1}, \mathcal{M}^{\mathrm{CBO}}, \widetilde{\mathcal{M}}^{\mathrm{CBO}}, \mathcal{M}^{\mathrm{CH}}, \widehat{\mathcal{M}}^{\mathrm{CH}})$ allows to ensure $\mathbb{P}(\Omega_M^c) \leq \delta$. To conclude the proof, let us denote by $K_\varepsilon \subset \Omega$ the set, where (13) does not hold and abbreviate

$$\epsilon = \varepsilon^{-1} C\left(\left|\lambda - \frac{1}{\Delta t}\right|^2 + \sigma^2\Delta t + \widetilde{\sigma}^2 + c_3 N^{-1}\right).$$

For the probability of this set we can estimate

$$\begin{aligned}\mathbb{P}(K_\varepsilon) &= \mathbb{P}(K_\varepsilon \cap \Omega_M) + \mathbb{P}(K_\varepsilon \cap \Omega_M^c) \leq \mathbb{P}(K_\varepsilon \,|\, \Omega_M)\,\mathbb{P}(\Omega_M) + \mathbb{P}(\Omega_M^c) \\ &\leq \mathbb{P}(K_\varepsilon \,|\, \Omega_M) + \delta \leq \epsilon^{-1}\,\mathbb{E}\left[\left\|x_k^{\mathrm{CBO}} - x_k^{\mathrm{CH}}\right\|_2^2 \,\middle|\, \Omega_M\right] + \delta,\end{aligned} \tag{57}$$

where the last step is due to Markov's inequality. By definition of the conditional expectation we further have

$$\mathbb{E}\left[\left\|x_k^{\mathrm{CBO}} - x_k^{\mathrm{CH}}\right\|_2^2 \,\middle|\, \Omega_M\right] \leq \frac{1}{\mathbb{P}(\Omega_M)}\mathbb{E}\left\|x_k^{\mathrm{CBO}} - x_k^{\mathrm{CH}}\right\|_2^2 \mathbb{1}_{\Omega_M} \leq 2\mathbb{E}\left\|x_k^{\mathrm{CBO}} - x_k^{\mathrm{CH}}\right\|_2^2 \mathbb{1}_{\Omega_M}.$$

Inserting now the expression from (56) concludes the proof. □

F. Proof details for Proposition C.2 and Theorem C.3

Proposition C.2 and Theorem C.3 are centered around the observation that the CH scheme (10) behaves gradient-like. To establish this, Proposition C.2 exploits, by using the quantitative nonasymptotic Laplace principle (see Section F.1 and in particular Proposition F.2 for a review of (Fornasier et al., 2021b, Proposition 18)), that one step of the implicit CH scheme (11) can be recast into the computation of a consensus point $x_\alpha^{\widetilde{\mathcal{E}}}$ for an objective function of the form $\widetilde{\mathcal{E}}(x) = \frac{1}{2\tau}\|\bullet - x\|_2^2 + \mathcal{E}(x)$. To prove Theorem C.3, this is combined with a stability argument for the MMS (12), which relies on the Λ-convexity of $\mathcal{E}$ (Assumption A4).

F.1. A quantitative nonasymptotic Laplace principle

The Laplace principle (Dembo & Zeitouni, 1998; Miller, 2006) asserts that for any absolutely continuous probability measure $\varrho \in \mathcal{P}(\mathbb{R}^d)$ it holds

$$\lim_{\alpha\to\infty}\left(-\frac{1}{\alpha}\log\left(\int \exp\left(-\alpha\widetilde{\mathcal{E}}(x)\right)d\varrho(x)\right)\right) = \inf_{x\in\operatorname{supp}(\varrho)}\widetilde{\mathcal{E}}(x).$$

This suggests that, as $\alpha\to\infty$, the Gibbs measure $\eta_\alpha^{\widetilde{\mathcal{E}}} = \omega_\alpha^{\widetilde{\mathcal{E}}}\varrho/\|\omega_\alpha^{\widetilde{\mathcal{E}}}\|_{L_1(\varrho)}$ converges to a discrete probability distribution (i.e., a convex combination of Dirac measures) supported on the set of global minimizers of $\widetilde{\mathcal{E}}$. However, even in the case that such minimizer is unique, it does not permit to quantify the proximity of $x_\alpha^{\widetilde{\mathcal{E}}}(\varrho) = \int x\,d\eta_\alpha^{\widetilde{\mathcal{E}}}$ (see also Equation (3)) to the minimizer of $\widetilde{\mathcal{E}}$ without the following assumption (see also Remark B.1).

Definition F.1 (Inverse continuity property)**.** A function $\widetilde{\mathcal{E}}\in\mathcal{C}(\mathbb{R}^d)$ satisfies the ℓ^2-inverse continuity property globally if there exist constants $\eta,\nu>0$ such that

$$\|x-\widetilde{x}^*\|_2 \le \frac{1}{\eta}\big(\widetilde{\mathcal{E}}(x)-\underline{\widetilde{\mathcal{E}}}\big)^\nu \quad \text{for all } x\in\mathbb{R}^d, \tag{58}$$

where $\widetilde{x}^*\in\mathbb{R}^d$ denotes the unique global minimizer of $\widetilde{\mathcal{E}}$ with objective value $\underline{\widetilde{\mathcal{E}}} := \inf_{x\in\mathbb{R}^d}\widetilde{\mathcal{E}}(x)$.

As elaborated on in Remark B.1 for the (ℓ^∞-)inverse continuity property, it is usually sufficient if (58) holds locally around the global minimizer $\widetilde{x}^*$. In the following Proposition F.2, however, we recall the quantitative Laplace principle in the slightly more specific form, where the ℓ^2-inverse continuity property holds globally as required by Definition F.1. For the general version, namely in the case of functions which satisfy (58) only on an ℓ^2-ball around $\widetilde{x}^*$ (see (Fornasier et al., 2021b, Definition 8 (A2)) for the details), we refer to (Fornasier et al., 2021b, Proposition 18).

Proposition F.2 (Quantitative Laplace principle)**.** *Let $\widetilde{\mathcal{E}}\in\mathcal{C}(\mathbb{R}^d)$ satisfy the ℓ^2-inverse continuity property in form of Definition F.1. Moreover, let $\varrho\in\mathcal{P}(\mathbb{R}^d)$. For any $r>0$ define $\widetilde{\mathcal{E}}_r := \sup_{x\in B_r(\widetilde{x}^*)}\widetilde{\mathcal{E}}(x)-\underline{\widetilde{\mathcal{E}}}$. Then, for fixed $\alpha>0$ it holds for any $r,q>0$ that*

$$\big\|x_\alpha^{\widetilde{\mathcal{E}}}(\varrho)-\widetilde{x}^*\big\|_2 \le \frac{\big(q+\widetilde{\mathcal{E}}_r\big)^\nu}{\eta} + \frac{\exp(-\alpha q)}{\varrho(B_r(\widetilde{x}^*))}\int\|x-\widetilde{x}^*\|_2\,d\varrho(x). \tag{59}$$

Proof. W.l.o.g. we may assume $\underline{\widetilde{\mathcal{E}}}=0$ since a constant offset to $\widetilde{\mathcal{E}}$ neither affects the definition of the consensus point in (3) nor the quantities appearing on the right-hand side of (59).

By Markov's inequality it holds $\|\exp(-\alpha\widetilde{\mathcal{E}})\|_{L_1(\varrho)} \ge a\varrho\big(\{x\in\mathbb{R}^d : \exp(-\alpha\widetilde{\mathcal{E}}(x))\ge a\}\big)$ for any $a>0$. With the choice $a=\exp(-\alpha\widetilde{\mathcal{E}}_r)$ and noting that

$$\varrho\left(\left\{x\in\mathbb{R}^d : \exp(-\alpha\widetilde{\mathcal{E}}(x))\ge\exp(-\alpha\widetilde{\mathcal{E}}_r)\right\}\right) = \varrho\left(\left\{x\in\mathbb{R}^d : \widetilde{\mathcal{E}}(x)\le\widetilde{\mathcal{E}}_r\right\}\right) \ge \varrho(B_r(\widetilde{x}^*)),$$

we obtain $\|\exp(-\alpha\widetilde{\mathcal{E}})\|_{L_1(\varrho)} \ge \exp(-\alpha\widetilde{\mathcal{E}}_r)\varrho(B_r(\widetilde{x}^*))$. Now let $\widetilde{r}\ge r>0$. With the definition of the consensus point in (3) and by Jensen's inequality we can decompose

$$\begin{aligned}\big\|x_\alpha^{\widetilde{\mathcal{E}}}(\varrho)-\widetilde{x}^*\big\|_2 \le{}& \int_{B_{\widetilde{r}}(\widetilde{x}^*)}\|x-\widetilde{x}^*\|_2\,\frac{\exp\left(-\alpha\widetilde{\mathcal{E}}(x)\right)}{\big\|\exp(-\alpha\widetilde{\mathcal{E}})\big\|_{L_1(\varrho)}}d\varrho(x)\\ &+\int_{(B_{\widetilde{r}}(\widetilde{x}^*))^c}\|x-\widetilde{x}^*\|_2\,\frac{\exp\left(-\alpha\widetilde{\mathcal{E}}(x)\right)}{\big\|\exp(-\alpha\widetilde{\mathcal{E}})\big\|_{L_1(\varrho)}}d\varrho(x).\end{aligned}$$

The first term is bounded by $\widetilde{r}$ since $\|x-\widetilde{x}^*\|_2\le\widetilde{r}$ for all $x\in B_{\widetilde{r}}(\widetilde{x}^*)$. For the second term we use the formerly derived

$\|\exp(-\alpha\widetilde{\mathcal{E}})\|_{L_1(\varrho)} \geq \exp(-\alpha\widetilde{\mathcal{E}}_r)\varrho(B_r(\widetilde{x}^*))$ to get

$$\begin{aligned}
&\int_{(B_{\widetilde{r}}(\widetilde{x}^*))^c} \|x-\widetilde{x}^*\|_2 \frac{\exp\left(-\alpha\widetilde{\mathcal{E}}(x)\right)}{\left\|\exp(-\alpha\widetilde{\mathcal{E}})\right\|_{L_1(\varrho)}} d\varrho(x) \\
&\qquad \leq \frac{1}{\exp(-\alpha\widetilde{\mathcal{E}}_r)\varrho(B_r(\widetilde{x}^*))} \int_{(B_{\widetilde{r}}(\widetilde{x}^*))^c} \|x-\widetilde{x}^*\|_2 \exp\left(-\alpha\widetilde{\mathcal{E}}(x)\right) d\varrho(x) \\
&\qquad \leq \frac{\exp\left(-\alpha\left(\inf_{x\in(B_{\widetilde{r}}(\widetilde{x}^*))^c}\widetilde{\mathcal{E}}(x)-\widetilde{\mathcal{E}}_r\right)\right)}{\varrho(B_r(\widetilde{x}^*))} \int \|x-\widetilde{x}^*\|_2 \, d\varrho(x).
\end{aligned}$$

Thus, for any $\widetilde{r} \geq r > 0$ we obtain

$$\left\|x_\alpha^{\widetilde{\mathcal{E}}}(\varrho)-\widetilde{x}^*\right\|_2 \leq \widetilde{r} + \frac{\exp\left(-\alpha\left(\inf_{x\in(B_{\widetilde{r}}(\widetilde{x}^*))^c}\widetilde{\mathcal{E}}(x)-\widetilde{\mathcal{E}}_r\right)\right)}{\varrho(B_r(\widetilde{x}^*))} \int \|x-\widetilde{x}^*\|_2 \, d\varrho(x). \tag{60}$$

We now choose $\widetilde{r} = \left(q+\widetilde{\mathcal{E}}_r\right)^\nu/\eta$, which satisfies $\widetilde{r} \geq r$, since (58) with $\underline{\widetilde{\mathcal{E}}} = 0$ implies

$$\widetilde{r} = \frac{\left(q+\widetilde{\mathcal{E}}_r\right)^\nu}{\eta} \geq \frac{\widetilde{\mathcal{E}}_r^\nu}{\eta} = \frac{\left(\sup_{x\in B_r(\widetilde{x}^*)}\widetilde{\mathcal{E}}(x)\right)^\nu}{\eta} \geq \sup_{x\in B_r(\widetilde{x}^*)} \|x-\widetilde{x}^*\|_2 = r.$$

Using again (58) with $\underline{\widetilde{\mathcal{E}}} = 0$ we thus have

$$\inf_{x\in(B_{\widetilde{r}}(\widetilde{x}^*))^c} \widetilde{\mathcal{E}}(x) - \widetilde{\mathcal{E}}_r \geq (\eta\widetilde{r})^{1/\nu} - \widetilde{\mathcal{E}}_r = q + \widetilde{\mathcal{E}}_r - \widetilde{\mathcal{E}}_r = q.$$

Inserting this and the definition of $\widetilde{r}$ into (60) gives the statement. □

F.2. The auxiliary function $\widetilde{\mathcal{E}}_k$

Let us now show that the function $\widetilde{\mathcal{E}}_k(x) := \frac{1}{2\tau}\left\|x_{k-1}^{\mathrm{CH}}-x\right\|_2^2 + \mathcal{E}(x)$, which appears later in the proofs of Proposition C.2 and Theorem C.3, satisfies the ℓ^2-inverse continuity property in form of Definition F.1 if $\mathcal{E}$ is Λ-convex and the parameter τ sufficiently small. As we discuss in Remark F.4 below, the condition on the parameter τ vanishes if $\mathcal{E}$ is convex, i.e., $\Lambda \geq 0$.

Lemma F.3 ($\widetilde{\mathcal{E}}_k$ satisfies the ℓ^2-inverse continuity property)**.** *Let $\widetilde{\mathcal{E}}_k$ be defined as above with $\tau > 0$ and with $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfying A4. Moreover, if $\Lambda < 0$, assume further that $\tau < 1/(-\Lambda)$. Then, $\widetilde{\mathcal{E}}_k$ satisfies the ℓ^2-inverse continuity property (58) with parameters*

$$\nu = \frac{1}{2} \quad \text{and} \quad \eta = \sqrt{\frac{1}{2\tau}+\frac{\Lambda}{2}}.$$

I.e., denoting the unique global minimizer of $\widetilde{\mathcal{E}}_k$ by $\widetilde{x}_k^{\mathrm{CH}}$, it holds

$$\left\|x-\widetilde{x}_k^{\mathrm{CH}}\right\|_2 \leq \frac{1}{\eta}\left(\widetilde{\mathcal{E}}_k(x)-\widetilde{\mathcal{E}}_k(\widetilde{x}_k^{\mathrm{CH}})\right)^\nu \quad \text{for all } x \in \mathbb{R}^d. \tag{61}$$

Proof. We first notice that $\widetilde{\mathcal{E}}_k$ is $2\eta^2 = \left(\frac{1+\Lambda\tau}{\tau}\right)$-strongly convex ($2\eta^2 > 0$ by assumption), since

$$\begin{aligned}
\widetilde{\mathcal{E}}_k(x) - \frac{1}{2}\left(\frac{1+\Lambda\tau}{\tau}\right)\|x\|_2^2 &= \frac{1}{2\tau}\left(\left\|x_{k-1}^{\mathrm{CH}}-x\right\|_2^2 - \|x\|_2^2\right) + \mathcal{E}(x) - \frac{\Lambda}{2}\|x\|_2^2 \\
&= \underbrace{\frac{1}{2\tau}\left(\left\|x_{k-1}^{\mathrm{CH}}\right\|_2^2 - 2\left\langle x_{k-1}^{\mathrm{CH}}, x\right\rangle\right)}_{\text{convex since linear}} + \underbrace{\mathcal{E}(x) - \frac{\Lambda}{2}\|x\|_2^2}_{\text{convex by A4}}
\end{aligned}$$

is convex by being the sum of two convex functions. By strong convexity of $\widetilde{\mathcal{E}}_k$, $\widetilde{x}_k^{\mathrm{CH}}$ exists, is unique and for all $\xi \in [0,1]$ it holds

$$\begin{aligned}\frac{1}{2}\left(\frac{1+\Lambda\tau}{\tau}\right)\xi(1-\xi)\left\|x-\widetilde{x}_k^{\mathrm{CH}}\right\|_2^2 &\leq \xi\widetilde{\mathcal{E}}_k(x)+(1-\xi)\widetilde{\mathcal{E}}_k(\widetilde{x}_k^{\mathrm{CH}})-\widetilde{\mathcal{E}}_k(\xi x+(1-\xi)\widetilde{x}_k^{\mathrm{CH}})\\ &\leq \xi\left(\widetilde{\mathcal{E}}_k(x)-\widetilde{\mathcal{E}}_k(\widetilde{x}_k^{\mathrm{CH}})\right),\end{aligned}$$

where we used in the last inequality that $\widetilde{x}_k^{\mathrm{CH}}$ minimizes $\widetilde{\mathcal{E}}_k$. Dividing both sides by ξ, letting $\xi \to 0$ and reordering the inequality gives the result. □

Remark F.4. In the case that $\mathcal{E}$ is Λ-convex with $\Lambda < 0$ (i.e., potentially nonconvex), Lemma F.3 requires that the parameter τ is sufficiently small, in order to ensure that $\widetilde{\mathcal{E}}_k$ is strongly convex and therefore has a unique global minimizer $\widetilde{x}_k^{\mathrm{CH}}$. On the other hand, if $\mathcal{E}$ is convex, i.e., $\Lambda \geq 0$, $\widetilde{\mathcal{E}}_k$ is strongly convex and therefore such constraint is not necessary, i.e., τ can be chosen arbitrarily.

Next, we give technical estimates on the quantities $(\widetilde{\mathcal{E}}_k)_r$, $\nu_k\big(B_r(\widetilde{x}_k^{\mathrm{CH}})\big)$ and $\int \left\|x-\widetilde{x}_k^{\mathrm{CH}}\right\|_2 d\nu_k(x)$, which appear when applying Proposition F.2 in the setting of the function $\widetilde{\mathcal{E}}_k$ and the probability measure $\nu_k = \mathcal{N}\big(x_{k-1}^{\mathrm{CH}}, 2\widetilde{\sigma}^2\mathrm{Id}\big)$. This allows to keep the proof of Proposition C.2 more concise.

Lemma F.5. *Let $\widetilde{\mathcal{E}}_k \in \mathcal{C}(\mathbb{R}^d)$ be as defined above with $\mathcal{E} \in \mathcal{C}(\mathbb{R}^d)$ satisfying A2. Then for the expressions $(\widetilde{\mathcal{E}}_k)_r$, $\nu_k\big(B_r(\widetilde{x}_k^{\mathrm{CH}})\big)$ and $\int \left\|x-\widetilde{x}_k^{\mathrm{CH}}\right\|_2 d\nu_k(x)$ appearing in Equation (59) the following estimates hold. Namely,*

$$\begin{aligned}(\widetilde{\mathcal{E}}_k)_r &\leq \left(\frac{1}{2\tau}\Big(r+4\tau C_1\left(\left\|x_{k-1}^{\mathrm{CH}}\right\|_2+\left\|\widetilde{x}_k^{\mathrm{CH}}\right\|_2\right)\Big)+C_1\left(1+r+2\left\|\widetilde{x}_k^{\mathrm{CH}}\right\|_2\right)\right)r,\\ \nu_k\big(B_r(\widetilde{x}_k^{\mathrm{CH}})\big) &\geq \frac{1}{(2\widetilde{\sigma})^d}\exp\left(-\frac{1}{2\widetilde{\sigma}^2}\left(r^2+12\tau^2C_1^2\Big(1+\left\|x_{k-1}^{\mathrm{CH}}\right\|_2^2+\left\|\widetilde{x}_k^{\mathrm{CH}}\right\|_2^2\Big)\right)\right)\frac{1}{\Gamma\left(\frac{d}{2}+1\right)}r^d,\\ \int \left\|x-\widetilde{x}_k^{\mathrm{CH}}\right\|_2 d\nu_k(x) &\leq 2\tau C_1\left(1+\left\|x_{k-1}^{\mathrm{CH}}\right\|_2+\left\|\widetilde{x}_k^{\mathrm{CH}}\right\|_2\right)+\sqrt{2d}\widetilde{\sigma}.\end{aligned}$$

Proof. Let us start by investigating the expressions $(\widetilde{\mathcal{E}}_k)_r$, $\nu_k\big(B_r(\widetilde{x}_k^{\mathrm{CH}})\big)$ and $\int \left\|x-\widetilde{x}_k^{\mathrm{CH}}\right\|_2 d\nu_k(x)$ individually.

Term $(\widetilde{\mathcal{E}}_k)_r$: By definition (see Proposition F.2) and under A2 it holds

$$\begin{aligned}(\widetilde{\mathcal{E}}_k)_r &= \sup_{x\in B_r(\widetilde{x}_k^{\mathrm{CH}})}\widetilde{\mathcal{E}}_k(x)-\widetilde{\mathcal{E}}_k(\widetilde{x}_k^{\mathrm{CH}})\\ &\leq \frac{1}{2\tau}\sup_{x\in B_r(\widetilde{x}_k^{\mathrm{CH}})}\left(\left\|x_{k-1}^{\mathrm{CH}}-x\right\|_2^2-\left\|x_{k-1}^{\mathrm{CH}}-\widetilde{x}_k^{\mathrm{CH}}\right\|_2^2\right)+\sup_{x\in B_r(\widetilde{x}_k^{\mathrm{CH}})}\mathcal{E}(x)-\mathcal{E}(\widetilde{x}_k^{\mathrm{CH}})\\ &\leq \frac{1}{2\tau}\left(r+2\left\|x_{k-1}^{\mathrm{CH}}-\widetilde{x}_k^{\mathrm{CH}}\right\|_2\right)r+C_1\left(1+r+2\left\|\widetilde{x}_k^{\mathrm{CH}}\right\|_2\right)r\\ &\leq \left(\frac{1}{2\tau}\left(r+2\left\|x_{k-1}^{\mathrm{CH}}-\widetilde{x}_k^{\mathrm{CH}}\right\|_2\right)+C_1\left(1+r+2\left\|\widetilde{x}_k^{\mathrm{CH}}\right\|_2\right)\right)r.\end{aligned}$$

Term $\nu_k\big(B_r(\widetilde{x}_k^{\mathrm{CH}})\big)$: Using the density of the multivariate normal distribution $\nu_k = \mathcal{N}\big(x_{k-1}^{\mathrm{CH}}, 2\widetilde{\sigma}^2\mathrm{Id}\big)$ we can directly compute

$$\begin{aligned}\nu_k\big(B_r(\widetilde{x}_k^{\mathrm{CH}})\big) &= \frac{1}{(4\pi\widetilde{\sigma}^2)^{d/2}}\int_{B_r(\widetilde{x}_k^{\mathrm{CH}})}\exp\left(-\frac{1}{4\widetilde{\sigma}^2}\left\|x-x_{k-1}^{\mathrm{CH}}\right\|_2^2\right)d\lambda(x)\\ &\geq \frac{1}{(4\pi\widetilde{\sigma}^2)^{d/2}}\int_{B_r(\widetilde{x}_k^{\mathrm{CH}})}\exp\left(-\frac{1}{2\widetilde{\sigma}^2}\left(\left\|x-\widetilde{x}_k^{\mathrm{CH}}\right\|_2^2+\left\|\widetilde{x}_k^{\mathrm{CH}}-x_{k-1}^{\mathrm{CH}}\right\|_2^2\right)\right)d\lambda(x)\\ &\geq \frac{1}{(4\pi\widetilde{\sigma}^2)^{d/2}}\exp\left(-\frac{1}{2\widetilde{\sigma}^2}\left(r^2+\left\|\widetilde{x}_k^{\mathrm{CH}}-x_{k-1}^{\mathrm{CH}}\right\|_2^2\right)\right)\int_{B_r(\widetilde{x}_k^{\mathrm{CH}})}d\lambda(x)\\ &= \frac{1}{(2\widetilde{\sigma})^d}\exp\left(-\frac{1}{2\widetilde{\sigma}^2}\left(r^2+\left\|\widetilde{x}_k^{\mathrm{CH}}-x_{k-1}^{\mathrm{CH}}\right\|_2^2\right)\right)\frac{1}{\Gamma\left(\frac{d}{2}+1\right)}r^d,\end{aligned}$$

where we used in the last step that the volume of a d-dimensional unit ball is $\pi^{d/2}/\Gamma\left(\frac{d}{2}+1\right)$. Here, Γ denotes Euler's gamma function. We recall for the readers' convenience that by Stirling's approximation $\Gamma(x+1) \sim \sqrt{2\pi x}\,(x/e)^x$ as $x \to \infty$.

Term $\int \left\|x - \widetilde{x}_k^{\mathrm{CH}}\right\|_2 d\nu_k(x)$: A straightforward computation gives

$$\begin{aligned}\int \left\|x - \widetilde{x}_k^{\mathrm{CH}}\right\|_2 d\nu_k(x) &= \int \left\|x - \widetilde{x}_k^{\mathrm{CH}}\right\|_2 d\mathcal{N}\left(x_{k-1}^{\mathrm{CH}}, 2\widetilde{\sigma}^2\mathrm{Id}\right)(x)\\ &= \int \left\|x + x_{k-1}^{\mathrm{CH}} - \widetilde{x}_k^{\mathrm{CH}}\right\|_2 d\mathcal{N}\left(0, 2\widetilde{\sigma}^2\mathrm{Id}\right)(x)\\ &\leq \left\|x_{k-1}^{\mathrm{CH}} - \widetilde{x}_k^{\mathrm{CH}}\right\|_2 + \int \|x\|_2 \, d\mathcal{N}\left(0, 2\widetilde{\sigma}^2\mathrm{Id}\right)(x)\\ &\leq \left\|x_{k-1}^{\mathrm{CH}} - \widetilde{x}_k^{\mathrm{CH}}\right\|_2 + \sqrt{2d}\widetilde{\sigma}.\end{aligned}$$

Concluding step: To conclude the proof, we further observe that since $\widetilde{x}_k^{\mathrm{CH}}$ is the minimizer of $\widetilde{\mathcal{E}}_k$, see (11), a comparison with x_{k-1}^{CH} yields

$$\frac{1}{2\tau}\left\|x_{k-1}^{\mathrm{CH}} - \widetilde{x}_k^{\mathrm{CH}}\right\|_2^2 + \mathcal{E}(\widetilde{x}_k^{\mathrm{CH}}) \leq \mathcal{E}(x_{k-1}^{\mathrm{CH}}).$$

With A2 it therefore holds

$$\left\|x_{k-1}^{\mathrm{CH}} - \widetilde{x}_k^{\mathrm{CH}}\right\|_2^2 \leq 2\tau\left(\mathcal{E}(x_{k-1}^{\mathrm{CH}}) - \mathcal{E}(\widetilde{x}_k^{\mathrm{CH}})\right) \leq 2\tau C_1\left(1 + \left\|x_{k-1}^{\mathrm{CH}}\right\|_2 + \left\|\widetilde{x}_k^{\mathrm{CH}}\right\|_2\right)\left\|x_{k-1}^{\mathrm{CH}} - \widetilde{x}_k^{\mathrm{CH}}\right\|_2,$$

or rephrased

$$\left\|x_{k-1}^{\mathrm{CH}} - \widetilde{x}_k^{\mathrm{CH}}\right\|_2 \leq 2\tau C_1\left(1 + \left\|x_{k-1}^{\mathrm{CH}}\right\|_2 + \left\|\widetilde{x}_k^{\mathrm{CH}}\right\|_2\right).$$

Exploiting this estimate in the former bounds, gives the statements. □

F.3. Proof of Proposition C.2

We now have all necessary tools at hand to present the detailed proof of Proposition C.2.

Proof of Proposition C.2. By using the quantitative Laplace principle F.2, we make rigorous and quantify the fact that x_k^{CH} approximates the minimizer of $\widetilde{\mathcal{E}}_k$, denoted by $\widetilde{x}_k$, for sufficiently large α.

To obtain the probabilistic formulation of the statement, let us again denote the underlying probability space by $(\Omega, \mathcal{F}, \mathbb{P})$ (note that we can use the same probability space as in Section E since the stochasticity of both schemes (10) and (11) is solely coming from the initialization) and introduce the subset $\widetilde{\Omega}_M$ of Ω of suitably bounded random variables according to

$$\widetilde{\Omega}_M := \left\{\omega \in \Omega : \max_{k=0,\dots,K} \max\left\{\left\|x_k^{\mathrm{CH}}\right\|_2, \left\|\widetilde{x}_k^{\mathrm{CH}}\right\|_2\right\} \leq M\right\}.$$

For the associated cutoff function (random variable) we write $\mathbb{1}_{\widetilde{\Omega}_M}$.

We first notice that by definition of the consensus point $x_\alpha^{\mathcal{E}}$ in (3) it holds

$$\begin{aligned}x_k^{\mathrm{CH}} = x_\alpha^{\mathcal{E}}(\mu_k) &= \int x \frac{\exp(-\alpha\mathcal{E}(x))}{\|\exp(-\alpha\mathcal{E})\|_{L^1(\mu_k)}} d\mu_k(x)\\ &= \int x \frac{\exp(-\alpha\mathcal{E}(x))\exp\left(-\frac{1}{4\widetilde{\sigma}^2}\left\|x - x_{k-1}^{\mathrm{CH}}\right\|_2^2\right)}{\int \exp(-\alpha\mathcal{E}(x'))\exp\left(-\frac{1}{4\widetilde{\sigma}^2}\left\|x' - x_{k-1}^{\mathrm{CH}}\right\|_2^2\right) d\nu_k(x')} d\nu_k(x)\\ &= \int x \frac{\exp(-\alpha\widetilde{\mathcal{E}}_k(x))}{\|\exp(-\alpha\widetilde{\mathcal{E}}_k)\|_{L^1(\nu_k)}} d\nu_k(x)\\ &= x_\alpha^{\widetilde{\mathcal{E}}_k}(\nu_k),\end{aligned} \tag{62}$$

which introduces the relation $\tau = 2\alpha\widetilde{\sigma}^2$ and where we chose $\nu_k = \mathcal{N}\big(x_{k-1}^{\mathrm{CH}}, 2\widetilde{\sigma}^2\mathrm{Id}\big)$, which is globally supported, i.e., $\mathrm{supp}(\nu_k) = \mathbb{R}^d$. Since, according to Lemma F.3, $\widetilde{\mathcal{E}}_k$ satisfies the inverse continuity property (61) with $\nu = 1/2$ and $\eta = \sqrt{\frac{1}{2\tau} + \frac{\Lambda}{2}} > 0$, the quantitative Laplace principle, Proposition F.2, gives for any $r, q > 0$ the bound

$$\big\|x_k^{\mathrm{CH}} - \widetilde{x}_k^{\mathrm{CH}}\big\|_2 = \big\|x_\alpha^{\widetilde{\mathcal{E}}_k}(\nu_k) - \widetilde{x}_k^{\mathrm{CH}}\big\|_2 \leq \frac{\big(q + (\widetilde{\mathcal{E}}_k)_r\big)^\nu}{\eta} + \frac{\exp(-\alpha q)}{\nu_k\big(B_r(\widetilde{x}_k^{\mathrm{CH}})\big)} \int \big\|x - \widetilde{x}_k^{\mathrm{CH}}\big\|_2 \, d\nu_k(x), \tag{63}$$

where $(\widetilde{\mathcal{E}}_k)_r := \sup_{x \in B_r(\widetilde{x}_k^{\mathrm{CH}})} \widetilde{\mathcal{E}}_k(x) - \widetilde{\mathcal{E}}_k(\widetilde{x}_k^{\mathrm{CH}})$. We further notice that by the assumption $\tau < 1/(-2\Lambda)$ if $\Lambda < 0$ it holds $\eta \geq 1/(2\sqrt{\tau})$ (in the case $\Lambda \geq 0$ the same bound holds trivially). Combining (63) with the technical estimates of Lemma F.5 and the definition of the cutoff function $\mathbb{1}_{\widetilde{\Omega}_M}$ allows to obtain

$$\begin{aligned}
&\mathbb{E}\big\|x_k^{\mathrm{CH}} - \widetilde{x}_k^{\mathrm{CH}}\big\|_2^2 \mathbb{1}_{\widetilde{\Omega}_M} \\
&\quad\leq 2\mathbb{E}\left[\frac{\big(q + (\widetilde{\mathcal{E}}_k)_r\big)}{\eta^2}\mathbb{1}_{\widetilde{\Omega}_M}\right] + 2\mathbb{E}\left[\frac{\exp(-2\alpha q)}{\nu_k\big(B_r(\widetilde{x}_k^{\mathrm{CH}})\big)^2}\left(\int \big\|x - \widetilde{x}_k^{\mathrm{CH}}\big\|_2 \, d\nu_k(x)\right)^2 \mathbb{1}_{\widetilde{\Omega}_M}\right] \\
&\quad\leq 8\tau\Big(q + \big(\tfrac{r}{2\tau} + C_1 + C_1 r + 6C_1 M\big)\, r\Big) \\
&\qquad + 4\exp\left(-2\alpha q + \frac{1}{\widetilde{\sigma}^2}\left(r^2 + 12\tau^2 C_1^2(1 + 2M^2)\right)\right) \frac{2^d(2\widetilde{\sigma}^2)^d}{r^{2d}}\Gamma\big(\tfrac{d}{2} + 1\big)^2 \left(4\tau^2 C_1^2(1 + 2M)^2 + 2d\widetilde{\sigma}^2\right) \\
&\quad= 8\tau\Big(q + \big(\tfrac{r}{2\tau} + C_1 + C_1 r + 6C_1 M\big)\, r\Big) \\
&\qquad + 4\exp\left(-2\alpha\left(q - \left(\frac{r^2}{\tau} + 12\tau C_1^2(1 + 2M^2)\right)\right)\right) \frac{2^d\tau^d}{\alpha^d r^{2d}}\Gamma\big(\tfrac{d}{2} + 1\big)^2 \left(4\tau^2 C_1^2(1 + 2M)^2 + d\frac{\tau}{\alpha}\right),
\end{aligned} \tag{64}$$

where in the last step we just replaced $2\widetilde{\sigma}^2$ by τ/α according to the relation. We now choose

$$r = \tau, \quad q = \frac{3}{2}\tau + 12\tau C_1^2(1 + 2M^2) \quad \text{and} \quad \alpha \geq \alpha_0 := \frac{1}{\tau}\Big(d\log 2 + \log(1 + d) + 2\log\Gamma\big(\tfrac{d}{2} + 1\big)\Big),$$

where Γ denotes Euler's gamma function, for which, by Stirling's approximation, it holds $\Gamma(x + 1) \sim \sqrt{2\pi x}\,(x/e)^x$ as $x \to \infty$. With this we can continue the computations of (64) with

$$\begin{aligned}
\mathbb{E}\big\|x_k^{\mathrm{CH}} - \widetilde{x}_k^{\mathrm{CH}}\big\|_2^2 \mathbb{1}_{\widetilde{\Omega}_M} &\leq 8\Big(2 + C_1 + C_1\tau + 6C_1 M + 12C_1^2(1 + 2M^2)\Big)\tau^2 \\
&\quad + 4\exp(-\alpha\tau)\frac{2^d}{\alpha^d\tau^d}\Gamma\big(\tfrac{d}{2} + 1\big)^2\left(4\tau^2 C_1^2(1 + 2M^2) + d\frac{\tau}{\alpha}\right) \\
&\leq 8\Big(3 + C_1 + C_1\tau + 6C_1 M + 14C_1^2(1 + M^2)\Big)\tau^2 \\
&\leq c\tau^2
\end{aligned} \tag{65}$$

with a constant $c = c(C_1, M)$. Notice that to obtain the next-to-last inequality one may first note and exploit that one has $\alpha\tau \geq 1$ as well as $1/\alpha \leq \tau$ as a consequence of $\alpha \geq 1/\tau$.

Probabilistic formulation: We first note that with Markov's inequality we have the estimate

$$\begin{aligned}
\mathbb{P}\big(\widetilde{\Omega}_M^c\big) &= \mathbb{P}\left(\max_{k=0,\dots,K} \max\left\{\big\|x_k^{\mathrm{CH}}\big\|_2, \big\|\widetilde{x}_k^{\mathrm{CH}}\big\|_2\right\} > M\right) \\
&\leq \frac{1}{M^4}\left(\mathbb{E}\max_{k=0,\dots,K}\big\|x_k^{\mathrm{CH}}\big\|_2^4 + \mathbb{E}\max_{k=0,\dots,K}\big\|\widetilde{x}_k^{\mathrm{CH}}\big\|_2^4\right) \\
&\leq \frac{1}{M^4}\big(\mathcal{M}^{\mathrm{CH}} + \widetilde{\mathcal{M}}^{\mathrm{CH}}\big),
\end{aligned}$$

where the last inequality is due to Lemmas D.3 and D.6. Thus, for any $\delta \in (0, 1/2)$, a sufficiently large choice $M = M(\delta^{-1}, \mathcal{M}^{\mathrm{CH}}, \widetilde{\mathcal{M}}^{\mathrm{CH}})$ allows to ensure $\mathbb{P}\big(\widetilde{\Omega}_M^c\big) \leq \delta$. To conclude the proof, let us denote by $\widetilde{K}_\varepsilon \subset \Omega$ the set, where (14) does not hold and abbreviate

$$\epsilon = \varepsilon^{-1} c\tau^2.$$

For the probability of this set we can estimate

$$\begin{aligned}\mathbb{P}\big(\widetilde{K}_\varepsilon\big) &= \mathbb{P}\big(\widetilde{K}_\varepsilon \cap \widetilde{\Omega}_M\big) + \mathbb{P}\big(\widetilde{K}_\varepsilon \cap \widetilde{\Omega}_M^c\big) \leq \mathbb{P}\big(\widetilde{K}_\varepsilon \,\big|\, \widetilde{\Omega}_M\big)\,\mathbb{P}\big(\widetilde{\Omega}_M\big) + \mathbb{P}\big(\widetilde{\Omega}_M^c\big) \\ &\leq \mathbb{P}\big(\widetilde{K}_\varepsilon \,\big|\, \widetilde{\Omega}_M\big) + \delta \leq \epsilon^{-1}\,\mathbb{E}\left[\left\|x_k^{\mathrm{CH}} - \widetilde{x}_k^{\mathrm{CH}}\right\|_2^2 \,\middle|\, \widetilde{\Omega}_M\right] + \delta,\end{aligned} \tag{66}$$

where the last step is due to Markov's inequality. By definition of the conditional expectation we further have

$$\mathbb{E}\left[\left\|x_k^{\mathrm{CH}} - \widetilde{x}_k^{\mathrm{CH}}\right\|_2^2 \,\middle|\, \widetilde{\Omega}_M\right] \leq \frac{1}{\mathbb{P}\big(\widetilde{\Omega}_M\big)}\mathbb{E}\left\|x_k^{\mathrm{CH}} - \widetilde{x}_k^{\mathrm{CH}}\right\|_2^2 \mathbb{1}_{\widetilde{\Omega}_M} \leq 2\mathbb{E}\left\|x_k^{\mathrm{CH}} - \widetilde{x}_k^{\mathrm{CH}}\right\|_2^2 \mathbb{1}_{\widetilde{\Omega}_M}.$$

Inserting now the expression from (65) concludes the proof. □

F.4. Proof of Theorem C.3

We now have all necessary tools at hand to present the detailed proof of Theorem C.3.

Proof of Theorem C.3. We combine in what follows Proposition C.2 with a stability argument for the MMS (12).

To obtain the probabilistic formulation of the statement, let us denote, as in the proof of Proposition C.2, the underlying probability space by $(\Omega, \mathcal{F}, \mathbb{P})$ (note that we can use the same probability space as in Section E since the stochasticity of the three schemes (10), (11) and (12) is solely coming from the initialization) and introduce the subset $\widetilde{\Omega}_M$ of Ω of suitably bounded random variables according to

$$\widetilde{\Omega}_M := \left\{\omega \in \Omega : \max_{k=0,\dots,K} \max\left\{\left\|x_k^{\mathrm{CH}}\right\|_2, \left\|\widetilde{x}_k^{\mathrm{CH}}\right\|_2\right\} \leq M\right\}.$$

For the associated cutoff function (random variable) we write $\mathbb{1}_{\widetilde{\Omega}_M}$.

We can decompose the expected squared discrepancy $\mathbb{E}\left\|x_k^{\mathrm{MMS}} - x_k^{\mathrm{CH}}\right\|_2^2 \mathbb{1}_{\widetilde{\Omega}_M}$ between the MMS (12) and the CH scheme (10) for any $\vartheta \in (0,1)$ as

$$\mathbb{E}\left\|x_k^{\mathrm{MMS}} - x_k^{\mathrm{CH}}\right\|_2^2 \mathbb{1}_{\widetilde{\Omega}_M} \leq (1+\vartheta)\,\mathbb{E}\left\|x_k^{\mathrm{MMS}} - \widetilde{x}_k^{\mathrm{CH}}\right\|_2^2 \mathbb{1}_{\widetilde{\Omega}_M} + (1+\vartheta^{-1})\,\mathbb{E}\left\|\widetilde{x}_k^{\mathrm{CH}} - x_k^{\mathrm{CH}}\right\|_2^2 \mathbb{1}_{\widetilde{\Omega}_M}. \tag{67}$$

In what follows we individually estimate the two terms on the right-hand side of (67).

First term: Let us first bound the term $\mathbb{E}\left\|x_k^{\mathrm{MMS}} - \widetilde{x}_k^{\mathrm{CH}}\right\|_2^2 \mathbb{1}_{\widetilde{\Omega}_M}$. By definition of x_k^{MMS} and $\widetilde{x}_k^{\mathrm{CH}}$ as minimizers of (12) and (11), respectively, and with the definition $\mathcal{E}_\Lambda(x) := \mathcal{E}(x) - \frac{\Lambda}{2}\|x\|_2^2$ it holds

$$\frac{(1+\tau\Lambda)x_k^{\mathrm{MMS}} - x_{k-1}^{\mathrm{MMS}}}{\tau} \in -\partial\mathcal{E}_\Lambda(x_k^{\mathrm{MMS}}) \quad \text{and} \quad \frac{(1+\tau\Lambda)\widetilde{x}_k^{\mathrm{CH}} - x_{k-1}^{\mathrm{CH}}}{\tau} \in -\partial\mathcal{E}_\Lambda(\widetilde{x}_k^{\mathrm{CH}}).$$

Since $\mathcal{E}_\Lambda$ is convex due to A4 and as consequence of the properties of the subdifferential we have

$$\left\langle -\frac{(1+\tau\Lambda)x_k^{\mathrm{MMS}} - x_{k-1}^{\mathrm{MMS}}}{\tau} + \frac{(1+\tau\Lambda)\widetilde{x}_k^{\mathrm{CH}} - x_{k-1}^{\mathrm{CH}}}{\tau}, x_k^{\mathrm{MMS}} - \widetilde{x}_k^{\mathrm{CH}} \right\rangle \geq 0,$$

which allows to obtain by means of Cauchy-Schwarz inequality

$$(1+\tau\Lambda)\left\|x_k^{\mathrm{MMS}} - \widetilde{x}_k^{\mathrm{CH}}\right\|_2^2 \leq \left\langle x_{k-1}^{\mathrm{MMS}} - x_{k-1}^{\mathrm{CH}}, x_k^{\mathrm{MMS}} - \widetilde{x}_k^{\mathrm{CH}}\right\rangle \leq \left\|x_{k-1}^{\mathrm{MMS}} - x_{k-1}^{\mathrm{CH}}\right\|_2 \left\|x_k^{\mathrm{MMS}} - \widetilde{x}_k^{\mathrm{CH}}\right\|_2$$

or, equivalently,

$$\left\|x_k^{\mathrm{MMS}} - \widetilde{x}_k^{\mathrm{CH}}\right\|_2 \leq \frac{1}{1+\tau\Lambda}\left\|x_{k-1}^{\mathrm{MMS}} - x_{k-1}^{\mathrm{CH}}\right\|_2. \tag{68}$$

Second term: For the term $\mathbb{E}\left\|\widetilde{x}_k^{\mathrm{CH}} - x_k^{\mathrm{CH}}\right\|_2^2 \mathbb{1}_{\widetilde{\Omega}_M}$ we obtained in (65) in the proof of Proposition C.2, for suitable choices of $\widetilde{\sigma}$ and α, the bound

$$\mathbb{E}\left\|x_k^{\mathrm{CH}} - \widetilde{x}_k^{\mathrm{CH}}\right\|_2^2 \mathbb{1}_{\widetilde{\Omega}_M} \leq c\tau^2 \tag{69}$$

with a constant $c = c(C_1, M)$.

Concluding step: Combining this with the estimate (68) yields for (67) the bound

$$\mathbb{E}\left\|x_k^{\mathrm{MMS}} - x_k^{\mathrm{CH}}\right\|_2^2 \mathbb{1}_{\widetilde{\Omega}_M} \leq \frac{1+\vartheta}{(1+\tau\Lambda)^2}\mathbb{E}\left\|x_{k-1}^{\mathrm{MMS}} - x_{k-1}^{\mathrm{CH}}\right\|_2^2 \mathbb{1}_{\widetilde{\Omega}_M} + c(1+\vartheta^{-1})\,\tau^2. \tag{70}$$

An application of the discrete variant of Grönwall's inequality (9) shows that

$$\mathbb{E}\left\|x_k^{\mathrm{MMS}} - x_k^{\mathrm{CH}}\right\|_2^2 \mathbb{1}_{\widetilde{\Omega}_M} \leq c(1+\vartheta^{-1})\,\tau^2 \sum_{\ell=0}^{k-1}\left(\frac{1+\vartheta}{(1+\tau\Lambda)^2}\right)^{\ell} \tag{71}$$

for all $k = 1, \dots, K$, where we used that both schemes are initialized by the same x_0.

Probabilistic formulation: We first note that with Markov's inequality we have the estimate

$$\begin{aligned}
\mathbb{P}\big(\widetilde{\Omega}_M^c\big) &= \mathbb{P}\left(\max_{k=0,\dots,K} \max\left\{\left\|x_k^{\mathrm{CH}}\right\|_2, \left\|\widetilde{x}_k^{\mathrm{CH}}\right\|_2\right\} > M\right) \\
&\leq \frac{1}{M^4}\left(\mathbb{E}\max_{k=0,\dots,K}\left\|x_k^{\mathrm{CH}}\right\|_2^4 + \mathbb{E}\max_{k=0,\dots,K}\left\|\widetilde{x}_k^{\mathrm{CH}}\right\|_2^4\right) \\
&\leq \frac{1}{M^4}\big(\mathcal{M}^{\mathrm{CH}} + \widetilde{\mathcal{M}}^{\mathrm{CH}}\big),
\end{aligned}$$

where the last inequality is due to Lemmas D.3 and D.6. Thus, for any $\delta \in (0, 1/2)$, a sufficiently large choice $M = M(\delta^{-1}, \mathcal{M}^{\mathrm{CH}}, \widetilde{\mathcal{M}}^{\mathrm{CH}})$ allows to ensure $\mathbb{P}\big(\widetilde{\Omega}_M^c\big) \leq \delta$. To conclude the proof, let us denote by $\widetilde{K}_\varepsilon \subset \Omega$ the set, where (15) does not hold and abbreviate

$$\epsilon = \varepsilon^{-1} c(1+\vartheta^{-1})\,\tau^2 \sum_{\ell=0}^{k-1}\left(\frac{1+\vartheta}{(1+\tau\Lambda)^2}\right)^{\ell}.$$

For the probability of this set we can estimate

$$\begin{aligned}
\mathbb{P}\big(\widetilde{K}_\varepsilon\big) &= \mathbb{P}\big(\widetilde{K}_\varepsilon \cap \widetilde{\Omega}_M\big) + \mathbb{P}\big(\widetilde{K}_\varepsilon \cap \widetilde{\Omega}_M^c\big) \leq \mathbb{P}\big(\widetilde{K}_\varepsilon \,\big|\, \widetilde{\Omega}_M\big)\,\mathbb{P}\big(\widetilde{\Omega}_M\big) + \mathbb{P}\big(\widetilde{\Omega}_M^c\big) \\
&\leq \mathbb{P}\big(\widetilde{K}_\varepsilon \,\big|\, \widetilde{\Omega}_M\big) + \delta \leq \epsilon^{-1}\,\mathbb{E}\left[\left\|x_k^{\mathrm{MMS}} - x_k^{\mathrm{CH}}\right\|_2^2 \,\middle|\, \widetilde{\Omega}_M\right] + \delta,
\end{aligned} \tag{72}$$

where the last step is due to Markov's inequality. By definition of the conditional expectation we further have

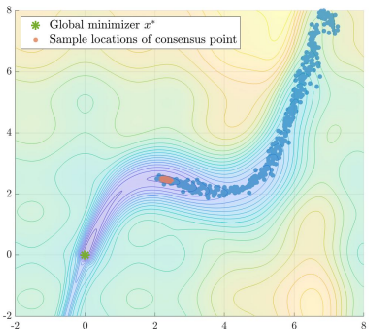
$$\mathbb{E}\left[\left\|x_k^{\mathrm{MMS}} - x_k^{\mathrm{CH}}\right\|_2^2 \,\middle|\, \widetilde{\Omega}_M\right] \leq \frac{1}{\mathbb{P}\big(\widetilde{\Omega}_M\big)}\mathbb{E}\left\|x_k^{\mathrm{MMS}} - x_k^{\mathrm{CH}}\right\|_2^2 \mathbb{1}_{\widetilde{\Omega}_M} \leq 2\mathbb{E}\left\|x_k^{\mathrm{MMS}} - x_k^{\mathrm{CH}}\right\|_2^2 \mathbb{1}_{\widetilde{\Omega}_M}.$$

Inserting now the expression from (71) concludes the proof. □

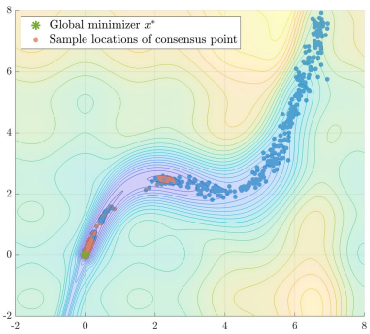
G. Additional numerical experiments

G.1. Comparison of the CH scheme (10) for different sampling widths $\widetilde{\sigma}$

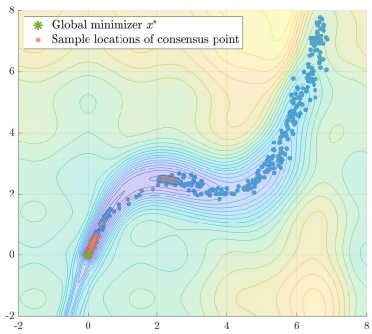
To complement Figure C.1a, we visualize in Figure G.1 the influence of the sampling width $\widetilde{\sigma}$ on the behavior of the CH scheme (10).



(a) The CH scheme (10) with sampling width $\widetilde{\sigma} = 0.4$ gets stuck in a local minimum of $\mathcal{E}$.



(b) The CH scheme (10) with sampling width $\widetilde{\sigma} = 0.6$ can occasionally escape local minima of $\mathcal{E}$.

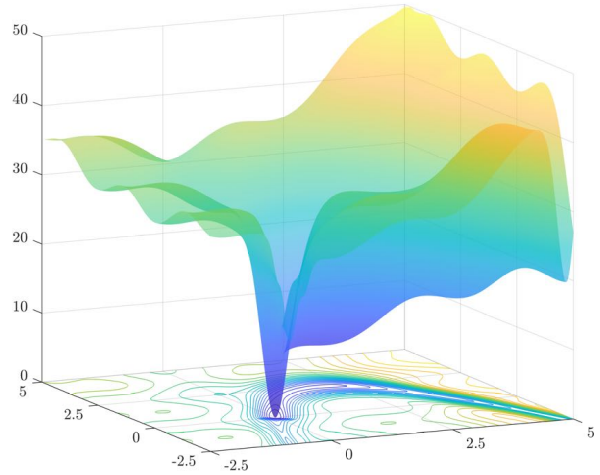


(c) The CH scheme (10) with sampling width $\widetilde{\sigma} = 0.7$ can escape local minima of $\mathcal{E}$.

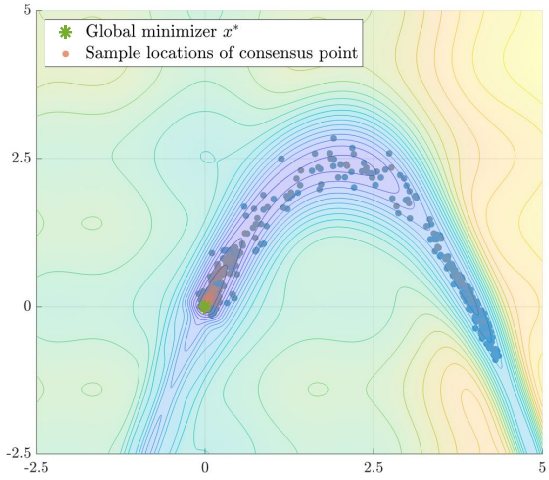
Figure G.1: A visual comparison of the CH scheme (10) for different sampling widths $\widetilde{\sigma}$. We depict the positions of the consensus hopping scheme (10) for different values of $\widetilde{\sigma}$ (0.4 in (a), 0.6 in (b) and 0.7 in (c)) in the setting of Figure C.1a. While for small $\widetilde{\sigma}$ the numerical scheme gets stuck in a local minimum of the objective, the ability to escape such critical points improves with larger $\widetilde{\sigma}$. Notice that (b) coincides with Figure C.1a.

G.2. The numerical experiments of Figures 1 and C.1 for a different objective

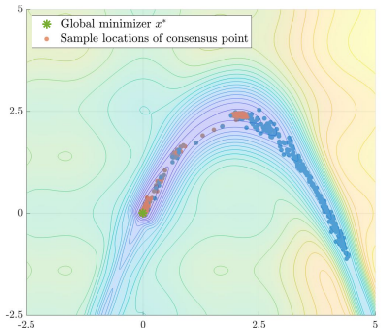
In the style of Figures 1 and C.1 we provide in Figure G.2 an additional set of illustrations of the behavior of the different algorithms analyzed in this work for a noisy Canyon function with a valley shaped as a second degree polynomial.



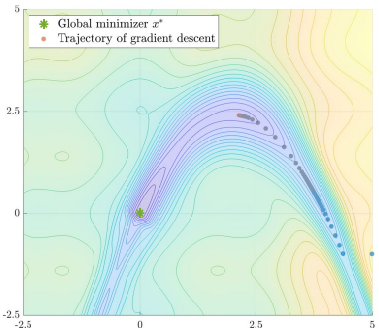
(a) A noisy Canyon function $\mathcal{E}$ with a valley shaped as a second degree polynomial



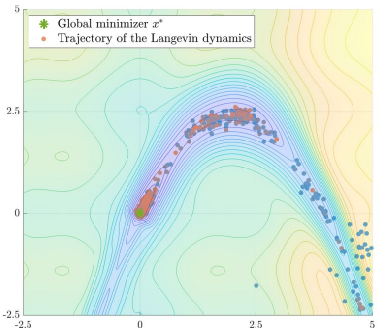
(b) The CBO scheme (4) (sampled over several runs) follows on average the valley while passing over local minima.



(c) The CH scheme (10) (sampled over several runs) follows on average the valley of $\mathcal{E}$ and can occasionally escape local minima.



(d) GD gets stuck in a local minimum of $\mathcal{E}$.



(e) The Langevin dynamics (6) (sampled over several runs) follows on average the valley of $\mathcal{E}$ and escapes local minima.

Figure G.2: An additional numerical experiment illustrating the behavior of the CBO scheme (4) (see (b)), the consensus hopping scheme (10) (see (c)), GD (see (d)) and the overdamped Langevin dynamics (6) (see (e)) in search of the global minimizer x^* of the nonconvex objective function $\mathcal{E}$ depicted in (a). The experimental setting is the one of Figures 1 and C.1 with the only difference of the particles being initialized around $(5, -1)$.